Wirkungsorientierte Evaluation im Rahmen der Armuts- und Reichtumsberichterstattung: Perspektivstudie

Beywl, Wolfgang; Speer, Sandra; Kehr, Jochen

Forschungsbericht / research report

Wirkungsorientierte Evaluation

im Rahmen der

Armuts- und Reichtumsberichterstattung

Perspektivstudie

Wolfgang Beywl, Sandra Speer, Jochen Kehr

2004

Auftraggeber:
Bundesministerium für Gesundheit und Soziale Sicherung (BMGS)
Rochusstr. 1
53123 Bonn

Durchführung und Berichterstattung:
Univation – Institut für Evaluation

Hohenstaufenring 63
50674 Köln
Telefon: 0221-4248071 – E-Mail: info@univation.org

Wolfgang Beywl, Sandra Speer, Jochen Kehr

Wirkungsorientierte Evaluation

im Rahmen der Armuts- und Reichtumsberichterstattung

Erstellt unter Mitarbeit von

Susanne Mäder

Melanie Borgmann

Dagmar Schreier

Köln

2004

Wirkungsorientierte Evaluation im Rahmen der Armuts- und Reichtumsberichterstattung

Bonn – BMGS (Hrsg.) 2003

Management Summary

Die durch das Bundesministerium für Gesundheit und Soziales veröffentlichte Studie weist auf: Evaluationen können einen wichtigen Beitrag zur Armuts- und Reichtumsberichterstattung und damit zur zielgeführten Ausgestaltung von Politik und Programmen der sozialen Integration und Vermeidung von Armut leisten.

Der Gesamtbericht umfasst folgende Kapitel: (1) Evaluationstheoretische Grundlegung, (2) Modelle der Evaluation, (3) Anforderungen von Experten/-innen an Evaluationen im Bereich Armut bekämpfender Politik, (4) Datenlage für Evaluationen sowie (5) Perspektiven und Empfehlungen.

Evaluationen und ihre Ergebnisse sind für einen breiten Adressatenkreis bedeutsam: Für Parlament, Regierung und Behörden werden die Grundlagen für Beratung und Entscheidung sowie für die Kontrolle der öffentlichen Mittelverwendung verbessert. Die Öffentlichkeit kann sich ein gesichertes Urteil darüber bilden, in welchem Maße soziale Problemlagen erfolgreich gemindert werden. In der Armutsbekämpfung aktiv Tätige erhalten Hinweise für die Steuerung der von ihnen durchgeführten Programme und Maßnahmen. Der Wissenschaft werden mit empirischen Evaluationen zusätzliche Datenquellen erschlossen.

Um von den Adressaten/-innen genutzt zu werden, müssen Evaluationen regelmäßig, unabhängig und fachkundig durchgeführt werden:

- Gesetze, Politiken und Programme von größerer Bedeutung sollten frühzeitig, ggf. in Form von Pilotvorhaben evaluiert werden.
- Die Ausschreibung von Evaluationen soll Zwecksetzung und Ausgangsfragestellungen klar benennen, dabei im Hinblick auf die Auswahl von Evaluationsmethoden offen gestaltet sein.
- Die „Standards für Evaluation“ der Deutschen Gesellschaft für Evaluation sollten sowohl für die Beauftragung als auch für Planung und Steuerung der Evaluation zugrunde gelegt werden.
- Im von Wertspannungen gekennzeichneten Politikfeld der Armutsvermeidung und sozialen Integration ist Unabhängigkeit der Evaluation von herausragender Bedeutung. Zu sichern sind daher hohe Transparenz über die Grundlagen, Methoden, Datenquellen und das Zustandekommen von Schlussfolgerungen und Empfehlungen.
- Ergebnisse einer Evaluation sollen klar und verständlich aufbereitet sein. Es ist wünschenswert, dass öffentlich finanzierte Studien binnen kurzer Frist einer breiten Öffentlichkeit zugänglich werden.
- Evaluationen können in die Armuts- und Reichtumsberichterstattung verstärkt integriert werden durch leitfadengestützte Projektgestaltung, erweitertes Evaluationsdokumentationswesen, auf Lebenslagen-Dimensionen zugeschnittene Metaanalysen sowie systematische Nutzung von Evaluations-Fachwissen, dabei auch aus politischen Systemen mit einer längeren Evaluationskultur.

Wolfgang Beywl, Sandra Speer, Jochen Kehr

Wirkungsorientierte Evaluation im Rahmen der Armuts- und Reichtumsberichterstattung

Bonn – BMGS (Hrsg.) 2003

Kurzfassung

Die 2003 durch das Bundesministerium für Gesundheit und Soziales veröffentlichte Perspektivstudie weist auf, dass Evaluationen einen wichtigen Beitrag zur Armuts- und Reichtumsberichterstattung und damit zur zielgeführten Ausgestaltung von Politik und Programmen der sozialen Integration und zur Vermeidung von Armut leisten können.

Die Kurzfassung soll Interesse wecken für eine detaillierte Auseinandersetzung mit dem Gesamtbericht* und formuliert abschließend Kernpunkte für die künftige Anlage von wirkungsorientierten Evaluationen im Kontext der Armuts- und Reichtumsberichterstattung.

Der Gesamtbericht gliedert sich in folgende Kapitel: (1) Evaluationstheoretische Grundlegung, (2) Modelle der Evaluation, (3) Anforderungen von Experten/-innen an Evaluationen im Bereich Armut bekämpfender Politik, (4) Datenlage für Evaluationen sowie (5) Perspektiven und Empfehlungen.

Definition von Evaluation – Kap. 1.1

Gegenstand der Perspektivstudie sind Evaluationen, die Programme, Maßnahmen und Politiken im Umfeld der Armuts- und Reichtumsberichterstattung wissenschaftlich fundiert vorbereiten, begleiten oder beurteilen. Bei Evaluationen handelt es sich um systematische Untersuchungen, die empirisch gewonnene Informationen bereitstellen, so dass der Wert eines Gegenstandes nachvollziehbar eingeschätzt werden kann.

Standards für Evaluation – Kap. 1.2 und Anlage 7.2

Diese und andere Anforderungen an Evaluationen werden in den „Standards für Evaluation“ und ihren begleitenden Materialien konkretisiert, die 2001 durch die Deutsche Gesellschaft für Evaluation (DeGEval) verabschiedet wurden. Diese kommentierten und im internationalen Raum bewährten Qualitätsleitlinien richten sich an

---

* Die Kapitelnummern in den Randnoten beziehen sich auf den Gesamtbericht.

Evaluatoren/-innen und auch an Auftraggebende. 25 Einzelstandards sollen sicherstellen, dass Evaluationen nützlich, durchführbar, fair und genau sind. Unter anderem geben sie Hinweise dazu, ob eine Evaluation durchgeführt werden soll, wie das Evaluationsthema eingegrenzt und wie das Evaluationsdesign erstellt werden soll. Auch zu Budgetierung, Vertragsschluss und Personalausstattung wird ein Referenzrahmen angeboten. Die Standards sind auch für Evaluationen im Kontext der Armuts- und Reichtumsberichterstattung geeignet.

„Programm“ als typischer Gegenstand – Kap. 1.3

Politische Maßnahmen, Initiativen, Modellprojekte, Kampagnen u. v. m. werden in der Evaluation als „Programme“ bezeichnet. Programme im Kontext der vorliegenden Perspektivstudie sind Ausdruck staatlichen oder staatlich gerahmten Handelns, das sich auf die Verbesserung von Rahmenbedingungen und die Behebung von Notlagen armer bzw. von Armut bedrohter Menschen richtet oder deren sozialer Ausgrenzung entgegen wirken soll. Programme sollen spezifizierte Ziele erreichen, und zwar mit Hilfe konzeptionell zugeschnittener Interventionen, die bei Einsatz von Ressourcen zu den gewünschten Resultaten führen. Vielfach weisen Gesetze und politische Reformen lediglich Ansätze zielgerichteten und planvollen Vorgehens auf. Die „Neue Steuerung“, verbunden mit einer Reorganisation öffentlicher Verwaltung, dringt aktuell stärker auf ziel- bzw. wirkungsorientierte Politiken. Dies kann zu einer verbesserten Arbeitsgrundlage für effektive und effiziente Programmevaluationen beitragen.

Resultate und Wirkungen im Fokus – Kap. 1.4

Gemessen und bewertet werden sollen zunehmend die Resultate der Programme, d.h. die bereit gestellten Leistungen (Outputs), die bei den Zielgruppen intendierten Veränderungen/Stabilisierungen (Outcomes/Outcome-Wirkungen) und die Wirkungen auf die sozialen Systeme (Impacts).

„Wirkungen“ liegen dann vor, wenn die gemessenen Resultate auf das Programm zurückgeführt sind, also nachvollziehbare theoretische Annahmen sowie empirische Daten über die Verbindung von Maßnahmen/Aktivitäten und Resultaten beigebracht sind.

Ausgangsfragestellungen der Perspektivstudie

Können Wirkungen ausschließlich durch strenge experimentelle Evaluationsstudien erfolgen? Ist diese aus den Naturwissenschaften entlehnte Experimentallogik bei der Evaluation sozialpolitischer, oft

personenbezogener Dienstleistungsprogramme angemessen? Welche Alternativen oder sinnvollen Ergänzungen dazu bestehen? Dies sind Ausgangsfragestellungen der Perspektivstudie.

Aufgaben wirkungsorientierter Evaluation

Mit dem Begriff der wirkungsorientierten Evaluation wird eine Vielzahl von Ansätzen und Modellen der Evaluation zusammengefasst, die eine oder mehrere der folgenden Aufgaben übernehmen: Sie stellen Auftraggebenden, Programmverantwortlichen oder anderen wichtigen Beteiligten von Informationen zur Verfügung: um entweder die Programmwirkungen fundiert zu planen (Wirkungsmodellierung), um eine Abschätzung sowohl der negativen wie der positiven Auswirkungen vornehmen zu können (Wirkungsidentifizierung) oder um zu belegen, dass die Programmaktivitäten die gewünschten Resultate ausgelöst haben (empirischer Wirkungsnachweis).

Funktionen wirkungsorientierter Evaluation – Kap. 1.5

Wirkungsorientierte Evaluation kann zu verschiedenen Zeitpunkten eingesetzt werden: bereits vor der Programmentwicklung (proaktive Evaluation); sie kann die Konzeption (klärende Evaluation) oder Umsetzung des Programms (interaktive Evaluation) begleiten, sie kann Kennzahlensysteme für die laufende Programmdokumentation bereitstellen (dokumentierende Evaluation) oder sie kann mit/nach Programmschluss feststellen, ob angemessene Resultate/Wirkungen erreicht sind, und ggf. ob sie die eingesetzten Ressourcen rechtfertigen (wirkungsfeststellende Evaluation).

Leistungsarten der Evaluation

Formative Evaluationen unterstützen, dass das Programm so ausgestaltet und verbessert wird, dass es seine beabsichtigten Ziele in möglichst großem Maße erreicht. Summative Evaluationen messen Resultate und Wirkungen, um eine Entscheidung über das Programm empirisch zu fundieren und eine Basis für die im demokratischen politischen System gebotene Rechenschaftslegung zu schaffen.

Vorsprung der Evaluation in den USA – Kap. 2.1.1

Da in Deutschland – wie in den meisten europäischen Ländern – die Evaluation ein noch recht junger Ansatz ist, werden zur Beantwortung der Fragestellungen zunächst Erfahrungen aus Nordamerika ausgewertet. Es gibt im sozialpolitischen System der USA eine ausgeprägte Tradition mit wirkungsorientierter Programmgestaltung und -evaluation. Bei gänzlich

unterschiedlichem System der sozialen Sicherung gibt es eine ausgeprägte programmförmige Politik zur Verbesserung kritischer Lebenslagen, und zwar insbesondere bezüglich Erziehung/Bildung, Gesundheit, Wohnen und Beschäftigung. In diesen Feldern führten intensive Einsätze von Evaluationen, dabei sowohl Erfolge als auch Enttäuschungen, seit Mitte der 60er Jahre zur Expansion der Evaluationsprofession und zur Ausdifferenzierung theoretischer Ansätze und Modelle. Mitte der 70er Jahre kam es – bei vorangegangenem Evaluations-Boom – zu einer Krise der Profession, da die Ergebnisse der oft groß angelegten Studien von der Politik kaum genutzt wurden. Seit Mitte der 90er Jahre, die einhergingen mit einer starken Einschränkung sozialpolitischer Maßnahmen, thematisiert die Evaluationsprofession verstärkt politische Macht und soziale Werte. Diesen Kontext sollen Programm-Evaluatoren/-innen als Bezugspunkt für die Erarbeitung von Evaluationsplänen, für Interpretationen von Daten wie für die Formulierung von Schlussfolgerungen und ggf. Empfehlungen nutzen.

In Deutschland vorrangig: Problembeschreibung - Kap. 2.1.2

In Deutschland besteht eine Tradition dichter, oft durch innovative Forschungsansätze geleisteter Beschreibungen und Erklärungen des *Problems* „soziale Desintegration“, „Ausschluss“ oder „Armut“. In den letzten Jahrzehnten entstanden z.B. die „Sozialindikatorenbewegung“ oder die „differentielle Arbeitslosenforschung“, die in die Armutsberichte der Wohlfahrtsverbände und Gewerkschaften und schließlich in die Armuts- und Reichtumsberichterstattung der Bundesregierung einmünden.

Erfahrungen mit wirkungsorientierten Programmen und ihrer Evaluation sind eher selten und erst neuerlich werden – etwa im Bund-Länder-Programm „Die Soziale Stadt“ oder im Kontext der Reform der Sozialhilfe – innovative Ansätze der Evaluation genutzt. Bei sich womöglich verschärfenden sozialen Problemlagen ist die gezielte Stärkung staatlichen und gesellschaftlichen Lösungspotentials zur Verbesserung und Stabilisierung gefährdeter Lebenslagen anzustreben. Hierzu kann wirkungsorientierte Evaluation einen Beitrag leisten.

Werte als zentrale Bezugspunkte der Evaluation – Kap. 2.2

Besonders in sozialpolitischen Themenfeldern müssen Evaluationen klar ausweisen, wie sie mit den oft starken Wertspannungen umgehen, die sowohl in Bezug auf die Ziele einer Politik der Armutsvermeidung und -minderung als auch in Bezug auf die einzusetzenden Maßnahmen bestehen. Was als Ziel, als Erfolg und schließlich als positive Wirkung gilt, hängt von der Wertposition ab, welche der/die Beurteilende einnimmt. Daher muss im Rahmen von Evaluationen ausgewiesen werden, welche Instanz für die Werteklärung verantwortlich ist, insbesondere in welchem Maße sie Bestandteil des Evaluationsprozesses selbst ist.

12 Modelle der Evaluation nach Werteberücksichtigung – Kap. 2.3

Gemäß ihrer Vorgehensweise bezüglich sozialer Werte lassen sich vier Haupttypen der Evaluation unterscheiden, denen in der Perspektivstudie insgesamt zwölf kommentierte Evaluationsmodelle zugeordnet sind:

- „Wertedistanzierte“ Evaluation klammert Werturteile aus und betrachtet sie als durch demokratische Verfahren evaluationsextern vorentschieden/zu entscheiden. Evaluationsspezialisten sollen nach strikten Regeln möglichst objektive Daten, Ergebnisse und Schlussfolgerungen zu sozialen Tatsachen bereitstellen, die in jedem beliebigen Werterahmen identisch zustande kämen.

  (1) Programmziel-gesteuerte, (2) Experimentaldesign-gesteuerte, (3) Quasi-Experimentaldesign-gesteuerte, (4) Programmkosten/-nutzen-gesteuerte, (5) Kontext-Mechanismus-gesteuerte, (6) Programmtheorie-gesteuerte Evaluation.

- „Werterelativistische“ Evaluation stellt soziale Werte bei Planung, Durchführung und Nutzung in den Mittelpunkt. Sie macht die bestehenden Spannungen transparent, ohne Partei zu nehmen. Evaluationsspezialisten klären im Dialog mit den Programmbeteiligten Werte, was Motivation und soziale Energie für die aktive Mitwirkung an der Evaluation einschließlich der Nutzung ihrer Ergebnisse freisetzen soll.

  (7) Spannungsthemen-gesteuerte, (8) Dialoggesteuerte Evaluation.

- „Wertepriorisierende“ Evaluation will differierende Werte rangordnen. Ziel ist ein möglichst breiter Wertekonsens als Basis für die Evaluation und die Nutzung ihrer Ergebnisse. Geklärte

Werte sind Voraussetzung für relevante Fragestellungen, Daten, Interpretationen und Schlussfolgerungen. Dabei wird ggf. hingenommen, dass Partizipation und Einflussnahme unterschiedlicher Gruppen auf die Bildung von Prioritäten gemäß den realen Machtverhältnissen differieren.

(9) Entscheidungsgesteuerte, (10) Nutzungsgesteuerte
(11) Stakeholder-Interessen-gesteuerte Evaluation.

- „Wertepositionierte“ Evaluation geht davon aus, dass soziale Systeme durch starke Machtungleichgewichte und Ungleichheit geprägt sind. Sie will durch Einnahme einer bestimmten Wertposition ein Gegengewicht bilden gegen die bestehende Wertehegemonie im politischen, sozialen und kulturellen Raum, und so die Position der Einflussschwachen stärken.
  (12) Selbstorganisations-gesteuerte Evaluation.

Klärungsbedarf für Evaluation in Deutschland – Kap. 2.4

Für die Mehrzahl der zwölf vorgestellten Evaluationsmodelle werden einzelne deutschsprachige Anwendungsfälle aus dem Umfeld der Armuts- und Reichtumsberichterstattung identifiziert und kurz dargestellt. Der Ertrag der Recherche nach einschlägigen aktuellen Evaluationsberichten aus Deutschland ist gering. In den meisten aufgefundenen Studien sind evaluationstheoretische Grundlagen nicht offen gelegt; Werteberücksichtigung und Nutzungsvorbereitung werden kaum erörtert. Ein Veröffentlichungsgebot öffentlich finanzierter Evaluationen wäre ein Schritt, dem evaluationstheoretischen Defizit entgegenzuwirken und damit verbesserte Grundlagen für nützliche Evaluationen zu schaffen. Auch könnte die Qualität dieser veröffentlichten Evaluationen mit Hilfe der Standards für Evaluation bewertet werden (Meta-Evaluation). Das Glossar im Anhang der Perspektivstudie erläutert Fachbegriffe der Evaluation und soll einer breiteren Leserschaft das Verständnis besonders der evaluationstheoretischen Teile des Berichtes erleichtern.

Empirische Erhebung von Qualitätsanforderungen an Evaluationen – Kap. 3

Die in der Evaluationstheorie und in den Modellen identifizierten – teils differierenden – Merkmale guter Evaluationen werden den Anforderungen von Experten/-innen gegenübergestellt, die mit Evaluationen im Arbeitsbereich der Armuts- und Reichtumsberichterstattung befasst sind. Dies

geschieht durch Interviews mit ausgewiesenen Wissenschaftlern/-innen sowie Gruppendiskussionen mit Verantwortlichen aus Bundes- und Landesministerien. Diese beiden Erhebungen machen deutlich, dass viele theoretische Überlegungen ähnlich lautend sowohl von der Wissenschaft als auch von der Praxis als Anforderungen und konzeptionelle Vorstellungen formuliert werden.

- Viele Experten/-innen unterstützen ein partizipatives und pluralistisches Vorgehen in der Evaluation, dabei auch eine Auswahl von geeigneten Evaluationsmodellen und Vorgehensweisen nach Zweck und Fragestellungen des Evaluationsauftrags. Von einem solchen, gegenstandsangepassten Vorgehen erwarten sie in besonderem Maße wirklichkeitsnahe Ergebnisse und Anstöße für die Weiterentwicklung von Programmen und Politiken.

  Partizipation und Pluralität als Grundprinzipien der Evaluation – Kap. 3.5.4

- Aus ihrer Sicht ist der systematische Wirkungsnachweis sozialpolitischer Programme eine herausragende Aufgabe für Evaluationen – gleichzeitig weisen sie auf die schwer bis kaum lösbaren konzeptionellen und methodischen Probleme hin, die damit verbunden sind, z. B. bezüglich der operationalen Messung, der Datenbasis oder kontrollierter Erhebungsdesigns. Vorgestellte Lösungsvorschläge decken sich mit den Überlegungen in verschiedenen Evaluationsmodellen.

  Relevanz wirkungsorientierter Evaluation – Kap. 3.5.3

- Viele Experten/-innen sehen eine Chance darin, Evaluation bei der Vorklärung, Vorbereitung und Einführung von Programmen als Planungs- und Steuerungsinstrument einzusetzen, um Bedarfsgerechtigkeit und Wirksamkeit armutsbekämpfender Maßnahmen sicherzustellen. Sie befürworten es, Aufgaben der Modellierung, der Identifizierung und des Nachweises von Wirkungen frühzeitig vor und während der Startphase von Programmen zu bearbeiten und insofern evaluatorisches Know-how in den gesamten Programm- und Politikzyklus einzubeziehen.

- Die Frage, welchen Beitrag die Evaluation zur Klärung von Programmzielen und zur Klärung von Wertperspektiven leisten soll, die den Interpretationen und Schlussfolgerungen zugrunde gelegt

  Dissens zur Rolle der Evaluation bei Ziel- und Werteklärung – Kap. 3.5.1

werden, wird kontrovers diskutiert.

- Es besteht ein großes Interesse, den evaluationstheoretischen Diskurs und die Auswertung vorliegender Evaluationsstudien umfassend zu nutzen, um zu nützlichen, glaubwürdigen und methodisch genauen Evaluationen zu kommen. Insgesamt erhoffen sich die Gesprächspartner/-innen auch fruchtbare Beiträge für die öffentliche und politische Konkretisierung von Strategien zur Armutsvermeidung und -minderung. Hierin sollen Politik, Wissenschaft und Öffentlichkeit einbezogen werden.

Diskurs über Evaluation gewünscht – Kap. 3.5.4.6

- Eine zentrale Anforderung besteht darin, Evaluationsstudien möglichst zeitnah zu ihrer Fertigstellung zu veröffentlichen, so dass ihr Beitrag zur Klärung von Zielen und Strategien im demokratischen Prozess erhöht wird und darüber hinaus auch die Evaluation als methodisches Konzept weiter entwickelt werden kann. Als wichtig bezeichnet wird die integrierte Auswertung von Evaluationsstudien gerade auch im Hinblick auf divergierende Ergebnisse einzelner Studien.

Evaluationsberichte öffentlich zugänglich machen – Kap. 3.5.4.7

Im Rahmen von Programmen und Politiken zu Armutsvermeidung und -minderung stellt sich die Frage, wie die in den letzten Jahren verbesserte Datengrundlage im Bereich der Armuts- und Reichtumsberichterstattung systematisch und effizient für wirkungsorientierte Evaluationen genutzt werden kann. Dies wird durch Auswertung von Literatur und durch Befragung ausgewiesener Experten/-innen beantwortet.

Nutzbarkeit vorhandener Datenquellen für Evaluationen – Kap. 4

Monitoring und Evaluation sind eng aufeinander verwiesen – dabei stößt ein allein stehendes Monitoring im Kontext einer Politik der Armutsvermeidung und –minderung schnell auf Grenzen. In aller Regel ist zur Erstellung eines gültigen und nützlichen Monitoring-Systems evaluatorische Vorarbeit zu leisten.

Zum Verhältnis von Monitoring und Evaluation; Kap. 4.1

Die Nutzbarkeit vorhandener Datenquellen wird insgesamt positiv eingeschätzt, wobei Verbesserungen bei den Datengrundlagen und bei der Integration unterschiedlicher Datentypen eingefordert werden:

- Administrative Daten stehen kostengünstig zur Verfügung. Oft können sie besonders im Zusammenhang von Evaluationen in regionalen Kontexten herangezogen werden, um die Ausgangssituation bei den Zielgruppen und die Rahmenbedingungen des jeweiligen Programms genau zu beschreiben sowie interregionale Vergleiche zu ermöglichen. Dabei wird genaue Kenntnis über den Entstehungsprozess und die Einschränkungen beim Gebrauch von administrativen Daten als unverzichtbar angesehen, um zu gültigen Schlussfolgerungen zu kommen. Teilweise wird auch eine Vereinheitlichung der EDV-Systeme, mit denen administrative Daten bei den verschiedenen Gebietkörperschaften erzeugt werden, als wünschenswert bezeichnet.

  Gültigkeit kostengünstiger Daten sichern – Kap. 4.2

- Zentral zusammengeführte statistische Daten werden insbesondere als Vergleichsgrundlage für bundesweit oder landesweit durchgeführte Programme angesehen und sollten noch stärker ausgeschöpft werden. Wie bei den administrativen Daten wären auch für statistische Daten teilweise Längsschnitte wünschenswert. Da bestimmte Teilgruppen der von Armut betroffenen/bedrohten Personen nicht genügend umfangreich und trennscharf ausweisbar sind, müssen statistische Daten durch administrative und Paneldaten ergänzt werden.

  Statistische Daten konsequent nutzen – Kap. 4.3

- Paneldaten werden als wichtige Bezugsgröße für Verlaufsuntersuchungen genannt. Wegen ihrer oft relativ kleinen Stichproben aus Bevölkerungsgruppen, die besonders von Armut betroffen oder bedroht sind, sind sie nur eingeschränkt nutzbar. Mit erhöhtem Finanzeinsatz sind derartige Stichproben schrittweise zu vergrößern und die vorhandenen Einschränkungen, spezielle Evaluationsfragestellungen zu beantworten, können gemindert werden.

  Bei Panelstudien Themenschwerpunkte ausbauen – Kap. 4.4

Die Verknüpfung vorhandener Datenquellen wird verstärkt empfohlen, um einerseits die Datenbasis zu verbreitern bzw. andererseits, um Längsschnittdaten bereit zu stellen.

Verknüpfung und Zugänglichkeit der Datenquellen erhöhen – Kap. 4.5

Teilweise wird es als notwendig erachtet, bei den Ämtern oder den Instituten, welche die verschiedenen Datentypen erzeugen bzw. vorhalten,

Ansprechpartner/-innen für Evaluatoren und Evaluatorinnen auszuweisen, so dass in Zukunft die Integration derartiger Daten in Evaluationen erleichtert wird.

Adressaten/-innen wirkungsorientierter Evaluation – Kap. 5

Der verstärkte und systematische Einbezug von Evaluation und von ihren Ergebnissen ist für einen breiten Kreis von Adressaten/-innen von Bedeutung: Für Parlament und Regierung schafft er eine erweiterte Beratungs- und Entscheidungsgrundlage und ermöglicht die Kontrolle der öffentlichen Mittelverwendung. Der weiteren interessierten Öffentlichkeit werden abgesicherte Informationen geliefert um zu beurteilen, in welchem Maße soziale Problemlagen mit Erfolg bearbeitet werden. Professionell und ehrenamtlich in der Armutsbekämpfung Tätige erhalten konkrete Hinweise für eine verbesserte Steuerung der von ihnen durchgeführten Programme und Maßnahmen. Den an der Armuts- und Reichtumsberichterstattung arbeitenden Wissenschaftlern/-innen werden mit empirischen Evaluationen zusätzliche Datenquellen erschlossen.

Empfehlungen für Beauftragung, Planung und Berichterstattung von Evaluationen

Um von den Adressaten/-innen genutzt zu werden, müssen Evaluationen regelmäßig, unabhängig und fachkundig durchgeführt werden. Hieraus lassen sich nachfolgende Empfehlungen ableiten:

- Gesetze, Politiken und Programme mit größerer und langfristiger Bedeutung sollen frühzeitig, ggf. im Rahmen von Pilotvorhaben evaluiert werden.
- Die Ausschreibung von Evaluationen soll einerseits Zwecksetzung und zu beantwortende Ausgangsfragestellungen sowie das zu evaluierende Programm klar benennen, andererseits für die Ausführung hinsichtlich zu wählender Evaluationsmodelle und –methoden hohe Flexibilität einräumen.
- Die „Standards für Evaluation“ der Deutschen Gesellschaft für Evaluation sollten sowohl für die Beurteilung von Angeboten durch den Auftraggeber als auch für die Planung und Steuerung der Evaluationsleistungen durch den Auftragnehmer zugrunde gelegt werden.
- In dem besonders stark von Wertspannungen gekennzeichneten Politikfeld der Armutsvermeidung und sozialen Integration ist die Unab-

hängigkeit der Evaluation von herausragender Bedeutung und erfordert hohe Transparenz über die theoretischen Grundlagen, angewandten Methoden, genutzten Datenquellen und besonders das Zustandekommen von Schlussfolgerungen und Empfehlungen.

- Die Ergebnisse einer Evaluation sollen klar und verständlich für die interessierte Öffentlichkeit aufbereitet werden und es ist wünschenswert, dass öffentlich beauftragte und finanzierte Studien binnen kurzer Frist einer breiten interessierten Öffentlichkeit zugänglich gemacht werden.
- Die Integration von Evaluationsergebnissen in die Armuts- und Reichtumsberichterstattung kann gefördert werden durch frühzeitigen Beizug evaluatorischen Sachverstands, eine leitfadengestützte Auftragsvergabe, ein erweitertes Evaluationsdokumentationswesen sowie auf Lebenslagen-Dimensionen zugeschnittene Metaanalysen von Evaluationen. Hierzu sollte ein Leitfaden mit Checklisten für die Vorbereitung und Vergabe von Evaluationsaufträgen bereitgestellt werden, dem ausgewählte, beispielhaft für den Armuts- und Reichtumsbericht veranschaulichte Evaluationsstandards zugrunde liegen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

ARB	Armuts- und Reichtumsberichterstattung
BA	Bundesanstalt für Arbeit
BMA	Bundesministerium für Arbeit und Soziales
BMGS	Bundesministerium für Gesundheit und Soziale Sicherung
CMO	Context-Mechanism-Outcomes
DEGEVAL	Deutsche Gesellschaft für Evaluation
DJI	Deutsches Jugendinstitut
ECHP	Haushaltspanel der Europäischen Gemeinschaft
E.E.	Empowerment Evaluation
EVS	Einkommens- und Verbrauchsstichprobe
EU-SILC	Survey on Income and Living Conditions
GG	Grundgesetz
ILO	International Labour Organisation
JC	Joint Committee on Standards for Educational Evaluation
KVI	Kommission zur Verbesserung der informationellen Infrastruktur
NIEP	Niedrigeinkommens-Panel
MASQT	Ministerium für Arbeit, Soziales und Qualifikation; Nordrhein-Westfalen
MAUT	Multiattribute Utility Technology
MZ	Mikrozensus
ReNoMo	REsponsive to Stakeholder`s Interests by Working with NOminal Groups using the MOderation Method
SEVAL	Schweizerische Evaluationsgesellschaft
SGB	Sozialgesetzbuch
SOEP	Sozioökonomisches Panel
IAB	Institut für Arbeitsmarkt und Berufsforschung

0 Zu diesem Bericht

Der erste Armuts- und Reichtumsbericht für Deutschland wurde am 25. April 2001 vorgestellt und vom Bundeskabinett gebilligt. Mit dem Beschluss des Deutschen Bundestages vom 19. Oktober 2001 zur Verstetigung der Armuts- und Reichtumsberichterstattung wurde die Bundesregierung dazu aufgefordert, „den zweiten Bericht als Instrument zur Überprüfung von Politik gegen Armut und soziale Ausgrenzung einerseits und Förderung von Teilhabegerechtigkeit andererseits in Deutschland zu nutzen, indem die Wirksamkeit der Maßnahmen überprüft und neue Maßnahmen angeregt werden.“ Hierzu wurden vom damaligen Bundesministerium für Arbeit und Soziales (BMGS) verschiedene Projekte initiiert; u. a. wurde das Forschungsprojekt „Wirkungskontrolle in der Armuts- und Reichtumsberichterstattung – Eine Perspektivstudie" am 08.07.2002 beim Kölner Evaluationsinstitut Univation in Auftrag gegeben.

Die vorliegende Perspektivstudie entwickelt eine theoretische und konzeptionelle Basis für Evaluationen im Rahmen der Armuts- und Reichtumsberichterstattung. Durch Auswertung evaluationstheoretischer Literatur, der Standards für Evaluation sowie veröffentlichter Evaluationsstudien wurden Evaluationsmethoden, Messkonzepte, Zugänge zur Werteberücksichtigung und Nutzungskonzepte der wirkungsorientierten Evaluation aufgearbeitet. Diese Erkenntnisse und auch Fragen zur Datenlage für Evaluationen wurden durch Interviews und Gruppendiskussionen mit Wissenschaftlern/-innen sowie Experten/-innen aus Bundesministerien erörtert und konkretisiert.

Die Perspektivstudie verfolgt folgende Zielsetzung: Die vorhandenen Erfahrungen mit der Durchführung wirkungsorientierter Studien und konzeptionelles wie methodisches Wissen zur Anlage wirkungsorientierter Evaluationen in Feldern der Armuts- und Reichtumsberichterstattung sollen systematisiert dargestellt und für die interessierte politische und Fachöffentlichkeit verfügbar gemacht werden. Dies soll auch die Ergänzung kommender Armuts- und Reichtumsberichte durch Einbezug von Evaluationsergebnissen vorbereiten. Dabei bemüht sich die Studie bezüglich wissenschaftstheoretischer Positionen und sozialpolitischer Werte um Neutralität, wohl wissend, dass dies begrenzt möglich ist.

Es werden zum einen verschiedene Positionen zu Modellen und Methoden dargelegt, wie Wirkungen staatlichen Handelns nachprüfbar festgestellt werden können. Eine Bandbreite bislang in Deutschland kaum aufgearbeiteter Modelle der Evaluation wird dargestellt und synoptisch verglichen. Auf dieser Grundlage können künftig für bestimmte Fragestellungen in der Armuts- und Reichtumsberichterstattung geeignete Evaluationsmodelle ausgewählt werden. Diese theoretische und konzeptionelle Ausarbeitung wird durch Stellungnahmen von Experten und Expertinnen überprüft und ergänzt. Dabei zeichnen allein die Autoren, denen für ihre Bereitschaft zur Mitwirkung an dieser Stelle gedankt sei, verantwortlich für die Ausführungen in diesem Bericht. Es ist nicht auszuschließen, dass die im Anhang 7.5 aufgeführten Personen abweichende oder gegensätzliche Positionen dazu einnehmen. Ein dritter Schwerpunkt befasst sich mit der Möglichkeit, bestehende Datenquellen und Datenerhebungsverfahren in wirkungsorientierte Evaluationen zu integrieren. Abschließend werden perspektivisch Thesen und Empfehlungen formuliert.

Die Perspektivstudie gliedert sich wie folgt: Im Anschluss an eine Einführung in die wirkungsorientierte Evaluation *(Kapitel 1)* werden in *Kapitel 2* allgemeine Modelle der Evaluation dargestellt. Dies geschieht geordnet nach unterschiedlichen Zugängen zur Berücksichtigung sozialer Werte. Sie werden durch ausgewählte Beispiele von abgeschlossenen Evaluationen aus Deutschland illustriert. *Kapitel 3* analysiert die Anforderungen an Evaluationen insbesondere bezüglich ihrer Nutzung durch Politik, Wissenschaft und Öffentlichkeit. Die Ausführungen basieren auf den Stellungnahmen von Experten/-innen aus dem wissenschaftlichen Gutachtergremium des Bundesministeriums für Gesundheit und Soziale Sicherung und aus Bundesministerien, die mit Fragen der Armuts- und Reichtumsberichterstattung bzw. der Evaluation in diesem Themenbereich befasst sind. In *Kapitel 4* wird die Verfügbarkeit von Daten aus Verwaltung, amtlicher Statistik und Umfrageforschung für Evaluationen dargelegt. Kapitel *5* erörtert die Perspektive wirkungsorientierter Evaluationen in der Armuts- und Reichtumsberichterstattung und gibt Anregungen dafür, wie die Gültigkeit und Glaubwürdigkeit von Evaluationen für eine produktive Nutzung ihrer Ergebnisse durch Politik und Öffentlichkeit gestärkt werden kann.

1 Evaluationstheoretische Grundlegung

Dieses erste Kapitel führt theoretische Grundbegriffe der Evaluation ein, verdeutlicht diese mit Hilfe der „Standards für Evaluation“ und weist differenziert Anwendungsmöglichkeiten und -grenzen wirkungsorientierter Evaluationen in Feldern der Politik zur Armutsvermeidung und -verminderung auf.

Zunächst wird eine Arbeitsdefinition von Evaluation gegeben. Externe und interne, summative und formative Evaluation werden unterschieden. Die „Standards für Evaluation“ als Bezugsgrundlage professioneller Evaluationspraxis werden vorgestellt und die Spannungen insbesondere zwischen dem Qualitätsmerkmal der „Nützlichkeit“ und dem der „Genauigkeit“ von Evaluationen werden heraus gearbeitet. Als typische Evaluationsgegenstände im Umfeld der Armuts- und Reichtumsberichterstattung werden Politiken und insbesondere „Programme“ identifiziert, deren Resultate und Wirkungen systematisch beschrieben und bewertet werden sollen. Dies geschieht unter zu beschreibenden Rahmenbedingungen, auf Basis eines Programmkonzeptes, durch einen Programmprozess, der Maßnahmen und Interventionen umsetzt. Die verschiedenen Arten von Resultaten und Wirkungen werden systematisch bestimmt und in Beziehung zueinander gesetzt, außerdem veranschaulicht für Programme zur Armutsbekämpfung. Abschließend werden in einem Überblick die verschiedenen Zugänge zur wirkungsorientierten Evaluation aufgezeigt, die für die unterschiedlichen Entwicklungsphasen, in denen sich ein Programm befindet, differentiell geeignet sind.

1.1 Definition und Formen der Evaluation

Die Evaluation als professioneller Ansatz der Politikberatung und -begleitung ist in Deutschland noch recht jung. Erst 1997 wurde die „Deutsche Gesellschaft für Evaluation“ gegründet. Seit Herbst 2002 erscheint die „Zeitschrift für Evaluation“. Eine akademische, z.B. postgraduale, Ausbildung in Evaluation steht in Deutschland kurz vor der Einführung; der Ausbildungsgang in Bern befindet sich 2003 in seinem zweiten Durchgang.

In dieser Aufbruchssituation sind Begriffsklärungen wichtig für die Weiterentwicklung der Theorie und Praxis der Evaluation.

Folgende Arbeitsdefinition liegt dieser Perspektivstudie zugrunde:

> *Evaluation bezeichnet die Summe systematischer Untersuchungen, die empirische, d.h. erfahrungsbasierte, Informationen bereit stellen, so dass es möglich wird, den Wert eines (in der Regel sozialen) Evaluationsgegenstandes nachvollziehbar einzuschätzen.*

Empirisch gewonnene Informationen bilden die Basis für Beschreibungen und Bewertungen, welche durch Evaluationen systematisch und intersubjektiv nachvollziehbar geleistet bzw. vorbereitet werden. Die Frage, wer schließlich bewertet, wird in den verschiedenen Modellen der Evaluation (vgl. Kap. 2) unterschiedlich beantwortet: die Evaluatoren/-innen, die Entscheider/-innen oder auch andere Beteiligte und Betroffene.

Die Abstützung auf bewährte empirische Methoden der Datenerhebung und -aufbereitung zu gehaltvollen Informationen ist unverzichtbarer Bestandteil von Evaluationen – nicht selten wird Evaluation sogar damit gleichgesetzt in Wendungen wie „Empirische Forschung mit dem Ziel, Wirkungen zu kontrollieren“. Wie diese Perspektivstudie deutlich macht, ist die systematische Beschreibung und Bewertung – hier von staatlichem Handeln in den breiten Feldern, die in Deutschland mit der Armuts- und Reichtumsberichterstattung angesprochen sind – ein komplexes und vielfach angreifbares Unterfangen.

Dieses exponierte Agieren gegenüber Politik und Öffentlichkeit, wirtschaftlichen und sozialen Interessengruppen sowie die dabei gemachten und zu lehrbarem Wissen verarbeiteten Erfahrungen haben in den vergangenen Jahrzehnten einen ausgedehnten Fundus theoretischen Wissens über Evaluation bis hin zu Modellen der Evaluation und Standards für Evaluationen hervorgebracht.

Evaluation wird – wie die Arbeitsdefinition festhält – zu Recht mit „Bewerten“ assoziiert. Menschen in ihren verschiedenen Rollen – Politiker/-innen, soziale Fachkräfte, Programmverantwortliche ... – reagieren auf dieses Ansinnen nicht selten ablehnend, ja aggressiv: Fachkräfte, deren Leistungen für ihre Zielgruppen künftig regelmäßiger und intensiver bewertet werden sollen, sehen sich und ihre professionelle Kompetenz kritisch geprüft, z.B. wenn „fremde“ Evaluationsinstitute in ihre Einrichtungen „eindringen“ und Daten abfordern, diese dann auswerten und vielleicht zu Schlussfolgerungen gelangen, welche berufliche Perspektiven verändern oder gar in

Frage stellen können. Wenn dies in Arbeitsfeldern geschieht, die – nicht nur in Bezug auf die Charakteristika der Zielgruppen, sondern auch auf die Ressourcen der dort arbeitenden Träger betrifft – durch Armut gekennzeichnet oder zumindest bedroht sind, wird das Spannungsfeld erahnbar, in dem sich Evaluation in Feldern der Armuts- und Reichtumsberichterstattung bewegt.

Um in diesem Spannungsfeld zu sozial nützlichen, glaubwürdigen und überprüfbaren Evaluationsergebnissen zu kommen, gibt es in der Evaluationstheorie eine Reihe von Vorschlägen und Arrangements, die Einführung von Evaluation, die Durchführung der eigentlichen Erhebungen bis hin zur Verbreitung von Ergebnissen betreffend. Einen zentralen Punkt bildet darin die Frage, wie Evaluationen mit sozialen Werten umgehen müssen, damit sie einen angemessenen Beitrag zur Lösung sozialer Probleme in der demokratischen Gesellschaft leisten können (vgl. Kap. 2.2).

Ein verantwortlicher Umgang mit dem Gegenstand der Bewertung ist ein Angelpunkt der professionellen Evaluationsethik. Daher gilt es zunächst deutlich zu machen, „was“ Gegenstand der Evaluation sein soll und welche Vermischungen von Evaluationsgegenständen (z.B. Programme, Organisationen und Personen) kontraproduktiv sind. Diese können zu erheblichen Widerständen oder zur Fehlleitung investierter öffentlicher Mittel führen oder politische Angriffe gegen die Evaluation, ihre Durchführer/-innen und Auftraggeber/-innen auslösen.

Evaluationsgegenstände sollen spezifisch bestimmt sein, also nicht etwa allgemein „Die Sozialpolitik“ oder „Die Steuerpolitik“ betreffen. Es geht um Elemente der Politik, die konkret abgegrenzt und benannt sein sollen. In der Regel handelt es sich um staatliches Handeln im öffentlichen, gesetzlich regulierten Auftrag, und dort vorrangig um geplante, durch Budgets abgegrenzte und über Ziele, Umsetzungen und Resultate bestimmbare Maßnahmen, Instrumente, Vorgehensweisen (vgl. Bussmann /Knoepfel 1997). In der Evaluationstheorie hat sich zur Bezeichnung derartiger „intentional gerichteter, mit Ressourcen ausgestatteter Interventionsbündel“ der Terminus „Programm“[1] durchgesetzt. „Programm“ eignet sich besonders zur Abgrenz-

1 Definitionen bei Smith (1989), Rossi, Freeman, Lipsey (1999, S. 23), Bundesamt für Gesundheit (1997, S. 12).

ung und darüber hinaus Konzeptualisierung[2] der zu evaluierenden Gegenstände im Rahmen unterschiedlicher Felder staatlicher Politik (vgl. detaillierter das Kap.1.3).

Im Unterschied dazu können auch Personen, z.B. Mitarbeiter/-innen der öffentlichen Verwaltung oder Mitglieder von Zielgruppen, „Gegenstand“ der systematischen Bewertung sein. Teilweise geschieht dies auch im Zusammenhang mit Programmevaluationen, kann dann erhebliche Widerstände auslösen und schnell auf Schutzbereiche des Datenschutzes treffen.

Die Ausführungen in diesem Bericht beziehen sich ausschließlich auf Programmevaluationen. Es ist eine Erfahrung aus vielen Evaluationsstudien, dass die Beurteilung von Mitarbeitenden oder die wertende Einschätzung von Zielgruppenmitgliedern strikt vom Vorhaben der Programmevaluation getrennt werden soll.[3]

Dies gilt sowohl für externe wie für interne Evaluationen. In aller Regel dürften in den Politikfeldern von Armut und Reichtum externe Evaluationen durchgeführt werden, d.h. externe Institute oder Einzelevaluatoren/-innen, die keine Verantwortung tragen für das zu evaluierende Programm, werden mit den Evaluationsaufgaben betraut. Zunehmend entstehen wegen der gesetzlichen Finanzierungsbestimmungen[4] oder auf dem Hintergrund des trägerinternen Qualitätsmanagements auch interne Evaluationsansätze (vgl. Heiner 1998). Das heißt, Positionen oder Abteilungen werden neu eingerichtet, um die eigenen Programme systematisch zu beschreiben und zu bewerten. Teilweise ist dies mit Qualitätsmanagement oder mit Monitoring-Systemen verbunden, die soziale Dienstleistungsunternehmen auf sich schnell verändernden Märkten zunehmend zur Unternehmenssteuerung benötigen. Dabei spielt schließlich

---

2 Vgl. die Ausführungen zum Programmtheorie-gesteuerten Evaluationsmodell in Kap. 8.6.

3 Die schließt selbstverständlich nicht aus, dass Daten aus Mitarbeiterbefragungen oder Testergebnisse von Maßnahmenteilnehmenden für Zwecke der Programmevaluation genutzt werden, wenn ein „informiertes Einverständnis“ seitens der Datengeber/-innen vorliegt. Der umgekehrte Weg – Nutzung von personenbezogenen Daten, die für Zwecke der Programmevaluation zugänglich gemacht oder erzeugt worden sind, für Zwecke der Beurteilung oder Auswahl von Personen – muss hingegen durch geeignete Vorkehrungen versperrt sein.

4 Exemplarisch veranschaulicht sei dies für die Arbeitsförderung, in deren Rahmengesetz - SGB III – bei der neuesten Novelle die folgende Bestimmung eingefügt wurde: „Zugelassen für die Förderung sind Träger, bei denen eine fachkundige Stelle festgestellt hat, dass ....(Punkt 1 bis 3 ausgelassen) der Träger ein System zur Sicherung der Qualität anwendet.“(§ 84 – Anforderungen an Träger).

auch der Ansatz der Selbstevaluation eine Rolle, der in diesem Papier ebenso wie die interne Evaluation im Hintergrund bleibt.[5]

In dem Maße, in dem Evaluation zur Überprüfung von Qualität und Wirkungen staatlicher Politik verstärkt angefragt und damit selbst zum Objekt des öffentlichen Interesses wird, ist sie gefordert, die eigenen Qualitätsmaßstäbe transparent darzulegen, um damit ihre eigenen Leistungen und Vorgehensweisen einer kritischen Prüfung zugänglich zu machen. Die 2001 von der Deutschen Gesellschaft für Evaluation verabschiedeten „Standards für Evaluation“ sind selbst wiederum als kodifizierter Fundus kumulierten Erfahrungswissens darüber anzusehen, wie Evaluationen (besonders in politisch umstrittenen Themenfeldern) professionell, ethisch und sozial verantwortlich anzulegen sind.

1.2 Standards der Evaluation als Bezugspunkt

In den USA sind seit den 70er Jahren mehrere Standard-Sets zur Erfassung und Steuerung der Qualität von Evaluationen entwickelt worden. Am verbreitetsten sind die Evaluationsstandards des „Joint Committee on Standards for Educational Evaluation“, das nach der 1981er Erstauflage 1994 die *Program Evaluation Standards* herausgegeben hat. Ihr Geltungsbereich erstreckt sich nun auf Evaluationsfelder über den Sozial- und Bildungsbereich hinaus.

Diese Standards wurden ins Deutsche übersetzt (Joint Committee 2000) und zunächst durch die Schweizerische Evaluationsgesellschaft (SEVAL 2001) aufgearbeitet. Dies geschah in engem Zusammenhang mit dem Nationalen Forschungsprogramm Nr. 27, das sich die Erprobung und Verbesserung von Methoden zur Erfassung und Beurteilung der Wirksamkeit staatlicher Maßnahmen zum Ziel gesetzt hatte (Bussmann 1996).

Die 1997 gegründete Deutsche Gesellschaft für Evaluation entschloss sich ebenfalls zu einem Standard-Setzungs-Prozess, der an die Vorarbeiten von Joint Committee und SEVAL anschloss. Im Herbst 2001 wurden die Standards für Evaluation von der Deutschen Gesellschaft für Evaluation verabschiedet (DeGEval 2002).

---

5 Vgl. zur internen und zur Selbstevaluation Heil/Heiner/Feldmann 2001 sowie den Ansatz der selbstorganisationsgesteuerten Evaluation im Kap. 8.12 sowie die dort genannte Literatur.

Die Evaluationsstandards richten sich sowohl an Evaluatoren und Evaluatorinnen als auch an Auftraggeber/-innen sowie an Beteiligte und Betroffene des zu evaluierenden Programms oder der Politik. Sie sind als Dialoginstrument und fachlicher Bezugspunkt für Evaluationen angelegt. Für alle Phasen der Evaluation liefern die Standards wichtige Hinweise. Dabei kann unterschieden werden nach phasenbezogenen Aufgaben im Ablauf des Evaluationszyklus und nach Querschnittsaufgaben, die mehrfach oder fortlaufend im Rahmen einer Evaluation zu bearbeiten sind.[6]

Abbildung 1: Hauptaufgaben in einer Evaluation

Phasenbezogene Aufgaben	A. Entscheidung über die Durchführung einer Evaluation
	B. Definition des Evaluationsproblems
	C. Planung der Evaluation
	D. Informationsgewinnung
	E. Informationsauswertung
	F. Berichterstattung zur Evaluation
Querschnittsaufgaben	G. Budgetierung der Evaluation
	H. Evaluationsvertrag
	I. Steuerung der Evaluation
	J. Personelle Ausstattung der Evaluation

Quelle: eigene Darstellung

Außerdem geben die Standards Hinweise für die Aus- und Weiterbildung in Evaluation und können bei Evaluationen von Evaluationen (Meta-Evaluationen) eingesetzt werden sowie schließlich Transparenz über Evaluation als professionelle Praxis gegenüber der Öffentlichkeit schaffen. Damit bieten die DeGEval-Standards einen wichtigen Referenzrahmen auch für Evaluationen in Feldern der Armuts- und Reichtumsberichterstattung.

Evaluationen sollen gemäß dieser Standards vier grundlegende Eigenschaften aufweisen: Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit. Den so benannten

6 Für eine Übersicht vgl. DeGEval (2002), S. 38-41.

vier Gruppen sind die 25 Einzelstandards zugeordnet.[7] Zu den mit drei Druckseiten knapp gehaltenen Standards hinzu treten Erläuterungen, Arbeitshilfen und Checklisten sowie ein Anhang (DeGEval 2002). Eine Transformationstabelle ermöglicht es den Anwendern/-innen der noch jungen DeGEval-Standards, auf die Materialien des Joint-Committee (1994/2000) zurückzugreifen.

Die vier Attribute Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit fassen die jeweilige Stoßrichtung der den vier Gruppen zugeordneten Standards pointiert zusammen. Es wird als wünschenswert dargestellt, dass eine Evaluation gleichzeitig alle vier Eigenschaften aufweist.

Die Standardgruppe „Nützlichkeit“ bezeichnet das zentrale Ziel von Evaluationen: Die bereitgestellten Informationen und Schlussfolgerungen sollen von den am evaluierten Programm Beteiligten und von ihm Betroffenen (Stakeholder) tatsächlich genutzt werden (N8)[8]. Aus der Forschung über Evaluation werden insgesamt sieben Anforderungen abgeleitet, deren Erfüllung Nutzung und Nutzen von Evaluationen gewährleisten soll: identifizierte und angemessen einbezogene Beteiligte und Betroffene, geklärte Evaluationszwecke, glaubwürdige und kompetente Evaluatoren/-innen, geeignete Informationsauswahl, transparent gemachte Werte, vollständige und klare Berichterstattung sowie Rechtzeitigkeit der Evaluationsaktivitäten.

Die Standardgruppe „Durchführbarkeit“ markiert, dass in der praktischen Durchführung von Evaluationen – im Unterschied zu wissenschaftlicher Grundlagenforschung – immer wieder Rücksicht zu nehmen ist und ggf. auch Einschränkungen zu machen sind aufgrund der ökonomischen, sozialen und politischen Bedingungen, in denen die zu evaluierenden Programme umgesetzt werden. Die Standards D1 bis D3 halten fest, dass immer wieder Kompromisse und Anpassungen – z.B. in Bezug auf den Umfang und die Tiefe der Datenerhebungen – erfolgen müssen: Die Evaluationsverfahren müssen der zu beschreibenden und zu bewertenden Praxis angemessen sein. Sie sollen diplomatisch eingeführt werden und sie sollen effizient im Sinne von Kosten-Nutzen-Relationen sein, um von der Praxis akzeptiert zu werden.

---

7 Vgl. die im Anhang I abgedruckte Fassung der DeGEval-Standards; die Zitierweise in diesem Kapitel – Anfangsbuchstabe der Standardgruppe sowie Ordnungsnummer innerhalb der Standardgruppe (z.B. N8) – folgt dieser Vorlage.

8 Die Ordnungsnummer – Anfangsbuchstabe der Standardgruppe plus Nummer des Standard innerhalb der Gruppe – verweist auf die im Anhang abgedruckten Standards.

Die Standardgruppe „Fairness“ enthält Ansprüche, wie sie aus der Wissenschaftsethik bekannt sind (F2 „Schutz individueller Rechte“ sowie F5 „Offenlegung der Ergebnisse“) und weitere, welche sich aus dem Spannungsfeld von Evaluation als soziale Praxis bewertende, wissenschaftlich fundierte Vorgehensweise ergeben (F1 „Formale Vereinbarungen“, F3 „Vollständige und faire Überprüfung“ sowie F5 „Offenlegung der Ergebnisse“). Da es sich bei den Zielgruppen von Programmen zur Vermeidung und Verminderung von Armut vielfach um ausgeschlossene, stigmatisierte Menschen handelt, kommt dem Gebot der Fairness bei Evaluationen in diesen Politikfeldern hohe Bedeutung zu.

Die Standards der Gruppe „Genauigkeit“ bezeichnen wissenschaftliche Methoden als unverzichtbare Grundlage der Evaluation. Sie fordern, den Geltungsbereich der Evaluation und ihrer Ergebnisse präzise anzugeben (G1, G2) sowie das Vorgehen und die genutzten Informationsquellen nachvollziehbar und überprüfbar darzulegen (G3, G4). G5 bis G7 verweisen mit Validität, Reliabilität, systematischer Fehlerprüfung sowie qualitativer und quantitativer Datenanalyse auf gängige Anforderungen der empirischen Sozialforschung. G8 fordert, dass Schlussfolgerungen nachvollziehbar aus der empirischen Datengrundlage hergeleitet sind. G9 schließlich verweist darauf, dass sich Evaluationen selbst der systematischen Evaluation stellen sollen (Meta-Evaluation).

Bei den insgesamt 25 Einzelstandards handelt es sich um knappe, verdichtete Texte, die oft einen, selten mehr als drei Sätze umfassen. Sie sind explizit als „Maximalstandards“ konzipiert. Eine ideale Evaluation würde jedem Einzelstandard entsprechen, der für diese Evaluation grundsätzlich anwendbar ist. Die Möglichkeit, dass Standards von vornherein auf ein konkretes Evaluationsvorhaben nicht anwendbar sind, wird insbesondere durch die JC-Standards explizit vorgesehen.[9] Dieses – bereits eingeschränkte – Ideal ist in der Praxis kaum zu erreichen, sei es, dass sich die Anforderungen von zwei oder mehr Standards als einander widersprechend herausstellen oder dass die finanziellen Ressourcen nicht hinreichen, um alle Standards erfüllen zu können. Auch wenn eine konkrete Evaluation kaum alle Standards gleichermaßen erfüllen kann, soll angestrebt werden, jeden einzelnen – soweit anwendbar – soweit wie möglich zu berücksichtigen.

---

9 Vgl. die Checkliste, welche den DeGEval-Standards beiliegt.

Die Evaluationsstandards enthalten in ihrer Konzeption keine Gewichtung – weder der Standardgruppen noch der Einzelstandards. Diese ist anhand der konkreten Evaluation vorzunehmen. Da die einzelnen Standards z.T. konkurrierende Ansprüche formulieren, ist es Aufgabe des Evaluators/der Evaluatorin zu entscheiden, welche Standards gegenüber anderen Vorrang genießen, dies zu begründen und auszuweisen.

Die Evaluationsstandards sind so angelegt, dass sie für unterschiedlichste Evaluationsansätze offen sind. So sind sie grundsätzlich anwendbar sowohl für „formative" Evaluationen, welche die Gestaltung eines Evaluationsgegenstandes begleiten und auf Verbesserungen abzielen, als auch für „summative" Evaluationen, die insbesondere zu einem Evaluationsgegenstand zusammenfassend Bilanz ziehen. Die „Standards für Evaluation“ beanspruchen, für die gesamte Bandbreite von Evaluationsmodellen Gültigkeit zu haben, und sind insofern wissenschaftstheoretisch, disziplinär und methodisch pluralistisch angelegt. Einerseits favorisieren sie kein bestimmtes Modell oder eine Gruppe von Evaluationsmodellen. Andererseits erweist sich, dass manche Modelle – insbesondere wenn sie „rein“ angewandt werden – mit Nützlichkeitsstandards im Konflikt stehen.[10] In der Evaluationspraxis findet in aller Regel ein „Mix“ in der Anwendung von Evaluationsmodellen statt, wenn es darum geht, ein konkretes Evaluationsdesign zu entwickeln und umzusetzen. Hierbei werden die Anforderungen der Evaluationsstandards vielfach umgesetzt, oft auch dann, wenn diese den Evaluatoren/-innen nicht bekannt sind. Damit soll nicht gesagt sein, dass alle oder nur ein Großteil der im Bereich der für die Armuts- und Reichtumsberichterstattung relevanten Politikfelder durchgeführten Evaluationen von hoher Qualität im Sinne der Evaluationsstandards sind – dies zu beurteilen erforderte systematische Meta-Evaluationen.[11]

Bemerkenswert ist, dass die Nützlichkeitsstandards an erster, die Genauigkeitsstandards an letzter Stelle angeordnet sind. Dies soll keine Rangfolge bezeichnen, sondern das in der Evaluationspraxis starke Spannungsfeld verdeutlichen, welches

---

10 Stufflebeam (2001), der einen systematischen Vergleich von insgesamt 22 Evaluationsmodellen entlang der JC-Standards vornimmt, spricht in diesen Fällen – etwas polemisch – von Quasievaluationen.

11 Vgl. die für Evaluationen in verschiedensten Schweizer Politikfeldern durchgeführte Meta-Evaluation von Widmer (1996).

sich aus wissenschaftlichen Gütekriterien einerseits und den Anforderungen der Evaluationsnutzer/-innen andererseits ergibt.

> *„In practice, therefore, the evaluator must struggle to find a workable balance between the emphasis to be placed on procedures that help ensure the validity of the evaluation findings and those that make the findings timely, meaningful, useful to the consumers". (Rossi/Freeman/Lipsey 1999, p. 31).*

Aus diesem Spannungsfeld resultiert auch, dass diese Perspektivstudie besonderen Wert legt auf die begriffliche Klärung von „Wirkung", „Wirkungskontrolle" und „Wirkungsorientierung", wie sie im Kap. 1.5 ausgeführt wird: Einerseits sind offensichtlich Evaluationen, welche die Wirksamkeit von Programmen und Politiken belegen wollen, von hohem potentiellen Nutzen für Entscheider/-innen und andere Beteiligte. Beschreibungen von „Wirkungen" stellen somit einen häufig zentralen „Informationsbedarf des Auftraggebers und anderer Adressaten und Adressatinnen" (N4) der Evaluation dar. Andererseits ist es aufgrund der Spezifik der sozialpolitischen Evaluationsgegenstände (vgl. das folgende Kapitel) ausgesprochen schwierig, die Verursachung von Wirkungen durch ein Programm/eine Politik durch „valide und reliable Informationen" (G5) zu beweisen, zumal nicht wenige Programme gar nicht konsequent auf Wirkungen hin geplant und durchgeführt werden (G1).

1.3 Evaluation von Politiken und Programmen

„Programm" ist ein generischer Begriff der Evaluationsfachsprache mit einer Vielzahl methodologischer Implikationen, die in der intensiven Auseinandersetzung mit der Evaluationstheorie deutlich werden.

Je nach Disziplin oder Politikfeld wird „Programm" in Deutschland mit unterschiedlichen Bedeutungen verknüpft. Intention dieses Abschnitts ist, die Reichweite des Programmbegriffs für Evaluationen im Bereich der Armuts- und Reichtumspolitik deutlich zu machen.[12]

Programme im Kontext dieser Perspektivstudie werden aufgefasst als Ausdruck staatlichen oder staatlich gerahmten Handelns, das auf die Verbesserung von Rah-

---

12 Die folgende Argumentation versucht einen Programm-Begriff zu umreißen, der für Programm-Evaluationen aller Art anwendbar ist. Er kollidiert vielfach mit disziplinär gebundenen Programm-Begriffen (etwa der Politikwissenschaft oder der Ökonomie); die begriffliche Klärung muss vorläufig bleiben und bedarf intensiver interdisziplinärer Diskussion.

menbedingungen und Behebung von Notlagen armer bzw. von Armut bedrohter Menschen bzw. ihrer sozialen Ausgrenzung gerichtet ist. Diese Programme sind verortet in verschiedensten Politikfeldern, wie sie durch den Lebenslagenansatz in der Armuts- und Reichtumsberichterstattung der Bundesregierung angesprochen werden: einerseits die auf öffentliche Transferleistungen bezogene Politik (z.B. Sozialhilfe, Wohngeld, Kindergeld), andererseits die Fachpolitiken, die für die verschiedenen Lebenslagen-Dimensionen relevant sind (z.B. Wohnungs-, Bildungs-, Gesundheitspolitik). Das Konzept „Programm" fordert, derartiges staatliches oder staatlich gerahmtes/finanziertes Handeln durch Konzepte zu konkretisieren und dabei insbesondere die angestrebten Resultate oder Wirkungen so zu beschreiben – möglichst operational – dass sie messbar werden.

Michael Scriven, der u.a. mit seinem „Evaluation Thesaurus" viele Begriffe in der Evaluationstheorie geprägt hat, charakterisiert ein Programm pointiert als „*the general effort that marshals staff and projects toward some (often poorly) defined and funded goal*".[13] Andere Autoren/-innen betonen ebenfalls die Zielgerichtetheit eines Maßnahmenbündels, welches im Zusammenspiel von Planung und Durchführung von Aktivitäten und zu erreichenden Resultaten ein Programm ausmacht.[14] Auch hier werden auf der Basis von Ressourcen, ausgerichtet auf bestimmte (im Idealfall operationalisierte) Ziele, Aktivitäten entfaltet, die zu Resultaten führen bzw. führen sollen.

Auf den ersten Blick unterscheiden sich Programme zunächst nach ihrer Größe, gemessen in eingesetzten oder umgesetzten Finanzmitteln, der Anzahl von in das Programm Einbezogenen oder von ihm Betroffenen oder des beteiligten Fachpersonals. Dabei ist die „Logik" von Programmen – setzen sie wenige Tausend oder mehrere Milliarden Euro um – grundsätzlich gleich. Auf der Ebene internationaler oder nationaler Politik bspw. bezeichnet „Programm" größere Bündel von Maßnahmen oder Teilstrategien, wie z.B. die insgesamt ca. 30 Milliarden Euro umfassenden Programme der strukturpolitischen Maßnahmen der Europäischen Union.

Der Komplexität eines Programms werden dabei nach oben Grenzen gesetzt durch den Übergang in eine *policy*. Diese bezeichnet das staatliche und parastaatliche

---

13 Scriven (1991, 285) – Programm als übliche Anstrengung, die Mitarbeitende und Projekte in Richtung eines (oft bescheiden) festgelegten und finanzierten Ziels führt.

14 Vgl. z.B. Rossi, Freeman, Lipsey (1999); Worthen, Sanders, Fitzpatrick (1997).

Handeln in einem bezüglich der Ziele lediglich grob umrissenen Politikfeld, z.B. der Sozial-, Bildungs- oder Gesundheitspolitik. In einem Politikfeld handeln Parlamente, Verwaltungen, Sozialversicherungsträger, Kammern, Innungen, Berufsverbände, Verbraucherzusammenschlüsse usw. arbeitsteilig und z.T. autonom voneinander. In einem Programm hingegen gibt es idealtypisch oberste Entscheider/-innen (dies kann auch ein demokratisch legitimiertes Gremium sein), Programmverantwortliche oder -manager, Mitarbeiter/-innen und Zielgruppen (Einzelpersonen, Unternehmen, Kommunen). Faktisch werden besonders große „Makro-Programme“ (z.B. die Strukturfondsprogramme der EU oder der Kinder- und Jugendplan des Bundes) zwar zentral budgetiert, aber in einem mehrstufigen Verfahren schließlich regional bei dezentraler Steuerung alloziiert, was ihre zielgeführte Steuerung und auch Evaluation erheblich kompliziert.

Vedung (1999, S. 124) weist darauf hin, dass einem Programm in aller Regel mehrere Programmtheorien zu Grunde liegen (die überdies meist wenig expliziert und vollständig sind), also bspw. die der politisch Entscheidenden und die der Umsetzer/-innen, letzteres z.B. durch Kommunen oder auch durch die freien Träger. Ähnlich differenziert Patton (1997, S. 221) zwischen (1) *espoused theories* (angetragene Theorien) – was die Beteiligten vorgeben oder glauben, auf welcher theoretischen Basis sie handeln und (2) *theories in use* – die konzeptionelle Basis, auf der sie tatsächlich handeln. Fallen diese Programmtheorien mit ihren Zielen und Bewertungskriterien stark auseinander, ergeben sich für die Evaluierbarkeit von Programmen ähnliche Schwierigkeiten wie für die von *policies*.[15]

Eine nicht geringe Zahl politischer Instrumente, die sich in Deutschland auf die Verminderung und Überwindung von Armut richten, liegen als Gesetze, insbesondere Leistungsgesetze, vor, und haben kaum „programmförmigen“ Charakter. D.h. dass es zwar grobe Zielvorgaben gibt, doch sind diese an keiner Stelle, auch nicht in Verordnungen oder Ausführungsbestimmungen, so weit operationalisiert, dass sie orientierend für das Handeln der Beteiligten sind. Ein prominentes Beispiel ist das Bundessozialhilfegesetz, das ursprünglich als Ausfallbürge (auf Zeit) für eine überschaubare Gruppe Betroffener gedacht war und heute für Millionen, teilweise auf

---

15 Als Vorgehensweise steht die Evaluierbarkeitsabschätzung zur Verfügung, die oft zu dem Ergebnis führen wird, zunächst im Rahmen einer formativen Evaluation einen Klärungsprozess der zu Grunde zu legenden Programmtheorie durchzuführen. Vgl. Patton (1997, S. 152f).; grundlegend vgl. Wholey/Hatry/Newcomer (1994).

mehrere Jahre, die zentrale Quelle des Lebensunterhalts darstellt. Auf Grund dieser strukturellen Funktionsverschiebung – beschleunigt durch fiskalpolitische Zuspitzungen bei Bund, Ländern und Kommunen – wird die im Bundessozialhilfegesetz lange Jahrzehnte „schlummernde“ Zielsetzung, Hilfeberechtigte durch geeignete Maßnahmen wieder unabhängig von diesem Transfereinkommen zu machen, zunehmend relevant.

Damit dies gelingt, müssen gegenwärtig auf lokaler Ebene eigenständige Programme entwickelt werden, die z.B. mit Stichworten wie „Hilfe zur Arbeit“, „Auswegberatung“, „Case Management“ angesprochen sind. Dabei ist die Verbindung zwischen bundesgesetzlicher Grundlage und lokalen Aktivitäten oft unklar, widersprüchlich und Gegenstand politischer Kontroversen.

Die beiden in Nordrhein-Westfalen durchgeführten Modellvorhaben „Pauschalierung der Sozialhilfe“ und „Sozialagenturen“ stellen den Versuch dar, für Teilelemente des mit dem BSHG geregelten staatlichen Handlungsinstrumentariums eine stärkere „Programmförmigkeit“ zu erreichen und damit auf den umsetzenden (kommunalen) Ebenen erhöhte Handlungssicherheit und verbesserte Erfolgsaussichten zu schaffen.

Mit der modellhaften Einführung der Pauschalierung von Sozialhilfe wird das Ziel verbunden, die Verwaltungsaufwendungen für Routinetätigkeiten im Rahmen der Leistungsgewährung zu reduzieren und auf diese Weise Ressourcen in den Sozialämtern für persönliche Hilfen, insbesondere Beratungsangebote, Hilfeplanung und Case-Management, freizusetzen. Die Evaluation erfolgt auf der Basis eines multiperspektivischen und multimethodischen Vorgehens. Diese wissenschaftliche Begleitung der Modellvorhaben ist so angelegt, dass im Rahmen einer klärenden und interaktiven Startphase die lokalen Modellvorhaben dabei unterstützt werden, Ziele der Maßnahmen und Projekte zu konkretisieren und auf den jeweiligen Bedarf vor Ort abgestimmte Maßnahmekonzepte zu entwickeln. Weitere Evaluationsaufgaben sind es zu überprüfen, in welchem Umfang die entwickelten Maßnahmen und Konzepte umgesetzt werden (Umsetzungsevaluation) und in welchem Maße die Modellvorhaben die gesetzten Ziele erreichen (Wirkungsevaluation) (MASQT 2001).

Das Modellvorhaben „Sozialagenturen“ verfolgt die Leitziele, Personen in prekären materiellen Lebenslagen bei der Bewältigung ihrer Lebenssituation zu unterstützen und Bewältigungsressourcen durch individuell zugeschnittene Förderangebote und aktive Beteiligung zu erschließen. Zum einen wird damit die Entwicklung eines abge-

stimmten Dienstleistungsangebotes gefördert, das an § 8 des Bundessozialhilfegesetzes „persönliche Hilfen“ anknüpft, zum anderen unterstützt die wissenschaftliche Begleitung die Verzahnung der lokalen Akteure, u.a. durch eine Steuerung der Angebotsstruktur mittels ämter- und trägerübergreifender Kooperationsformen.

In den letzten Jahren wird immer wieder versucht, die Steuerungsfähigkeit der Politik zur Armutsvermeidung dadurch zu erhöhen, dass Organisationen umstrukturiert, fusioniert oder dezentralisiert werden, um die Effektivität und Effizienz der öffentlichen Dienstleistungserstellung zu steigern. Es besteht die Gefahr, dass mit der Konzentration auf organisationale *Strukturen* die Konzepte, insbesondere Ziele des eigentlichen Dienstleistungprozesses nicht, nicht klar genug, unangemessen oder widersprüchlich gesetzt werden. Die oft komplexen Prozesse (Maßnahmebündel/ Interventionen) sind dann nur bruchstückhaft geplant und es bleibt entgegen der Intention bei bürokratischer Verwaltung von Armut und Armutsrisiken. Die Fassung des Gegenstands von Evaluationen als „Programm“ – mit Bedingungen, Konzept, Prozess und Resultaten/Wirkungen – bildet insofern ein Gegengewicht gegen die Verengung sozialpolitischer Reformen auf strukturelle Aspekte.

Im Rahmen von größeren Programmen ist es besonders wichtig, dass Teilprogramme aufeinander abgestimmt sind, was zentral über die Klärung und die Transparenz von Zielen erfolgt. So sind beispielsweise Image- und PR-Kampagnen, öffentliche Veranstaltungen, Qualifizierungsmaßnahmen, niedrigschwellige Angebote oder individuelle Beratungs- und Betreuungsangebote für sich genommen kleinere Programme, deren systematische Einordnung in übergreifende Programme erst zu Wirksamkeit und Effizienz auf gesellschaftlicher Ebene führen kann. Wie die Evaluationen mehrerer lokaler Programme/Projekte, die im Rahmen eines übergreifenden Programms durchgeführt werden, zusammengeführt werden können, wird neuerdings verstärkt unter dem Stichwort „Cluster-Evaluation“ diskutiert.[16]

Programme kommen schließlich auch auf einem Mikro-Level etwa therapeutischer oder intensivpädagogischer Prozesse zwischen zwei Personen zustande. Das *Handbuch der Evaluation psychologischer Interventionsmaßnahmen* nutzt „Programm“ als

---

16 Vgl. Sanders (1997) (und vgl. auch die im Kap. 2.4 dargestellte Studie des DJI zum Bundesmodellprogramm „Mobile Jugendsozialarbeit für junge Menschen ausländischer Herkunft“)

Synonym für „Maßnahme“ und „Intervention“ im Bereich der klinischen, der pädagogischen und der Organisations-Psychologie (Hager, Patry, Brezing, 2000). Auch für diese Bereiche unterscheiden die Autoren zwischen „flächendeckenden“ und „lokalen“ Programmen. Kennzeichnend für viele – auch große/nationale – Programme im Bereich der Armutsvermeidung und -verminderung ist, dass sie sich aus tausenden derartiger Mikro-Programme zusammensetzen, wobei die Logik der Programmplanung und -durchführung bis hin zur nationalen Ebene weitgehend gleich bleibt – ein zentraler Ansatzpunkt zur effektiven und effizienten Programmsteuerung auch auf der Makro-Ebene.

Für die Evaluation in den USA (und auch in Australien und Kanada) gilt, dass Gebrauch wie Verstehenskonsens des Begriffes *program* höher sind als in Deutschland. Dies ist vor der Kulisse der Politikmuster in diesen Ländern nicht verwunderlich. Die US-amerikanischen Regierungen nähern sich seit den 60er Jahren (im Rahmen von Kennedys *War Against Poverty* und Johnsons *Great Society*) vielen sozialen Problemen über Programme, die spezifische Zielgruppen (z.B. Minoritäten, Bewohner/-innen strukturschwacher Gebiete etc.) ansprechen und deren Weiterfinanzierung von ihrem Erfolg in Bezug auf gesetzte Ziele abhängig ist. In Deutschland war die Problembearbeitung hingegen institutionalisiert, d.h. die öffentliche Hand bzw. die von ihr beauftragten Einrichtungen (Parafisci wie Gesetzliche Renten-, Kranken- Unfallversicherungen, Knappschaft; außerdem Wohlfahrtsverbände), erbrachten die Leistungen auf Basis einer dauerhaften Finanzierung, oft in der Art eines Monopols. Ihr Aufgabenbereich war abgesteckt; ausgewiesene Ziele fehlten, waren wenig präzisiert oder wurden zumindest nicht auf ihre Erreichung überprüft. Mit „Neuer Steuerung“, der Erfordernis, Kontrakte (intern und extern) auf der Basis von Leistungs-, Entgelt- und Qualitätsvereinbarungen zu treffen,[17] wandelt sich institutionalisierte Leistungserbringung vielfach in programmförmige um.

„Programm“ wird in dieser Perspektivstudie als generischer Begriff der Evaluationsfachsprache genutzt, der von der Mikro-Ebene zielgerichteter, z.B. therapeutischer, Interventionsbündel bis zur Makro-Ebene internationaler Politikarchitekturen anwendbar ist:

---

17 So z.B. geregelt in §78a - f des Kinder- und Jugendhilfegesetzes für die stationäre und teilstationäre Erziehungshilfe.

Programme sind beschriebene und durchgeführte, intentional aufeinander bezogene Bündel von Interventionen, Maßnahmen, Projekten oder Teilprogrammen, die aus einer Folge von auf ausgewiesene Ziele hin ausgerichteten Aktivitäten bestehen, welche auf der Basis von verfügbaren Ressourcen durchgeführt werden und darauf gerichtet sind, vermittels bereitgestellter Leistungen (Outputs) bestimmte, bei bezeichneten Zielgruppen intendierte Zustände (Outcomes) oder Wirkungen oder darüber hinaus im sozialen System zu erreichende Wirkungen auszulösen. Gegenstand der Evaluation können sowohl das Konzept des Programms als auch die Durchführung des Programms (Prozess) und seine Resultate sein.

Ein Programm besteht also aus einem fixierten (verschriftlichten) Plan oder Entwurf *und* dessen Umsetzung in politische Praxis oder staatliches Handeln. Im Programmentwurf sind die Ziele, Vorgehensweisen usf. der Praxis konkret vorweggedacht.

Ein elaboriertes Verfahren, die Programmplanung mit der Programmevaluation zu verschränken, ist das Konzept der „Programmtheorie“ (Chen 1994; Weiss 1998, S. 55ff; Rossi, Freeman/Lipsey 1999, S. 98ff; Patton 1997, S. 215ff). im Kern sind dies logisch miteinander verknüpfte, begründete Annahmen darüber, in welcher gedachten „einfachen“ (chronologisch) angeordneten Verkettung bestimmte Programminterventionen bzw. Zwischenresultate dazu beitragen, die einzelnen Programmziele zu erreichen. Gleichzeitig werden relevante Bestandteile des Programmkontextes dargestellt (vgl. auch das Modell der Programmtheorie-gesteuerten Evaluation in Kap. 8.6).

1.4 Resultate und Wirkungen als Fokus der Programmevaluation

„Wirkung“ soll hier als besondere Form von Programmresultaten aufgefasst werden, also Veränderungen (oder Stabilisierungen)[18], die festzustellen sind, nachdem ein Programm oder eine Phase eines Programms stattgefunden hat. Resultate qualifizieren sich dann als Wirkungen, wenn sie auf das Programm „ursächlich“ rückführbar sind, also nachvollziehbare theoretische Annahmen sowie empirische Daten über die Verbindung von Maßnahmen/Aktivitäten *und* Resultaten beigebracht werden

18 Es ist eine verbreitete Figur, im Rahmen z.B. von sozialpolitischen Programmen bei angestrebten Resultaten/Wirkungen stets von „Veränderungen“ zu sprechen, auch um öffentliche Legitimation für solches „innovatives“ staatliches Handeln zu erhöhen. Dabei kommt es – gerade auch in Feldern der Armuts- und Reichtumspolitik – vielfach auf Stabilisierungen von Lebenslagen an, z.B. von chronisch kranken, von alten, von behinderten Menschen. Dies halten wir in diesem Text dadurch fest, dass wir zumindest gelegentlich den Stabilisierungstopos erwähnen.

(vgl. ausführlicher Kromrey 2000). In diesem Fall sprechen wir von einem „empirischen Wirkungsnachweis“.[19]

Entgegen verbreiteter Annahmen sind Evaluationen nicht ausschließlich auf den Nachweis von Wirkungen verpflichtet. So konzentrieren sich z.B. Programmziel-gesteuerte Evaluationen (vgl. Kap. 2.3) darauf, Zielerreichung (Effektivität) festzuhalten. Es wird empirisch überprüft, ob die angezielten Resultate tatsächlich vorliegen. Ein empirischer Ursachennachweis ist in diesem Evaluationsmodell nicht verpflichtend. Auch das *Outcome-Measurement* konzentriert sich auf die Frage, ob und in welchem Maße Veränderungen bei Zielgruppen eingetreten sind. Empirisch zu belegen, dass man von einer Outcome-Wirkung sprechen kann, wäre eine zusätzliche, nicht eine Pflicht-Leistung im Rahmen dieses Evaluationskonzeptes (Plantz/Greenway/Hendricks 1997). Dabei werden zunehmend Ansätze weiterentwickelt, Wirkungen von Programmen in theoretische/logische Argumentationsmuster einzubetten (Programmtheorie/logisches Modell) und so argumentativ plausibel zu machen, dass die gemessenen Resultate auf die dokumentierte Programmdurchführung rückführbar sind. Dies wird hier als „Wirkungsmodellierung“ bezeichnet.[20]

Auch beschränkt sich in einer wirkungsorientierten Evaluation die Perspektive nicht zwangsläufig auf solche Wirkungen, die durch Programmziele beschrieben sind. Dieses sind die so genannten „intendierten Wirkungen“. Hingegen betrachten bestimmte Evaluationsansätze auch (wie die zielfreie Evaluation in besonderem Maße) nicht-intendierte Wirkungen von Programmen, die als Nebenresultate bezeichnet werden oder – noch allgemeiner – als Folgen.[21] Oft wird in diesem Zusammenhang auch von ‚Nebenwirkungen’ gesprochen, woran jedoch – ursächlicher Zusammenhang zu Programm muss belegt sein – höhere Anforderungen zu richten wären. Wir sprechen daher von Nebenresultaten. Das Bemühen darum, nicht vorhergesehene Resultate aufzufinden wird hier als „Wirkungsidentifizierung“ bezeichnet.

---

19 Die Bezeichnung „Wirkungskontrolle“ hingegen wird wegen ihrer Mehrdeutigkeit in dieser Perspektivstudie nicht verwandt. Im nachfolgenden Kap. 1.5 wird diese Entscheidung ausführlicher begründet.

20 Ein der naturwissenschaftlichen Forschungslogik entlehnter Ursachenbeweis wie er z.B. bei agrarwissenschaftlich-biologischen Versuchsanlagen oder in der medizinischen Medikamentenwirkungsforschung Standard ist, unterbleibt hier aufgrund der Annahme/Erfahrung, dass er bei personenbezogenen Dienstleistungen nicht führbar ist.

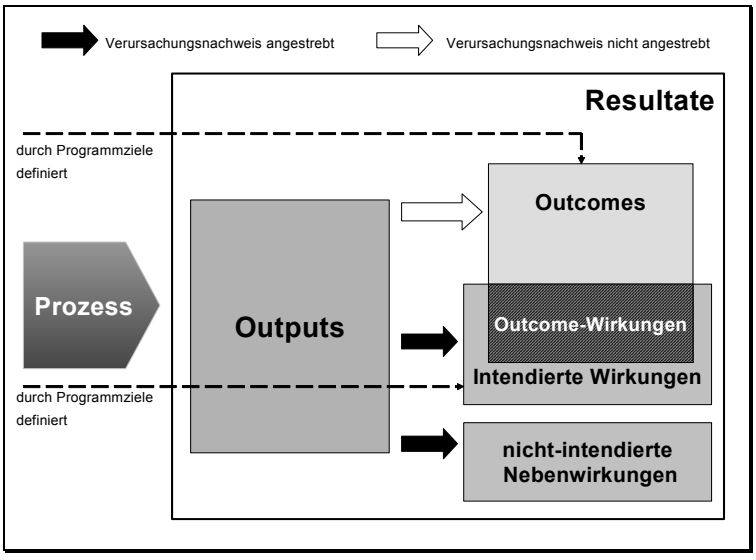
21 Vgl. eine ausführlichere Darstellung – mit leicht abweichender Terminologie – bei Klöti (1997).

Schließlich können Evaluationen sich auf die Beschreibung und Bewertung der Programmumsetzung (Prozessevaluation) konzentrieren, was besonders bei Modellprogrammen angemessen ist oder bei sich vergleichsweise rasch und gravierend ändernden Programmkontexten (wie z.B. nach einem historischen Ereignis wie der deutschen Vereinigung). Andere Gründe sind, dass ein Programm noch zu jung oder instabil ist oder dass es primärer Evaluationszweck ist, kurzfristig gesicherte Hinweise zu geben, wie anlaufende Programme optimiert werden können.

„Wirkungsorientierte“ Evaluationen stellen Auftraggebenden, Programmverantwortlichen oder anderen wichtigen Beteiligten solche Informationen zur Verfügung, die sie nutzen können, um die Programmwirkungen fundiert zu planen, sie auf Basis der Rückmeldungen aus der Evaluation zu optimieren, um sich von der Wirksamkeit der Programmaktivitäten zu überzeugen oder um sowohl die negativen wie die positiven Auswirkungen abzuschätzen.

Auch die Überprüfung der Zielerreichung von Programmen (u.a. *Outcome-Measurement*) und die auf die Optimierung von Prozessen fokussierten Evaluationen können einen Beitrag zur wirkungsorientierten Evaluation leisten.

Abbildung 2: Resultate und Wirkungen



Quelle: eigene Darstellung

Die Abbildung 2 veranschaulicht, welche Schwerpunkte eine Evaluation – über die Beschreibung und Bewertung des Programmprozesses hinaus – setzen kann und welche Arten von Resultaten in den Mittelpunkt der Untersuchungen gestellt werden

können.[22] Es wird dabei deutlich, dass es nicht in jedem Falle um den Nachweis von Wirkungen geht:

- **Outputs** im Sinne von produzierten Leistungen eines Programms bilden den Schwerpunkt in dokumentierenden Evaluationen, die dem Programm-Monitoring nahe stehen. Es wird festgehalten, wie viele Einheiten eines bestimmten Outputs (z.B. Beratungsstunden, Trainingstage, Geld- oder Sachleistungen) mit den Programminputs bereitgestellt werden. Eine andere Form von Outputs sind z.B. realisierte Klientenzahlen oder gezählte Erstkontakte bei niedrigschwelligen Angeboten. Die Grenze zu Outcomes ist fließend, insofern z.B. Personen ohne festen Wohnsitz sich in Übergangswohnheimen nicht nur „aufhalten“ (Output) sondern darüber hinaus auch gesundheitlich stabilisiert werden, ruhig schlafen, sicher sind vor Angriffen und vieles mehr (= Outcomes).
- **Outcomes** werden durch Evaluationen fokussiert, welche die bei Zielgruppen auftretenden Veränderungen/Stabilisierungen z.B. in Wissen, Einstellung, Verhalten oder Status als zentral erachten. Typische Outcomes sind vermehrtes Wissen (etwa über gesetzlich verbürgte Ansprüche als Unterhaltsberechtigte), veränderte Einstellung (bspw. zur Erwerbs- oder Familienarbeit), verändertes Verhalten (wie im Umgang mit begrenzten finanziellen Mitteln) oder bei Bewerbungen um Erwerbsarbeitsplätze. Gewinn von Sicherheit, z.B. durch Eintritt in die Grundsicherung oder in eine unbefristete, sozialversicherungspflichtige Beschäftigung sind Beispiele für Outcomes im Sinne der Statusverbesserungen. Üblicherweise werden die anzustrebenden Outcomes im Sinne von Zielen des Programms als wünschenswert ausgewiesen (im Programm-Konzept).
- **Intendierte Wirkungen** sind dann besonders relevant, wenn der Nachweis gefordert ist, dass das Programm in einem quantifizierbaren Maße verursachend ist für die festgestellten Veränderungen/Stabilisierungen. Wenn die im Zielsystem des Programms bezeichneten Outcomes ausgelöst sind und deren Entstehen nachweislich auf die Programmaktivitäten rückführbar ist, bezeichnen wir dies als Outcome-Wirkungen.

---

22 Die Resultatsart „Impact“ wird hier aus Vereinfachungsgründen nicht dargestellt; vgl. Abb.3.

Sie sind der Teil der erreichten Ziele, für die in der Evaluation überzeugend dargelegt ist, dass ihr Erreichen auf die Programmumsetzung zurückgeht.

Weitere intendierte Wirkungen gehen über die Zielgruppe hinaus, z.B. indem das soziale Klima eines Sozialraums friedlicher ist, sein Image verbessert ist, seine soziale Durchmischung breiter ist. Im Unterschied zur Literatur, die solche Wirkungen teilweise auch als Outcomes kennzeichnet, sprechen wir hier allgemein von intendierten Wirkungen. Der Terminus Outcomes bleibt reserviert für Veränderungen/Stabilisierungen bei den Zielgruppen von Programmen.

- **Nicht-intendierte Wirkungen** sind im Unterschied zu allen voranstehenden Resultatsarten nicht im Zielsystem des Programms enthalten. Sie sind dann von herausragendem Interesse, wenn in einer großen Wirkungsbandbreite ein Resümee zur sozialen/ökonomischen Wertigkeit eines Programms gezogen oder wenn Programmalternativen umfassend verglichen werden sollen. Nicht-intendierte Wirkungen können sowohl zu einer negativeren sowie zu einer positiveren Bewertung des Programms führen. Es kommt darauf an, ob sie nachträglich als Unterstützung der Programmziele oder als konkurrierend mit diesen eingeschätzt werden, oder ob sie außerhalb des Zielsystems als Beeinträchtigung oder als Erhöhung des Programmwertes insgesamt interpretiert werden. Ein bekanntes Beispiel für eine negative Nebenwirkung ist die Einführung eines (zu) niedrigen Dosenpfandes in den USA, mit der die Verschmutzung von Strassen und Nachbarschaften durch Dosenmüll reduziert werden sollte. Daraufhin wurden die Mülleimer von den ganz Armen nach Pfanddosen durchsucht und dabei zum Teil ganz ausgeleert, was genau den gegenteiligen Effekt in Bezug auf die Programmziele hatte: Die Verschmutzung der Strassen und Nachbarschaften nahm zu. Auf der anderen Seite gibt es auch nachträglich positiv gewertete Nebenresultate, die bisweilen sehr lange, schwer nachweisbare Wirkungsketten aufweisen. Neben den methodischen Schwierigkeiten, systematisch Nebenwirkungen zu identifizieren, ggf. auch mit einem Vollständigkeitsanspruch, kommt ein zweites Problem hinzu: Wenn vorab nicht geklärt

werden kann, ob (da ja noch unbekannt) Nebenwirkungen als positiv oder negativ bezeichnet werden, kann in der Phase der Berichterstellung noch einmal erheblicher Streit zwischen den Programmbeteiligten aufkommen, wie Nebenwirkungen einzuordnen sind und wie diese in eine Gesamtbewertung des Programms eingehen. Dies ist sicher einer der Gründe dafür, dass derartige „zielfreie Evaluationen“ vergleichsweise selten umgesetzt werden.

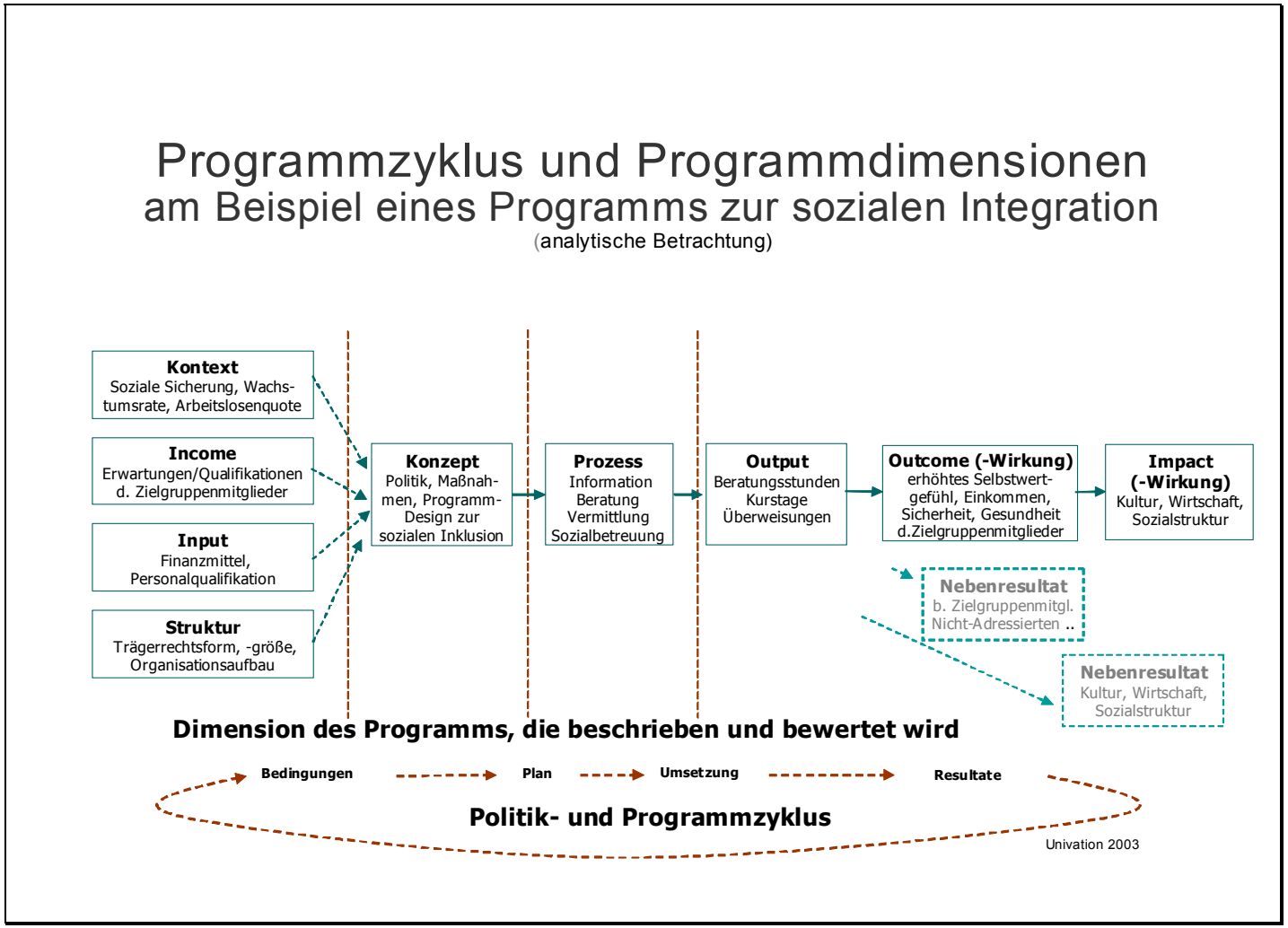
Ausdrücklich sei empfohlen, sich bei der Entwicklung von Evaluationsdesigns auf eine oder wenige dieser Resultatsarten zu konzentrieren. Evaluationen können auch dann schwer durchführbar sein und ggf. die vorgegebenen Zeitlinien verfehlen, wenn zu umfangreich neben einer resultatsbezogenen Evaluation auch die Programmprozesse oder z.B. das breite Feld seines sozioökonomischen oder organisatorischen Kontextes thematisiert werden. Solche „umfassenden Evaluationen“ (*comprehensive evaluations*) überfordern nicht nur die ihnen zur Verfügung stehenden Budgets, sondern auch die Untersuchungslogistik. Sie führen zu bruchstückhaften Ergebnissen, die nachträglich „vom grünen Tisch“ aus argumentativ verbunden werden müssen. Die Nachvollziehbarkeit der gezogenen Schlussfolgerungen im Hinblick auf die empirische Basis wird oft nicht plausibel herstellbar sein. Die Glaubwürdigkeit der Evaluationsergebnisse steht dann im Zweifel.

Im Zuge der Evaluation von Maßnahmen zur Armutsvermeidung oder -verhinderung sind Outcomes bzw. Outcome-Wirkungen von besonderer Bedeutung: Die Armuts- und Reichtumsberichterstattung hat dadurch eine besondere Erweiterung erfahren, dass sie ihre Untersuchungsinstrumentarien auf die Lebenslagen ausrichtet, in der sich die Zielgruppen von solchen Programmen befinden, die im Rahmen von Evaluationen entworfen, verbessert oder beurteilt werden sollen. Gleichzeitig sollte nicht gänzlich vernachlässigt werden, welche Nebenresultate auftreten. Hierzu eigenen sich ggf. gezielte Interviews mit Experten und Expertinnen, die Stakeholder mit verschiedenen und ggf. konfligierenden Interessen in Bezug auf das Programm vertreten. Oftmals wird sich aber herausstellen, dass sich hieraus im Wesentlichen Einschätzungen und Meinungen über Nebenwirkungen ergeben und nur selten ein genügend sicherer empirischer Nachweis erbracht werden kann. Jedenfalls können solche (Neben-) „Wirkungseinschätzungen“ als begrenzende und relativierende Aspekte in Gesamtbewertungen von Programmen einbezogen werden, die sich im

Schwerpunkt auf die Outcomes, d.h. die Verbesserungen bzw. Stabilisierungen bei den Zielgruppen von Programmen zur Armutsvermeidung oder -verminderung feststellen lassen.

Die nachfolgende Abbildung 3 gibt – veranschaulicht am Beispiel eines Programms zur sozialen Integration – einen Überblick über die möglichen Schwerpunkte, die Datenerhebungen und -auswertungen in Bezug auf die aufgeführten Dimensionen eines Programms setzen können.

Abbildung 3: Programmzyklus und Programmdimensionen am Beispiel sozialer Integration



Quelle: eigene Darstellung

Der vereinfachte Programmzyklus gliedert sich in die vier Hauptteile „Bedingungen“, „Plan“, „Umsetzung“ und „Resultate“, woraufhin der Zyklus erneut durchlaufen werden kann. Die differenzierte Darstellung von (a) Bedingungsfeldern, (b) Konzept, (c) Prozess und (d) Resultatsfeldern dient drei möglichen Zwecken:[23]

- In der proaktiven, klärenden und interaktiven Evaluation wird sie als Steuerungs- und Navigationshilfe zur Planung und Justierung des Programms eingesetzt.
- In der dokumentierenden Evaluation dient sie als Orientierungsraster, um die Evaluationsfragestellungen zu verorten; oft fokussieren diese Input, Prozess und Outcomes.
- In der wirkungsfeststellenden Evaluation gliedert sie einerseits die Resultatsarten, die als Wirkungen nachgewiesen werden sollen, und liefert gleichzeitig ein Modell für die Zuschreibung von Ursachen (Prozess, Rahmenbedingungen) zu Resultaten.

Es lassen sich vier **Bedingungsfelder** unterscheiden, die sowohl Randbedingungen für die Programmumsetzung und den Programmerfolg darstellen, als auch bei der Programmplanung zu beachten sind:

- Der **Kontext** umfasst die Umgebungsbedingungen eines Programms. Je nach Reichweite des Programms können diese auf lokaler, nationaler oder internationaler Ebene liegen und soziale, ökologische, politische und kulturelle Aspekte betreffen, die sich nur langfristig und unabhängig vom Programm selbst ändern. In gewissem Umfang kann die Gesetzgebung auf den Kontext einwirken. Bis zum Eintreten der intendierten Veränderungen vergehen in aller Regel längere Zeiträume. Beispiele sind Wirtschaftswachstum, soziale Schichtung, Arbeitslosenquote oder öffent-

23 Diese Darstellung kombiniert das CIPP- Modell von Stufflebeam (1972) (Context-Input-Process-Product) sowie das seit den späten 80er Jahren in Deutschland diskutierte und im SGB aufgenommene Struktur-Prozess-Ergebnis-Modell, das von Maja Heiner (2001, S. 43) um die Konzept-Dimension erweitert worden ist. Wir sprechen von „Resultaten“ und nicht „Ergebnissen“, um der immer wieder aufkommenden Verwechslung mit Evaluationsergebnissen (i.S.v. Befunden) entgegenzuwirken.

liche Meinung. Identische Programmkonzepte führen bei stark differenten Kontexten zu unterschiedlichen und ggf. gegensätzlichen Resultaten.[24]

- Der **Income** umfasst, was die Zielgruppenmitglieder an Voraussetzungen in das Programm „einbringen“, insbesondere an Wissen, Einstellungen, Bedarfen, Werten etc. So gibt es z.B. starke Unterschiede im Income bezüglich des Alters, des Bildungsstandes oder des Geschlechtes der potentiellen Programmteilnehmenden. Identische Programmkonzepte können auf Zielgruppenmitglieder mit unterschiedlichen „Incomes“ unterschiedlich und auch gegensätzlich wirken. Dies betrifft z.B. die Motivation zur Teilnahme an einer Maßnahme, etwa von Aufstiegsorientierten vs. Statusindifferenten. Mittels geeigneter Strategien, z.B. der Teilnehmendenauswahl (Profiling), kann der Income eines Programms in gewissem Maße gesteuert werden. Dies ist jedoch für schwer messbare Teilnehmendenmerkmale wie Akzeptanz oder Motivation mit hohen Unsicherheiten behaftet. Da es sich bei einer Vielzahl der Programme im Umfeld der Armuts- und Reichtumsberichterstattung um Programme vom Typus „personenbezogene Dienstleistungen“ handelt[25], kommt dieser Dimension häufig besondere Bedeutung zu (Spiegelfeld des „Income“ bei den Resultaten ist der „Outcome“). [26]
- Der **Input** bezeichnet finanzielle, personale oder andere Ressourcen, die in ein Programm investiert werden. Diese stellen eine vergleichsweise variable Bedingung dar, insofern z.B. Kostenarten (Personal- vs. Sachausgaben) oder Personalqualifikationen (durch gezielte Rekrutierung oder Fortbildung) beeinflusst werden können. Die Input-Bedingungen können standardmäßig im Rahmen des Programm-Controlling oder des Monitoring erfasst und möglicherweise vollständig in Geldwerten dargestellt werden (Spiegelfeld zum Input bei den Resultaten ist der Output).

---

24 Hierauf hebt das Kontext-Mechanismus-gesteuerte Evaluationsmodell ab (vgl. Kap. 2.3.1.5).

25 Stichworte: Ko-Produktion, uno-actu-Prinzip, Wertebestimmtheit (vgl. Haller 1998, Beywl 1999, Müller-Kohlenberg/Münstermann 2000).

26 Selbstorganisationsgesteuerte Evaluationsmodelle setzen mit Schwerpunkt bei den Income-Bedingungen an, die z.B. im Rahmen eines Empowerment im Durchlauf des Programms gestärkt werden sollen (vgl. Kap. 2.3.1.5).

Kosten-Nutzungsgesteuerte Evaluationsmodelle rekurrieren zentral auf den Input (vgl. Kap. 2.3.4)

- **Struktur** meint Bedingungen, die als Merkmale des Trägers eines Programms vorliegen, wie z.B. seine Rechtsform, seine Kapitalausstattung und sein Finanzierungsmix, seine Personalstruktur, Qualitätsmanagementsystem usf.. Die mittelfristig veränderbare Struktur ist eine besonders relevante Durchführungsbedingung: Wenn staatliche Programme auf schlecht/unter Zeitdruck vorbereitete/angepasste Strukturen z.B. auf kommunaler Ebene treffen, kommt es zu erheblichen Umsetzungsproblemen. Geldgeber können bei Modellprogrammen durch eine gezielte Trägerauswahl bei der Vergabe von Programmdurchführungen in begrenztem Umfang Einfluss auf dieses Passungsproblem nehmen. Bei der Regelumsetzung von Gesetzen hingegen sind vielfach die Körperschaften im föderalen Aufbau der Bundesrepublik Deutschland pflichtgemäß zuständig und mit der Leistungserbringung zu betrauen. Dann kommt der Abstimmung der verschiedenen beteiligten Ebenen eine herausragende Rolle zu, was durch begleitende Evaluationen unterstützt werden kann.[27]

- Das **Konzept** enthält die Festlegungen von Auftraggebenden, Programmplanenden/-verantwortlichen darüber, was das Programm bis wann bei welchen Zielgruppen in welchen Kontexten auslösen soll (Zielsetzungen), welche Aktivitäten zur Zielerreichung eingesetzt werden sollen (Interventionsplanung) und wie der Programmprozess insgesamt gesteuert und überwacht werden soll (Qualitätssicherung/-management). Wünschenswert ist, dass Ziele, die in Zielsysteme eingebettet sind, leitend sind für die Entwicklung eines Konzeptes. Diese Ziele müssen bestimmten formalen und fachlichen Anforderungen genügen, damit sie orientierend sind für die Programmplanung insgesamt und damit sie eine Grundlage für die Entwicklung von Bewertungskriterien für das Programm bieten (insbesondere in der zielgeführten Evaluation). Derartige Zielsysteme sind

27 Das Modell der entscheidungsgesteuerten Evaluation nimmt die Struktur-Bedingungen einer Programmlandschaft/von Programmträgern als Orientierungspunkt, um möglichst ummittelbar Medikamentenwirkungsforschung Standard ist, unterbleibt hier aufgrund der Annahme/Erfahrung, dass er bei personenbezogenen Dienstleistungen nicht führbar ist.

über mehrere Programmebenen hinweg – z.B. von der europäischen über die Bundes-, die Landes- bis zur lokalen Ebene – weiter auszudifferenzieren, bis sie auf der Maßnahmenebene in weitgehend operationale Zielbeschreibungen einmünden. Dies ermöglicht auch eine systematische Entwicklung von Interventionen, von denen erwartet wird, dass sie einen maßgeblichen und nachhaltigen Beitrag zur Zielerreichung leisten. Hierbei sind fachliche Annahmen über die Zusammenhänge zwischen bestimmten Interventionen und bestimmten Zielen leitend, wie sie z.B. aus der sozialen Arbeit, der Sozialmedizin oder aus anderen Professionen stammen, die sich beruflich mit der Thematik der Armutsbewältigung beschäftigen. Kommt es zu einer starken Explikation von Zielen, zugeordneten Interventionen und dahinter stehenden fachlichen Annahmen, kann von einer Programmtheorie oder einem logischen Modell gesprochen werden. Hierzu hat sich ein eigenständiger Evaluationsansatz entwickelt. Nicht selten sind Konzepte auch bei groß angelegten Programmen nur rudimentär entwickelt, was sowohl die Programmsteuerung und -überwachung als auch die Programmevaluation erschwert.[28]

- Mit dem **Prozess** des Programms ist die Durchführung der Interventionen/ Maßnahmen gemeint, die im Konzept zur Zielerreichung vorgesehen sind. Innerhalb von personenbezogenen Dienstleistungsprogrammen finden hier die zahlreichen Kontakte und Handlungen statt, in denen Programmpersonal und Zielgruppenmitglieder zusammen an der Erstellung der Dienstleistung arbeiten (Ko-Produktion). Es ist weitgehend anerkannt, dass gewünschte Veränderungen/Stabilisierungen bei Zielgruppen sich dann ergeben, wenn diese mit den gesetzten Zielen zumindest teilweise übereinstimmen (Akzeptanz) und wenn ihre Ressourcen systematisch in die Prozessgestaltung einbezogen werden. Es ist daher fehlleitend, bei Zielpersonen von Programmen zur Vermeidung oder Verminderung von Armut von „Kunden“ oder „Konsumenten“ zu sprechen. Vielmehr handelt es sich – soll er denn gelingen – um aktiv Mitgestaltende des Prozesses.

---

28 Der programmtheoriegesteuerte Evaluationsansatz (vgl. Kap. 2.3.1.6) hält Instrumente und Vorgehensweisen bereit, um die Konzeptklärung zu unterstützen und damit Programme auch besser „evaluierbar“ zu machen.

In manchen Evaluationsansätzen führt dies dazu, dass die Zielgruppenmitglieder bzw. deren Vertreter auch zu Mit-Gestaltern der Evaluation werden. Dem unterliegt die Annahme, dass gültige Informationen gerade in Evaluationsfeldern mit hohen Wertspannungen den Einbezug der „Betroffenen“ erfordern. In professionellen Systemen etwa der Berufsberatung, der Beschäftigungsförderung oder der Sozialen Arbeit bestehen bezüglich der Prozesse routinisierte Handlungsabläufe, die im Rahmen von Programmumsetzungen immer wieder auf die Bedingungsfelder und Zielsetzungen rückzubinden sind.[29]

- Der **Output** bezeichnet sämtliche Leistungen wie Materialien, Waren, Aktivitäten, Publikationen und insbesondere Dienstleistungen, die durch den Programmprozess (eine Kampagne, ein Projekt etc.) direkt produziert werden, wie z.B. Broschüren, Profilings, Assessments, Qualifizierungskurse, Beratungsgespräche, Leistungsstunden, Hilfepläne, Vermittlungen, Auszahlungsfälle (Buchungen) von Transferleistungen etc. Controlling und Monitoring sind bewährte Verfahren, um Systeme zur Messung von Outputs bereit zu stellen. Übersehen wird von Vertretern derartiger Verfahren oft, dass die Erfassung und Quantifizierung von anderen Resultaten im Sinne von Outcomes oder Outcome-Wirkungen weitaus komplexer sind und sich z.B. über Kennzahlsysteme nicht hinreichend darstellen lassen. Bisweilen ist die Abgrenzung zu solchen weiter führenden „Resultaten“ schwierig, wie im Beispiel der Vermittlung: Gilt diese bereits als intendierte Wirkung oder muss z.B. nachgewiesen sein, dass diese Vermittlung auch nachhaltig erfolgte (bestimmt etwa über die Mindestdauer des Verbleibs am vermittelten Arbeitsplatz). Im Rahmen des Programmablaufs wird bereits in der Festlegung des Programm-Konzepts entschieden, bei welchen später eintretenden Ereignissen von „Outputs“ oder „Outcomes“ gesprochen wird. Eine kritische Bewertung von Konzepten auf dem Hintergrund fachlicher Theorien und Annahmen ermöglicht es auch, die Plausibilität der Zurechnung zu Outputs einerseits und zu Outcomes andererseits zu überprüfen.

---

29 Dialoggesteuerte, Spannungsthemen-gesteuerte und Stakeholder-Interessen-gesteuerte Evaluationsmodelle haben hier besondere Stärken.

- **Outcome** umfasst die Resultate der Interventionen/Aktivitäten eines Programms, wie veränderte Einstellungen/verändertes Verhalten bei Zielgruppenmitgliedern oder Vorteile für die Zielgruppen (z.B. erhöhtes Haushalteinkommen). Von der Konzeptebene betrachtet, handelt es sich bei Outcomes um angestrebte Zustände, in denen sich Mitglieder der Zielgruppen des Programms nach Durchführung der Programmaktivitäten befinden sollen. Die Feststellung von Outcomes erfordert gewöhnlich einen relativ hohen Aufwand bei der Datenerhebung, besonders wenn die Zielgruppen räumlich stark verstreut sind. Erfahrungen zeigen, dass auch die Zielgruppen von Programmen der Armutsvermeidung und -verminderung Befürchtungen gegenüber Datenerhebungen hegen, ggf. die Teilnahme an diesen verweigern oder gar falsche Auskünfte geben, um sich vor (vermeintlichen) Nachteilen durch unbedachte Aussagen zu schützen. Während ein Outcome wie „freie Verfügbarkeit von (Transfer-) Einkommensbestandteilen von mindestens X Euro pro Monat und Haushaltsmitglied“ vergleichsweise leicht zu messen ist, sind Veränderungen oder Stabilisierungen in den Lebenslagen-Dimensionen wie Gesundheit oder Bildung nur mit auf die jeweiligen Programmziele speziell zugeschnittenen und entwickelten Erhebungsinstrumentarien messbar. Von Outcome-Wirkungen spricht man, wenn die Outcomes durch Anwendung eines entsprechenden theoretischen Wirkungsmodells und eines darauf abgestimmten Untersuchungsdesigns nachvollziehbar auf das Programm zurückgeführt werden.

- **Wirkungen** eines Programms sind Veränderungen oder Stabilisierungen bei den Zielgruppen, in deren persönlichem Umfeld, ihrer Nachbarschaft, im Stadtteil, in der Stadt oder in der Gesellschaft, die ursächlich auf das Programm zurückgeführt werden können. Im Vordergrund der Wirkungsfrage stehen üblicherweise die vorher, in der Phase der Programmkonzipierung, als wünschenswert ausgewiesenen Wirkungen. Der Wirkungsbegriff sollte jedoch nicht auf diese erwünschten Wirkungen beschränkt werden; damit würde verstellt, dass es auch unerwünschte Wirkungen gibt. Im Rahmen der Zwecksetzung bzw. der Fragestellungen einer Evaluation sollte klar ausgewiesen sein, in welchem Maße erwünschte oder auch unerwünschte, vorhergesehene oder auch nicht-vorher-

gesehene Wirkungen untersucht werden sollen. Alle diese Typen qualifizieren sich dann als „Wirkungen“, soweit diese durch theoretische Annahmen und/oder entsprechende Erhebungsdesigns (vgl. die experimentellen und quasi-experimentellen Evaluationsmodelle) auf die Programminterventionen zurückgeführt sind. Wenn Wirkungen im Konzept des Programms nicht vorhergesehen sind, spricht man auch von **Nebenwirkungen**. Auch diese können – nachträglich bewertet – sowohl erwünscht als auch unerwünscht sein.

Die Abbildung 4 stellt die dynamischen Programmdimensionen dar. Die beiden Rahmenbedingungen „Struktur“ und „Kontext“ sind nicht aufgeführt; sie werden als weitestgehend statisch vorausgesetzt. Im Rahmen des „Konzeptes“ des Programms wird die konkrete Verkettung der verbleibenden – nachfolgenden exemplarisch dargestellten – Programmdimensionen geplant.

Abbildung 4: Programme in der Armutsbekämpfung als Ereignisketten

Programmdimensionen		Beispiele für zu messende Merkmale
Impact/Gesamtwirkung:	**9**	Intendierte soziale, kulturelle und ökonomische Einwirkungen auf Kommune, Region u. Gesellschaft
Outcome III: Lebenslage und Status	**8**	Bewirkte personale, berufliche, soziale Position / Integration der Zielgruppen-Mitglieder
Outcome II: Handeln und Verhalten	**7**	Übernahme neuer✷ Handlungsweisen und Zeigen veränderten✷ Verhaltens bei Zielgruppen
Outcome I: Wissen, Einstellungen, Werte, Fertigkeiten	**6**	Ausgelöste kognitive oder affektive Veränderungen✷/Kompetenzen bei Zielgruppen
Output III: Reaktionen	**5**	Einschätzung des Programms (Interesse, Zufriedenheit, Stärken, Schwächen) durch Teilnehmende
Output II: Teilnahme	**4**	Anzahl, Dauer, Intensität v. Zielgruppen sowie Passgenauigkeit (bzgl. Lebenslage/Ressourcen)
Output I: Aktivitäten	**3**	Umsetzungsgrad des Programms (Anzahl Maßnahmen, Praktika, Kurse, Leistungsstunden ...)
Income: Ressourcen/Kompetenzen der TN	**2**	Fähigkeiten zum Disponieren mit Geld und Zeit, Arbeits- und Selbsthilfevermögen, Selbstwertgefühl
Input: Geld, Personal, Zeit	**1**	Aufgewendete Finanzmittel, Anzahl & Qualifikationen Personal, Zeitverbrauch

✷ Je nach Ziel muss nicht immer „Veränderung“ bewirkt oder „Neues“ hervorgebracht werden; es kann auch um die Stabilisierung von Erreichtem und die Verhinderung von Verschlechterung gehen.

W. Beywl, Univation 2003

Zur Veranschaulichung sind die Programmdimensionen als Glieder einer chronologisch angeordneten Ereigniskette aufgeführt (linke Seite der Abbildung). Auf der rechten Seite werden Beispiele für zu messende Merkmale gegeben. Die Abbildung macht insgesamt deutlich, dass für jedes Glied der „Ereigniskette“ passende Messinstrumente zu entwickeln sind. In aller Regel werden sich Evaluationen auf eine begrenzte Zahl dieser Kettenglieder beschränken und anderen weniger Aufmerksamkeit widmen.

Es ist wünschenswert, dass Programmkonzeptionen im Bereich der Vermeidung und Verminderung von Armut alle neu aufgeführten Teilglieder konzeptionell konkretisieren und deren logische Verknüpfung explizieren. Wird dies gründlich und umfassend geleistet, ist die Grundlage für eine Programmtheorie-gesteuerte Evaluation geschaffen. Auch für andere Evaluationsmodelle ist es nützlich, zu verorten, welche der aufgeführten neun Glieder der Ereigniskette durch die Evaluation fokussiert werden und zu welchen empirische Datenerhebungen durchgeführt werden.

Die Benennung als Glieder einer „Ereigniskette“ soll den Anspruch eines ursächlichen Wirkungsnachweises absenken. Es geht nicht in jedem Fall darum, die Verursachung des nachfolgenden Gliedes durch das vorangehende empirisch zu „beweisen“. Vielmehr ist es eine Frage der Wahl, ob dies über begründete Annahmen, plausible Modelle oder auf Ursachen nachweisende Untersuchungsdesigns geführt wird.

1. *Inputs* im Sinne von Geld, Personal und Zeit, die in das Programm investiert werden, sollten im Rahmen öffentlich geförderter Programme pflichtgemäß durch ein geeignetes Kostenrechnungs-/Controllingsystem festgehalten und ausgewiesen werden. Allerdings erfordert es erhebliche Anstrengungen sowohl auf Seite des Geldgebers, des Programmträgers und der Programmumsetzer und ist in dem Maße zu rechtfertigen, wie es die Durchführung der Programme fordert und nicht durch bürokratische Lasten über Bedarf einschränkt.

2. Die Messung der *Incomes* im Sinne der Ressourcen und Kompetenzen von Zielgruppenmitgliedern wird in manchen Fällen über repräsentative Stichprobenziehung möglich sein. Wenn jedoch lokale Programme spezifisch ausgestaltet werden und unter unterschiedlichen Kontextbedingungen ablaufen, wird es erforderlich sein, die lokalen Programmträger mit der Erhebung von

Income-Merkmalen zu beauftragen. Effizient durchführbar ist es dann, wenn es in die Programmdurchführung integriert wird, z.B. in Form einer Automatisierung der Erfassung von Teilnehmendendaten. Sehr bald gibt es Grenzen einer bundesweiten Standardisierung. Eine andere Grenze besteht in datenschutzrechtlichen Bestimmungen.

3. *Outputs* I im Sinne von durchgeführten *Aktivitäten* sind heute oftmals wichtigste Grundlage für die Rechenschaftslegung bezüglich einer sachlich angemessenen Verwendung der eingesetzten Mittel („Unterschriftenlisten“ der Teilnehmenden). Im Rahmen eines Programm-Monitoring werden diese heute oft standardmäßig erhoben; bei hoher lokaler Differenziertheit von Programmen ist es wiederum schwierig, derartige Outputdaten regional und national zu aggregieren.

4. *Outputs II* im Sinne der *Teilnahme* von Zielgruppenmitgliedern an den Maßnahmen werden durch Programmziele mitgesteuert. Es stellt sich nicht nur die Frage, wie viele Zielgruppenmitglieder erreicht werden, sondern auch, ob die Mischung der Teilnehmendenschaft den Vorgaben entspricht (z.B. mindestens x% Jugendliche im Alter zwischen .... und/oder aus benachteiligten Stadtteilen) oder ob die Verbleibdauer im Rahmen der gewünschten Zeitspanne liegt (weder zu lange Teilnahme noch unerwünschte Abbrüche). Qualifizierte Teilnahmedaten setzen ein entsprechend ausdifferenziertes Zielsystem und ein darauf aufbauendes programmintegriertes Monitoring voraus.

5. *Outputs III*: Vielfach konzentrieren sich empirische Erhebungen im Rahmen von Evaluation auf die Erhebung von *Reaktionen* seitens der Teilnehmenden. Die Zielgruppen beurteilen dabei Aspekte oder Elemente des Programms. Wie diese Urteile begründet sind bzw. in welchem Zusammenhang sie zu den Programmzielen stehen, wird oftmals nicht genügend expliziert und es fehlen entsprechende Messverfahren. Keinesfalls sollte die Erhebung von Reaktionen auf Programme mit einer konzeptionell und empirisch zureichenden Evaluation des Programms vermengt werden. Wohl liefern Reaktionsdaten Hinweise auf die Akzeptanz von Programmen durch die Zielgruppen und können somit auf Stärken und Schwächen des Programmprozesses hinweisen.

Die folgenden drei Ereignisglieder umfassen Outcomes im Sinne von Veränderungen und Stabilisierungen bei den Teilnehmenden.

6. Es werden zunächst *Outcomes I* wie die Veränderungen/Stabilisierungen von *Wissen, von Werten oder Einstellungen* aufgeführt, die weitgehend bereits während oder zum Abschluss der Durchführung von Programmen bei Zielgruppen auftreten und messbar sind. Bei Programmen zur Vermeidung oder Verhinderung von Armut bereiten diese kurzfristig erreichbaren Outcomes auf die nächsten beiden Ereignisglieder vor, nämlich dass sich die Zielgruppenmitglieder tatsächlich *anders verhalten und neue Handlungsstrategien* umsetzen, was schließlich zu einem veränderten Status bzw. einer verbesserten Lebenslage in ausgewiesenen Dimensionen führen kann.
7. *Outcome II*: Bereits während der Durchführung eines Programms, besonders aber auch im Nachhinein, kann überprüft werden, ob die Zielgruppenmitglieder wie gewünscht handeln: Entfalten sie Aktivitäten der Selbsthilfe, handeln sie so, dass sie unabhängiger werden von staatlichen Transferzahlungen, organisieren sie ihre Interessen als sozial Marginalisierte? Das Attribut des „Wünschenswerten“ macht ab dieser zweiten Outcome-Ebene sehr deutlich, dass es in der Gesellschaft und den verschiedenen Interessengruppen und sozialen Milieus unterschiedliche Vorstellungen über das gibt, was konkret „wünschbar“ ist und verweist auf die Frage der Werteberücksichtigung sowohl bei der Programmplanung als auch bei deren Evaluation.
8. Die *Outcome-Stufe III: „Lebenslage und Status“* verweist direkt auf die Lebenslage-Dimensionen und die Ressourcen-Dimension in der Armuts- und Reichtumsberichterstattung. Hier finden sich die gewünschten „Situationen oder Lebensbedingungen“, in die Mitglieder armer oder marginalisierter Bevölkerungsgruppen, unterstützt durch das Programm, einmünden sollen. Die Messungen können sowohl auf der Mikroebene, also in Bezug auf die Veränderungen oder Stabilisierungen, die bei einzelnen Personen der Armutsbevölkerung feststellbar sind, oder auf der Makroebene, in Bezug auf die

soziale Positionierung und Zusammensetzung von Segmenten armer Bevölkerungsgruppen, vorgenommen werden.[30]

9. Unter „*Impacts*“ oder „*Gesamtwirkungen*“ sollen solche (hier: gewünschten) Wirkungen von Programmen zur Verminderung und Vermeidung von Armut gefasst werden, die auf der Ebene von Sozialsystemen auftreten. Beispiele sind (a) Nachbarschaften / Stadtquartiere, in denen Netzwerke und Ressourcen zur Minderung/Vermeidung von Armut geschaffen sind oder ein gesellschaftliches Meinungsklima, das Bekämpfung von Armut hoch auf der politischen Agenda verortet oder das bestimmte Lebensformen, die mit dauerhafter oder zeitweiser Abhängigkeit von gesellschaftlicher Solidarität/Transferleistungen einhergehen, toleriert.

   Deren Messung erfordert besonders umfassende empirische Designs; der Ursachennachweis auf einzelne Programme oder Programmbündel dürfte in aller Regel sehr schwierig, wenn überhaupt leistbar sein.

Obwohl die grafische Darstellung der neungliedrigen Ereigniskette in Abb. 4 bereits starke Vereinfachungen enthält – vielfach verlaufen Beeinflussungsprozesse nicht unlinear, sondern sind eher multipel; es gibt auch rekursive Prozesse und gegenläufige Beeinflussungen – so macht sie deutlich, wie anspruchsvoll umfassende Evaluationen von nationalen Programmen der Vermeidung und Verhinderung von Armut sind. Andererseits dürfte es für lokale und regionale Programme oder Programmsegmente hilfreich sein, wenn sich sowohl die Planung der Programme als auch die Evaluation an der entfalteten Systematik orientieren. In einem längerfristig zu sehenden Forschungsprozess können dabei sowohl Schlussfolgerungen für die Evaluation umfassender, nationaler Programme gezogen werden als auch systematisch Möglichkeiten genutzt werden, lokale und regionale Programmumsetzungen zusammenzuführen und so zu einer systematischen Wertung nationaler Programme zu kommen.

Der nachfolgende Abschnitt 1.5 stellt dar, welche unterschiedlichen Funktionen Evaluation bei der Steuerung, Bewertung und Rechenschaftslegung von Pro-

---

30 Hierauf wird im Kapitel 4 „Datenlage für Evaluationen“ dieses Berichts ausführlich eingegangen.

grammen zur Vermeidung/Verminderung von Armut haben kann und wie diese im Programmzyklus gezielt genutzt werden können.

1.5 Funktionen und Ansatzpunkte wirkungsorientierter Evaluation

Am Ausgangspunkt dieser Perspektivstudie stand die Frage, wie Evaluationen dazu beitragen können, die Wirkungen staatlicher Politik und Programme nachvollziehbar zu machen und zu überprüfen. In der Darstellung von Grundbegriffen, der Vorstellung der Standards für Evaluation und der Diskussion des Wirkungsbegriffes ist deutlich geworden, dass Evaluationen sowohl rückwärts gewandt bereits durchgeführte Programme als auch vorwärts gewandt noch nicht entwickelte/in der Planungs- oder Erprobungsphase befindliche Programme zum Gegenstand haben können.

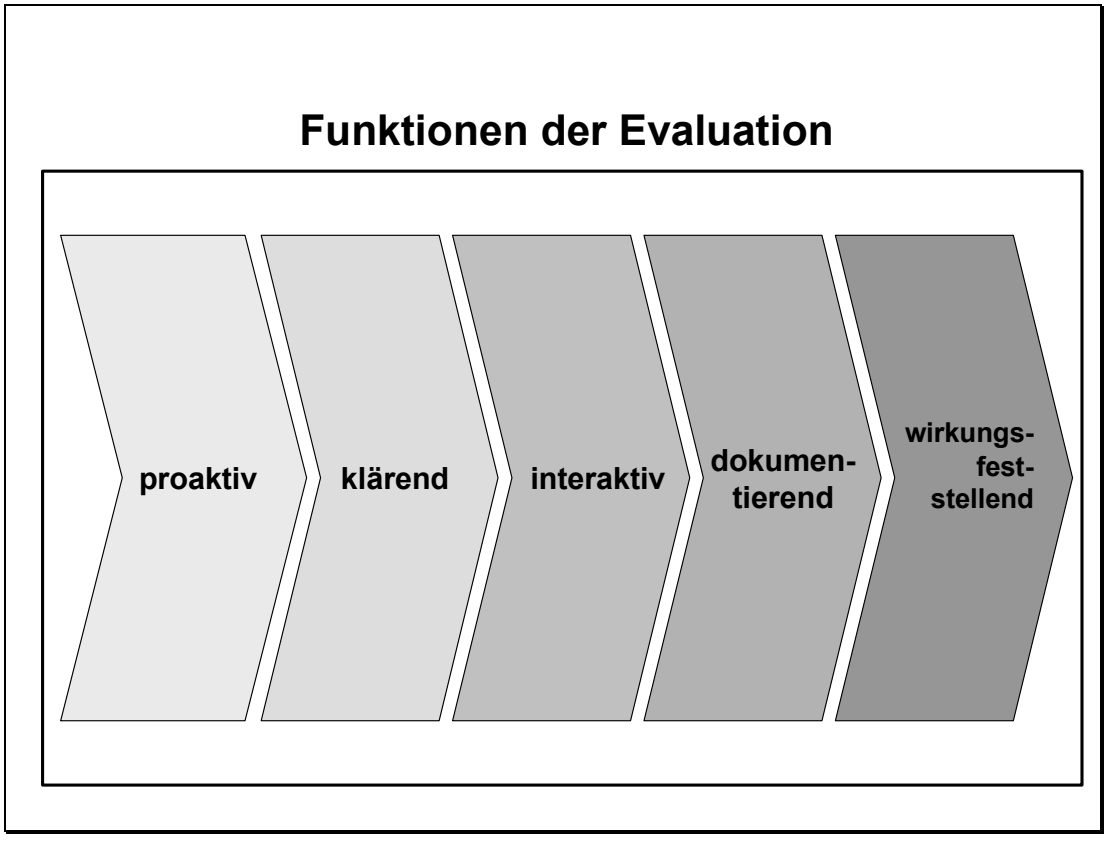
Abbildung 5: Zur Ausklammerung des Begriffs der „Wirkungskontrolle“ aus der Perspektivstudie

Der in der Alltagssprache weit verbreitete Terminus der Wirkungskontrolle ist in der sozialwissenschaftlichen und evaluationstheoretischen Literatur nicht einheitlich definiert. Er unterstellt die (leichte) empirische Nachweisbarkeit von Ursache-Wirkungs-Relationen und kann darüber hinaus die Assoziation auslösen, dass dies (immer) durch „kontrollierte“, d.h. (quasi-) experimentelle Designs zu geschehen habe. Die Theoriegeschichte der Evaluation im Bereich sozialer Programme, insbesondere der personenbezogenen Dienstleistungen – und die Ausdifferenzierung von Modellen der Evaluation machen deutlich, dass die Durchführung von Wirkungskontrollen im strengen Sinne vielfach nicht möglich ist, dass sie nur unter günstigen Umständen (Gleich-Halten der Rahmenbedingungen eines stark standardisierten Programms) und mit hohem Aufwand durchführbar sind und dass darüber hinaus die Nutzung ihrer Untersuchungsergebnisse oft unterbleibt. Schließlich unterstellt dieser Zugang, dass Programme zur Armutsverminderung und -vermeidung in aller Regel klare Ziele, darauf zugeschnittene Interventionen usw. aufweisen, also hinreichend beschrieben, wie geplant umgesetzt, stabil über die Zeit und damit grundsätzlich übertragbar sind. Dies ist hingegen selten der Fall und stattdessen eine vielfach noch zu leistende Aufgabe für die wirkungsorientierte Evaluation. Außerdem trägt eine reine Wirkungskontrolle im experimentellen Sinne in aller Regel wenig an empirisch gesicherten Informationen bei, welche die (formative) Verbesserung von Programmen abstützen könnten. Angesichts sich rasch wandelnder Rahmenbedingungen in einer globalisierten Ökonomie und kurzen technologischen und sozialen Innovationszyklen greift eine Reduktion der Evaluation auf ihre summative Leistung zu kurz. „Wirkungsorientierung“ soll die innovativen, lösungsorientierten Potentiale der Evaluation für eine Politik der Armutsverminderung und -vermeidung hervor heben gegenüber einer ausschließlich um Überprüfung und Nachweis bemühten, nachträglich kontrollierenden Vorgehensweise.

Um die gesamte Leistungsbreite von Evaluationen präsent zu machen, sprechen wir von „wirkungsorientierter“ Evaluation. Diese ist darauf verpflichtet, ihre theoretischen Erkenntnisse und methodischen Vorgehensweisen in allen Phasen eines Programms auf „Wirkungen“ auszurichten – optional von der vorgängigen Machbarkeitsstudie zu einem Programm bis hin zur Kosten-Nutzen-Analyse nach Programmschluss.

Welche möglichen Funktionen Evaluation dabei haben kann, steht im Mittelpunkt dieses Kapitels.

Abbildung 6: Funktionen der Evaluation



Quelle: eigene Darstellung

Je nach Zeitpunkt im Programmablauf, zu dem Evaluation eingesetzt wird, kann sie eine oder – zeitlich nacheinander – mehrere der folgenden Funktionen wahrnehmen (vgl. Owen/Rogers 1999):

- **Proaktive Evaluation** ermittelt vor dem Start eines Programms anhand von Vergleichsuntersuchungen oder eigens für diesen Zweck durchgeführten Erhebungen, welche Bedarfe bei den Zielgruppen vorliegen (Bedarfsermittlung), ob die vorhandenen Rahmenbedingungen die Durchführung des Programms ermögli-

chen (Machbarkeitsabschätzung) bzw. welche entwickelbaren Programmalternativen voraussichtlich zu welchen unterschiedlichen Resultaten führen (Simulationen).

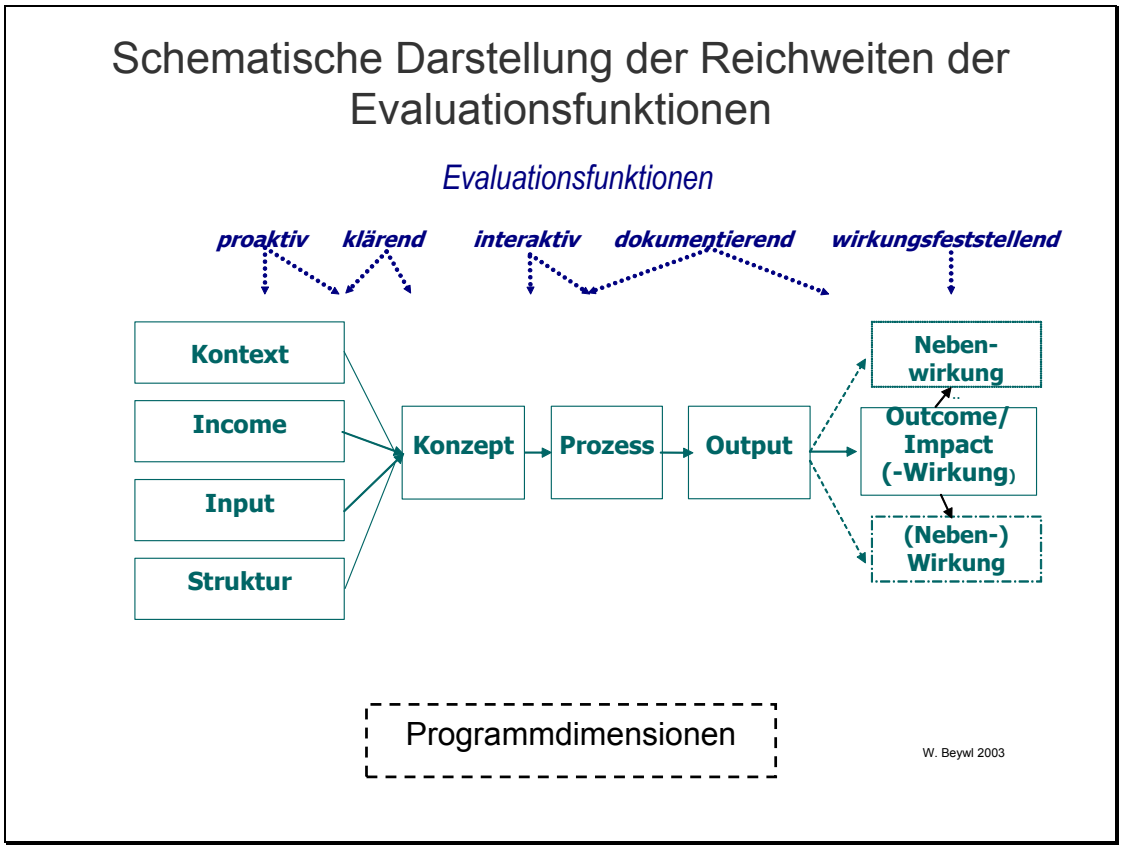
- **Klärende Evaluation** überprüft in der Konzeptphase eines Programms und vor Beginn seiner Umsetzung die Prägnanz und Stimmigkeit der Programmziele, ggf. auch deren Passung auf Bedarfs- und soziale Problemlagen. Um die Bewertung des Programms anhand der Ziele vornehmen zu können, werden die Ziele in Fragestellungen überführt. Klärende Evaluation bereitet so die Messbarkeit der Programmziele vor, unterstützt die Abstimmung von Interventionen auf die Ziele, gewinnt Informationen über die Ausgangssituation und die Umsetzungsbedingungen (Zielklärung, Konzeptentwicklung und Programmtheorie).
- **Interaktive Evaluation** läuft parallel zur Programmumsetzung, besonders bei Pilot- und Modellprogrammen, liefert Zwischenergebnisse zur Prozessqualität und unterstützt die Feinabstimmung von Zielen, Interventionen sowie Anpassung und Umsetzungen an verschiedenen Standorten. Sie überprüft die Akzeptanz des Programms bei verschiedenen Beteiligten (z.B. Kommunen, Wohlfahrtsverbänden). Sie arbeitet Stärken und Schwächen des Programms heraus, gibt ggf. Hinweise zur Neuausrichtung von einzelnen Programmschritten (Programmbegleitung/-optimierung) und prüft Möglichkeiten einer z.B. bundesweiten Übertragung. Dabei stimmt sie empirisches Vorgehen beständig mit den am Programm Beteiligten ab und gibt möglichst kurzfristig Rückmeldungen, so dass die Programmverantwortlichen zeitnah Verbesserungsmaßnahmen einleiten können.
- **Dokumentierende Evaluation** (auch: „Monitoring“, vgl. Kap. 4.1.1) stellt laufend zentrale Kennzahlen zu Programmen / Teilbereichen bereit. In der Regel handelt es sich um ein schlankes Indikatorenset, das auf der Basis bestehender Datenbanken/Statistiken/des Programm-Monitorings der Träger zu wiederkehrenden Zeitpunkten (monatlich, quartalsweise, jährlich) vollständig bereitgestellt werden kann. Im Mittelpunkt stehen dabei Output-Indikatoren. Auch kurzfristig auftretende sowie leicht zu messende Outcome-Indikatoren können integriert werden. Die Erhebungsverfahren werden möglichst in den Programmablauf eingebaut und eine kurzfristige Rückkopplung an Programmverantwortliche/Finanziers ist vorgesehen.(Programmlegitimierung/Feinabstimmung/Überprüfung der Zielerreichung).

- **Wirkungsfeststellende Evaluation** zeichnet nach, welche Ziele das Programm in welchem Umfang erreicht hat und ob die Zielerreichung tatsächlich auf die Durchführung des Programms zurückzuführen ist. Im Mittelpunkt können die intendierten Wirkungen bei den Zielgruppen (Outcome-Wirkungen) oder Wirkungen etwa auf die natürliche, gebaute oder soziale Umwelt stehen. Im Rahmen „zielfreier Evaluationen“ ist es schließlich möglich, Wirkungsüberprüfungen unabhängig von den Programmzielen anzulegen, was es insbesondere ermöglichen soll, nicht-intendierte Nebenwirkungen offen zu legen (Rechenschaftsregelung/Erfolgskontrolle/Auswahl von reifen Programmalternativen, Entscheidungen über Einstellung, Weiterführung oder Auswertung von Programmen).

Die Aufgabe der Wirkungsfeststellung wird durch die verschiedenen Ansätze der Evaluation höchst unterschiedlich angegangen. Allgemein geht es darum zu belegen und plausibel zu machen, dass ein empirisch nachgewiesenes Resultat (z.B. durch vorher langfristig Arbeitslose gelungene Arbeitsaufnahmen) auf das Programm zurückzuführen ist. Dies erfolgt je nach gewähltem Evaluationsmodell, verfügbaren Evaluationsressourcen, Art der Umsetzung des Programms (lokal/national – zentral/dezentral) auf unterschiedlichen Wegen, z.B. durch Befragung von Beteiligten und Betroffenen oder von Experten/-innen (Wirkungseinschätzung), durch Nachzeichnung im Rahmen eines differenzierten, ggf. wissenschaftlich begründeten Wirkungsmodells (Wirkungsmodellierung) oder durch Isolation der Programmwirkungen im Rahmen (quasi-)experimenteller Designs (Wirkungsnachweis). Dabei können diese Strategien kombiniert werden, allerdings mit Folgen für den zu treibenden Evaluationsaufwand (vgl. die Darstellung der Evaluationsmodelle im Kap. 2.3).

Die Reichweite der fünf Evaluationsfunktionen ist unterschiedlich: So setzen interaktive, dokumentierende und wirkungsfeststellende Vorgehensweisen ein bereits laufendes Programm voraus. Wirkungsfeststellung kann bei Programmen mit einem Mindestmaß an routinisierter Durchführung gültige Ergebnisse liefern, doch nur mit Einschränkungen bei Pilotprogrammen, die ein neues oder neu identifiziertes soziales Problem erstmals bearbeiten.

Abbildung 7: Evaluationsfunktionen im Programmzyklus



Quelle: eigene Darstellung

Wie im Kap. 2.1 im historischen Rückblick ausführlicher dargestellt, laufen Evaluationen, die sich ohne Prüfung der Programmqualität unmittelbar auf die Wirkungsfeststellung konzentrieren, Gefahr, Blindleistungen zu produzieren. Wenn es den politischen Entscheidern/-innen auf die Identifikation wirksamer oder zumindest wirkungsfähiger Programme ankommt, ist ihnen wenig damit gedient, wenn z.B. die Erklärung für aufwändig gemessene *fehlende* Wirkungen lautet: Das Programm ist von falschen Rahmenbedingungen ausgegangen; sein Konzept weist keine operationalisierungsfähigen Ziele aus; die Implementationstiefe des Programms ist an der Mehrzahl der Programmstandorte nicht befriedigend. Die häufige Meldung des *no effect* und die wachsende Unzufriedenheit von Politik und öffentlicher Meinung führten in den frühen 70er Jahren in den USA zu einer Nutzungskrise der Evaluation und damit zu einer Umorientierung in den Modellen und Vorgehensweisen (vgl. Kap. 2.1.1).

Bezogen auf die fünf dargestellten Evaluationsfunktionen bedeutet dies, dass vor der Entscheidung, überzugehen von der proaktiven in die klärende, von der klärenden in die interaktive Funktion usw. jeweils zu prüfen ist, ob das Programm die erforderliche

Strukturiertheit und Reife aufweist, um mit der nächst voraussetzungsvollen Evaluationsfunktion bearbeitet zu werden.

Dabei kann jede der fünf Evaluationsfunktionen von vornherein „wirkungsorientiert“ angelegt werden. So kann bereits mit der proaktiven oder der klärenden Funktion das Schwergewicht auf Wirkungen gelegt werden, wie die folgenden möglichen Evaluationsfragestellungen verdeutlichen:

- Mit welcher Wahrscheinlichkeit kann das zu entwickelnde Programm unter den gegebenen Rahmenbedingungen die ihm von der Politik angetragenen Wirkungen erreichen? Wie sollen die übergreifenden Ziele des zu entwickelnden Programms mit Prioritäten versehen werden, so dass „Zielüberfrachtung“ vermieden und die Auslösung der zentralen Wirkungen möglichst sicher gebahnt wird? Welche Ressourcen sind schätzungsweise erforderlich, um eine gewünschte Quantität und Qualität von Wirkungen auszulösen? Welche der Wirkungen können mit hoher, welche mit geringerer Wahrscheinlichkeit ausgelöst werden? Welche Schlussfolgerungen sind zu ziehen? (proaktiv)
- Wie sind die Programmziele angesichts des politischen Auftrages anzulegen, damit sie auf den Kern der beabsichtigten Wirkungen ausgerichtet sind? Wie ist eine Passung mit den Ressourcen der verschiedenen Zielgruppen und ihren Eingangsbedingungen (Income) als wichtige Bedingung für Programmerfolg sicherzustellen? Wie können die Programmaktivitäten so geplant und abgestimmt werden, dass möglichst geringe Blindleistungen und ein möglichst hoher Prozentsatz an Leistungen bereit gestellt wird, die unmittelbar in intendierte Wirkungen/ Outcomes münden? Wie kann der Zusammenhang zwischen den genannten Elementen so konstruiert und dargestellt werden, dass sich eine schlüssige und übertragbare Programmtheorie ergibt? (klärend)

Zusammenfassend ist eine Evaluation dann als „wirkungsorientiert“ zu bezeichnen, wenn sie bei der jeweiligen Evaluationsfunktion darauf achtet, dass intendierte Wirkungen im Evaluationsplan zentral angesprochen werden.

Formative wirkungsorientierte Evaluationen unterstützen es, dass das Programm so ausgestaltet und verbessert wird, dass es seine beabsichtigten Wirkungen in möglichst großem Maße erzielt.

Summative wirkungsorientierte Evaluationen messen Resultate und Wirkungen, um eine Entscheidung über das Programm empirisch zu fundieren und eine Basis zu schaffen für die im demokratischen politischen System erforderliche Rechenschaftslegung.

Die Auswahl des jeweils passenden Evaluationsmodells hängt somit stark davon ab, *wie* die Programme/Politiken, deren Wirkungen zu überprüfen sind, im Detail ausgestaltet sind. Wir gehen nicht davon aus, dass es für alle Programmarten, Evaluationszwecke und zu beantwortende Fragestellungen ein universell passendes „Goldmodell“ der Evaluation gibt, sondern dass dies stets eine Frage der angemessenen Wahl zwischen unterschiedlich geeigneten Alternativen ist. Einen Einblick in die Wahlmöglichkeiten und eine Orientierung für Auswahlen gibt das folgende Kapitel.

2 Modelle der Evaluation

In diesem Kapitel wird zunächst die Entwicklung der Evaluation – mit Schwerpunkt USA – im Kontext von sozialpolitischen und insbesondere Programmen zur Bekämpfung der Armut dargestellt. Im nächsten Abschnitt werden verschiedene Typologien von Evaluationsmodellen vorgestellt. Schließlich wird die für diese Perspektivstudie speziell entwickelte Modelltypologie eingeführt und erläutert. Daran schließen sich die Vorstellungen der insgesamt zwölf ausgewählten Modelle an.

2.1 Evaluation und soziale Politik: USA – Deutschland im Vergleich

Dieses Kapitel verweist auf das in Deutschland bestehende evaluationstheoretische Defizit in Bezug auf die systematische Beschreibung und Bewertung von armutsbezogenen Politiken und Interventionsprogrammen.

2.1.1 USA: Entwicklung der Evaluation im Kontext sozialpolitischer Programme

„Like many poor people, evaluation in the United States has grown up in the „projects“ – federal projects spawn by the Great Society legislation of the 1960s.”
(Patton 1997, S. 10)

Die Ausdifferenzierung von Evaluation aus der empirischen Sozial- und Wirtschaftsforschung ist eng verbunden mit staatlichen Programmen zur Vermeidung von Armut. Schon früh wurden wirkungsorientierte Studien im Kontext der Politik zur Verbesserung von kritischen Lebenslagen insbesondere bezüglich Erziehung/Bildung und Gesundheit durchgeführt. Die „Geburt“ und Reifung der Evaluation zur entwickelten Methodologie und Profession steht im Kontext staatlicher Politik, insbesondere der Bildungspolitik und der Sozialpolitik in den USA.

> *„Vor dem ersten Weltkrieg waren die wichtigsten Anstrengungen darauf gerichtet, einerseits Qualifizierungsprogramme einzuschätzen, in denen es um Alphabetisierung und Beschäftigung geht, andererseits Initiativen im öffentlichen Gesundheitswesen, die auf die Verringerung der Sterblichkeit und von infektiösen Erkrankungen gerichtet waren.“ (Rossi 1999, S. 19)*

In der Weltwirtschaftskrise engagierte sich die US-amerikanische Bundesregierung im Rahmen nationaler Wohlfahrtsprogramme und forderte Informationen über deren Ablauf und Auswirkungen an, die oftmals sehr unsystematisch generiert wurden.

> *„Während der Depression der 30er Jahre übernahm die Bundesregierung der Vereinigten Staaten eine Hauptrolle dabei, Armut, Hunger und Arbeitslosigkeit zu mindern; das Naheliegende in Bezug auf Evaluation war es, ein paar arbeitslose Akademiker zu beschäftigen, damit diese Programm-Geschichten (program histories) aufschreiben." (Patton 1997, S. 10)*

Einen großen Schritt vorwärts tat die Evaluationstheorie und -praxis im Zusammenhang mit den Programmen im Umfeld des *New Deal* des US-amerikanischen Präsidenten Roosevelt. Damals wurden erstmals umfassende Möglichkeiten gesehen, breit angelegte experimentelle Wirkungsstudien durchzuführen.

> *„Diese Laboratorien, die durch die Planungsagenturen des New Deal eingerichtet sind, ermöglichen einen effektiveren Einsatz der experimentellen Methode in den Forschungsprojekten von Sozialwissenschaftlern/-innen. Diese Forschung wiederum würde nicht nur einen Zuwachs für die Wissenschaft bedeuten, sondern würde auch eine Form des sozialen Auditing darstellen für die Planungsbehörden, indem sie die Veränderungen aufzeichnen und berechnen, die durch die Programme errungen worden sind." (Stephan 1935, S. 518)*

Obwohl sich Programmevaluation bereits Ende der 50er Jahre in den USA etabliert hatte, erfolgte unter den Präsidentschaften von Kennedy und Johnson der eigentliche Durchbruch: Einen Meilenstein markiert das Jahr 1965, in dem die *War on Poverty/Great Society*-Programmatik beschlossen und das Planning-Programming-Budgeting-System (PPBS) durch Verordnung eingeführt wurde:

> *„Angesprochen waren Fachleute, welche ihre Fähigkeiten und Interessen bereit sind einzusetzen, um die Effizienz zu untersuchen, mit denen öffentliche Maßnahmen ihre Ressourcen alloziieren, deren Auswirkungen auf das individuelle Verhalten, deren Effektivität beim Erreichen der Ziele, für die sie zugeschnitten wurden, und deren Effekte in Bezug auf das Wohlergehen der Reichen gegenüber den Armen, der Minoritäten gegenüber der Majorität und des Nordens gegenüber dem Süden. Diesen Fachleuten boten beide Programme Standing, Legitimation und finanzielle Unterstützung. ... Eine präsidentielle Executive Order verschaffte Tausenden Beschäftigung und finanzielle Unterstützung, die ihre analytischen Fähigkeiten zu solchen Fragen der*

Effizienz, der Effektivität und der Gleichheit einzusetzen bereit waren.“ (Havemann 1987, S. 120)

Die Evaluation dieser nationalen Programme war in der Regel verpflichtend und die hierfür eingestellten Budgets lagen oft bei 1% bis 3% der Programmgesamtkosten.

Vielfach prägend und geradezu synonym mit Evaluation war zu diesem Zeitpunkt das Programmziel-gesteuerte Evaluationsmodell (vgl. Kap. 2.3.1.1): Aus einer idealerweise vollständigen, gegliederten Zielhierarchie des Programms, die in operationalisierten Feinzielen (*objectives*) mündet, werden die Bewertungsmaßstäbe und damit auch die weitgehend standardisierten Erhebungsinstrumente abgeleitet. Der Grad der Abweichung bzw. Übereinstimmung der Programmresultate von/mit den Zielen gilt als Maßstab für Erfolg und Misserfolg (vgl. z.B. Provus 1971).

Als Kontrapunkt wurde von Scriven (1973, 1991) das zielfreie Evaluationsmodell in die Diskussion gebracht. Obwohl dieses Modell selten in Reinform realisiert wird, hat es doch die methodologische Weiterentwicklung der Evaluation maßgeblich angeregt, weshalb seine Argumentation hier ausführlicher dargestellt sei: Es kritisiert am Programmziel-gesteuerten Modell, dass zwar die Zielerreichung überprüft werde, damit aber in keiner Weise eine umfassende Bewertung des Programms geleistet werden könne: Selbst wenn die zentralen Ziele erreicht werden, könnten die (nicht gemessenen) negativen Nebenwirkungen des Programms die (ausschließlich gemessenen) positiven Wirkungen übersteigen. Beispielsweise könne dies dann geschehen, wenn die Kosten des Programms nicht beachtet werden und z.B. Programmalternativen nicht auf ihre Kostenwirksamkeit hin betrachtet werden. Fehlentscheidungen seien durch eine zielgebundene Evaluation vorprogrammiert. Scriven vertritt die Auffassung, dass die Einschätzung der Güte und Verwendbarkeit eines Evaluationsgegenstandes auch ohne Kenntnis der Programmziele möglich und sogar besonders fruchtbar sei. Die Evaluation solle gegenüber der gesamten Bandbreite der Auswirkungen offen sein. Es wird somit die Unterscheidung in Haupt- und Nebeneffekte (sekundäre Effekte) eines Programms aufgehoben und damit auch die Vorab-Unterscheidung zwischen vorhergesehenen und nicht vorhergesehenen sowie die Unterscheidung zwischen erwünschten und unerwünschten Auswirkungen. Dies soll verhindern, dass die Aufmerksamkeit allein den beabsichtigten Haupteffekten gilt und die drei anderen Wirkungsarten vernachlässigt werden.

Scriven vergleicht stattdessen die Auswirkungen des betrachteten Programms mit den Bedürfnissen/Bedarfen (*needs*) der angesprochenen Zielgruppen und denen der ausgeschlossenen bzw. anderweitig negativ betroffenen Bevölkerungsgruppen.

Vedung (1999) geht noch einen Schritt weiter: Er schließt auch Bedürfnisse aus dem Evaluationsmodell aus, da diese noch schwieriger zu bestimmen seien als die Ziele. Vedung stellt die zielfreie Evaluation in den Kontext der neuen Steuerungstechnik des öffentlichen Sektors (Resultatsanalyse). Es sollen statt der Evaluationsverantwortlichen die Adressaten der Evaluation Bewertungen vornehmen, auf der Basis einer umfassenden Information über die festgestellten (Aus-)Wirkungen. Dies ist vergleichbar mit dem Vorgehen des vergleichenden Warentestes, der über verschiedenste Dimensionen eines Gegenstandes (Produkte oder Dienstleistungen) Mess- und Einschätzdaten zur Verfügung stellt. Die schließliche Auswahl des passenden Produktes – z.B. über individuelle Gewichtung der verschiedenen Bewertungsdimensionen – bleibt dem Kunden überlassen.

Während auf einem Markt Käufer/-innen und Nutzer/-innen des Produkts/der Dienstleistung meist identisch sind, fallen bei sozialen Programmen beide Rollen auseinander: So „kauft“ die Kommune eine Beratungsdienstleistung bei einem freien Träger ein, welche Mitglieder der Zielgruppe kostenlos (ggf. mit Gutscheinen ausgestattet) nutzen. In gewisser Weise kann die „zielfreie Evaluation“ als analoge Anwendung des Warentestes im Bereich Evaluation von (sozialen) Gegenständen (Programmen) angesehen werden.

Dieser Evaluationsansatz bietet sich bei unklaren Programmzielen und räumlich weit verstreuten Zielgruppen an. Er ist in der Praxis selten durchgeführt worden, gleichwohl als Denkmodell zur Schärfung von Annahmen im Rahmen von Wirkmodellen anregend. Da das Modell trotz seiner Konzeption in den 60ern und seiner intensiven Rezeption in der Evaluationsliteratur kaum praktische Anwendung findet, ist es in die Modelldarstellung in Kap. 2.3 nicht aufgenommen.

Eine weitere Kritik am ‚Urmodell’ der zielgesteuerten Evaluation lautet, dass Zielzustände als Resultate zwar festgehalten sind, aber nicht geklärt ist, ob diese ursächlich auf das Programm oder auf andere Faktoren zurückgehen. Die so aufgeworfene Frage der Wirkungen (belegt durch bestimmte Erhebungsdesigns) trat mit Nachhalt bei den großen, auf Armutsverminderung gerichteten Programmen der 60er Jahre in den Mittelpunkt. Die Vision von der „Experimentierenden Gesellschaft“ war

geboren, die vielfach als das Ideal der zu etablierenden Zusammenarbeit von Politik und Evaluation galt:

> *„Die experimentelle Gesellschaft wird eine sein, die energisch vorgeschlagene Lösungen für beständige Probleme ausprobiert. Diese Lösungen gehen hervor aus „harten" und multidimensionalen Evaluationen der Outcomes. Sie entwickeln sich zu anderen Alternativen hin, wenn die Evaluation zeigt, dass eine Reform ineffektiv oder schädlich gewesen ist. Wir haben eine solche Gesellschaft heute noch nicht." (Donald T. Campbell in seiner Eröffnungsansprache zur Jahrestagung der American Psychological Association 1971, S. 223)*

Die auf kausale Erklärungen zielende (quasi-) experimentelle Forschung (vgl. Kap. 2.3.1.2 sowie 2.3.1.3) steht in der Tradition des auf John Stuart Mill (1806-1873) zurückgehenden empirisch-analytischen Wissenschaftsverständnisses bzw. des kritischen Rationalismus (Popper 1989), von Gegnern als „Positivismus" bezeichnet. Evaluation wird aus diesem Verständnis als eine Form angewandter empirischer Forschung verstanden, die denselben wissenschaftlichen Gütekriterien – insbesondere dem Kriterium der Theoriegeleitetheit – wie die Grundlagenforschung zu genügen hat.

Dieses enge Verständnis von Evaluation erweiterte sich in den USA – in Reaktion auf die ausgebliebenen Nutzungen von Evaluationsergebnissen im Evaluations-Boom zwischen 1965 und 1975 – und führte zur bis heute fortschreitenden Ausdifferenzierung verschiedenster Evaluationsmodelle. In Deutschland hingegen war bis in die 90er Jahre hinein das Verständnis von Evaluation als Unterform angewandter Sozialforschung dominant (vgl. Kap. 2.1.2).

Einen Meilenstein in der Entwicklung der Evaluation aus dem Bereich der Armutsbekämpfung stellen die Ende der 60er Jahre begonnenen Feldexperimente zu den Wirkungen einer negativen Einkommenssteuer dar. Es sollte hierbei insbesondere überprüft werden, ob diese Form eines garantierten Mindesteinkommens die Mitglieder der Empfänger-Haushalte von der Aufnahme einer Erwerbsarbeit abhält. In das erste einer Serie von fünf aufwändigen Feldexperimenten, das *New Jersey-Pennsylvania Income Maintenance Experiment*, wurden vollständige Familien unterhalb der Armutsgrenze mit männlichen Haushaltsvorständen zwischen 18 und 58 Jahren einbezogen. Dabei wurden acht verschiedene Ausgestaltungen des Transfereinkommens ausprobiert. Die Kontrollgruppe bestand aus Haushalten, die keinerlei Unter-

stützung erhielten. Im Anschluss an eine Haushaltsbefragung wurden diejenigen Haushalte, die sich zur Teilnahme daran bereit erklärt hatten, nach Zufall der Experimental- und der Kontrollgruppe zugewiesen. Nach Beginn der Zahlungen wurden die Familien vierteljährlich interviewt. Durch Verweigerungen und Interviewausfälle blieben von den 1.300 zu Beginn nach Ablauf des dreijährigen Untersuchungszeitraums 700 vollständig erfasste Fälle übrig. Ein – nicht unumstrittenes – Ergebnis war, dass die Haushalte der Experimentalgruppe ihre Erwerbsanstrengungen um 5% reduziert hatten (Kershaw/Fair 1976). Die anderen Studien spezifizierten das Ergebnis insofern, als die erwerbsmindernden Effekte besonders für junge Erwachsene und Mütter kleiner Kinder auftreten (Rossi/Lyall 1976; SRI International 1983, Robins et. al 1980, Mathematica Policy Research 1983).

Die Erfahrungen mit diesen groß angelegten Programmen, die z.T. übereilt beschlossen und schlecht vorbereitet umgesetzt wurden, waren vielfach ernüchternd. In der Konsequenz wurden auch auf anderen Feldern weniger sozialpolitische Programme finanziert. Zum anderen wurde auch die *sunset legislation* eingeführt, d.h. Leistungsgesetze hatten von vorneherein eine terminierte Gültigkeit, die zu ihrer Verlängerung einen positive (Aus-)Wirkungen bescheinigenden Evaluationsbescheid erforderte.

Nicht nur die Sozialpolitik, sondern auch die Evaluation geriet in den 70er Jahren in Misskredit. Zum einen war dies extern bedingt. Im Rahmen der beginnenden restriktiven Fiskalpolitik gewannen betriebswirtschaftliche Verfahren an Boden. Diese bemühen sich, die messbaren (Aus-)Wirkungen in Relation zu den aufgewendeten öffentlichen Mitteln in monetären Größen zu bestimmen (vgl. Kap. 2.3.1.4). Im General Accounting Office wurde die Funktion der Programmüberprüfung unter dem Stichwort *accountability* aufgewertet und *performance-auditing* sowie *performance measurement* wurden in den kommenden Jahrzehnten weiter entwickelt (vgl. Wisler 1996).

> *„Kurz gesagt verändert sich der politisch institutionelle und der politisch inhaltliche Kontext, in dem Evaluationen durchgeführt werden. Die Bewegung geht weltweit in Richtung verminderter Besteuerung, verminderter Staatsdefizite, verminderter Transferleistungen von den Reichen zu den Armen und verminderter Größe von Verwaltung und Regierung. Wegen dieser Bewegung nimmt Evaluation eine größere öffentliche*

Bedeutung ein insofern der Bedarf steigt, die Kosten-Effektivität von Politiken und Programmen zu messen." (Chelimsky 1997, S.15)

Zum anderen war die zu Beginn der 70er Jahre einsetzende Krise der Evaluation auch hausgemacht. Die Kernproblematik wurde darin gesehen, dass viele oder sogar die meisten Evaluationsergebnisse nicht befriedigend genutzt wurden, weder für Verwaltungs- und Haushaltsentscheidungen, noch für politische Richtungsentscheidungen oder zur Programmverbesserung. Dies wurde durch mehrere Studien empirisch nachgewiesen (vgl. Beywl 1988, S. 32ff). Im Mittelpunkt der Kritik standen damals die von Politik und Wertentscheidungen distanzierten, rein quantitativen, oftmals sehr kostenintensiven Evaluationen, die mit (quasi-) experimentellen Designs arbeiteten.

„Vor 25 Jahren nahmen viele Evaluatoren/-innen Naiverweise an, dass ihre Ergebnisse routinemäßig als zentraler Input für politische Entscheidungen genutzt würden. Die Parteinahme für das Experimentelle dieser Zeit nährte vielleicht diese Naivität, weil die Entscheidungslogik, die den experimentellen Designs unterliegt, offensichtlich das Modell des rationalen Akteurs aus der öffentlichen Politik spiegelt." (Cook 1997, S. 40)

Eine Reaktion auf diese „Krise" war die Entwicklung neuer Evaluationsansätze, insbesondere der Entscheidungsorientierten, aber auch der responsiven oder emergenten Modelle.[31]

Die Reagan-Administration stützte die Einschränkung des damals einzigen flächendeckend angelegten sozialen Grundsicherungsprogramms in den USA, des AFDC (Aid to Families with Dependent Children, datierend aus dem Jahr 1935), welches hauptsächlich Kindern allein erziehender Frauen Geld- und Sachleistungen gewährte, auf eine damals breit rezipierte Meta-Analyse vorliegender Evaluationen von Charles Murray (1984). Dieser kam zu dem Schluss, dass diese Programme der hauptsächliche Mechanismus für die Armutsfalle (*culture of dependency*) seien, in welche die Zielgruppen gerieten und empfahl die vollkommene Einstellung aller Sozialleistungen an außerhalb von Ehen geborenen Kinder (vgl. auch die Darstellung von Heclo 2001, S. 181-183). In der Folge berief Präsident Reagan eine Kommission

---

31 Eine aktuelle Diskussion zwischen Vertretern verschiedener Modelle findet sich in Donaldson/ Scriven 2003 (Stewart I. Donaldson/Michael Scriven (Hrsg.)) Evaluation Social Programs and Problems. Visions for the New Millennium. Mahwah/London 2003.

zur Überprüfung der nationalen Sozialprogramme, deren Bericht den Umbau in Richtung der *Welfare to Work* Programme einleitete (Domestic Policy Council 1986).

In der Folge avancierte das Thema „Evaluation und Soziale Gerechtigkeit“ zu einer zentralen Thematik unter US-amerikanischen Evaluatoren/-innen: 1992 war Debra Rog Herausgeberin eines Themenheftes *Evaluating programs for the homeless*. Kenneth Sirotnik gab 1990 ebenfalls ein Themenheft der führenden Publikationsreihe *New Directions for Program Evaluation*[32] zu *Evaluation and Social Justice* heraus. Die Jahrestagung der Amerikanischen Evaluationsgesellschaft trug 1994 ebendiesen Titel.

Ein Produkt dieser Auseinandersetzung ist die stärkere Thematisierung von politischer Macht und sozialen Werten als zu berücksichtigendem Kontext, in dem Evaluatoren/-innen insbesondere im Rahmen nationaler und bundesstaatlicher Programme tätig werden:

> *„Wir haben schließlich gelernt, die vertrackten normativen Fragen nicht mehr zu vermeiden. Sollten sozialwissenschaftliche Informationen solch eine instrumentelle, Entscheidung generierende Rolle in einer Demokratie spielen, wobei Wertfragen als allgegenwärtig gelten und Management-Effizienz ein weniger bedeutsames Desiderat ist?“ (Cook 1997, S. 40)*

Cook führt eine Argumentation fort, die sein prominenter Koautor, der 1996 verstorbene Donald D. Campbell, bereits in Absetzung zum „objektivistischen“ empirisch-analytischen Ansatz begonnen hatte: Campbell bezweifelt die Existenz objektiver sozialer Tatsachen, da deren Beobachtung immer schon konfundiert sei durch wissenschaftliche Theorien oder auch Alltagstheorien der Beobachtenden. Insofern würden in der Empirie „Annahmen über die Wirklichkeit“ nicht mit Ergebnissen „reiner Beobachtung“, sondern mit bereits „voraussetzungsvollen Beobachtungen“ verglichen (Campbell 1982). Allerdings stünden soziale Werte – gefasst in Leitlinien der Politik, Programmziele oder auch operationalisierte Handlungsziele der Programmdurchführenden – außerhalb der Reichweite von Evaluation. Sie seien als Basis und Bezugsrahmen der Evaluation zu akzeptieren und zentrale Quelle für Maßstäbe (*criteria*), an denen der Wert von Programmen beurteilt werden solle.

---

32 Heute: New Directions for Evaluation.

House hat sich als Evaluationstheoretiker bereits seit Ende der 70er Jahre intensiv mit der Relevanz und Funktion von Werten (*values*) in der Evaluation beschäftigt und geht einen Schritt weiter: Er bezeichnet im Anschluss an Campbell „Behauptungen über Tatsachen" (*fact claims*) als zentral für die Steuerung von Evaluationen und ergänzt diese um *value claims* - also „Behauptungen über Werte", die sich in der Evaluation immer wieder mit „Behauptungen über Tatsachen" vermischten. Was in einem Kontext – „Christoph Columbus entdeckte Amerika" – als „pure" Tatsachenbehauptung erscheint, enthält in einem anderen Kontext – dem ethnisch sensiblen Diskurs etwa über die Darstellung US-amerikanischer Geschichte in einschlägigen Lehrbüchern – eine starke Wertespannung. House betont im Unterschied zu den wertedistanzierten Evaluationstheoretikern die Möglichkeit und Notwendigkeit, auch „Wertebehauptungen" dem rationalen Diskurs zugänglich zu machen. Er sieht dabei allerdings erheblichen methodologischen Entwicklungsbedarf:

> *„Genau so, wie wir über die Jahre ausgeklügelte Verfahren zur Überprüfung von Tatsachenbehauptungen entwickelt haben, müssen wir Verfahren entwickeln, um Behauptungen zu sammeln und zu verarbeiten, die starke Wertaspekte enthalten, so dass unsere evaluativen Schlussfolgerungen auch in Bezug auf diese Behauptungen unverzerrt sind. Gegenwärtig vermischen sich diese beiden Arten von Behauptungen im Rahmen von Evaluations-Studien." (House 1999, S. 313)*

Noch weiter geht die konstruktivistische Evaluation (sie nennt sich selbst auch *Fourth Generation Evaluation,* vgl. Guba/Lincoln 1989): Sie will gerade Werten derjenigen Gruppen, die nicht durch die Finanziers oder die Programmleiter/-innen repräsentiert werden, eine Stimme und damit Durchsetzungskraft geben:[33]

> *„Fourth Generation Evaluation versucht einige Machtungleichgewichte zu korrigieren, indem Stakeholder einbezogen werden, die in früheren Generationen der Evaluationspraxis an die Seitenlinie gestellt worden sind; und sie will ihnen eine Stimme geben." (Lincoln 2003, S. 80)*

Mit den Evaluationsmodellen der 80er Jahre werden die ‚Stakeholder' zu den zentralen Akteuren in der Evaluation: Alle diejenigen, welche an der Finanzierung, an Fortführungsentscheidungen, an der Planung oder der Durchführung von Pro-

33 Vgl. als Beispiel aus dem Bereich ethnischer Minoritäten Clayson u. a. (2002).

grammen beteiligt sind und diejenigen, die von ihren Auswirkungen – sei es positiv oder negativ – betroffen sind, sind Adressaten/-innen der Evaluationsergebnisse und deren potentielle Nutzer/-innen. Nutzung von Evaluation – so die These der 90er Jahre – kann durch angemessenen Einbezug der Programm-Stakeholder in die Evaluation erreicht werden (vgl. exponiert Stufflebeam 1972 und Patton 1997).

> *„Prinzipien sind der Einbezug aller relevanten Stakeholder-Perspektiven, -Werte und Interessen in die Studie; extensiver Dialog zwischen den Evaluatoren/-innen und den Stakeholdern, und manchmal auch zwischen den Stakeholdern untereinander; und extensive Abwägungen und Beratungen, um zu gültigen Schlussfolgerungen zu kommen." (House 1999, S. 314)*

Parallel dazu breiten sich partizipative Evaluationsmodelle aus, die Partei nehmen für die einflussschwachen Gruppen, namentlich für die auf soziale Leistungsprogramme angewiesenen armen Bevölkerungsgruppen.

Cousins und Whitmore (1998, S. 6) sehen eine Bandbreite partizipativer Evaluationsmodelle, von den Stakeholder- oder nutzungsorientierten, die sie der *Practical Participatory Evaluation* zuordnen, bis zur *Transformative Participatory Evaluation*, der es stärker um die Demokratisierung sozialen Wandels geht. Dieser wird insbesondere dadurch vorbereitet, dass die (sozial benachteiligten) Zielgruppen zu Akteuren/-innen in der Evaluation werden. Für deren persönliche Kompetenzen und Machtposition in der Gesellschaft soll Evaluation stärkend wirken. Ein herausragendes Modell ist die *Empowerment Evaluation*, zu Beginn der 90er Jahre von David Fetterman (1993) entwickelt (vgl. Kap. 2.3.4.1; vgl. auch Fetterman/Kaftarian/Wandersman 1996; Fetterman 2003; Mertens 2003).

Eine zweite Entwicklung, die bereits Ende der 60er Jahre einsetzte (Suchman 1967), ist die Steuerung von Evaluationen durch „Programmtheorien". Diese verstehen sich als Alternative zu den aus der wissenschaftlichen Grundlagenforschung hervor gehenden Theorien, die allgemein gültige, auf einem System widerspruchsloser, empirisch geprüfter Hypothesen beruhende soziale Gesetzmäßigkeiten darstellen wollen, welche eine wissenschaftliche und wertneutrale Steuerung von Programmen ermöglichen sollen (vgl. Friedrichs 1973, S. 50ff).

Mit Programmtheorien soll ein Modell der Handlungs- oder Ablauf-Logik konstruiert werden, die dem Programm zu Grunde liegt. Dieses ist der Kern jedes Programm-

Konzeptes, das in vielen Fällen nicht explizit vorliegt sondern in den impliziten Annahmen, den Routinehandlungen und dem selbstverständlichen Fachwissen der Akteure/-innen (besonders der Programmmanager/-innen und Mitarbeiter/-innen) enthalten ist. Logische Modelle formalisieren und veranschaulichen Annahmen darüber, wie die Programm-Inputs und der Programm-Prozess die angezielten Programm-Outputs und -Outcomes „herstellen“ (vgl. Abb. 8).

Abschließend sei herausgestellt, dass die Entstehung und Entwicklung von Evaluationsmodellen vielfach verbunden ist mit Programmen zur Bekämpfung von Armut. Diese Programme richteten sich schon immer auf Ursachen, Ausprägungen, Begleiterscheinungen und Folgen von Armut in verschiedenen Lebenslagen:

- lokale soziale Integration als Strukturbedingung für die Entstehung und Verfestigung von Armutslagen (Evaluation von Community Building Programmen),
- schicht- und ethnienspezifische Benachteiligungen im Erziehungs- und Bildungswesen als Auslöser für Armut (z.B. Evaluation der Programme Head Start),
- Einkommens- und Erwerbsdefizite als unmittelbarer Ausdruck von Armut (Evaluation der AFDC-Programme / der „welfare to work” Programme),
- gesundheitliche Beeinträchtigungen als Folge von Einkommensarmut und sozialer Deprivation (Evaluation der public-health-Programme).
- Obdachlosigkeit, Beschaffungskriminalität oder Prostitution als extreme Ausprägungen verfestigter Armuts-Lagen (z.B. Evaluation von Programmen gegen Obdachlosigkeit).

Strukturelle Grundlage ist ein System der sozialen (Grund-)Sicherung, das sich vom deutschen stark unterscheidet und vergleichsweise instabil und kurzfristigen Politikwechseln gerade auch in den Bundesstaaten unterworfen ist. Dies löst immer wieder neue Initiativen und Politikansätze aus, deren Wirkungen durch Evaluationen zu überprüfen sind.

2.1.2 Deutschland: Fortgeschrittene Sozialberichterstattung – Entwicklungsbedarf bei Evaluation

„Eine genaue Analyse der sozialen Wirklichkeit in Deutschland ist notwendig, um Armut zielgenauer entgegenwirken und gesellschaftspolitische Reformmaßnahmen zur Stärkung sozialer Gerechtigkeit und gleicher Chancen für die Menschen ergreifen zu können."
(Armuts- und Reichtumsbericht 2001, S. XXXV)

In Deutschland gibt es keine vergleichbar intensive Entwicklung von Evaluationsmodellen, weder im Allgemeinen noch bezogen auf den Bereich der Armutsvermeidung und -verminderung. Dabei gibt es hier eine Tradition dichter, nicht selten durch stark innovative Forschungsansätze geleisteter Beschreibungen und Erklärungen des *Problems* soziale Desintegration, Ausschluss oder Armut. Diese Geschichte kann und soll hier nicht nachgezeichnet werden, doch sei exemplarisch auf einige Studien verwiesen (vgl. ausführlicher Hauser/Neumann 1992).[34]

Erste umfassende Untersuchungen zur Armut in Deutschland, z.B. die Enquêten des Vereins für Socialpolitik, nahmen die an Karl Marx und Max Weber orientierten Klassenanalysen auf und bemühten sich um die Beantwortung der „Sozialen Frage" nach den gesellschaftlichen Folgen des Industrialisierungsprozesses.

Die Studie über die „Arbeitslosen von Marienthal" von Marie Jahoda (1933) stellt mit ihrem multimethodischen Vorgehen (Befragung, Beobachtung, non-reaktive Verfahren) einen Qualitätssprung in der empirischen Erfassung deprivierter Lebenslagen dar[35]. Ihr Werk regt die Forschung insbesondere über die psychischen Folgen der Arbeitslosigkeit bis heute an. Aus der kritischen Auseinandersetzung mit Jahodas Pionierwerk ist u.a. die „differentielle Arbeitslosenforschung" hervorgegangen, zu der Alois Wacker (2000) bemerkt:

> *„Entsprechend finden sich in der Mehrzahl der neueren Studien mehr oder minder kompliziert gebaute typologische Ordnungssysteme zur Beschreibung der unter-*

---

34 Hauser/Neumann unterscheiden fünf Phasen sozialwissenschaftlicher Beschäftigung mit Armut seit dem II. Weltkrieg: Armut der Nachkriegszeit; Latenzphase der Armutsforschung in den 60ern, Diskussion um Randgruppen; Neue Soziale Frage im Zusammenhang mit der Geißler-Studie; Konzepte der multiplen Deprivation und Lebenslagen.

35 Die Studie – in Österreich durchgeführt und in Deutschland veröffentlicht – fand hier breite Rezeption.

schiedlichen Belastungsgrade und Bewältigungsstile. ... Je nach Konstellation bedeutsamer belastender und entlastender Faktoren und in Abhängigkeit von der berufsbiographischen Position können die subjektiv erfahrenen Belastungen erheblich differieren. ... Die differentielle Arbeitslosenforschung macht zum einen das Gelände der Arbeitslosigkeitsfolgen vielgestaltiger und damit unübersichtlicher; sie erlaubt aber zugleich eine genauere Eingrenzung der Problemgruppen des Arbeitsmarktes und eine passgenauere Planung von Reintegrationsmaßnahmen als bisher." (S. 56).

Mit dieser Schlussfolgerung verweist Wacker implizit auf das bestehende evaluationstheoretische Defizit.

Renate Mayntz legte 1952 einen innovativen Ansatz zur empirischen Gemeindesoziologie. Es handelt sich um ihr Stadtportrait „Wandel einer Industriegemeinde", in dem sie den ökonomischen und demographischen Wandel einer Mittelstadt nachzeichnete und die soziale Integration der Bewohner sowie die soziale Ungleichheit in der Stadt untersuchte. Jürgen Friedrichs würdigt dies 2002 mit der Veröffentlichung der „Replikationsstudie", die wiederum einen Schwerpunkt setzt beim Thema „Soziale Ungleichheit und Lebensstile in Euskirchen". Renate Mayntz (2002, S. 204) resümiert in ihrem Nachwort:

„Trotz einer erkennbaren Tendenz der Konzentration auf mittlere Statuslagen hat sich die Kurve der Statusverteilungen in ihrer Form nicht grundsätzlich verändert ... Damit hängt auch die empirisch belegte Tatsache zusammen, dass Niveauveränderungen durch politische Interventionen leichter zu bewerkstelligen sind als Strukturveränderungen."

Der Schlusssatz von Mayntz enthält eine für Strategien sozialer Integration verfolgenswerte These. Bemerkenswert ist die Verbindung, die zwischen sozioökonomischer Benachteiligung und sozialer Segmentation („Verinselung sozialer Netzwerke") insbesondere für Aussiedler sowie an- und ungelernte Arbeiter festgestellt wird. Diese „... verfügen nur über ein geringes ökonomisches, instrumentelles, soziales und kulturelles Kapital, ... (was es ihnen schwer macht) ... sich selbst aus einer marginalen Position zu befreien." Friedrichs/Kecskes/Wolf (2002, S. 200) sehen Handlungsbedarf auf kommunaler Ebene, um „... einem weiteren Abrutschen von Teilen dieser Bevölkerungsgruppe entgegenzuwirken". Hinweise, wie derartige Programme zu gestalten wären, liefert die Studie nicht.

Nach der Umsetzung des am 1. Juni 1962 in Kraft getretenen Bundessozialhilfegesetzes nahm die sozialwissenschaftliche Thematisierung von Armut deutlich ab (Situation der „bekämpften Armut“). Mitte der 70er Jahre kommt es in der Folge der mit der ersten „Ölkrise“ verbundenen sozialen Folgen wiederum zu einer Intensivierung der beschreibenden Armutsforschung, teils unter veränderter Perspektive. So richtet Geißler (1975) den Blick auf bestimmte Teilgruppen, die, obwohl oder da die alte „Soziale Frage“ weitgehend gelöst zu sein schien, von Armut betroffen sind (kinderreiche Familien, unverheiratete oder verwitwete Frauen im Rentenalter). Gleichzeitig wird die materielle Armut im Zusammenhang mit dem sozialwissenschaftlichen Interesse an „Randgruppen“ wieder entdeckt. Dies geht einher mit der Thematisierung der „vielfältigen Dimensionen von Ungleichheit ... (z.B. in den Bereichen Bildung, Wohnen, Gesundheit)“ (Hauser/ Neumann 1992, S. 240), einer Wurzel des Lebenslagenansatzes in der Armutsforschung.

Neue Technologien der elektronischen Datenverarbeitung befördern schließlich die „Sozialindikatoren-Bewegung“, die in den 70er Jahren in Deutschland floriert (Ballerstedt/Glatzer 1979). Mit der umfassenden und differenzierten Entwicklung einzelner objektiver und subjektiver Indikatoren sowie der Konstruktion von Indikatorensystemen und Indizes können Armutsphänomene in Deutschland sowohl regional differenziert als auf hoher Aggregationsebene beschrieben und in ihrem Zeitablauf analysiert werden (vgl. Noll 2002).

Angestoßen durch die Sozialindikatorenforschung, kommt es zu weiteren Differenzierungen in der Armutsforschung, stichwortartig benannt mit „Einbezug subjektiver Indikatoren“, der „Multidimensionalität“ der betrachteten Lebensbereiche über die Einkommensarmut hinaus sowie „multiple Verursachung“ von Armut (Arbeitslosigkeit, Trennung und Scheidung, Krankheit, Mangel an sozialem Kapital ...). Dies mündet ein in eine intensivierte „Sozialberichterstattung“ und schließlich die vorrangig von den Wohlfahrtsverbänden bzw. dem Deutschen Gewerkschaftsbund initiierten „Armutsberichte“ der 90er Jahre (Hanesch u.a. 1992, Hauser/Hübinger 1993; Hübinger /Neumann 1997 sowie AWO 2000). Die von der AWO und dem Institut für Soziale Arbeit und Sozialpädagogik in Frankfurt a.M. durchgeführte Studie mit besonderem Schwerpunkt bei der Lebenslage von Kindern steht exemplarisch für die Tendenz, differenzierte Analysen für bestimmte armutsbetroffene Teilgruppen vorzunehmen (Kinder und Jugendliche, Frauen, alte Menschen, Alleinerziehende, Familien,

Migranten/-innen, Behinderte, Obdachlose ...). Dies ist eine Betrachtungsweise, die in den USA – einem Land ohne eine allgemeine soziale Grundsicherung – bereits länger verbreitet ist.

Gemäß dem Beschluss des Bundestages vom 27. Januar 2000 wird nach ca. zweijähriger Vorbereitungszeit am 25. April 2001 der Bericht „Lebenslagen in Deutschland – der erste Armuts- und Reichtumsbericht“ durch die Bundesregierung verabschiedet (Kuck-Schneemelcher 2001).

> *(Die Berichterstattung) „... hat das Ziel, ein differenziertes Bild über die soziale Lage in Deutschland zu geben. ... Mit ihrer Gesamtschau der sozialen Wirklichkeit eröffnet sie eine systematische Verzahnung verschiedener Politikbereiche. Sie hat die Aufgabe, materielle Armut und Unterversorgung sowie Strukturen der Reichtumsverteilung zu analysieren und Hinweise für die Entwicklung geeigneter politischer Instrumente zur Vermeidung und Beseitigung von Armut, zur Stärkung der Eigenverantwortlichkeit sowie zur Verminderung von Polarisierungen zwischen Arm und Reich zu geben.“ (Armuts- und Reichtumsbericht 2001, S. XIV)*

Diese Formulierung lässt erkennen, dass auf der Basis dieses ersten Berichtes Programme und Maßnahmen entwickelt und verstärkt zum Einsatz kommen sollen, die gezielt Vermeidung und Überwindung von Armut sowie Integration in Deutschland unterstützen und dabei Tendenzen zur Verstärkung sozialer Ungleichheit entgegenwirken. Welcher Stellenwert dabei den vorgelegten tiefgehenden und differenzierten Analysen zukommt, um Erfolg versprechende Programme systematisch zu konzipieren, bleibt offen.

Auf Länderebene werden speziell auf Armutslagen – im Sinne des Lebenslagenkonzeptes – gerichtete Programme durchgeführt. („Hamburger Armutsbekämpfungsprogramm“, „Hessisches Projektnetz Wohngebiets- und Stadtteilmanagement“). Das neu geschaffene Bund-Länder-Programm „Stadtteile mit besonderem Entwicklungsbedarf – die soziale Stadt“ greift diesen Ansatz auf.

Dabei ist die „Programmförmigkeit“ im Sinne einer entwickelten „Programmtheorie“ ansatzweise ausgeprägt.

Abbildung 8: Ziele von „Soziale Stadt“

a) die Kooperation der verschiedenen Fachressorts der Gemeindeverwaltung, um durch den abgestimmten bzw. koordinierten Einsatz der finanziellen, planerischen und intervenierenden Instrumente bauliche, wirtschaftliche und soziale Verbesserungen in den Problemquartieren zu ermöglichen,

b) eine Beschreibung der Defizite der ausgewählten Quartiere und die Formulierung realistischer Ziele für die angestrebte Quartiersentwicklung,

c) die Schaffung eines leistungsfähigen Stadtteilmanagements zur Bündelung und zum zielgenauen Einsatz der verfügbaren Ressourcen durch Heranziehung qualifizierter Gebietsmanager und Entwicklungsträger sowie

d) die Sicherstellung einer aktiven Bürgerbeteiligung, da erst durch diese das angestrebte aktivierende Moment des Entwicklungsprozesses in Gang kommen kann.

Quelle: nach Hanesch 2001, S. 43ff

Diese groben Leitziele eines nationalen Programms müssen auf der lokalen Ebene konkretisiert und in konkrete Aktionspläne umgesetzt werden. Hierfür macht der Zielkatalog hauptsächlich *prozedurale* Vorgaben (Prozess-Dimension): Es ist auf lokaler Ebene ein Konzept zu entwickeln, das auf einer Bedarfsanalyse beruht (b). Außerdem sollen als neue Elemente der Strukturqualität die kommunalen Ressourcen mit ihren jeweiligen Instrumenten verstärkt kooperieren (a) und es soll ein Stadtteilmanagement eingerichtet werden (c). Auf Basis der (theoretischen) Annahme, dass Partizipation der Bürger und Bürgerinnen ein wichtiges Prozesselement ist, soll Bürgerbeteiligung hierfür sichergestellt werden (d).

An diesem Beispiel wird deutlich, dass Evaluation für die dezentrale Umsetzung eines nationalen Programms zunächst in ihrer klärenden Funktion angefragt ist: Sie kann auf dem Hintergrund der grob orientierenden strukturellen und prozeduralen Prinzipien die lokalen Akteure unterstützen, eine je angepasste Programmtheorie zu entwickeln, welche die verschiedenen Programmdimensionen – angefangen von Konzept und Struktur über Prozess bis hin zu Resultaten/Wirkungen – differenziert beschreibt. Dabei ist zentral, dass diese Programmtheorie sowohl durch die fachlichen Annahmen und Handlungsprinzipien der dort tätigen Gebietsmanager und weiterer Fachkräfte als auch durch empirisch gewonnene Informationen über die konkreten Bedarfslagen vor Ort, darunter auch zum Income, also zu den durch die Bewohner und Bewohnerinnen mitgebrachten Ressourcen, fundiert ist.

Ein solches Herunterarbeiten von nationalen auf dezentrale Programmebenen mit Hilfe der Programmtheorie ist für die Bundesrepublik Deutschland noch recht neu. An dem Beispiel „Soziale Stadt“ wird deutlich, dass Wirkungsorientierung in der Evaluation bereits bei der Konzeptentwicklung und Prozessbegleitung angelegt sein muss. Einerseits kann damit die Wahrscheinlichkeit erhöht werden, dass die Programmumsetzung die spezifizierten Ziele erreicht. Andererseits kann die wirkungsfeststellende Evaluation auf die im klärenden und interaktiven Begleitprozess gewonnene Bewertungskriterien aufbauen. Nur für Programme, die von vornherein wirkungsorientiert konzipiert und umgesetzt werden, macht Wirkungsfeststellung Sinn, es sei denn, man beschränkt das Nutzungsinteresse auf ein einfaches „erfolgreich“ oder „erfolglos“. In der Regel gibt es ein ausgeprägtes sozialpolitisches Interesse, vorhandene Mittel mit hoher Sicherheit zur Auslösung gewünschter Wirkungen einzusetzen und aus den jeweils festzustellenden Stärken und Schwächen für künftige Programmentwicklungen und -umsetzungen zu lernen.

In diesem Sinne ist auch der in der Ausschreibung zu dieser Perspektivstudie fixierte Auftrag aus dem Kontext der Armuts- und Reichtumsberichterstattung zu verstehen:

> *„In Vorbereitung des zweiten Armuts- und Reichtumsberichts und der weiteren zukünftigen Berichterstattung sind auch auf diesem Gebiet grundlegende Forschungsaktivitäten notwendig. Die bisherige Konzeption sieht vor, im Rahmen der Grundlagenforschung das „fundamentale Evaluationsproblem“ zu thematisieren. Ziel eines solchen Forschungsprojekts muss es sein, die tragfähige theoretische und konzeptionelle Basis für eine fortlaufende und begleitende Wirkungsforschung bei der Armuts- und Reichtumsberichterstattung zu entwickeln. Zu klären ist hierbei insbesondere, wie die Wirkungsforschung zu operationalisieren ist (Messkonzepte, Evaluation) und inwieweit verwendbares Datenmaterial zur Verfügung steht.“ (Quelle: BMGS)*

Einer derartigen Evaluationstheorie, der es auf möglichst unmittelbar nützliche Beiträge bei der Findung Erfolg versprechender Lösungen für praktische sozialpolitische Probleme ankommt, wird in Deutschland erst seit Mitte der 90er Jahre wieder verstärktes Interesse gewidmet. Lange dominierte ein von den akademischen Einzeldisziplinen, insbesondere der Soziologie und der Psychologie, geprägtes Evaluationsverständnis, das vorrangig an kumulativem theoretischem Wissen über den jeweiligen Gegenstandbereich interessiert ist.

Jürgen Friedrichs nimmt in seinem – bis heute weiterhin verbreiteten – Lehrbuch zur „Empirischen Sozialforschung“ (1973, S. 375) die damalige in den USA noch starke empirisch-analytische Evaluation bereits wahr und skizziert sie in einem kurzen Abschnitt. Sein Werk, das im Rahmen der Stadtsoziologie immer auch Armutsthemen zum Gegenstand hatte, hat für Jahrzehnte Standards gesetzt, gerade auch für die angewandte Sozialforschung. Darin kommt dem Begründungszusammenhang (Theorie, Begriffe, Hypothesen, Variablen, Datenerhebung, Auswertung etc.) die zentrale Stellung zu. Die beiden für Evaluationen mindestens gleichrangigen Komponenten – der Entdeckungszusammenhang (Wie wird festgelegt, was die relevanten Fragestellungen sind? Wie werden dabei soziale Werte und Konflikte berücksichtigt?) und der Verwertungszusammenhang (Wie werden die gewonnenen Informationen und Erkenntnisse genutzt? Wie wird konkret die Umsetzung des gewonnenen Wissens für die Optimierung sozialer Programme vorbereitet?) – werden eher knapp behandelt (Friedrichs 1973, S. 50ff).

In dieser Tradition steht das insbesondere für die Psychologie führende Lehrbuch von Jürgen Bortz und Nicola Döring „Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler“. Sie lokalisieren die Evaluation wie folgt: „Die Evaluationsforschung befasst sich als ein Teilbereich der empirischen Forschung mit der Bewertung von Maßnahmen und Interventionen (2002, S. 101). Sie nehmen Bezug auf die Definition von Rossi und Freeman. In der aktuellen 6. Auflage heißt es:

> *„Programmevaluation ist der Einsatz von sozialwissenschaftlichen Forschungsmethoden, um systematisch die Effektivität von sozialen Interventionsprogrammen zu untersuchen ... Evaluationsforscher (Evaluatoren) nutzen soziale Forschungsmethoden, um soziale Programme zu untersuchen, zu bewerten und sie verbessern zu helfen – und das in all ihren relevanten Aspekten, einschließlich der Diagnose des sozialen Problems, auf das sie sich beziehen, ihre Konzipierung und ihr Design, ihre Umsetzung, ihre Outcomes und ihre Effizienz.“ (Rossi/Freeman/Lipsey 1999, S. 4)*

Auch Rossi und Kollegen schreiben der Evaluation sowohl gestaltende als auch bilanzierende Funktionen in den verschiedenen Phasen des Programmzyklus“ zu. Ebenso wie diese amerikanischen Autoren beharren auch Bortz/Döring auf dem Terminus „Evaluationsforschung“ (vs. Evaluation), da aus ihrer Sicht darin der unbedingt einzuhaltende Anspruch aufgehoben ist, „... dass Evaluationen wissenschaftlichen

Kriterien genügen müssen, die auch sonst für empirische Forschungsarbeiten gelten“ (Bortz/Döring 2002, S. 102).[36]

Das folgende Kapitel 2.2 legt dar, dass sich aus der Spannung zwischen einer starken „Wissenschaftsorientierung“ auf der einen, einer „Nützlichkeitsorientierung“ auf der anderen Seite sehr unterschiedliche Evaluationsmodelle mit jeweils relativen Stärken und Schwächen ausdifferenziert haben. Die systematische Beschreibung und Bewertung von Programmen und Maßnahmen zur Armutsvermeidung und -minderung wird in Zukunft auf diese erweiterte Palette durch unterschiedliche Evaluationstheorien gesteuerter Vorgehensweisen zurückgreifen können.

2.1.3 Von der Problembeschreibung zur Konkretisierung staatlicher Ansätze zur Lösung von Armut

Der Auftrag des Bundestages, in Zukunft verstärkt die Wirkungen von Politiken und Programmen zu betrachten, welche die Ressourcenverteilung und die Lebenslagen armer und von Armut bedrohter Menschen in Deutschland verbessern wollen, macht eine lösungsorientierte Erweiterung der bisherigen Perspektive der Armuts- und Reichtumsberichterstattung erforderlich.

Bislang stehen die Definition von Armut und die Messung der so definierten Armut mittels fortgeschrittener Analysemodelle und Datenzugänge im Vordergrund der sozialwissenschaftlichen Bemühungen. Wenn es nun in einem weiteren Schritt um eine wissenschaftlich fundierte Beschreibung und Bewertung von Lösungen der erforschten Armutsprobleme geht, ist eine positive Zielbeschreibung in den zu ergreifenden Politiken und Programmen unabdingbar:

„Armutsvermeidung“ oder „Armutsverminderung“ kennzeichnen Vermeidungsziele, die einen (nicht operationalisierbaren) Zustand angeben, der *nicht* eintreten soll oder *nicht* mehr vorliegen soll. Es bleibt unklar, wie der alternativ gewünschte und angestrebte Zustand aussehen soll. „Armutsvermeidung“ lässt z.B. offen, ob die jetzt „armen“ Erwerbsfähigen durch staatliche Transfereinkommen über die Armutsschwelle gehoben werden sollen oder – zu einem in der Zielformulierung festzulegenden Anteil – durch Aufnahme einer Erwerbsarbeit oder durch nicht-erwerbsförmige Selbsthilfe. Das Vermeidungsziel lässt offen, ob Alleinerziehende kleiner

36 Vgl. hierzu relativierend die Standards für Evaluation, Kap. 1.2.

Kinder allein durch einen wie auch immer gestalteten Lastenausgleich oder durch parallele Teilzeit-Erwerbsarbeit die Armutsschwelle überwinden sollen. Es lässt offen, ob nicht-deutsche Arme durch geeignete Maßnahmen aus ihrer Situation bei Verbleib im Lande heraus geholfen werden soll oder ob die Zuwanderung armutsbedrohter Ausländer/-innen verringert oder die Auswanderung armutsbetroffener Ausländer gefördert werden soll.

Die Wertkonflikte, die bei politischen Entscheidungen über derartig spezifizierte Zielsetzungen auftreten, sind augenfällig und spielen in den aktuellen sozialpolitischen Auseinandersetzungen eine gravierende Rolle. Politische Programme und Maßnahmen bleiben jedoch orientierungslos, wenn ihre angezielten Wirkungen nicht als Ziele klar beschrieben werden, die selbst wiederum eine transparent zu machende Wertgrundlage haben.

Das Augenmerk ist somit darauf zu richten, welche positiven, möglichst operational beschreibbaren Ausprägungen in den verschiedenen Lebenslagen durch die einschlägigen Programme und Politiken angestrebt werden. Es ist ganz offensichtlich, dass die eigentliche Konkretisierung, der Zuschnitt auf die besonderen Kontextvariablen z.B. bestimmter Stadtquartiere nur auf der dezentralen Ebene möglich sind. Hier muss vorrangig die Konkretisierung der Programme geleistet werden; auf dieser Ebene findet auch vorrangig die Empirie von Evaluationen statt.

Programmentwickler/-innen sind daher darauf zu verpflichten, die auf Vermeidung und Überwindung von Armut gerichteten Ziele der von ihnen konzipierten Programme präzise zu formulieren und die Umsetzung der Programme daran auszurichten.

Die Politik, insbesondere auf Bundes- und Landesebene, ist gefordert, Leitziele vorzugeben und auf der Konkretisierung der Ziele auf den dezentralen Ebenen zu bestehen. Perspektive sollte es sein, derartige verknüpfte Zielsysteme von der Bundespolitik über die Landesebene bis zu deren Konzeption und schließlich Umsetzung auf der lokalen Ebene sichtbar zu machen – als Grundlage einer Politik der sozialen Integration.

2.2 Typologie von Evaluationsmodellen gemäß der Dimension „Werteberücksichtigung“

„Eine empirische Wissenschaft vermag niemanden zu lehren, was er soll, sondern nur was er kann und – unter Umständen – was er will.“ (Weber 1951, S. 151).

Wie bereits angesprochen, sind die durch die Armuts- und Reichtumsberichterstattung betrachteten Politikfelder in besonderem Maße durch Wertdifferenzen geprägt, z.B. bezüglich der Festlegung der Höhe des Lebensstandards, der für eine Armutsgrenze zu setzen ist. Auch die Festlegung, welche Arten von Wirkungen im Rahmen empirischer Evaluationsstudien mit Vorrang zu überprüfen sind, ist wertegeprägt: Sollen vorrangig subjektive Zufriedenheiten der Zielgruppen, ihre „objektiven“ Lebenslagen oder die Wirkungen des zu evaluierenden Programms auf die öffentlichen Haushalte oder die Volkswirtschaft untersucht werden? Wirkungsdimensionen wie „subjektive Befindlichkeit der Zielgruppen“, ihre Lern- und Entwicklungsprozesse in verschiedenen Lebenslagen-Dimensionen, ihr Eingliederungsverhalten in den Arbeitsmarkt oder die Beeinflussung der Wirtschaftsleistung von Regionen stehen je nach sozialer Wertposition der Fragenden oder Urteilenden mehr oder weniger stark im Vordergrund und können diesen entsprechend auch ganz unterschiedlich operationalisiert werden. Was für die einen Beteiligten eine positiv besetzte Wirkung ist – z.B. ein häufiger Arbeitsplatzwechsel bei verschiedenen Formen der Beschäftigung (geringfügig, sozialversicherungspflichtig, selbstständig) – mag für eine andere Beteiligtengruppe negativ sein, weil aus ihrer Sicht Grundwerte sozialer Sicherheit als Arbeitnehmer/-in verletzt werden. Die zu leistende Interpretation von Daten geschieht immer wertegebunden – Daten allein geben keine Orientierung.

> *„Werte sind im beschreibenden Sinne allgemeine Neigungen von Menschen, bestimmte Umstände anderen Umständen vorzuziehen, und sind im normativen Sinne Aufforderungen an Menschen, dies zu tun. ... Mit dem Begriff Werte werden sowohl die tatsächlichen Dispositionen von Menschen beschrieben, wie auch solche, die mit unterschiedlichen Begründungsansprüchen als verbindlich gelten oder angefordert werden.“ (Meyer 1999, S. 259)*

Vielfach wird die große Bedeutung von Werten erst in interkultureller Betrachtung offensichtlich, etwa was das Geben und Nehmen sozialer Sicherheit zwischen den

Generationen untereinander oder den Umgang mit erwerbsunfähigen Minderheiten betrifft. Die Definition von Meyer geht auf den Kulturanthropologen Clyde Kluckhohn (1967) zurück, der Werte als langfristig treibende Faktoren bei der Entscheidung für Handlungsziele und Handlungsarten identifiziert.

Werte liegen sozialpolitischen Interessen und Handlungen zu Grunde und orientieren diese langfristig. Sowohl auf individueller wie auf kollektiver Ebene sind sie relativ stabil und verändern sich schrittweise und langsam, es sei denn, soziale Krisensituationen bewirken einen plötzlichen Umschlag (wobei sich nach Abklingen der Krise die ursprüngliche Wertekonstellation wieder herstellen kann). Geteilte soziale Werte rekurrieren vielfach auf Formulierungen in Verfassungen („Jeder hat das Recht auf freie Entfaltung seiner Persönlichkeit ... “) oder auf den traditionalen Normenbestand einer nationalen oder regionalen Kultur.

Soziale Werte werden oft schlagwortartig formuliert (Freiheit, Gerechtigkeit, Sicherheit) und haben dann geringe Differenzierungskraft zwischen unterschiedlichen politischen oder sozialen Milieus.

Wenn im Rahmen von Evaluationen Werte geklärt werden, ist es erforderlich, diese zu spezifizieren, um zu prüfen, ob tatsächlich Konsens oder doch (starker) Dissens in Bezug insbesondere auf Interpretationen (= wertebasierte Erklärungen) von Daten vorliegt. Werte sollten dabei (ebenso wie Ziele) „positiv“ formuliert werden, um erkennbar und damit orientierend für die Planung und Durchführung von Programmen wie Evaluationen zu sein. „Vermeidungswerte“ wie „Armut ist eine Schande für die Gesellschaft“ lassen keine Richtung erkennen, auf die sich sozialpolitisches Handeln beziehen soll. Der Grundwert des Art. 1 (1) des Grundgesetzes für die Bundesrepublik Deutschland, „Die Würde des Menschen ist unantastbar“, erhält erst durch den Nachsatz „Sie zu achten und zu schützen ist die Verpflichtung aller staatlichen Gewalt“ seine Richtung.

An Politiken und Programmen, die sich auf Vermeidung und Überwindung von Armut richten, gibt es ganz unterschiedliche Ansprüche, die auf verschiedenen bis hin zu antagonistischen Werten basieren. So können z.B. politische Maßnahmen einer sozialen Grundsicherung auf Basis folgender beider konkurrierender Werte konzipiert und auch evaluiert werden: „Das Recht auf kulturelle Teilhabe hat jeder und jede unabhängig von seinem Beitrag für das Ganze“ oder „Leistungen der Gemeinschaft

erfordern Gegenleistungen von denen, die unterstützt werden, damit die Gesellschaft lebensfähig bleibt“.

Da die Frage, was als anzustrebendes Ziel, was als Erfolg und schließlich was als positive Wirkung politischen Handelns zu bezeichnen ist, von der Wertposition abhängt, welche der/die Beurteilende einnimmt, ist es erforderlich, die Verantwortlichkeit für die Werteklärung im Rahmen von Evaluationen auszuweisen. Dies gilt in ganz besonderem Maße für Evaluationsgegenstände, die durch Interessengegensätze und Verteilungskonflikte geprägt sind. Die Frage des Ob und Wie der Bezugnahme auf soziale Werte durch Evaluationsverfahren ist daher besonders im Bereich von Programmen zur Vermeidung von Armut von großer Bedeutung.

Systematische Verfahren zur Bestimmung und Ordnung der Wertperspektiven der Stakeholder sind verfügbar. Offen bleibt dabei, ob diese Verfahren durch die Politik/ die Programmverantwortlichen oder (auch) durch die Evaluationsverantwortlichen zu nutzen sind.

Bereits seit Ende der 70er Jahre wird die quantifizierende *Multiattribute Utility Technology* (MAUT) eingesetzt (Edwards/Newman 1982). An die Identifikation der Stakeholder schließt sich eine Phase an, in welcher deren Wertdimensionen oder Wertzuschreibungen (*attributes*) herausgearbeitet werden. Diese werden in eine logische hierarchische Struktur gebracht (sog. Wertebaum). Für jede Beteiligtengruppe wird sodann die relative Wichtigkeit eines jeden Wertes empirisch festgestellt. Damit steht für Interpretationen und Bewertungen des Programms eine gesicherte – in Teilen nach den verschiedenen Stakeholdern divergierende – Basis zur Verfügung. Zentraler Vorteil dieses relativ aufwändigen Verfahrens ist, dass es auch für den Einsatz in nationalen Programmen mit einer Vielzahl lokaler Programmstandorte geeignet ist.

Beywl/Potter (1998) schlagen – basierend auf der in Deutschland entwickelten Moderationsmethode – mit ReNoMo ein Werteklärungsverfahren vor, das einer ähnlichen Logik folgt wie MAUT, aber eher für lokale Standorte und Modellprogramme geeignet ist: In moderierten Gruppendiskussionen arbeiten Vertreter und Vertreterinnen der Stakeholder Wertpositionen heraus, die durch diese in eine Rangfolge gebracht werden und als Grundlage dienen, um die primären Evaluationsfragestellungen festzulegen. Besonders geeignet sind Fragestellungen mit geringer oder mittlerer Werte-

diskrepanz (ähnliche *value claims*[37]) bei hoher Unsicherheit über die Datenlage (starke Diskrepanz bezüglich der *fact claims*).

In den nicht seltenen Fällen, in denen Werte den Akteuren nicht oder nur teilweise bewusst sind, eignet sich schließlich die Fokusgruppen-Methode, um diese diskursiv in homogenen Gruppen heraus zu arbeiten (grundlegend: Krueger 1994).

Im nächsten Kapitel wird der Frage nachgegangen, welche Vorschläge verschiedene evaluationstheoretische „Schulen“ zum Umgang mit Werten in der Evaluation machen. Die Darstellung erfolgt entlang einer Typologie von Evaluationsmodellen. Ausschlaggebend für deren Konstruktion ist, ob Werte und deren Klärung als zu leistender Bestandteil der Evaluation angesehen werden oder ob die Verantwortung dafür außerhalb des Evaluationsprozesses verortet wird. Diese Frage selbst ist eine zentrale wertgeladene Streitfrage, welche die modernen Sozialwissenschaften seit ihrer Entstehung begleitet. [38]

Damit ist die Unterscheidung danach, ob Werte – sei es explizit oder implizit – aus dem Evaluationsprozess ausgeklammert werden oder ob sie als ein zentraler Bestandteil angesehen werden, leitend für die Darstellung und Diskussion verschiedener für den Bereich der Armuts- und Reichtumspolitik relevanter Evaluationsmodelle.

Unter einem Modell der Evaluation wird eine ausformulierte, theoretisch begründete und durch praktische Evaluationserfahrungen gesättigte Anleitung verstanden, wie praktische Evaluationen geplant und durchgeführt werden sollen. Modelle sind oft mit Namen bestimmter Evaluationstheoretiker/-innen verbunden, welche diese erstmals konkretisiert und weiter entwickelt haben. Sie machen insbesondere Aussagen darüber, wie Steuerungsentscheidungen zur Evaluation getroffen werden sollen (wann,

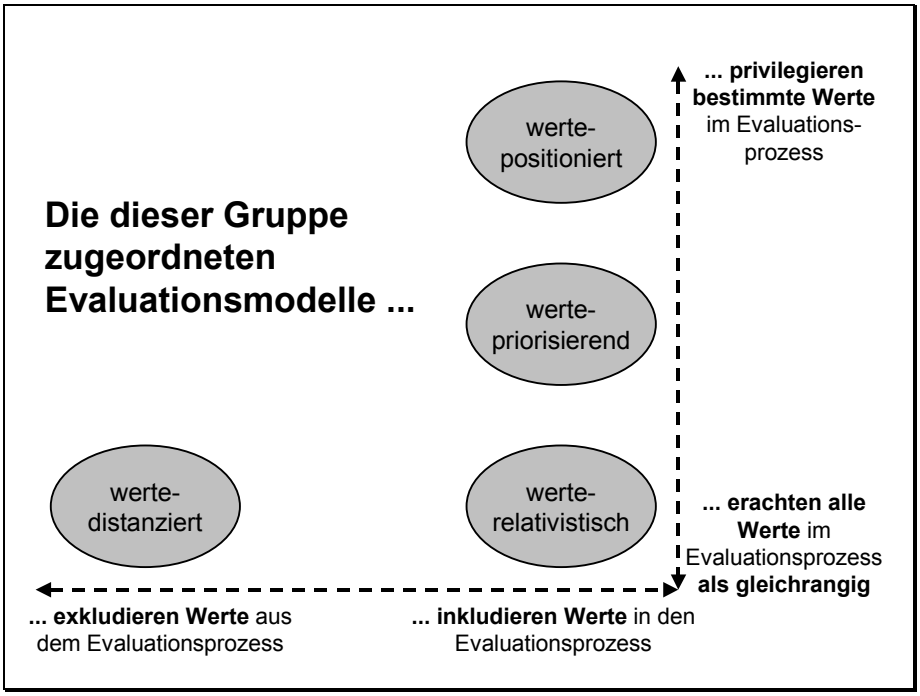
---

37 Vgl. die im Kap. 2.1.1 referierten Ausführungen von House (2001).

38 Es würde hier zu weit führen, soll jedoch zumindest angedeutet sein, da die verbreiteten Widerstände, Werte zum Ausgangs- und Bezugspunkt von Evaluationen zu nehmen, gerade in den deutschen Sozialwissenschaften, wissenschaftshistorisch geprägt sind: Der erbitterte, polemische und bis zur gegenseitigen persönlichen Diffamierung getriebene „Positivismusstreit in der deutschen Soziologie“ (Adorno u.a. 1969) blockiert bis heute die Entwicklung einer eigenständigen, auf die Bundesrepublik Deutschland passenden Evaluationskultur, vielleicht einschließlich der Bereitschaft, die methodologischen Entwicklungen im angelsächsischen Raum gerade im brisanten Feld der Armuts- und Reichtumspolitik zur Kenntnis zu nehmen. Letztlich blieb und bleibt von beiden Konfliktparteien unbeantwortet, wie empirische Wissenschaft praktisch verfahren soll, damit ihre Ergebnisse von Politik und Gesellschaft so angemessen wie möglich interpretiert und genutzt werden

durch wen, mit welchen Verfahren), welche Anforderungen an Design und Methoden der empirischen Untersuchungen zu richten sind und welche Rolle der Evaluation bei der Verbreitung der gewonnenen Ergebnisse zukommt. Darüber hinaus enthalten Modelle unterschiedliche wissenschaftstheoretische Setzungen und wissenschaftsethische Grundorientierungen.[39] Modelle selbst sind gebildete Typen, insofern sie vielfach Ansätze mehrerer Autoren/-innen zusammenfassen, sowohl deren konzeptionellen Aussagen wie ihr praktisches Handeln umgreifen als auch Ähnliches und dabei Unterschiedliches subsumieren.

Abbildung 9: Haupttypen von Evaluationsmodellen gemäß ihrer Berücksichtigung von Werten



Quelle: eigene Darstellung

Von ihrem Selbstverständnis bzgl. der Frage der Wertberücksichtigung her können Evaluationsmodelle wie folgt zu vier Gruppen zusammengefasst werden:

- **Wertedistanzierte** Modelle klammern –z.B. in der Tradition von Max Weber (1951)[40] und Karl Popper (1969) – die Werturteilsfrage i.S. der Bestimmung des sozialpolitisch Wünschbaren aus dem Evaluationsprozess

39 Vgl. Beywl (1988), S. 45 und die dort zitierte Literatur sowie Stufflebeam (2001).

40 Dabei könnte die oben zitierte Passage von Max Weber von Vertretern aller vier hier unterschiedenen Typen auf Akzeptanz treffen; der historisch gebundene „Werturteilsstreit“ ist heute weitgehend obsolet, wirkt jedoch in der Methodologie wertdistanziert geprägter Evaluationen und im immer wieder aufflammenden „Paradigmenstreit“ in der Evaluation weiter fort.

aus. Die theoretische Rahmung der Evaluation und die Umsetzung in empirische Untersuchungen verlaufen nach strikten Regeln, deren Einhaltung einer als wertfrei unterstellten intersubjektiven Überprüfung zugänglich gemacht wird. Wertfragen werden als z.B. durch die Verfahren der parlamentarischen Demokratie oder des demokratisch kontrollierten Verwaltungshandelns als vorentschieden oder außerhalb von/nachgängig zu Evaluationen zu entscheiden gesetzt. Letzteres gilt z.B. in besonderem Maße für weltanschaulich gebundene Trägerorganisationen, die Evaluationen in Auftrag geben.

Evaluationsspezialisten sollen möglichst objektive Daten, Ergebnisse und Schlussfolgerungen zu sozialen Tatsachen bereitstellen, die in jedem beliebigen Werterahmen identisch zustande kämen. Die sozialpolitische Bewertung von Ergebnissen wird dem vor- oder nachgelagerten Entscheidungsprozess durch Programmfinanziers und –entscheider/-innen, Programmleiter/-innen oder Programmkunden/-innen überlassen. Die Nutzung oder Verwendung der Evaluationsergebnisse wird von den Evaluatoren/-innen selbst nicht beeinflusst. Die Kontakte zu Beteiligten werden auf die Phase vor der Datenerhebung und die nach der Erstellung des Berichtes konzentriert, verbunden mit der Absicht, den Einfluss der Beteiligten auf die theoretische und empirische Arbeit zu minimieren. Allerdings gibt es auch wertneutrale Evaluationsmodelle, in denen – aus rein pragmatischen Erwägungen – eine engere Zusammenarbeit, sei es mit den Entscheidern/-innen oder der Programmleitung, abgestrebt wird. Diese Kooperation wird als „technisch“ und nicht interessengeleitet angesehen.

- **Werterelativistische** Modelle messen sozialen Werten die zentrale Bedeutung bei der Planung, Durchführung und Nutzung von Evaluationen zu. Sie arbeiten Wertgemeinsamkeiten und besonders Wertekonflikte in allen Phasen der Evaluation heraus und halten die bestehenden Spannungen aufrecht, ohne Partei zu nehmen. Dabei sollen die Werte der vollen Breite der Beteiligten und Betroffenen einbezogen werden, was eine gewisse Stärkung ansonsten einfluss- oder artikulationsschwacher Beteiligter impliziert. Es ist originäre Aufgabe von Evaluationsspezialisten, im

Dialog mit diesen Gruppen Werte zu identifizieren, zu klären und für die Planung des Evaluationsprozesses zu nutzen. Vornehmstes Ziel ist es zu ermöglichen, dass über Werte verhandelt wird, dass Fragestellungen auf die Verhandlungsergebnisse abgestimmt formuliert werden, dass Daten auf dem Hintergrund unterschiedlicher Wertpositionen interpretiert und alternative Schlussfolgerungen wertbezogen formuliert werden. Wertkonflikte werden dabei grundsätzlich mehrdimensional verortet, also z.B. nicht nur zwischen „Arm“ und „Reich“, sondern auch „Jung“ und „Alt“, „Mann“ und „Frau“, „Inländer/-in“ und „Migrant/-in“. Dabei lenken die Evaluierenden die Aufmerksamkeit der Adressaten/-innen auf Wertspannungen, etwa indem Berichte nach wertgeladenen Spannungsthemen (*issues*) gegliedert werden. Vertreter/-innen möglichst unterschiedlicher Wertpositionen bekommen Berichte und Ergebnisse zur Verfügung gestellt. Die offene und idealerweise herrschaftsfreie Kommunikation über Werte und Interessen soll Motivation und soziale Energie für die aktive Gestaltung der Evaluation einschließlich der Nutzung ihrer Ergebnisse bei den Stakeholdern freisetzen.

- **Wertepriorisierende** Modelle gehen ebenfalls von starken Wertdifferenzen in Gesellschaft, Organisationen, Netzwerken oder Kooperationen aus, bemühen sich dabei, diese einem pragmatischen Prozess der methodisch vorbereiteten Prioritätensetzung zuzuführen: Auf welcher Basis eines möglichst breiten Wertekonsenses werden Evaluationen von der Breite der Beteiligten und Betroffenen aktiv mitgetragen oder zumindest akzeptiert? Welche Themen, die durch stark umstrittene Werte und Interessen gekennzeichnet sind, können in welcher Form Erfolg versprechend durch die Evaluation verfolgt werden? Die systematische Erhebung von Werten und deren Klärung zwischen den am Programm Beteiligten, insbesondere zu Beginn einer Evaluation, ist Voraussetzung dafür, dass relevante und nützliche Fragestellungen, Daten und Ergebnisse bereitgestellt werden. Die aktive Beteiligung relevanter Vertreter verschiedener Wert- und Interessenpositionen soll zu einem breiten Wertkonsens führen, auf dessen Basis die verschiedenen Phasen der Evaluation möglichst störungsfrei bewältigt werden können. Kompromissbildung ist ein Mittel, um nützliche Evaluationen zu ermöglichen. Dabei wird – um die Durchführbarkeit der

Evaluation zu gewährleisten – ggf. auch hingenommen, dass der Grad der Partizipation bzw. der Einflussnahme gemäß den realen Machtverhältnissen differiert, z.B. Auftraggeber/-innen, Programmfinanziers oder Programmleiter/-innen mehr Einfluss auf die Steuerung des Evaluationsprozesses haben. Die Vertreter/-innen dieser Modelle gehen davon aus, dass Partizipation der relevanten Beteiligten und Betroffenen die Bereitschaft zur Unterstützung der sowie die Teilnahme an der Evaluation fördert und daher grundsätzlich wünschenswert ist. Auch die Nutzung und Nützlichkeit der Evaluationsergebnisse mag vom – gleichwohl angestrebten – Ideal einer Gleichverteilung auf alle Stakeholder abweichen.

- **Wertepositionierte** Ansätze gehen explizit davon aus, dass Gesellschaften, Verbände und Organisationen durch starke strukturelle Machtungleichgewichte, soziale und ökonomische Ungleichheit geprägt sind. Jedem auf die Verteilung von materiellen Ressourcen und die Situation in Lebenslagen gerichteten Programm liegen somit vitale soziale und ökonomische Interessen zu Grunde. Wertpositionierte Evaluationen sollen ein Gegengewicht bilden gegen die bestehende Wertehegemonie im politischen, sozialen und kulturellen Raum (gefasst z.B. im Menschenbild des *homo oeconomicus*), indem die Einflussschwachen gestärkt und ihnen eine Stimme im politischen Prozess gegeben wird. Deshalb entscheiden sich Vertreter/-innen dieser Ansätze für eine bestimmte Wertposition (z.B. mehr Verteilungsgerechtigkeit, ggf. auch zu Ungunsten der Leistungsgerechtigkeit). Sie tun dies insbesondere advokatorisch für Benachteiligte, also z.B. von Armut bedrohte oder arme Menschen. Sie sind sensibel für existentielle Wertfragen (z.B. soziale Benachteiligung und Gerechtigkeit). Dabei treffen die Vertreter/-innen dieser Ansätze ihre Wertentscheidung auf Basis sozialer Leitbilder, die sie transparent machen und dabei nicht zur Disposition stellen. Vorrangige vorgesehene Nutzer/-innen ihrer Evaluationsergebnisse (und auch des Evaluationsprozesses selbst) sind die sozial Benachteiligten bzw. Organisationen und Verbände, die für deren Interessen eintreten. Oft wird deren Aktivierung durch partizipative Elemente bis hin zur selbstständigen Übernahme der Evaluation angestrebt.

Mit dieser Typologie werden vorrangig heuristische Ziele verfolgt: In der Weiterentwicklung und Diskussion von Evaluationsmodellen, die auf die deutsche Politik im Bereich von Armut und Reichtum angewandt werden, soll eine Bezugsfolie angeboten werden, welche die Auseinandersetzungen und Positionsklärungen in den deutschen und europäischen Sozialwissenschaften des 20. Jahrhunderts aufnimmt.

2.3 Vorstellung relevanter Evaluationsmodelle

Die im vorangegangenen Abschnitt skizzierte Typologie nutzend, sind im Folgenden insgesamt zwölf Evaluationsmodelle dargestellt.[41] Die Modelle werden so beschrieben, wie dies wichtige Entwickler/-innen des jeweiligen Modells in ihren grundlegenden Texten tun.

Abbildung 10: Überblick über die exemplarisch dargestellten Evaluationsmodelle

Werteberücksichtigung	Kapitel	Name des Modells	Synonyme
1. wertedistanziert	2.3.1.1	Programmziel-gesteuerte Evaluation	Objectives-Based Studies, Effektivitätsstudien
	2.3.1.2	Experimentaldesign-gesteuerte Evaluation	---
	2.3.1.3	Quasi-Experimentaldesign-gesteuerte Evaluation	Nicht-Experimentelle Wirkungskontrolle
	2.3.1.4	Programmkosten/-nutzen-gesteuerte Evaluation	Kosten-Nutzen-Analysen, Kosten-Effektivitäts-Analysen
	2.3.1.5	Kontext-Mechanismus-gesteuerte Evaluation	Realistic Evaluation, Realist Evaluation
	2.3.1.6	Programmtheorie-gesteuerte Evaluation	Programm Theorie (*program theory*), logisches Modell (*logic model*), Theorie der Veränderung (*theory of change*), theoriebasierte Evaluation, oft verbunden mit Outcome-Measurement.
2. werterelativistisch	2.3.2.1	Spannungsthemen-gesteuerte Evaluation	Responsive (Fallstudien-) Evaluation
	2.3.2.2	Dialoggesteuerte Evaluation	Konstruktivistische Evaluation: Fourth Generation Evaluation
3. werteprioriserend	2.3.3.1	Entscheidungsgesteuerte Evaluation	Decision/Accountability-Oriented Studies (Entscheidungsgesteuerte/Rechenschaftslegungsorientierte Evaluation)
	2.3.3.2	Nutzungsgesteuerte Evaluation	Utilization focused Evaluation
	2.3.3.3	Stakeholder-Interessen-gesteuerte Evaluation	Deliberative Democratic Evaluation
4. wertepositioniert	2.3.4.1	Selbstorganisations-gesteuerte Evaluation	Empowerment Evaluation Inklusive Evaluation

41 Die Zuordnung zur Dimension der „Werteberücksichtigung“ ist in der ausführlicheren Darstellung der Modelle dargestellt, die im Anhang dieses Berichtes abgedruckt ist.

Die Präsentation dieser Breite unterschiedlicher Evaluationsmodelle erfolgt erstmalig für den deutschsprachigen Raum.[42] Auch besteht nicht der Anspruch auf Vollständigkeit.[43] Wegen des beschränkten Platzes sind Vereinfachungen nicht zu vermeiden.

Die hier genutzten Benennungen der jeweiligen Modelle weichen vielfach von den Namen ab, welche die jeweiligen Entwickler/-innen ihnen gegeben haben. Folgende zwei Überlegungen waren dabei leitend:

- Die Benennungen folgen sämtlich der Dimension „Steuerungsfaktoren“[44]. Diese bezeichnen die für das Modell typischen Bezugspunkte, welche die Evaluation in der Planungsphase orientieren und in ihrem Fortgang immer wieder ausrichten. Sie können Bestandteile des Programms sein (Programmziele, Kosten-Nutzen), sind aber zumeist Charakteristika der Evaluation, und bezeichnen die den Evaluationsplan strukturierenden Methoden (Experimente oder Quasi-Experimente), Prozessmerkmale der Evaluation („Spannungsthemen“) oder Ergebnisarten („Nutzung“ der Evaluationsergebnisse, „Entscheidungen“ auf Basis der Evaluation).
- Die durch die Entwickler/-innen selbst verwendeten Benennungen haben vielfach einen normativen Unterton, werben mit einem Schlagwort für die Überlegenheit des eigenen Ansatzes. Dies klingt an in Bezeichnungen wie „realistische“, „demokratische“, oder *Empowerment -Evaluation*. Hier werden dem gegenüber möglichst neutrale[45] Benennungen gewählt.

---

42 Vorarbeiten siehe Beywl (1988).

43 Das Konzept der Mikrosimulation wird nicht dargestellt. Es handelt sich dabei um ein Design für ex ante-Evaluationen. Es kann zur Wirkungs-Abschätzung eines möglichen Einsatzes eines Programms bzw. von Programmalternativen genutzt werden. Es kann auch ex-post verwendet werden, um zu simulieren, was passiert wäre, wenn an Stelle des tatsächlichen realisierten Programms ein anderes realisiert worden wäre. Die Verwendbarkeit, sowohl unter Aspekten der Durchführbarkeit (Datenlage ...) wie unter solchen der Nützlichkeit, lässt sich nicht abschätzen. Eine Positionierung des Ansatzes im Rahmen von Modellen der Evaluation ist durch die Vertreter/-innen dieses Ansatzes noch zu leisten. Die Mikrosimulation ist in Verbindung mit Gesetzesfolgeabschätzungen einzusetzen (vgl. BMI, 2002). Für den theoretischen Hintergrund der Mikrosimulation vgl. Merz (1991) für einen konkreten Anwendungsfall vgl. BMGS (2003) und für Nutzungsmöglichkeiten vgl. Citro/Hanushek (1991).

44 Der entsprechende englischsprachige Begriff ist der des *advance organizers* (Vorrück-Organisierers). Die Bennennung des Modells setzt sich zusammen aus dem jeweiligen Steuerungsfaktor – z.B. „Dialog“ und dem Partizip „gesteuert“. Durch Großschreibung – im Beispiel: „Dialoggesteuert“ wird der Eigenname für das jeweilige Modell kenntlich gemacht.

45 Die Autoren/-innen der Perspektivstudie sind sich bewusst, dass diese Neutralität nicht wirklich erreicht werden kann, da je nach disziplinärer Herkunft des Lesers/der Leserin und

Unter einem Modell sind in der Regel mehrere Ansätze verschiedener Autoren und Autorinnen zusammengefasst,[46] die sich bei genauerer Analyse und insbesondere in ihrem Selbstverständnis wiederum unterscheiden. Die Perspektivstudie verfolgt die Absicht, die große Differenziertheit der zeitgenössischen Evaluationstheorie deutlich zu machen und die darin enthaltene Komplexität so zu reduzieren, dass auch für Interessierte, die keine oder kaum Kenntnisse in Evaluationstheorie haben, eine überschaubare Typologie entsteht.

Es handelt sich bei dieser Vorstellung um idealtypische Modelle. In der Evaluationspraxis liegen sie in Reinform selten vor. Sie stellen einen Orientierungsrahmen auch für Studien dar, die sich gar nicht explizit auf eines der Modelle beziehen oder sich nicht einmal als „Evaluation" bezeichnen. Einige Modelle nehmen für sich in Anspruch, in andere integrierbar zu sein oder umgekehrt andere Modelle integrieren zu können. Eine derartige Hierarchisierung von Modellen wird explizit nicht angestrebt.

Die Darstellung der verschiedenen Modelle versucht, Unterschiede zu pointieren und spezifische Stärken und Schwächen herauszuarbeiten, welche insbesondere bei einer reinen Anwendung des Modells entstehen können. Vertreter/-innen von einzelnen Modellen werden oft auf dem Standpunkt stehen, sie könnten Schwächen vermeiden und ihre Modelle schlössen die Stärken der anderen Modelle ein, da sie bestimmte Vorgehensweisen dieser anderen Modelle bei Bedarf integrieren könnten. Dies steht außer Frage. In der Evaluationspraxis werden jedoch auf Grund knapper materieller und zeitlicher oder auch Kompetenzressourcen immer wieder Kompromisse geschlossen und Prioritäten gesetzt, so dass sich in der Praxis die relativen Stärken und Schwächen des jeweils gewählten/bevorzugten Modells durchsetzen.

Die Beschreibung erfolgt jeweils nach der in umseitig abgedruckten Abbildung 11 abgedruckten Gliederung:[47]

---

ihrer beruflichen Biographie mit Begriffen wie „Nutzung" oder „Spannungsthemen" ganz unterschiedlich gerichtete (positiv, neutral, negativ) Konnotationen verbunden sein können.

46 In der Zeile „Synonyme" und über die Titel der angegeben Publikationen können die Ursprungsbenennungen erschlossen werden.

47 Im Anhang sind für interessierte Leser/-innen ausführlichere Modelldarstellungen abgedruckt; hier finden sich auch Ausführungen zu den beiden Dimensionen „Werteberücksichtigung" und „Wichtige Quellen", die in den nachfolgenden Zusammenfassungen nicht dargestellt sind.

Abbildung 11: Gliederungsschema für die Modell-Steckbriefe

Synonyme	Hier werden verschiedene Synonyme in englischer oder auch deutscher Sprache und auch unterschiedliche Übersetzungen angegeben.
Charakterisierung Modell	Was kennzeichnet dieses Modell, welches Verständnis haben seine Vertreter/-innen von Evaluation und wie würde eine solche Evaluation (in groben Zügen) geplant werden?
Steuerungsfaktoren	(engl.: *advance organizers*) Welches sind die zentralen Bezugspunkte, welche die Evaluatoren/-innen heranziehen, um die Evaluation zu planen und während ihrer Durchführung zu steuern?
Hauptzwecke	Welches sind die intendierten Verwendungen der Evaluation bzw. ihrer Ergebnisse?
Quellen der Fragestellungen	Wer legt die Fragestellungen, welche die Evaluation beantworten soll fest, bzw. aus welchen Quellen werden sie entnommen?
Typischerweise eingesetzte Methoden	Welche Methoden werden zur Datenerhebung innerhalb einer solchen Evaluation eingesetzt? Sind sie festgelegt oder variabel? Gibt es bevorzugte Methoden?
Stärken	Welche Charakteristika des Modells sind als besonders Erfolg versprechend/anwendungsfreundlich/gut durchdacht hervorzuheben bzw. zu welchen positiven Effekten kann die Durchführung einer solchen Evaluation führen?
Schwächen	Welche Probleme können bei der Umsetzung einer solchen Evaluation auftreten, bzw. an welchen Stellen bestehen tendenziell blinde Flecke? Sind bestimmte negative Effekte von Evaluationen nach diesem Modell zu erwarten?
Werteberücksichtigung	Wie verhält sich das Modell zu Werten der Beteiligten und Betroffenen und welcher Umgang damit ist vorgesehen?
Wichtige Quellen	Welches sind grundlegende Werke oder Lehrbücher für das Modell oder haben es richtungsweisend weiterentwickelt bzw. wo wird es besonders gut dargestellt?

Während die Modelle in den theoretischen Schriften in „reiner“ Form vorkommen, werden sie in der Evaluationspraxis vielfach kombiniert und gemischt. So ist es z.B. häufig anzutreffen, dass quantitative, quasi-experimentelle Studien mit qualitativen Fallstudien kombiniert werden. Auf derartige empirisch vorfindbare Kombinationen wird sowohl in der ausführlichen Modelldarstellung wie im Abschnitt 2.4 eingegangen.

Es ist in der evaluationstheoretischen Literatur nicht unumstritten, welche Modelle welcher Autoren/-innen (wobei diese über mehrere Auflagen im Verlauf von 15 oder gar 25 Jahren auch grundlegende Veränderungen vornehmen) ähnliche Modelle darstellen und insofern zusammenzufassen sind. Auf derartige Fein-Differenzierungen wird in der Darstellung verzichtet.

2.3.1 Wertedistanzierte Modelle

Als Ansätze, die Wertfragen gezielt aus dem Evaluationsprozess ausklammern, werden insgesamt sechs Modelle vorgestellt:

2.3.1.1 Programmziel-gesteuerte Evaluation

Diese auch *Objectives-Based Studies* oder „Effektivitätsstudien“ genannten Modelle stammen ursprünglich aus dem Bereich Bildung/Erziehung. Es soll überprüft werden, ob die Ziele eines Programms erreicht werden. Der Grad der Zielerreichung ist ausschlaggebendes Kriterium für seine Bewertung. Als Vergleichsmaßstäbe werden auch Resultate ähnlicher Programme oder Aussagen der Grundlagenforschung heran gezogen.

Intendierte Ziele des Programms müssen operationalisiert werden, damit sie mit geeigneten Methoden gemessen werden können. Programmziel-gesteuerte Evaluationen sind häufig intern, in den Organisationen der Programmverantwortlichen, angelegt.

Steuerungsfaktoren sind die operationalen Lern-, Einstellungs- oder Verhaltensziele, die durch das Programm bei den Zielgruppen ausgelöst werden sollen.

Hauptzweck ist es, gesicherte Informationen bereit zu stellen, so dass über die Ausweitung/Fortführung von Programmen entschieden oder Maßnahmen zur optimierten Zielereichung ergriffen werden können.

Fragestellung ist typischerweise, in welchem Maße die identifizierten Ziele eines Programms erreicht werden.

Es können insbesondere die verschiedensten Methoden der oft standardisierten (Lern-)Ziel- oder Leistungsmessung angewendet werden.

Der Ansatz ist gut geeignet, wenn es darum geht, klar zugeschnittene Projekte mit expliziten operationalisierten Zielen zu beurteilen. Die Logik dieser Evaluationsstudien – Gegenüberstellung von Zielen und Resultaten – ist leicht zugänglich. Häufig wird „Evaluation“ mit Programmziel-gesteuerter Evaluation gleichgesetzt.

Probleme bestehen, wenn zu Beginn der Evaluation keine expliziten Programmziele vorliegen. Sie müssen innerhalb einer Zielklärung nachträglich formuliert werden. Es ist unsicher, ob das betrachtete Programm tatsächlich durch diese Ziele gesteuert ist. Es besteht auch die Gefahr, dass schlechte Zielerreichung gemessen wird, obwohl

das Programm Gutes leistet. Die Ansprüche, die an die Validität der Erhebungsinstrumente angelegt werden, sind sehr hoch. Erhebliche Ressourcen werden in die Entwicklung und Prüfung der Instrumente investiert.

2.3.1.2 Experimentaldesign-gesteuerte Evaluation

Dieses Vorgehen kontrolliert, ob/in welchem Ausmaß ein Programm (und nicht andere Faktoren) die operational bestimmten Zielgrößen bei der Zielgruppe ursächlich ausgelöst hat. Zu diesem Zweck werden die Zielgrößen für die am Programm teilnehmende Gruppe (Experimentalgruppe) gemessen und denen einer Kontrollgruppe, die nicht teilgenommen hat, gegenüber gestellt. Die Zuweisung zu diesen beiden Gruppen geschieht streng nach dem Zufallsprinzip.

Steuerungsfaktoren sind, wie in der zielgesteuerten Evaluation, angestrebte operationale Ziele. Darüber hinaus auch die Vollständigkeit, Zugänglichkeit und Aktualität der Datensätze für die Ziehung von Zufallsstichproben.

Es geht darum, diejenigen Arten von Programmen/Programmvarianten zu bestimmen, die tatsächlich zur Erreichung der gesetzten Ziele beitragen. Dies ist Grundlage für Entscheidungen über die Ausbreitung/Regeleinführung von Modellprogrammen.

Zentrale Fragestellung ist, ob/in welchem Umfang das Programm in Bezug auf die explizierten Ziele „einen Unterschied macht“.

Benötigt werden vorrangig quantitative, möglichst metrisch skalierte Daten, welche fortgeschrittene statistische Kontroll- und Auswertungsprozeduren erlauben.

Experimentelle Designs bieten sich für Programme an, die in hohem Umfang standardisiert, unter stabilen Rahmenbedingungen und mit hohen Teilnehmerzahlen durchgeführt werden (können). Die Strategie der Gruppenbildung nach Zufall weist -in Analogie zu naturwissenschaftlichen Experimenten – hohe Plausibilität in Bezug auf eine unabhängige Messung der Wirkung auf. Die experimentelle Wirkungskontrolle ist für Modellversuche oder sehr deutliche Veränderungen in bereits existierenden Programmen geeignet.

Da Ziele zentrale Steuerungsfaktoren sind, gelten dieselben häufigen Umsetzungsschwierigkeiten wie für Programmziel-gesteuerte Evaluationen. Vorausgesetzt wird eine weitgehende Planbarkeit/sichere Umsetzbarkeit von Programmprozessen.

Personale Faktoren des Programmpersonals oder Eigenheiten der tragenden Organisation spielen keine Rolle bzw. sollen keine spielen (Ausschluss von Srövariablen). Die Übertragbarkeit der Ergebnisse auf andere sozioökonomische (lokale) Kontexte und Programmausgestaltungen ist problematisch (Problem „externer Validität“). Es gelingt oft, eine Grundlage für Grundsatzentscheidungen zu schaffen; dabei ergeben sich kaum Anhaltspunkte, *wie* Programme zu verbessern sind. Ethische Probleme: Es wird oft abgelehnt, Kontrollgruppenmitglieder von Vorteilen oder speziellen Fördermaßnahmen fern zu halten. Andererseits gibt es Widerstand gegen eine „zwangsweise“ Zuweisung von Programmgruppenmitgliedern. Bei Freiwilligkeit der Erstmeldung kommt es zu erheblichen Selektionswirkungen und Verzerrungen.

2.3.1.3 Quasi-Experimentaldesign-gesteuerte Evaluation

Dieser Ansatz, auch „nicht-experimentelle Wirkungskontrolle“ genannt, benötigt keine randomisierte Kontrollgruppe. Genauso wie das experimentelle Design fokussiert er auf den Unterschied zwischen einer Situation *mit* und einer *ohne* Intervention. Es wird eine Vergleichssituation konstruiert, die beschreibt, was sich ohne die betrachtete Intervention entwickelt hätte (kontrafaktische Situation). Es werden bspw. für die Programm-Teilnehmer/-innen „statistische Zwillinge“ zu einer Referenzgruppe zusammengefasst (*Matching*). Idealerweise unterscheidet die Vergleichssituation sich von der tatsächlichen Situation nur in dem Umstand der fehlenden Intervention. Zu diesem Zweck müssen viele Einflussgrößen für beide Situationen gleich gehalten werden, wozu abgesicherte theoretische Kenntnisse zu Programm und Feld herangezogen werden. Wenn sich Experimentalgruppe und Vergleichsgruppe dennoch unterscheiden, kann nachträglich eine statistische Gewichtung vorgenommen werden. Bei einer quasi-experimentellen Wirkungskontrolle ohne Vergleichsgruppe können verschiedenste Pre- und Posttest-Designs durchgeführt werden. Verfahren mit Vergleichsgruppe kommen dem „Ideal“ des Experiments näher.

Steuerungsfaktoren sind die der Programmziel-gesteuerten Evaluation. Hinzu kommen methodische Faktoren (bspw. vollständige Datensätze über Programmteilnehmer /-innen und Nicht-Programmteilnehmer/-innen).

In Hauptzwecken und Fragestellungen besteht große Ähnlichkeit mit der Experimentaldesign-gesteuerten Evaluation.

Es können verschiedenste statistische und ökonometrische Verfahren eingesetzt werden. So ist das Matching ein prominentes und dabei im Datenmanagement ausspruchsvolles Verfahren. Es wird von manchen Autor/-innen auch mit dem quasi-experimentellen Design gleichgesetzt.

Bei in den letzten Jahren verbesserter Datenlage können mit Hilfe fortgeschrittener statistischer Methoden abgesicherte Ergebnisse produziert werden. Das Modell ist vor allem bei längerem, großflächigem, standardisiertem (unverändertem) Programmeinsatz sinnvoll. Ethische Probleme stellen sich weniger als in den Experimentaldesign-gesteuerten Evaluationen.

Matching-Verfahren erfordern große, möglichst vollständige bzw. repräsentative Datensätze, die ausreichende Informationen über Programm-Teilnehmer/-innen sowie Nicht-Teilnehmer/-innen enthalten. In gewichtenden Berechnungen müssen die Maßnahme-Heterogenität und die regionale Heterogenität ausreichend berücksichtigt werden, was die Transparenz und Nachvollziehbarkeit von Ergebnissen und Schlussfolgerungen für Außenstehende Nicht-Statistiker/-innen einschränkt (Gefahr der Nicht-Nutzung der Evaluationsergebnisse).

2.3.1.4 Programmkosten/-nutzen-gesteuerte Evaluation

Kosten-Nutzen-Analysen oder Kosten-Effektivitäts-Analysen stellen idealerweise alle (monetären) Kosten und Nutzen eines Programms gegenüber bzw. vergleichen Kosten und Nutzen verschiedener Programmalternativen. Es können ökonomische Messgrößen im Vordergrund stehen, aber auch psycho-soziale Effekte betrachtet werden. Es lassen sich verschiedene Analyseperspektiven unterscheiden: individuelle und soziale (institutionelle oder volkswirtschaftliche) Bilanzierungen. Wenn sich die (Aus-)Wirkungen nicht unmittelbar in monetären Werten messen lassen und es hierfür auch keine klaren Marktsubstitute gibt – was vielfach für sozialpolitische Programme zutrifft – ist eine ausführliche Beschreibung der Effekte erforderlich (Kosten-Effektivitäts- oder Kosten-Wirksamkeits-Analysen). Ein Programm wird dann als effizient bezeichnet, wenn der Nutzen die Kosten übersteigt bzw. wenn es gegenüber einem anderen Programm eine bessere Kosten-Nutzen-Relation aufweist.

Zentrale Steuerungsfaktoren sind die Programmziele; anfallenden Input-Kosten eines Programms wird der Nutzen im Sinne der Zielerreichung gegenüber gestellt.

Zeit- und Alternativenvergleich sowie die Bestimmung der Produktivität eines Programms werden durchgeführt, um kostenwirksame Programme auszuwählen bzw. Programme mit schlechter Kosten-Nutzen-Relation einzustellen.

Fragestellungen beziehen sich auf den Vergleich von Kosten und Nutzen, differenziert nach bestimmten Kosten-/ Nutzenarten, ggf. unterschieden nach Kontext-Bedingungen verschiedener Programmstandorte oder nach Programmvarianten.

Für die Kostenseite kann auf bestehende Kostenrechnungsinstrumente (z.B. aus der Neuen Steuerung, aus Controllingsystemen …) zurückgegriffen werden. Bei Kosten-Nutzen-Analysen i.e.S. müssen die Nutzen in Geldeinheiten ausgedrückt werden, was aufwändige Bestimmungsverfahren erfordert.

Kosten- und Nutzenkomponenten werden transparent gemacht und möglichst beziffert. Im Gegensatz zu vielen anderen Evaluationsmodellen werden Kosten explizit als zentrale Analyseeinheit und Bewertungsdimension genutzt (im Sinne einer Aussage über die optimale Nutzung von Ressourcen, insbesondere öffentlichen Finanzmitteln).

Die Grenzen der Programmziel- sowie der (Quasi-) Experimentaldesign-gesteuerten Modelle gelten auch hier (z.B. eindeutige Zuordnung der Effekte zum Programm). Die Ermittlung von Kosten-Nutzen-Relationen beruht auf teilweise bezweifelbaren Annahmen oder Operationalisierungen.

2.3.1.5 Kontext-Mechanismus-gesteuerte Evaluation

Dieses relativ junge Modell – auch *Realistic* oder *Realist Evaluation* –ist bislang kaum umgesetzt. Es weckt großes Interesse wegen seines Anspruches, Ursachen gültig und weniger aufwändig als (quasi-) experimentelle Designstudien zu bestimmen. Programme werden nicht als „Dinge“ angesehen, die funktionieren oder nicht, sondern vielmehr als Bündel von Annahmen über Mechanismen, die für bestimmte Subjekte in bestimmten Kontexten (*context*) „arbeiten“ und bei Einsatz entsprechender Ressourcen in der Lage sind, Outcomes auszulösen (Context-Mechanism-Outcome - Konfigurationen): „What works for whom in what circumstances?“ Wirkmächtige Programme haben dann ein kausales Potenzial, welches das (gewünschte) Handeln der Zielgruppen auslösen *kann*: Es ist schließlich das Handeln der Beteiligten, welches die Programme zum Funktionieren bringt. Ein

Programm stellt somit Ideen und Ressourcen breit, welche die Programmteilnehmenden befähigen, sich/etwas zu verändern. Ihr Entscheidungshandeln ist ausschlaggebend für den Programmerfolg. Im Unterschied zur Annahme des (quasi-) experimentellen Modells löst das Programm (dort: *treatment*) die Wirkung nicht unmittelbar aus, sondern aktiviert einen Mechanismus, der regelhaft Wahlentscheidungen und Ressourcen von Individuen oder sozialen Aggregaten (wie z.B. Nachbarschaften) miteinander verbindet, so dass Wirkungen als Lösungen sozialer Probleme zustande kommen. Entscheidend für Auslösen/Nicht-Auslösen des Mechanismus“ ist auch der soziale Kontext, in dem das Programm stattfindet, z.B. in unterschiedlichen sozial benachteiligten Stadtteilen oder in strukturschwachen ländlichen Regionen usf.

Steuerungsfaktoren sind hier theoretische, empirisch schrittweise abgesicherte Annahmen über spezifische soziale Mechanismen, die in bestimmten sozialen Kontexten Wirkungen auslösen.

Hauptzweck ist die Entwicklung differenzierter, jeweils auf enge Politikfelder spezifisch zugeschnittener Theorien über CMO-Konfigurationen, die von Politikern/-innen, Verwaltung und Programmverantwortlichen mit Aussicht auf Erfolg ausgewählt und eingesetzt werden können. Es geht darum, heraus zu finden, wie die eingeführten Veränderungen die begrenzten Wahlmöglichkeiten (*constrained choices*) der Teilnehmenden kognitiv ausweiten und letztlich verändern.

Fragestellungen resultieren aus den Theorien (gefasst in CMO-Konfigurationen) der feldkundigen und spezialisierten Evaluatoren/-innen.

Typischerweise eingesetzte Methoden sind, neben vielen anderen, strukturierte Interviews, um die CMO-Konfigurationen aufzudecken.

Stärke dieses Ansatzes ist, dass die Komplexität sozialen Handelns auf hohem Niveau erhalten und in feldspezifischen Theorien abgebildet wird, was kumulatives Wissen darüber verspricht, wie Programme wirksam angelegt werden können. Die zentrale Erkenntnis der personenbezogenen Dienstleistungstheorie, dass Outcomes sozialer Programme wesentlich durch Koproduktion von Fachkräften und Zielgruppenmitgliedern zustande kommen, ist in das Modell integriert.

Einzuschränken ist, dass der Ansatz bislang vorwiegend für Programme/Maßnahmen erprobt wurde, die sehr sensible Werte betreffen, die auf einem außerge-

wöhnlich breiten gesellschaftlichen Konsens beruhen.[48] Typisch für soziale Programme sind hingegen ausgeprägte Wertkonflikte zwischen den verschiedenen Stakeholdern. Das Modell liefert kaum praktische Anweisungen, wie Evaluationen geplant und durchgeführt werden sollen.

2.3.1.6 Programmtheorie-gesteuerte Evaluation

Dieses Modell ist in den letzten zehn Jahren vermehrt eingesetzt worden und ist unter den Benennungen Programm Theorie (*program theory*), logisches Modell (*logic model*) oder Theorie der Veränderung (*theory of change*) bekannt.

Im besten Fall besteht eine geprüfte Theorie, welche die Praxis des zu evaluierenden Programms und auch die Evaluation anleitet. Diese „Programmtheorie“ benennt Bedingungen, Umsetzungselemente und angestrebte Resultate des Programms und ordnet diese in einem logischen Ablaufschema an.

Wenn eine entwickelte Theorie fehlt, ist es Aufgabe der Evaluation, diese zusammen mit den Beteiligten zu erstellen. Tatsächlich ist diese Theorieentwicklung häufig zu leisten. Dies erfordert von Seiten der Evaluatoren/-innen ausgeprägte Kenntnisse des Politikfeldes und des aktuellen Standes der wissenschaftlichen Forschung. Die Evaluatoren/-innen müssen die Dimensionen (Kontext, Ausgangslage, Interventionen, Resultate) sowie Richtung und Intensität der Beziehungen zwischen diesen identifizieren und darlegen. Zentral sind die Outcomes, also die bei den Zielgruppen ausgelösten Veränderungen bzw. Stabilisierungen in Wissen, Einstellung, Verhalten sowie die Veränderungen/Stabilisierungen in ihrer sozialen Umwelt Die Explikation der Programmtheorie sollte vor Programmstart erfolgen. So können Konstruktionsfehler des Programms vermieden werden. Die explizierte Programmtheorie stellt sich bildlich als Diagramm mit Pfeilen (als Symbole für kausale Verbindungen) zwischen den einzelnen Dimensionen dar.

Steuerungsfaktoren sind in der Programmtheorie-gesteuerten Evaluation die Mechanismen, die von Interventionen zu Outcomes führen sowie Kontextbedingungen.

---

48 z.B. Rückfall-Prävention von Strafgefangenen oder Vermeidung von Einbrüchen in Wohnungen – beide tragen dazu bei Gefährdungen abzuwenden, die Sicherheit und Unversehrtheit der (potentiellen) Opfer so stark verletzen, dass sie auf hohe Akzeptanz stoßen.

Hauptzweck ist es zu verstehen, wie das Programm funktioniert, zu erklären, warum es Erfolg hat oder/und Vorschläge für die Weiterentwicklung des Programms zu machen.

Die Fragestellungen werden von der programmleitenden Theorie abgeleitet; insofern diese von (bzw. gemeinsam mit) den Beteiligten entwickelt ist, nimmt sie deren Perspektiven auf.

Das Modell verfährt multimethodisch, arbeitet während der Theoriekonstruktion eher mit qualitativen, in der Programmüberprüfung mit quantitativen Daten.

Der unmittelbare Nutzen des Modells besteht darin, dass im Falle von Implementationsfehlern Strategien zur Überwindung entwickelt werden können. Wenn das Programm Schwächen/Misserfolge aufweist, kann zwischen einem „Theorie-Fehler“ und einem „Umsetzungs-Fehler“ unterschieden werden. Bei jungen Programmen ist es möglich abzuschätzen, ob sich das Programm entsprechend den Vorhersagen in der Praxis bewähren wird.

Eine fehlende Theorie kann große Arbeitsbelastungen für die Evaluatoren/-innen hervorrufen. Ihre eigentliche Aufgabe, eine Bewertung der Programmgüte und –Verwendbarkeit vorzunehmen, kann darunter leiden. Auch können sich die Evaluatoren/-innen für die Konzeption des Programms verantwortlich fühlen (Rollenkonflikt). Mit vielfacher Anwendung Programmtheorie-gesteuerter Evaluationen wächst das methodologische Wissen darüber an, wie solche Theorien konstruiert und für Programm und Evaluation genutzt werden können, doch kommt es nicht zu kumulativem i.S.v. gesetzmäßigem Wissen; für jeden Evaluationsfall ist eine spezifische Programmtheorie zu konstruieren, was den Aufwand für dieses Modell erhöht.

2.3.2 Werterelativistische Modelle

Die folgenden beiden Modelle haben einen gemeinsamen Ursprung, die mit dem Werk von Robert Stake Mitte der 70erJahre eingeleitete Erweiterung des Evaluationsparadigmas. Während sein eigenes, inzwischen ausdifferenziertes „responsives“ Modell die Person des Evaluators/ der Evaluatorin als Instrument einsetzt, um wertgeladene Spannungsthemen transparent zu machen, verbindet der Ansatz der Dialoggesteuerten Evaluation die grundsätzlich werterelativistische Position mit einem ausgeprägt demokratisch-partizipativen Vorgehen. Zwischen beiden Modellen besteht eine Spannung zwischen „Individualismus“ und „Kollektivismus“. Von ihrer

philosophisch-wissenschaftstheoretischen Grundlage her messen beide Modelle Werten eine entscheidende Rolle in der Evaluation zu. Implizit (Stake) oder explizit (Guba/Lincoln) gehen sie davon aus, dass ein Herausheben und Gleichbehandeln der Werte verschiedener Beteiligtengruppen die Position Marginalisierter in Programm und Evaluation stärkt.

2.3.2.1 Spannungsthemen-gesteuerte Evaluation

Dieser Ansatz hat Mitte der 70er Jahre die Erweiterung des evaluationstheoretischen Paradigmas wesentlich beeinflusst und ist auch unter der Bezeichnung „responsive Evaluation“ bekannt geworden. Eine solche Evaluation antwortet (am. *respond*) auf Informationsanliegen der verschiedenen Beteiligten und Betroffenen. Ein Programm wird dicht und vertieft beschrieben, zusammen mit seiner lokalen Umsetzung („Fall“). Der Kontext und der Prozess der Umsetzung des Programms werden intensiv beschrieben, auch dann, wenn der Fokus der Evaluation auf den Resultaten resp. Wirkungen liegt. Mit der hieraus hervorgehenden Beschreibung (in einer ganz spezifischen Berichtsform repräsentiert) erhalten die Beteiligten relevante Informationen, die sie gemäß ihrer jeweiligen Werte und Interessen beurteilen.

Steuerungsfaktoren sind wertgeladene „Spannungsthemen“ (*issues*). Dies sind von den Beteiligten als problematisch, konfliktreich oder ungelöst wahrgenommene Programmbestandteile.

Es geht primär darum, ein Programm intensiv zu schildern und seine Wirkungsweise zu erhellen, vorrangig um es zu verbessern. Daten, Informationen und Interpretationen können durch die Programmbeteiligten für eine intensive Auseinandersetzung mit dem Programm genutzt werden.

Die Konkretisierung der Fragestellungen erfolgt fortlaufend (*emergent*) danach, was für die vorgesehenen Adressaten der Evaluation am interessantesten ist. Damit „antwortet“ der responsive Ansatz beständig auf deren evaluative Bedarfe.

Multiperspektivität bedeutet, mehrere möglichst verschiedenartige Methoden und Datenquellen parallel einzusetzen (mit einer Präferenz für qualitative Daten), etwa Beobachtung, dichte Beschreibung, narrative Verfahren, Interviews sowie Auswertung von Dokumenten. Um „natürliche“ Alltagssituationen beobachtbar und für Außenstehende nachvollziehbar zu machen, wurde eine spezifische Fallstudien-Methode entwickelt (Stake 1995).

Die responsive Evaluation eignet sich besonders zur Entwicklung und Unterstützung neuer Programme, da sie ihren Betrachtungsfokus flexibel und zeitnah an die oft schnellen Veränderungen in der Programmdurchführung anpassen kann. Eine große Chance von responsiven Fallstudien liegt in ihrer konzeptionellen Nutzungsmöglichkeit als erfahrungsbasierte Grundlage für die Meinungsbildung über Themen von Armut und Reichtum und damit auch zur diskursiven Festlegung von Bewertungskriterien für Maßnahmen und Programme.

Die methodische Offenheit – die sich z.B. auch im Fehlen von Verfahrensanweisungen, operational beschreibendem Vorgehen und systematischen Forschungen über diesen Fallstudienansatz manifestiert – kann leicht zu einer nicht angemessenen Anwendung insbesondere durch unerfahrene Evaluatoren/-innen führen. Breit angelegte Programme auf Landes- oder Bundesebene lassen sich nur bedingt responsiv evaluieren, da zu viele Beteiligte und Betroffene in die Abstimmungsprozesse einzubeziehen wären und die Auswahl der wenigen bearbeitbaren Fälle nur schwer schlüssig und für wichtige Beteiligte akzeptabel begründet werden kann. [49]

Eine große Chance von responsiven Fallstudien liegt in ihren konzeptionellen Nutzungsmöglichkeiten als erfahrungsbasierte Grundlage für die Meinungsbildung über Themen von Armut und Reichtum.

2.3.2.2 Dialoggesteuerte Evaluation

Dieses Modell nimmt explizit Bezug auf die konstruktivistische Wissenschaftstheorie. Es bezeichnet sich selbst auch – eine Paradigmenfolge von den Programmziel-gesteuerten über die (Quasi-) experimentellen bis zu den pragmatischen Modellen der Entscheidungs- und Nutzungsgesteuerten Evaluation unterstellend – als *Fourth Generation Evaluation*. Es basiert auf der Grundannahme von Erkenntnis als individueller Deutung eines einzelnen Menschen auf der Grundlage seiner Weltsicht und überträgt diese auf die Evaluation. Eine intersubjektive Einigung auf bestimmte Deutungen – also Bewertungen – sei nach einem Prozess des dialogischen Austausches von Perspektiven möglich, dabei gebunden an räumlich-zeitlich spezifizierte Kontexte. Universell gültige Bewertungen wie „richtig“ oder „wahr“ seien hin-

---

49 Einen Ausweg hierzu hierfür bietet die Cluster-Evaluation; vgl. Anm. 16 in Kap. 1.

gegen ausgeschlossen, d.h. eine raumzeitlich überdauernde Bewertung eines sozialen Programms als „gut“, „überlegen“ oder „beste Praxis“ sei grundsätzlich nicht möglich.

In der Dialoggesteuerten Evaluation fungieren Beteiligten und Betroffenen wie auch die Evaluatoren/-innen als ‚Erkenntnisinstrumente’, die zu einem geteilten, informierten Verständnis durch Auseinandersetzung untereinander und mit Informationen beitragen. Die Stakeholder sollen die Evaluation gemäß ihrer Interessenlagen mit steuern, was ihnen die Evaluatoren/-innen ermöglichen sollen.

Der Dialog zunächst innerhalb der einzelnen Gruppen, dann zwischen den Gruppen wird fortlaufend rückgekoppelt. Es soll soviel Konsens wie möglich hergestellt werden. Durch dauerhaft ungelöste Gegensätze ergeben sich dabei ggf. immer wieder neue Ansatzpunkte für Evaluation.

Steuerungsfaktoren sind die im Dialog geklärten Annahmen, Anliegen und Spannungsthemen der Stakeholder (*claims, concerns and issues)*.

Hauptzweck ist es, die Vielfalt der Deutungen von Daten und Informationen über das Programm transparent zu machen und eine dialogische Auseinandersetzung der Beteiligten und Betroffenen zu initiieren.

Fragestellungen werden mit den Beteiligten und Betroffenen gemeinsam entwickelt und im Verlauf der Evaluation evtl. verändert oder spezifiziert.

Eine ausgewogene Mischung aus qualitativen und quantitativen Methoden und angemessen aufbereiteten Daten und Berichten soll die Kommunikation zwischen allen Beteiligten und Betroffenen unterstützen.

Stärke des Modells ist, dass der Evaluationsprozess und seine Ergebnisse völlig transparent sind, was die Akzeptanz und Nutzung der Ergebnisse begünstigt. Dadurch, dass Stakeholder als „Instrumente“ aktiv werden, müssen ggf. keine oder weniger eigenständige Instrumente entwickelt werden. Die Evaluation identifiziert keine Verantwortlichen für Erfolg oder Misserfolg eines Programms und ist so weniger bedrohlich.

Achillesferse einer Dialoggesteuerten Evaluation ist, dass Auftraggeber/-innen, Stakeholder und Evaluatoren/-innen darin zu Beginn und fortlaufend übereinstimmen müssen, dass ein solches dialogisches, bestehende Machtungleichgewichte nivellierendes Vorgehen angemessen ist und dass sie in diesem Prozess zusammen-

arbeiten. Die Stakeholder müssen akzeptieren, dass Ergebnisse möglicherweise widersprüchlich, in jedem Fall wenig eindeutig sein werden. Ein gefundener Konsens ist nicht übertragbar auf andere Settings und Programme.

2.3.3 Werteprioritisierende Modelle

Auch für dieses Modell sind individuelle, organisatorische und soziale Werte/ Interessen von zentraler Bedeutung. Sie verfahren in Anerkenntnis eines politischen / ökonomischen Kontextes, in dem Programme durchgeführt werden und Evaluationen stattfinden, „pragmatisch“ mit diesen Werten, in dem sie Prozesse der Prioritätensetzung einleiten und dabei auch in Kauf nehmen, dass sich „einflussreichere“ Werte stärker durchsetzen als andere. Anspruch ist dabei, dass dieser Prozess der Priorisierung – etwa im Evaluationsbericht – transparent gemacht wird. Das – letztlich wert-/interessenbasierte – Rangordnen von Fragestellungen, Bewertungskriterien sowie der Einsatz von Untersuchungsressourcen dient auch dazu, Evaluationen zu fokussieren und auf neuralgische Punkte, seien dies Entscheidungen oder – weitergefasst – Nutzungen, schlank auszurichten.

2.3.3.1 Entscheidungsgesteuerte Evaluation

Dieser oftmals auch *Decision/Accountability-Oriented* (Entscheidungsgesteuert/Rechenschaftslegungs-orientiert) genannte Ansatz soll präzise so geplant und terminiert werden, dass für im Voraus bestimmte Entscheidungssituationen vor, während oder nach der Programmdurchführung rechtzeitig die erforderlichen Informationen bereit stehen. Er reagiert darauf, dass von Entscheidungssituationen entkoppelte Evaluationen häufig nicht genutzt werden. Typische Entscheidungsgegenstände sind: Sollen „Pilote“ in ein Bundes-Regelprogramm überführt werden? War es gerechtfertigt, die öffentlichen Mittel einzusetzen? Zur Evaluationssteuerung dient z.B. das CIPP-Schema (Stufflebeam 1972), das verschiedenen Klassen von Entscheidungssituationen im Programmverlauf (*countenances*) eine „Evaluationsart“ zuordnet. Das passende Design wird anschließen von den Evaluationsverantwortlichen entwickelt und umgesetzt. Dieser Evaluationsansatz verlangt, auf die Kräfte im Feld des Programms flexibel zur reagieren und passende Untersuchungspläne bereit zu stellen. Evaluatoren/-innen arbeiten mit den künftigen Entscheidern/-innen, den Adressaten/-innen der Evaluationsergebnisse zusammen, um heraus zu finden, worin deren Unsicherheiten und ihre abweichenden oder gegensätzlichen Einschätzungen der empirischen Realität bestehen. Sie sollen auch solche Ent-

scheidende einbeziehen, die relativ weit vom Programm entfernt sind, aber längerfristig erhebliche Einwirkungsmöglichkeiten haben.

Steuerungsfaktoren sind die Unsicherheiten, untereinander abweichende und konträre Realitätssichten der Entscheidenden (und weiterer Beteiligter), Tagesordnungspunkte vorauszusehender Entscheidungspunkte und (z.B. gesetzliche) Erfordernisse der Programmrechenschaftslegung.

Evaluationsergebnisse sollen v.a. dazu genutzt werden, Entscheidungen über die Verbesserung oder Grundsatzentscheidungen zu Programmen zu treffen.

Fragestellungen stammen insbesondere von denjenigen, die Entscheidungen treffen werden. Sie können von den Evaluierenden im Sinne von Optionen formuliert und eingebracht werden und sollen schließlich von den künftigen Entscheidern verabschiedet werden. Gefragt wird typischerweise nach Resultaten von Programmalternativen, nach Defiziten oder Stärken von Programmelementen, nach ihren Nettokosten oder -erlösen.

Es können alle empirischen Methoden eingesetzt werden – nur (quasi-)experimentelle Versuchspläne sind wegen ihrer Inflexibilität, auf sich verändernde Entscheidungssituationen zu reagieren, tendenziell ausgeschlossen.

Ein zentraler Vorteil des Entscheidungsgesteuerten Modells besteht darin, dass es wahrscheinlicher wird, dass Evaluationen aufmerksam verfolgt und ihre Ergebnisse tatsächlich genutzt werden. Durch die Fokussierung auf nachgewiesenermaßen angefragte Informationen und Ausrichtung der Zeitpläne auf die Zeitpunkte, zu dem sie verfügbar sein müssen, erbringt die Evaluation idealerweise ausschließlich nutzbare Leistungen, so dass sie selbst eine hohe Kostenwirksamkeit erreicht. Die Stakeholder erlangen so mehr Klarheit über das Programm und dessen Steuerung.

Die Standardkritik an diesem Evaluationsmodell lautet, Entscheidungssituationen seien selten planbar. Außerdem wird angenommen, dass Entscheidungen weniger rational, auf Grundlage von fundierten Informationen, sondern vielmehr beeinflusst durch Routinen, Machtgefüge oder andere außerhalb der Sache liegende Einflüsse getroffen werden. Dieser Einwand betrifft alle anderen Evaluationsmodelle gleichermaßen. Sind die Interessenkonflikte – insbesondere die unausgesprochenen – zu groß oder gibt es andere Gründe für taktische Manöver bis hin zu gezielten Fehlinformationen und Zurückhaltung von Wissen über Entscheidungssituationen, droht

die Evaluation ins Leere zu laufen. Interessen von Personen, die qua sozialem Status, Alter, Nationalität usw. von Entscheidungen ausgeschlossen sind, werden in diesem Ansatz nicht berücksichtigt. Daher wird ihm auch vorgeworfen, eine zu geringe Unabhängigkeit gegenüber machtvollen Interessen aufzuweisen.

2.3.3.2 Nutzungsgesteuerte Evaluation

Dieses Modell legt Planung, Durchführung und Ergebnisvermittlung von Evaluationen so an, dass ihre Nutzung durch die vorgesehenen Nutzer/-innen optimiert wird. Unter Nutzung wird sowohl die Verwendung der Ergebnisse als auch des Evaluationsprozesses selbst verstanden. Die Nutzung wird während des gesamten Evaluationsablaufes vorbereitet, indem informelle Kontakte und formelle Abstimmungsgespräche mit den Beteiligten stattfinden und beständig schriftliche Informationen (z.B. in Form von periodisch versandten „Neuigkeiten aus der Evaluation“) oder mündliche Zwischenpräsentationen stattfinden. Damit sucht dieses Modell grundsätzlich eine breitere Öffentlichkeit als der entscheidungsorientierte Ansatz. Die Nutzbarkeit einer Evaluation hängt wesentlich davon ab, ob die Evaluatoren/-innen die primär vorgesehenen Nutzer/-innen (*primary intended users*) dafür gewinnen können, die Voraussetzungen der Nutzung vorab zu klären. Nutzung wird definiert als eine Handlung, die durch *konkrete* Personen ausgeführt wird. Diese müssen bekannt sein oder vorausschauend konstruiert werden, ebenso die von ihnen beabsichtigten Verwendungen der Evaluation und ihrer Ergebnisse. Um Nutzung zu erreichen, muss diese gezielt vorbereitet werden, wozu insbesondere Workshops, Simulationen und andere interaktive Verfahren dienen.

Steuerungsfaktoren sind die intendierten Nutzungen der intendierten Hauptnutzer/-innen. Diese Nutzungen können, müssen aber nicht „Entscheidungen“ sein.

Alle Zwecke, von der Entscheidungsfindung und Programmverbesserung bis hin zur Erweiterung der Wissensgrundlage, sind möglich. Der nutzungsorientierte Ansatz ist insofern hoch adaptiv.

Fragestellungen gehen hervor aus den im Dialog mit den vorgesehenen Nutzern/-innen identifizierten Informationsinteressen. Diese werden in der Regel von den Evaluationsverantwortlichen formuliert und anschließen kommunikativ validiert. Fragestellungen können im Ablauf der Evaluation an Relevanz gewinnen oder verlieren oder es können auch neue Fragestellungen hinzukommen.

Als Evaluationsmethoden kommen grundsätzlich alle Verfahren in Betracht, dabei besonders diejenigen, die für die vorgesehenen Nutzer/-innen nachvollziehbar und plausibel darstellbar sind (Anforderung hoher „augenscheinlicher“ Validität).

Da gemäß diesem Modell Evaluationen durchgängig entlang der präzisierten und gegengeprüften Nutzungserwartungen der Adressatengruppen ausgerichtet werden, ist die Wahrscheinlichkeit der tatsächlichen Nutzung hoch. Widerstände gegen Evaluation können gezielt gemindert werden. Evaluation unterstützt ein zielgerichtetes Wissensmanagement und das Lernen in Organisationen. Bei denjenigen, die sich beteiligten, wird ein „evaluativer“ Blick auf die von ihnen verantworteten Programme und Maßnahmen geschult, so dass die Evaluation hohe Nachhaltigkeit in Politikfeldern und Organisationen erreichen kann.

Substantielle Zeit- und materielle Ressourcen müssen in die *Steuerung* dieser Evaluationen investiert werden. Für umfangreiche Datenerhebungen (große Stichproben/ aufwändige Instrumentenentwicklung) werden damit die Ressourcen eingeschränkt. Eine starke Nutzung der Evaluation wird oft erkauft durch eine stark fokussierte, oft qualitativ/deskriptiv geprägte Datengrundlage. Der Evaluationsprozess ist anfällig gegenüber Versuchen von Nutzergruppen, Druck auf seine Anlage oder die Darstellung von Ergebnissen auszuüben. Nutzungsfokussierte Evaluatoren/-innen müssen über ein breites Kompetenzprofil verfügen; sie müssen starke Kommunikatoren/-innen sein und Abstimmungsprozesse beteiligtenorientiert und gleichzeitig effizient führen können. Nur wenige Evaluatoren/-innen bringen ein so breites Kompetenzprofil mit, wie für dieses Evaluationsmodell erforderlich, was ggf. aufwendigere Teamlösungen erforderlich macht, wodurch evtl. auch Abhängigkeiten eines/-r einzelnen stark involvierten Evaluators/-rin vorgebeugt werden kann.

2.3.3.3 Stakeholder-Interessen-gesteuerte Evaluation

Das Modell der *Deliberative Democratic Evaluation* verlangt, demokratische Verfahrens-Prinzipien im gesamten Evaluationsprozess zu berücksichtigen. Dadurch soll Evaluation auch dort zu ausgewogenen Ergebnissen führen, wo es gegensätzliche Wert-/Interessen-Positionen gibt. Dies geschieht durch den gleichberechtigten Einbezug der Betroffenen und Beteiligten und aller Adressatengruppen in allen Phasen des Evaluationsprozesses.

House und Howe – die Begründer dieses noch selten umgesetzten, dabei viel diskutierten Ansatzes – sehen *Deliberative Democratic Evaluation* nicht als isoliertes Evaluationsmodell. Wenn andere Evaluationen folgenden drei Anforderungen genügen, könnten auch diese als *deliberative democratic* bezeichnet werden:

Der *Einbezug* der relevanten Interessen aller verschiedenen Stakeholder-Gruppen wirkt einseitiger Voreingenommenheit/Beeinflussung von Prozess und Ergebnis der Evaluation entgegen. Es müssen ggf. Vertreter/-innen für solche Gruppen bestimmt werden, die sich selbst nicht vertreten können.

Die Auseinandersetzung mit und unter den Stakeholdern ist notwendig, um diejenigen Interessen zu identifizieren, die sie tatsächlich zu bestimmten Äußerungen oder Handlungen bewegen.

Abwägung bedeutet, Daten, Informationen und Interessen (auch Werte) miteinander in Verbindung zu bringen. Die Evaluatoren/-innen sollen selbst nicht werten, aber einen ausgewogenen, diskursiven Prozess anstoßen und begleiten, der zu wertenden Ergebnissen führt.

Die Evaluatoren/-innen müssen alle Stakeholder-Gruppen identifizieren und sicher gehen, dass zumindest ein Repräsentant jeder Gruppe in die Evaluation aktiv einbezogen wird. Sollte es dabei große Unterschiede in der Macht der Beteiligten geben, werden machtnivellierende Verfahren eingesetzt: Moderation, Mediation, Ombudslösungen. Die Evaluatoren/-innen sollen es unterstützen, dass die Stakeholder-Vertreter sich möglichst authentisch und angemessen intensiv engagieren. Demokratische Partizipation soll genutzt werden, um schließlich zu sicheren und ausgewogenen Schlussfolgerungen zu kommen.

Steuerungsfaktoren sind die transparenten und einem demokratischen Abwägungsprozess zugänglich gemachten Interessen aller Beteiligten und Betroffenen.

Die Fragestellungen werden im gleichberechtigten Dialog möglichst aller erarbeitet. Kernfragestellungen beziehen sich dabei immer auf die Verwendbarkeit des Programms im Hinblick auf die Interessen der Betroffenen und Beteiligten.

Es werden über den ganzen Prozess der Evaluation hinweg solche Methoden bevorzugt, die Inklusion, Dialog und Abwägung erlauben und fördern.

Der Ansatz sieht Evaluation als Beitrag zur Demokratisierung auf lokaler bis nationaler Ebene. Verzerrungen, z.B. durch einseitig interessegeleitete Vorgaben der

Auftraggeber/-innen, kann entgegengewirkt werden. Akzeptanz der Evaluation, ihre Umsetzung und Nutzung können gefördert werden.

Es kann den Evaluatoren/-innen schwer fallen, die tatsächlichen Informationsinteressen der Stakeholder zu identifizieren. Der Dialog kann so viel Zeit beanspruchen, dass es zu Verzögerungen und gar zur Nicht-Durchführung vorgesehener Datenerhebungen kommt. Durchführbar ist eine solche Evaluation nur, wenn die Auftraggeber/-innen bereit sind, einen großen Teil ihrer Macht zu teilen und vorläufige Ergebnisse zu relativ frühen Zeitpunkten an eine Öffentlichkeit weiterzugeben.

Der Ansatz erscheint als eine idealtypische Vorstellung von Evaluation, wie sie zumindest in vollem Umfang in der Realität – besonders bei bundesweiten Programmen – kaum durchführbar sein wird. Auch die Übertragbarkeit auf andere Modelle muss bezweifelt werden. Die Autoren erkennen diese Schwierigkeit an und plädieren für eine Durchführung „so gut wie möglich“.

2.3.4 Wertepositioniertes Modell

Der nachfolgend charakterisierte Evaluationsansatz wird eingesetzt, um Beteiligten und Betroffenen mit Mitteln der Evaluation dabei zu helfen, sich selbst zu helfen und die Programme zu verbessern. Als Werte und Interessen werden insbesondere die der Benachteiligten, aber auch der Programmmitarbeitenden zu Grunde gelegt, welche Gruppen durch die Evaluation „gestärkt“ werden sollen.

2.3.4.1 Selbstorganisationsgesteuerte Evaluation

Empowerment Evaluation oder *Inclusive Evaluation* wendet Evaluations-Konzepte, -techniken und -ergebnisse an, um das Programm zu verbessern und Selbstorganisation der Beteiligten und Betroffenen zu fördern. Die Programmverantwortlichen, -mitarbeitenden und -nutzer/-innen einschließlich der Auftraggebenden führen *gemeinsam* die Evaluation durch, wobei die externen Evaluatoren/-innen wesentlich als Moderierende, Beratende oder Ausbildende in Evaluation fungieren oder Dienstleistungen bei der praktischen Durchführung der Evaluation leisten (Datenerhebung, Auswertung).

Die *Empowerment Evaluation* umfasst folgende Hauptschritte:

1. Programm-Betroffene und -beteiligte werden dabei unterstützt, die Leitziele (*mission*) des Programms zu formulieren;

2. sie identifizieren die wichtigsten Programmaktivitäten und beschreiben deren Durchführung systematisch mit Hilfe von Datenerhebungsmethoden; sie interpretieren die Daten, und bewerten das Programm in gemeinsamer Diskussion;
3. sie entwickeln Ziele und Strategien für eine künftige Verbesserung und vereinbaren ein evaluatives Vorgehen für die künftige Betrachtung und Bewertung des Programms.

Dieser Prozess soll sich idealerweise zirkulär verstetigen. Die Unparteilichkeit der Schlussergebnisse und -bewertungen soll dadurch gesichert werden, dass in deren Diskussion alle Beteiligten und Betroffenen einbezogen werden.[50]

Steuerungsfaktoren sind die länger- und mittelfristigen Ziele, Bedürfnisse und Motive der Stakeholder.

Der übergeordnete Zweck ist die Förderung der Selbstbestimmung (*Empowerment*). Die Programmbeteiligten sollen zu einer systematischen und andauernden (selbst-) evaluativen Tätigkeit befähigt werden als Instrument zur fortlaufenden Verbesserung des Programms.

Die Gemeinschaft von Praktiker- und Nutzergruppen des Programms entwickelt die Fragestellungen, die in fortwährender Interaktion geschärft und ausdifferenziert werden.

Es können verschiedenste – vorzugsweise leicht zu erlernende und einfach zu beherrschende – qualitative oder quantitative Methoden eingesetzt werden.

Idealerweise entsteht eine Atmosphäre gemeinsamen Lernens in gemeinsamer Verantwortung. Durch den starken Einbezug in den Prozess der Evaluation kann soziale Benachteiligung im Bereich der Handlungskompetenzen ausgeglichen werden. Die Evaluation ist sehr kostengünstig. Den Stundeneinsätzen beratender Fachkräfte stehen sowohl Qualifizierungseffekte als auch unmittelbare Qualitätsverbesserungen des Programms gegenüber.

Das Modell ist nur dann umsetzbar, wenn eine hoch rezeptive und motivierte Gruppe von Beteiligten gebildet werden kann. Die starke Unterstützung von Benachteiligten

---

50 Dieses Modell steht dem der Selbstevaluation in Deutschland nahe, die als eine Methode der Qualitätsentwicklung im Bereich der sozialen Arbeit entwickelt wurde; (vgl. Heiner 1988, 1998 sowie v. Spiegel 1993, 2001).

durch externe Evaluatoren/-innen mag die Glaubwürdigkeit der Evaluation beeinträchtigen. In der Praxis kann es aber auch dazu kommen, dass die Evaluatoren/-innen vorwiegend mit den Fachkräften zusammen arbeiten, deren Interessen und Werte damit privilegiert werden. Besonders dann, wenn die Fachkräfte auch die Interessen der sie beschäftigenden Organisationen mit vertreten, kann die Glaubwürdigkeit der Evaluationsergebnisse und die Unabhängigkeit vorgenommener Bewertungen bezweifelt werden. Dem kann dadurch entgegen gewirkt werden, dass die Evaluationsberater/-innen darauf drängen, auch Externe (z.B. Vertreter/-innen von Praktikumbetrieben oder potentielle Arbeitgeber/-innen der Maßnahmeteilnehmer/-innen) in die Evaluationssteuerung einzubeziehen.

2.4 Realisationen der Modelle in deutschen Evaluationen

Mit der Identifikation und Auswertung deutscher Evaluationsstudien in Feldern der Armuts- und Reichtumsberichterstattung wird geprüft, in welchem Maße dort auf die im Kapitel 2.3 dargestellten Evaluationsmodelle Bezug genommen wird.

Beispielhafte Studien aus der Bundesrepublik Deutschland, die Programme zu armutsrelevanten Themen zum Gegenstand haben, wurden auf verschiedenen Wegen gesucht (Internet-Recherchen, Telefonate mit Experten/-innen und Auswertung einer sozialwissenschaftlichen Datenbank).

Insgesamt war der Ertrag mit knappen Dutzend einschlägigen Evaluationsstudien, die substantiell etwas zur verwendeten Evaluationskonzeption aussagen, gering. Ein Grund dafür könnte darin bestehen, dass gerade Auftragsevaluationen im Bereich der Bundes- und Landespolitik – die avisierte, selten tatsächlich getroffene Ebene unserer Suche – mit Verspätung oder auch gar nicht veröffentlicht werden. Schließlich könnte es sein, dass Evaluationen seltener an sozialwissenschaftliche Forschungsdatenbanken gemeldet werden als z.B. freie Forschungsvorhaben.

Die Recherche im Bereich Evaluation von Projekten des Bund-Länderprogramms „Stadtteile mit besonderem Entwicklungsbedarf – Die soziale Stadt“ ergab, dass es hier intensive Überlegungen zur Theorie und Methodologie der Evaluation gibt und es auch versucht wird, ein für die Vielzahl der Projekte handhabbares Evaluationskonzept zu entwickeln. Dies ist dabei lediglich in Ausnahmenfällen auf die aktuellen Diskussionen über Modelle der Evaluation im internationalen Raum bezogen. Außerdem steht bei „Die soziale Stadt“ – wie auch in anderen Bundesprogrammen – ein Monitoring basierend auf der Erhebung von Kennzahlen („Indikatoren“) stark im Vordergrund, das im wesentlichen Outputs erfasst, nicht aber Outcomes oder gar (Outcome-)Wirkungen.

Einige Studien sind einem klassischen sozialwissenschaftlichen Forschungsdesign verpflichtet, verstehen sich selbst als „Evaluation“ oder „Evaluierung“, thematisieren jedoch kaum Fragen der Entstehung und Prioritätensetzung von Evaluationsfragestellungen und lassen nicht erkennen, wie Informationen wann zu welchen Zweck an welche Adressaten/-innen-Gruppen vermittelt werden sollen.

Insgesamt lässt sich feststellen, dass bei den identifizierten Studien zwar häufig ein Kapitel oder Abschnitt dem methodischen Vorgehen (im Sinne von: Methoden zur Er-

hebung von Daten) gewidmet wird, die Darstellung des Evaluationsmodells, dem gefolgt wird, jedoch sehr selten erfolgt.

Insofern bieten die nachfolgend steckbriefartig skizzierten Studien positive Ansätze. Sie kommen aus den Bereichen Beschäftigungsförderung, Kinder- und Jugendhilfe, Reform der Sozialhilfe, Integration ausländischer Mitbürger und Stadtentwicklung und haben eine Bedeutung im Kontext von Armutsprävention oder -bekämpfung. [51]

Abbildung 12: Beispiel für eine Evaluation ohne Bezug auf Theorie und Modelle der Evaluation:

ELSES (*Evaluation of Local Socio-Economic Strategies in Disadvantaged Urban Areas*) untersucht jene Programme und Projekte, die lokale Arbeit schaffen und sichern, Erwerbsquellen vermehren sowie sozialen und ökonomischen Infrastruktur gegen den Ausschluss von benachteiligten Nachbarschaften/Stadtteilen revitalisieren wollen. Zentrales Ergebnisdokument der wissenschaftlichen Begleitung sind im Verlauf von zwei Jahren angefertigte Fallstudienberichte von sechs Nachbarschaften/Stadtteilen aus sechs Mitgliedsländern der Europäischen Union.

Dem Untersuchungsansatz ist im Schlussbericht der kurze Abschnitt *The research design* (ELSES 2000, S. 7-10) gewidmet. Die Evaluationsfragestellungen sind skizziert wie folgt:

- *„descriptive: What is happening, who are the actors, what approaches are currently adopted?*
- *causal: to understand and assess the relation between institutional and organisational structures and effects and impact produced,*
- *normative: applying a common set of evaluation criteria to assess whether impacts and results of policy intervention have been satisfactory."*

Nachfolgend werden die eingesetzten Datenerhebungsinstrumente skizziert. Als methodische Bezugspunkte werden die beiden Praxishilfen der MEANS-Collection (EC MEANS 1999) und der International Labour Organisation (ILO 1999) genannt. Bezüge zu evaluationstheoretischer Literatur fehlen. Die Kategorisierung beschreibend-kausal-normativ wirkt ad hoc gebildet. Das Thema des Wertebezugs wird unter Verweis auf *a common set* ausgeblendet.

Beispiel 1: Experimentaldesign-gesteuerte Evaluation

Mit der Evaluation des **Einstiegsgelds in Baden-Württemberg** wird der gleichnamige Modellversuch bewertet, der einen zeitlich befristeten Zuschuss zur Sozialhilfe in Form eines erhöhten Freibetrags bei der Anrechnung von Erwerbseinkommen auf die Sozialhilfe gewährt.

---

51 Sie werden im Anhang II, S. 225ff. in einem Kurzportrait beschrieben, welches das Programm, Evaluationszweck und -fragestellungen, das grundlegende Evaluationsmodell, die verwendeten Methoden, den Umgang mit Werten von Stakeholdern und das vorgesehene Nutzungskonzept skizziert. Hier wird ein Auszug daraus in neuen Kurz-Steckbriefen abgedruckt.

Dies soll die Aufnahme eines Beschäftigungsverhältnisses am ersten Arbeitsmarkt fördern.

„Die Evaluation soll beantworten, inwieweit der mit dem Einstiegsgeld-Konzept angezielte Abbau der ‚Sozialhilfefalle' zu positiven Beschäftigungseffekten führen kann." (Dann/Kirchmann, Spermann/ Volkert 2002, S. 15), um über eine evtl. Übertragung dieser Maßnahme entscheiden zu können.

Genutzt wird das Experimentaldesign-gesteuerte Evaluationsmodell (in Kombination mit deskriptiver Teilnehmerstatistik und Leitfadenbefragung). Unterschiedliche Werte Beteiligter werden in der Studie nicht aufgegriffen.

Beispiel 2: Quasi-Experimentaldesign-gesteuerte Evaluation

Eine mikroökonometrische Evaluation (Jerger/Pohnke/Spermann 2001) richtet sich auf die **Mannheimer Arbeitsvermittlungsagentur (MAVA),** die mit einem integrierten Fallmanagement arbeitet und einen Beratungsdienst zur Unterstützung der Sachbearbeiter/-innen des Sozialamtes eingerichtet hat.

Die Wahrscheinlichkeit einer erfolgreicheren Vermittlung von arbeitsfähigen Sozialhilfeempfängern/-innen durch die MAVA gegenüber einer „üblichen" Betreuung durch das Sozialamt soll abgeschätzt und Effekte auf die Nachhaltigkeit des Beschäftigungsverhältnisses sollen ermittelt werden.

Die Quasi-Experimentaldesign-gesteuerte Evaluation vergleicht die von der MAVA betreute Programm-Teilnehmenden-Gruppe mit einer nachträglich gebildeten Kontrollgruppe aus arbeitsfähigen Hilfebeziehern/-innen. Die soziale Wertigkeit der Maßnahmenziele wird als gegeben vorausgesetzt; alternative Ziele werden nicht diskutiert.

Beispiel 3: Programmkosten/-nutzen-gesteuerten Evaluation

Die fiskalische und soziale Kosten-Nutzen-Analyse **örtlicher Beschäftigungsförderung** für arbeitslose Sozialhilfeempfänger **(**Trube, 1995) will klären, wie die Aufwendungen, die der Kommune durch Finanzierung und Durchführung von Maßnahmen örtlicher Beschäftigungsförderung entstehen, in Relation zu erwartbaren fiskalischen und sozialen Nutzeneffekten stehen. Zudem werden die Treffgenauigkeit der Beschäftigungsförderung und Akzeptanz des Maßnahmenangebots und seiner Zugänglichkeit für die Teilnehmenden untersucht.

Es wird eine Programmkosten/-nutzen-gesteuerte Evaluation unter Einsatz von experimentellen Designelementen durchgeführt.

Es werden nicht nur die fiskalischen Aufwendungen und Erträge/Einsparungen, sondern auch psychosoziale Be- und Entlastungseffekte untersucht. Insbesondere letztere enthalten Wertentscheidungen (soziale, gesundheitliche, psychische und gesellschaftliche Integration und Desintegration), deren Zustandekommen nicht thematisiert wird.

Beispiel 4: Programmtheorie-gesteuerte Evaluation

Bei den **Projekten im Rahmen der EU-Gemeinschaftsinitiative URBAN I** in Saarbrücken[52] handelt es sich um die lokale Umsetzung einer Gemeinschaftsinitiative der Europäischen Union, die zu einer ausgewogenen wirtschaftlichen und sozialen Entwicklung in benachteiligten städtischen Gebieten beitragen soll. Wichtiger Bestandteil ist eine kleinräumige Bürger/-innenbeteiligung, die sicherstellen soll, dass die Interessen der Bewohner/-innen berücksichtigt werden.

Die Evaluation soll eine Verlaufskontrolle und Bewertung der Zielerreichung der Projekte leisten.

Zu Grunde gelegt wird das Modell der Programmtheorie-basierten Evaluation: Basierend auf der Analyse der sozioökonomischen Ausgangslage werden Ziele des Programms festgelegt und Projekte zugeschnitten. Diese textlich-qualitativen Zusammenhänge von Projektplanung und -umsetzung und erreichten Wirkungen (*logical framework*) sind mit Hilfe von Projektplanungsübersichten schematisiert und Grundlage der Bewertung der Programmumsetzung.

Beispiel 5: Spannungsthemen-gesteuerte Evaluation

Beim Pilotprojekt „**Qualitätsentwicklung und frühkindliche Erziehung in türkischen Vereinen**“ handelt es sich um einen mehrstufig geplanten bundesweiten Aktivierungs-, Begleit- und Weiterbildungsprozess. Junge türkische Eltern sollen in der frühkindlichen Erziehung unterstützt werden, indem Schulungen zum Thema entwickelt und angeboten werden. Letztlich soll ein Forum geschaffen werden, in dem Eltern in einer durch Migration und kulturelle Differenz geprägten Lebenslage sich mit einer Kursleitung zu Erziehungsfragen austauschen können. (vgl. Kalscheuer, 2001, S. 69)

Zweck der Evaluation ist die Unterstützung der Verbesserung der Maßnahmen auf Basis einer kontinuierlichen Reflexion. Eine ursprünglich angelegte Programmtheorie-gesteuerte Evaluation war wegen Instabilität des Programms nicht durchführbar. Das alternativ genutzte responsive Modell nach Stake (1995) arbeitete die Spannungsthemen – besonders auch auf der Ebene kultureller Werte – heraus, die zwischen verschiedenen Beteiligten und Betroffenen bestehen. Verschiedene Stakeholder wirkten an der Steuerung der Evaluation mit. Werte wurden explizit gemacht und dabei nicht abschließend geklärt.

Beispiel 6: Dialoggesteuerte Evaluation

Die Jugendhilfeeinrichtung Forstenwalde in Brandenburg sieht sich als eine „innovative, stärker als üblich institutionalisierte Alternative zur U-Haft“. Die Jugendlichen sollen in einem pädagogisch strukturierten Kontext ihre sozialen und kognitiven Kompetenzen stärken.

Die Evaluation will dazu beitragen, eine von den Akteuren geteilte Einstellung zu Zielen und Formen der Arbeit in der Einrichtung herzustellen und die Zusammenarbeit und das Konzept zu verbessern.

---

52 Ein Bericht der Evaluation liegt uns nicht vor, die Darstellung basiert auf Mitteilungen der Evaluierenden sowie Informationen aus Darstellungen der Evaluation bzw. des Programms im Internet auf den Seiten des isoplan-Instituts und der Stadt Saarbrücken: www.isoplan.de/europa/Eval.htm, http://urban.saarbruecken.de oder www.saarbruecken.de/sbnet/08/urban/urb_anf.htm.

Die Dialoggesteuerte Evaluation „ist ausdrücklich 'konstruktivistisch' angelegt“ (Pleiger/Weißmann/ Friedrich/Klawe, 2002, S. 49). Die Einrichtung wird als ein soziales und kulturelles System gesehen, an deren Funktionieren zahlreiche Akteure beteiligt sind. Es soll beschrieben werden, in welchem situativen Kontext die sozialen Konstruktionen der Beteiligten, z.B. über abweichendes Verhalten, zu Stande gekommen sind, bzw. die gefundenen Deutungen sollen in das System rückgekoppelt und diskutiert werden. Werte der Stakeholder sind Gegenstand des andauernden Dialogs.

Beispiel 7: Nutzungsgesteuerte Evaluation

Mit den Modellvorhaben zur Pauschalierung von Sozialhilfe (Ministerium für Arbeit und Soziales, Qualifikation und Technologie NRW 2002) wird angestrebt, in der Sozialverwaltung die Aufwendungen für Routinetätigkeiten zu reduzieren und Ressourcen in den Sozialämtern für persönliche Hilfen, insbesondere Beratungsangebote, Hilfeplanung und Case Management, frei zu setzen. In verschiedenen Kreisen und kreisfreien Städten in NRW wird das bundesweite Modellprojekt umgesetzt.

Die Evaluation soll Informationen bereitstellen, die gesicherte Entscheidungen des Gesetzgebers zur künftigen Ausgestaltung der Sozialhilfe, insbesondere über die Ausgestaltung weiterer Pauschalierungen und persönlicher Hilfen ermöglichen und darüber hinaus die öffentliche Diskussion mit Informationen versorgt. Zentrale Aufgaben sind u.a. Einleitung eines Erfahrungsaustausches, die Untersuchung von Umsetzung und Ergebnissen der Modellvorhaben sowie die Unterstützung bei Organisationsentwicklung und Qualifizierung der Träger.

Dabei sollen den vorgesehenen Nutzerinnen und Nutzern in Politik, Verwaltung und der Öffentlichkeit nützliche Informationen für ihre jeweiligen Interessen bezüglich der Weiterentwicklung der Sozialhilfe zur Verfügung gestellt werden.

Perspektiven und Werte der Beteiligten und Betroffenen sollen in die Evaluation einbezogen werden, um eine Basis für glaubwürdige und nützliche Evaluationsergebnisse zu schaffen.

Beispiel 8: Stakeholder-Interessen-gesteuerte Evaluation

Das **Bundesmodellprogramm „Mobile Jugendsozialarbeit für junge Menschen ausländischer Herkunft**“ zielt darauf, über eine aufsuchende Strategie neue Zugangswege zu Jugendlichen ausländischer Herkunft zu erproben und für sie weiterführende Angebote und Hilfen zur Unterstützung der sozialen und beruflichen Integration zu entwickeln. Bestehende Hilfs- und Angebotsstrukturen sollen für die Jugendlichen geöffnet und untereinander vernetzt werden.

Die Evaluation will die Prozesse des Bundesmodellprogramms beschreiben und dessen Effekte und Nebeneffekte bewerten. Daneben geht es um eine Verbesserung der laufenden Modellprojekte.

Das Modell der Stakeholder-Interessen-gesteuerten Evaluation wird mit dem Design einer Cluster-Evaluation kombiniert. Im „offen” und partizipativ angelegten Vorgehen sollen die Anliegen und Konfliktthemen der am evaluierten Programm beteiligten Gruppen zu Fragestellungen und einem konkreten Vorgehen führen, indem sie im Austausch mit Projektmitarbeitern/-innen und Bundestutoren/-innen entwickelt werden (Haubrich/Frank, 2000, S 13). Der Umgang mit Werten der Stakeholder wird im Bereich von intendierten Effekten thematisiert.

Beispiel 9: Selbstorganisationsgesteuerte Evaluation

Das Programm „Beratungsstellen und Arbeitslosenzentren für Langzeitarbeitslose und von Langzeitarbeitslosigkeit bedrohte Personen des Landes NRW“ umfasst rund 50 Beratungsstellen und 100 Arbeitslosenzentren. Diese bieten Beratung, Fortbildung, Freizeitgestaltung und Austauschmöglichkeiten mit dem Ziel an, die Zielpersonen zu stabilisieren und aktivieren; auch soll die Vernetzung mit anderen lokalen Akteuren der Arbeitsmarktpolitik gefördert werden.

Zweck der Evaluation ist die Beschreibung und Bewertung der Arbeit von Beratungsstellen und Arbeitslosenzentren im Sinne einer Rechenschaftslegung gegenüber der Öffentlichkeit sowie deren Verbesserung auf der Basis der gefundenen Ergebnisse.

Die Evaluation wird an den Fragen der Stakeholder orientiert (Vorgehen: emergent). Gleichzeitig sind grundlegende Fragestellungen vom Auftraggeber vorgegeben. Stakeholder-Gruppen werden durch intensive Kommunikation, u.a im Rahmen von Gremien, in die Entwicklung des Evaluationsdesigns beteiligt. Zusätzlich wird ein Empowerment-Ansatz verfolgt, indem Mitarbeitende an lokalen Projekten zur Selbstevaluation ihrer Arbeit angeregt und geschult werden (Arbeitsstelle für Evaluation der Universität zu Köln, 1999, S. 100).

Werte verschiedener Beteiligter sind aufgenommen; es besteht eine Tendenz, die Interessen der Mitarbeitenden in den Einrichtungen bei Planung und Auswertung besonders zu berücksichtigen.

Es werden insgesamt neun beispielhafte Studien aus Deutschland in der Modelltypologie vorgestellt, die sich je einem evaluationstheoretischen Modell zuordnen lassen. Keine Beispiele fanden sich für die Modelle „Programmziel-gesteuerte Evaluation“, „Entscheidungsgesteuerte Evaluation“ sowie „Kontext-Mechanismus-gesteuerte Evaluation“. Die vorgestellten Studien stellen eine Auswahl dar, die Übersicht ist nicht als repräsentativ anzusehen. In vielen anderen Berichten zu themenrelevanten Studien finden sich keine oder lediglich andeutungsweise Bezüge zur evaluationstheoretischen Diskussion. Zentrale Merkmale dieser Evaluationen, wie z.B. der Ausweis des Evaluationszweckes, die Herkunft und Priorisierung von Fragestellungen, die Haltung zur Frage der Wertberücksichtigung oder des Einbezugs oder Nicht-Einbezugs von Beteiligten und Betroffenen in die Evaluationssteuerung werden nicht angesprochen. Das Zustandekommen methodischer Grundentscheidungen der Untersuchungen – z.B. die Auswahl der Methoden oder die Festlegung der Bewertungskriterien betreffend – ist vielfach nicht nachvollziehbar und erscheint daher „ad hoc“ getroffen.

Diese Aussagen beruhen auf einem Ausschnitt von Evaluationsstudien im Bereich der für die Armuts- und Reichtumsberichterstattung relevanten Themenfelder, der keine Repräsentativität für die Gesamtheit der z.B. in den letzten 12 Jahren durch-

geführten Untersuchungen beanspruchen kann. Ein erstes Problem besteht darin, dass Berichte nicht öffentlich zugänglich sind oder ihr Fundort nicht identifiziert werden kann. Es ist wünschenswert sicherzustellen, dass insbesondere von öffentlichen Auftraggebern finanzierte Evaluationsstudien kurzfristig veröffentlicht und z.B. in Datenbanken zugänglich und die Berichtstexte – etwa durch Download von Dateien – leicht abrufbar gemacht werden.

Die Auswertung der inhaltlichen Erkenntnisse und Ergebnisse dieser Studien erleichtert die Kumulation empirisch basierten Wissens zu Maßnahmen im Bereich der Armuts- und Reichungsberichterstattung. Dies kann genutzt werden, um neue Programme und Maßnahmen stärker erfahrungsbasiert zu planen mit dem Ziel, ihre Wirksamkeit zu erhöhen.

Eine systematische Auswertung der methodischen Berichtsteile entlang von Dimensionen wie Evaluationszweck, zentrale Fragestellungen, genutztes Evaluationsmodell kann zur methodologischen Reifung der Evaluationen beitragen und durch eine verbesserte Vorbereitung gezielter Nutzung durch Auftraggebende und andere Stakeholder einen weiteren – indirekten – Beitrag zur Verbesserung der evaluierten Programme leisten.

3 Anforderungen von Experten/-innen an Evaluationen im Bereich Armut bekämpfender Politik

3.1 Einleitung

Im folgenden Berichtsteil werden die Ergebnisse der Interviews mit Experten/-innen aus dem Bereich der Armuts- und Reichtumsberichterstattung dargestellt. Es wurden Mitte Oktober bis Anfang November 2002 neun *face-to-face*-Interviews mit Forschern/-innen und Evaluatoren/-innen im Umfeld der Armuts- und Reichtumsberichterstattung geführt sowie im November 2002 zwei Fokusgruppen mit Mitarbeitenden aus Ministerien veranstaltet, die (potentielle) Auftraggeber/-innen für Evaluationen im Kontext der Armut bekämpfenden Politik sind. Zunächst wird dargestellt, welche Zielsetzungen und Fragestellungen die Erhebungen verfolgten und wie bei ihrer Durchführung und der Auswertung der Ergebnisse vorgegangen wurde. Danach werden die Ergebnisse der Erhebungen dargestellt.

Die Ergebnisse bieten einen Überblick über die Evaluationspraxis und die Qualitätsanforderungen an Evaluation im Kontext der Armuts- und Reichtumsberichterstattung aus Sicht von Auftraggebern/-innen und Forschern/-innen. Mit ihrer Darstellung soll die Anschlussfähigkeit der im ersten Teil vorgestellten Modelltypologie an den aktuellen Diskurs im Bereich der Evaluation von Maßnahmen zur Armutsbekämpfung sichergestellt werden.

3.2 Zielsetzung und Fragestellungen

Zielsetzung der Fokusgruppenerhebung war die Konkretisierung der Nutzungsbedingungen von wirkungsorientierten Evaluationen aus Sicht der Auftraggeber/-innen sowie die Prüfung und gegebenenfalls Präzisierung der Standards zum Thema Nützlichkeit aus den „Standards für Evaluation“ (DeGEval 2002). Die Teilnehmer/-innen wurden hinsichtlich ihrer Anforderungen an Evaluationen, bezüglich deren Durchführung, Nutzen, Verwendungskontext und Zwecken befragt. Insbesondere wurde auch nach dem von Auftraggebern/-innen erwarteten Umgang mit Werten bzw. Wertespannungen bei Evaluationen gefragt. Gegenstand der Befragung waren Programmevaluationen in Feldern der Politik zur Vermeidung von Armut und zur sozialen Integration von Armut bedrohter oder betroffener Menschen. Der Schwerpunkt sollte dabei auf Programmen mit hohem Interventionscharakter liegen.

Die Fokusgruppenmethode wurde gewählt, da angenommen wurde, dass Wissen für die Beantwortung der genannten Fragestellungen oft implizit vorliegt und in der themenorientierten Diskussion der Fokusgruppe Raum für den freien Austausch und die Weiterentwicklung von Erwartungen, Meinungen und Positionen besteht (Krueger 1994). Damit eine freie Meinungsäußerung stattfindet, sollte die Gruppe möglichst homogen zusammengesetzt sein. Die Kommunikation der Teilnehmenden sollte nicht durch ein Machtgefälle oder persönliche Verflechtungen in der Gruppe gestört werden, was weitestgehend realisiert werden konnte.

Die an einem thematischen Leitfaden orientierten Interviews sollten dazu dienen, die Anschlussfähigkeit der in diesem Bericht vorgestellten Modelltypologie an den beginnenden Fachdiskurs zum Thema wirkungsorientierte Evaluation im Kontext der Armut bekämpfenden Politik zu prüfen. Das Wissen der Forscher/-innen wurde genutzt, um eine Grundlage für die Erstellung der Typologie zu bilden und notwendige Veränderungen und Ergänzungen zu konkretisieren. Die Fragestellungen der Forscher/-innen-Interviews bezogen sich schwerpunktmäßig auf konzeptionelle und methodische Aspekte von Evaluationen, es wurden aber auch – genau wie in den Fokusgruppen – Fragen bezüglich Zweck und Nutzen von Evaluationen sowie zu Wertespannungen im Rahmen von Evaluationen gestellt.

3.3 Vorgehen bei der Durchführung der Erhebungen

Die Teilnehmer/-innen an den Fokusgruppen und Interviews wurden in Abstimmung mit den Auftraggebern/-innen und auf Grund folgender Kriterien ausgewählt: Gesprächspartner/-innen für die Interviews sollten insbesondere solche Forscher/-innen sein, die einen breiten Überblick über die Evaluationspraxis und die eingesetzten Evaluationsmodelle im Bereich der in Frage kommenden, Armut bekämpfenden Programme haben. Sie wurden identifiziert über Fachliteratur, Wissenschafts- und Fachgesellschaften sowie Tagungsprogramme. Die Befragten sollten eine möglichst hohe Heterogenität in Bezug auf ihren disziplinären Hintergrund und – soweit feststellbar – ihrer bevorzugten Evaluationsmodelle aufweisen. Die Teilnehmer/-innen der Fokusgruppen stammen aus mehreren Bundesministerien und sind im Bereich der Armuts- und Reichtumsberichterstattung sowie angrenzenden Politikfeldern tätig. Sie haben bereits Evaluationen in Auftrag gegeben oder werden in Zukunft damit befasst sein.

Es sind im Zeitraum von zwei Monaten insgesamt neun Interviews mit Forschern/-innen geführt worden. An den zwei Fokusgruppen nahmen acht bzw. zwölf Mitarbeiter/-innen aus Ministerien teil.[53] Die Aussagen der Interviewpartner/-innen und der Fokusgruppenteilnehmer/-innen wurden sowohl sinngemäß protokolliert als auch mit Einverständnis der Befragten auf Tonband aufgezeichnet. Die Tonbandabschriften wurden den Interviewteilnehmern/-innen zur Stellungnahme und gegebenenfalls Korrektur zugeschickt.

3.4 Vorgehen bei der Auswertung der Ergebnisse

Die Auswertung sowohl der Ergebnisse der Interviews als auch der Fokusgruppen erfolgte systematisch unter Anwendung der Methode der Text-Sortier-Technik (TST), die durch Univation entwickelt wurde und langjährig im Einsatz ist.[54] Bei diesem Verfahren werden in einem ersten Schritt die Erhebungsbögen und Fragen codiert. Anschließend werden die Textpassagen des Datenmaterials in einzelne Sinneinheiten zerlegt mit dem Ziel, ähnliche Textelemente in inhaltlich homogenen Kategorien zusammenzufassen. Die gebildeten Kategorien werden mit einem Code versehen, so dass ein Kategoriensystem entsteht, welches in einer „Legende“ dokumentiert wird.

Für Dritte ist es im Nachhinein überprüfbar und nachvollziehbar, wie es zu den Aussagen und Schlussfolgerungen aus dem qualitativen Datenmaterial kommt. Die im Ergebnisteil veröffentlichten Zitate von Experten/-innen sind im vorliegenden Bericht zwecks Sicherstellung der Anonymität der Teilnehmenden nur begrenzt durch Quellenangaben ausgewiesen. Um Rückschlüsse auf einzelne Personen zu vermeiden, wurde lediglich deutlich gemacht, ob die Aussagen aus einem der Interviews (§ A) oder aus einer Fokusgruppe (§ B) stammen.

Die Aussagen der Interviewpartner/-innen und Fokusgruppenteilnehmer/-innen wurden innerhalb eines gemeinsamen Kategoriensystems ausgewertet, da sie sich auf gleiche Themenbereiche erstrecken und werden im folgenden Ergebnisteil entlang der vorgefundenen Sinnstrukturen zusammen abgebildet. Anhand der Quellenangabe lässt sich eine Zuordnung der Aussagen zu den beiden Gruppen vor-

---

53 Eine Liste der Teilnehmer/-innen an den Fokusgruppen und Interviews befindet sich im Anhang des Berichts.

54 Beywl/Schepp-Winter (2000), Materialien zur Qualitätsentwicklung, Zielgeführte Evaluation von Programmen, Heft 29, S. 62f.

nehmen, wodurch Schwerpunktsetzungen und Akzente nachvollziehbar sind. Darüber hinaus werden in den Schlussfolgerungen Gemeinsamkeiten und Unterschiede zwischen den Ergebnissen der Interviews und der Fokusgruppen angesprochen.

3.5 Ergebnisse aus den Erhebungen

Im Folgenden werden die Aussagen aus den Fokusgruppen und Interviews entlang der Kategorien „Werteberücksichtigung“, „Wirkungsorientierung“, „Aufgabenfelder von Evaluation“ und „Nutzenentstehung“ dargestellt.

3.5.1 Stellenwert und Umgang mit Werten in der Evaluation

Die einbezogenen Experten/-innen sprechen die Bedeutung von Werten in den folgenden Phasen des Evaluationsprozesses an:

- Festlegung der Politik- bzw. Programmziele
- Bestimmung des Evaluationsgegenstandes
- Auswahl der Methode
- Ergebnisinterpretation

Es wird intensiv diskutiert, durch welche Personen (Forscher/-innen, Auftraggeber/-innen) oder gesellschaftliche Konventionen die Wertevorgabe erfolgt und welche Rolle der Evaluation bei der Festlegung bzw. Klärung von Werten zukommt.

Die Forscher/-innen weisen darauf in, dass die **Ziele der Politikbereiche bzw. Maßnahmen**, an welchen sich die Evaluation bei der Beschreibung und Bewertung orientiert, normativ festgelegt sind. Als Ziele von Programmen im Bereich der Armutsbekämpfung oder -prävention werden u.a. genannt, Armut zu reduzieren oder bestimmte Personenkreise in Erwerbstätigkeit zu bringen. Es wird davon ausgegangen, dass dahinter bestimmte Werte liegen, z.B. „Armut ist kein angenehmer Zustand“ oder „Erwerbstätigkeit ist ein erstrebenswerter Zustand“. Gleichzeitig wird hervorgehoben, dass die Auswahl der Maßnahmen zur Bekämpfung von Armut an die Werte und Vorgaben des Grundgesetzes und der Wirtschaftsordnung gebunden ist.

Ob und inwieweit die Klärung von Werten Bestandteil der Evaluation sein soll, sehen die einbezogenen Experten/-innen unterschiedlich. Einerseits wird es als Aufgabe der Politik gesehen, die zu verfolgenden, wertgebundenen Ziele vor Beginn des Forschungsprozesses festzulegen. Die Normen- und Wertegebung sollte in einer poli-

tischen Debatte erfolgen. Einzelne Forscher/-innen weisen darauf hin, dass die Evaluation keinen Einfluss auf die Ziele der Programme nehmen dürfe, deren Erreichung sie überprüft, vielmehr sei es eine Frage der sozialpolitischen Auseinandersetzung, die richtigen Ziele des Kampfs gegen die Armut zu definieren.

Andererseits wünschen sich einige einbezogene Experten/-innen, dass bereits die Werteklärung Bestandteil der Evaluation sein soll. Obwohl die Zielfestlegung bereits im Vorfeld der Wirkungsforschung liege, müssten die Zieldefinition und die Vorgaben selber eigentlich Gegenstand kritischer Überprüfung sein. Es könne eine Aufgabe der Evaluation sein, auf die impliziten Annahmen einzugehen, also auf die unterschiedlichen Werturteile, die der Diskussion zu Grunde liegen.

Zudem könne durch eine Differenzierung der Zielebenen, indem Ziele nach Erreichbarkeit unterschieden werden, eine Klärung der Werte unterstützt werden. Auch der Evaluationsprozess könne dazu beitragen, vage und ungenau formulierte Zielvorgaben für die Beteiligten zu präzisieren und bei der Konkretisierung der Ziele die damit verbundenen Werturteile zu identifizieren. Es wird angemerkt, dass es möglicherweise keinen Konsens bezüglich der Umsetzung bestimmter Leitziele gibt.

> *„Meistens geschieht die Werteklärung punktuell und nicht durch eine allgemeine Definition. Man kann natürlich sagen, dass man die Gesellschaft gerechter machen möchte. Aber das ist eine sehr vage Formel. Punktuell genauer ist aber, zu sagen, wir wollen die Gesellschaft in der Weise gerechter machen, dass der Anteil derjenigen, die eine Bildung erhalten, die wenigstens zum Hauptschulabschluss führt, größer ist. Die Startchancen sind dann weniger ungleich. Dies kann man unter das Ziel der Verbesserung der Startchancengleichheit subsumieren. Dagegen dürfte es kaum möglich sein, allgemein und konsensfähig das Ziel der Startchancengleichheit zu definieren.“ (§ A)*

Die einbezogenen Experten/-innen halten Evaluation für besonders aussichtsreich in Politikbereichen, in denen bereits ein tragfähiger Konsens über die anzustrebende Zielrichtung besteht. Darüber hinaus sollten die Ziele von den Zielgruppen der Programme mitgetragen werden. Tatsächlich besteht nach den Aussagen der Experten/-innen in einzelnen Politikbereichen der Armuts- und Reichtumsberichterstattung auf Grund unterschiedlicher politischer Wertsetzung teilweise Uneinigkeit bezüglich der anzustrebenden Ziele. Zusätzlich erschwerten Zielkonflikte

unterschiedlicher Politikbereiche (z.B. der Wirtschafts- und Sozialpolitik) die Festlegung einer einheitlichen Zielstruktur.

Bezüglich der Frage, durch wen oder was die Werte vorgegeben werden, diskutieren die Experten/-innen verschiedene Modelle: Festlegung der Werte durch Forscher/-innen (z.B. technokratisches Modell), durch Auftraggeber/-innen (Politik), gemeinsame Festlegung durch Auftraggeber/-innen und Wissenschaftler/-innen oder durch gesellschaftliche Konventionen. Bei der technokratischen Beratung legten die Forscher/-innen ihre Wertannahmen zu Grunde, z.B. über die Grenzwerte von Strahlenintensität. Obwohl es sich dabei um Entscheidungen von großer gesellschaftlicher Reichweite handele, bestimme in diesen Fällen weniger die Politik als die Forschung über die zu verfolgenden Vorgaben. Im Gegensatz dazu existierten auch Beratungsmodelle, in denen die Werte der Forscher/-innen hinter den politisch fixierten Wertvorstellungen zurückbleiben und nicht in den Beratungsprozess einfließen. Als dritte Variante wird das prozessorientierte Beratungsmodell genannt, wobei eine Aushandlung zwischen den relevanten Werten der Forscher/-innen und Auftrageber/-innen stattfindet. Zudem wird angesprochen, dass die Auftraggeber/-innen teilweise keine eindeutigen Wertpositionen in den Forschungsprozess einbringen. In diesem Fall werden die Wertvorgaben maßgeblich durch die beteiligten Forscher/-innen getroffen, obwohl die Entscheidung als eine politisch gefällte erscheine. Daran schließt sich die Forderung einiger Forscher/-innen an, im Verlaufe des Evaluationsprozesses müsse eine Klärung der Wertannahmen der Auftraggeber/-innen erfolgen.

Wertpositionen beeinflussen neben der Festlegung der Politik - bzw. Programmziele die **Bestimmung des Evaluationsgegenstandes**, was die Forscher/-innen und Auftraggeber/-innen deutlich herausstellen. So sei zu Beginn eine Entscheidung darüber zu treffen, welches Programm, welche Dimensionen eines Programms und unter welcher Perspektive (z.B. Genderaspekt) der Programmprozess oder die Programmwirkungen betrachtet werden sollen.

Darüber hinaus sei die **Auswahl der Methode** im Evaluationsprozess durch normative Setzung bestimmt. Die einbezogenen Experten/-innen – vor allem die Forscher/-innen – weisen darauf hin, dass es notwendig sei, die Werturteile und Annahmen, die der Methodenwahl zu Grunde liegen, offen zu legen. Konkret heißt das, es müssen die Vor- und Nachteile der zur Verfügung stehenden Methoden diskutiert werden und die Auswahl einer bestimmten Vorgehensweise mit ihren bestehenden Implikationen

müsse (wie z.B. bei ökonometrischen Analysen implizit unterstellte Wahrscheinlichkeitsverteilungen) begründet und dargestellt werden.

> *„Eine alternative Vorgehensweise bestünde darin, dass man sagt, es gibt eine Palette von Methoden für dieses Problem, A, B, C, D. Aus den und den Gründen wähle ich C, im Vergleich zu B hat C folgende Vorteile, im Vergleich zu A jene Vorteile usw. Warum eine bestimmte ökonometrische Methode für adäquat gehalten wird, muss begründet werden. Dies ist die Forderung nach Transparenz der Methodenwahl." (§ A)*

Zudem stellen einige Experten/-innen heraus, dass die **Interpretation der Ergebnisse** maßgeblich durch zu Grunde liegende Wertesysteme gesteuert wird. Demnach könnten aus einem Ergebnis – ausgehend von unterschiedlichen Wertpositionen – divergierende Schlussfolgerungen gezogen werden bzw. kämen verschiedene Forschungsansätze auf Grund unterschiedlicher Herangehensweisen zu voneinander abweichenden Aussagen über die Wirkung von Maßnahmen. Letzteres könne dazu führen, dass sich die Politik die zu ihren Überzeugungen passenden Studien heraussucht.

Da normative Setzungen den gesamten Forschungsprozess begleiten, fordern viele Experten/-innen deren Identifikation und Thematisierung durch die Evaluation. Durch die Transparenz der Wertpositionen sollten Interpretationsvorgänge nachvollziehbar werden. Eine wertfreie Haltung der Wissenschaftler/-innen erscheint einigen Forscher/-innen teilweise illusorisch.

3.5.2 Aufgabenfelder der Evaluation

Nach den Aussagen der Befragten können Evaluationen zu verschiedenen Zeitpunkten im Programmzyklus stattfinden. Evaluation übernimmt, abhängig davon, zu welchem Zeitpunkt der Programmplanung und -durchführung sie einsetzt, unterschiedliche Aufgaben. Ein Teil der Experten/-innen beschreibt **ex-ante-Evaluationen**, die den Zweck haben, vor der Entscheidung für eine Maßnahme Programmalternativen bezüglich ihrer Umsetzbarkeit, Passgenauigkeit etc. zu überprüfen. Evaluationen, die zeitlich vor dem Programmstart liegen, könnten helfen „Felder für zukünftigen Handlungsbedarf offen zu legen" (§ A) und „verschiedene Konzepte noch vor der Anwendung zu bewerten, das wäre also eher eine Konzeptbewertung" (§ A). Darüber hinaus können sich die Experten/-innen vorstellen, dass

die Evaluation zur Zielklärung der Maßnahmen beiträgt. Dies kann vorab geschehen, indem die Zielsetzungen der beteiligten Akteure/-innen thematisiert werden.

Es wird angemerkt, dass sich die Funktion von Evaluation nicht auf die Ermittlung von Programmwirkungen beschränkt, sondern z.B. durch die Unterstützung bei der Bedarfsermittlung, der Zielklärung oder der Konzeptentwicklung dazu beitragen sollte, Maßnahmen zu gestalten und weiterzuentwickeln.

Nachfolgendes Zitat zeigt den Bedarf an klärender Evaluation, deren Aufgabe darin liegt, die Messbarkeit der Programmziele vorzubereiten und die Abstimmung von Interventionen zu unterstützen.

> *„Es ist notwendig, eine eindeutige Zuordnung vorzunehmen zwischen Zielstruktur und Mitteleinsatz.“ (§ A)*

Als sinnvoll werden auch Evaluationen **begleitend** zur Programmdurchführung eingeschätzt. Hierbei werden die erhobenen Daten als Zwischenergebnisse zur Prozessqualität für eine laufende Optimierung der Programmumsetzung rückgemeldet.

Ex-post-Evaluationen setzen zeitlich nach der Programmdurchführung ein und richten ihr Augenmerk überwiegend auf die Ergebnisebene. Sie legen offen, welche Ziele das Programm in welchem Umfang erreicht hat und ob tatsächlich die intendierten Wirkungen bei den Zielgruppen aufgetreten sind. In diesem Zusammenhang thematisieren viele Experten/-innen das Zurechnungsproblem, das heißt die ursächliche Rückführung der festgestellten Outcomes auf die Maßnahme, als eine große methodische Schwierigkeit im Bereich der Wirkungsanalyse. Die Probleme bei der Messung von Wirkungen werden im folgenden Kapitel näher dargestellt.

3.5.3 Das Konzept „Wirkung“

Nachfolgend werden die verschiedenen Arten von Wirkungen, bspw. Output, Outcomes, sonstige intendierte Wirkungen und Nebenwirkungen, die Schwierigkeiten bei der Wirkungsmessung sowie Auswege aus der Problematik diskutiert.

3.5.3.1 Wirkungsarten

Von den Befragten wird betont, dass eine Evaluation auf der Resultats-Ebene sowohl den Output als auch den Outcome eines Programms oder einer Politikmaßnahme beschreiben soll. Als Output können alle Aktivitäten, Produkte etc. beschrieben werden, die das Programm selbst herstellt (wie z.B. Besuchszahlen von sozialen Ein-

richtungen). Der Output stellt die erste, unmittelbar zu beschreibende Ebene der Resultate eines Programms dar.

Letztlich ist mit einem zufrieden stellenden Output (z.B. hohe Teilnahme an einem Programm zur Armutsbekämpfung) aber noch keine Aussage über die Resultate auf der Outcome-Ebene, also die erreichten Zielzustände bei den Zielgruppen, möglich. Eine hohe Nachfrage eines Programms hat nicht unbedingt eine Veränderung bei den Zielgruppen zur Folge.

Neben den Outputs sollten daher die Outcomes, ein breites Spektrum intendierter und auch nicht intendierter, ggf. unerwünschter, Wirkungen umfangreich erfasst werden. Je nachdem, welche Wirkungsebene in den Blick genommen werde, müssten passende Indikatoren gewählt werden.

> *„Die Untersuchung muss umfassend sein, in allen Aspekten der Änderung. Ich habe genannt: Kosteneinsparung, Änderung der Verwaltung, Einfluss auf die Betroffenen usw. Nicht die enge Vorstellung: Bei Pauschalierung geben wir jetzt 3 Millionen aus und vorher haben wir bei Einzelfallprüfung nur 2,5 Mill. ausgegeben. Das wäre eine unzulässige Verengung. Auch den Nutzen oder Schaden für die beteiligten Hilfeempfänger muss man messen. Wenn sie z.B. alle drei Monate zum Sozialamt gehen und dort einen halben Tag warten müssen, um einen Wintermantel oder anderes zu beantragen, dann entsteht Aufwand. Das muss in eine umfassende Untersuchung einbezogen werden.“ (§ A)*

Resultate können 1. nach der Zeitdimension, 2. nach der Dimension der Reichweite des Programms (Mikro-, Meso- oder Makroebene), 3. nach den beabsichtigten bzw. unbeabsichtigten Wirkungen (den so genannten Nebenwirkungen) und 4. nach den Wechselwirkungen mit anderen Programmen bzw. Rahmenbedingungen unterschieden werden.

Die Mehrzahl der Befragten problematisiert, dass Wirkungen je nach Anlage des Programms zu unterschiedlichen Zeitpunkten entstehen können. Demnach mache es einen großen Unterschied, ob die Resultate eines Programms ein halbes Jahr nach Programmende oder etwa nach zwei Jahren betrachtet werden.

Wirkungen von Programmen und Politik werden in ihrer **Zeitdimension** von zwei Parametern bestimmt: 1. vom Entstehungszeitpunkt der Wirkung und 2. von der Dauer der Wirkung. Deshalb ist es entscheidend, die zeitliche Wirkungsebene im Vorfeld festzulegen. Einerseits können Wirkungen an mehreren Zeitpunkten ent-

stehen z. B. unmittelbare Primärwirkungen mit eventuellen Folgewirkungen, andererseits können Programme oder Politik ausschließlich mittelfristig oder langfristig wirken. Der zweite Parameter der Zeitdimension, die Dauer der Wirkungen, wird auch als die Nachhaltigkeit von Wirkungen beschrieben.

Die **Reichweite eines Programms** wird von den Experten/-innen meist nach Mikro-, Meso- und Makroebene unterschieden. Weil ein Programm der Mikroebene zu Auswirkungen auf der Makroebene führen kann und umgekehrt häufiger Programme der Makroebene auf die Mikroebene wirken, sei es umso wichtiger die Zusammenhänge dieser unterschiedlich weit reichenden Wirkungsebenen zu analysieren.

> *„Mein Ansatz ist daher, die Wirkung von einer Ebene auf die andere Ebene zu analysieren, z.B. bei der Sozialhilfe. Also die Konsequenzen des Handelns einer Ebene für die anderen Ebenen zu berücksichtigen. Die Frage: Wer bewirkt was auf welcher Ebene?“ (§ A)*

Weiterhin wird angemerkt, dass bei **Modellprojekten auf Mikroebene**, wenn sie in der Regelpraxis verallgemeinert werden sollen, zusätzlich die weiteren Wirkungen auf der Makroebene mit bedacht werden müssen.

Einige Befragte thematisieren, dass Wirkungen **in beabsichtigte und unbeabsichtigte Wirkungen**, sog. Nebenwirkungen unterteilt werden können. Anhand der beabsichtigten Wirkungen könne das Programm auf Zielerreichung überprüft werden, indem man die Diskrepanz zwischen den Zielen und den Wirkungen eines Programms untersucht. Beabsichtigte Outcome-Wirkungen sind Feststellungen von Veränderungen und Stabilisierungen bei den Zielgruppen, zu denen formulierte Ziele vorliegen. Weitere beabsichtigte Wirkungen seien in Bezug auf Organisationen oder Sozialräume der Wirtschaftsregionen analysierbar. Nebenwirkungen sind Wirkungen, die außerhalb der formulierten Ziele des Programms liegen.

Nach Aussagen der Experten/-innen können Wechselwirkungen eines Programms sowohl mit anderen Programmen als auch mit sonstigen Faktoren bzw. Veränderungen von Rahmenbedingungen entstehen. Wenn mehrere Programme gleichzeitig durchgeführt werden, dann können die Wirkungen eines Programms ganz aufgehoben, abgeschwächt oder verstärkt werden oder es kann ein zusätzlicher Synergieeffekt auftreten.

3.5.3.2 Probleme der Wirkungsmessung

Die meisten Experten/-innen stellen einerseits die Probleme, andererseits die Wichtigkeit der Wirkungsmessung heraus. Zwei Hauptprobleme der Messung sind festzustellen:

Die Wirkungsfeststellung, auch Identifikationsproblem genannt. Die Aufgabe der **Wirkungsidentifizierung** wird oft im Zusammenhang mit dem – weiter gehenden - Wirkungsnachweis thematisiert. Das Problem, die gesamte Breite von Wirkungen einer Maßnahme zu erfassen entsteht durch die Vielzahl der verschiedenen Wirkungsarten, die bei einem Programm auftreten können, aber nicht müssen. Dementsprechend bestehen vier Messprobleme:

A. Zu welchem Zeitpunkt wird die Wirkungsmessung vorgenommen?

B. Auf welcher Ebene (Mikro, Meso und/ oder Makro) wird die Wirkung gemessen?

C. Welche Outcome-Stufen (z.B. Veränderungen im Wissen, Einstellungen oder Verhalten) bzw. über die Individuen hinaus gehende Wirkungen werden erfasst?

D. Wie bekommt man Informationen über Nebenwirkungen?

Als eine weitere Herausforderung wird der **Wirkungsnachweis,** auch Zuordnungsproblem genannt, von nahezu allen Befragten angesprochen. Das Problem des Wirkungsnachweises besteht darin, die festgestellte Wirkung auf den/die ursächlichen Faktor/-en zurückzuführen. Wenn mehrere Programme gleichzeitig ablaufen und sich evtl. die Rahmenbedingungen verändern, dann können verschiedene Wirkungen miteinander konfundieren.

> *„Bei der Ex-post-Evaluation sind alle mitwirkenden Faktoren historisch bestimmt. Da besteht das Hauptproblem daraus, als Ursache den der betrachteten Maßnahme zuzurechnenden Beitrag für das Ergebnis herauszufiltern. Wir wissen aber nie mit Sicherheit, wie groß der Beitrag anderer mitwirkender Ursachen ist. Dies ist das berühmte Zurechnungsproblem, das wegen der verschiedensten Scheinkorrelationen nur schwer lösbar ist.“ (§ A)*

Folgende Probleme des Wirkungsnachweises beschreiben viele Befragte: Welche Programme interagieren miteinander? Dies beinhaltet insbesondere die Frage, inwieweit mehrere Programme bei den gleichen Zielgruppen Veränderungen und Stabilisierungen bezüglich der formulierten Ziele bewirken. Gerade in der Armuts- und

Sozialpolitik sei die Wirkung häufig nicht auf ein einziges Instrument zurückzuführen, sondern in der Regel auf eine Kombination von vielen Faktoren, die ein bestimmtes Ergebnis produzieren.

Welche Randbedingungen interagieren mit dem Programm? Neben der oben beschriebenen Interaktion zwischen verschiedenen Programmen können auch veränderte Randbedingungen, die entweder zum Teil ebenfalls durch die Wirkung von Programmen entstehen oder durch eine gesellschaftliche Weiterentwicklung, wie z.B. die Demographie , auf die Resultate einwirken.

3.5.3.3 Auswege aus der Problematik

Von den Experten/-innen werden drei Auswege aus der Messproblematik aufgezeigt:

- Hinsichtlich des Problems der Festlegung des Zeitpunktes der Messung aus der Wirkungsidentifizierung wird von einem Experten vorgeschlagen, die Festlegung des Messzeitpunktes von der Irreversibilität einer Maßnahme abhängig zu machen: Wenn eine Maßnahme irreversibel sei, dann müsse man auch langfristige Messungen durchführen, ansonsten nicht.
- Ein Ausweg aus der Nachweisproblematik besteht nach Ansicht einiger Experten/-innen darin, im Vorfeld der Programmdurchführung Annahmen über die Wirkungsweise des Programms zu Grunde zu legen, so dass zum späteren Zeitpunkt die festgestellte Wirkung mit einer hohen Plausibilität, d.h. Wahrscheinlichkeit, ursächlich auf das Programm zurückgeführt werden kann. Annahmen über die Wirkungsweise könnten innerhalb des logischen Modells der Programmtheorie formuliert werden, welches die Programmdimensionen in einen sachlogischen Zusammenhang stellt. Dieser Zusammenhang sollte folgende Programmdimensionen umfassen: Die Bedingungen eines Programms, die Ausgangssituation der Zielgruppe; das Konzept und den Prozess, das heißt die Intervention, sowie das Resultat, bestehend aus Output, Outcome, intendierten Wirkungen und Nebenwirkungen. Wenn Evaluation diese Rückführung der Wirkung auf das Programm leisten soll, müsste Evaluation nicht nur die Resultate eines Programms in den Blick nehmen, sondern die oben beschriebenen anderen Dimensionen des Programms ebenfalls. Die Aufgabe des Wirkungsnachweises müsste im Zusammenhang mit den anderen Aufgaben der Evaluation erfolgen (vgl. 3.5.2).

- Ein weiterer Ausweg verläuft ähnlich: Wenn das Wirkungsnachweisproblem nicht lösbar wäre, könnte stattdessen in allen Phasen des Programms wirkungsorientiert auf eine Verbesserung der Zielerreichung hingewirkt werden.

3.5.4 Nutzenentstehung im Evaluationszyklus

Die Optimierung der Nutzenentstehung ist nicht erst nach Ablauf der empirischen Untersuchungen oder dar der Evaluation insgesamt vorzubereiten, sondern sollte bereits in den vorgelagerten Evaluationsphasen mitbedacht werden. Inwieweit die Nutzenentstehung an den verschiedenen Stellen im Evaluationszyklus unterstützt werden kann, wird im folgenden Kapitel dargestellt.

3.5.4.1 Einbezug von Beteiligten und Betroffenen

Mehrere Experten/-innen geben an, dass sich ein dialogischer bzw. interaktiver Evaluationsansatz als nützlich erwiesen hat. Durch den Einbezug der Beteiligten und Betroffenen in die verschiedenen Phasen des Evaluationsprozesses könne eine höhere Motivation, das Evaluationsvorhaben zu unterstützen und eine größere Akzeptanz gegenüber den Evaluationsergebnissen ausgelöst werden. Auch die Betroffenen von Maßnahmen sollten bei der Identifikation und Ausgestaltung von Veränderungen beteiligt werden. Auf diese Weise könne ein Zuwachs an selbstverantwortlichem Handeln und eine größere Akzeptanz gegenüber den Maßnahmen erreicht werden.

> *„Es hat sich häufig bewährt, dass man die Evaluationsinstrumente mit diesen Akteuren diskutiert, das heißt, dass die Instrumente intersubjektiv entwickelt werden, bevor sie zum Einsatz kommen. Das wäre eine Besonderheit von Evaluationen im gesellschaftlichen Bereich. Man könnte diese Herangehensweise interaktiv bzw. dialogisch nennen. Auf diese Weise wollen wir erreichen, dass die Akteure zu einer optimalen Mitwirkung motiviert werden.“ (§ A)*

3.5.4.2 Evaluationszwecke festlegen

Nach den Aussagen der Befragten sollen Evaluationsstudien neben dem reinen Erkenntnisgewinn wissenschaftlich abgesicherte Grundlagen für das politische Handeln bereitstellen, indem sie Informationen liefern über die Wirkung bestimmter wirtschaftlicher bzw. sozialpolitischer Instrumente. Durch die gewonnenen Informationen können die Konsequenzen der Politik besser abgeschätzt werden.

Zum einen könne die wirkungsorientierte Evaluation Informationen darüber liefern, ob und inwieweit die angesetzten Maßnahmen die angestrebten Ziele erreichen. Zum anderen könnten die Evaluationsergebnisse aber auch zur Verbesserung und passgenauen Ausgestaltung von Programmen verwendet werden. Einige Experten/-innen machen deutlich, dass Evaluationsergebnisse für die Planung und Steuerung von Programmen bzw. Interventionen nützlich sein können. Indem begleitend zur Einführung der Maßnahme Daten erhoben werden, könnten Hemmnisse und Fehlentwicklungen frühzeitig erfasst werden, so dass Korrekturen im Umsetzungsprozess eine bessere Zielerreichung ermöglichen.

Einzelne Experten/-innen weisen auf die Gefahr hin, dass eine Evaluation seitens der politischen Akteure nicht immer mit dem Zweck vergeben wird, umfassende Informationen zu erhalten, sondern um auf Ergebnisse zurückgreifen zu können, die eine Rechtfertigung der eigenen Politik erlauben. Möglicherweise stimmen die vorgegebenen Evaluationszwecke nicht mit den tatsächlichen Interessenlagen der Politikakteure überein. In diesem Fall sei die zweckgemäße Nutzung der Evaluationsergebnisse seitens der Auftraggeber/-innen begrenzt.

3.5.4.3 Zielklärung der Programme

Es besteht weitgehend Konsens seitens der teilnehmenden Experten/-innen, dass die Ziele der zu evaluierenden Maßnahmen transparent und differenziert formuliert sein sollen. Je kleinschrittiger die angestrebten Zielebenen expliziert würden, desto passgenauer könne auch das Evaluationsdesign zugeschnitten werden.

> *„Es wäre weiterhin wichtig, dass der Begriff Armut so konkretisiert wird, dass man sich auf ganz spezifische Zielgruppen, Fragestellungen bezieht und für die Zielgruppen ganz spezifische, überschaubare Zielformulierungen entwickeln kann. Je eingegrenzter die Zielgruppe, Fragestellungen und Zielformulierungen sind, um so besser kann man die Evaluation durchführen.“ (§ A)*

Es zeichnet sich eine Spannungslinie entlang der Frage ab, ob die als notwendig betrachtete Zielklärung *vor* Beginn der Evaluation abgeschlossen sein sollte oder als eine Teilaufgabe der Evaluation anzusehen ist. Einerseits wird gefordert, dass die Zielklärung im Rahmen einer politischen Debatte vorab geleistet wird. Andererseits erhoffen sich Experten/-innen durch die Klärung der Zielvorstellung der beteiligten Akteure im Rahmen des Evaluationsprozesses präzisere Zieldefinitionen. Die Eva-

luation solle so zu einem Austausch über die zu Grunde liegenden Ziele beitragen und damit das Politikhandeln transparenter machen.

3.5.4.4 Evaluationsgegenstände und -dimensionen

Die Befragung der Experten/-innen ergibt: Evaluationen beschreiben und bewerten Programme, Projekte, Initiativen und andere Maßnahmen der Politik. Sie können Modellprojekte begleiten, d.h. ihre Einführung unterstützen und erste Informationen über ihre Wirkungen bereitstellen. Weiterhin können politisch beabsichtigte oder bereits umgesetzte gesetzliche Veränderungen sowie einzelne Maßnahmen, die im Bereich der Politik zu Vermeidung und Verminderung von Armut durchgeführt werden, Gegenstand sein. Analog zu den einzelnen Phasen eines Programms kann sich die Evaluation stärker auf die Dimension der Rahmenbedingungen, des Konzepts, des Prozesses oder der Resultate beziehen (vgl. Kap. 1.4).

> *„Man könnte sich natürlich auch vorstellen, dass man, wenn man sich wirklich konkrete Projekte nimmt und dann noch mal den Gedanken aufgreift, ich will nicht nur das Ergebnis evaluieren, sondern ich will auch evaluieren, ob möglicherweise die Durchführung, der Prozess, dazu geführt hat, dass ich kein Ergebnis erzielen konnte.“ (§ B)*

Es wird betont, dass es wichtig sei, den Kontext der Programme im Blick zu haben, um mögliche Einflussfaktoren auf die angestrebten Zielgrößen zu identifizieren. Dies betonen besonders die befragten Auftraggebern/-innen, um gegenläufige Effekte des Programms plausibel erklären zu können. Möglicherweise könne der Einfluss von Randbedingungen auch dazu führen, dass sich die erzielten ‚positiven' und ‚negativen' Effekte gegenseitig aufheben. Darüber hinaus wirkten äußere Faktoren auf den Ablauf des Programms. Dies müsse daher bei der Prognose zukünftiger Resultate des Programms mitbedacht werden. Dieser Zusammenhang wurde am Beispiel der Bewertung der Rentenversicherung dargestellt:

> *„Wenn man die Entwicklung der Rentenversicherung prognostizieren will, müsste man zunächst prognostizieren, wie sich die Bevölkerung entwickelt. Auch wenn man sich nicht für die Bevölkerungsentwicklung interessiert, sondern z.B. nur für die finanzielle Stabilität der Rentenversicherung, so bilden derartige Prognosen doch die Voraussetzung für eine möglichst gute Vorhersage über die Zielvariable. Da taucht die Unsicherheit nicht beim Zurechnungsproblem auf, sondern bei der Prognose der Randfaktoren, d.h. des Datenkranzes.“ (§ A)*

3.5.4.5 Evaluationsansätze und Methoden

Die einbezogenen Experten/-innen beschreiben die **intendierten Programmziele** als Steuerungsmittel für eine Evaluation. Dabei solle anhand von Vorher-Nachher-Messungen überprüft werden, inwieweit die Resultate der Maßnahme den vorab festgelegten Zielen entsprechen. Ausgehend von den Zielsetzungen bzw. Fragestellungen werden die geeigneten Datenerhebungsinstrumente bestimmt. Im Vordergrund steht die Resultatsebene: der Grad der Zielerreichung. Um das Resultat eines Programms als Erfolg bzw. Misserfolg charakterisieren zu können, müssen Wertentscheidungen getroffen werden, wie die Veränderungsspannen zwischen Ausgangs- und Endzeitpunkt bewertet werden. Darüber hinaus wird es als sinnvoll angesehen, die Auswahl des Evaluationsgegenstandes ausgehend von den Zieldefinitionen vorzunehmen. Entsprechend dem Schaubild „Programmzyklus und Programmdimensionen am Beispiel sozialer Integration“ (Abb. 3 in diesem Bericht) kann die Evaluation verschiedene Elemente des Programms, z.B. Bedingungen, Prozesse, Resultate etc. fokussieren.

Das folgende Zitat verdeutlicht, dass sich die Resultate implementierter Maßnahmen negativ auf die Erreichung parallel existierender Leitziele auswirken können. Bei einer zielgeführten Evaluation sei es wesentlich zu überprüfen, inwieweit die eigenen Leitziele erreicht worden sind. Eine nähere Betrachtung der Auswirkungen auf Leitziele *anderer* Politikbereiche wird von einzelnen als nachrangig eingeschätzt.

> *„Wenn eine Evaluation über die Wirkung des Pfändungsfreigrenzenrechts in Auftrag gegeben wird, dann sagt mir der Evaluator, ein Nebeneffekt ist, dass die Leute sich jetzt größere Autos kaufen und die Leute so die Umweltverschmutzung verstärken. Dies ist im Rahmen einer gezielten Evaluation nicht das vorrangige Interesse des Auftraggebers.“ (§ B)*

Einige Experten/-innen sehen das **randomisierte Realexperiment** als idealen Ansatz für eine Wirkungsanalyse. Da Realexperimente in sozialen Anwendungsfeldern auf Grund des aufwändigen Designs und ethischer Bedenken häufig nicht durchführbar seien, wird ein **quasi-experimenteller Ansatz** befürwortet. In diesem Fall werden eine Experimentalgruppe – Gruppe, die an einer Maßnahme teilgenommen hat – und eine Kontrollgruppe – „statistischer Zwilling“ der Experimentalgruppe, die keiner Maßnahme unterzogen wurde – im Hinblick auf die Veränderung in bestimmten Interventionsdimensionen verglichen (vgl. ausführlicher

Kapitel 2.3.1.3). Diese methodische Herangehensweise bietet die Möglichkeit, den Einfluss von Randfaktoren, z.B. gesellschaftlicher Bedingungen, auf die gemessenen Effekte zu neutralisieren, so dass die festgestellten Ausprägungen in der Experimentalgruppe eindeutig auf die Teilnahme an den Maßnahmen zurückzuführen sind.

Es wird kritisch angemerkt, dass sich das experimentelle bzw. quasi-experimentelle Design nur bedingt für den Nachweis von kumulativen Wirkungen – wie sie im Bereich der Armuts- und Reichtumsberichterstattung auftreten – eignet. Darüber hinaus solle bedacht werden, dass die Ergebnisse Gültigkeit beanspruchen unter der Annahme der Konstanz der Verhaltensparameter. Bei diesem Verfahren würden die unerwünschten Wirkungen ausgeblendet. In der Realität spielten diese jedoch eine wesentliche Rolle. Demnach müssten bei einer Betrachtung von Armut als mehrdimensionalem Phänomen die Instrumente der Wirkungsforschung einer klassischen Evaluation, die an rein mathematisch statistischen Verfahren orientiert seien, durch qualitative Verfahren ergänzt werden.

Als ein Modell im Rahmen der ex-ante Evaluation wird die **Mikrosimulation** genannt. Um unterschiedliche Maßnahmen bereits *vor* ihrer realen Einführung auf ihre Auswirkungen – z.B. soziale Verträglichkeit, Finanzierbarkeit oder auf Verteilungswirkungen hin – zu untersuchen, werden zukünftige Entwicklungen auf der Basis vorhandener Datensätze in unterschiedlichen Szenarien simuliert. Die Mikrosimulation ist abhängig vom prognostizierten Zeithorizont und der damit einhergehenden Anzahl der veränderten Parameter unterschiedlich komplex angelegt. Es bestehe einerseits die Möglichkeit, die Veränderung eines Parameters – z.B. Erhöhung des Kindergeldes – mit dem Status quo zu vergleichen oder andererseits mit einer Alternativsimulation – z.B. Schaffung von Kindergartenplätzen. Zunächst werden auf diese Weise Primärwirkungen simuliert bei gleich bleibender Ausgangsituation. Darüber hinaus könnten Folgewirkungen eingeschätzt werden, wenn bestimmte Verhaltensannahmen im Zusammenhang mit der Veränderung eines Parameters unterstellt werden. Letztendlich basiere die Tragfähigkeit der Zukunftsszenarien damit auf der Gültigkeit der Annahmen, die in das Modell investiert wurden. Die unterstellten Verhaltensannahmen – z.B. Zunahme der Erwerbsarbeit von Frauen in Folge vermehrter Kindergartenplätze – sollten mit Hilfe von qualitativen Verfahren überprüft werden.

Als ein weiterer Evaluationsansatz wird die **Kosten–Nutzen–Analyse** erwähnt, wobei die monetären Kosten und der Nutzen eines Programms gegenübergestellt werden (vgl. ausführlicher Kapitel 2.3.1.4).

Die Experten/-innen wenden verschiedene **Datenerhebungs- und Auswertungsmethoden** an. Genannt werden u.a. Feldforschung, Literaturanalyse und die Auswertung von Datensätzen. Hierbei nutzen sie explorative, kinästhetische, quantitative und qualitative Verfahren. Die Entscheidung für die passende Methode bzw. das geeignete Design wird durch die Zwecksetzung, die Fragestellungen und den Kontext der Evaluation bestimmt. Der Vorteil von quantitativen Verfahren wird von einigen Forschern/-innen in der Möglichkeit gesehen, Daten zu gewichten und zu bewerten. Qualitativ ließen sich vor allem Zusammenhänge *zwischen* den Ebenen erfassen. Während einerseits überwiegend quantitative Ansätze als Methode der Wahl betrachtet werden, wird andererseits der verstärkte Einsatz von qualitativen Verfahren gefordert, um Wirkungszusammenhänge aufzuzeigen. Hier gibt es auch deutliche Differenzen, da die Befürworter quantitativer Methoden den qualitativen Verkürzungen unterstellen und umgekehrt. Die Mehrheit der Forscher/-innen spricht sich jedoch für eine Kombination unterschiedlicher Ansätze aus und fordert, eine Vielfalt von Erhebungsmethoden zuzulassen.

> *„Generell würde ich dafür plädieren, nach Möglichkeit immer mehrere Verfahren oder Methoden zu verwenden bzw. verschiedene Modell-Annahmen zu Grunde zu legen, und dann zu untersuchen, ob vergleichbare Ergebnisse rauskommen. Ein Methodenmix hat sich in meinen eigenen Forschungen immer bewährt.“ (§ A)*

Darüber hinaus sollte die Wirkung von Einflussfaktoren auf das Ergebnis kontrolliert werden und der Unschärfebereich der quantitativen Messungen angegeben werden. Der Unsicherheitsspielraum der Ergebnisse sollte auf Seiten der Politik jedoch nicht zur Handlungsunfähigkeit führen. Es sei wichtig, um Unschärfen zu wissen, aber häufig ausreichend, dass die Makrodaten Tendenzen wiedergeben.

An dieser Stelle zeigt sich folgendes Dilemma: Zum einen wird erklärt, dass die methodischen Ansprüche an die Studien erhöht werden sollten, um eine bessere Überzeugungskraft gegenüber den Entscheidungsträgern/-innen aufzuweisen. Zum anderen sind komplexe methodische Verfahren für die Auftraggeber/-innen kaum nachzuvollziehen, was die Akzeptanz der Ergebnisse erschwert.

3.5.4.6 Ergebnisdarstellung

Ein Teil der Experten/-innen schlägt vor, die Ergebnisinterpretation zusammen mit den Beteiligten vorzunehmen. Diese interaktive Vorgehensweise könne so schon vor der Veröffentlichung der Studie zu einer Ergänzung und Absicherung der Ergebnisse beitragen. Durch die Kenntnis lokaler Besonderheiten seien vor allem die Feldexperten/-innen in der Lage, unerwartete Effekte zu erklären bzw. die vorgeschlagenen Interpretationsmuster zu bestätigen oder abzulehnen. Demnach sei es die Hauptaufgabe der Evaluatoren/-innen, die Ergebnisse zusammenzufassen und Schlussfolgerungen zu formulieren. Die derart aufbereiteten Ergebnisse sollten vorgestellt und anschließend durch wichtige Beteiligte aus verschiedenen Perspektiven betrachtet und schließlich in einer Endfassung festgehalten werden.

Der Evaluationsbericht sollte sich nach Auffassung vieler Antwortender nicht auf die Ergebnisdarstellung beschränken, sondern weiterführend Empfehlungen und Hinweise zur Verbesserung und Veränderung des Programms enthalten, welche den Adressaten/-innen der Studie Handlungsoptionen eröffnen. Einige Auftraggeber/-innen fordern, dass die Empfehlungen politikberatend formuliert werden und in ihrer Form konkret und feldspezifisch ausgestaltet sind. Weiterhin sollten die Konsequenzen der vorgeschlagenen Handlungsmöglichkeiten für die politischen Akteure/-innen abschätzbar sein und den politischen Prozess berücksichtigen. Das Aufzeigen von hemmenden bzw. fördernden Faktoren für die Zielerreichung sowie möglicher unbeabsichtigter Nebenwirkungen biete einen Ansatzpunkt für ein systematisch begründetes politisches Handeln. Darüber hinaus sollten bei der Ergebnisdarstellung normative Implikationen und die eingesetzten methodischen Verfahren mit den zu Grunde liegenden Annahmen offen gelegt werden.

> *„Außerdem muss die Auswahl der verwendeten Methoden transparent sein. Im Prinzip müssten auch ein komplexes Simulationsmodell und dessen Datenbasis veröffentlicht werden, denn ein Teil der Politikbeeinflussung, geschieht dadurch, dass bestimmte Alternativen gerechnet werden und bestimmte nicht. Die Palette der Möglichkeiten ist immer viel größer als das, was bei begrenztem Aufwand gerechnet werden kann.“ (§ A)*

Es wird als wichtig erachtet, dass eine anwendungsorientierte Berichtsform gewählt wird, welche die Nachvollziehbarkeit der Ergebnisse für die Adressaten/-innen ermöglicht. Zu diesem Zweck wird eine kurze Zusammenfassung mit den wichtigsten

Ergebnissen und den daraus abgeleiteten Handlungsempfehlungen als ein unverzichtbarer Berichtsteil angesehen.

> *„Transferfähigkeit der Projektergebnisse ist wichtig. Es muss eine Summary geben, in dem auf 20 Seiten die wichtigsten Ergebnisse dargestellt werden und was die daraus resultierenden Politikberatungsempfehlungen sind.“ (§ A)*

3.5.4.7 Ergebnisverwendungsprozess

Um eine größtmögliche Nutzung der Evaluationsergebnisse zu erreichen, sollten die Evaluationsstudien nach Auffassung von Experten/-innen veröffentlicht werden. Einige fordern, dass auch das verwendete Datenmaterial anderen Wissenschaftlern/-innen zur Verfügung gestellt wird, um eine intersubjektive Überprüfbarkeit der Ergebnisse zu ermöglichen.

Einer unvermittelten und beliebigen Nutzung von Evaluationsergebnissen stehen einige Experten/-innen ablehnend gegenüber. Sie wünschen sich einen vorgeschalteten Austausch zwischen Wissenschaft, Politik und Fachöffentlichkeit, um verschiedenen Risiken vorzubeugen. Ansonsten bestehe die Gefahr, dass sich politische Akteure/-innen je nach Interessenlage passende Evaluationsstudien auswählen und die Ergebnisse womöglich selektiv für die Verfolgung ihrer eigenen politischen Absichten nutzen. Auf diese Weise würden andere Evaluationsstudien systematisch ausgeblendet. Vor allem die Tatsache, dass wirkungsorientierte Evaluationen im selben thematischen Bereich zu gegensätzlichen Ergebnissen kommen, mache einen Diskurs erforderlich. Viele Experten/-innen befürworten daher eine kritische Bewertung von Ergebnissen, bevor eine entsprechende politische Schlussfolgerung gezogen wird. In der Diskussion solle eine Auseinandersetzung über die Angemessenheit der methodischen Verfahren erfolgen und eine damit einhergehende kritische Überprüfung der Ergebnisse der unvermittelten Nutzung vorangestellt werden. In den Evaluationsstudien würden vorrangig Teilsegmente der vorhandenen Informationen vermittelt werden. Insofern sei es wichtig, die Ergebnisse in eine Fachdiskussion einzubringen und sie dann dadurch zu ergänzen oder zu überprüfen.

Um eine zweckgemäße Nutzung der Evaluationsergebnisse sicher zu stellen, wird vorgeschlagen, die Evaluatoren/-innen in die Begleitung des politischen Umsetzungsprozesses mit einzubeziehen. Es wird befürwortet, die aus den gezogenen

Schlussfolgerungen abgeleiteten Maßnahmen wiederum bei der Implementation wissenschaftlich zu begleiten. Auf diese Weise könne die Praxisrelevanz sichergestellt werden.

Das Verhältnis zwischen Politik und Evaluation wird als ein ambivalentes beschrieben. Dies zeige sich daran, dass die politischen Akteure/-innen einerseits an den Wirkungen der eingesetzten Instrumente interessiert sind, andererseits aber die Konfrontation mit unerwünschten Evaluationsergebnissen fürchteten.

Einige einbezogene Experten/-innen wünschen sich, dass sozialpolitische Schlussfolgerungen zukünftig auf der Basis von wirkungsorientierten Evaluationsstudien gezogen werden. Zu diesem Zweck sollte bei der Auswahl von Maßnahmen darauf geachtet werden, dass diese evaluierbar seien bzw. bei der Implementation der Programme die Überprüfbarkeit mitgedacht wird.

> *„Das ist genau die Aufgabe der Wirkungsforschung, dass die Aussagen zu den sozialpolitischen Schlussfolgerungen künftig begründet werden müssen, auf der Grundlage der Bewertung des Einsatzes der Instrumente.“ (§ A)*

> *„Ich denke, dass es deutlich wird, dass man sich, wenn man es ernst nimmt und den Armuts- und Reichtumsbericht und seine Maßnahmen evaluieren will, dass schon bei der Auswahl der Maßnahmen und Projekte, die man dort hineinschreibt, einfach klar sein muss: Ist es evaluierbar? Was ist notwendig, damit es evaluierbar ist?“ (§ B)*

Seitens der Evaluatoren/-innen sollte eine kritische Einschätzung erfolgen, inwieweit eine Beantwortung geforderter Fragestellungen möglich ist – auch unter der Berücksichtigung der Datenlage. Eine vorherige kritische Überprüfung der Evaluierbarkeit und eine Prognose über die Güte der Ergebnisse machten die Chancen und Grenzen einer Evaluation transparent. Trotz bestehender Begrenzungen, denen wirkungsorientierte Evaluationen unterworfen seien , bestehe die Möglichkeit, systematisch gewonnene Informationen bereitzustellen, die ein transparentes und zielorientiertes Politikhandeln fördern. Auch wenn nicht immer ganz präzise Aussagen gemacht werden könnten, entstehe doch ein öffentlicher Diskurs, der dazu beitragen könnte, Politik bedarfsgerechter zu gestalten. In diesem Rahmen sollte seitens aller Beteiligten Verständnis für bestehende Restriktionen aufgebracht werden. Das gelte sowohl für die Forscher/-innen, die sich mit bestimmten methodischen

Problemen auseinander setzen müssten als auch für Politiker/-innen, welche ihrer Partei und den Wählern/-innen verpflichtet sind.

Als Adressaten von Evaluationsstudien werden die unmittelbaren Auftraggeber/-innen genannt, die Fach- bzw. allgemeine Öffentlichkeit, die Evaluatoren/-innen und die Betroffenen, die Zielgruppen von Maßnahmen sind.

3.6 Schlussfolgerungen

Die Experten/-innen begrüßen die in Gang gesetzte Debatte über die Evaluation von Programmen im Bereich der Armuts- und Reichtumsberichtserstattung sehr. Sie sehen sich an einem Entwicklungsprozess beteiligt, der transparentes, systematisch vorbereitetes und begleitetes Politikhandeln zum Ziel hat. Sowohl die Auftraggeber/-innen als auch die Forscher/-innen verbinden mit Evaluationen die Erwartung, mittels wissenschaftlich fundierter Informationen die Konsequenzen der Politik besser abschätzen zu können und damit die Ziele der Armutsvermeidung und –überwindung besser zu erreichen. Zu diesem Zweck sollte bei der Konzeption von Maßnahmen darauf geachtet werden, dass diese evaluierbar sind bzw. es sollte bei der Implementation der Programme ihre Überprüfbarkeit mitgedacht werden.

In den Antworten der Experten/-innen finden sich – mit zum Teil unterschiedlichen Schwerpunktsetzungen – die im Kapitel 1.2 beschriebenen Standards für Evaluationen zumindest implizit als Qualitätsanforderungen an Evaluationen wieder. Es wird deutlich, dass viele Experten/-innen ein partizipatives und pluralistisches Evaluationsverständnis unterstützen. Dies zeigt sich vor allem an ihrem Wunsch nach einer vorbereiteten und regelmäßigen Erörterung zwischen Politik, Wissenschaft und Öffentlichkeit über die gewonnenen Evaluationsergebnisse. Als unbedingte Voraussetzung für den fachlichen Austausch wird die Veröffentlichung der Evaluationsstudien angesehen. Der Einbezug von Beteiligten – vorrangig bei der Ergebnisinterpretation, aber auch in anderen Evaluationsphasen – wird von vielen Forschern/-innen als nutzensteigernd beschrieben. Sie wünschen sich eine Diskussion über die Nützlichkeit der Studien für die Beschreibung und Beurteilung von wirtschaftlichen und sozialpolitischen Programmen auf der Basis aller zur Verfügung stehender Evaluationsstudien – gerade auch vor dem Hintergrund divergierender Evaluationsergebnisse (Evaluationssynthesen/Meta-Analysen).

Beide Gruppen – Auftraggeber/-innen und Forscher/-innen – betonen, dass die Perspektiven und Annahmen, auf denen die Anlage der jeweiligen Evaluation und die Interpretation der Ergebnisse beruhen, offen gelegt werden sollen. Bezüglich der Frage, inwieweit Werte und deren Klärung als Bestandteil bzw. Aufgabe der Evaluation angesehen werden oder ob die Verantwortung dafür außerhalb des Evaluationsprozesses verortet wird, gibt es auf Seiten der Experten/-innen unterschiedliche Auffassungen. Einige sehen in der Klärung der teilweise impliziten

Wertpositionen die Chance, eine kritische Erörterung und Abwägung der meist wertgebundenen Ziele der beteiligten Akteure/-innen zu fördern. Vor allem die Auftraggeber/-innen sprechen die Problematik an, dass auf Grund unklarer und zum Teil gegenläufiger Ziele in den Politikbereichen der Armuts- und Reichtumsberichterstattung die Bewertung der Programmresultate als Erfolg oder Misserfolg erschwert wird bzw. es zu konkurrierenden und damit unbefriedigenden Interpretationen kommt. Dem vorzubeugen eignen sich aus unserer Sicht besonders werteberücksichtigende Evaluationsmodelle.

Es findet sich auch die gegenteilige Position, wonach die Festlegung der Werte vorab in Form einer politischen Auseinandersetzung erfolgen soll, welche dem Evaluationsprozess vorangestellt wird. Dieses Evaluationsverständnis beruht auf einem „wertedistanzierten“ Modell, wonach Wertfragen z.B. durch die Verfahren der parlamentarischen Demokratie vorentschieden werden oder zumindest als außerhalb von Evaluationen zu entscheiden gesetzt werden. Es wird allerdings seitens der Forscher/-innen darauf hingewiesen, dass die von den Programminitiatoren/-innen vertretenen Werte teilweise nicht hinreichend konkret vorliegen, so dass Forscher/-innen diese zum Teil durch eigene Wertpositionen ergänzen. Eine werteklärende Vorgehensweise scheint bei schwacher Werteexplikation zu evaluierender Programme daher vielfach notwendig.

Das im Kap. 1.2 beschriebene Spannungsverhältnis zwischen Nützlichkeits- und Genauigkeitsstandards, welches sich aus wissenschaftlichen Gütekriterien einerseits und den Anforderungen der Evaluationsnutzer/-innen andererseits ergibt, wird von den Forschern/-innen als Dilemma thematisiert. Demnach haben elaborierte methodische Verfahren mit der Bereitstellung von möglichst genauen Ergebnissen den Vorteil, dass die Aussagen der Studien durch die Auftraggeber/-innen ernster genommen werden. Die Komplexität der eingesetzten Verfahren erschwert andererseits die Nachvollziehbarkeit der Ergebnisse, so dass es zu einem Akzeptanzproblem und damit auch Nutzungsproblem kommen kann. Eine Einigung zwischen Evaluatoren/-innen und Auftraggebern/-innen muss innerhalb dieses Spannungsfelds immer wieder neu ausgehandelt werden.

Was die einzusetzenden Datenerhebungs- und Auswertungsmethoden betrifft, befürworten die Experten/-innen überwiegend eine Vielfalt an methodischen Zugängen. Teilweise werden entweder quantitative oder qualitative Verfahren vorgezogen.

Einen großen Stellenwert in der Diskussion nimmt die Bedeutung der Wirkungsmessung ein. In diesem Zusammenhang thematisieren die Experten/-innen das Problem der Wirkungsidentifizierung und des Wirkungsnachweises.

Die Schwierigkeit, Wirkungen zu identifizieren, entsteht durch die Vielzahl verschiedener Wirkungsarten (Nebenwirkungen, unterschiedliche Zeitpunkte) sowie Reichweiten von Wirkung, die bei einem Programm auftreten können. Das Problem des Wirkungsnachweises besteht darin, die festgestellte Wirkung auf das Programm als ursächlichen Faktor zurückzuführen oder aber eine Aussage dazu zu machen, einen wie großen Anteil der Verursachung das Programm gegenüber weiteren Faktoren hat.

Ein Ausweg besteht nach Ansicht einiger Experten/-innen darin, im Vorfeld der Programmdurchführung spezifizierte Annahmen über die Wirkungsweise des Programms zu formulieren, so dass zu einem späteren Zeitpunkt die festgestellte Wirkung mit einer hohen Plausibilität auf das Programm zurückgeführt werden kann. Eventuell kann dies auch kombiniert werden mit experimentellen bzw. quasi-experimentellen Designs. Demnach sollten nicht ausschließlich die Resultate eines Programms in den Blick genommen werden sondern ebenfalls die übrigen Dimensionen eines Programms – Bedingungen, Konzept und Prozess.

Darüber hinaus besteht die Möglichkeit, in *allen* Phasen des Programms wirkungsorientiert auf eine Verbesserung der Zielerreichung hinzuarbeiten. Eine Erweiterung des Aufgabenfelds von Evaluation – von der alleinigen Betrachtung der Resultate zur Beschreibung und Bewertung weiterer Programmdimensionen – ermöglicht zudem, dass bereits die Entwicklung und Steuerung von politischen Maßnahmen systematischer und zielgeführter erfolgt. Dies wird von den Experten/-innen als eine wichtige, bislang noch nicht genügend genutzte Funktion von Evaluation geführt.

4 Datenlage für Evaluationen

Die Qualität einer Evaluation ist deutlich von der Qualität der verwendeten Daten abhängig (s. dazu die Genauigkeits-Standards der DeGEval 2001). Zwei Hauptarten von Daten können – nach Zeitpunkt ihres Entstehens – unterschieden werden: erstens Daten, *die bereits vorliegen*, welche im Rahmen der Evaluation also zugänglich gemacht, aufbereitet, ggf. geprüft, ausgewertet und damit zur Beantwortung von Fragestellungen genutzt werden; zweitens Daten, die nach Festlegung der Evaluationsfragestellungen eigens erhoben werden, da keine ausreichende Datenbasis vom ersten Typ mit der erforderlichen Spezifität vorliegt. Auf folgende Datentypen wird in diesem Kapitel eingegangen: „administrative oder prozessgenerierte Daten", „statistische Daten" und „Panel-Erhebungsdaten".

Zunächst werden die bestehende Datenlage und Dateninfrastruktur für Evaluationen bzw. daraus resultierende Stärken und Schwächen dargestellt und Möglichkeiten für zukünftige Strategien zur Verbesserung der Datenlage aufgezeigt. Es gibt Überschneidungen mit den allgemeinen Problemen bzgl. der Datenlage für die Armuts- und Reichtumsberichterstattung.[55] Der Fokus liegt auf den für Evaluationen spezifischen Daten-Problemen im Kontext der Armuts- und Reichtums-berichterstattung.

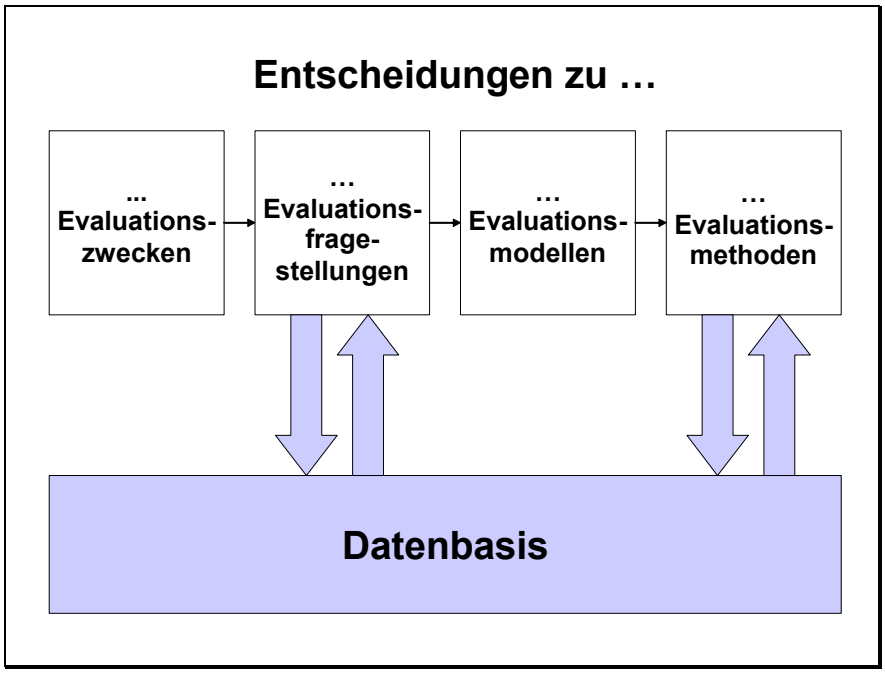
Die dargestellten Ergebnisse basieren einerseits auf Literaturanalysen, andererseits auf Ansichten und Meinungen von Experten/-innen, die befragt wurden. Deren Aussagen sind anonymisiert in die folgende Darstellung eingeflossen, die auch Zitate enthält.[56] Es wird zunächst auf den Zusammenhang zwischen Evaluation und Datenlage eingegangen, anschließend werden die unterschiedlichen Datentypen einzeln besprochen.

---

55 Für eine Übersicht über die Datenlage für die Armuts- und Reichtumsberichterstattung vgl. Müllenmeister-Faust (2002), S. 182ff.

56 Die für zum Thema Datenlage befragten Experten/-innen werden hier mit „§C" bezeichnet. Weitere Angaben unterbleiben zur Sicherung der Anonymität. Die Interviewten haben vor ihrem jeweiligen Erfahrungshintergrund unter den Datentypen Schwerpunkte gesetzt.

Abbildung 13: Datenbasis für Evaluationen und Evaluationsdesign



Quelle: eigene Darstellung

4.1 Datenlage und Evaluationstheorie

4.1.1 Monitoring und Evaluation

Monitoring und Evaluation stehen in engem Zusammenhang. Bei Fehlen einer verbindlichen Definition wird unter Monitoring allgemein die *laufende* Erfassung und Dokumentation von Daten sowie die Berichterstattung darüber verstanden. In einem festgelegten zeitlichen Rhythmus werden vorab festgelegte Daten erhoben und dokumentiert. Dieser Datenkranz muss im Ablauf stabil bleiben, um die angezielten Verlaufsmuster darstellen zu können. Monitoring ist vielfach fester Bestandteil von Verwaltungsaufgaben. Dabei entstehen administrative Daten, die wichtige Grundlage für Evaluationen sein können. Monitoring im Kontext der Armutsvermeidung und -verminderung erfolgt oft durch die programmtragenden Organisationen.[57]

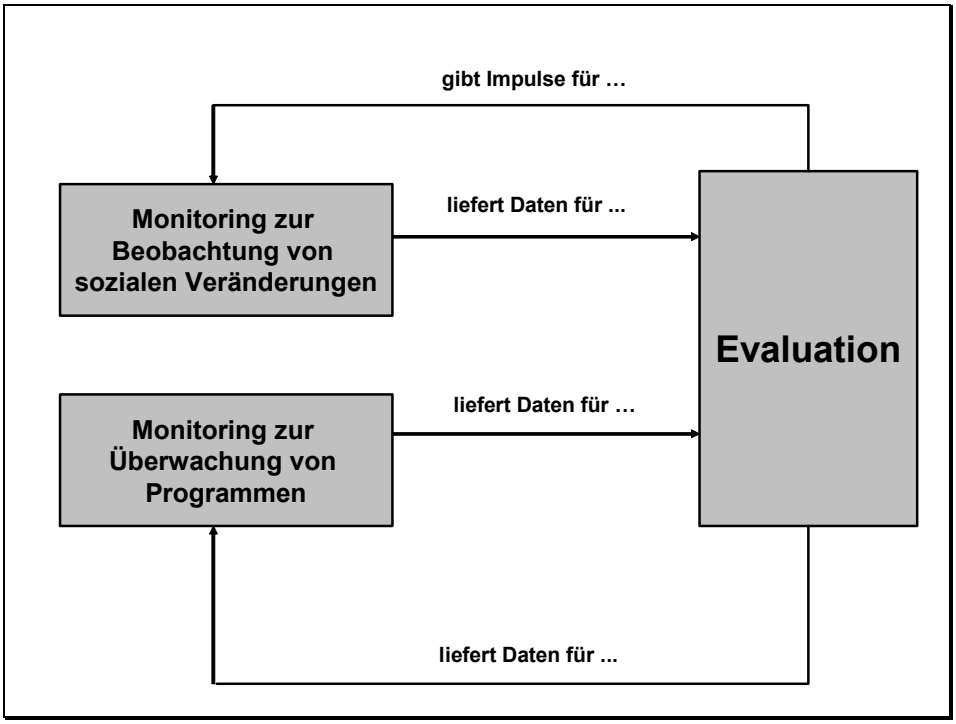
Typische Gegenstände des Monitoring sind Finanzdaten (z.B. Kosten und Umsätze), Output-Kennzahlen (z.B. Anzahl Teilnehmender – differenziert nach soziodemographischen Merkmalen - oder Ausgaben pro Programm-Teilnehmer /-in sowie leicht

57 Die Einbeziehung der Träger in den Evaluationsprozess ist u. a. auch deshalb wichtig, da vor Ort regelmäßig Daten zu den Programmen anfallen. Ausschlaggebend ist, das Monitoring im Vorhinein bezüglich Inhalten wie Verfahrensweise genau zwischen Trägern und Finanziers abzustimmen. Als Beispiel aus der Evaluierung Kommunaler Vermittlungsagenturen vgl. den Abschlussbericht von Eisentraut/Wagner (2002), S. 16.

messbare Outcomes, z.B. Quoten kurzfristiger Wiedereingliederung. Es können zwei Arten des Monitoring unterschieden werden:

Erstens kann Monitoring allein stehend soziale Veränderungen beobachten. Hier werden vor allem Bedürfnisse und Bedarfe erhoben, auf welche die Politik mit konkreten Maßnahmen reagiert. So kann Monitoring feststellen, wie sich die Lebensqualität von Personengruppen im Niedrigeinkommenssektor entwickelt. Durch Monitoring können soziale Zustände und Problemfelder im Zeitverlauf abgebildet werden. Diese Art des Monitoring kann unabhängig von Evaluationen erfolgen.

Abbildung 14: Monitoring und Evaluation



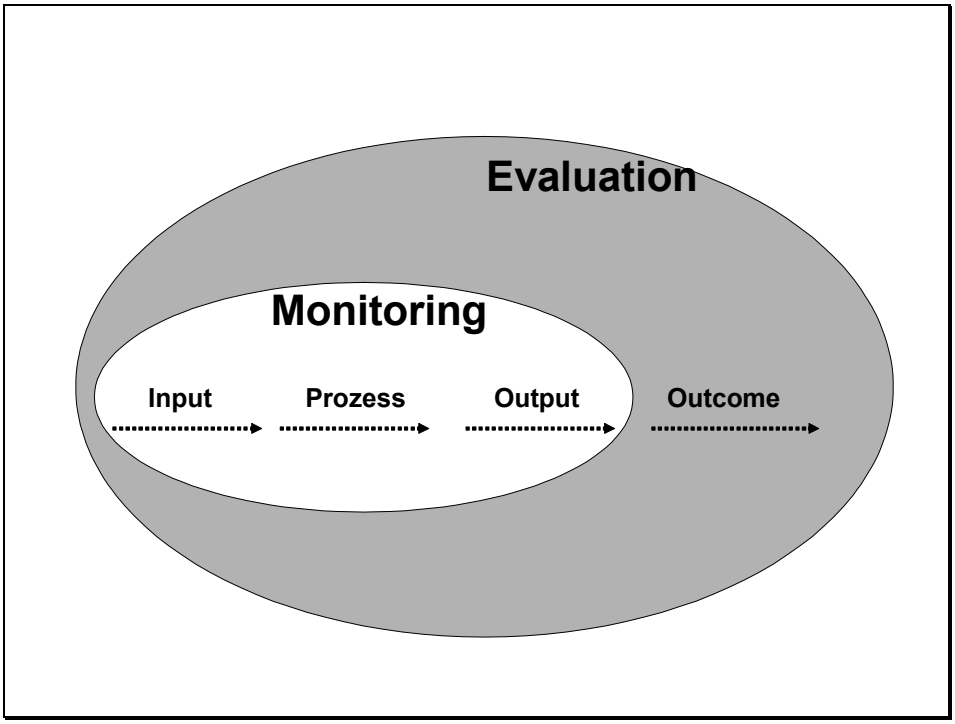
Quelle: eigene Darstellung

Zweitens kann Monitoring Teilaufgabe von Evaluationen sein (oder eine von fünf Funktionen; vgl. Kap. 1.5). Es kann insbesondere Input, Prozessmerkmale sowie Output eines Programms erfassen. Vielfach stehen Programmversorgung und Bedarfsdeckung im Mittelpunkt. Dieses Monitoring kann zur Erfolgskontrolle durch Evaluation für Programmmanager beitragen, reicht hierfür jedoch allein nicht aus.[58]

58 Leistungssteuerung allein über Inputs, Outputs und kurzfristige Outcomes – dies sind die Erfassungsstärken des Monitoring - kann kontraproduktiv sein: Mittel- und längerfristige Outcomes und damit die eigentlichen Programmziele werden dann wegen des Zwanges zu kurzfristigen ‚Monitoring-Erfolgen' in Verfolgung wie Messung vernachlässigt.

Im Evaluationsdesign wird festgelegt, zu welchen Zeitpunkten welcher Datenkranz zyklisch erhoben werden soll. Die Entwicklung des Monitoring ist damit Bestandteil der Evaluationsplanung. Eine Evaluation und deren Ergebnisse verändern vielfach wiederum eine gewählte Monitoring-Strategie. Evaluation und Monitoring sollten idealerweise zirkulär verknüpft sein (vgl. Auer/Kruppe 1996, S. 908).

Abbildung 15: Monitoring als Bestandteil von Evaluationen



Quelle: eigene Darstellung

Im Rahmen einer Politik der Armutsvermeidung und –minderung genügt ein allein stehendes Monitoring nicht: Neben der Vernachlässigung relevanter Outcomes und seiner geringen Flexibilität, auf Verschiebungen in Programmen und deren Kontext durch Designanpassung zu reagieren, kann Monitoring keine Nebenwirkungen erfassen, insbesondere nicht solche, die unerwartet *und* unerwünscht sind. In aller Regel ist zur Erstellung eines gültigen und nützlichen Monitoring-Systems erhebliche evaluatorische Vorarbeit zu leisten. Diese führt je nach „Werteberücksichtigung“ im gewählten Evaluationsmodell (vgl. Kap. 2.2) zu unterschiedlichen Monitoring-Systemen und Indikatorenauswahlen.

4.1.2 Verwendung bestehender Indikatorensysteme

Der Datenbedarf in der Evaluation bestimmt sich gemäß der jeweiligen, die Untersuchung leitenden Evaluationsfragestellungen. Diese durch die Fragestellungen angesprochenen Sachverhalte sollten möglichst exakt und eindeutig beschrieben werden. Hierfür kann teilweise auf Operationalisierungen in Form bestehender (sozialer) Indikatoren zurückgegriffen werden. Dies muss in jedem einzelnen Falle geprüft werden. So könnte zum Beispiel eine rigide Verpflichtung von Evaluationen auf die Nutzung von Indikatorensystemen dazu führen, dass sich die beantworteten Fragestellungen nach den Indikatorensystemen richten und nicht umgekehrt: Nur die Daten aus bestehenden Indikatorensystemen werden im Rahmen von Evaluationen genutzt, die unmittelbar und passgenau zur Beantwortung fest gelegter, relevanter Fragestellungen beitragen. Diese Prüfung der Gültigkeit von genutzten Kennzahlen oder Indikatoren für den zu beschreibenden Sachverhalt hat genauso intensiv zu erfolgen, wie dies auch bei einer Neukonzeption von Datenerhebungsinstrumenten erforderlich ist.

Durch Monitoring können wiederum Daten für den Gebrauch als Indikatoren generiert werden.

> *„Will ich z.B. die Wirkung auf die Lebenslage der Betroffenen untersuchen oder will ich die Wirkung auf der Kostenebene untersuchen oder will ich die Wirkung für das Verwaltungshandeln untersuchen oder geht es eher um makro-ökonomische Effekte? Für jede Wirkungsebene müssen geeignete Indikatoren ausgewählt werden.“ (§ A).*

So gibt es für jeden Teilschritt im Politik- bzw. Programmzyklus entsprechend unterschiedliche Indikatoren:[59] Kontext-, Input-, Income-, Struktur-, Prozess-, Output-, Outcome- und Nebenwirkungsindikatoren. Brauchbare Indikatoren zu formulieren ist insbesondere bei diffusen Zielen schwierig. Wenn also Programme ohne klare und spezifizierte Ziele in ihre Umsetzungsphase gehen, ist eine wirkungsorientierte Evaluation sehr aufwändig, wenn nicht unmöglich. Aus diesem Grund besteht gegenwärtig ein starker Trend, Evaluation von Beginn an – mit der Funktion Klärung und Interaktion – in den Programmzyklus einzubauen. In der Regel müssen für jedes Pro-

59 Siehe Abbildung 3 „Programmzyklus und Programmdimensionen“.

gramm-Ziel ein oder mehrere Outcome-Indikatoren formuliert werden. Dies wird umso schwieriger, desto länger die betrachtete Wirkungskette eines Programms ist.[60]

In Deutschland gibt es zurzeit kein festgelegtes Indikatoren-Set in Bezug auf die Armuts- und Reichtumsberichterstattung.[61] „Für die Armutserfassung und -erforschung gibt es nur ein beschränktes Indikatorentableau zur Beschreibung und Analyse von Armutsentwicklungen, Armutspopulationen und Armutsregionen."
(S. VSOP, 1995, S. 121). Es werden die verschiedenen Lebenslagen betrachtet, in deren Zusammenhang gängige Indikatoren gebraucht werden, diese wurden jedoch nicht grundsätzlich im Rahmen der Armuts- und Reichtumsberichterstattung festgelegt.[62]

> *„Indikatoren sind natürlich die entscheidende Grundlage für die Bewertung der Zielerreichung in der Armutspolitik, wie auch in anderen sozialpolitischen Bereichen der Politik. Es kann nicht sein, dass wir uns auf eine oder wenige Indikatoren beziehen, wir brauchen ein breites Set von Indikatoren. Auf der europäischen Ebene gibt es qualitative und quantitative Indikatoren. Wir haben ein Set von Primär- und Sekundärindikatoren, die auf europäischer Ebene einheitlich sind. Wir haben die Möglichkeit in den Mitgliedsstaaten quasi auf ein tertiäres Bündel von Indikatoren, die selbst von Land zu Land unterschiedlich sind, zuzugreifen. Der erste Prozess der Standardisierung hat schon stattgefunden. Von daher müssen wir nicht von vorne anfangen. Der Prozess läuft schon seit längerer Zeit. Aber dieser Prozess erfordert auch – sei es auf der EU- oder nationalen Ebene – eine kritische Überprüfung und gegebenenfalls eine Weiterentwicklung, z.B. bei den quantitativen Indikatoren, die sich auf Bereiche wie Einkommen, Versorgung auf dem Arbeitsmarkt, Versorgung auf dem Wohnungssektor, auf einige wenige Bereiche beziehen. Als erste Orientierung sind sie sicherlich hilfreich, aber sie reichen nicht aus, wenn es um differenzierte Fragestellungen geht. Sie reichen nicht aus, darüber ausreichende Antworten zu geben. Indikatoren sind sehr wichtig, aber wir sind zur Zeit mittendrin in dem Prozess der Bestimmung der Indikatoren und der Definition von Indikatoren, die empirisch gehaltvoll sind.“ (§ A)*

---

60 Siehe Kap. 3.5.3 Das Konzept „Wirkung

61 Es kann natürlich auch auf andere bereits existierende oder in der Diskussion befindliche Indikatorensysteme zurückgegriffen werden, wie solche der OECD, der EU, ILO.

62 Für eine Diskussion von verschiedenen Sozial-Indikatoren in Europa vgl. Atkinson (2002); im Kontext der NAP incl. auch European Commission (2002), S. 88ff.

Hier wurden die Indikatoren des *National Action Plan Against Poverty and Social Inclusion* (NAPIncl), der auf EU-Ebene koordiniert wird, angesprochen. Eine Abstimmung von Indikatoren und Berichtssystemen zwischen Armuts- und Reichtumsbericht und NAPIncl ist als erstrebenswert anzusehen.[63]

> *„Ein Teil der Gesundheitsindikatoren [auf Bundesebene] ist auch nach sozialer Lage aufgeschlüsselt, d.h. speziell ... das Verhalten zur Situation als auch zum Inanspruchnahmeverhalten. Sie werden nach Möglichkeit auch immer nach sozialer Schicht gegliedert und ausgewertet und langfristig beobachtet. In diesem Kontext wird auch ein Konzept für den Indikatorensatz für den Bezug Armut, Gesundheit oder Reichtum/Gesundheit entwickelt, der dauerhaft mit Daten gespeist werden kann." (§ B)*

Die Indikatoren im Gesundheitsbereich sind zu einem großen Teil Selbsteinschätzungen (§ A).

In der Diskussion um feststehende Indikatoren für die Armuts- und Reichtumsberichterstattung sind auch die unterschiedlichen Armutsdefinitionen sowie -konzepte entscheidend. Die Auswahl von Indikatoren im Rahmen von Evaluationen sollte hiervon losgelöst stattfinden können, da diese sich an den Programmzielen orientieren.[64] Die Auswahl und Definition von Indikatoren ist ein wichtiger Bestandteil des Evaluationsprozesses. Bei einer Ausklammerung dieser Funktion besteht die Gefahr, dass die Evaluation auf die Funktion des Monitoring zurückgeschnitten wird. Aus den Programmzielen sollte ableitbar sein, ob eher monetäre oder nichtmonetäre, objektive oder subjektive Indikatoren, materielle und soziale Deprivationsindikatoren bzw. Kombinationen hiervon für Evaluationen sinnvoll sind.[65]

Damit Evaluationen sinnvoll auf bestehende Indikatoren zurückgreifen können, wäre ein sehr breites Indikatoren-Set notwendig. Da einerseits die Effekte zielgruppen- oder problemspezifischer Programme kaum an allgemein definierten Indikatoren abzulesen sein dürften und umgekehrt Veränderungen von der Ausprägung eines auf Armutslagen in der Gesellschaft bezogenen Indikatorwertes kaum durch bestimmte (kleinere) Programme zu erklären sind, müssen im Rahmen von Evalua-

---

63 Vgl. Semrau/Müllenmeister-Faust (2002) zum Verhältnis von NAPIncl und Armuts- und Reichtumsberichterstattung bzw. konzeptionelle Unterschiede beider Berichtssysteme.

64 Siehe auch Ausführungen zur Werteklärung in Kapitel 2.2, da auch die Auswahl der Indikatoren durch verschiedenste Werte geprägt sein kann.

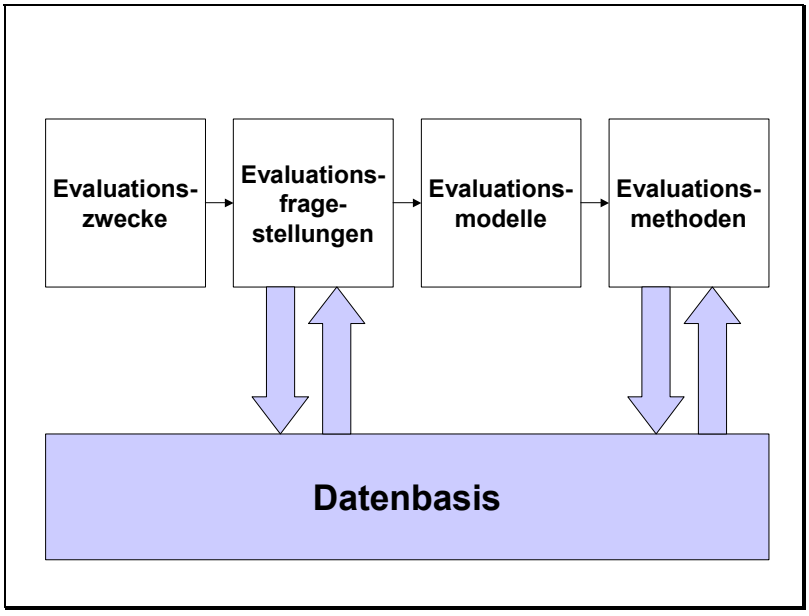
65 Für eine Übersicht vgl. bspw. Vranken et al. (2001), S. 36.

tionen eigene Indikatoren gebildet werden. Je mehr Indikatoren bereits definiert sind und im Rahmen der Armuts- und Reichtumsberichterstattung beobachtet werden, desto größer ist die Wahrscheinlichkeit, diese im Rahmen von Evaluationen als Bezugssystem nutzen oder sogar direkt in eine Evaluation einfließen lassen zu können. Außerdem werden derzeit in den Berichtsystemen von Bund, Ländern und Kommunen verschiedene Indikatoren verwendet. Es wäre wünschenswert, dass die verwendeten Indikatoren angeglichen werden bzw. kompatibel sind. So müssten für jede Lebenslagendimension entsprechende Indikatorensets festgelegt sein.[66] Diese Art von Indikatoren-Set kann für Evaluationen hilfreich sein. Andererseits können bereits existierende Indikatoren aber auch den Blick für unterschiedlichste Wirkungsmechanismen einengen und sollten aus diesem Grund auch im Rahmen von Evaluationen andauernd hinterfragt werden.

4.1.3 Evaluationsmodelle und Datenlage

Für die verschiedenen in Kap. 2.3 vorgestellten Evaluationsmodelle sollen in diesem Abschnitt die dort jeweils eingesetzten Methoden und damit auch die Relevanz bzw. Anforderungen an die Datenlage kursorisch dargestellt werden.

Abbildung 16: Datenbasis für Evaluationen – schematischer Überblick



Quelle: eigene Darstellung

66 Für Anforderungen an Indikatoren vgl. Noll 2003.

Die dargestellten Evaluationsmodelle zeichnen sich durch eine unterschiedliche Haltung zur Auswahl von Methoden aus: Manche sind von vornherein auf ein Methoden-Set festgelegt (z.B. die Programmziel-gesteuerte Evaluation), z.T. in Kombination mit einem bestimmten Erhebungsdesign wie die (quasi-)Experimentaldesign-gesteuerten Evaluationsmodelle. Andere sind völlig offen in der Methodenwahl, darunter einige mit einer Tendenz z.B. zu qualitativen Methoden.

Zumeist werden durch die Modelle bestimmte Phasen des Ablaufs und der Durchführung einer Evaluation besonders betont. Für einzelne Phasen sind wiederum bestimmte Methoden besonders zweckmäßig einzusetzen. Für die Analyse der Bedürfnisse bestimmter Zielgruppen werden oftmals qualitative Methoden eingesetzt. Für die Weiterentwicklung von Programmen dominieren ebenfalls qualitative Methoden. Bei eher abschließenden Programm-Bewertungen stehen häufig quantitativ orientierte Methoden im Vordergrund.

Gerade die Problemstellungen der Evaluation im Kontext der Armuts- und Reichtumsberichterstattung legen eine Kombination quantitativer und qualitativer Verfahren nahe, um seinen situationsspezifischen Anforderungen gerecht werden zu können. Es wird immer wieder die Wichtigkeit von qualitativen Methoden zur Bereicherung von quantitativen Ansätzen betont.[67] Nur quantitative Verfahren sind in der Lage, bspw. die Auswirkungen von Fortbildungsprogrammen auf Beschäftigung und Einkommen zu demonstrieren. Qualitative Studien sind hingegen notwendig, um erstens die Gründe zu erklären, warum manche Programme erfolgreicher sind als andere und zweitens Möglichkeiten zur Verbesserung von Programmen aufzuzeigen.

Summative Evaluationen werden eher mit quantitativen Methoden und formative Evaluationen mit Hilfe qualitativer Methoden vorgenommen, auch wenn hier eine strikte Trennung weder möglich noch sinnvoll erscheint. Außerdem wird von Evaluatoren/-innen vielfach sowohl ein formatives als auch ein summatives Vorgehen verlangt, wobei unterschiedlichste methodische Zugänge miteinander verbunden werden müssen. Eine Diskussion des Einsatzes von qualitativen und quantitativen Methoden ist häufig mit dem Spannungsfeld von Mikro- und Makroperspektive verknüpft.

---

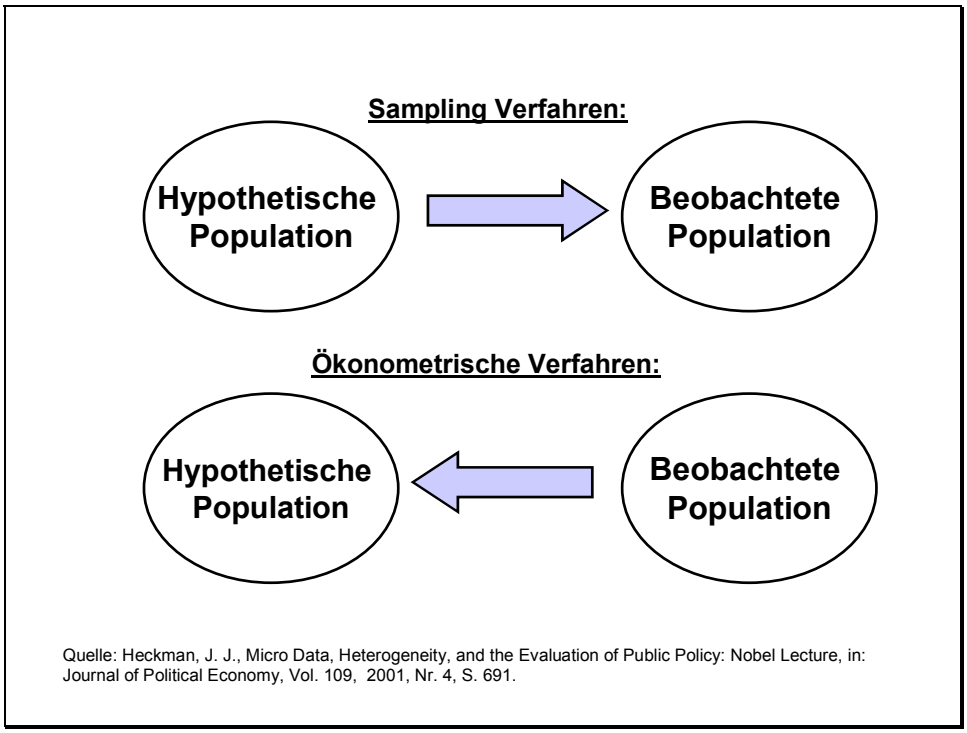
67 Vgl. auch DeGEval-Standard G7 „Analyse qualitativer und quantitativer Informationen“ im Anhang. Vgl. Rossi/Freeman/Lipsey 1999, S. 269ff.

Die empirischen Methoden und Instrumente müssen auf die Ziele und Gegenstände der Evaluation angepasst sein. Im Kommentar des DeGEval-Standards zur Analyse qualitativer und quantitativer Informationen wird gefordert, auf die Aussagekraft der Methoden und auch auf ihre Begrenzungen hinzuweisen. Die unterschiedlichen Methoden benötigen bzw. erzeugen unterschiedliche Arten von Daten - mit ihren jeweiligen Stärken und Schwächen. Im Folgenden wird ein Bezug zwischen Evaluationsmodellen, eingesetzten Methoden und Datenlage hergestellt[68].

Experimentaldesign-gesteuerte Evaluationen basieren auf der klassischen Versuchsanordnung mit einer Experimental- und einer (oder mehreren) Kontrollgruppen, die zunächst möglichst in allen Aspekten identisch sind. Die Bestimmung von Experimental- und Kontrollgruppe muss durch die Evaluatoren/-innen steuerbar sein. Es werden erst im Verlauf des „Experiments“ die für dieses Evaluationsmodell notwendigen Daten hergestellt. Experimentaldesign-gesteuerte Evaluationen sind somit vor allem im Vorfeld der Teilnehmer/-innen-Zuweisung für die Experimental- und Kontrollgruppe auf die externe Datenlage angewiesen, da diese teilweise Grundlage der Auswahl (Randomisierung) bzw. Überprüfung von identischen Eigenschaften der Experimental- mit der Kontrollgruppe sein kann. Hierzu können sowohl aussagekräftige Individualdaten gehören als auch sozioökonomische Daten über einzelne Kommunen oder Stadtbezirke. Aber auch nach Abschluss des Experiments wird meist auf Daten aus dem Programmprozess zurückgegriffen, um Wirkungen abschätzen zu können. Vielfach ist es notwendig, zusätzliche Daten zu erheben – insbesondere für die Zeit nach der Programmteilnahme –, um Aussagen über mögliche Programmwirkungen machen zu können.

68 Siehe hierzu als Hintergrund die ausführlichen Modelldarstellungen im Anhang.

Abbildung 17: Hypothetische Daten in der Evaluation



Quelle: Heckman, J. J., Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture, in: Journal of Political Economy, Vol. 109, 2001, Nr. 4, S. 691.

Im Rahmen eines quasi-experimentellen Evaluationsdesigns sind einerseits Daten für die betreffenden Teilnehmer/-innen eines Programms notwendig und andererseits Daten für Nicht-Teilnehmer/-innen des Programms, die wiederum ähnliche Charakteristika wie die Teilnehmer/-innen aufweisen müssen. Zur Überprüfung dieser Charakteristika müssen die Datensätze zumindest einige vergleichbare Variable beinhalten. Nur so kann bei Anwendung dieses Evaluationsmodells mit Hilfe geeigneter statistischer Matching-Methoden eine kontrafaktische Situation geschätzt werden. Es werden also möglichst Individualdatensätze der Teilnehmer/-innen und einer geeigneten Kontrollgruppe über einen längeren Zeitraum benötigt. Diese Evaluationsstrategie konnte bisher in Deutschland wenig angewendet werden (Hujer/Caliendo 2002). Die Anwendung dieses Evaluationsmodells ist von der existierenden Datenlage abhängig, die kaum im Nachhinein hergestellt werden kann.

Auch für den Einsatz von Mikrosimulationen ist eine fundierte Datenlage notwendige Durchführungsvoraussetzung. Querschnittsdaten sind die Basis für statische und dynamische Mikrosimulationen (Merz, 1991, S. 79). Diese Mikrodatenbasis entsteht dabei erst durch die Verknüpfung von verschiedenen bestehenden Datenbasen (Merz, 1991, S. 95). Es werden auch aus anderen vorherigen Zeitpunkten weitere aggregierte Kontrolldaten – also Makrodaten – benötigt. Mikrosimulationen können

auch auf der Basis von Daten aus Panels durchgeführt werden.[69] Diese Art von Evaluation bezieht sich bisher weitestgehend auf einkommensabhängige Betrachtungen. Deshalb ist dieser Evaluationsansatz für andere Interventionsarten, die über den monetären Transfer bzw. Veränderungen von Steuergesetzen hinausgehen, weniger geeignet.

Bei Kosten-Nutzen-Analysen müssen desto weniger Daten im Rahmen von Evaluationen erhoben werden, je besser verwaltungsinterne Controlling-Systeme aufgebaut und gepflegt sind, je besser also auf bestehende Daten zurückgegriffen werden kann. Im Rahmen der Ausbreitung der Neuen Steuerung werden zunehmend und zeitnah relevante Kostendaten erzeugt.[70] Diese werden jedoch je nach Kostenbegriff und Bezugsebene in den seltensten Fällen ausreichen. Insbesondere bei einer psycho-sozialen Nutzenerfassung müssen zumeist eigenständige Erhebungen durchgeführt werden.

Für experimental- bzw. Quasi-Experimentaldesign-gesteuerte Evaluationen oder auch Programmkosten/-nutzen-gesteuerte Evaluationen ist es im Vergleich zu anderen Modellen relativ eindeutig, wie die Datenlage als Voraussetzung zur Modellnutzung aussehen muss. Bei anderen Evaluationsmodellen sind die einzusetzenden Methoden weniger im Vorhinein festgelegt und damit die Erfordernisse an die Datenlage weniger verallgemeinerbar. Des Weiteren gibt es Modelle, die eher partizipativ orientiert sind und bei denen die eigene Datenerhebung ein fester Bestandteil ist. Diese können relativ unabhängig von der bestehenden Datenlage eingesetzt werden.

Bei Programmziel-gesteuerten Evaluationsstudien sind die Anforderungen an die Datenlage von den eingesetzten Methoden abhängig. Es sind zumeist spezifische Daten über Zielgruppen und Kontext notwendig. Die Evaluationsfragestellungen von Programmziel-gesteuerten Evaluationen lassen sich selten befriedigend mit existierenden Daten beantworten. Dies gilt auch für den Einsatz von quantitativen, auf operationalisierten Zielen beruhende Tests bei „harten“ Outcome-Indikatoren.

---

69 Vgl. Otto (2002) als ein Beispiel für eine Mikrosimulation im Kontext der Armutsbekämpfung auf der Basis des SOEP.

70 Für Controlling in der Sozialhilfe vgl. BMGS (1999b), S. 142ff.

Bei der Programmtheorie-gesteuerten Evaluation hängen die Anforderungen an die Datenlage von der Differenziertheit der gewählten bzw. entwickelten Theorie(n) ab. In aller Regel müssen hierfür eigens neue Datenerhebungsinstrumente entwickelt werden. In ihrem Rahmen ist eine differenzierte Beschreibung des Kontextes des Programms bzw. seiner lokalen Umsetzungen erforderlich. Dies wird durch eine Rückgriffmöglichkeit auf vorhandene Indikatorensysteme erheblich vereinfacht.

Bei der Entscheidungsgesteuerten Evaluation wird nicht nur ein konkretes Programm, sondern es werden evtl. auch andere Möglichkeiten sowie Einschätzungen konkurrierender Programme einfließen. Die allgemeine Datenlage kann ebenfalls zur Kontextbeschreibung wichtig sein. Die Probleme hierbei sind jenen der Programmziel-gesteuerten Evaluationen und Programmkosten/-nutzen–gesteuerten Evaluationen ähnlich. Im Kern fordern Entscheider/-innen für Entscheidungssituationen sehr spezifische Daten an, zu denen gesonderte Instrumente entwickelt werden müssen, die zudem schnell einsetzbar und auswertbar sein müssen (Zeitdruck in Entscheidungssystemen). Hieran wird besonders klar ersichtlich, dass zwischen den auf „Nützlichkeit“ – hier Nutzung in Entscheidungen – und den auf „Genauigkeit“ – Datenqualität – zielenden Evaluationsstandards erhebliche Spannungen bestehen (vgl. Beywl/Speer im Erscheinen).

Für Evaluationen mit beteiligtenorientierten Methoden werden Daten vorrangig von den Evaluatoren/-innen erhoben, da diese auf spezifische Informationsanliegen der Beteiligten und Betroffenen antworten wollen. Welche Art der Datenlage für die Nutzungsgesteuerte Evaluation gegeben sein sollte, ist vom jeweiligen Evaluationszweck und den konkreten vorgesehenen Nutzungen der Evaluationsergebnisse abhängig. Es besteht eine Selbstverpflichtung, schon um die Kostenwirksamkeit der Evaluation zu optimieren, vorhandene Daten zuerst auszuwerten und diese weitestmöglich auszuschöpfen.

Im Rahmen von Spannungsthemen-gesteuerten Evaluationsverfahren wird weitestgehend mit offenen, qualitativen Verfahren gearbeitet. Da eher auf narrative Beschreibungen als auf die Sammlung von Messdaten über Fälle hinweg zurückgegriffen wird, ist die bestehende Datenlage weniger wichtig. Sie kann zur Kontextbeschreibung herangezogen werden, die aber ebenfalls von dichten qualitativen Informationen lebt, zu der einige zentrale sozioökonomische Kennzahlen ergänzend hinzugezogen werden können.

Die Selbstorganisationsgesteuerte Evaluation setzt auf Zusammenarbeit mit konkreten Betroffenen. Es findet selten eine übergreifende unabhängige Datenanalyse statt, es wird vielmehr mit qualitativen und quantitativen Methoden ein Bezug zum unmittelbaren Lebensumfeld der Handelnden hergestellt. Der augenfällige lebensrelevante Gehalt von Daten (face-validity) spielt in diesem Evaluationsmodell eine herausragende Rolle. Diese Art der Evaluation stützt sich kaum auf bestehende Daten.

Die Dialoggesteuerte Evaluation hat die stärkste Tendenz dazu, möglichst viele unterschiedliche Datenquellen zu nutzen: Von Sozialindikatorensystemen bis hin zu narrativen Einzel-Intensivinterviews. Wahrheitswert von Evaluationsergebnissen kann hier gemäß den Autor/-innen des Modells nicht aus einer vorgeblich besonders validen Quelle entstehen, sondern vielmehr aus dem Abgleich mehrerer Quellen, die zur Aufklärung eines Sachverhaltes genutzt werden. Gegenüber den in vorhandene Indikaktorensysteme eingebauten kulturellen Verzerrungen ist die Dialoggesteuerte Evaluation sehr kritisch eingestellt

Die Kontext-Mechanismus-gesteuerte Evaluation fordert eine sehr enge Abstimmung der Datenerhebungsinstrumente auf die drei für Programmevaluationen leitenden Theorie-Elemente: Kontext, Mechanismen und Outcomes. Die Instrumente müssen eigens entworfen werden; grundsätzlich sind alle Erhebungsmethoden nutzbar. Bestehende Indikatorensysteme dienen hier ausschließlich zur Problembeschreibung und werden zur Wirkungsfeststellung nicht herangezogen.

Nicht nur für die jeweiligen Evaluationsmodelle gibt es – wie hier dargestellt – spezifische Anforderungen an die Datenlage, auch zur Beschreibung des Kontextes eines Evaluationsgegenstandes ist die bestehende Datenlage von hoher Bedeutung. Eine Kontextbeschreibung ist Teil der meisten Evaluationsmodelle, bei einigen Methoden fließen Kontextvariablen in weitere Berechnungen ein. Insbesondere für die Kontextbeschreibung kann im Rahmen von Evaluationen nicht zu viel Aufwand betrieben werden. So merkte ein Experte an:

> *„Es kann wichtig sein zu wissen, wie auf Kreisebene die aktuelle Arbeitsmarktlage charakterisiert werden kann. Solche Daten, wie kommunenbezogene oder stadtteilspezifische Angaben über Sozialhilfequoten, Arbeitslosenquote, Teilnahmen an bestimmten Maßnahmen der aktiven Arbeitsmarktpolitik, müssten einfacher verfügbar sein“.*

4.1.4 Kosten-Nutzen-Aspekte von Evaluationen

Der DeGEval-Standard D 3 besagt: „Der Aufwand für Evaluation soll in einem angemessenen Verhältnis zum Nutzen der Evaluation stehen.“ Im Erläuterungstext zu diesem Standard heißt es, dass es sowohl zu Beginn als auch bei Abschluss einer Evaluation oft schwierig ist, genaue Aussagen zu Kosten und Nutzen eines Evaluationsvorhabens zu machen. Dennoch sollten insbesondere bei der Entscheidung zu einzelnen Evaluationen Kosten und Nutzen abgeschätzt werden. Die Kosten umfassen den monetären Wert aller eingesetzten Ressourcen. Die direkten Kosten können in der Regel genau beziffert werden. Indirekte Kosten hingegen sind schwierig zu quantifizieren. Dies gilt umso mehr für die Nutzenseite. Diese kann nur geschätzt werden.

In diesem Abschnitt werden einzelne Kriterien angeführt, die für die Entscheidung zur Evaluation im Kontext der Armuts- und Reichtumsberichterstattung relevant sein können.[71] So sollte bereits in der Budgetierung im Zusammenhang mit der Entscheidung für eine Programmdurchführung eine Programm-Evaluation entsprechend eingeplant werden. Es ist bei einer Entscheidung für eine Evaluation zu bedenken, welche Reichweite diese hat. So kann z.B. für ein Modellprojekt eine besonders intensive Evaluation sinnvoll sein, da bei einer späteren Übertragung des Modellversuchs eine große Anzahl an teilnehmenden Personen betroffen sein bzw. ein entsprechend großes Budget eingesetzt wird. Des Weiteren dürfte ausschlaggebend sein, wie viele Personen bereits von dem Programm betroffen sind, wie entscheidend die von dem Programm erwartete Zielerreichung für die Vermeidung bzw. Verminderung von Armut ist, und wie wahrscheinlich es ist, dass das Programm dauerhaft Bestand hat. Außerdem dürfte auch eine besondere Neuartigkeit eines Programms den Bedarf an Evaluation erhöhen. Wie zentral ist das Programm zur Vermeidung/Beseitigung von Armut in Bezug auf die verschiedensten betroffenen Politikfelder bzw. wie stark sind diese miteinander verknüpft? Wird die Evaluation als Grundlage für bestimmte Entscheidungen besonders dringend gebraucht? Außerdem kann die Entscheidung für die Durchführung einer Evaluation von den Stärken bzw. Schwächen der Programmkonzeption bzw. -implementation abhängig gemacht

---

71 Für diesen Abschnitt vgl. die Kommentierung des Standards „F 3 Evaluation Efficiency“ in Beywl/Speer (2004) bzw. siehe DeGEval-Standard D3 „Effizienz von Evaluation“ und auch U4 „Auswahl und Umfang der Information“ im Anhang.

werden. Wenn es also sowohl Probleme im Programm-Design als auch bei der Implementation gibt, erscheint eine früh einsetzende „interaktive“ Evaluation besonders dringlich. Diese Überlegungen gelten für die Auswahl von insgesamt zu evaluierenden Programmen genauso wie für die Auswahl von zu evaluierenden Teil-Aspekten solcher Programme.

4.2 Administrative Daten

Unter **„administrativen oder prozessgenerierten Daten“** verstehen wir solche, die im Rahmen der Verwaltungsvorgänge, bspw. der Sozial- und Arbeitsverwaltungen, erzeugt werden oder aus dem Monitoring von Programmen oder Buchhaltungssystemen von Programmträgern stammen. Administrative Daten werden allein zum Zwecke der Ausführung von Gesetzen und Vorschriften erhoben. Die Auswahl der Daten ist von Gesetzen und Vorschriften abhängig. Administrative Daten können unterschieden werden in solche, die in unmittelbarem Zusammenhang stehen mit der Bewilligung von Geldtransfers bzw. mit der Programmteilnahme, und Daten, die hierfür keine Relevanz haben.

4.2.1 Vollständigkeit und Inhalte

Administrative Daten geben vielfach genaue Informationen über Programmteilnahme, d.h. darüber, wer welche Leistungen in welcher Höhe und wie lange erhalten hat. Es werden im Regelfall Daten für alle Programm-Teilnehmenden erhoben.

> *(Es sei sehr positiv,) „... dass umfassend Aspekte erhoben werden, die für Evaluationen wichtig sind, z.B. im Falle der Sozialhilfeakten sind Einkommensarten und deren Höhe enthalten; es ist vermerkt, wie diese Einkommensarten angerechnet werden und welche Hilfemaßnahmen den jeweiligen Sozialhilfeempfängern empfohlen werden.“*
> *(§ C)*[72]

Es sind somit Einzelfallauswertungen, Analysen von Fallverläufen und Bezugsdauern von Transferleistungen möglich. Des Weiteren sind Auswertungen auf kommunaler Ebene möglich, was für die Evaluation von Modellprojekten besonders bedeutsam

72 Dies gilt für die Bearbeitung vor Ort. In der amtlichen Statistik (s.u. Kapitel 4.3 „Statistische Daten“) der Sozialhilfe wird die Höhe der anzurechnenden Einkommen nicht erhoben.

ist, aber auch für kommunale Vergleiche im Rahmen von Evaluationen. Administrative Daten sind in der Regel sehr zeitnah verfügbar.

Die Angaben, die für die Bewilligung von Geldtransfers oder Programm-Teilnahmen von den Teilnehmer/-innen gemacht werden, dürften zu einem hohen Grad valide sein.[73] Ein weiterer Vorteil von administrativen Daten liegt somit auch in ihrer empirischen Überprüfbarkeit – insbesondere was die in Interviews oder Surveys eher tabuisierten Angaben zum Einkommen angeht. Hierin besteht ein großer Vorteil dieses Datentyps gegenüber Daten aus Umfragen.

> *„Daten werden häufig in den Prozessen generiert, die man evaluieren will, z.B. Historikdatei der Bundesanstalt für Arbeit; auf deren Basis können Erwerbsbiografien abgebildet werden." (§ B)*

Außerdem liefern bspw. Daten aus dem Sozialhilfebezug auch haushaltsbezogene Daten. Fehlen Informationen, kann dies zwei Gründe haben: Entweder soll diese Information nicht erhoben werden, weil sie für die Erhebungsstelle für die Zahlbarmachung nicht von Interesse ist, oder die Information kann nicht erhoben werden, weil es auf dem Verwaltungswege kaum möglich ist, hierüber Information zu erhalten, wie z.B. bei Schwarzarbeit.

Auf die Programm-Teilnehmer/-innen bezogen handelt es sich bei den administrativen Daten um Vollerhebungen. Sie liefern eine vollständige Datenbasis und es gibt somit keine Stichprobenprobleme. Es wäre insbesondere bei relativ geringem Vorkommen der Programm-Zielgruppe in der Gesamt-Population teuer, ähnliche Informationen durch eigene Erhebungen wie Umfragen zu generieren. Die administrativen Daten sind unabhängig vom Evaluationsprozess vorhanden und sind deshalb grundsätzlich eine sehr „kostengünstige“ Datenquelle. Allerdings hat der Versuch, Sozialhilfegeschäftsdaten in den nordrhein-westfälischen Modellversuchen zur Pauschalierung von Sozialhilfe gezeigt, dass es einerseits eine Vielzahl von Fehlerquellen gibt (z.B. Verbuchung unter einem falschen, ähnlich klingenden sozialhilferechtlichen Sachverhalt) und dass insbesondere zwischen verschiedenen Kommu-

---

73 Vgl. Engels/Sellin (2000) zur Evaluation des automatisierten Datenabgleichs in der Sozialhilfe nach §117 Abs. 1 und 2 BSHG und Aufdeckung von Missbrauchsfällen. Wie sich im Rahmen des Datenabgleichs in jüngster Zeit gezeigt hat, sind die Angaben der Hilfeempfänger/-innen in den meisten Fällen wahrheitsgetreu, es gibt nur eine sehr geringe Quote von Missbrauch bzw. Falschangaben.

nen sehr unterschiedliche Buchungsroutinen bestehen, was einen interkommunalen Vergleich nur nach sehr aufwändigen Umcodierungsarbeiten ermöglicht.

Administrative Daten sind im Hinblick auf die „Adäquation“ zu einer Forschungsfrage bzw. zu einzelnen Evaluationsfragestellungen relativ unflexibel. Die Daten sind vielfach um weitere Informationen zu ergänzen. So werden nur programmspezifische Angaben erhoben, insbesondere im Kontext der Armuts- und Reichtumsberichterstattung wären dagegen Angaben zu mehreren Lebenslagendimensionen wünschenswert.

Das Fehlen dieser Informationen ist ein Grund dafür, dass sub-zielgruppenspezifische Analysen schwierig durchzuführen sind. Wenn beispielsweise die Situation von Aussiedlern/-innen betrachtet werden soll bzw. Wirkungen von Programmen auf diese spezielle Sub-Zielgruppe betrachtet werden sollen, dann ist dies mit Hilfe von administrativen Daten nach der Meinung der Experten/-innen schwierig oder unmöglich zu erreichen.

> *„Zwar sind Angaben über die Einreise und die Verteilung auf die Bundesländer verfügbar, aber eine weiter gehende aussiedlerspezifische Datenerhebung ist mit Schwierigkeiten verbunden, da Aussiedler Deutsche im Sinne des Artikels 116 Abs.1 GG sind und somit nicht gesondert statistisch erfasst werden. In der Arbeitsmarktstatistik werden sie für einen Zeitraum von fünf Jahren nach der Einreise gesondert dokumentiert, in der Sozialhilfestatistik gar nicht.“*[74] *(§ C)*

Des Weiteren fehlen viele für Evaluationen wichtige Informationen, z.B. ist es auf der Basis der administrativen DV-Daten der Sozialämter nicht erkennbar, ob eine Person Arbeitslosengeld oder Arbeitslosenhilfe bezieht, hierfür wäre eine gesonderte Auswertung der Papier-Akten erforderlich. Dies liegt z.T. auch daran, dass die Daten für Bedarfsgemeinschaften und nicht für Personen vorliegen.

Zusammengefasst besteht ein Spannungsfeld zwischen dem Interesse der Evaluatoren/-innen und dem der Kommune bzw. des Sachbearbeiters/der Sachbearbeiterin existiert. Für die Evaluation wäre eine Erhebung von relativ vielen Variablen auf dem administrativen Wege wünschenswert, weil damit die Wahrscheinlichkeit steigen würde, dass für unterschiedliche Evaluationsfragestellungen bereits

---

74 S. Engels et al. (2000), S. 4. Dies ist ein Aspekt, der sowohl auf administrative als auch auf statistische Daten zutrifft.

administrative Daten zur Verfügung stehen. Für die Sachbearbeiter/-innen hingegen würde dies eine zusätzliche Arbeitsbelastung bedeuten, die für sie zunächst keinen Ertrag bringt.

Da in den administrativen Daten nur die tatsächlichen Programmteilnehmer/-innen erfasst werden können, enthalten sie keine Informationen über Anspruchsberechtigte, die nicht am Programm teilnehmen. So kann die Angabe, wie viele Personen z.B. in Notunterkünften leben, nur ein Hinweis (unter anderen) für die tatsächliche Zahl Obdachloser sein. Die ausschließliche Verwendung administrativer Daten würde also zu einer Nichterfassung der so genannten verdeckten Armut führen. Administrative Daten sollten aus diesem Grund durch andere Datentypen ergänzt werden, um Kenntnis über Anspruchsberechtigte unabhängig von tatsächlicher Programm-Teilnahme zu erhalten und darauf aufbauend die Zielerreichung eines bestehenden Programms beurteilen oder auch Wirkungen von Programmveränderungen einschätzen zu können.

Hinzu kommt, dass die Vergleichbarkeit kommunaler Daten u. U. nicht gegeben ist, da Kommunen selbst bestimmen, wie Daten erhoben und insbesondere abgelegt werden. Eine Verbesserung der Daten-Vergleichbarkeit von Kreisen oder Städten ähnlicher Größenklassen wird z. Zt. durch so genannte Vergleichsringe in Angriff genommen.[75] Allerdings handelt es sich hierbei entweder bereits um Auswertungen administrativer Daten, was einen gesonderten aufwändigen Arbeitsgang erfordert, oder der Vergleich wird auf einen sehr beschränkten Datenkranz eingegrenzt, der sich für Programmevaluationen vielfach als zu eng erweisen wird.

Wird im Rahmen von Evaluationen mit administrativen Daten aus unterschiedlichen Quellen gearbeitet, gibt es weitere Probleme bei der Kompatibilität regionaler Bezugsgrößen. Bspw. werden Angaben zur Sozialhilfe auf kommunaler Ebene erhoben, Arbeitsamtbezirke gehen jedoch z.T. über Stadtgrenzen hinaus. Dies kann sowohl bei Kontextbeschreibungen als auch bei der Analyse zentraler Effekte Schwierigkeiten bereiten.

---

75 Diese Vergleichsringe werden von der Bertelsmann-Stiftung, Gütersloh, unterstützt.

4.2.2 Datenqualität

Einerseits wünschen sich Evaluatoren/-innen und Wissenschaftler/-innen, dass administrative Daten aufbereitet zur Verfügung gestellt werden sollten. Andererseits wird argumentiert, dass für die Analyse der Daten auch die Kenntnis ihrer Genese notwendig sei.[76] Die eigene Gewinnung und Aufbereitung der Daten wäre hilfreich. Aus Kosten-Nutzen-Aspekten heraus kann jedoch alternativ über eine sehr vollständige und präzise eingehaltene Dokumentation der Erhalt des erforderlichen Detailwissens gewährleistet werden. De facto werden solche Dokumentationen in den Ämtern häufig nicht vollständig geführt bzw. teilen sich auf eine Vielzahl von Verwaltungsanweisungen auf, die z.B. in einem größeren Sozialamt nur wenige Personen vollständig im Überblick haben. Die Experten/-innen betonen mehrfach, wie wichtig die Berücksichtigung der sachlichen, zeitlichen und sozialen Zusammenhänge der Datenentstehung für die Datenanalyse sei. Zu den Informationen über die Datenentstehung gehört das Wissen über die „Regeln" und „Eingabeanweisungen" der Ämter, genauso wie das Wissen über deren Umsetzung und institutionelle Regelungen.[77]

Die Qualität der Datenerhebung ist wiederum von den unterschiedlichen Institutionen, deren Anreizmechanismen sowie der Gründlichkeit und dem Belastungsgrad der Personen, die erheben, abhängig.[78] So wird von mehreren Experten/-innen berichtet, dass für die Sachbearbeiter/-innen vor Ort der Fokus stärker auf der Erhebung der für die Programm-Teilnahme bzw. -vergabe relevanten Daten liegt und weitere administrative Daten vernachlässigt werden. Administrative Daten werden nicht unter sozialwissenschaftlichen und auch nicht unter statistischen Gesichtspunkten erhoben. Sie sind somit nur für den Geschäftsgang valide, für den sie erhoben werden. Für viele Daten, die für Evaluationsfragestellungen bedeutsam sind, ergeben sich jedoch andere Anforderungen an den Dateninhalt. Dieses Validitätsproblem von aus Sicht des „Zahlbarmachungsprozesses der Kommune" unwichtigen Daten fällt bei Plausibilitätsüberprüfungen durch die Landesämter für Statistik auf, wenn bspw. 90% der Hilfeempfänger/-innen als ungelernt geführt werden oder die Schulaus-

---

76 Vgl. Brinkmann (2002), S. 14.

77 Dies gilt auch bei dem Vorliegen von verbindlichen Vorgaben zur Datenerhebung.

78 Objektivität ist ein zentrales Gütekriterium im Bereich der quantitativen Methoden. Es besagt, dass Messungen unabhängig von der jeweiligen Person sein sollte, die das Messinstrument anwendet. Siehe auch DeGEval-Standard G5 „Valide und reliable Informationen" im Anhang.

bildung durchgängig nicht bekannt ist. Außerdem kann es sein, dass bspw. bei Beginn einer Programm-Teilnahme bestimmte Daten erhoben werden – z.B. Familienstand oder Bildungsniveau – diese sich jedoch im Zeitverlauf ändern, die administrativen Daten allerdings nicht weiter aktualisiert werden.

Vorgänge im Rahmen der Vermittlungsstatistik der Bundesanstalt für Arbeit haben gezeigt, welche Art von Fehldeutungen möglich sind, wenn zu wenig über den Entstehungsprozess und damit auch die Qualität der Daten bekannt ist. So kann es bspw. praktische Probleme aus Gründen der Zeitknappheit geben. Ein/-e Teilnehmer/-in einer Fördermaßnahme der beruflichen Weiterbildung erhält Unterhaltsgeld in der Höhe des Arbeitslosengeldes. Deshalb kommt es nicht immer zu einer sofortigen Umbuchung von Arbeitslosengeld auf Unterhaltsgeld durch die betreffenden Vermittler/-innen.[79] Würden jedoch durch Computerprogramme die Veränderungen bestimmter Umstellungen „erzwungen“, könnte die Zuverlässigkeit der Daten erhöht werden. Andererseits stoßen Neueinführungen und Ergänzungen von DV-Systemen immer wieder auf große Hindernisse, da sie den Sachbearbeitern/-innen immer wieder Umstellungen abverlangen, was bei vielen tausend Buchungsvorgängen pro Monat und Beschäftigtem leicht eine zu hohe Belastung bedeuten kann.

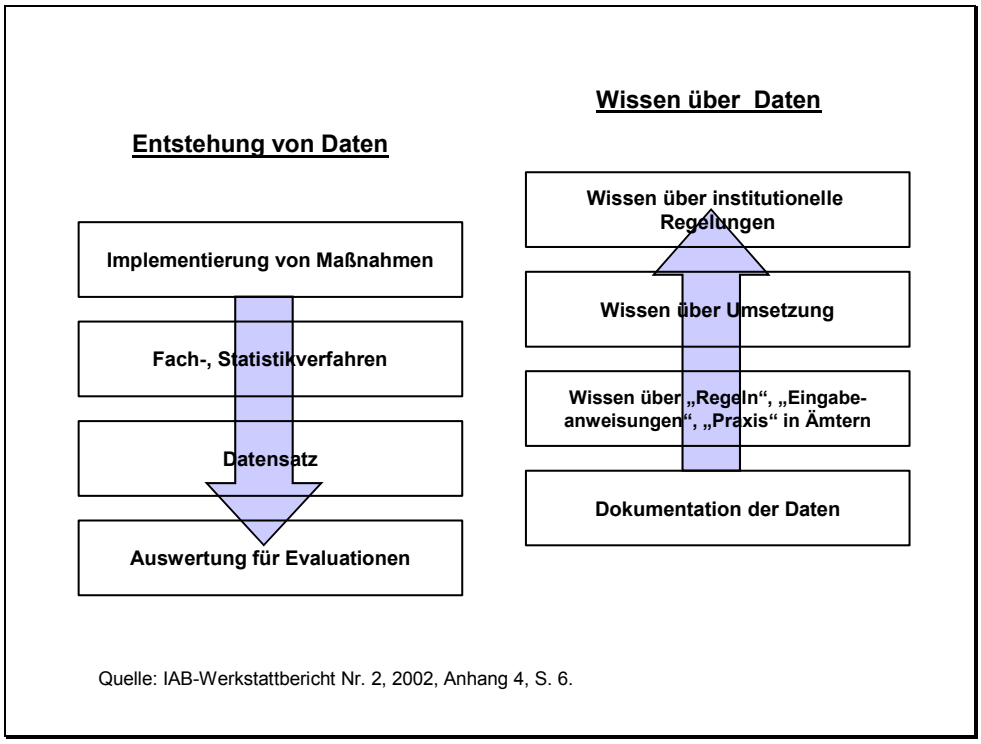
Die nicht befriedigende Datenqualität von administrativen Daten hängt auch damit zusammen, dass es keine Anreize für Kommunen zur Erhebung von bestimmten Daten gibt. Auf kommunaler Ebene besteht häufig wenig Interesse an der Aufbereitung sowie Nutzung der administrativen Daten, die für die Zuweisungen z.B. durch den Kreis als Sozialhilfeträger nicht entscheidend sind. Die Erhebung dieser Daten wird eher als lästige Pflicht betrachtet. Hierzu gehören bspw. Angaben über Schulbildung oder der Ausbildung von Programm-Teilnehmer/-innen, die für die Kommunen nur bei Maßnahmen wie Hilfen zur Arbeit interessant sind.[80]

---

79 Vgl. Bender (2002), S. 19.

80 Dies könnte sich im Rahmen der neuen Grundsicherung und der Anwendung anderer Zuweisungsmodalitäten ändern.

Abbildung 18: Wissen über die Entstehung von Daten



Quelle: IAB-Werkstattbericht Nr. 2, 2002, Anhang 4, S. 6.

Die Experten/-innen machen folgende Vorschläge: Ein/-e Ansprechpartner/-in für die Datenlage auf Kreisebene wäre wünschenswert. Diese/-r „Datenexperte/-in auf Kreisebene“ sollte Auskünfte geben können z.B. über Art der Datenerhebung, des Datenstandes und der Datenqualität. Im Idealfall würde diese Person auch die Daten aus dem System für Evaluationen zur Verfügung stellen. Auch ein weiterer Einsatz des Controllings – im Sinne der Neuen Steuerung – in den öffentlichen Verwaltungen dürfte die Datenqualität erhöhen. So wird von Experten/-innen vorgeschlagen, entsprechende Controller/-innen, die eine große Nähe zu den Praxisproblemen haben, in den Evaluationsprozess mit einzubinden.

Auch ein Audit-Sytem für die Genauigkeit von Daten könnte deren Qualität weiter erhöhen, wozu auch in die EDV eingebaute Kontrollen gehören könnten.

4.2.3 Datenschutz

Administrative Daten sind aus Datenschutzgründen für Verwaltungsexterne schwierig zu erhalten. Vielfach müssen intime private Informationen geliefert werden, damit eine Programm-Teilnahme gewährleistet wird. Deshalb herrscht Vertrauensschutz, ein Schutz des Privaten bzw. der Intimsphäre (vgl. Brady et al., 2002). Insbesondere im Kontext der Evaluation von Programmen zur Überwindung bzw. Vermeidung von Armut wird deshalb als Gefahr der stark erhöhten Datentransparenz eine soziale

Brandmarkung (Stigmatisierung) gesehen. Es sollte aus diesem Grund fortlaufend eine Abwägung zwischen umfassender Analyse der administrativen Daten und Verletzung der Privatsphäre stattfinden.

Nach der Entscheidung, welche Daten erhoben werden sollen, ist festzulegen, wer Zugang zu den Daten haben soll bzw. unter welchen Umständen Zugang gewährt werden sollte.[81] Es ist zu überlegen, ob es nicht sinnvoll ist, längerfristig überlokale Clearingstellen zu errichten, in denen die verschiedenen Interessen in Bezug auf die Evaluation sozialer Programme vertreten sind (Verwaltungsleiter/-innen, Datenschützer/-innen, DV-Experten/-innen, Evaluatoren/-innen, ...), die nach exemplarischen Lösungswegen suchen. Ansonsten finden Evaluationen vielfach in einem rechtlich spannungsreichen Feld statt – mit erheblichen Risiken für alle Beteiligten.

Grundsätzlich ist es unter der Voraussetzung der Anonymisierung möglich, administrative Daten für Evaluationen zu nutzen.

> *„Die Übermittlung von personenbezogenen Sozialdaten für die wissenschaftliche Forschung und Planung im Sozialleistungsbereich ist grundsätzlich möglich, um den unabweisbaren Bedürfnissen der Forschung und Planung Rechnung zu tragen. Sie steht jedoch gem. § 75 SGB X unter dem Vorbehalt, dass schutzwürdige Interessen von Betroffenen nicht beeinträchtigt werden oder das öffentliche Interesse an der Forschung oder Planung das Geheimhaltungsinteresse des Betroffenen erheblich überwiegt."*[82]

> *„Dennoch gehe ich unter Abwägung der Interessen der Wissenschaft auf freie Forschung und dem Recht auf informationelle Selbstbestimmung davon aus, dass im Rahmen des Modellvorhabens zunächst grundsätzlich von den Trägern der Sozialhilfe nur anonymisierte oder aggregierte Daten an die wissenschaftliche Begleitung bzw. das beauftragte EDV-Unternehmen weitergegeben werden. In diesem Zusammenhang weise ich darauf hin, dass Daten auch dann noch als anonymisiert gelten, wenn die Feststellung der Identität der betroffenen Person zwar möglich erscheint, der Aufwand*

---

81 Die Entscheidung darüber, welche Daten zu erheben sind, wird in Deutschland relativ unabhängig von der betreffenden Programmevaluation gefällt.

82 Siehe MASQT Datenschutzerklärung, in: Booklet für die Hilfeberechtigtenbefragung im Modellprojekt Pauschalierung in der Sozialhilfe 2001.

hierfür aber, gemessen an einem etwaigen Interesse Dritter, unverhältnismäßig hoch wäre."[83]

Im Rahmen des neu gefassten § 118 Bundessozialhilfegesetz zur Wissenschaftlichen Forschung im Auftrag des Bundes heißt es:

„Der Träger der Sozialhilfe darf einer wissenschaftlichen Einrichtung, die im Auftrag des Bundesministeriums für Arbeit und Sozialordnung ein Forschungsvorhaben durchführt, das dem Zweck dient, die Erreichung der Ziele von Gesetzen über soziale Leistungen zu überprüfen oder zu verbessern, Sozialdaten übermitteln, soweit 1. dies zur Durchführung des Forschungsvorhabens erforderlich ist, insbesondere das Vorhaben mit anonymisierten oder pseudonymisierten Daten nicht durchgeführt werden kann, und 2. das öffentliche Interesse an dem Forschungsvorhaben das schutzwürdige Interesse des Betroffenen an einem Ausschluss der Übermittlung erheblich überwiegt ..."

Für die Bundesanstalt für Arbeit ist es zur gesetzlichen Pflicht geworden, wissenschaftlichen Einrichtungen auf Ersuchen anonymisierte Daten für Zwecke der Arbeitsmarkt- und Berufsforschung zur Verfügung zu stellen.[84]

Aus Datenschutzgründen werden in der Praxis bestimmte Informationen ausgeblendet bzw. nicht für die Evaluation weitergegeben.[85] Hierzu gehört bspw. das Geburtsdatum von Sozialhilfeempfänger/-innen. Ohne diese Angabe – zumindest in verkürzter Form – ist es jedoch nicht möglich „Alterskohorten“ zu bilden. Hier wäre nach Experten/-innen-Meinung eine vorherige Einordnung des Geburtsdatums in eine entsprechende Altersklasse eine andere, wenngleich nicht ohne zusätzlichen Aufwand leistbare Alternative. Die Art und Weise der Anonymisierung sollte deshalb möglichst in Absprache mit den Evaluatoren/-innen stattfinden.

---

83 ebenda

84 Dies gilt gemäß § 282 SGB III, neuer Absatz 7. Ist ein Rückgriff auf nicht anonymisierte Daten unverzichtbar, ist auch weiterhin das Verfahren nach §75 SGB X anzuwenden.

85 Im Rahmen des Datenschutzes sind im Kontext der Armuts- und Reichtumsberichterstattung neben den Regelungen des Bundesdatenschutzgesetzes vielfach die Besonderheiten des Schutzes von „Sozialdaten“ relevant. Hierzu gehören Daten, die von einem Leistungsträger im Sinne des SGB zur Erfüllung seiner Aufgaben erhoben, verarbeitet oder genutzt werden.

4.2.4 Technischer Datenzugang

Auf Grund der kommunalen Selbstverwaltung können Kommunen grundsätzlich selbst bestimmen, wie sie den Dokumentationsprozess von Programmen gestalten bzw. mit welcher EDV sie erhobene Daten verwalten. Gesetzgebungskompetenz und Durchführungskompetenz liegen in unterschiedlichen Händen. Auf Grund dieser Tatsache herrschen vielfältige Inkompatibilitäten verschiedener EDV-Systeme. Auch innerhalb von Kreisen verwenden einzelne Kommunen unterschiedliche Software, so lassen sich administrative Daten auf Kreisebene ggf. nur schwierig zusammenfassen. In kreisfreien Städten wird hingegen einheitlich erhoben. Nach Experten/-innen-Meinung haben größere (kreisfreie) Städte (ab ca. 100.000 Einw.) einen wesentlich besseren Datenstand als Kreise. So ist auch im Rahmen von Modellprojekten zu beobachten, dass hieran gerade größere Städte oder einzelne Kreise teilnehmen, die besonders fortgeschritten in ihrer Steuerung sind.

In Nordrhein-Westfalen ist für das Modellprojekt der Sozialagenturen comp.ASS Casemanager eine Software für das Fallmanagement entwickelt worden. Dieses EDV-Programm wird den Sozialagenturen kostenlos zur Verfügung gestellt. Andere Kommunen und Städte, die nicht an dem Modellprojekt teilnehmen, erhalten diese Software mit einem Rabatt, der mit den folgenden Jahren immer geringer wird. Durch diese und ähnliche Anreize könnte es zu einer zunehmenden Vereinheitlichung von EDV-Systemen zwischen den Kommunen kommen und damit könnte auch das Zusammenführen von Daten erleichtert werden – das Einverständnis der Kommunen vorausgesetzt. Diese Daten aus dem Fallmanagement könnten für Evaluationen hoch relevant sein. Inwiefern die Vorteile des Rabatts bereits von Nicht-Modellkommunen wahrgenommen worden sind, ist derzeit nicht bekannt. Mit Regelungen wie der geschilderten können in jedem Fall nur Anreize geschaffen werden, ein Zwang zur Nutzung bestimmter EDV kann auf die Kommunen nicht ausgeübt werden.

Außerdem wäre denkbar, dass auf Bundesebene – von dem für das jeweilige Programm zuständige Bundesministerium – geeignete Dokumentationsverfahren entwickelt werden und diese den Maßnahmenträgern angeboten werden. Dies könnte den Erhalt von vereinheitlichten Daten begünstigen. Die Dokumentationsverfahren sollten dabei zwei Kriterien erfüllen: erstens die Geschäftsprozesse, also die Belange der Leistungserbringer, unterstützen und zweitens valide Ergebnis-

se produzieren. Gleichzeitig ist nicht von der Hand zu weisen, dass solche Angebote nur in Zusammenhang mit (kostenfreien/-günstigen) Schulungsangeboten und einem Einführungscoaching auf breitere Akzeptanz treffen dürften. Sie können nur den Charakter von Empfehlungen haben.

Da die Erhebung von Daten in der kommunalen Verantwortung liegt, wäre des Weiteren daran zu denken, einen einheitlichen „Exportstandard“ für die Datenverarbeitung zu vereinbaren. Dies würde bedeuten, dass administrative Daten weiterhin mit verschiedenster Software verwaltet werden, jedoch die Überführung der Daten in eine statistische Software (bspw. SPSS, SAS oder Sphinx) vereinfacht würde. Von dieser wünschbaren Situation ist die aktuelle Praxis noch weit entfernt.

4.2.5 Erzeugung von Längsschnittdaten

Ein Problem bei der Verwendung von administrativen Daten besteht darin, dass bei der verwaltungsinternen Aktualisierung von Daten oftmals die Daten des jeweiligen Vormonats überschrieben werden und damit auch verloren gehen. Dies ist in der Regel eine technische Eigenheit des DV-Systems. Derzeit müssen Längsschnittdaten auf der Basis von administrativen Daten durch Evaluatoren/-innen umständlich erzeugt werden. D.h. die Geschäftsdaten der verschiedenen relevanten Städte und Landkreise werden monatlich heruntergeladen und archiviert, solange bis extern ein Längsschnitt erzeugt werden kann.[86] Auch hier würde es interne Lösungsmöglichkeiten geben. Aus Sicht der Evaluatoren/-innen wäre es wünschenswert, wenn Längsschnittdaten in einer verwaltungsinternen Datenbank selbst erzeugt würden. Es müsste also eine entsprechende gesetzliche Regelung geben, die diese Verläufe festlegt. Der Aufbau solcher internen Datenbanken würde einerseits der Kommune selbst ohne großen Aufwand möglich sein und damit aufwändige Datenaufbereitungen der Evaluatoren/-innen ersparen und andererseits die Möglichkeit verwaltungsinterner Evaluationen erhöhen.

---

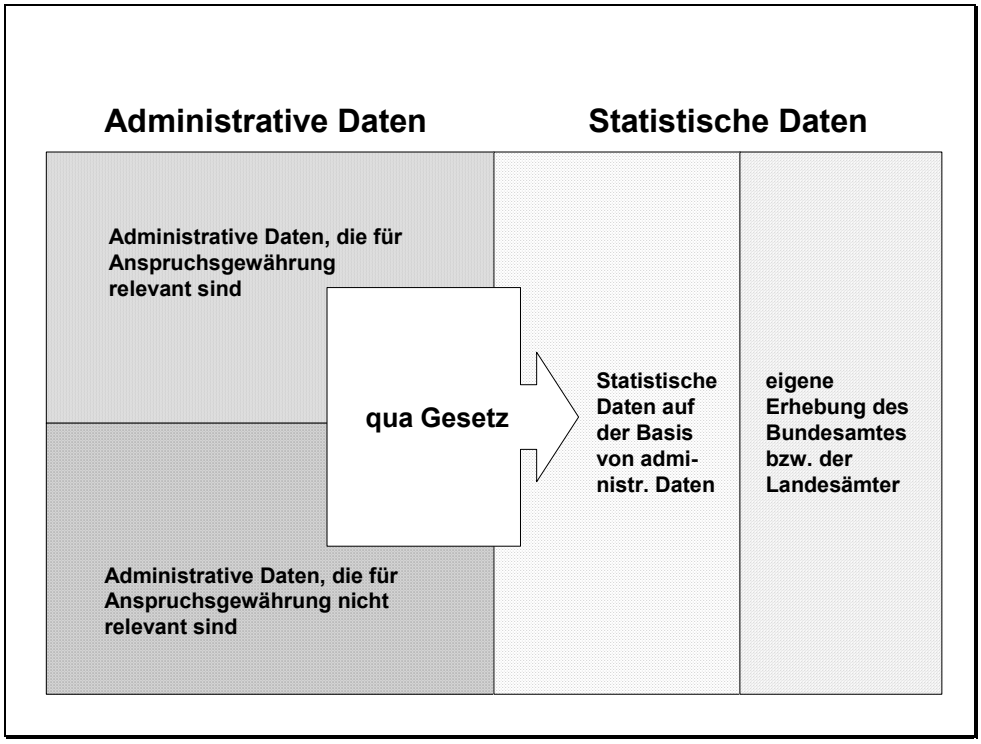
86 Dieses Vorgehen wurde bspw. für die Erzeugung von Längsschnittdaten bei der Evaluation der Pauschalierung der Sozialhilfe (PASO) in NRW von Univation e.V. gewählt, aber auch bei anderen Evaluationen durchgeführt.

4.3 Statistische Daten

Unter **„statistischen Daten“** verstehen wir solche, die durch Unternehmen, Behörden und Organisationen für Zwecke der amtlichen Statistik bereitgestellt werden oder die durch die statistischen Ämter selbst erhoben werden.

Es gibt zwischen den administrativen und den statistischen Daten große Überschneidungen. Viele der statistischen Daten basieren auf administrativen Daten. Letztere werden in einigen Fällen der Öffentlichkeit vom Statistischen Bundesamt in aggregierter Form bereitgestellt.[87] Des Weiteren werden vom Statistischen Bundesamt eigene statistische Erhebungen durchgeführt.

Abbildung 19: Zusammenhang von administrativen und statistischen Daten



Quelle: eigene Darstellung

Zu statistischen Datenquellen mit Relevanz für Evaluationen im Kontext der Armuts- und Reichtumsberichterstattung gehören: Sozialhilfestatistik, Wohngeldstatistik, Krankenkassen- und Pflegestatistik, Kinder- und Jugendhilfestatistik, Arbeitslosenstatistik, Asylbewerberleistungsstatistik, Kreditstatistiken, Einkommenssteuer-

87 Für eine Übersicht über prozessproduzierte Statistikdaten aus Verwaltungen, vgl. Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001), S. 93ff.

statistik und Bildungsstatistik.[88] Die allgemeine Zugänglichkeit zu den von der amtlichen Statistik erhobenen Daten ist sehr gut. Die Datengrundlage (Definitionen, Abgrenzungen) ist gut dokumentiert und nach außen hin transparent. Statistische Daten sind weitgehend als glaubwürdig und neutral anerkannt, da das Statistische Bundesamt ohne Interesse an speziellen Evaluationsfragestellungen diese Daten zur Verfügung stellt. Spezifizierte, ggf. einschränkende Aussagen über die Datenqualität werden vom Bundesamt mit den Daten zusammen zur Verfügung gestellt. Die statistischen Daten beruhen auf einer breiten, d.h. repräsentativen bzw. Vollerhebungs-Datenbasis. Wenn nach professionellen Kriterien erhoben und aufbereitet wird, erlaubt sie den Einsatz unterschiedlichster statistischer Verfahren zur analytischen Auswertung. Die hier erwähnten statistischen Daten stammen meist aus reinen Querschnittserhebungen, d.h. es können mit ihrer Hilfe keine Verläufe analysiert werden. Wie bei administrativen Daten wird auch im Rahmen von statistischen Daten die Perspektive der Erhebungspersonen ausgeblendet. Es werden meist objektive Sachverhalte erfasst und keine individuellen Problemwahrnehmungen. Die Daten sind nicht auf der Ebene der Erhebungseinheit verfügbar. Statistische Daten können in Kombination mit anderen Quellen als wichtige Datenbasis für Evaluationen dienen. Für Evaluationen sind derartige Statistiken zudem als Kontextinformationen sehr wichtig. Des Weiteren können diese Daten zur Validierung von Daten aus eigenen Erhebungen dienen.

Bei statistischen Daten sind häufig nicht die Originaldaten, sondern Sekundäranalysen für evaluative Zwecke verfügbar. Dadurch, dass lediglich die publizierten Ergebnisse verwendet werden können, fehlt ein Vielfaches an Informationen, insbesondere an Tiefengliederung. So gibt es für bestimmte Daten ausschließlich arithmetische Mittel, die aber schwerlich auszuwerten sind, wenn ihre Verteilung nicht normal, sondern z.B. linksschief ist oder im unmittelbaren Bereich des arithmetischen Mittels z.B. Nullfälle enthalten sind.

Die Qualität vieler statistischer Daten ist von der Qualität der zu Grunde liegenden administrativen Daten abhängig, das bedeutet, dass die Datenqualität dieser beiden Datentypen ähnliche Gültigkeits- und Zuverlässigkeitsprobleme hat. Gemäß Experten/-innen-Meinung leidet die Datenqualität der statistischen Daten an den vielen

88 Vgl. auch die Übersicht zu den Themen Erwerbstätigkeit/Einkommen und Vermögen/Verbrauch/Wohnen sowie Gesundheit/Soziale Sicherung, KVI (2001), S. 86f.

Schnittstellen zwischen Datenerhebung und Weitergabe der Daten durch das Statistische Bundesamt. In den Kommunen gibt es unterschiedliche personelle Zuständigkeiten im Rahmen verschiedener Themenbereiche, wie Jugend, Arbeit, etc. sowie verschiedenes Fach- und EDV-Personal. Von der Kommune werden Daten vielfach zunächst zu einer kommunalen Datenzentrale übermittelt, um dann an das zuständige Landesamt für Statistik weitergeleitet zu werden. Dann werden diese Daten wiederum an das Statistische Bundesamt weitergegeben. Es wird berichtet, dass es (minimale) Unterschiede zwischen den Daten der primären Datenerheber/-innen und der amtlichen Statistik gibt. Dies würde bspw. bei Daten wie Einwohnerzahl oder auch Arbeitslosenquote vorkommen. Sollten Kommunen oder Städte feststellen – so Experten/-innen Meinung weiter –, dass gemeldete Daten zu ihrem Nachteil sind während ihnen andere Daten vorliegen, melden sie diese nach und legen damit andere Daten offen. Auch dies ist ein Indiz dafür, dass Ungenauigkeiten auch bei statistischen Daten vorkommen können. Auch Strohmeier et al. (1999, S. 17) berichten von Einschränkungen bei der Datenqualität am Beispiel der amtlichen Statistik in Bezug auf Sozialhilfeempfänger/-innen: Berichtsstichtage weichen von der Vorgabe der amtlichen Statistik ab; verfahrensbedingt gibt es Probleme bei der monatlichen Zuordnung von Personen; die statistische Erfassung wird nicht gemäß der Vorgaben durchgeführt bzw. die Vorgaben sind vielleicht zu unpräzise; bei der Umstellung auf automatisierte Verfahren kann es organisatorische Schwierigkeiten geben.

Ähnlich wie bei den administrativen Daten wird über die kommunale Ebene von wenig Interesse an der Aufbereitung sowie Nutzung von statistischen Daten berichtet.[89] Hier könnte es hilfreich sein, den Verwendungszusammenhang der Daten für die Datenerheber/-innen stärker herauszustellen, immer wieder Feedback zu geben und über exemplarische Auswertungsmöglichkeiten für die kommunale Ebene zu berichten. Zur Verbesserung der Qualität von (wichtigen) statistischen Daten wird von einem Experten vorgeschlagen, zusätzlich zu den Plausibilitätsprüfungen die statistischen Einzelangaben stichprobenartig zu kontrollieren. So kann es bisher theoretisch auch „erfundene“ Angaben geben, die jedoch bei Plausibilitätsprüfungen nicht auffallen würden.

---

89 Vgl. auch bspw. BMGS (1999b) für Ausführungen zur Datenqualität im Rahmen der Steuerung von Sozialhilfe, S. 162f.

Experten/-innen beklagen, dass in einigen Fällen veraltete Erhebungskonzepte verwendet würden. So werden bspw. Angestellte und Arbeiter/-innen unterschieden, nicht jedoch, ob es sich um eine befristete Stelle handelt oder mit welcher genauen Wochenstundenzahl gearbeitet wird. In diesem Zusammenhang wird der Wunsch geäußert, Kategorien der Erhebungskonzepte an soziale und wirtschaftliche Veränderungen anzupassen. Es sollten regelmäßige Beratungsrunden mit den Anwendern/-innen durchgeführt werden, um den Datenkranz weiter zu entwickeln bzw. an eine veränderte soziale Realität anzupassen, da Anwender/-innen an konkreten Fragestellungen arbeiten, die Schwächen von Daten zu Tage bringen können.

Neben den aggregierten Daten aus einzelnen Statistik-Bereichen sind auch so genannte Public-Use-Files mit absolut anonymisierten und Scientific-Use-Files mit faktisch anonymisierten Mikrodaten-Files verfügbar. Für den Kontext der Armuts- und Reichtumsberichterstattung sind dies Daten aus dem Mikrozensus, EVS, ECHP sowie Lohn- und Einkommenssteuerstatistik als Scientific-Use-File und die Sozialhilfestatistik als Public-Use-File.[90] Der Zugang zu den statistischen Mikrodaten ist durch die Schaffung eines so genannten Forschungsdatenzentrums am Statistischen Bundesamt und den Landesämtern weiter verbessert worden.[91] Die hier erwähnten Files sind bisher mit relativ großen Zeitverzögerungen zur Verfügung gestellt worden. Die Mikrodaten-Files für die Sozialhilfestatistik der letzten Jahre werden zukünftig in kürzeren Zeitabständen zur Verfügung gestellt.

Bei den Mikrodaten-Files handelt es sich um repräsentative und konsistente Datensätze, bei welchen die Daten nicht nachbearbeitet werden müssen. Auch bei Hilfen in besonderen Lebenslagen, für Wohnungslose oder Eingliederungshilfen aber auch der Pflegestatistik wäre es wichtig, für Evaluationen nicht nur aggregierte statistische Daten, sondern auch Mikrodaten-Files – also anonymisierte Individualdaten – zur Verfügung gestellt zu bekommen. Außerdem wird von den Experten/-innen ein Angebot von vertieft regionalisierten Daten durch die statistischen Landesämter gewünscht. Auf diese Weise könnte die Nutzbarkeit der Daten für kleinräumigere Programm-Evaluationen gesteigert werden.

---

90 Diese Angaben beziehen sich auf verschiedenste verfügbare Jahrgänge; vgl. Zwick (2003).

91 Es gibt auch noch weitere Forschungsdatenzentren in Deutschland, wovon das am IAB in Nürnberg auch für den Kontext der Armuts- und Reichtumsberichterstattung relevant ist. Siehe hierzu Kapitel 4.5 Verknüpfung verschiedener Datenquellen.

Es wurde auch von den Experten/-innen erwähnt, dass es sinnvoll sei, Mikrodaten für (fast) alle Betroffenen an Stelle von Stichproben zur Verfügung gestellt zu bekommen. Hierbei wäre es besonders wichtig zu vermeiden, dass von einem Datensatz auf eine bestimmte Person geschlossen werden kann.[92] Derzeit wird an der faktischen Anonymisierung von wirtschaftsstatistischen Einzeldaten gearbeitet (vgl. Sturm, 2002). Aus Sicht von Evaluationsexperten wäre es wünschenswert, ein ähnliches Projekt für Sozialdaten zu initiieren.

In Bezug auf die EVS wird als problematisch angesehen, dass die Befragten eine Selbsteinstufung vornehmen (vgl. Hauser/Becker 2001, S. 46-60). Wie bei den administrativen Daten sind auch für statistische Daten Längsschnittanalysen ergiebiger als die Analyse von Querschnittsdaten. Einige der befragten Experten/-innen geben an, dass sie es auch wünschenswert fänden, Inhalte des EVS als Längsschnittbefragung anzulegen, dies jedoch sehr aufwändig wäre.

Gemäß § 7 BundesstatistikG dürfen in besonderen Fällen Bundesstatistiken auch ohne Auskunftspflicht durchgeführt werden.[93] Für die Anwendung dieses Paragrafen müssen zwei wesentliche Bedingungen erfüllt sein: Erstens besteht keine Auskunftspflicht und zweitens dürfen höchstens 10.000 Personen befragt werden. Bei einem hypothetischen Rücklauf von 50 bis 60 % – was als sehr hoch anzusehen wäre – würden nur von 5.000 bis 6.000 Personen Daten vorliegen. Von den bisher nach diesem Gesetz durchgeführten Erhebungen sind für die Armuts- und Reichtumsberichterstattung bspw. die Zeitbudgeterhebung 2001/02 oder Testerhebungen zur statistischen Erfassung von Wohnungslosigkeit in NRW von Relevanz. Dieser Paragraf ist bisher im Rahmen von konkreten Evaluationen nicht genutzt worden.

Auf europäischer Ebene werden Statistiken über Einkommen und Lebensbedingungen in Europa (EU-SILC) ab 2005 das ECHP ersetzen.[94] Das EU-SILC soll in Europa

---

92 Hierfür können augenfälliger Identifikatoren unterdrückt werden, geographische Angaben limitiert, die Anzahl bestimmter Datenelemente reduziert und mit Hilfe von Intervallen und Rundungen gearbeitet werden (vgl. Brady et al. 2002). Vgl. auch KVI (2001), S. 161 ff. sowie S. 285 für den Vorschlag eines so genannten Forschungsdatengeheimnis.

93 Dies gilt für kurzfristig auftretenden Datenbedarf einer obersten Bundesbehörde oder zur Klärung wissenschaftlich-methodischer Fragestellungen auf dem Gebiet der Statistik.

94 Vgl. Europäische Kommission (2002), S. 91. Zum Zeitpunkt der Erstellung dieses vorliegenden Berichtes war noch nicht abschließend geklärt, welche Variablen im Themenbereich *social exclusion* in den EU-SILC Eingang finden werden. Deutschland wird erst ab 2005 nach Start des EU-SILC daran teilnehmen.

als Referenzquelle für die Analyse in den Bereichen Einkommen und soziale Ausgrenzung sowie zur Überwachung der Fortschritte, die bei der Umsetzung von Strategien zur sozialen Eingliederung erreicht worden sind, herangezogen werden. Je nach Ausgestaltung kann hieraus eine neue, für die Evaluation wichtige Datenquelle entstehen.

4.4 Panel-Erhebungsdaten

Unter **„Panel-Erhebungsdaten“** verstehen wir solche, die im Rahmen längerfristig geplanter Wiederholungserhebungen erzeugt werden. Diese basieren normalerweise auf einem fixen Datenkranz und festgelegten Operationalisierungen, wie dies z.B. beim SOEP realisiert wird.

Panel-Daten liefern Verlaufsinformationen und können somit Veränderungen auf Individual- bzw. Haushaltsebene abbilden. Diese Daten sind informationsreicher als die Querschnittsdaten der Verwaltungen und der Statistik, da sich auf Basis dieses Datentyps Zusammenhänge zwischen Ereignissen und ihren Ursachen analysieren lassen. Diese Art von Daten kann einerseits direkt zur Wirkungsanalyse von neuen Programmen oder Programmänderungen sowie andererseits als Vergleichsmaßstab für Daten anderer Quellen dienen.

Für die direkte Programm-Evaluation ist ihre sinnvolle Nutzung jedoch abhängig von der Größe der Maßnahme und der Anzahl der betroffenen Personen im Panel. So können nach Experten/-innen-Meinung bspw. Effekte einer Kindergelderhöhung nur bei einer sehr deutlichen Steigerung (z.B. Verdoppelung) im Rahmen eines Panels zuverlässig beobachtet werden.[95] Außerdem kann von einer sinnvollen Nutzung nur bei flächendeckenden Programmen ausgegangen werden, da ansonsten die Fallzahlen im Panel zu gering sind. Es ist anzumerken, dass es selbst bei flächendeckenden Maßnahmen bei geringem Anteil an Anspruchsberechtigten zu geringen Fallzahlen im Panel kommen kann. In diesem Fall können die Panel-Daten nicht sinnvoll für Evaluationen genutzt werden.

Wie vielfach in der Literatur angemerkt, werden durch das SOEP die oberen und unteren Ränder einer Gesellschaft nur unzureichend erfasst (vgl. Strohmeier et al.,

---

95 Für eine Analyse möglicher Effekte einer Einführung eines einkommensabhängigen Kindergeldzuschlags auf der Basis des SOEP vgl. Otto (2002). Hierbei handelt es sich jedoch um eine Mikrosimulation und nicht um eine Evaluation eines realisierten Programms.

1999, S. 12; vgl. Isengard, 2002, S. 36ff.). Hierzu gehört auch die Gruppe der Sozialhilfebezieher/-innen. Deshalb sind die Fallzahlen der von Armut Betroffenen und Bedrohten für differenzierte Analysen oftmals zu gering. Dem könnte nur durch ein so genanntes „Oversampling“ dieser Ränder begegnet werden. Hierbei wäre es entscheidend, mit welcher Art der Auswahl von Personen dieses Oversampling vollzogen wird, denn die Repräsentativität wird hierdurch entscheidend beeinflusst. Im schlechteren Falle wären zwar vertiefte Informationen über die Zielgruppen der „Armen“ und der von Armut Bedrohten erhältlich, dies jedoch um den Preis einer mangelhaften Repräsentativität. Dies ist für Evaluationen insofern entscheidend, als dann das SOEP in diesem unteren Einkommensrand nicht mehr als Vergleichsmaßstab zu benutzen wäre.

Wenn im Rahmen des SOEP Subzielgruppen differenzierter betrachtet werden sollen, wie bspw. die Gruppe der Aussiedler/-innen, trifft man auf Schwierigkeiten bzw. Grenzen. Die Fallzahlen von Spätaussiedlern/-innen sind zu gering (vgl. Engels et al., 2000, S. 5). Somit sind auch keine repräsentativen Aussagen über Sozialhilfe- oder Wohngeldbezug dieser Teilzielgruppe möglich. Es können lediglich Schätzungen auf Basis dieser Daten vorgenommen werden. Diese sind jedoch als Basis für konkrete Evaluationen kaum zu verwenden. Eine weitere Vergrößerung der Stichprobe des SOEP würde die Nutzbarkeit für Evaluationen erheblich verbessern können. So wären Auswertungen im Hinblick auf regionale Effekte auch eher möglich.

Die Flexibilität der Panels wird u.a. durch die mögliche Aufnahme neuer Fragen gewährleistet. So ist z.B. das Thema Pflegeversicherung in das SOEP neu aufgenommen worden. Diese auch relativ kurzfristige Flexibilität ist jedoch nur für große und langfristige Programme sinnvoll zu nutzen.

Wenn Panel-Daten nicht direkt zur Programm-Evaluation geeignet sind, dann können sie dennoch evtl. Referenzgrößen hierfür liefern. Panel-Daten geben damit ein Referenzszenario von Personen und Haushalten ab, die nicht an einer Maßnahme teilgenommen haben. Wenn z.B. Eingliederungsquoten in den ersten Arbeitsmarkt nach einer Programm-Teilnahme betrachtet werden sollen, ist eine Referenzgröße wichtig. Es kann sinnvoll sein, für einen Vergleichsmaßstab aus Panel-Daten jene herauszuziehen, die Zeiten nach gewöhnlichen Arbeitsplatzwechseln angeben. Auch kleinräumige Umfragen können mit denen aus dem SOEP hinsichtlicht Repräsentativität verglichen werden. Hier würden Daten des SOEP im Sinne von nationalen

Durchschnittswerten verwendet. Auch zur Abschätzung von Nicht-Inanspruchnahmen von Leistungen kann das SOEP (aber auch EVS) genutzt werden. Es sind bspw. einige Studien zur Berechnung von Nicht-Inanspruchnahmequoten der Sozialhilfe auf Basis dieser Daten durchgeführt worden (vgl. Übersicht bei Engels/Sellin, 2001, S. 24).

Angesprochen wurden auch einige Themenbereiche und konkrete Fragestellungen, für die es wünschenswert wäre, sie in das SOEP aufzunehmen. Bei der Inanspruchnahme von sozialen Dienstleistungen wird im SOEP der Sozialhilfebezug erfasst, nicht jedoch, ob Beratungsmaßnahmen in Anspruch genommen wurden oder an Fördermaßnahmen teilgenommen wurde. Ebenso wird die Erwerbsbeteiligung erhoben, aber nicht, ob es sich um öffentlich geförderte Erwerbsbeteiligung handelt. Maßnahmelandschaften, die interessieren könnten, werden in keiner Weise angesprochen.

Im Rahmen des NIEP wurden ca. 2000 Haushalte im unteren Einkommensbereich erfasst (vgl. Kortmann/Sopp/Thum, 2002). Das NIEP wird zunächst vermutlich nicht fortgeführt und ist damit für aktuelle bzw. zukünftige Evaluationen zunehmend weniger direkt zu verwenden. Auswertungen und Erkenntnisse aus dem NIEP können in Zukunft nur noch durch eventuell aufgedeckte Wirkungszusammenhänge in Evaluationen einfließen. Es wird sich dabei voraussichtlich eher um Armutsforschung handeln, die mit den gesammelten Daten durchgeführt werden kann, als um tatsächliche Evaluationen. Wenn sich nicht zu viele Kontextfaktoren in naher Zukunft ändern, dürften die Daten aus dem NIEP auch weiterhin für ex-ante-Evaluationen nutzbar sein.

Im NIEP werden Personen erfasst, die zu Beginn der Erhebung über ein Einkommen verfügten, welches das 1,5fache des jeweiligen haushaltsspezifischen Sozialhilfebedarfs nicht überstieg, bzw. Personen, die aus dieser Situation herausgekommen sind. Hierdurch können auch über das Ende des Sozialhilfebezugs die Lebensbedingungen ehemaliger Bezieher/-innen nach dem Ausstieg aus der Sozialhilfe im Zeitverlauf untersucht werden (vgl. Buhr 2002). Diejenigen, die erst vor kurzem in die Situationen geringen Einkommens geraten sind („Absteiger/-innen“), wurden jedoch nicht erfasst. Jedoch kann auch diese Gruppe für Evaluationen sehr relevant sein.

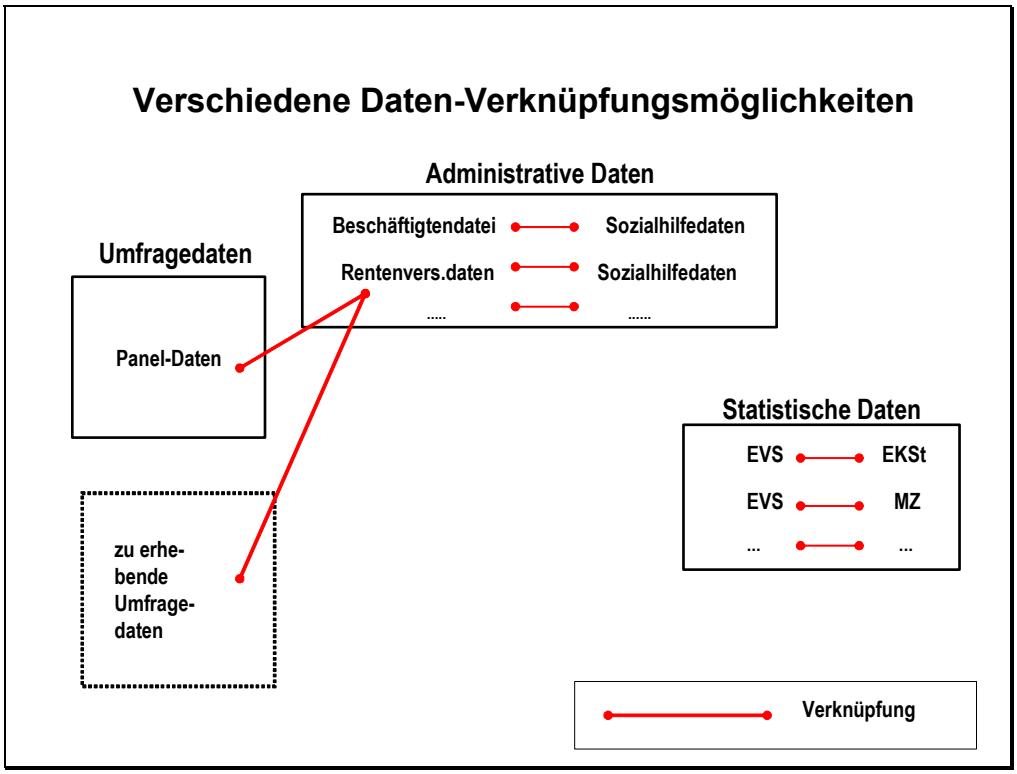
Bilanzierend wird eine Fortführung des NIEP von Experten/-innen als sinnvoll erachtet. Für den unteren Bereich wird vorgeschlagen, noch zusätzliche Teilnehmer/-innen hinzuzunehmen und eine Ergänzungsstichprobe mit Ausländer/-innen durchzuführen. Auch wird angemerkt, dass im Falle einer Fortführung die Stichprobe verdoppelt werden sollte. Dies vor allem vor dem Hintergrund einer relativ großen Panel-Sterblichkeit. Ein Experte merkte an, dass noch einige Erhebungswellen notwendig seien, um die Qualität und Nutzung des NIEP beurteilen zu können. Alternativ wäre auch an eine Vereinigung von NIEP und SOEP zu denken. Auch bei einer Zusammenlegung sollte eine Zusatzbefragung für den Niedrigeinkommenssektor durchgeführt werden. Umgekehrt würden durch die Fragen aus dem SOEP interessante Zusatzinformationen zum NIEP geliefert werden können. Außerdem könnten Fragen des NIEP ins SOEP integriert werden, z.B. Fragen zu subjektiven Bewertungen der sozialen Lage. Es sollten die Fragen aus dem NIEP-Fragenprogramm zum überwiegenden Teil erhalten bleiben.

4.5 Verknüpfung verschiedener Datenquellen

Es gibt verschiedene Möglichkeiten, um die Daten zu verknüpfen: Innerhalb eines Programmbereichs können unterschiedliche Datensätze über den gesamten Zeitverlauf miteinander verknüpft werden. So entstehen Längsschnittdaten (vgl. Kapitel 4.2.5).

Des Weiteren können Datensätze aus verschiedenen administrativen Datenquellen miteinander verknüpft werden. Auf der Basis von administrativen Datensätzen ist es auch möglich, Stichproben zu ziehen und mit den so ermittelten Personen – deren zentrale sozio-demographische Daten bereits bekannt sind – Umfragen durchzuführen. Außerdem können auf der Basis von einmaligen oder auch wiederkehrenden Befragungen (Panels) Verknüpfungen zu administrativen Daten der Befragten hergestellt werden.

Abbildung 20: Verknüpfung von Daten



Quelle: eigene Darstellung

Administrative und statistische Daten können auf Teilbereiche von Armut Licht werfen. So können Daten aus der Arbeitsverwaltung Informationen über Erwerbstätigkeiten bzw. Arbeitslosigkeit geben, Daten der Krankenkassen geben Informationen über Gesundheitszustände und Hilfsbedürftigkeiten etc. Mit der Verknüpfung von Daten aus verschiedenen Quellen besteht die Möglichkeit, ein Gesamtbild über die „Armut“ in verschiedenen Lebenslagen einzelner Personen zu zeichnen (vgl. Levecque/ Vranken, 2000). So können – neben der Erhebung von Daten im Rahmen einer konkreten Evaluation – Informationen über die Kumulation von Armut in verschiedenen Lebenslagen und deren Wechselwirkungen auch in Bezug auf konkrete Programme beurteilt werden. Biografische Angaben in Rentenversicherungsdaten, zu anderen Daten hinzugefügt, können für eine Analyse von Ist-Situationen besonders wichtig sein oder aber auch für Matching-Verfahren bei Vergleichsgruppen-Untersuchungen genutzt werden.

Durch die Verknüpfung von verschiedenen Datenquellen können bspw. Längsschnittdaten bzw. biographische Daten erzeugt werden und dies auf eine Weise, welche die

betrachteten Personen kaum bzw. nicht weiter als Datengeber/-innen belastet.[96] Reischmann (2003) führt an, dass zunächst zu überprüfen sei, ob Materialien oder Dokumentationssysteme nicht bereits einen Teil der notwendigen Informationen enthalten und deshalb nicht gesondert für Evaluationen erhoben werden müssen. Hierfür bieten Verknüpfungsmöglichkeiten bestehender Datenquellen Ansatzpunkte und machen diese besser nutzbar. Dies kann auch durch eine Verknüpfung von bestehenden Daten und neu erzeugten Daten geschehen. Seit einigen Jahren wird durch Wissenschaftler die Strenge einiger Regelungen des „Datenschutz“ kritisiert und größere „Forschungsfreiheit“ diskutiert.[97]

Neue Technologien können diese Transparenz durch ihre besseren Möglichkeiten, Datensätze, Evaluation und Programmsteuerung zu verknüpfen, herstellen. Gleichzeitig können bei unsachgemäßer Verwendung die Risiken für einzelne Personen – im Sinne eines gläsernen Menschen oder sozialer Stigmatisierung – größer werden. Der Datenschutz, der für von Armut Betroffene besonders relevant sein kann, da er zu einem Minimum an Lebenssicherheit beiträgt, muss beachtet werden.[98] Deshalb ist der Evaluationsstandard „F2 Schutz individueller Rechte“ bei Datenverknüpfungen besonders relevant. So wird bspw. in vielen Fällen nicht davon ausgegangen, dass eine bestimmte Geschlechtszugehörigkeit zu besseren oder schlechteren Ergebnissen der Programm-Teilnahme führt. Dennoch gibt es bspw. Berufsgruppen, in denen eher Männer oder eher Frauen arbeiten. Innerhalb einer Evaluation müssen diese Aspekte berücksichtigt und es muss sich für oder gegen eine Erhebung des Geschlechts der Teilnehmenden entschieden werden. Ähnliches gilt für Angaben über das Alter der Teilnehmenden. Ältere Arbeitnehmer/-innen haben in bestimmten Berufen besondere Schwierigkeiten, eine Anstellung zu finden. Die Nationalität bzw. Zugehörigkeit zu einer Minoritätengruppe eines Programms kann ebenfalls Auswirkungen auf den Erfolg der Teilnahme haben. Bei diesen sensiblen Daten sollten die Evaluatoren/-innen nur dann Erhebungen bzw. Auswertungen bzgl. dieser Aspekte durchführen, wenn sie zur Beantwortung der zentralen Evaluationsfragestellungen

---

96 Siehe auch DeGEval-Standard D1 „Angemessene Verfahren“ im Anhang.

97 Vgl. hierzu das Memorandum von Hauser/Wagner/Zimmermann (1998) und Reaktionen hierauf bspw. Metschke/Wellbrock (1999).

98 Vgl. Bundesministerium für Arbeit und Sozialordnung, 1999a, S. 85.

unabdingbar und mit den Datenschutz- und Persönlichkeitsrechten zweifelsfrei vereinbar sind.

Wie bei der allgemeinen Verwendung von personengebundenen/auf Personen rückführbaren Daten im Rahmen von Evaluationen gibt es auch bei Datenverknüpfungen Datenschutzaspekte, die besonders beachtet werden müssen. „... die derzeitige Philosophie der Produktion amtlicher Statistiken [erlaubt] nur so genannte "Einbahnstraßenregelungen" in die statistischen Ämter hinein [...]. Es gibt also zur Zeit keine Möglichkeit, Einzelinformationen aus dem Bereich der amtlichen Statistik wieder in den Verwaltungsvollzug hinaus zu transferieren“.[99] Diese hier zitierte Regelung geht auf das Jahr 1953 zurück. Gemäß dem Volkszählungsurteil aus den 80er Jahren dürfen Daten, die für statistische Zwecke erhoben wurden und nach der gesetzlichen Regelung auch dafür vorgesehen sind, nicht für den Verwaltungsvollzug eingesetzt werden.[100] Auch ein für viele Register sowie Dateien geltendes Personenkennzeichen oder Substitut wäre ein Schritt in die Richtung, den Bürger in seiner Gesamt-Persönlichkeit zu registrieren. Dies widerspricht jedoch dem Recht auf Persönlichkeitsschutz. Es gibt somit grundsätzliche ethische und demokratietheoretisch begründete Vorbehalte gegenüber der Verknüpfung von verschiedenen Datenquellen.

Umfangreiche Vorarbeiten zur Frage des Datenschutzes sind in Bezug auf den Einsatz von pädagogischen und psychologischen Tests geleistet worden. Diese sind kodifiziert in den *Standards for Educational and Psychological Testing*, die mittlerweile in der fünften Ausarbeitung vorliegen.[101] Diese enthalten ein eigenes Kapitel zum Thema *Testing in Program Evaluation and Public Policy* (S. 163-169). Besonders die Standards 15.7 (Offenlegung der Testzwecke) und 15.10 (informed consent) zeigen auf, wie Themen des Datenschutzes in diesen Bereichen geregelt sein sollen. Sie könnten als „Ergänzungsstandards“ zu den Evaluations-Standards der DeGEval gesehen werden. Dies gilt insbesondere für die Fälle, in denen Daten von den Evaluatoren/-innen selbst erhoben werden.

---

99 Eichler, 2000, S. 87.

100 Für eine zusammenfassende Darstellung der Relevanz des Volkszählungsurteils für die Datenverknüpfung vgl. Eichler (2000), S. 87-89.

101 AERA/APA/NCME 1999. Diese sind von der American Educational Research Association, der American Psychlogical Association und dem National Council on Measurement in Education gemeinsam verabschiedet worden. Sie sind von den entsprechenden deutschen und anderen europäischen Fach- und Wissenschaftsgesellschaften anerkannt.

Abbildung 21: Ergänzende Standards zu Datenschutz und Persönlichkeitsrechten

Standard 15.7

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to identify and monitor their impact and to minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.

Comment*:* Mandated testing programs are often justified in terms of their potential benefits for teaching and learning. Concerns have been raised about the potential negative impact of mandated testing programs, particularly when they affect important decisions for individuals or institutions. To the extent possible, students, parents, and staff should be informed of the domains, on which the students will be tested, the nature of the item types, and the standards for mastery. Effort should be made to document the provision of instruction in tested content and skills, even though it may not be possible or feasible to determine the specific content of instruction for every student. An example of negative impact is the use of strategies to raise performance artificially.

Standard 15.10

Those who have a legitimate interest in an assessment should be informed about the purposes of testing, how tests will be administered and scored, how long records will be retained, and to whom and under what conditions the records may be released.

Comment*:* Those with a legitimate interest may include the test takers, their parents or guardians, or personnel who may be affected by results (teachers, program staff).

4.5.1 Administrative und statistische Daten

Seit 1998 werden Sozialdaten zwar nicht direkt miteinander verknüpft, jedoch nach §117 BSHG zur Ermittlung des Parallelbezugs von Leistungen miteinander abgeglichen.[102] Daten der Sozialhilfeträger werden bei dem Datenabgleich an die Datenstelle der Rentenversicherungsträger übermittelt und diese meldet die Namen der Sozialhilfeempfänger/-innen zurück, die Renten oder Arbeitslosengeld bezogen haben.[103] Der automatisierte Sozialhilfedatenabgleich nach §117 BSHG erscheint dem Bundesbeauftragten für Datenschutz im Sinne seiner aufgezeigten Ergebnisse und Präventivwirkungen vertretbar. Er fordert jedoch gleichzeitig eine weitere Beobachtung, ob der Datenabgleich auch weiterhin im Interesse des Gemeinwohls ge-

102 Für ein kurze Übersicht zum Datenabgleich in anderen Staaten vgl. Eichler 2000, S. 91-92.

103 Vgl. Engels/Sellin, 2000.

rechtfertigt ist.[104] Die Sozialämter haben zur Verhinderung von Missbrauch oder zur Erlangung sonstiger für sie wichtiger Informationen Interesse an einem Datenabgleich.[105] Zu den für die Sozialämter weiterhin interessierenden Informationen gehören jene der Kfz-Meldestellen, über Bezug von Wohngeld, Bezug von Unterhaltsleistungen, Informationen der Gewerbemeldestellen, Einwohnermeldeamt, Ordnungsamt, Ausländerbehörden, Bauämter, Katasterämter, Finanzämter, Krankenkassen, Arbeitsämter.

Es könnten Erfahrungen des Datenabgleichs für die Datenverknüpfung genutzt werden. Zugleich werden durch diesen Datenabgleich validere Daten erzeugt. Es ist zu überlegen, wer die Datenverknüpfungen vornimmt. So ist zu vermuten, dass der Datenabgleich zwischen Verwaltungen größere Risiken des Verstoßes gegen den Datenschutz birgt als die Verknüpfung von Datensätzen durch Evaluatoren/-innen, wobei auch bei diesen in vorstellbaren Grenzfällen Interessenkonflikte mit dem Datenschutz entstehen können. Die bisher in der Berufsethik getroffenen Regelungen weisen zwar in diese Richtung, sind jedoch vergleichsweise allgemein formuliert und rechtlich nicht verbindlich.[106]

Administrative Daten müssen zunächst bearbeitet werden, um aus verschiedenen Datensätzen einen analytischen Datensatz zu schaffen. Nach Goerge/Lee (2002) sind Standardisierung der Daten vielfach notwendig, da bspw. „männlich“ in einem Datensatz „M“ und in einem anderen „1“ genannt wird. Auch *missing values* müssen als solche erkannt werden. Wenn Wort für Wort verglichen wird, passen auf Grund von Tippfehlern oder Abkürzungen weniger Datensätze zueinander. Zudem sollte die Validität und Genauigkeit von Verknüpfungen überprüft werden. Eine Angleichung der Struktur verschiedenster Datensätze würde eine Verknüpfung von solchen Daten erleichtern.

Im Rahmen des ersten Armuts- und Reichtumsberichts wurden bereits Möglichkeiten der Verknüpfung von EVS und Einkommensteuer-Statistik diskutiert. Im Rahmen des EVS werden vor allem Personen mit mittleren und kleineren Einkommen befragt. Durch die mikroanalytische Verknüpfung von EVS und Einkommensteuer-Statistik

---

104 Vgl. BfD, 2001, S. 137.

105 Vgl. BMGS 1999a, S. 85f. für Informationsbedarf der Sozialämter.

106 Vgl. z.B. die Standards für Evaluation im Anhang; KVI-Gutachten (2001).

sollen vor allem Informationen über Personen mit höheren Einkommen geschätzt werden. Auch die Verknüpfung von MZ und Sozialhilfestatistik wurde angedacht. Diese Arten der Verknüpfungen wären synthetischer Natur und keine Verknüpfung von tatsächlichen Datensätzen einer Person.

Eine Verknüpfung von Sozialhilfe- und Rentenversicherungsdaten wird allein dadurch erschwert, dass es sich um unterschiedliche Verantwortungsbereiche handelt. Es wäre einfacher, Daten zu verknüpfen, die innerhalb eines Verantwortungsbereiches wie des Statistischen Bundesamts liegen. So käme eher eine Verknüpfung von bspw. Daten über Sozialhilfeempfänger/-innen und Daten aus der Pflegestatistik in Betracht.

In der amtlichen Statistik wird derzeit jedem Fall eine Kennnummer zugeordnet. Diese Kennnummer wird bei Rückfragen durch die statistischen Ämter genutzt. Sie könnte jedoch kaum für eine Verknüpfung von Daten verwendet werden, da grundsätzlich in der Statistik Hilfsmerkmale nur solange wie notwendig mit den Erhebungsmerkmalen verbunden bleiben sollen.[107] Es gibt jedoch auch Möglichkeiten, über andere Merkmale wie Geburtsdatum, Geschlecht, Adresse, Beruf etc. verschiedene Datenquellen miteinander zu verknüpfen. Es bleibt ein Restrisiko der nicht eindeutigen Zuordnung (vgl. Brady et al., 2002). Gleichzeitig sind diese Verfahren sehr aufwändig und erfordern insbesondere bei großen Fallzahlen einen erheblichen Ressourceneinsatz.

In der Arbeitsmarktpolitik waren bisher viele Informationen zu einer Person in verschiedenen Datensätzen enthalten: Diese werden seit kurzem personenbezogen zusammengeführt. Bisher tauchte eine Person mit längerer „Karriere“ in verschiedenen Datensätzen auf, ohne dass hier ein Abgleich durchgeführt wurde. Die verschiedenen Datenquellen werden weiterhin nebeneinander existieren, es gibt jedoch seit kurzem eine so genannte Maßnahme-Teilnehmer-Grunddatei.[108] Diese Datei enthält zentral verfügbar Individualdatensätze über Maßnahmeteilnehmer/-innen und über Maßnahmen, an denen diese teilgenommen haben. Zu den Maßnahmen-Daten gehören jene aus CoSach ABM (einschließlich SAM, Lohnkostenzuschüsse und Überbrückungsgeld), freie Förderung, Förderung der beruflichen Weiterbildung

---

107 Vgl. Eichler, 2000, S. 86.

108 Vgl. Kellner, 2002.

(einschl. Trainingsmaßnahmen) sowie ESF-BA. Nach Kellner (2002) können evtl. zu einem späteren Zeitpunkt auch weitere Maßnahmen-Daten, wie solche aus der beruflichen Rehabilitation, in die Datei integriert werden. Die enthaltenen Datensätze sind über Identifizierer mit weiteren administrativen Daten verknüpfbar, z.B. mit der Arbeitslosendatei und mit der Beschäftigtendatei, wenn die Sozialversicherungsnummer bekannt ist. Auf diese Weise werden auch Daten für die Zeit vor und nach der Maßnahme-Teilnahme, also Verlaufsdaten, erschlossen. Für Befragungszwecke können auch Name, Adresse und Telefonnummer ermittelt werden. Zur Nutzung dieser Maßnahme-Teilnehmer-Grunddatei wird eine Genehmigung des BMWA nach § 75 SGB X benötigt.[109] Bei dem Aufbau solch einer Grunddatei handelt es sich um Basisarbeit, die nicht nur für aktuelle Evaluationen hoch relevant ist, sondern eine grundsätzliche Daten-Infrastruktur für zukünftige Evaluationen im Bereich der Arbeitsmarktpolitik zur Verfügung stellt. Dieses Vorgehen dürfte dadurch erleichtert worden sein, dass es sich um unterschiedliche Datenquellen in ein und derselben Behörde handelt.

Wenn im Kontext der Armuts- und Reichtumsberichterstattung Datenverknüpfungen angedacht werden sollten, wären Verknüpfungen zwischen Rentenversicherungsdaten und weiteren Daten einfacher zu gestalten als bspw. mit Daten aus dem Bereich der Krankenversicherung. Die Daten aus der Krankenversicherung sind ungleich komplizierter und die Zahl der betroffenen Akteur/-innen ist wesentlich größer. Im Rahmen des Alterssicherungsberichts der Bundesregierung sind Rentenversicherungsdaten für Verknüpfungszwecke bereits genutzt worden.

Es ist auch denkbar, dass Datensätze nur sehr dezentral verfügbar sind und somit der Aufwand, um diese zusammenzuführen, größer wäre als der für eine eigene Erhebung notwendige. Bei Daten über Kindertagesstätten könnte dies zutreffen. Des Weiteren ist diese Entscheidung auch von der Stichprobengröße abhängig. Auch müssen die verknüpften Daten entsprechenden statistischen Qualitätsmaßstäben gerecht werden, da ansonsten die Daten nicht für die angedachten Verfahren genutzt werden können.

---

109 Über eine Anonymisierung der Daten für Evaluationszwecke durch IAB-Externe wird nachgedacht. Vgl. Kellner (2002).

4.5.2 Verknüpfung von prozessgenerierten Daten mit Umfragen

Da administrative Daten relativ wenige Informationen bspw. über Familienprozesse oder persönliche Netzwerke enthalten, können komplementäre Befragungen diese zusätzlich erforderlichen Daten liefern. So wird auch im KVI-Gutachten gefordert,

> *„... Vorteile der Daten aus der bisherigen Umfrageforschung (Theoriebezug, Messsensibilität, Aktualität der Fragestellungen und des Datenzugangs, multivariate Mikrodatenanalyse, Längsschnitt) mit den Vorteilen der Daten der amtlichen Statistik (Vollerhebungen bzw. große Stichproben, Kohortenserien, regionale Gliederung, kleine Gruppen, zuverlässige historische Vergleiche, Kontextbezug zu Haushalten, Wohnbezirken, Wohngemeinden und Arbeitsstätten, Nutzbarkeit als Hochrechnungsrahmen) verknüpfen zu können."*[110]

Und weiter:

> *„In den Fällen, in denen eine Übermittlung von persönlichen Sozialdaten unbedingt erforderlich ist (z.B. für die im Rahmen des Modellvorhabens vorgesehenen Interviews und Hilfeempfängerbefragungen durch die wissenschaftliche Begleitung), ist die vorherige Einwilligung der Betroffenen einzuholen. Unter diesen Voraussetzungen bestehen gegen eine Übermittlung von Sozialdaten keine Bedenken und ist eine Genehmigung nach § 75 Abs. 2 SGB X nicht erforderlich."*[111]

Diese Regelung wird bei einer Hilfeempfänger/-innen-Befragung im Rahmen der Evaluation eines Modellprojekts in NRW derzeit praktiziert. Die zu Befragenden erhalten bzgl. des Interviews eine schriftliche Ankündigung vom Sozialamt und werden nur im Anschluss an ihre Zustimmung befragt. Diese Art der Verknüpfung ist datenschutztechnisch somit relativ einfach zu realisieren. Verknüpfungen dieser Art können zudem relativ flexibel und kurzfristig durchgeführt werden.

4.5.3 Verknüpfung von Umfragedaten mit prozessgenerierten Daten

Auch umgekehrt kann vorgegangen werden: Daten aus Umfragen werden mit administrativen Daten verbunden. Derzeit arbeitet das MPI, Berlin, zusammen mit dem FIAB an einer Verknüpfung der Daten aus der Beschäftigtendatei mit den Er-

---

110 KVI-Gutachten, 2001, S. 25.

111 Siehe MASQT Datenschutzerklärung, in: Booklet für die Hilfeberechtigtenbefragung im Modellprojekt Pauschalierung in der Sozialhilfe 2001. Siehe auch § 118BSHG, wie schon oben zitiert.

hebungsdaten aus der German Life Course Study (GHLS) gearbeitet. Dieses Forschungsvorhaben steht nicht im direkten inhaltlichen Kontext zur Armuts- und Reichtumsberichterstattung und beinhaltet auch keine konkrete Programm-Evaluation. Dennoch sind die derzeitigen Erfahrungen mit der Verknüpfung von Umfrage-Daten und administrativen Daten auch für zukünftige Evaluationen von Bedeutung. Von den Befragten haben ca. 80% der Verknüpfung ihrer Daten zugestimmt (*informed consent*). Dies ist eine absolut notwendige Voraussetzung für dieses Verfahren. Von den Personen, die zugestimmt haben, kannten wiederum ca. 25% ihre Sozialversicherungsnummer, die idealerweise hierbei als Identifikator für die Verknüpfung dient. Bei denjenigen, die ihre Sozialversicherungsnummer nicht angeben konnten, besteht die Möglichkeit, die Daten über andere Angaben, wie das Geburtsdatum und Geschlecht, miteinander zu verknüpfen. Diese Art von Verfahren ist dann problematisch, wenn Daten zur Identifikation genutzt werden, die jedoch in einem späteren Schritt eigentlich zum Vergleich der Daten aus den verschiedenen Datenquellen genutzt werden sollten (methodisches Problem). Es wurde berichtet, dass die Sozialhilfenummer als Identifikator problematisch ist, da diese Nummer auch den Anfangsbuchstaben des Namens enthält. Der Name kann sich jedoch – insbesondere bei Frauen – geändert haben und wird bei der Sozialversicherungsnummer nicht entsprechend angepasst.

Im Rahmen des SOEP wird die Sozialversicherungsnummer bisher nicht erhoben, dies wird jedoch derzeit angedacht.[112] Es wäre an Anreize zur Einwilligung zur Verknüpfung von Panel-Teilnehmer/-innen zu denken, wie Gratifikationen oder Lose für wohltätige Zwecke. Durch diese Anreize würde das Besondere an einer solchen Einwilligung betont werden. Wenn Panel-Teilnehmer/-innen nach einer Einwilligung zur Verknüpfung gefragt werden, kann Misstrauen hervorgerufen und schlechtestenfalls die Teilnahme aufgekündigt werden, was wiederum Effekte auf die Qualität der eigentlichen Umfrage hätte.

4.5.4 Abschließende Überlegungen

Die Verknüpfung von Daten ist im Spannungsfeld verschiedener Aufgaben des Staates und der Grundrechte weiterhin zu diskutieren. Im Hintergrund steht die Datenschutzdebatte der 80er Jahre mit dem Unterschied, dass zwischenzeitlich durch

112 Vgl. Schupp/Wagner 2002, S. 171f.

elektronische Datenverarbeitung die technischen Möglichkeiten der Verknüpfung von Daten um ein Vielfaches gestiegen sind. Insofern hat sich der mögliche Nutzen der Verknüpfung von personengebundenen/auf Personen rückführbaren Daten für Wissenschaft und Politik potenziert, gleichzeitig aber auch die Gefahr des Missbrauchs und damit die Verletzung des Persönlichkeitsschutzes.

Neue Verknüpfungen von Daten, insbesondere wenn dadurch auf längere Zeit verfügbare Datensätze entstehen, sollten dieses Spannungsfeld ständig im Blick haben und ihm durch eine Abwägung zwischen der Notwendigkeit, Daten für Politik und Forschung bereitzustellen, und den Persönlichkeitsschutz zu wahren, gerecht werden.

Eine Möglichkeit könnte in Wenn-Dann-Lösungen bestehen: Wenn mit Hilfe der Datenverknüpfung die Möglichkeiten der Datennutzung verbessert werden, dann müssen gleichzeitig auf der anderen Seite die Sicherungsmechanismen zur Bewahrung der Datenschutzes ausgebaut werden. Gleichzeitig wäre es eine Möglichkeit, ständig zu hinterfragen, ob der Nutzen der Verknüpfung der Daten wirklich angemessen hoch ist. Eine Evaluation von Programmen benötigt Daten, die sich über die Fragestellungen auf Ziele eines Programms beziehen, insofern sind die Verwendungsmöglichkeiten und der Nutzen von vorhandenen Daten in Abwägung zum Aufwand der Verwendung und der Gefahr des Datenmissbrauchs präzise abzuschätzen.

Psychologische Untersuchungen und Tests dürfen nur mit Einwilligung der Beteiligten durchgeführt werden. Eine Möglichkeit besteht darin, diese Regelung auf die Verknüpfung von Daten zu übertragen, so dass Verknüpfungen nur mit Einwilligung der Beteiligten durchgeführt werden.[113]

Die Diskussion über die Nutzung insbesondere statistischer und administrativer Daten und ihrer systemüberschreitenden Verknüpfung hat gerade erst begonnen. In dem Maße, in dem Veränderungen und Vorteile bei Zielgruppen (Outcomes) zu den vorrangig betrachtenden Wirkungen bei Programmen und politischen Maßnahmen werden, steigt die aufgezeigte Spannung zwischen dem öffentlichen Interesse an aussagekräftigen Daten und dem grundgesetzlich garantierten Interesse am Schutz der Persönlichkeit und der Privatsphäre an. Da die technologische Entwicklung in den Möglichkeiten der DV-gestützten Datensysteme in den kommenden fünf bis

---

113 *Informed consent*-Prinzip; siehe das geschilderte Beispiel MPI/IAB.

zehn Jahren noch erhebliche Produktivitätssteigerungen verspricht, wird das Thema um so dringlicher. Es ist – wie ein Blick in das Kapitel 2 dieses Berichtes mit den unterschiedlichen Evaluationsmodellen und ihren Positionen zum Umgang mit Werten deutlich macht – ein Kernthema der Evaluationstheorie selbst.

4.6 Weitere Datenlücken

Im Rahmen der Expertenbefragung zu dem Thema Datenlage wurde auch nach Themenkomplexen bzw. Lebenslagedimensionen gefragt, zu denen besonders wenige Daten vorliegen.[114] Einige Anmerkungen wurden im Zusammenhang mit den obigen Datentypen besprochen. Einzelne weitere Anregungen werden hier im Folgenden aufgeführt und sind sicherlich nicht vollständig. Zudem ist darauf hinzuweisen, dass Anforderungen an die Datenlage immer von konkreten Programmen und deren Evaluation abhängig sind. Dieser Zusammenhang konnte bei der Befragung jedoch nicht hergestellt werden, da mögliche Evaluationsgegenstände und -zwecke im Kontext der Armuts- und Reichtumsberichterstattung nicht festgelegt sind.

Bei Stichproben-Erhebungen muss fast immer eingeschränkt werden, dass die Ränder der gesellschaftlichen Schichtung kaum in repräsentativer Weise vertreten sind. Dies ergibt sich einerseits aus der mathematischen Logik der Stichprobenziehung: Kleine Untergruppen brauchen eine unvergleichlich größere Teil-Stichprobe als große Hauptgruppen, um im statistischen Sinne „repräsentativ“ vertreten zu sein. Zum anderen ergibt sich dies wegen lückenhafter Erfassung dieser Personen in Dateien, welche die Grundgesamtheit enthalten: Personen, die in Einrichtungen leben, wie z.B. im Justizvollzug, in Kasernen oder Pflegeheimen, sind ebenso wie Wohnungslose kaum in administrativen Daten und noch viel weniger in Umfragedaten vertreten. Da für diese Zielgruppe kaum auf bestehende Daten bzw. Datensysteme zurückgegriffen werden kann, müssen vorrangig eigene Erhebungen durchgeführt werden und es liegen kaum Vergleichsdaten vor.

Es wird von mehreren befragten Experten/-innen weiterer Datenbedarf zum Querschnittsthema Migration konstatiert. Hierzu würden Informationen, wie Herkunftsland der betreffenden Personen – durch doppelte Staatsangehörigkeiten teilweise schwierig zu ermitteln – Ausgrenzung bzw. Integration sowie Gesundheitsversorgung

---

114 Für Vorschläge zu einer Verbesserung der Datenlage im Zusammenhang Aspekten der Einkommensverteilung vgl. Hauser/Becker (2001), S 183f.

dieses Personenkreises gehören. Erhebungen stehen vielfach Sprachprobleme im Wege. Oft sei das Geschlecht nicht zweifelsfrei aus dem Vornamen zu rekonstruieren. Auch würde es über die Zielgruppe der Menschen mit Behinderung (in Werkstätten) zu wenige aussagekräftige Daten geben, dies sei u.a. durch die unterschiedlichsten Kostenträgerstrukturen bedingt. Diese Zielgruppe habe ein sehr niedriges Einkommen. Des Weiteren sei funktionaler Analphabetismus wenig dokumentiert. Auch die PISA-Studie würde hierzu nicht die notwendige Datengrundlage bieten.

Einzelne Projekte im Rahmen großer Programme – wie z.B. Konzeptionen von Beratungsstellen – seien nicht ausreichend dokumentiert. Da diese Informationen jedoch Teil einer Beschreibung des Evaluationsgegenstandes (siehe Standard G1) sein sollten, müsste auch hier eine stärkere Verpflichtung zur Dokumentation eingeführt werden.

Über die Lebenslagendimension „Wohnen“ sei relativ wenig im Kontext der Armuts- und Reichtumsberichterstattung bekannt, beklagte ein Experte. So gäbe es Daten über den sozialen Wohnungsbau und Angaben über Miethöhen von Sozialhilfeempfänger/-innen, es sei jedoch sehr wenig über die Wohnungsqualität und deren Zumutbarkeit bekannt.[115] Es gibt zwar eine Wohnungsstichprobe, die Auskunft über Wohnraumversorgung und Qualität von Gebäuden gibt, diese enthält jedoch bisher keine weiteren, angrenzenden Lebenslagendimensionen.

Ein weiteres Themenfeld, das beleuchtet werden sollte, sei die Überschuldung. Hierüber gibt es nach Meinung der Experten/-innen nur wenige administrative Daten. Die Daten aus dem EVS seien im Hinblick auf die Überschuldungsproblematik unplausibel, da die Verschuldung im EVS ungefähr dreimal so niedrig sei wie auf der Basis von einzelnen Umfragen. In diesem Themenfeld sollte die Datenlage weiter verbessert werden.[116]

Zu innerfamiliären Ungleichheiten können mit der bestehenden Datenbasis kaum Aussagen getroffen werden, da nur haushaltsbezogene Daten vorliegen. Hier empfiehlt sich ein qualitativer Zugang zu dem Thema, wenn auch diesbezügliche

---

115 Vgl. für die Relevanz dieses Aspektes im Rahmen einer konkreten Evaluation Speer (2001).

116 Zu weiterem Forschungsbedarf bezüglich Längsschnittdaten zu Überschuldungssituationen von Haushalten vgl. Korczak (2001), S. 171 f.

Wirkungen betrachtet werden sollten. Auch sei bei der Situation der Alleinerziehenden zu vermuten, dass die Kinder gut versorgt seien, um ihnen eine soziale Ausgrenzung zu ersparen, die Mütter/Väter jedoch um so weniger an Ressourcen zur Verfügung hätten. Auch wurde darauf hingewiesen, dass es bei „ärmeren" Jugendlichen (16-25) auf Grund häufigerer Umzüge und evtl. nur loser Kontakte zu den Eltern besonders schwierig sei, längerfristig Kontakt zu halten und damit Längsschnittdaten zu erhalten.

4.7 Zusammenfassende Darstellung der Datenlage

Die allgemeine Datenlage für die Armuts- und Reichtumsberichterstattung ist bereits vor dem ersten Bericht der Bundesregierung aufgearbeitet worden und zu einem großen Teil dafür auch verwendet worden. In dieser Perspektivstudie wird die Datenlage unter dem Blickwinkel der Evaluation analysiert.

Bestehende Indikatoren und Indikatoren-Systeme erweisen sich als hilfreich für Evaluationen. Je breiter ein Indikatoren-System angelegt ist, desto höher ist die Wahrscheinlichkeit, dass hierauf bei konkreten Evaluationen rekurriert werden kann. Ein Zwang zur Verwendung bestimmter Indikatoren sollte jedoch nicht eingeführt werden, da dieser den Blick für mögliche zu analysierende erwünschte und unerwünschte Wirkungen einengen könnte.

Monitoring und Monitoring-Systeme sowie deren produzierte Daten können – neben ihrer eigentlichen Funktion – sinnvoll Evaluationen unterstützen. Eine wechselseitige Abstimmung und Ergänzung ist wünschenswert.

Viele Fragen, die im Rahmen von Armutspolitik von hoher Relevanz sein können, sind bisher nicht untersucht worden. Für die Beantwortung einiger Evaluationsfragestellungen ist die Datenlage nicht ausreichend. Zudem können aus Kosten-Nutzen-Gründen nur einzelne wichtige Evaluationsfragestellungen bearbeitet werden.

Die erläuterten Evaluationsmodelle sind in unterschiedlichem auf eine breite bereits vorhandene Datenlage angewiesen. Die Quasi-Experimental-gesteuerte Evaluation hat hier ganz deutlich die größten Anforderungen. Andere Modelle arbeiten teilweise multi-methodisch bzw. weniger auf bestimmte Methoden festgelegt, so dass sie flexibler im Hinblick auf unterschiedliche Datenlagen sind. Einige Modelle arbeiten

vorzugsweise mit eigenen Erhebungen und werden durch Datenlücken in ihrer Anwendung nur marginal eingeschränkt.

Administrative Daten sind wichtiger Bestandteil vieler Evaluationen. Es liegen mehrere Probleme mit der Datenqualität und dem Datenzugang vor, die teils nur längerfristig und bei konzertiertem Handeln der beteiligten Verwaltungsebenen behoben werden können.

Daten des Sozioökonomischen Panels oder der Bundesstatistik könnten auch für Evaluationen auf lokaler oder Landesebene stärker genutzt werden. Dies könnte durch entsprechende Servicestellen/Ansprechpartner/-innen für Evaluatoren/-innen in den ausführenden Organisationen begünstigt werden. Letzteres gilt ebenso für die technisch oft sehr komplexe Aufgabe der Verknüpfung von Daten aus unterschiedlichen Quellen.

Existierende Datenquellen könnten durch verschiedene Daten-Verknüpfungsmöglichkeiten noch besser für Evaluationen genutzt werden. Hier gibt es einerseits bereits Erfahrungswerte, andererseits besteht an dieser Stelle noch großes Potential. Die umfangreiche Verknüpfung von Daten mit Bezug zu Personen sollte sorgsam mit ihren möglichen Gefahren abgewogen werden. Der Datenschutz (wie die Schaffung eines Forschungsdatengeheimnisses) sollte nach angemessen Lösungen zur Wahrung des Persönlichkeitsschutzes einerseits, zum Einbezug der verschiedenen Typen relevanter Daten von armen oder von Armut bedrohten Personen andererseits suchen. Auch die Auswirkungen eines unterschiedlich ausgestalteten Datenschutzes sollten evaluiert werden.

5 Perspektiven und Empfehlungen

Die Studie zur „Wirkungsorientierten Evaluation im Rahmen der Armuts- und Reichtumsberichterstattung“ für das Bundesministerium für Gesundheit und Soziales verbreitert die Basis für die Fachdiskussion und bezeichnet einen Ausgangspunkt für die Weiterentwicklung der theoretischen und konzeptionellen Basis von Evaluationen im Umfeld der Armuts- und Reichtumsberichterstattung. In diesem abschließenden Kapitel werden für eine weitere Auseinandersetzung mit dem Thema Thesen und Empfehlungen formuliert.

Ein verstärkter und systematischer Einbezug von Evaluationsergebnissen in die Armuts- und Reichtumsberichterstattung ist für mehrere Adressaten/-innen-Gruppen von Interesse:

- Den am politischen Prozess Beteiligten wird empirisch abgesichertes Wissen über die Qualität und Wirksamkeit der zur Armutsbekämpfung durchgeführten Programme sowie die Auswirkungen von gesetzlichen Regelungen vermittelt. Damit entsteht eine verbesserte Beratungs- und Entscheidungsgrundlage in Bezug auf die Konzipierung und Finanzierung von politischen Maßnahmen. Die für die Kontrolle der öffentlichen Mittelverwendung zuständigen Instanzen, insbesondere Parlamente und Rechnungshöfe, erhalten belastbare Auskünfte.
- Der weiteren Öffentlichkeit, Bürgern/-innen und Steuerzahlern/-innen sowie den Medien werden Informationen über staatliches Handeln zur Verfügung gestellt, so dass sie beurteilen können, ob wichtige soziale Problemlagen mit Erfolg bearbeitet werden. Dies kann auch die Legitimität einer auf die Vermeidung und Verminderung von Armut gerichteten Politik erhöhen.
- Fach- und Führungskräfte, die beruflich in Programmen zur Armutsbekämpfung tätig sind, öffentliche, gemeinwirtschaftliche und private Unternehmen, welche solche Programme ausführen und für deren effektive und effiziente Durchführung Verantwortung tragen, erhalten abgesicherte Hinweise dazu, wie Programme Erfolg versprechend geplant, gesteuert und verbessert werden können.

- Die mittels Evaluationen gewonnenen qualitativen und quantitativen Daten können im Rahmen der Armuts- und Reichtumsberichterstattung genutzt werden. Dies betrifft besonders Ergebnisse zu Programmen, die einen Schwerpunkt in einer der Lebenslagen-Dimensionen haben, z.B. Wohnen oder Bildung, wenn diese Daten – möglichst in bereits aufbereiteter Form – für wissenschaftliche Zwecke freigegeben werden.
- Da das Thema Armutsvermeidung und -verminderung ein Querschnittsthema ist, und damit verschiedenste Politikbereiche und Ressorts berührt sind, kann eine Initialwirkung für einen verstärkten Einsatz wirkungsorientierter Evaluationen in breiten Politikfeldern ausgelöst werden.

Evaluationen können einen maßgeblichen Beitrag zur Armuts- und Reichtumsberichterstattung und längerfristig zur zielgeführten Ausgestaltung von Politik und Programmen der sozialen Integration und zur Vermeidung von Armut leisten. Zu diesem Zweck müssen sie regelmäßig, unabhängig und fachkundig durchgeführt werden. Sie sollen evaluationsfachliche Standards einhalten, Datenquellen optimal nutzen und Ergebnisse klar und verständlich für die interessierte Öffentlichkeit aufbereiten. Die Integration von Evaluation in die Armuts- und Reichtumsberichterstattung kann durch orientierende evaluationsfachliche Leitlinien, ein erweitertes Evaluationsdokumentationswesen, Metaanalysen von Evaluationen und frühzeitigen Beizug von evaluatorischem Sachverstand gefördert werden.

Um dies zur ermöglichen, werden die folgenden Empfehlungen formuliert:

1. Programme, Maßnahmen oder Gesetze mit größeren Kosten bzw. Folgekosten im Umfeld der Politik zur Armutsbekämpfung sollen im Regelfall evaluiert werden. Je nach Zahl der Beteiligten, der Zielgruppenmitglieder, der lokalen Programmstandorte, der Anzahl der an der Umsetzung beteiligten (z.B. föderalen) Ebenen usw. sollen angemessene Budgets für die Evaluation bereitgestellt werden. Besonders intensiv sollen Pilotvorhaben evaluiert werden, deren z.B. bundesweite Übertragung geplant ist. Rechtzeitig vorliegende Pilotevaluationen können maßgeblich zur Erkennung und Minderung von Implementationsdefiziten auf den verschiedenen Ebenen des föderalen Systems beitragen. Gesetze mit erheblichen Folgewirkungen für die betroffenen Be-

völkerungsgruppen sowie komplexen beabsichtigten sowie evtl. unbeabsichtigten Wirkungen sollten – wann immer möglich – eine vorgeschaltete Pilotphase mit interaktiver, dokumentierender und wirkungsfeststellender Evaluation aufweisen. Um Fehlentwicklungen frühzeitig zu erkennen, sind bei besonders kostenintensiven politischen Vorhaben proaktive und klärende Evaluationen wünschenswert/zu empfehlen (s. S. 42ff, S.114ff). Dabei kann Evaluation die politisch und operativ Verantwortlichen dabei unterstützen, „positive“ und klare Zielbeschreibungen zu erstellen, so dass daran anschließende Programme und Maßnahmen operationale und damit auch überprüfbare Ziele setzen können (s. S. 63ff).

2. Evaluationen sollen in Ausschreibungsverfahren mit konkurrierenden Angeboten vergeben werden. Haushaltsrechtlich zu prüfen ist die Möglichkeit, ein Kostendach (Höchstsumme zu veranschlagender Kosten) bekannt zu geben, wenn dadurch die Beurteilbarkeit der Kosten-Leistungsrelationen konkurrierender Angebote bei komplexen Aufträgen verbessert werden kann. Als Grundlage der Vergabeentscheidung sind insbesondere auch evaluationsfachliche Ausweise zu den eingesetzten Verfahren heranzuziehen. In Ausschreibungen sollen der Zweck und die zu beantwortenden Fragestellungen durch die Auftraggebenden möglichst präzise benannt werden. Beim Zweck sollte klar formuliert sein, ob vorrangig die Programmentwicklung/-verbesserung unterstützt oder aber eine Entscheidungsgrundlage geschaffen werden soll. Insbesondere im Falle von neuen, experimentierenden oder Programmen in dynamischem politisch-ökonomischem Kontext ist es empfehlenswert, den Bietern/-innen Spielraum bei der Konkretisierung bezüglich der Fragestellungen einzuräumen. Schließlich sollte aus der Ausschreibung hervorgehen, ob ein solcher Einbezug von weiteren Beteiligten und Betroffenen in die Steuerung der Evaluation erwünscht, möglich oder nicht gewünscht ist. Die Auswahl des Evaluationsmodells und der geeigneten Erhebungsmethoden und eine explizite Begründung dafür sollten hingegen den Anbietern abgefordert werden.

   Bei Ausschreibungen und Bewertungen der Angebote sollten die Auftraggebenden Evaluationsfachwissen nutzen, das sie intern vorhalten oder

extern in Form von Evaluationsberatung hinzuziehen. So könnte ein/-e Evaluationsexperte/-expertin beauftragt werden, der/die sich nicht beworben hat und auch in unmittelbar benachbarten Politikfeldern keine öffentlichen Aufträge durchführt, die Evaluierbarkeit des Gegenstandes mit dem vorgeschlagenen Evaluationsmodell abzuschätzen. Diese Fachperson könnte auch die Auftraggebenden bei der Ausarbeitung der Ausschreibung beraten. Dem sollten die „Standards für Evaluation“ zu Grunde gelegt werden (s. S. 7ff, S. 106ff).

3. Evaluationsanbieter/-innen sollen Evaluationen so planen und steuern, dass die Standards für Evaluation eingehalten werden. Dabei sollen sie im technischen Bericht transparent machen, welche Entscheidungen sie bei der Evaluationsplanung bei kaum/nicht vereinbaren Anforderungen, etwa zwischen den Nützlichkeitsstandards und den Genauigkeitsstandards, wie begründet getroffen haben. Evaluationen sollen auch klare Schwerpunkte ausweisen, sich auf vorrangige Fragestellungen und Programmdimensionen, z.B. den Prozess *oder* die Resultate beziehen und insbesondere nachvollziehbare, für die Politik und weitere Beteiligte relevante Schlussfolgerungen zur Verfügung stellen (s. S. 23ff, S. 115ff). Zu aufwändigeren Evaluationen sollten follow-up Studien durchgeführt werden, welche den Grad der Nutzung der bereit gestellten Evaluationsergebnisse analysieren und Möglichkeiten für eine intensivere Nutzung aufzeigen (s. S 7ff, S. 106ff).

4. In Politikfeldern, die von ausgeprägten Wert- und Interessenunterschieden gekennzeichnet sind – wie es Themen von Armut und Reichtum ganz besonders sind –, ist von besonderer Bedeutung, dass Evaluationen von unabhängigen Fachleuten durchgeführt werden. Diese sollen auch offen legen, wie sich der von ihnen angewandte Evaluationsansatz in Bezug auf soziale Werte verhält. Klammert er Werte aus oder greift er sie im Rahmen der Evaluation ausdrücklich auf? Wenn letzteres: Unterstützt er speziell Werte von einflussschwachen Gruppen, behandelt er alle Werte gleich oder ordnet er Werte nach Wichtigkeit? Schließlich soll im Evaluationsbericht leicht nachvollzogen werden können, wie die Daten gewonnen, wie sie aufbereitet und wie ggf. wertende Schlussfolgerungen

und Empfehlungen aus den Daten hergeleitet und ggf. auf ausgewiesene Wertgrundlagen bezogen sind. So kann der Anspruch, dass Evaluationen faire und ausgewogene Bewertungen vornehmen oder vorbereiten sollen, nachweislich eingelöst werden (S. S. 66ff, S.106ff).

5. Das Wissen bei Evaluatoren/-innen über vorhandene Datenquellen und Datenentstehungsprozesse sowie ihr Zugang zu Datenquellen sollen verbessert werden. Bei Datenlücken können bzw. müssen im Rahmen von Evaluationen eigene Erhebungen durchgeführt werden. Existierende Datenquellen könnten jedoch durch verschiedene Daten-Verknüpfungsmöglichkeiten noch besser für Evaluationen genutzt werden. Hier gibt es einerseits neuere Erfahrungen und andererseits bei der Verknüpfung von verschiedenen Datenquellen und -typen noch ein großes Potenzial (s. S. 131ff).

6. Evaluationen gewinnen an Glaubwürdigkeit, wenn die erstellten Berichte öffentlich zugänglich sind. Besonders öffentlich beauftragte und finanzierte Studien sollten binnen kurzer Frist einer breiten interessierten Öffentlichkeit zugänglich gemacht werden, z.B. durch Verfügbarkeit als herunterladbare Dateien im Internet. Darüber hinaus sollten auch die technischen Berichte einer evaluationsfachlichen und wissenschaftlichen Öffentlichkeit zugänglich gemacht werden. Solche Berichte enthalten z.B. Überlegungen zur Entscheidung für ein Evaluationsmodell, zu Auswahl, Entwicklung und ggf. Validierung von Erhebungsinstrumenten, zur Prüfung, Aufbereitung und Gewichtung von Daten, zu Entscheidungen für bestimmte Auswertungsprozeduren und Darstellungsmethoden, zum Prozess des Interpretierens und der Erarbeitung von Schlussfolgerungen usw.. Im Rahmen von Meta-Evaluationen kann die Güte solcher Evaluationen überprüft werden und Politik sowie interessierte Öffentlichkeit können dabei unterstützt werden, die Gültigkeit und Glaubwürdigkeit der Ergebnisse einzuschätzen (s. S. 122ff).

7. Für eine verstärkte Integration von Evaluationsergebnissen in die Armuts- und Reichtumsberichterstattung wird empfohlen, einen Leitfaden mit Checklisten für die Vorbereitung und Vergabe von Evaluationsaufträgen bereit zu stellen, dem ausgewählte, beispielhaft für den Armuts- und

Reichtumsbericht veranschaulichte Evaluationsstandards zugrunde liegen. Dieser Leitfaden kann sowohl den Auftraggebern wie den Evaluatoren/-innen klare Orientierungen geben.

8. Der Armuts- und Reichtumsbericht sollte zusätzlich zu den bisherigen Anlagen eine geordnete und kommentierte Liste der abgeschlossenen und laufenden Evaluationen enthalten, die Politiken und Programme untersuchen, welche für Themen des Berichtes von Relevanz sind. Auch sollten die Lücken benannt werden, für die keine/nicht genügend Evaluationen vorliegen bzw. welche in Planung befindlich sind.

9. Maßnahmen, Programme und Modellprojekte lassen sich in aller Regel in ihrem Schwerpunkt einer der Lebenslagendimensionen bzw. der Einkommensdimension zuordnen. Wünschenswert wäre, Evaluationsergebnisse in die Darstellungen der Lebenslagendimensionen bzw. der jeweiligen Politikfelder systematisch zu integrieren, darüber hinaus die Aussagen zur Wirksamkeit von Programmen in den Lebenslagendimensionen zu bündeln, d.h. jeweils auf die Breite der vorliegenden Evaluationen zu stützen. Dabei soll auf Übereinstimmungen und auf Divergenzen aufmerksam gemacht werden. Um die hierfür erforderlichen Synthesen durchführen zu können, müssen laufende Evaluationen auf den verschiedenen Ebenen des föderalen Systems frühzeitig identifiziert und deren Ergebnisdarstellungen zusammen getragen werden.

10. Für die Weiterentwicklung der Armuts- und Reichtumsberichterstattung ist es wünschenswert, dass die im Rahmen von Evaluationen erhobenen Daten für die Berichterstattung nutzbar gemacht werden. Perspektivisch können auf diese Weise – als wünschbarer Nebeneffekt von Evaluationen – Datenlücken geschlossen werden. Voraussetzung dafür sind Bestimmungen in den Evaluationsaufträgen: dass – i.d.R. anonymisierte – Rohdaten für wissenschaftliche Zwecke aufbereitet (also auch dokumentiert) werden sollen und weitergegeben – werden können. So können – sobald für die jeweilige Lebenslagendimension genügend aufbereitete Daten vorliegen - Meta-Analysen/Synthesen durchgeführt werden. Hierfür ist es wünschenswert, dass in der Evaluations-Datenbank Schlüsselwörter

vorgesehen sind, welche die Zuordnung der erhobenen Untersuchungsdaten zu den Lebenslagen-Dimensionen erleichtern (vgl. Empfehlung 12).

11. Die systematische Integration von Evaluationen und ihren Ergebnissen in die Armuts- und Reichtumsberichterstattung führt zu einer dichten Beschreibung der Politik zur Armutsbekämpfung in Deutschland. Diese ist insofern nicht vollständig, als nicht alle Programme evaluiert und nicht alle vorliegenden Evaluationsergebnisse zugänglich sind. Es können Lebenslagendimensionen bzw. spezifische Segmente identifiziert werden, zu denen wenig dokumentierte Beschreibungen vorliegen. Darüber hinaus kann festgestellt werden, in welchen Dimensionen/Regionen für welche Zielgruppen es wirkungsorientierte Programme gibt und welche Lücken bestehen, für die neue/veränderte Programme entwickelt werden müssten.

12. Es ist wünschenswert, die Finanzierung von Evaluationen für die Armuts- und Reichtumsberichterstattung frühzeitig zu planen, was es fördert, Evaluation zum selbstverständlichen Bestandteil von Berichterstattung und Politikformulierung zu machen (Mainstreaming). Insbesondere öffentlich beauftragte Evaluationsstudien sollten zentral im Rahmen einer Datenbank erfasst werden. Öffentliche Auftraggeber sollten die Auftragnehmenden auffordern, die übernommenen Studien kurzfristig bei der Datenbank anzumelden. Ausschlaggebend ist, dass ein für Auftrags-Evaluationen inhaltlich geeignetes Erfassungssystem bereit steht, welches z. B. Angaben zum Evaluationszweck, zur Quelle der Evaluationsfragestellung, zur Wahl des Evaluationsmodells und zur Berücksichtigung von Werten, zur Art des Einbezugs von Beteiligten und Betroffenen enthält.

13. Es sollte das Wissen über Evaluationen in der Armutspolitik aus anderen Staaten mit einer längeren Evaluationskultur systematisch aufgearbeitet werden. Zu diesen Staaten gehören insbesondere die USA, Kanada, Großbritannien, Schweden und die Niederlande sowie Australien. Aus der Fachliteratur und aus Evaluationsberichten sollten Erkenntnisse über Theoriebildung sowie kumulatives Wissen über Wirkungsketten und Kontextbedingungen einer Politik der Armutsvermeidung und sozialen Integration zusammengetragen werden.

6 Literatur

Adorno, Th. W., Albert, H., Dahrendorf, R., Habermas, J., Pilot, H., Popper, K.R., (1969), Der Positivismusstreit in der deutschen Soziologie, Neuwied.

Alkin, M. C. (1969), Evaluation Theory Development, in: Evaluation Comment, Vol. 2, S. 2-7.

AERA/APA/NCME - American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for educational and psychological testing, Washington.

Arbeitsstelle für Evaluation der Universität zu Köln (1999), Endbericht zur Evaluation des Programms ”Förderung von Beratungsstellen und Arbeitslosenzentren für Langzeitarbeitslose und von Langzeitarbeitslosigkeit bedrohte Personen” des Landes Nordrhein-Westfalen (Beywl, W., Potter, P., Schmidt K. u.a.), Köln.

Atkinson, T., Cantillon, B., Marlier, E., Nolan, B. (2002), Social Indicators – The EU and Social Inclusion, Oxford.

Auer, P., Kruppe, Th. (1996), Labour Market Monitoring in EU Countries, in: Schmid, G., O´Reilly, J., Schömann, K. (Hrsg.), International Handbook of Labour Market Policy and Evaluation, Cheltenham, S. 899-922.

AWO (2000) Sozialbericht. Gute Kindheit – Schlechte Kindheit, Bonn.

Ballerstedt, E. und Glatzer, W., unter Mitwirkung von Mayer K. U. und Zapf W., (1979), Soziologischer Almanach. Handbuch gesellschaftspolitischer Daten und Indikatoren für die Bundesrepublik Deutschland, Frankfurt, New York.

Bartelheimer, P. (2000), Kommunale und staatliche Sozialberichterstattung – Ansätze für eine Integration – Thesen für die Tagung des MASQT zur Weiterentwicklung der Sozialberichterstattung in NRW, Düsseldorf, 2.11.2000. www.masqt.nrw.de

Bechthold, S. (2002), Ein Access-Panel für die amtliche Statistik – Weiterentwicklung des methodischen Instrumentariums, in: Allg. Statistisches Archiv, Vol. 86, S.203-212.

Bender, S. (2002), Forschungsdatenzentrum in der BA, in: IAB-Werkstattbericht Nr. 2, S. 16-19.

Beywl, W. (1998), Zur Weiterentwicklung der Evaluationsmethodologie, Frankfurt (Reprint Univation, Köln).

Beywl, W. (1999), Nutzenfokussierte Evaluation von Humandienstleistungen. Plädoyer für eine sozialwissenschaftliche Rückbesinnung in der Qualitätsdebatte, in: Sozialwissenschaften und Berufspraxis, 2/1999, S. 143-156.

Beywl, W., Joas, S. (2000), Evaluation ist unnatürlich! Eine Einführung in die nutzenfokussierte Evaluation entlang eines Seminars von Michael Q. Patton, in: Strübing J. u.a. (Hrsg.) Empirische Sozialforschung und gesellschaftliche Praxis. Bedingungen und Formen angewandter Forschung in den Sozialwissenschaften, Opladen 2000, S. 83-100.

Beywl, W. & Müller-Kohlenberg, H. (2001), Perspektiven der Evaluation in der Kinder- und Jugendhilfe, Materialien zur Qualitätssicherung in der Kinder- und Jugendhilfe, Nr. 35, Berlin (BMFSFJ).

Beywl, W., Potter, P. (1998), RENOMO – A Design Tool for Evaluations: Designing Evaluations REsponsive to Stakeholder`s Interests by Working with NOminal Groups using the MOderation Method, in: Evaluation – The International Journal for Theory, Research an Practice, 4, no.1 (Jan. 1998), pp. 53-72

Beywl, W., Schepp-Winter, E. (2000), Materialien zur Qualitätsentwicklung, Zielgeführte Evaluation von Programmen, Heft 29, S. 62 f.

Beywl, W., Speer, S. (im Erscheinen), Standards for Evaluation Practices. On the Way to Develop Standards for Program Evaluation in Vocational Education and Training Contexts, erscheint in: Descy, P., Tessaring, M. (Hrsg.) Third Research Report on Vocational Education and Training, CEDEFOP, Luxembourg.

Bickman, L. (1990), Using Program Theory to Describe and Measure Program Quality, in: Bickman, L. (Hrsg.), Advances in Program Theory. New Directions for Program Evaluation.

Blank, R. M., Hastings R. (Hrsg.) (2001), The New World of Welfare, Washington D.C.

Bortz, J., Döring N. (2002), Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler, Berlin et al.

Brady, H.E., Grand, S.A., Powell, M.A., Schink, W. (2002), Access and Confidentiality Issues with Administrative Data, in: Ver Ploeg, M., Moffitt, R.A., Citro, C.F. (Hrsg.), Studies of Welfare Populations, Data Collection and Research Issues, Washington D.C., S. 220-274.

Brinkmann, C. (2002), Pilotprojekt „Beschleunigte Datenbereitstellung für externe Institute zum Zwecke der Evaluation arbeitsmarktpolitischer Instrumente“, in: IAB-Werkstattbericht Nr. 2, S. 13-15.

Brinkmann, C. et al. (2002), Dreifache Heterogenität von ABM und SAM und der Arbeitslosigkeitsstatus der Teilnehmer sechs Monate nach Programm-Ende, IAB-Werkstattbericht Nr. 18, 18.12.2002.

Buhr, P. (2002), Ausstieg wohin? Erwerbssituation und finanzielle Lage nach dem Ende des Sozialhilfebezugs, Zes-Arbeitspapier Nr. 4, Zentrum für Sozialpolitik, Universität Bremen.

Bundesamt für Gesundheit (BAG) (Hrsg.) (1997), Leitfaden für die Planung von Projekt- und Programmevaluationen, Bern.

Bundesministerium des Inneren (BMI) (Hrsg.) (2002), Moderner Staat – Moderne Verwaltung, Praxistest zur Gesetzesfolgenabschätzung, Berlin.

Bundesministerium für Gesundheit und Soziale Sicherung.(BMGS) (2003), Lebenslagen, Indikatoren, Evaluation – Weiterentwicklung der Armuts- und Reichtumsberichterstattung Dokumentation des 1. Wissenschaftliches Kolloquium am 30. und 31. Oktober 2002 im Wissenschaftszentrum Bonn.

Bundesministerium für Arbeit und Soziales (Hrsg.) (1999a), Maßnahmen zur Erfolgskontrolle im Bereich der Sozialhilfegesetzgebung. Zusammenfassung des Abschlußberichts, Forschungsbericht Nr. 284-1, Bonn.

Bundesministerium für Arbeit und Soziales (Hrsg.) (1999b), Maßnahmen zur Erfolgskontrolle im Bereich der Sozialhilfegesetzgebung. Ergebnisse der Fachkonferenzen, Forschungsbericht Nr. 284-2, Bonn.

Bundesregierung (Hrsg.) (2001), Daten und Fakten – Materialband zum ersten Armuts- und Reichtumsbericht der Bundesregierung, Berlin.

Bussmann, W. (1996), Evaluationen – mehr Transparenz über die Wirkungen staatlichen Handelns; Ergebnisse aus dem Nationalen Forschungsprogramm „Wirksamkeit staatlicher Maßnahmen“ NFP 27, Bern.

Bussmann, W., Knoepfel, P. (1997), Typische Entstehungszusammenhänge von Evaluationen; in: Bussmann, W., Klöti, U., Knoepfel, P. Einführung in die Politikevaluation, Basel, S. 119-132.

Campbell, D. T. (1982), Experiments as Arguments, in: House, E. R., Mathison, S., Pearsol J. A. & Preskill H. (Eds.) Evaluation studies review annual Vol 7, S. 117-128, Beverly Hills, CA.

Campbell, D. T. [1971] (1991), Methods for the Experimenting Society. Evaluation Practice 12 (3), S. 223-260. Reprint of the 1971 presentation to the American Psychological Association.

Chelimsky, E. (1997), The coming transformations in evaluation, S. 1-26 in: Chelimsky, E. und Shadish W. R., (Hrsg.). Evaluation for the 21st century. A handbook. Thousand Oaks.

Chelimsky, E., Shadish W. R. (Hrsg.), (1997), Evaluation for the 21st century. A handbook. Thousand Oaks.

Chen, H-T. (1994), Theory-Driven Evaluations, Thousand Oaks.

Citro, C. F., Hanushek E. A. (Hrsg.) (1991), Improving Information for Social Policy Decisions – The Uses of Microsimulation Modeling, National Research Council, Washington D.C.

Clayson, Z. C., Castaneda, X., Sanchez, E., Brindis C. (2002), Unequal Power-Changing Landscapes: Negotiations between Evaluation Stakeholders in Latino Communities, American Journal of Evaluation Vol. 23, No. 1, S. 33-44

Cook, T. D. (1997), Lessons Learned in Evaluation Over the Past 25 Years; in: Chelimsky, E., Shadish W. R., (Hrsg.) (1997), Evaluation for the 21st century. A handbook, S. 30-52. Thousand Oaks.

Cousins, J. B., Whitmore, E. (1998), Framing Participatory Evaluation. In: Whitmore, E. (Hrsg.) (1998), Understanding and Practicing Participatory Evaluation, Jossey-Bass (New Directions for Evaluation, No. 80), San Francisco.

Cronbach, L. J. (1963), Course Improvement through Evaluation, in: Teachers College Record, Vol. 64, S. 672-683.

Dann, S., Kirchmann, A., Spermann, A., Volkert, J. (Hrsg.) (2002): Einstiegsgeld in Baden-Württemberg. Schlussbericht, Sozialministerium Baden-Württemberg.

Deeke, A. und Wiedemann E. (2002), Evaluierung aktiver Arbeitsmarktpolitik und Datengrundlagen, IAB-Werkstattbericht Nr. 2, 20.03.02.

Der Bundesbeauftragte für den Datenschutz (BfD) (2001), Tätigkeitsbericht 1999-2000, Bonn.

Deutsche Gesellschaft für Evaluation (DeGEval) (Hrsg.) (2002). Standards für Evaluation. Köln.

Haubrich, K., Frank, K. (2000), Vom Aufsuchen zur beruflichen Integration. Evaluationsstudie zum Bundesmodellprogramm „Mobile Jugendsozialarbeit für junge Menschen ausländischer Herkunft“, DJI-Arbeitspapier Nr. 1-156, April 2000.

Domestic Policy Council, Office of the President (1986), Up from Dependency. 8 vols. Government Printing Office.

Donaldson, S. I.; Scriven, M. (Hrsg.) (2003). Evaluation Social Programs and Problems. Visions for the New Millennium, Mahwah / London.

EC MEANS (European Commission, EC Structural Funds) (1999), MEANS Collection. Evaluating socio-economic programs, Vol. 1-6, Luxembourg.

Edwards, W., Newman, R. J. (1982), Multiattribute Evaluation. Series Quantiative Applications in the Social Sciences No 26, Newbury Park: Sage.

Eichler, U. (2000), Probleme der Verknüpfungen personenbezogener Einzeldaten aus verschiedenen Registern, in: Allg. Statistisches Archiv, Vol. 84, S. 83-93.

Eisentraut, R., Wagner, G. (2002), Evaluierung von Vermittlungsagenturen auf kommunaler Ebene, „Förderung von Maßnahmen zur Erprobung zusätzlicher Wege in der Arbeitsmarktpolitik“, BMGS (Hrsg.), Forschungsbericht Nr. 293, Bonn.

ELSES (2000), Evaluation of Local Socio-Economic Strategies in Disadvantaged Urban Areas (ELSES) – Final Report – Institut für Landes- und Stadtentwicklungsforschung des Landes Nordrhein-Westfalen (Research Institute for Regional and Urban Development of the Federal State of North Rhine-Westphalia), Dortmund.

Engels, D., Hägele, H., Machalowski, G., Sellin, C. (2000), Aussiedlerinnen und Aussiedler in der Sozialhilfe, Kurzfassung, www.isg-institut.de [Januar 2003].

Engels, D., Sellin, C. (2000), Die Praxis des automatisierten Datenabgleichs in der Sozialhilfe nach § 117 Abs. 1 und 2 BSHG, BMGS (Hrsg.), Forschungsbericht Nr. 283, Bonn.

Engels, D., Sellin, C. (2001), Forschungsprojekt Vorstudie zur Nichtinanspruchnahme zustehender Sozialhilfeleistungen, BMGS (Hrsg.), Bonn.

Europäische Kommission (2002), Gemeinsamer Bericht über die soziale Eingliederung, Luxembourg.

European Communities (2002), Quality in the European Statistical System - The Way Forward, Luxembourg.

European Commission (EC) (1999). The MEANS Collection, European Communities: Luxembourg, Vol. 1-6.

Fetterman, D. M. (1993), Speaking the language of power, Communication, Collaboration and Advocacy (Translating ethnography into Action), London Falmer.

Fetterman, D. M., Kaftarian, S. J, Wandersman, A. (Hrsg.) (1996), Empowerment Evaluation: Knowledge and Tools for Self-Assessment and Accountability, Thousand Oaks.

Fetterman, D. M. (2000), Foundations of Empowerment Evaluation. Thousand Oaks.

Fetterman, D. M. (2003), Empowerment Evaluation Strikes a Responsive Cord; in: Donaldson/Scriven 2003, S. 63-76.

Friedman, V. J. (2001), Designed Blindness: An Action Science Perspective on Program Theory, in: American Journal of Evaluation Vol. 22, No. 2, S. 161-181.

Friedrichs, J. (1973), Methoden der empirischen Sozialforschung. Reinbek: Rowohlt.

Friedrichs, J.; Kecskes, R., Wolf, C. (2002), Struktur und sozialer Wandel einer Mittelstadt. Eine empirische Untersuchung in Euskirchen, Opladen: Leske + Budrich.

Geißler, H. (1975), Neue Soziale Frage. Zahlen, Daten und Fakten. Mainz.

Glaser, B. G., Strauss, A. L. (1967), The Discovery of Grounded Theory, Chicago: Aldine.

Goerge, R. M., Lee, B. J. (2002), Matching and Cleaning Administrative Data, in: Ver Ploeg, M., Moffitt, R. A., Citro, C. F. (Hrsg.), Studies of Welfare Populations, Data Collection and Research Issues, Washington D.C., S. 197-219.

Greene, J. C., ABMGS, T. A. (2001), Responsive Evaluation, in: New Directions for Evaluation, San Franciso, No. 92.

Groves, R.M., Couper, M.P. (2002), Designing Surveys Acknowledging Nonresponse, in: Ver Ploeg, M., Moffitt, R.A., Citro, C.F. (Hrsg.), Studies of Welfare Populations, Data Collection and Research Issues, Washington D.C., S. 13-54.

Guba, E. G., Lincoln, Y. S. (1981), Effective Evaluation. Improving the Usefulness of Evaluation Results through Responsive and Naturalistic Approaches, San Francisco: Jossey-Bass.

Guba, E. G., Lincoln, Y. S. (1989), Fourth Generation Evaluation, Newbury Park, CA.

Hager, W., Patry J. L., Brezing, H. (2000): Handbuch Evaluation psychologischer Interventionsmaßnahmen. Standards und Kriterien, Bern.

Haller, S. (1998), Beurteilung von Dienstleistungsqualität. Wiesbaden.

Hanesch, W., Schmid-Urban, P., Dilcher, R., Feldmann, U., Spiegelberg, R. (1992), Kommunale Sozialberichterstattung, Arbeitshilfen Heft 41, Frankfurt: Eigenverlag des Deutschen Vereins für öffentliche und private Fürsorge.

Hanesch, W. u.a.(1994), Armut in Deutschland. Reinbeck bei Hamburg.

Hanesch, W. (2001), Armut und Integration in den Kommunen, in: DFK 2001/I, S. 27-47.

Hauser, R. (1997), Armutsberichterstattung, in: Noll, H. (Hrsg.), Sozialberichterstattung in Deutschland, S. 19-45, Weinheim.

Hauser, R., Becker, I. (2001), Forschungsprojekt Einkommensverteilung im Querschnitt und im Zeitverlauf 1973-1998, Bundesministerium für Arbeit und Soziales (Hrsg.), Lebenslagen in Deutschland, Bonn.

Hauser, R., Wagner, G. G., Zimmermann, K. F. (1998), Erfolgsbedingungen empirischer Wirtschaftsforschung und empirisch gestützter wirtschafts- und sozialpolitischer Beratung – Ein Memorandum, in: Allg. Statistisches Archiv, Vol. 82, S. 369-379.

Hauser, R., Hübinger, W. (1993) Arme unter uns.

Hauser, R., Neumann, U. (1992), Armut in der Bundesrepublik Deutschland. Die sozialwissenschaftliche Thematisierung nach dem Zweiten Weltkrieg, in: Leibfried, St. und Voges, W. (Hrsg.), Armut im modernen Wohlfahrtsstaat. Sonderheft 32 der Kölner Zeitschrift für Soziologie und Sozialpsychologie, Opladen, S. 237-271.

Haveman, R. H. (1987), Policy Analysis and Evaluation Research After Twenty Years, Policy Studies Journal, Vol. 16, S. 191-218.

Heckman, J. J., (2001), Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture, in: Journal of Political Economy, Vol. 109, Nr. 4, S. 673-748.

Heckman, J.J., LaLonde, R.J., Smith, J.A. (1999), The Economics and Econometrics of Active Labor Market Programs, in: Ashenfelter, O., Card, D. (Hrsg.), Handbook of Labor Economics, Vol III, Amsterdam et al., S. 1865-2097.

Heckman, J. J.; Smith J. A. (1996), Experimental and Nonexperimental Evaluation, in: Schmid, G., O´Reilly, J., Schömann, K. (Hrsg.) International Handbook of Labour Market Policy and Evaluation, Cheltenham, S. 37-88.

Heclo, H. (2001), The Politics of Welfare Reform, in: Blank, R. M., Hastings R. (Hrsg.), S. 169-200.

Heil, K., Heiner, M., Feldmann, U. (Hrsg.) (2001), Evaluation Sozialer Arbeit, Frankfurt a. M., S. 59-91.

Heiner, M. (2001), Planung und Durchführung von Evaluationen Anregungen, Empfehlungen, Warnungen, in: Heil, K., Heiner, M., Feldmann, U. (Hrsg.) (2001), Evaluation Sozialer Arbeit, S. 35-58, Frankfurt a. M.

Heiner, M. (Hrsg.) (1988), Selbstevaluation in der Sozialen Arbeit, Freiburg i. Breisgau.

Heiner, M. (Hrsg.) (1998), Experimentierende Evaluation. Ansätze zur Entwicklung lernender Organisationen, München.

House, E. R.; Howe, K. R. (1998), Deliberative Democratic Evaluation in Practice, Boulder, University of Colorado.

House, E.R.; Howe, K.R. (1999), Values in Evaluation and Social Research. Thousand Oaks, CA: Sage Publications.

House, E. R., Howe, K. R. (2000), Deliberative Democratic Evaluation, in: New Directions for Evaluation, Vol. 85, S. 3-12.

Hübinger, W., Neumann, U. (1997), Menschen im Schatten.

Hübler, O. (2001), Evaluation of policy interventions: measurement and problems, in: Allg. Statistisches Archiv, Vol. 85, S. 103-126.

Hujer, R., Caliendo, M. (2000), Evaluation of Active Labour Market Policy: Methodological concepts and empirical estimates, IZA Discussion Paper, No. 236, Bonn.

Hujer, R.; Caliendo, M (2002), Lohnsubventionen in Deutschland - Wie sieht eine optimale Evaluierungsstrategie aus? In: Vierteljahreshefte zur Wirtschaftsforschung, 4/2002 (im Erscheinen).

ILO (International Labour Organisation) (1995), Design, monitoring and evaluation of technical cooperation programs and projects. A training manual, Geneva.

International Labour Office (ILO) [10/01/01] Guidelines for the Preparation of Independent Evaluations of ILO Programmes and Projects [online]. Available from Internet: http://www.ilo.org/public/english/bureau/program/guides/evalmenu.htm.
[November 2002].

Isengard, B. (2002), Machbarkeitsstudie zur Erhebung einkommensschwacher und einkommensstarker Haushalte im Sozio-Oekonomischen Panel (SOEP), DIW Materialien Nr. 17, Juni 2002, Berlin.

Jahoda, M./Österreichische Wirtschaftspsychologische Forschungsstelle (Hrsg.), (1933), Die Arbeitslosen von Marienthal. Ein soziographischer Versuch über die Wirkungen langandauernder Arbeitslosigkeit. Mit einem Anhang zur Geschichte der Soziographie. Bearbeitet und herausgegeben von der Österreichischen Wirtschaftspsychologischen Forschungsstelle (Fünfter Band der Psychologischen Monographien, herausgegeben von Prof. Dr. Karl Bühler), Leipzig.

Jerger, J., Pohnke, Ch., Spermann, A.: Gut betreut in den Arbeitsmarkt? Eine mikroökonmetrische Evaluation der Mannheimer Arbeitsvermittlungagentur, in: MittAB,
Nr. 4, 2001, S. 567-576.

Joint Committee on Standards for Educational Evaluation (2000): Handbuch der Evaluationsstandards, 2. Aufl., Opladen. (Original: The Program Evaluation Standards, 1994. 2. Auflage. Thousand Oaks).

Kalscheuer, M. (2001), Responsive Evaluation „antwortet“ auf ein sich entwickelndes Pilotprojekt – Frühkindliche Erziehung und Qualitätsentwicklung in türkischen Elternvereinen, in: Materialien zur Qualitätsentwicklung, Perspektiven der Evaluation in der Kinder- und Jugendhilfe, Heft 35, S. 69 ff.

Kazi, M. A. F. (2003), Realist Evaluation in Practice, Thousand Oaks.

Kellner, E. (2002), Maßnahme-Teilnehmer-Grunddatei, Neue Basis für anspruchsvolle Wirkungsforschung., in: IAB-Materialien, Nr. 2, S. 8.

Kersting, V. (2000), Kinderarmut im Ruhrgebiet – Fakten eines Armutszeugnisses der Region, Referat auf der Konferenz der Gewerkschaft Erziehung und Wissenschaft am 2.2.2000 in Gelsenkirchen, Bochum.

Kluckhohn, C. (1967), Values and Value Orientations in the Theory of Action, in: Parsons, T. and Shils E., Towards a General Theory of Action, New York, S. 388-433.

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) (Hrsg.) (2001), Wege zu einer besseren informationellen Infrastruktur, Baden-Baden.

Kortmann, K., Sopp, P., Thum, M. (2002), Das Niedrigeinkommens-Panel, Methodenbericht, Infratest Sozialforschung, München (unveröffentlicht).

Korczak, D. (2001), Überschuldung in Deutschland zwischen 1988 und 1999, Bundesministerium für Familie, Senioren, Frauen und Jugend (Hrsg.), Stuttgart.

Kromrey, H. (2000), Die Bewertung von Humandienstleistungen. Fallstricke bei der Implementations- und Wirkungsforschung sowie methodische Alternartiven; in: Müller-Kohlenberg, H., Münstermann, K. (Hrsg.) (2000), Qualität von Humandienstleistungen, Opladen, S. 19-58.

Krueger, Richard A. (1994), Focus Groups. A Practical Guide for Applied Research; 2nd edition; Thousand Oaks, London, New Delhi: SAGE Publications.

Kuck-Schneemelcher, D. (2001), Der Armuts- und Reichtumsbericht der Bundesregierung. Bilanz des ersten Berichts und Perspektiven für die künftige

Berichterstattung, in: Archiv für Wissenschaft und Praxis der sozialen Arbeit, 32. Jg., Nr. 4, S. 70-83.

Lechner, M. (2002), Eine wirkungsorientierte aktive Arbeitsmarktpolitik in Deutschland und der Schweiz: Eine Vision – Zwei Realitäten, in: Perspektiven der Wirtschaftspolitik, Band 3, Heft 2, S. 159-174.

Levecque, K., Vranken, J. (2000), La Valorisation des Banques de Données socio-economiques dans l´Étude de la Pauvreté et de l´Exclusion Sociale, In: Revue Belge de Sécurité Sociale, Nr.1 , S. 193-214.

Levin, H.M., McEwan, P.J. (2001), Cost-Effectiveness Analysis. London.

Lincoln, Y.(2003), Fourth Generation Evaluation in the New Millenium; in: Donaldson/Scriven (2003), S. 77-90.

MacNeil, C. (2002), Evaluator as Steward of Citizen Deliberation, in: American Journal of Evaluation Vol. 23, No. 1, S. 45-54.

Mathematica Policy Research, Final Report of the Seattle-Denver Income Maintenance Experiment. Vol 2. Princeton: Mathematica Policy Research 1983.

Mayntz, R. (1958), Soziale Schichtung und sozialer Wandel in einer Industriegemeinde, Stuttgart.

Mayntz, R. (2002), Nachwort, in: Friedrichs, J., Kecskes, R.; Wolf, C. (2002), Struktur und sozialer Wandel einer Mittelstadt. Eine empirische Untersuchung in Euskirchen, S. 203-205.

Meier, U., Preuße, H., Sunnus, E. M. (2001), Armutsprävention und Milderung defizitärer Lebenslagen durch Stärkung der Haushaltsführungskompetenzen, Haushaltsführung im Versorgungsverbund der Daseinsvorsorge, Materialien zur Familienpolitik Nr. 13, Universität Gießen.

Mertens, D. M. (2003), The Inclusive View of Evaluation: Visions for the New Millennnium, in: Donaldson/Scriven (2003), S. 91-108.

Merz, J. (1991), Microsimulation – A survey of principles, developments and applications, in: International Journal of Forecasting, Nr. 7, S. 77-104.

Metschke, R., Wellbrock, R. (1999), Statistikgeheimnis und Datenschutz wieder die empirische (Wirtschafts-)Forschung? Anmerkungen zum Memorandum, in: Allg. Statistisches Archiv, Vol. 83, S. 152-157.

Meyer, T. (1999), Werte, in Richter, D., Weißeno, G. (Hrsg.) (1999), Lexikon der politischen Bildung, Band 1 Didaktik und Schule, Schwalbach/Ts. S. 259

Ministerium für Arbeit und Soziales, Qualifikation und Technologie MASQT (Hrsg.) (2002), Modellprojekt 'Pauschalierung von Sozialhilfe' NRW. Zwischenbericht der wissenschaftlichen Begleitung, Düsseldorf.

Ministerium für Arbeit und Soziales, Qualifikation und Technologie MASQT (2001), Datenschutzerklärung.

Ministerium für Arbeit, Soziales und Stadtentwicklung, Kultur und Sport des Landes Nordrhein-Westfalen (MASSKS) (1999), Sozialbericht '98 für das Land Nordrhein-Westfalen. Materialienband. Düsseldorf.

Moffit, R. A. und Ver Ploeg, M. (2001), Evaluating Welfare Reform in an Era of Transition, National Research Council, Washington DC.

Morgan, D. L., Krueger, R. A. (1998), The Focus Group Kit; 1st edition; Thousand Oaks, London, New Delhi: SAGE Publications.

Müllenmeister-Faust, U., Semrau, P (2002), The Poverty and Wealth Report and the National Action Plan (NAPincl): Mutual Coordination and Prospects, in: Hauser, R. und I. Becker (Hrsg.), Reporting on Income Distribution and Poverty, Berlin et al., S. 127-142.

Müllenmeister-Faust, U. (2002), Möglichkeiten und Grenzen der Armuts- und Reichtumsberichterstattung, in: Sell, S. (Hrsg.) Armut als Herausforderung, Schriften der Gesellschaft für Sozialen Fortschritt, Band 23, S. 169-192.

Müller, W. et al. (1991), Die faktische Anonymität von Mikrodaten, in: Schriftenreihe Forum der Bundesstatistik, Bd. 19, Wiesbaden.

Müller-Kohlenberg, H.; Münstermann, K. (Hrsg.) (2000), Qualität von Humandienstleistungen. Evaluation und Qualitätsmanagement in Sozialer Arbeit und Gesundheitswesen, Opladen: Leske + Budrich.

Murray, C. (1984), Losing Ground: American Social Policy, 1950-1980, New York, Basic Books.

Musgrave, R. A. and Musgrave, P. B. (1980), Public Finance in Theory and Practice, New York et al.

Newman, K. S. (2002), The Right (Soft) Stuff: Qualitative Methods and the Study of Welfare Reform, in: Ver Ploeg, M., Moffitt, R. A., Citro, C.F. (Hrsg.), Studies of Welfare Populations, Data Collection and Research Issues, Washington D.C., S. 355-383.

Noll, H.-H. (2003), Indikatoren einer mehrdimensionalen Armuts- und Reichtumsberichterstattung, in: Bundesministerium für Gesundheit und Soziale Sicherung (BMGS) (Hrsg.), Lebenslagen, Indikatoren, Evaluation - Weiterentwicklung der Armuts- und Reichtumsberichterstattung, Dokumentation des 1. Wissenschaftlichen Kolloquium, 30. und 31. Oktober 2002 im Wissenschaftszentrum Bonn (im Erscheinen).

Otto, B. (2002), Die sozioökonomischen Folgen eines einkommensabhängigen Kindergeldzuschlags. Eine Mikrosimulation der „Grünen Kindergrundsicherung“, DIW Diskussionspapier Nr. 273, Berlin, Februar 2002.

Owen, J. M., Rogers, P. J. (1999), Program Evaluation. Forms and Approaches. London, Thousand Oaks, New Dehli.

Patton, M. Q. (1997): Utilization-Focused Evaluation. The New Century Text. Thousand Oaks, London, New Delhi.

Pawson, R. and Tilley, K. (1997). Realistic Evaluation. London: Sage.

Petersen, D. M. (2002), The Potential of Social Capital Measures in the Evaluation of Comprehensive Community-Based Health Initiatives, American Journal of Evaluation Vol. 23, No. 1, S. 55-64.

Plantz, M. C., Greenway, M. T., Hendricks, M. (1997), Outcome Measurement: Showing Results in the Nonprofit Sector, in: New Directions for Evaluation 75, (Fall 1997).

Pleiger, D., Weißmann, R., Friedrich, J., Klawe, W., Menschen statt Mauern – Evaluation der Jugendhilfeeinrichtung Frostenwalde in Brandenburg, in: Materialien

zur Qualitätssicherung in der Kinder- und Jugendhilfe, Nr. 35, Berlin (BMFSFJ), S. 49 – 52 ff.

Popper, K. (1969), Die Logik der Sozialwissenschaften, in: Adorno u. a., S. 103-124.

Popper, K. (1989), Logik der Forschung (9. Auflage), Tübingen Mohr (Erstveröffentlichung 1934).

Potter, P., Evaluating advice and resource projects for the long-term unemployed in Germany: a mixed method approach, in: Klaus Künzel (Hrsg.): Internationales Jahrbuch der Erwachsenenbildung/International Yearbook of Adult Education, Band 27, 1999, S. 91-104.

Provus, M. N. (1971), Discreapancy Evaluation, Berkeley.

Reis, C. (2001), Das Rahmenkonzept der „Sozialagentur“ – Ziele, Aufgabenfelder, Organisationsvarianten, Frankfurt a. M.

Reischmann, J. (2002), Weiterbildungs-Evaluation, Neuwied, Kriftel.

Robins, P. K. (1980), Spiegelman, R. R., Weiner, S.; Mell, J. G. (eds.), A Guaranteed Annual Income: Evidence from a Social Experiment. New York: Academic Press.

Rog, D. J. (Hrsg.), (1992), Evaluating programs for the homeless. New Directions for Program Evaluation (Series No. 52), San Francisco.

Rogers, P. R. (2000), Program Theory, Not whether Programs work but how they work, in: Stufflebeam, D. L.; Madaus, G. F., Kellaghan, T. (Hrsg.), Evaluation Models, Boston: Kluwer Academic Publishers.

Rossi, P. H., Lyall, K. (1976), Reforming Public Welfare, New York: Russel Sage.

Rossi, P. H., Freeman, H. E., Lipsey, M. W. (1999), Evaluation. A Systematic Approach, 6th ed., Thousand Oaks.

Sanders, J. R. (1997): Cluster Evaluation. In: Chelimsky, E., Shadish, W. R. (Hrsg.), Evaluation for the 21st Century. A Handbook. Thousand Oaks, London, New Delhi, S. 396-404.

Schömann, K. (1996), Longitudinal Designs in Evaluation Studies, in: Schmid, G., O'Reilly, J., Schömann, K. (Hrsg.), International handbook of labour market policy and evaluation. Cheltenham, S. 115-142.

Schupp, J., Wagner, G. G. (2002), Maintenance of and inovation in long-term panel studies: The case of the German Socio-Economic Panel, in: Allg. Statistisches Archiv, Vol. 86, S. 163-175.

Scriven, M. (1973), Goal-Free evaluation, in: House, E. R. (Hrsg.), School Evaluation: The Politics and Process, Berkeley, CA.

Scriven, M. (1991), Evaluation Thesaurus, Newbury Park, CA.

SEVAL Standards (2001), Schweizerische Evaluationsgesellschaft (SEVAL) Evaluationsstandards. www.seval.ch [Januar 2003].

Shadish, W. R., Cook, T .D., Campbell, D. T. (2002), Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Boston, New York.

Singer, E., Kulka, R. A. (2002), Paying Respondents for Survey Participation, in: Ver Ploeg, M., Moffitt, R. A., Citro, C. F. (Hrsg.), Studies of Welfare Populations, Data Collection and Research Issues, Washington D.C., S. 105-128.

Sirotnik, K. A. (1990) (Hrsg.), Evaluation and Social Justice: Issues in public education, in: New Directions for Evaluation, Vol. 45, S. 23-36.

Smith, M.F. (1989), Evaluability assessment. A practical approach, Norwell.

Sozialgesetzbuch Drittes Buch – Arbeitsförderung – http://www.BMGS.de/download/gesetze_web/Sgb03/sgb03xinhalt.htm [Januar 2003].

Speer, S. (2002), Praxis der Unterkunftspauschalierung und Thesen zu Erfolgsfaktoren, Arbeitspapier, Univation e.V., Köln.

Spiegel, H. v. (2001), Leitfaden für Selbstevaluationsprojekte in 18 Arbeitsschritten, in: Heil, K., Heiner, M. und Feldmann U. (Hrsg.), Evaluation Sozialer Arbeit, Frankfurt a. M., S. 59-91.

Spiegel, H. v. (1993), Aus Erfahrung lernen. Qualifizierung durch Selbstevaluation, Münster.

SRI International (1983), Final report of the Seattle-Denver Income Maintenance Experiment Vol. 1. Palo Alto CA: SRI International.

Stake, R. (1995), The Art of Case Study Research, Thousand Oaks.

Stephan, A. S. (1935), Prospects and Possibilities: The New Deal and the New Social Research, Social Forces, May, 13, S. 515-521.

Strohmeier, K. P., Hank, K., Kersting, V., Langenhoff, G. (1999), Armut in Nordrhein-Westfalen. Umfang und Struktur des Armutspotentials. Forschungsbericht für den 8. Sozialbericht des Ministeriums für Arbeit, Soziales und Stadtentwicklung, Kultur und Sport des Landes Nordrhein-Westfalen. Bochum.

Stufflebeam, D. L. (1966), A depth Study of the Evaluation Requirement. in: Theory into Practice, Vol. 5, S. 121-134.

Stufflebeam, D. L. (1967), The Use and Abuse of Evaluation in Title III, in: Theory into Practice, Vol. 6, S. 126-133.

Stufflebeam, D.L. (1972), Evaluation als Entscheidungshilfe, in: Wulf, Ch. (Hrsg.) (1972), Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen, München.

Sturm, R. (2002), Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten, in: Allg. Statistisches Archiv, Vol. 86, S. 468-477.

Suchman, E. A. (1967), Evaluative research. Principles and Prctice in Public Service and Social Action Programs. New York.: Russel SagE Foundation.

Trube, A.: Fiskalische und soziale Kosten-Nutzen-Analyse örtlicher Beschäftigungsförderung – Eine exemplarische Untersuchung, Beiträge zur Arbeitsmarkt- und Berufsforschung, Nr. 189, Nürnberg, IAB, 1995.

Tyler, R. W. (1950), Basic Principles of Curriculum and Instruction. Chicago: University of Chicago Press.

Univation e.V. (2001), Modellprojekt „Pauschalierung von Sozialhilfe“ NRW, Zwischenbericht der wissenschaftlichen Begleitung, Köln. http://www.5-q.de/paso/ [Stand: 28.3.2003]

Vedung, E. (1999), Evaluation im öffentlichen Sektor, Wien.

Voges, W. (1992), Sozialhilfedaten als soziale Indikatoren. Wie die Sozialverwaltung Informationen zur Armutsbeseitigung liefern könnte, in: Johrendt, N. und Schneider H. R. (Hrsg.), Computerunterstützte Sozialberichterstattung und Sozialplanung, Bielefeld, AJZ 1992.

Vranken, J. et al. (2001), Towards an integrated approach of European policies on social exclusion and inclusion, University Antwerpen.

VSOP (1995), Standards der Armutsberichterstattung – Fachpolitische Stellungnahme des Vereins für Sozialplanung e.V., in: NDV, Heft 3, S. 120-123.

Weber, M., Die Objektivität sozialwissenschaftlicher und sozialpolitischer Erkenntnis,in: M. Weber. Gesammelte Aufsätze zur Wissenschaftslehre Tübingen Mohr (Erstveröffentlichung 1904).

Weber, T. (2002), Einführung der Statistiken über eine bedarfsorientierte Grundsicherung im Alter und bei Erwerbsminderung, in: Wirtschaft und Statistik 12/2002, S. 1076-1079.

Webster, W. J. (1975), The Organization and Functions of Research Evaluation in a large urban School District, Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Webster, W. J. (1995), The Connection between Personnel Evaluation and School Evaluation, in: Studies in Educational Evaluation, Vol. 20, S. 113-145.

Weiss, C. H. (1972), Evaluation Research. Methods for Assessing Program Effectiveness. Englewood Cliffs, NJ: Prentice-Hall.

Weiss, C. H. (1995), Nothing as practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families, in: Connell, J., Kubisch, A., Schorr, L. B., Weiss, C. H. (eds.), New Approaches to Evaluating Community Initiatives, New York: Aspen Institute.

Weiss, C. H. (1998), Evaluation. 2nd Edition. Upper Saddle River, Prentice Hall.

Weitzmann, B., Silver, D., Dillman, K.-N. (2002), Integrating a Comparison Group Design into a Theory of Change Evaluation: The Case of the Urban Health Initiative, American Journal of Evaluation, Vol. 23, No. 4, S. 371-385.

Whitmore, E. (1998), Understanding and Practicing Participatory Evaluation; New Directions for Evaluation. Nr. 80, San Francisco.

Wholey, J. S., Hatry, H. P., Newcomer, K. E. (Hrsg.) (1994), Handbook of practical program evaluation, San Francisco.

Widmer, T. (1996), Meta-Evaluation: Kriterien zur Bewertung von Evaluationen. Bern.

Wießner, F. (2002), Raum für Experimente, Frei fördern und forschen, in: IAB-Materialen, Nr. 2, S. 12-13.

Wisler, C. (Hrsg.) (1996), Evaluation and Auditing. Prospects for Convergence New Directions for Evaluation, No. 71, Jossey-Bass Inc., Publishers, San Francisco, CA.

Worthen, B. R./Sanders, J. R./Fitzpatrick, J., L. (1997) Program Evaluation. Alternative Approaches and Practical Guidelines. 2. Auflage. New York: Addison-Wesley

Worthen, B. R.; Schmitz, C. (1997), Conceptual Challenges Confronting Cluster Evaluation, in: Evaluation, Vol. 3, No. 3, S. 300-319

Wottawa, H., Thierau, H. (1998), Lehrbuch Evaluation, Bern.

Zwick, M. (2003), Das Forschungsdatenzentrum des Statistischen Bundesamtes, Folienpräsentation zum Vortrag, gehalten auf 1. Konferenz für Sozial- und Wirtschaftsdaten, am 13./14. Januar 2003 in Wiesbaden (unveröffentlicht).

7 Anhang I

7.1 Glossar

Adressaten und **Adressatinnen** der Evaluation (engl. *audiences*)	Die intendierten (nicht unbedingt tatsächlichen, also auch erreichten) Nutzer von Evaluationsergebnissen.
Auswirkungen eines Programms	„Auswirkungen“ wird wie in der Alltagssprache genutzt: Behauptet wird einen Zusammenhang zwischen einem Auslöser (z.B. Intervention) und einem Resultat (zum Beispiel verbesserter sozialer Status), ohne dass die Verbindung systematisch / empirisch nachgewiesen wäre. Oft werden Auswirkungen mit ➔ „Wirkungen“ gleichgesetzt. Diesen Terminus reservieren wir für systematisch auf den Auslöser zurückgeführte Resultate.
Beteiligte am und **Betroffene** durch das Programm (engl. s*takeholder*)	Beteiligte sind Personen, Gruppen oder auch Organisationen, die in Bezug auf den ➔ Evaluationsgegenstand eine aktive Rolle spielen, z.B. Finanziers eines ➔ Programms, Akteure/-innen, die im Programmkontext tätig sind. Betroffene sind insbesondere Personen mit wenig Einfluss, die sich oft in den Zielgruppen eines Evaluationsgegenstandes finden. Es können aber auch vom Gegenstand Ausgeschlossene oder Benachteiligte sein, die vorher u.U. nicht im Fokus des Programms standen. Die Trennlinie zwischen Beteiligten, Betroffenen, ➔ Adressaten/-innen und Nutzern/-innen kann nicht scharf gezogen werden, Personenkreise überschneiden sich häufig.
dokumentierende Evaluation	Eine ➔ Evaluationsfunktion. Während der Dokumentation werden laufend Kennzahlen über den Programmverlauf bereitgestellt. Erhebungsverfahren hierfür sind in den Programmverlauf fest integriert. Dies soll dazu dienen, die Qualität des Programmverlaufs kontinuierlich mit angemessenem Aufwand zu überprüfen, um ggf. steuernd eingreifen zu können (auch: ➔“Monitoring“).

Effekte des Programms	Beabsichtigte ➔ Outcomes und ➔ Wirkungen (für die formulierte Ziele vorliegen). Von „Nebeneffekten“ – im Sinne von vorab nicht vorausgesehenen Effekten eines Programms (=Zielzuständen) zu sprechen ist logisch unmöglich: Was nicht bekannt ist kann nicht vorausgesehen werden.
emergentes Design	Ein emergentes Evaluationsdesign hat keinen von vornherein festgelegten ➔ Evaluationsplan, sondern das weitere Vorgehen ergibt sich immer erst aus den Ergebnissen der vorhergegangenen Phase der Evaluation.
Ergebnisse der Evaluation	➔ Evaluationsergebnisse
Evaluation	Evaluation bezeichnet die Summe systematischer Untersuchungen, die empirische, d.h. erfahrungsbasierte, Informationen bereitstellen, so dass es möglich wird, den Wert (➔Güte und ➔Verwendbarkeit) eines (in der Regel sozialen) ➔ Evaluationsgegenstandes einzuschätzen.
Evaluationsergebnisse vs. Programmresultate	Der Terminus „Evaluationsergebnisse“ umfasst die durch eine konkrete Evaluation bereitgestellten Beschreibungen oder Bewertungen (Datenauswertungen, Schlussfolgerungen, Beurteilungen, Empfehlungen) und soll sprachlich von den ➔ „Resultaten“ des ➔ Programms unterschieden werden.
Evaluationsfragestellung vs. Fragen	Die Fragestellungen richten sich darauf, was die Auftraggeber/-innen der Evaluation bzw. je nach Modell andere ➔ Beteiligte und Betroffene über den ➔ Evaluationsgegenstand an Informationen anfordern. Mit den Fragestellungen wird eingegrenzt, zu welchen Aspekten des ➔ Programms systematisch Daten und Informationen zu gewinnen sind. Der ➔ Evaluationsplan soll präzise auf die Fragestellungen zugeschnitten sein. Zur sprachlichen Abgrenzung werden (komplexe) Fragen, die durch eine Evaluation beantwortet werden sollen, im Gegensatz zu „Fragen“, die z.B. in offener oder geschlossener Form in einem Fragebogen oder Interviewleitfaden vorkommen, mit „Fragestellungen“ bezeichnet.

Evaluationsfunktionen

Eine Evaluation kann, um die Optimierung und Stabilisierung des ➔ Programms bzw. eine optimale Feststellung seiner ➔ Wirkungen (➔ Wirkungsfeststellung) sicherzustellen, verschiedene Funktionen übernehmen. Diese Funktionen sind: ➔ proaktive Evaluation, ➔ klärende Evaluation, ➔ interaktive Evaluation, ➔ dokumentierende Evaluation und ➔ wirkungsfeststellende Evaluation.

Evaluationsgegenstand
(engl. *evaluand*;
auch *evaluation object*)

Das durch eine Evaluation zu Beschreibende und zu Bewertende. Dies kann z.B. ein ➔ Programm, Projekt, Produkt, eine Maßnahme, Leistung, Organisation, Politik, Technologie oder Forschung sein. Der Evaluationsgegenstand wird zum einen angemessen deskriptiv dargestellt (Programmbeschreibung mit Bestandteilen wie Träger, ➔ Zielgruppen, Interventionen, ...), zum anderen fokussiert durch ➔ Evaluationsfragestellungen in Bezug auf seine relevanten Elemente empirisch untersucht.

Evaluationsmodell
(auch Evaluationsansatz)

Ausformulierte, theoretisch begründete und durch praktische Evaluationserfahrungen gesättigte Anleitung, wie praktische Evaluationen geplant und durchgeführt werden sollen. In unterschiedlichen Evaluationsmodellen werden zumeist auch verschiedene ➔ Evaluationsfunktionen angesprochen. Gemäß der Berücksichtigung von Werten können Evaluationsmodelle als ➔ wertedistanziert, ➔ werterelativistisch, ➔ wertepriorisierend oder ➔ wertepositioniert Modelle eingeordnet werden.

Evaluationsplan
(auch Evaluationsdesign)

Der Evaluationsplan enthält das Vorgehen bei der Evaluation, wie es von den Evaluierenden, u.U. im Zusammenarbeit mit ➔ Stakeholdern, vorgesehen ist. Es kann z.B. beschrieben sein, welche Daten erhoben werden sollen, welche ➔ Fragestellungen beantwortet werden sollen oder in welcher Form wann ➔ Ergebnisse präsentiert werden sollen

Evaluationszwecke (engl. *purpose of the evaluation*) vs. (Programm-) Ziele	Die intendierten Verwendungen der Evaluation bzw. ihrer ➔ Ergebnisse, z.B. die Verbesserung eines ➔ Programms. Ein anderer häufiger Zweck ist die Grundlegung von Entscheidungen über Programme (Weiterführung, Ausbreitung). Schließlich ist ein möglicher Zweck der Gewinn von Erkenntnissen ohne unmittelbaren praktischen Nutzen für den Evaluationsgegenstand. Der Terminus „Zweck“ (d. Evaluation) soll zur Erleichterung der Verständigung abgegrenzt werden von „Zielen“, die im Bereich des ➔ Evaluationsgegenstandes zu finden sind (z.B. Lernziele, Vermittlungsziele). Evaluationszwecke sind leitend für die Ausrichtung der Evaluation.
formative Evaluation vs. summative Evaluation	Eine formative Evaluation soll vor allem die Leistung erbringen, die Gestaltung des ➔ Evaluationsgegenstandes anzulieten. Der Zweck der Evaluation liegt hier v.a. in einer Verbesserung des Gegenstands. Sie soll den Verantwortlichen und ➔ Beteiligten helfen, den Evaluationsgegenstand und seine ➔ Verwendbarkeit zu optimieren und Ressourcen möglichst effizient einzusetzen.
Fragestellung der Evaluation	➔ Evaluationsfragestellungen
Funktionen der Evaluation	➔ Evaluationsfunktionen
Gegenstand der Evaluation	➔ Evaluationsgegenstand
Güte (engl. *merit*) vs. Verwendbarkeit des Programms	Zurückgehend auf Guba/Lincoln (1981) wird mit Güte (im Gegensatz zu ➔ Verwendbarkeit) die „intrinsische“ Qualität eines ➔ Evaluationsgegenstandes bezeichnet. Das meint im Falle eines ➔ Programms beispielsweise die Stringenz und fachwissenschaftliche Absicherung des Konzepts. Güte ist zeitlich und räumlich vergleichsweise stabil.

Impacts

Hierbei handelt es sich um ➔ Resultate, die *nicht* bei den Zielgruppen auftreten / gemessen werden, sondern in sozialen Systemen, insbesondere Organisationen (Unternehmen, sozialen Dienstleistungsanbietern, Schulen) oder in Sozialräumen (Nachbarschaften, Kommunen, Regionen) oder im Netzwerk der personalen und organisationalen Akteure eines Politikfeldes (Weiterbildungssystem eines Bundeslandes, Gesundheitssystem einer Nation). So kann z.B. das soziale Klima eines Sozialraums friedlicher oder die Effektivität und Effizienz eines Dienstleistungssystems höher sein. In aller Regel ist die Messung von Impacts (noch) aufwändiger als die von Outcomes, da nicht Personen Merkmalsträger der zu erfassenden Veränderungen/Stabilisierungen sind sondern Beziehungen zwischen Personen sowie deren ökonomische, sozio-kulturelle, institutionelle, natürliche, technologische usw. Umwelt. Impacts sind oft Resultate einer oder mehrerer Ketten von Prozessen, Outputs, Outcomes und erschließen sich erst bei längerfristiger Betrachtung. Der empirische Nachweis, dass Impacts durch bestimmte Ursachen ausgelöst sind, dürfte selten gelingen (in diesen Fällen könnte man von Impact-Wirkungen sprechen).ö

Incomes

Bezieht sich auf das, was die ➔ Zielgruppenmitglieder an Voraussetzungen in das ➔ Programm „mitbringen", insbesondere an Wissen, Einstellungen, Bedarfen, Werten etc. So gibt es z.B. starke Unterschiede in den Incomes bezüglich des Alters, des Bildungsstandes oder des Geschlechtes der potentiellen Programmteilnehmenden. Incomes stellen eine zentrale Bedingung dar, die bei der Konzeptentwicklung und Umsetzung von Programmen berücksichtigt werden muss. Mittels der Teilnehmendenauswahl (profiling) können die Incomes eines Programms gesteuert werden.

Inputs

Bezeichnet finanzielle, personale oder andere Ressourcen, die in ein ➔ Programm investiert werden. Diese stellen eine (im Unterschied zu ➔ Kontext und ➔ Struktur) vergleichsweise variable Bedingung dar, insofern z.B. Kostenarten (Personal- vs. Sachausgaben) oder Personalqualifikationen (durch gezielte Fortbildung) beeinflusst werden können.

interaktive Evaluation — Eine ➔ Evaluationsfunktion. Die Evaluation liefert Zwischenergebnisse zur Prozessqualität während der Programmumsetzung und unterstützt dabei die Feinabstimmung im Programmverlauf.

klärende Evaluation — Eine ➔ Evaluationsfunktion. Hierbei wird die Klarheit und Stimmigkeit der Programmziele überprüft. Die Programmziele werden, wenn die Feststellung der ➔ Resultate oder ➔ Wirkungen eines ➔ Programms erwünscht ist, in ➔ Evaluationsfragestellungen überführt. Die Klärung unterstützt die Konzeptentwicklung, die Entwicklung einer Programmtheorie und die Abstimmung von Interventionen innerhalb des Programms und gewinnt Informationen über die Ausgangssituation und die Umsetzungsbedingungen.

Kontext — Beschreibt die Umgebungsbedingungen eines ➔ Programms. Je nach Reichweite des Programms können diese auf lokaler, nationaler oder internationaler Ebene liegen und soziale, politische und kulturelle Aspekte betreffen, die sich nur langfristig und unabhängig vom Programm selbst ändern. Beispiele sind Wirtschaftswachstum, soziale Schichtung, Arbeitslosenquote, öffentliche Meinung. Kontextbedingungen sind in Programmkonzepte einzubeziehen. Identische Programmkonzepte führen bei stark differenten Kontexten zu unterschiedlichen Resultaten.

Konzept des Programms — Das Konzept enthält die Annahmen von Auftraggebenden, Programmplanenden/-verantwortlichen darüber, was das ➔ Programm bis wann bei welchen ➔ Zielgruppen in welchen ➔ Kontexten auslösen soll (Zielsetzungen), welche Aktivitäten zur Zielerreichung (Interventionsplanung) eingesetzt werden sollen und wie der Programmprozess insgesamt gesteuert und überwacht werden soll (Qualitätssicherung).

Modell der Evaluation — Zu Typen zusammengefasste, ausformulierte, theoretisch begründete und durch praktische Evaluationserfahrungen gesättigte Anleitungen, wie praktische Evaluationen geplant und durchgeführt werden sollen.

Monitoring — Laufende Erfassung und Dokumentation und von Daten und die Berichterstattung darüber ➔ dokumentierende Evaluation

Nebenresultate des Programms	Ein Nebenresultat ist ein ➔ Resultat des ➔ Programms, das im ➔ Konzept nicht als ein angestrebtes Resultat vorgesehen ist. (Es liegt außerhalb des im Konzept festgehaltenen Zielsystems, ist insofern nicht-intendiert.) Nebenresultate können sowohl bei ➔ Zielgruppenmitgliedern, als auch bei anderen Personen (insbesondere bei Angehörigen oder bei vom Programm Ausgeschlossenen), in der Umwelt/dem Kontext des Programms oder bei ➔ Strukturen etc. auftreten. Eventuell sind sie bei Programmbeginn nicht voraussehbar. Wenn das Nebenresultat zwar vorhersehbar aber nicht gewünscht ist, ist dies ein unerwünschtes Nebenresultat. Nebenresultate können nur *nachträglich* als erwünscht bewertet werden. Wenn ihre Verursachung durch das Programm nachgewiesen ist (was methodisch schwierig zu bewerkstelligen ist) spricht man von Nebenwirkungen.
Nebenwirkungen	➔Nebenresultate
Nutzen der Evaluation/ Evaluationsergebnisse (engl. *use*)	Der Nutzen soll dadurch hergestellt werden, dass ➔ Evaluationsergebnisse gebraucht (i.S.v. benutzt) werden und einen Zugewinn (an Erkenntnis, Qualität, …) darstellen. Auch der Prozess der Evaluation selbst kann zu einem Nutzen führen, man spricht dann vom ➔ Prozessnutzen der Evaluation.
Nutzer und **Nutzerinnen** der Evaluationsergebnisse	Diejenigen Personen, die ➔ Evaluationsergebnisse oder den Prozess der Evaluation (➔ Prozessnutzen) tatsächlich nutzen.
Outcomes des Programms	Outcomes sind Merkmale, die bei Mitgliedern von Zielgruppen gemessen werden. Sie bezeichnen die intendierten Resultate der Interventionen/Aktivitäten eines Programms, wie z.B. veränderte Einstellungen oder verändertes Verhalten bei Zielgruppenmitgliedern oder Vorteile für die ➔Zielgruppen. Nicht-intendierte Resultate bei Zielgruppenmitgliedern werden nicht als Outcomes bezeichnet.
Outcome-Wirkungen des Programms	➔ Outcomes, also Veränderungen / Stabilisierungen bei den Zielgruppen, die ursächlich auf das Programm zurückzuführen sind.

Outputs des Programms	Bezeichnet sämtliche Leistungen wie Materialien, Waren, Aktivitäten, Publikationen und insbesondere Dienstleistungen, die durch den ➔ Evaluationsgegenstand (ein ➔ Programm, ein Projekt etc.) direkt produziert werden, wie z.B. Unterrichtsstunden, Leistungsstunden, Broschüren, Profilings, Assessments, Kurse, Beratungsgespräche, Hilfepläne, Vermittlungen, etc. Outputs sind „Mengen“, die sich vergleichsweise leicht zählen und damit quantitativ messen lassen. Sie stehen den ➔Inputs, also (ibs. finanziellen) Aufwendungen gegenüber. Ein Input-Output-Vergleich stellt z.B. dar, welche Bildungseinrichtung bei identischen Inputs (z.B. gemessen als monetäre Kosten) die meisten Outputs produziert.
Parafisci	Körperschaften des öffentlichen Rechts mit eigener Finanzhoheit. Sie sind mit den öffentlichen Haushalten über – oft enge – gesetzliche Vorgaben und Teilfinanzierung (z.B. Bundeszuschuss) verbunden: ibs. Gesetzliche Rentenversicherung, Gesetzliche Krankenversicherung, Knappschaften, Gesetzliche Unfallversicherung.
partizipative Evaluation	Eine Evaluation ist partizipativ, wenn ➔ Beteiligte und Betroffene des ➔ Programms einbezogen werden, einen Einblick in die Durchführung der Evaluation bekommen bzw. mit über ihre Durchführung bestimmen können und dadurch bestimmte Fähigkeiten erlangen oder erweitern können.
personenbezogene Dienstleistungen	Dienstleistungen, die im Rahmen von ➔Programmen in Feldern der Armuts- und Reichtumsberichterstattung erbracht werden, haben häufig die Veränderung/Stabilisierung von Wissen, Einstellungen/Handeln von Personen zum Ziel. Diese Personen (Klienten, Teilnehmende ...) müssen aktiv mitwirken, damit das Resultat zustande kommt (Gebot der Ko-Produktion). Die Zielqualität des Resultates wird bestimmt durch die konkret handelnden Personen auf dem Hintergrund ihrer individuellen Werte (vgl. Beywl 1999).

proaktive Evaluation	Eine ➔ Evaluationsfunktion. Vor dem Start eines ➔ Programms soll anhand von Vergleichsuntersuchungen oder eigens dafür durchgeführten Erhebungen ermittelt und beschrieben werden, wie die Bedingungen (➔ Kontext, ➔ Struktur und ➔ Income) für ein Programm in einem bestimmten Feld/an einem bestimmten Ort/zu einem bestimmten Zeitpunkt aussehen, welche Bedarfe bei ➔ Zielgruppen vorliegen und ob die vorhandenen Rahmenbedingungen die Durchführung eines Programms ermöglichen.
Programm	Ein Bündel von Maßnahmen, das aus einer Folge von Aktivitäten/Interventionen besteht, basierend auf einem Set von Ressourcen. Sie sind auf bestimmte (in der Regel bei bezeichneten ➔ Zielgruppen) zu erreichende ➔ Resultate gerichtet.
Prozess des Programms	Mit dem Prozess des ➔ Programms ist die Durchführung der Interventionen/Maßnahmen gemeint, die im ➔ Konzept zur Zielerreichung vorgesehen sind. Dies kann z.B. die Durchführung von Beratungsstunden sein.
Prozessnutzen der Evaluation	Nicht nur die Ergebnisse, die die Evaluation erbringt, sondern auch ihre Durchführung selbst kann positiven Nutzen für das ➔ Programm, Auftraggeber/-innen, am Programm ➔ Beteiligte und von ihm Betroffene haben. Diesen Nutzen der Evaluation nennt man Prozessnutzen. Beispielsweise können ➔ Zielgruppenmitglieder, die an der Durchführung der Evaluation, z.B. bei der Klärung der Programmziele, mitwirken, ihre Handlungsfähigkeiten erweitern.
Rechenschaftslegung (engl. *accountability*)	Die Darstellung von Erfolg und Effektivität eines ➔ Programms nach außen, zumeist eine Auflage an die Programmdurchführenden. Zu diesem Zweck wird häufig eine Evaluation herangezogen, die Daten für die Grundlegung der Bewertung liefern kann.
Resultate des Programms	Sammelbezeichnung für ➔ Outputs, ➔Effekte, ➔ Outcomes, ➔Impacts und die ➔ Wirkungen eines Programms.
Stakeholder des Programms	➔ Beteiligte und Betroffene

Steuerungsfaktoren
(engl. *advance organizer*)

Steuerungsfaktoren sind die orientierenden Elemente, die Evaluatoren/-innen heranziehen, um eine Studie in Gang zu setzen und schrittweise auszurichten. Diese können Teil des ➔ Evaluationsgegenstandes sein, z.B. Ziele oder Spannungsthemen, aus sozialwissenschaftlichen Theorien stammen (z.B. „Kontext-Mechanismus“-Verbindungen) oder aus der Evaluationstheorie (z.B. Nutzungen der Evaluation).

Struktur

Meint Bedingungen, die beim Träger (oder: darüber hinausgehend in einem Verbundsystem) eines ➔ Programms vorliegen, wie z.B. seine Rechtsform, seine Kapitalausstattung und seinen Finanzierungsmix, seine Personalstruktur, Qualitätsmanagementsystem usf. Die lediglich mittelfristig veränderbare Struktur ist eine relevante Durchführungsbedingung. Geldgeber/-innen können durch eine gezielte Trägerauswahl bei der Vergabe von Programmen gezielt Einfluss auf Strukturbedingungen nehmen.

summative Evaluation
vs. formative Evaluation

Die Leistung einer summativen Evaluation ist es, zu einem ➔ Evaluationsgegenstand eine zusammenfassende Bilanz zu ziehen. Ihr Zweck ist es häufig, grundlegende Entscheidungen über den Evaluationsgegenstand zu ermöglichen.

Verwendbarkeit des Programms
(auch: Nutzen des Programms, engl. *worth*)

Im Gegensatz zu ➔ Güte wird mit Verwendbarkeit der Gebrauchswert eines ➔ Evaluationsgegenstandes für bestimmte Anwender/-innen in bestimmten Situationen zu bestimmten Zeitpunkten bezeichnet. Die Verwendbarkeit eines Programms ist von der Homogenität oder Äquivalenz von ➔Kontext, ➔ Struktur und ➔ Income abhängig (eine Prüfung der Übertragbarkeit der ➔ Evaluationsergebnisse, z.B. auf andere Kontexte oder auf die Gesamtheit des Bundesgebietes, ist als eigenständiger Schritt erforderlich).

Werte (soziale)

Werte sind zunächst biographisch erworbene, kulturell basierte Dispositionen von Individuen, bestimmte Umstände/Lösungen/Handlungen anderen vorzuziehen. Deskriptiv beschreiben sie die tatsächlichen Dispositionen von Menschen, normativ solche (sozialen) Dispositionen, die mit unterschiedlichen Begründungen als verbindlich gelten oder eingefordert werden. Werte sind stärker auf längerfristiges Handeln bezogen und auch emotional verankert; sie können mit kurzfristigeren und rationaler begründeten (z.B. materiellen) Interessen in Widerspruch geraten. Werte im sozialen Kontext werden von Personenmehrheiten (Gruppen, Aggregaten, Teilkulturen) getragen. Für sozialpolitische ➔Programme sind sie relevant, da differierende Werte sozialer Gruppen sowohl zu differierenden Programmzielen (z.B. bezüglich eines wünschbaren Maßes der gesellschaftlichen Einkommensdifferenzierung) als auch zu differierenden Programminterventionen (z.B. zwischen Überredung, Anreiz oder Zwang) führen. Damit ergeben sich für die Evaluation von Programmen unterschiedliche, möglicherweise unvereinbare Bewertungskriterien, was die Zentralität von Werten für Evaluationstheorie- und Praxis unterstreicht.

wertedistanziert

Ein ➔ Evaluationsmodell kann wertedistanziert genannt werden, wenn die Werturteilsfrage aus dem Evaluationsprozess ausgeklammert wird. Die theoretische Rahmung der Evaluation und die Umsetzung in empirische Untersuchungen verlaufen nach strikten Regeln, deren Einhaltung einer als wertfrei unterstellten intersubjektiven Überprüfung zugänglich gemacht wird. Wertfragen werden als vorentschieden oder außerhalb von/nachgängig zu Evaluationen zu entscheiden gesetzt.

wertepositioniert

Wertepositionierte ➔ Evaluationsmodelle gehen explizit davon aus, dass Gesellschaften, Verbände und Organisationen durch starke strukturelle Machtungleichgewichte, soziale und ökonomische Ungleichheit geprägt sind. Wertpositionierte Evaluationen sollen gegen diese bestehende Wertehegemonie ein Gegengewicht bilden. Deshalb entscheiden sich Vertreter/-innen dieser Ansätze für eine bestimmte Wertposition. Sie tun dies insbesondere advokatorisch für Benachteiligte, also z.B. von Armut bedrohte oder arme Menschen.

wertepriorisierend

Ein ➔ Evaluationsmodell kann wertepriorisierend genannt werden, wenn im Evaluationsprozess den angenommenen starken Wertdifferenzen in Gesellschaft, Organisationen, Netzwerken etc. mit einem Prozess der methodisch vorbereiteten Prioritätensetzung begegnet wird. Dabei soll durch die ➔ Betroffenen und Beteiligten ein möglichst breiter Wertekonsens erarbeitet werden, der eine nützliche Evaluation fördert.

werterelativistisch

Werterelativistische ➔ Evaluationsmodelle messen sozialen Werten die zentrale Bedeutung bei der Planung, Durchführung und Nutzung von Evaluationen zu. Sie arbeiten Wertgemeinsamkeiten zwischen ➔ Beteiligten und Betroffenen und besonders Wertekonflikte in allen Phasen der Evaluation heraus und halten die bestehenden Spannungen aufrecht, ohne Partei zu nehmen. Die verschiedenen Werte sollen offen und idealerweise herrschaftsfrei verhandelt werden. Dies impliziert eine gewisse Stärkung ansonsten einfluss- oder artikulationsschwacher Beteiligter.

Wirkungen des Programms vs. Auswirkungen).

Veränderungen oder Stabilisierungen bei den ➔ Zielgruppen und in der Gesellschaft, die ursächlich auf das ➔ Programm zurückgeführt werden. Erwünschte und unerwünschte, vorhergesehene und nicht-vorhergesehene ➔Resultate, also z.B. ➔ Outputs und ➔ Outcomes, soweit diese ursächlich auf das Programm zurückgeführt werden. Wenn Wirkungen im ➔ Konzept des Programms nicht vorhergesehen sind und Ihr Entstehen nachweislich auf das Programm zurückgeht, spricht man auch von ➔Nebenwirkungen

Wirkungseinschätzung

Experten/-innen oder Beteiligte und Betroffene im Umfeld eines Programms werden gebeten, dessen Wirksamkeit unter Nutzung ihres Erfahrungswissens einzuschätzen. Streng genommen handelt es sich um subjektive Meinungen über Resultate/Wirkungen, die durch Interessen, fachliche Vorlieben oder andere Faktoren beeinflusst sein können, ohne dass dies kontrolliert ist. Wirkungseinschätzung ist somit die schwächste, manchmal jedoch einzig realisierbare und finanzierbare Form der Wirkungsfeststellung.

Wirkungsorientierte Evaluation

Eine Evaluation wird dann als wirkungsorientiert bezeichnet, wenn sie bei der jeweiligen ➔Evaluationsfunktion darauf achtet, dass intendierte➔ Wirkungen im ➔ Evaluationsplan zentral berücksichtigt werden. Formative wirkungsorientierte Evaluationen unterstützen es, dass das ➔ Programm seine beabsichtigten Wirkungen optimal erzielt. Summative wirkungsorientierte Evaluationen erheben vorrangig (intendierte und nicht-intendierte) Wirkungen, um eine Entscheidung über das Programm zu fundieren. Wirkungsorientierte Evaluationen stellen Auftraggebenden, Programmverantwortlichen oder anderen wichtigen Beteiligten solche Informationen zur Verfügung, die sie nutzen können, um entweder die Programmwirkungen fundiert planen zu können (Wirkungsmodellierung), um sich von der Wirksamkeit der Programmaktivitäten überzeugen zu können (Wirkungsfeststellung) oder um eine Bewertung des Programms in Abwägung der negativen wie der positiven Wirkungen vornehmen zu können (Wirkungsidentifizierung).

wirkungsfeststellende Evaluation Eine ➔ Evaluationsfunktion, die i.R. der wirkungsorientierten Evaluation die Resultate von Programmen untersucht. Die Evaluation soll in diesem Falle belegen, dass ein empirisch nachgewiesenes ➔ Resultat (z.B. durch vorher langfristig Arbeitslose gelungene Arbeitsaufnahmen) auf das ➔ Programm zurückzuführen ist. Dies erfolgt je nach gewähltem Evaluationsmodell und gewählten Evaluationsressourcen auf unterschiedlichen Wegen, z.B. durch Befragung von ➔ Beteiligten und Betroffenen oder von Experten/-innen (Wirkungseinschätzung), durch Rückführung im Rahmen eines differenzierten, ggf. wissenschaftlich begründeten Wirkungsmodells (Wirkungsmodellierung) oder durch Isolation der Programmwirkungen im Rahmen (quasi-) experimenteller Designs (Wirkungsnachweis). Im Unterschied zu den drei vorgenannten Ausprägungen bezieht die ➔Wirkungsidentifizierung besonders nicht-intendierte Wirkungen in die Betrachtung ein.

Wirkungsidentifizierung Neben den angezielten/vorhergesehenen Wirkungen lösen Programme auch nicht Vorhergesehenes aus. Dabei handelt es sich insbesondere um nicht-intendierte Wirkungen, die im Rahmen der politischen oder Programm-Leitzziele (nachträglich) entweder als positiv (erwünscht) oder negativ (unerwünscht) bewertet werden können. Soweit sich Evaluationen (auch) damit beschäftigen, derartige Nebenwirkungen aufzuspüren, leisten sie „Wirkungsidentifizierung“ (vgl. als primär auf Wirkungsidentifizierung gerichtete Evaluationen die „Zielfreien Evaluationsmodelle“).

Wirkungsmodellierung Gewünschte Wirkungen im Rahmen von „logischen Modellen“ der Programmumsetzung werden differenziert beschrieben und visualisiert. Durch theoretische/logische Argumentationsmuster wird ein plausibles Programmmodell hergestellt, welches den Weg von den ➔Inputs und ➔ Incomes über den Prozess des ➔Programms bis zu den empirisch messbaren Resultaten (z.B. ➔Outcomes) vorzeichnet. Dies kann – muss aber nicht – mit Vorher-Nachher-Messungen oder (quasi-)experimentellen Designs empirisch untermauert werden.

Wirkungsnachweis, empirischer	Für den Wirkungsnachweis müssen zum ersten die Programmresultate empirisch gemessen werden. Außerdem erfordert dies den *empirischen* Nachweis, dass diese Resultate ursächlich durch das Programm/seine Interventionen ausgelöst sind. Dies geschieht in der Regel durch Wahl eines (quasi-)experimentellen Designs, das den Vergleich von Resultaten zwischen einer Gruppe von Programmteilnehmenden (Experimentalgruppe) mit einer möglichst ähnlichen Gruppe, die nicht am Programm teilnimmt (Kontrollgruppe), ermöglicht.
Zielgruppen des Programms	Hiermit werden die Personen bezeichnet, an die sich ein ➔ Programm richtet, d.h. bei denen schließlich ➔ Outcomes ausgelöst werden sollen.
Zweck der Evaluation	➔ Evaluationszweck

7.2 Standards für Evaluation der Deutschen Gesellschaft für Evaluation (DeGEval-Standards)[117]

„Evaluationen sollen vier grundlegende Eigenschaften aufweisen: Nützlichkeit – Durchführbarkeit - Fairness - Genauigkeit.

N Nützlichkeit

Die Nützlichkeitsstandards sollen sicherstellen, dass die Evaluation sich an den geklärten Evaluationszwecken sowie am Informationsbedarf der vorgesehenen Nutzer und Nutzerinnen ausrichtet.

N1 Identifizierung der Beteiligten und Betroffenen: Die am Evaluationsgegenstand beteiligten oder von ihm betroffenen Personen bzw. Personengruppen sollen identifiziert werden, damit deren Interessen geklärt und so weit wie möglich bei der Anlage der Evaluation berücksichtigt werden können.

N2 Klärung der Evaluationszwecke: Es soll deutlich bestimmt sein, welche Zwecke mit der Evaluation verfolgt werden, so dass die Beteiligten und Betroffenen Position dazu beziehen können und das Evaluationsteam einen klaren Arbeitsauftrag verfolgen kann.

N3 Glaubwürdigkeit und Kompetenz des Evaluators / der Evaluatorin: Wer Evaluationen durchführt, soll persönlich glaubwürdig sowie methodisch und fachlich kompetent sein, damit bei den Evaluationsergebnissen ein Höchstmaß an Glaubwürdigkeit und Akzeptanz erreicht wird.

N4 Auswahl und Umfang der Informationen: Auswahl und Umfang der erfassten Informationen sollen die Behandlung der zu untersuchenden Fragestellungen zum Evaluationsgegenstand ermöglichen und gleichzeitig den Informationsbedarf des Auftraggebers und anderer Adressaten und Adressatinnen berücksichtigen.

N5 Transparenz von Werten: Die Perspektiven und Annahmen der Beteiligten und Betroffenen, auf denen die Evaluation und die Interpretation der Ergebnisse beruhen, sollen so beschrieben werden, dass die Grundlagen der Bewertungen klar ersichtlich sind.

---

117 Die Darstellung der Standards ist entnommen aus: Deutsche Gesellschaft für Evaluation (DeGEval), 2002, S. 8-11; http://www.degeval.de.

N6 Vollständigkeit und Klarheit der Berichterstattung: Evaluationsberichte sollen alle wesentlichen Informationen zur Verfügung stellen, leicht zu verstehen und nachvollziehbar sein.

N7 Rechtzeitigkeit der Evaluation: Evaluationsvorhaben sollen so rechtzeitig begonnen und abgeschlossen werden, dass ihre Ergebnisse in anstehende Entscheidungsprozesse bzw. Verbesserungsprozesse einfließen können.

N8 Nutzung und Nutzen der Evaluation: Planung, Durchführung und Berichterstattung einer Evaluation sollen die Beteiligten und Betroffenen dazu ermuntern, die Evaluation aufmerksam zur Kenntnis zu nehmen und ihre Ergebnisse zu nutzen.

D Durchführbarkeit

Die Durchführbarkeitsstandards sollen sicherstellen, dass eine Evaluation realistisch, gut durchdacht, diplomatisch und kostenbewusst geplant und ausgeführt wird.

D1 Angemessene Verfahren: Evaluationsverfahren, einschließlich der Verfahren zur Beschaffung notwendiger Informationen, sollen so gewählt werden, dass Belastungen des Evaluationsgegenstandes bzw. der Beteiligten und Betroffenen in einem angemessenen Verhältnis zum erwarteten Nutzen der Evaluation stehen.

D2 Diplomatisches Vorgehen: Evaluationen sollen so geplant und durchgeführt werden, dass eine möglichst hohe Akzeptanz der verschiedenen Beteiligten und Betroffenen in Bezug auf Vorgehen und Ergebnisse der Evaluation erreicht werden kann.

D3 Effizienz von Evaluation: Der Aufwand für Evaluation soll in einem angemessenen Verhältnis zum Nutzen der Evaluation stehen.

F Fairness

Die Fairnessstandards sollen sicherstellen, dass in einer Evaluation respektvoll und fair mit den betroffenen Personen und Gruppen umgegangen wird.

F1 Formale Vereinbarungen: Die Pflichten der Vertragsparteien einer Evaluation (was, wie, von wem, wann getan werden soll) sollen schriftlich festgehalten werden, damit die Parteien verpflichtet sind, alle Bedingungen dieser Vereinbarung zu erfüllen oder aber diese neu auszuhandeln.

F2 Schutz individueller Rechte: Evaluationen sollen so geplant und durchgeführt werden, dass Sicherheit, Würde und Rechte der in eine Evaluation einbezogenen Personen geschützt werden.

F3 Vollständige und faire Überprüfung: Evaluationen sollen die Stärken und die Schwächen des Evaluationsgegenstandes möglichst vollständig und fair überprüfen und darstellen, so dass die Stärken weiter ausgebaut und die Schwachpunkte behandelt werden können.

F4 Unparteiische Durchführung und Berichterstattung: Die Evaluation soll unterschiedliche Sichtweisen von Beteiligten und Betroffenen auf Gegenstand und Ergebnisse der Evaluation in Rechnung stellen. Berichte sollen ebenso wie der gesamte Evaluationsprozess die unparteiische Position des Evaluationsteams erkennen lassen. Bewertungen sollen fair und möglichst frei von persönlichen Gefühlen getroffen werden.

F5 Offenlegung der Ergebnisse: Die Evaluationsergebnisse sollen allen Beteiligten und Betroffenen soweit wie möglich zugänglich gemacht werden.

G Genauigkeit

Die Genauigkeitsstandards sollen sicherstellen, dass eine Evaluation gültige Informationen und Ergebnisse zu dem jeweiligen Evaluationsgegenstand und den Evaluationsfragestellungen hervorbringt und vermittelt.

G1 Beschreibung des Evaluationsgegenstandes: Der Evaluationsgegenstand soll klar und genau beschrieben und dokumentiert werden, so dass er eindeutig identifiziert werden kann.

G2 Kontextanalyse: Der Kontext des Evaluationsgegenstandes soll ausreichend detailliert untersucht und analysiert werden.

G3 Beschreibung von Zwecken und Vorgehen: Gegenstand, Zwecke, Fragestellungen und Vorgehen der Evaluation, einschließlich der angewandten Methoden, sollen genau dokumentiert und beschrieben werden, so dass sie identifiziert und eingeschätzt werden können.

G4 Angabe von Informationsquellen: Die im Rahmen einer Evaluation genutzten Informationsquellen sollen hinreichend genau dokumentiert werden, damit die Verlässlichkeit und Angemessenheit der Informationen eingeschätzt werden kann.

G5 Valide und reliable Informationen: Die Verfahren zur Gewinnung von Daten sollen so gewählt oder entwickelt und dann eingesetzt werden, dass die Zuverlässigkeit der gewonnenen Daten und ihre Gültigkeit bezogen auf die Beantwortung der Evaluationsfragestellungen nach fachlichen Maßstäben sichergestellt sind. Die fachlichen Maßstäbe sollen sich an den Gütekriterien quantitativer und qualitativer Sozialforschung orientieren.

G6 Systematische Fehlerprüfung: Die in einer Evaluation gesammelten, aufbereiteten, analysierten und präsentierten Informationen sollen systematisch auf Fehler geprüft werden.

G7 Analyse qualitativer und quantitativer Informationen: Qualitative und quantitative Informationen einer Evaluation sollen nach fachlichen Maßstäben angemessen und systematisch analysiert werden, damit die Fragestellungen der Evaluation effektiv beantwortet werden können.

G8 Begründete Schlussfolgerungen: Die in einer Evaluation gezogenen Folgerungen sollen ausdrücklich begründet werden, damit die Adressaten und Adressatinnen diese einschätzen können.

G9 Meta-Evaluation: Um Meta-Evaluationen zu ermöglichen, sollen Evaluationen in geeigneter Form dokumentiert und archiviert werden.“

7.3 Interviewleitfaden der Experten/-innen-Interviews

Allgemein gefragt: Welche Qualitätsmerkmale sollen wirkungsorientierte Evaluationen erfüllen? Wir werden nachher genauer einzelne Aspekte fokussieren. Bitte beschränken Sie sich erst einmal auf die methodischen Aspekte.

Fragestellung: Modelle der Wirkungsorientierten Evaluation in der Armuts- und Reichtumsberichterstattung

Was ist Ihr Ansatz der wirkungsorientierten Evaluation? Wie arbeiten Sie (Mit welchem Modell der wirkungsorientierten Evaluation?)?

Welche Besonderheiten/Eigenschaften hat dieser Ansatz?

Welche anderen deutlich zu unterscheidenden Ansätze sind Ihnen bekannt?

Nennen Sie 2 oder 3 spezifische Vorteile und Nachteile, die Sie bei einem oder mehreren dieser Ansätze sehen.

Fragestellung: Methoden

Welche Qualitätsanforderungen bezüglich der Methodik sollen wirkungsorientierte Evaluationen erfüllen?

Welche spezifischen methodischen Ansätze bei der Datenerhebung und -auswertung wenden Sie an? Welche sollten aus Ihrer Sicht in der Zukunft verstärkt angewendet werden? Aus welchem Grund?

Wovon hängt es ab, welche Methode Sie wählen?

Fragestellung: Nutzung der Studien

Was soll Ihrer Meinung nach in erster Linie durch die Studien der Wirkungsorientierten Evaluation erreicht werden?

Woran machen Sie fest, dass eine zweckgemäße Nutzung von Studien stattfindet?

Wie weit ist es anzustreben, dass Entscheidungen über Programme möglichst unmittelbar durch die Ergebnisse von Evaluationen beeinflusst werden (konzeptioneller – instrumenteller Nutzen)?

Welche Chancen, welche Risiken sehen Sie in der unmittelbaren Nutzung der Ergebnisse?

Welcher weitere Nutzen von Studien ist Ihnen wichtig?

Fragestellung: Adressaten der Ergebnisse

Wer sind die primären, wer sind weitere relevante Adressaten der Ergebnisse solcher Studien?

Wie erfolgt durch diese eine angemessene oder optimale Nutzung von Ergebnissen?

Fragestellung: Besonderheiten des Feldes Armuts- und Reichtumsberichterstattung

Welche spezifische Sichtweise auf die Armuts- und Reichtumsberichterstattung haben Sie?/Wenn es sie nicht geben würde in der Armuts- und Reichtumsberichterstattung, was würde dann fehlen?

Legen Sie auf eine spezifische Zielgruppe einen besonderen Schwerpunkt?

Welche speziellen Anforderungen an wirkungsorientierter Evaluation gibt es, die aus den Besonderheiten des Feldes der Armuts- und Reichtumsberichterstattung resultieren?

Welche Rollen spielen/welchen Zweck erfüllen gängige Indikatoren zur Messung von Armut (im Bereich Einkommen/Vermögen bzw. Lebenslagendimensionen bei Wirkungsorientierter Evaluation). Welche Indikatoren halten Sie für besonders relevant?

Fragestellung: Werteberücksichtigung

Die Festlegung von Armutsschwellen ist ja bekanntlich ein normativer Akt. Welche Rolle spielen darüber hinaus normative Aspekte bei wirkungsorientierten Evaluationen?

Gibt es Phasen der Evaluation, in denen diese eine besonders große Rolle spielen?

Wie weit soll auf den normativen Akt hingewiesen werden/soll er transparent gemacht werden?

2. Teil

Wirkungsanalysen können Maßnahmen untersuchen, die mehr oder weniger starken Interventionscharakter haben. Schätzen Sie bitte den Interventionscharakter der Maßnahmen, die Sie untersuchen, auf einer Skala von 1-10 ein. Wert 1 = geringer/kein Interventionscharakter, Wert 10 sehr starker Interventionscharakter.

Kennen Sie besondere Ansätze der wirkungsorientierten Evaluation für Maßnahmen mit stärkerem Interventionscharakter?

Welche speziellen Qualitätsanforderungen an wirkungsorientierter Evaluation gibt es bei Maßnahmen, die stärkeren Interventionscharakter haben?

Zum Nutzen: Was soll ihrer Meinung nach in erster Linie durch Studien der wirkungsorientierten Evaluation erreicht werden, die Maßnahmen mit stärkerem Interventionscharakter untersuchen?

Welche Rolle spielen normative Aspekte bei Evaluationen von Maßnahmen mit stärkerem Interventionscharakter? Spielen normative Aspekte bei Evaluationen von Maßnahmen mit stärkerem Interventionscharakter eine andere bzw. größere/kleinere Rolle?

7.4 Frageleitfaden der Fokusgruppen

Nennen Sie bitte Ihren Namen und erläutern Sie kurz Ihren Beitrag zur Armuts- und Reichtumsberichterstattung!

Wie sind Sie bisher mit Evaluationen in Berührung gekommen? An dieser Stelle interessiert uns die gesamte Bandbreite an Möglichkeiten: von Evaluationen gehört, Evaluationsberichte gelesen, für die eigene Arbeit genutzt, selbst in Auftrag gegeben, etc.

Welche Nutzenerwartungen waren mit diesen Evaluationen verbunden? Was kann Evaluation überhaupt für die Armuts- und Reichtumsberichterstattung leisten? Wie müssten Evaluationen optimalerweise angelegt sein, bzw. welche Qualitäten müssen Evaluationen aufweisen, damit dieser Nutzen erreicht wird?

Stellen Sie sich vor, Sie vergeben eine Evaluation im Bereich Armuts- und Reichtumsberichterstattung: Was müsste bei der Auftragsvergabe geklärt werden? Was ist aus Ihrer Sicht bei der Vergabe und Planung von wirkungsorientierten Evaluationen wichtig, damit die Ergebnisse für den Armuts- und Reichtumsberichterstattung optimal genutzt werden können?

Wie sollte der Prozess von der Ergebnisdarstellung bis zur politischen Nutzung optimal gestaltet werden?

Wie sollten die Ergebnisse aufbereitet sein bzw. vermittelt werden, damit für Sie als Auftraggeber größtmöglicher Nutzen entsteht?

Welche besonderen Kontextbedingungen /politische Besonderheiten/Rahmenbedingungen müssen bei der Evaluation von Maßnahmen zur Reduzierung von Armut beachtet werden?

Werte beeinflussen die Politik zur Armutsvermeidung bekanntermaßen maßgeblich (z.B. bei der Festlegung von Armutsgrenzen). Welche Wertespannungen gibt es Ihrer Meinung nach in der Armuts- und Reichtumsberichterstattung und wie sollte optimalerweise damit umgegangen werden? Wie sollen Ihrer Meinung nach Wertespannungen in der Evaluation berücksichtigt werden?

Fällt Ihnen etwas ein, was in der gesamten Diskussion bisher kaum angesprochen worden ist, bei der Durchführung von wirkungsorientierten Evaluationen im Bereich Armuts- und Reichtumsberichterstattung aber unbedingt beachtet werden sollte?

7.5 Übersicht über die an den Erhebungen beteiligten Personen

Nr.	Institution	Frau/Herr	Name
1.	Bundesministerium für Bildung und Forschung	Frau	Albrecht-Lohmar
2.	Soziologisches Forschungsinstitut Göttingen e.V	Herr	Dr. Bartelheimer
3.	Zentrum für Sozialpolitik der Universität Bremen	Frau	Dr. Buhr
4.	Institut für Sozialforschung und Gesellschaftspolitik	Herr	Dr. Engels
5.	Bundesministerium für Familie, Senioren, Frauen und Jugend	Herr	Fischer
6.	Bundesministerium für Gesundheit und Soziale Sicherung	Frau	Godschalk
7.	Bundesministerium für Bildung und Forschung	Herr	Gros
8.	Fachhochschule Darmstadt	Herr	Prof. Dr. Hanesch
9.	Universität Frankfurt	Herr	Prof. Dr. Hauser
10.	Evangelische Fachhochschule Rheinland-Westfalen-Lippe	Herr	Prof. Dr. Huster
11.	Bundesministerium für Bildung und Forschung	Frau	Heuermann-Busch

12.	Deutscher Verein für öffentliche und private Fürsorge, Frankfurt	Herr	Höft-Dzemski
13.	Bundesministerium für Gesundheit und Soziale Sicherung	Frau	Hommes
14.	Bundesministerium für Familie, Senioren, Frauen und Jugend	Frau	Dr. Icken
15.	Bundesministerium für Wirtschaft und Arbeit	Herr	Jülicher
16.	Bundesministerium für Gesundheit und Soziale Sicherung	Herr	Klebula
17.	Minsterium für Wirtschaft und Arbeit, NRW	Frau	Kocks
18.	Bundesministerium für Wirtschaft und Arbeit	Herr	Kolb
19.	Deutsches Institut für Wirtschaftsforschung e.V.	Herr	Dr. Krause
20.	Bundesministerium für Gesundheit und Soziale Sicherung	Herr	Lutz
21.	Hans Böckler Stiftung	Frau	Dr. Mezger
22.	Bundesministerium für Wirtschaft und Arbeit	Herr	Monse
23.	Bundesministerium für Gesundheit und Soziale Sicherung	Herr	Dr. Mozet
24.	Bundesministerium für Gesundheit und Soziale Sicherung	Herr	Münch
25.	Institut der deutschen Wirtschaft	Frau	Peter
26.	Fachhochschule Frankfurt	Herr	Prof. Dr. Reis
27.	Bundesministerium für Wirtschaft und Arbeit	Frau	Rüschkamp
28.	Bundesministerium der Justiz	Herr	Sabel
29.	Gesellschaft für Sozialwissenschaftliche Frauenforschung (GSF e.V.)	Frau	Prof. Dr. Sellach
30.	Bundesministerium für Gesundheit und Soziale Sicherung	Frau	Schmidt
31.	Deutsches Institut für Wirtschaftsforschung e.V.	Herr	Dr. Schupp
32.	Statistisches Bundesamt	Herr	Seewald
33.	Bundesministerium für Verkehr, Bau- und Wohnungswesen	Herr	Dr. Völker
34.	Fachhochschule Pforzheim	Herr	Prof. Dr. Volkert
35.	Deutsches Institut für Wirtschaftsforschung e.V.	Herr	Prof. Dr. Wagner
36.	Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit	Herr	Dr. Wehrspann
37.	Robert Koch Institut	Herr	Dr. Ziese

8 Anhang II: Modelle der Evaluation

Die in Kap. 2.3 bereits referierten Modelle der Evaluation werden im Folgenden noch einmal ausführlicher in Tabellenform vorgestellt.

8.1 Programmziel-gesteuerte Evaluation

Synonyme	Objectives-Based Studies, Effektivitätsstudien.
Charakterisierung Modell	Der klassische Evaluationsansatz, der vor allem aus dem Bereich Bildung/Erziehung stammt, will überprüfen, ob die Ziele eines Programms erreicht werden, indem Resultate des Programms mit den vorher festgelegten Zielen verglichen werden. Der Grad der Zielerreichung ist ausschlaggebendes Kriterium, um die Güte und Verwendbarkeit des Programms zu beurteilen. Es muss entschieden werden, welcher Grad der Zielerreichung als akzeptabel oder als ein Erfolg des Programms betrachtet wird. Zu diesem Zweck können Evaluationsergebnisse vergleichbarer Programme oder Aussagen der Grundlagenforschung herangezogen werden. Intendierte Ziele des Programms müssen operationalisiert werden (sie sollen beschreiben, wie die resultierenden Effekte z.B. bei Zielgruppen beobachtet werden können), damit sie schließlich mit geeigneten Methoden gemessen werden können. Idealerweise liegen explizite Zielformulierungen bei Beginn des Programms vor. Die Ziele werden üblicherweise von den Programmentwicklern/-innen formuliert, eventuell bei lokalen Umsetzungen noch stärker differenziert. Es ist jedoch nicht unüblich, dass Ziele auch bei etablierten Programmen nicht in geeignet operationalisierter Form vorliegen. Sie müssen dann zu Beginn der Evaluation von den Evaluatoren/-innen in Zusammenarbeit mit den Programmverantwortlichen und -mitarbeitenden formuliert werden. Es muss geprüft werden, zu welchem Grad das Programm selbst zur Erreichung der Ziele beigetragen hat und zu welchem Grad mögliche andere Einflüsse beigetragen haben. Dieses Problem kann mit methodischen Mitteln (s.u.) und/oder über den Versuch, kausale Zusammenhänge zwischen Programmaktivitäten und beobachtbaren Effekten zu formulieren (vgl. Programmtheoriebasierte Evaluation), angegangen werden. Möglichst sollen in das kausale Hypothesengerüst auch nicht zum Programm gehörige Kontextvariablen einbezogen sein. Vorteilhaft für die Qualität solcher Studien ist es, wenn zur Bewertung des Grads der Zielerreichung des Programms auch geprüft wird, inwieweit die Bedarfe der Zielgruppen getroffen werden, wenn auch nach evtl. Nebeneffekten gesucht wird. Programmziel-gesteuerte Evaluationen sind häufig intern (bei den Programmverantwortlichen selbst) angelegt.
Steuerungsfaktoren	Die operationalen Lern-, Einstellungs- oder Verhaltensziele, die durch das Programm bei den Zielgruppen ausgelöst werden sollen.
Hauptzwecke	Feststellen, ob mit der Durchführung des Programms die gesetzten Ziele erreicht werden.
Quellen der Fragestellungen	Fragestellung ist, in welchem Maße die identifizierten Ziele eines Programms (durch das Programm selbst) erreicht werden. Üblicherweise richten sich Hauptfragestellungen auf Ziele mittlerer Komplexität, die durch Teilfragestellungen konkretisiert werden, welche auf die operationalen Ziele Bezug nehmen.

Typischerweise eingesetzte Methoden	Es können die verschiedensten Methoden der (Lern-)Ziel- oder Leistungsmessung angewendet werden (klassisch: Klausuren oder Tests; vgl. AERA u. a., 1999). Es sind auch standardisierte Verhaltensbeobachtungen möglich, wenngleich sehr aufwändig und daher nur für höchst prioritäre Ziele leistbar. Idealerweise sind solche Tests validiert und standardisiert. Um die Zielerreichung den Programmaktivitäten selbst zuschreiben zu können, werden ggf. Vorher-Nachher-Vergleiche von Daten, Kontrollgruppen-Designs oder noch komplexere Aufbauten notwendig (vgl. die Modelle 0 und 2.3.1.2).
Stärken	Der Ansatz ist da besonders gut anwendbar, wo es darum geht, klar zugeschnittene Projekte mit expliziten, klaren Zielsetzungen zu beurteilen. Dies gilt bspw. dann, wenn ein geprüftes und als Pilot bereits evaluiertes Programm in Art vieler identischer „Klone“ national an vielen Standorten implementiert wird. Es sind dann sehr leicht Vergleiche der Resultate unterschiedlicher lokaler Umsetzungen möglich. Bei unterdurchschnittlichen Ergebnissen wird somit schnell Nachbesserungsbedarf in der Programmumsetzung identifiziert. Die Logik dieser Evaluationsstudien ist dem „gesunden Menschenverstand“ leicht zugänglich, es ist sogar in vielen Fällen so, dass unter „Evaluation“ nichts anderes als die Programmziel-gesteuerte Evaluation verstanden wird. Der Ansatz setzt keine Experimentierfreude oder ein großes Engagement der Stakeholder voraus.
Schwächen	Sollten zu Beginn der Evaluation keine expliziten Programmziele vorliegen, müssen sie innerhalb einer Zielklärung nachträglich formuliert werden. Dies kann ein langwieriger Prozess sein. Es ist unsicher, ob das Programm tatsächlich durch diese Ziele gesteuert ist. Programme ohne eine Kultur laufender Zielrevision können schnell an den aktuellen Bedarfen vorbeigehen. Es besteht auch die Gefahr, dass schlechte Zielerreichung gemessen wird, obwohl die Programme Gutes leisten. Eine Programmziel-gesteuerte Evaluation schließt eine Prüfung der Angemessenheit der intendierten Programmziele nicht zwingend ein. Aussageschwach sind Programmziel-gesteuerte Studien, in denen zur Bewertung weder der Programmprozess noch Bedarfe von Zielpersonen oder evtl. Nebenwirkungen des Programms erfasst werden. Eine Erfassung der Zielerreichung allein erscheint häufig nicht ausreichend, um die Güte und Verwendbarkeit des Programms zu beurteilen. Der „Versuchsaufbau“ zur Datensammlung gemäß diesem Modell kann sehr komplex werden und die Maßstäbe, die an die Validität der Erhebungsinstrumente angelegt werden, sehr hoch, weil die Anforderungen der empirischen psychologischen oder Sozialforschung gelten (AERA-Teststandards). Ein großer Teil der Zeit und Energie der Evaluation muss in die Auswahl und Prüfung der Instrumente und die Datensammlung gesteckt werden und geht für andere Aufgaben (z.B. Commitment der Stakeholder gewinnen, Ergebnisse vermitteln) verloren. Ergebnisse, die eindeutig einen Erfolg oder ein Versagen des Programms feststellen sollen, werden von Programmmitarbeitenden evtl. als bedrohlich empfunden, was zu Blockaden und bewussten Verzerrungen von Daten führen kann. Dies kann auch dazu führen, dass ausschließlich leicht erreichbare Ziele formuliert werden. Innovation wird evtl. gehemmt. Außer zur evtl. notwendigen Zielklärung werden Stakeholder nicht einbezogen, was die Nutzung der Evaluationsergebnisse wenig begünstigt.
Werteberücksichtigung	Wertedistanziert.
Wichtige Quellen	Tyler (1950), Weiss (1972, in D: 1974), Suchman (1967).

8.2 Experimentaldesign-gesteuerte Evaluation

Synonyme	---
Charakterisierung Modell	Es wird kontrolliert, ob/in welchem Ausmaß ein Programm (und nicht andere Faktoren) die operational bestimmten Zielgrößen bei der Zielgruppe (z.B. durchschnittlicher Einkommenszuwachs in €/erreichter Lebensstandardindex von x) ursächlich ausgelöst hat. Zu diesem Zweck werden die Zielgrößen für die am Programm teilnehmende Gruppe (Experimentalgruppe) gemessen und den gemessenen Werten für eine Kontrollgruppe, die am jeweiligen Programm/an einem ähnlichen Programm nicht teilgenommen hat, gegenüber gestellt. Die Zuweisung zur Experimentalgruppe einerseits, zur Kontrollgruppe andererseits geschieht nach dem Zufallsprinzip. Bei genügend großen Stichproben können die gemessenen Zielgrößen (z.B. Durchschnittswerte für beide Gruppen) unmittelbar verglichen und damit Aussagen über die Wirksamkeit des Programms gemacht werden.
Steuerungsfaktoren	[siehe Programmziel-gesteuerte Evaluation] Da Zufallsstichproben vollständige Datensätze über potentielle Programmteilnehmer/-innen voraussetzen, orientiert sich die Anlage des Experiments außerdem an der Vollständigkeit, Zugänglichkeit und Aktualität dieser Datensätze.
Hauptzwecke	Überprüfen, in welchem Maße das Programm tatsächlich zur Erreichung der gesetzten Ziele beiträgt.
Quellen der Fragestellungen	Die operationalen Ziele, die das Programm anstrebt, werden zu Fragestellungen umformuliert und bei der Programm- und der Kontrollgruppe gemessen (In welchem Ausmaß wird das Ziel X bei der Programmgruppe einerseits, der Kontrollgruppe andererseits erreicht?). Im Mittelpunkt steht die Fragestellung, ob/in welchem Umfang das Programm „einen Unterschied macht“.
Typischerweise eingesetzte Methoden	Benötigt werden vorrangig quantitative, möglichst metrisch skalierte Daten, welche fortgeschrittene statistische Kontroll-, Gewichtungs- und Auswertungsprozeduren erlauben. Diese können über standardisierte Erhebungsbogen oder Fragebogen gewonnen werden. Besonders zeit- und Kosten sparend kann sich auswirken, wenn für die Zuordnung von Personen zur Experimental- und Kontrollgruppe auf vorhandene Daten zugegriffen werden kann (z.B. die Sozialhilfegeschäftsdaten der Träger der Sozialhilfe; vgl. Kap. 4.2). Hinzu kommen standardisierte schriftliche Erhebungen, wenn statistisch nicht erfasste relevante Merkmale zusätzlich zu gewinnen sind.
Stärken	Experimentelle Designs bieten sich für Programme an, die in hohem Umfang standardisiert, unter stabilen Rahmenbedingungen und mit hohen Teilnehmendenzahlen durchgeführt werden. Die Strategie der Gruppenbildung nach Zufall weist bei potentiellen Nutzern hohe Plausibilität in Bezug auf eine unabhängige Messung der Programmwirkung auf („interne Validität“). Bei genügenden Stichprobengrößen repräsentieren die Stichproben ein verkleinertes Abbild der Grundgesamtheit der Teilnehmenden (evtl. auch der Nicht-Teilnehmenden) und erlauben somit eine Verallgemeinerbarkeit der Ergebnisse. Die experimentelle Wirkungskontrolle ist insbesondere für Modellversuche oder sehr deutliche Veränderungen in bereits existierenden Programmen geeignet. Diese Art des Evaluationsdesigns ist für Außenstehende relativ leicht verständlich und ein aus den Naturwissenschaften bekannter und bewährter Ansatz.

Schwächen	Aufwändige Wirkungskontrollen und Wirkungsvergleiche setzen von vornherein mit operationalisierten Zielen entwickelte Programme voraus, sind andernfalls schwer durchführbar. Die Fragestellungen werden stark auf die Resultatsvariablen konzentriert; der Einfluss sowohl der Teilnehmer/-innen-Merkmale, die Frage der Passung des Programmangebots auf die Bedarfe der Zielgruppe, die Merkmale des Programmprozesses sowie nicht-indendierte Wirkungen selbst bleiben weitgehend ausgeblendet. So ist oft nicht klärbar, ob das Programm mit seinem Konzept und/oder mit dessen Umsetzung, also seinem Prozess, gescheitert ist. Vorausgesetzt wird eine weitgehende Planbarkeit von Programmprozessen; der Einfluss personaler Faktoren des Programmpersonals (Motivation, Konfliktsteuerungskompetenz) oder der das Programm tragenden Organisation (Betriebsklima, Güte des Qualitätsmanagements) werden vernachlässigt. Experimentelle Studien sollten bereits mit Programmbeginn und nicht erst bei bereits implementierten Programmen durchgeführt werden, da es ansonsten vielfach zu Veränderungen der Programmdurchführung kommt (*Randomization Bias*). Die Übertragbarkeit der Ergebnisse auf andere sozioökonomische Kontexte und Programmausgestaltungen ist problematisch (Problem „externer Validität“). Dies zeigt, dass die Bedingungen eines „sozialen“ Experiments sich von Laborbedingungen der Naturwissenschaften deutlich unterscheiden und damit auch dieser „Goldstandard“ je nach Kontext fraglich sein kann. Dieser Evaluationsansatz bedarf der nachhaltigen Unterstützung der beteiligten politischen und verwaltungstechnischen Akteure, wofür im Modell keine Verfahren ausgewiesen sind. Bei kleineren Grundgesamtheiten ist es kritisch, eine ausreichend repräsentative Experimental- und Kontrollgruppe bilden zu können. Zwar wird eine sichere Entscheidungsgrundlage für vorhandene Programme geschaffen, doch gibt es kaum Anhaltspunkte dazu, wie Programme zu verbessern sind. Mit randomisierte Experimenten werden auch ethische Probleme verbunden: So wird es oft abgelehnt, Kontrollgruppenmitglieder von Vorteilen oder speziellen Fördermaßnahmen fern zu halten. Diesem Argument kann entgegengehalten werden, dass aber auch eine Programmteilnahme negative Effekte für die Teilnehmer haben kann, bspw. Stigmatisierungseffekte. Andererseits gibt es Widerstand gegen eine „zwangsweise“ Zuweisung von Programmgruppenmitgliedern in Maßnahmen, die nicht speziell auf ihre Bedarfe abgestimmt sind. Aus verschiedenen Gründen kommt es zu Teilnahmeverweigerung oder –abbruch bzw. die Mitglieder der Kontrollgruppe nehmen an anderen/ähnlichen Programmen teil (*Substitution Bias*), was die Vergleichbarkeit der beiden Gruppen und damit die Gültigkeit der Evaluationsergebnisse einschränkt. Faktisch wird oft so verfahren, dass die Aufnahme in die Gesamtgruppe auf freiwilliger Basis erfolgt und dann aus diesen Freiwilligen per Zufall zwei Gruppen gebildet werden. Dabei kommt es über die Freiwilligkeit der Erstmeldung bereits zu erheblichen Selektionswirkungen und Verzerrungen gegenüber der Gesamtpopulation, was die Gültigkeit der Ergebnisse mindert. Der Aussagegehalt von Feldexperimenten ist insbesondere in dynamischen, sich entwickelnden Bereichen, wie es auch die verschiedenen Interventionsfelder im Bereich der Armuts- und Reichtumspolitik betrifft, gefährdet. Sinnvoller Weise müssen experimentelle Wirkungskontrollen mit nicht-experimentellen Evaluationsmodellen kombiniert werden, um weitere Aussagen über die Verallgemeinerbarkeit der Ergebnisse treffen zu können. Dieses Evaluationsmodell ist nicht geeignet, wenn Programme mit einem gesetzlichen Anspruch zur Teilnahme flächendeckend eingeführt wurden.
Wertberücksichtigung	Wertedistanziert: Das (Feld-)Experiment gilt bei seinen Vertretern/-innen als das „objektive“, wertneutrale Vorgehen per se. Die mit der Auftragsvergabe oder der erfolgten/angestrebten, z.B. gesetzlichen, Regelung beabsichtigten sozialpolitischen Ziele werden nicht auf ihre Wertegebundenheit befragt. Programme gelten als (neutrale) Mittel für vorab, im demokratischen Prozess gesetzte Ziele.
Wichtige Quellen	Bortz/Döring (2002), Heckman/Smith (1996) Shadish/Cook/Campbell (2002),

	Moffit/Ver Ploeg (2001).

8.3 Quasi-Experimentaldesign-gesteuerte Evaluation

Synonyme	Nicht-Experimentelle Wirkungskontrolle.
Charakterisierung Modell	Im Unterschied zur experimentellen Wirkungskontrolle arbeitet die quasi-experimentelle Wirkungskontrolle ohne randomisierte Kontrollgruppe. Es gibt verschiedenste quasi-experimentelle Ansätze. Alle fokussieren auf den Unterschied zwischen einer Situation mit und einer ohne Intervention. Dieser Unterschied kann sowohl auf der Ebene der Individuen als auch auf höheren Aggregationsebenen, wie z.B. Kommunen, analysiert werden. In der Literatur werden folgende Haupttypen quasi-experimenteller Verfahren unterschieden: Verfahren ohne Vergleichsgruppe und solche mit Vergleichsgruppen. Im letzteren Fall wird eine Vergleichssituation, die beschreibt, was sich ohne die betrachtete Intervention entwickelt hätte, konstruiert (kontrafaktische Situation). Es werden bspw. für die Programm-Teilnehmer/-innen „statistische Zwillinge“ zu einer Vergleichsgruppe als Referenzgruppe zusammengefasst (Matching-Verfahren). Idealerweise unterscheidet die Vergleichssituation sich von der tatsächlichen Situation nur in dem Umstand der fehlenden Programmteilnahme. Zu diesem Zweck müssen viele Einflussgrößen für beide Situationen gleich gehalten werden. Um diese Einflussgrößen zu identifizieren bedarf es abgesicherter theoretischer Kenntnisse über diejenigen Faktoren (unanhängigen Variablen), welche auf die zu prüfende Wirkungen einen (besonders starken) Einfluss haben. Wenn sich Experimentalgruppe und Vergleichsgruppe dennoch in wichtigen Einflussvariablen unterscheiden, kann nachträglich eine statistische Gewichtung vorgenommen werden, um z.B. die gemessene Effektstärke in eine Effektstärke umzurechnen, die bei identischer Vergleichssituation aufgetreten wäre. Auch hierin gehen zu sichernde theoretische Vorannahmen ein. Bei einer quasi-experimentellen Wirkungskontrolle ohne Vergleichsgruppe können verschiedenste Pre- und Posttest-Designs durchgeführt werden. Verfahren mit Vergleichsgruppe kommen dem „Ideal“ des Experiments näher.
Steuerungsfaktoren	[siehe Programmziel-gesteuerte Evaluation] Insbesondere die Verfahren zur Berechnung der Vergleichssituation setzten vollständige Datensätze über Programmteilnehmer/-innen und Nicht-Programmteilnehmer/-innen voraus. Somit orientiert sich die Anlage von Quasi-Experimenten vielfach an der Vollständigkeit, Zugänglichkeit und Aktualität der relevanten Datensätze. Die Vollständigkeit der Datengrundlage steuert damit die Anlage der Untersuchung.
Hauptzwecke	[siehe Programmziel-gesteuerte Evaluation]
Quellen der Fragestellungen	[siehe Experimentaldesign-gesteuerte Evaluation]
Typischerweise eingesetzte Methoden	Es können verschiedenste statistische und ökonometrische Verfahren eingesetzt werden, wobei für eine Evaluationsstudie zumeist eine der folgenden Methoden gezielt eingesetzt wird: • Zeitreihenanalysen (Interventionsmodelle) • Querschnittsvergleiche • „difference-in-difference“-Methoden • Matching-Verfahren Letztere Methode ist die prominenteste und aufwändigste; wird von manchen Autoren auch mit dem quasi-experimentellen Design gleichgesetzt.
Stärken	Durch in jüngster Zeit verbesserte statistische Methoden sowie verbessertes Datenmaterial können mit diesem Evaluationsmodell zunehmend stabile Ergebnisse produziert werden. Kann auch zur Extrapolation bestehender Ergebnisse in andere Situationen (Zielgruppen-Income, Programm- und Kontextveränderungen) genutzt werden. Ist vor allem bei längerem, großflächigem standardisierten (unverändertem) Programmeinsatz sinnvoll.

	Dieser Evaluationsansatz wird als direkte Alternative zur Experimentaldesign-gesteuerte Evaluation gesehen; dies gilt vor allem, wenn die Voraussetzungen für letzteres Design nicht gegeben sind.
Schwächen	Eine der schwächsten Varianten dieses Ansatzes ist ein Vorher-Nachher-Vergleich, der kaum Zurechnungen der Effekte zur Intervention zulässt. Matching-Verfahren hingegen sind nur auf der Basis von großen Datensätzen möglich, die ausreichende Informationen über Programm-Teilnehmer/-innen sowie Nicht-Teilnehmer/-innen liefern. Eventuell aufwändige Arbeiten zur Vorbereitung der Datensätze für weitere statistische Verfahren sind notwendig. Ist somit auch nur für großflächige Programme sinnvoll einsetzbar. Diese Variante des Evaluationsansatzes kann jedoch nur dann eingesetzt werden, wenn kein gesetzlicher Anspruch auf Programmteilnahme besteht, da ansonsten keine Vergleichsgruppe gebildet werden kann. Es müssen in den Berechnungen jedoch auch die Maßnahme-Heterogenität und die regionale Heterogenität ausreichend berücksichtigt werden. Vielfach liegen keine ausreichenden Informationen für Nicht-Teilnehmer/-innen vor, dann ist es bei Matching-Verfahren nicht möglich „statistische Zwillinge“ zu finden. Gibt keine Hinweise zur Programmverbesserung. Die Aussagekraft der Ergebnisse ist von den Annahmen der Analysesituationen abhängig. Sehr strikte (auch unzutreffende) Annahmen führen zu geringeren verbleibenden Unsicherheiten. Es ist eine elaborierte und anerkannte Theorie über den Zusammenhang zwischen Randbedingungen, Programmprozess und Wirkungen erforderlich. Diese muss auch Differenzierungen für bestimmte Teilgruppen der Gesamtzielgruppe enthalten, da die Varianz zwischen Teilgruppen in der Experimentalgruppe höher sein kann als die zwischen Experimental- und Kontrollgruppe. Die teilweise komplizierten Annahmen sowie statistischen Verfahren und bestehende Unsicherheiten müssen explizit gemacht und erklärt werden; ansonsten bleiben verschiedene Ergebnisse von quasi-experimentellen Wirkungskontrollen auf derselben Datenbasis unverständlich und ihr Zustandekommen ist den vorgesehenen Nutzern/-innen der Evaluationsergebnisse nicht nachvollziehbar (Gefahr der Nicht-Nutzung). Die Existenz verschiedenster quasi-experimenteller Verfahren führt zu Unübersichtlichkeit. Die methodologischen Fragen können von den Ergebnissen ab lenken.
Wertberücksichtigung	Wertedistanziert.
Wichtige Quellen	Heckman/LaLonde/Smith (1999), Hujer/Caliendo (2000), Shadish/Cook/Campbell (2002).

8.4 Programmkosten/-nutzen-gesteuerte Evaluation

Synonyme	Kosten-Nutzen-Analysen, Kosten-Effektivitäts-Analysen.
Charakterisierung Modell	Es werden (monetäre) Kosten und Nutzen eines Programms gegenübergestellt bzw. Kosten und Nutzen verschiedener Programmalternativen analysiert und verglichen. Hierfür können ökonomische Messgrößen im Vordergrund stehen, aber auch psycho-soziale Effekte betrachtet werden. In einer idealen Kosten-Nutzen-Analyse werden alle realen Kosten und Nutzen vollständig berücksichtigt. Hierzu gehören somit direkte und indirekte sowie tangible und intangible Kosten und Nutzen. Es lassen sich verschiedene Analyseperspektiven unterscheiden: individuelle und soziale (institutionelle oder volkswirtschaftliche) Bilanzierungen. Die Entscheidung für die Perspektivenwahl der Kosten-Nutzen-Analyse wird zumeist durch die Auftraggeber bestimmt (bspw. individuelle/Programm-/staatliche Perspektive). Wenn sich die Wirkungen nicht unmittelbar in monetären Werten messen lassen und es hierfür auch keine klaren Marktsubstitute gibt **–** was vielfach für sozialpolitische Programme zutrifft **–** ist eine Aufzählung und ausführliche Beschreibung der Effekte notwendig. Bei dieser Variante handelt es sich dann um so genannte Kosten-Effektivitäts-Analysen oder Kosten-Wirksamkeits-Analysen. Ein Programm wird dann als effizient bezeichnet, wenn der Programmnutzen die Kosten übersteigt bzw. wenn ein Programm gegenüber einem anderen Programm eine bessere Kosten-Nutzen-Relation aufweist.
Steuerungsfaktoren	Zentral sind die Programmziele; Steuerungsfaktoren sind die anfallenden Input-Kosten eines Programms sowie die Nutzen im Sinne der Zielerreichung und Nebenwirkungen eines Programms.
Hauptzwecke	Entscheidungsfindung, Zeit- und Alternativenvergleich; die errechneten Kosten-Nutzen-Relationen können mit solchen von ähnlichen Programmen verglichen werden und die Produktivität eines Programms in ökonomischen Messgrößen bestimmt werden. Die Kosten-Nutzen-Analyse erhöht die Zurechenbarkeit der politischen Verantwortung und trägt zur Aufdeckung inkonsistenter Entscheidungen bei.
Quellen der Fragestellungen	Fragestellungen beziehen sich auf den Vergleich von Kosten und Nutzen, ggf. differenziert nach bestimmten Kosten- Nutzenarten; es besteht die Wahl zwischen den Varianten einer Kosten-Nutzen-Analyse i.e.S. und Kosten-Effektivitäts-Analyse; nach der Perspektivenwahl gilt es für das betreffende Programm möglichst vollständig alle Kosten und Nutzen zu erheben.
Typischerweise eingesetzte Methoden	Üblicherweise werden Gegenwartswerte berechnet. Für die Kostenseite kann teils auf bestehende Kostenrechnungselemente zurückgegriffen werden. Diese müssen vielfach ergänzt und auf die entsprechenden Bezugsgrößen hin berechnet werden. Bei Kosten-Effektivitäts-Analysen werden Outcomes nicht monetarisiert, müssen sich jedoch auf klare, messbare Programmziele beziehen (Zielerreichungsgrade). Bei Kosten-Nutzen-Analysen i.e.S. werden jedoch die Nutzen in Geldeinheiten ausgedrückt. Sozialwissenschaftliche Methoden sind insbesondere gefordert, um den Zusammenhang zwischen Prozessen, Outputs und Outcomes offen zu legen. Hierzu gehören bspw. experimentelle, quasi-experimentelle Designs und Korrelationsstudien.
Stärken	Eine Entscheidung kann auf der Basis von Kosten-Nutzen-Analysen fokussiert vorbereitet werden. Einzelne Kosten- und Nutzenkomponenten werden transparent gemacht und weitestgehend beziffert. Im Gegensatz zu vielen anderen Evaluationsmodellen werden hier die Kosten eines Programms explizit als zentrale Analyseeinheit genutzt und gehen mit hohem Gewicht in dessen Beschreibung und Bewertung ein. Neben dem Vergleich von alternativen Programmen können Kosten-Nutzen-Analysen auch zur besseren Nutzung von Ressourcen verwendet werden. Dieses Evaluationsmodell kann auch gut mit anderen Modellen kombiniert werden und ist gerade durch die Analyse der Kostenseite, die im Rahmen anderer Evaluationsmodelle vernachlässigt wird, zu empfehlen.

Schwächen	Unter Umständen beleuchtet eine Kosten-Nutzen-Analyse einseitig eine Perspektive. Zur Behebung dieser Schwäche können jedoch auch Kosten-Nutzen-Analysen aus verschiedenen Perspektiven durchgeführt und verglichen bzw. eine hohe Aggregationsebene gewählt werden, die vielfältige Kosten-Nutzen-Aspekte enthält. Nur dadurch können verschiedenste Opportunitätskosten integriert werden. Die Ermittlung von Kosten-Nutzen-Relationen beruht auf teilweise zweifelhaften Annahmen und unsicheren Realitäten. Zudem müssen die intendierten und die nicht intendierten Wirkungen eindeutig zugeordnet werden, so dass auch nur der tatsächlich durch das Programm erzielte Nutzen den entsprechenden Kosten gegenübergestellt wird. Diese Bedingung ist vielfach schwierig zu erfüllen. Des Weiteren ist es insbesondere bei sozialpolitischen Programmen nicht einfach, nicht-monetäre in monetäre Werte umzuwandeln. Hier kann alternativ der Einsatz von Kosten-Effektivitäts-Analysen bei mehreren Programmen mit ähnlichen Zielen sinnvoll sein. Allerdings setzt auch dies klare Zielformulierungen der Programme voraus, die – wie an anderer Stelle erörtert – vielfach fehlen (siehe die Ausführungen zur Programmziel-gesteuerten in Kap. 8.1 und Programmtheorie-gesteuerten Evaluation; Kap. 8.6). Der Vergleich von Programmen auf der Basis von Kosten-Nutzen-Analysen gestaltet sich schwierig, da auch entsprechende Kenntnisse über die Programme sowie über den wirtschaftlichen und sozialen Kontext bekannt sein müssen. Die Qualität des Kosten-Nutzen-Ansatzes ist vor allem von der Analyse der Programmwirkungen abhängig. Deshalb ist es vielfach sinnvoll, dieses Evaluationsmodell mit anderen zu kombinieren, die insbesondere auf die Feststellung der Wirkungen fokussieren sowie die Implementation betrachten, was wiederum zu einem sehr großen Aufwand führt.
Wertberücksichtigung	Als zentraler Wert ist die Kostenwirksamkeit von Programmen gesetzt, im Sinne ökonomischer Verausgabung vorhandener Mittel. Die Ziele, auf die Nutzen bzw. Effektivität des Programms in Relation rekurrieren, werden als gegeben (vom Auftraggebenden/Programmverantwortlichen) übernommen und nicht auf ihre soziale oder kulturelle Wertebasis hinterfragt – insofern ist die Kosten-Nutzen-Analyse eher wertedistanziert. Es werden zumeist vorrangig der Informationsbedarf der Auftraggeber/-innen bzw. der verantwortlich am Programm Beteiligten bedient. Hierbei unterstützen Kosten-Nutzen-Analysen deren Wert- und Interessenpositionen; bei einer Entscheidung über Fortführungen von Programmen ist ausschlaggebend, wie und durch wen Schwellenwerte für Kosten, die bspw. nicht überschritten werden dürfen und Nutzen, die mindestens erfüllt werden sollen, festgelegt werden (da es auch sinnvoll sein kann Programme fortzusetzen, bei denen die Kosten den Nutzen übersteigen). Bei der Ermittlung der Nutzenseite können bei einer partizipativ gestalteten Kosten-Nutzen-Analyse auch die Werte der Teilnehmer/-innen einer Maßnahme einfließen; implizite Werte werden dann expliziert.
Wichtige Quellen	Levin/Mc Ewans (2001); Musgrave/Musgrave (1980).

8.5 Kontext-Mechanismus-gesteuerte Evaluation

Synonyme	Realistic Evaluation, Realist Evaluation.
Charakterisierung Modell	Dieses relativ junge Modell versteht sich explizit als wissenschaftlicher Ansatz, verortet Evaluation im Gegensatz zum Gros der aktuellen Modelle als Form angewandter Forschung, nicht als eigenständigen professionellen Zugang. Das Modell setzt sich gleichzeitig ab von den herkömmlichen (quasi-) experimentell basierten Ansätzen, deren vorgebliche Unfähigkeit, kausale Muster zu identifizieren und damit hilfreich zu sein für Entscheidungen und Verbesserungen von Programmen, postuliert wird. Gegenüber dieser positivistischen Sicht seien Programme nicht „Dinge“, die funktionierten oder nicht, sondern vielmehr enthielten sie bestimmte Ideen die für bestimmte Subjekte in bestimmten Situationen arbeiten. „What works for whom in what circumstances?“ Die Erklärung, warum und wie soziale Programme die Fähigkeit haben, Wandel auszulösen, steht im Mittepunkt des Forschungsinteresses. In Absetzung zu den experimentellen Ansätzen wird „Verursachung“ intern konfiguriert. Wirkmächtige Programme haben ein kausales Potential, welches das (gewünschte) Handeln der Zielgruppen auszulösen im Stande ist: „Rather it is the action of Stakeholders that makes them work and the causal potential of an initiative takes the form of providing reasons and ressources to enable program participants to change.“ (S. 215) Pawson/Tilley entwerfen ein theoretisches Gerüst welches sicher stellen soll, dass das durch Evaluation bereit gestellte Wissen unmittelbar relevant für das betrachtete soziale Programm und das soziale Problem ist, welches dieses bearbeitet: Grundlegend ist, dass das meiste, was ein soziales Programm produziert und wie es dies produziert, durch oft nicht beobachtbare, generative Kräfte hervor gebracht ist, d.h. durch ein internes Potential eines Systems, welches zum richtigen Zeitpunkt unter günstigen Bedingungen aktiviert wurde. „… program outcomes are generated by a range of macro and micro social forces. In social life the choice making behaviour of individuals in their different situations is fundamental to understanding their manifest patterns of behaviour … Program evaluations need to grasp how the changes introduced inform and alter the balance of the constrained choices of participants.“ (S. 215) Im Unterschied zum experimentellen Denken löst nicht das Programm unmittelbar die Wirkung aus, sondern es aktiviert einen Mechanismus, der regelhaft Wahlentscheidungen und Ressourcen von Individuen oder sozialen Aggregaten (wie z.B. Nachbarschaften) miteinander verbindet, so dass Wirkungen (Outcomes) als Lösungen sozialer Probleme zustande kommen. Die Beziehung zwischen einem kausalen Mechanismus und seinen Wirkungen liegt nicht fest sondern ist variabel. Entscheidend für Auslösen/Nicht-Auslösen des Mechanismus ist der soziale Kontext, in dem das Programm stattfindet, also z.B. die Normen, Beziehungsmuster, die Eingangsbedingungen der Programm-Teilnehmenden, die räumlichen Ressourcen in unterschiedlichen, sozial benachteiligten Stadtteilen.
Steuerungsfaktoren	Theoretische, empirisch schrittweise abgesicherte Annahmen über spezifische soziale Mechanismen, die in bestimmten sozialen Kontexten Wirkungen auslösen.
Hauptzwecke	Dieser Evaluationsansatz will differenzierte, auf jeweils enge Politikfelder spezifisch zugeschnittene Theorien über CMO-Konfigurationen weiterentwickeln und schrittweise absichern. Diese sollen das Denken von Politikern und Praktikern sowie der Öffentlichkeit informieren (konzeptioneller Nutzen) und außerdem möglichst unmittelbar die Planung und Durchführung von Programmen anleiten (*process of program realization*; instrumenteller Nutzen) bzw. politische Entscheider darüber informieren. Primäre Adressaten der Evaluation sind (Fach-)Politiker und leitende Mitarbeiter/-innen der Exekutive

	und auch Programmleiter/-innen. Pawson/Tilley sehen selbst erhebliche Hindernisse gegen diesen gedachten Nutzungsweg – „Policy makers often want simple answers. Learning methodology is not normally a priority!“ (S. 211) und bieten dabei keine eigenständigen Lösungswege als integralen Bestandteil ihres Modells an.
Quellen der Fragestellungen	Fragestellungen resultieren aus den Theorien (gefasst in CMO-Konfigurationen) der Forscher/-innen. Sie folgen der allgemeinen Formulierung „What works for whom in what circumstances?“. Da die Forscher spezialisierte Feldkenntnisse aufweisen sollen, also z.B. in Bereichen wie Obdachlosigkeit oder Analphabetismus substantielle Erfahrungen mitbringen, sind sie in der Lage, spezifizierte leitende Fragestellungen zu formulieren.
Typischerweise eingesetzte Methoden	Es gibt keine Präferenzen für bestimmte Methoden – in der Mehrzahl der Untersuchungsbeispiele liegen mit Rückfallquoten oder Diebstahlquoten auffällig leicht messbare Outcomes vor. Um die CMO-Konfigurationen aufzudecken bedienen sich Pawson und Tilley typischerweise strukturierter Interviews, wobei sie klare Annahmen dazu haben, welche Stakeholder zu welchen Aspekten über Wissen verfügen.
Stärken	Die besondere Chance dieses Evaluationsansatzes liegt darin, dass übertragbares, kumulatives Wissen aufgebaut wird, das für spezifische Felder sozialer Politik empirisch geprüfte „Konfigurationen von Context-Mechanismus-Outcomes“ ausweist (CMO-Configurations), die künftige Programmentwicklung und -umsetzung anleiten können. Die Komplexität des sozialen Handelns soll auf hohem Niveau erhalten und in feldspezifischen Theorien abgebildet werden, was kumulatives Wissen darüber verspricht, wie Programme wirksam angelegt werden können und wie Wirkung/Nicht-Wirkung detailliert und theoretisch gehaltvoll erklärt werden kann.
Schwächen	Der Ansatz ist bislang vorwiegend auf engen Feldern ausprobiert; für die Wertantagonismen (z.B. Straftäter-Opfer) bzw. ein sehr breiter Wertkonsens bestehen (dass Einbrüche in Wohnungen Sicherheit und Unversehrtheit in Frage stellen). Typisch für soziale Programme sind jedoch graduelle Wertkonflikte, so dass die Übertragbarkeit des Kontext-Mechanismus-gesteuerten Ansatzes hierauf noch zu prüfen ist. Das Modell konzentriert sich auf eine programmnahe Theorie der Verursachung und liefert kaum praktische Anweisungen, wie Evaluationen konkret geplant und durchgeführt werden sollen. Weitgehend wird eine theorielastige Fachsprache genutzt, die bei Politikern/-innen und Praktikern/-innen auf Widerstand treffen könnte. „Übersetzungen“ in die Sprache von Politikern/-innen und Programmverantwortlichen fehlen.
Werteberücksichtigung	Wertedistanziert – zwar versteht sich der Ansatz als Beitrag zur *piecemeal social reform* und sieht die evaluierten Sozialprogramme in einem Kontext von Machtauseinandersetzungen und Interessenskonflikten eingebettet, doch bleiben Wert- und Interessendivergenzen aus dem Evaluationsprozess selbst ausgeblendet. Zwar sei das, was als soziales Problem definiert werde „politically coloured“ (, doch seien soziale Regeln, Normen und Werte als Begrenzungen für die Wirksamkeit von Programmmechanismen hinzunehmen. Schließlich stellen die Beispiele in Pawson/Tilleys Buch Programme der Kriminalitätsprävention (z.B. Selbsthilfeaktivitäten beim Schutz gegen Einbrecher in strukturell benachteiligten Stadtteilen) oder der Rehabilitation von Strafgefangenen dar, für die charakteristisch ist, dass in der Gesellschaft ein sehr breiter Wertkonsens besteht, auch darüber dass den sozial geächteten Interessen und Werten der Straftäter in keiner Weise entgegenkommen werden soll.
Wichtige Quellen	Pawson/Tilley (1997); Kazi (2003).

8.6 Programmtheorie-gesteuerte Evaluation

Synonyme	Programm Theorie (*program theory*), logisches Modell (*logic model*), Theorie der Veränderung (*theory of change*), theoriebasierte Evaluation, oft verbunden mit Outcome-Measurement.
Charakterisierung Modell	Im besten Fall gehen diese Evaluationen von einer vorliegenden, sorgfältig entwickelten und geprüften Theorie aus, die das zu evaluierende Programm anleitet. Diese „Programmtheorie“ trifft z.B. Aussagen über die Wirkungszusammenhänge in bestimmten Kontexten, die zu den erwünschten Programmzielen führen. Den Evaluatoren/-innen hilft sie dabei, Fragestellungen, Instrumente und Methoden für ihre Untersuchung auszuwählen oder zu entwickeln. In den Fällen, wo eine solche Programmtheorie nicht vorliegt, ist es häufig Aufgabe der Evaluation, eine solche zusammen mit den Beteiligten zu entwickeln. Tatsächlich ist dies häufig zu leisten, was hohe Anforderungen z.B. an die Kenntnisse der Evaluatoren/-innen zum Politikfeld und zum aktuellen Stand der wissenschaftlichen Forschung stellt. Sie müssen die Elemente (Kontext, Ausgangslage, Interventionen, Resultate) sowie Richtung und Intensität der Beziehungen zwischen diesen Elementen identifizieren, welche die Programmlogik darstellen und die Evaluation anleiten können. Zentrales Element jeder Programmtheorie sind die Outcomes, also die bei den Zielgruppen ausgelösten Veränderungen bzw. Stabilisierungen in Wissen, Einstellung Verhalten sowie die Resultate im Sinne von Veränderungen/Stabilisierungen in der sozialen Umwelt der Zielgruppen (Nachbarschaften, Stadtteile ...). Da diese Voraussetzungen in der Praxis oft nicht gegeben sind, werden die Ansprüche an eine Programmtheorie oft reduziert z.B. bezüglich Vollständigkeit, formaler Widerspruchslosigkeit, Absicherung von Kausalitäten u. a. Gebräuchlich sind daher auch Bezeichnungen wie *program design*, *action plan*, „Programmlogik“, „logisches Modell“, „Programmmodell“. Stakeholder sollen als „Experten/-innen“ in die Formulierung der Programmtheorie einbezogen werden, auch damit ihr Commitment für die Evaluation sowie die Nutzung der Evaluationsergebnisse erreicht werden. Chen, der dieses Modell in den 90er Jahren verbreitet hat, setzt zwecks Konstruktion und Darstellung der Programmtheorie die Methode des „Logischen Modells“ ein, in der in sechs Schritten unter Einbezug der Stakeholder zunächst Informationen zusammen gestellt, Kontexte und zu lösende Probleme beschrieben, Elemente des logischen Modells gesammelt, definiert und zusammengesetzt werden und schließlich das Modell verifiziert wird. Als Informationsquellen für die Entwicklung der Programmtheorie werden neben den Stakeholdern auch wissenschaftliche Literatur, Evaluationsberichte zu ähnlichen Programmen, die Beobachtung des Programms selbst oder Befragungen externer Experten/-innen (in allen möglichen Kombinationen) eingesetzt. Die Explikation der Programmtheorie sollte, wenn möglich, *vor* dem Programmstart erfolgen. Wird zu diesem Zeitpunkt auch das Evaluationsdesign angelegt, können Konstruktionsfehler des Programms rechtzeitig erkannt und vermieden werden. Dabei ist die Formulierung der Programmtheorie auch bei bereits laufenden Programmen nicht unüblich. Die explizierte Programmtheorie stellt sich bildlich im einfachsten Fall als Diagramm mit Pfeilen (als Symbole für kausale Verbindungen) zwischen den einzelnen Elementen dar, sie kann auch um einflussreiche Kontext-Variablen erweitert werden. Eine Darstellung der Programmtheorie durch eine non-kausale, systemische Theorie ist ebenso möglich.
Steuerungsfaktoren	Finden sich in der Programmtheorie: die Mechanismen, die von Interventionen zu Outcomes führen, Kontextbedingungen, unabhängige, vermittelnde und abhängige Variablen, Kausalbeziehungen zwischen Programmbestandteilen.

Hauptzwecke	Neben der Beurteilung von Güte und Verwendbarkeit des Programms ist es Ziel zu bestimmen, in welchem Ausmaß das Programm theoretisch stimmig ist, zu verstehen, wie es funktioniert, also zu erklären, warum es Erfolg hat oder fehlschlägt (also zum Verständnis der Programmlogik beizutragen) und Vorschläge für die Weiterentwicklung der Programmtheorie und damit des Programms zu machen.
Quellen der Fragestellungen	Die Fragestellungen werden von der programmleitenden (vorliegenden oder durch die Evaluatoren/-innen mit den Stakeholdern entwickelten) Theorie abgeleitet: zu jedem der Elemente.
Typischerweise eingesetzte Methoden	Es gibt verschiedenste Methoden, die Programmtheorie zu identifizieren, bzw. abzubilden, wobei z.B. grundlegende Elemente und Beziehungen von Programmen beschrieben und als Folie für die Entwicklung der konkreten Programmtheorie angeboten werden. Häufig eingesetzt wird das „Modell der Programmlogik“, bei dem z.B. mit Hilfe von Flussdiagrammen Prozesse und Zusammenhänge zwischen Inputs und Outcomes dargestellt werden. In einigen Fällen wird auch auf die stark qualitativ geprägte *Grounded Theory* (Glaser/Strauss, 1967) zurückgegriffen, wobei hier – im Gegensatz zum Modell der Programmlogik – ausgiebige und systematische empirische Beobachtung und Analyse (mit Hilfe der einschlägigen Methoden) notwendig sind, um auf induktivem Weg eine befriedigende Programmtheorie zu konstruieren. Um die eigentlichen Evaluationsfragestellungen zu beantworten, können dann Programmtheorie und -praxis verglichen werden. Das Modell ist multimethodisch angelegt, arbeitet in der Phase der Theoriekonstruktion eher mit qualitativen, in der Programmüberprüfung mit quantitativen Daten.
Stärken	Sollte bereits eine fundierte Theorie des Programms vorliegen, kann sich die Evaluation auf ein glaubwürdiges Fundament stützen und zu dessen weiterer Verbesserung beitragen, indem Fragestellungen und Vorgehen abgeleitet werden und Ergebnisse vor dem Hintergrund der bestehenden Kenntnisse interpretiert werden. Sollte keine Theorie vorliegen, ist der Ansatz eine Anregung dazu, eine solche zu entwickeln, was für die Konzeption jeder Evaluation und darüber hinaus auch für die Programmentwicklung und -konsolidierung hilfreich ist. Diese Bemühungen sollten jedoch eindeutig als eine Annäherung an eine Programmtheorie verstanden und benannt werden. Eine Verbindung zwischen Grundlagenwissenschaft und Programm kann hergestellt werden, wenn die Konstrukteure der Programmtheorie über entsprechende Qualifikationen und Ressourcen verfügen. Dies kann auch zu einer Vereinheitlichung von Fachsprache führen und so den Austausch zwischen ähnlichen Programmen/Projekten und den Austausch zwischen den verschiedenen Programmbeteiligen unterstützen. Im Falle von Implementationsfehlern können Strategien zur Überwindung entwickelt werden. An den Punkten, an welchen das Programm Schwächen/Misserfolge hat, kann eine Unterscheidung zwischen einem „Theorie-Fehler“ und einem „Umsetzungs-Fehler“ gemacht werden. Bei jungen, gerade erst in der Implementation befindlichen Programmen ist es durch einen Vergleich zwischen Theorie und Praxis möglich abzuschätzen, ob das Programm den Vorhersagen entsprechend sich in der Praxis bewährt. Die Beschäftigung mit der Programmtheorie und Implementation des Programms führt unter den Programm-Mitarbeitenden idealerweise zu einer reflexiven Arbeitsweise (Prozessnutzen) und sind förderlich für die Zusammenarbeit unter den Mitarbeitenden.

Schwächen	Sollte keine Theorie des Programms in ausgearbeiteter Form vorliegen, kann die Aufgabe, eine solche zu erstellen, eine große zeitliche und Arbeitsbelastung für die Evaluatoren/-innen sein. Die eigentliche Aufgabe, eine Bewertung der Programmgüte und -verwendbarkeit vorzunehmen kann darunter leiden. Eine weitere Gefahr besteht darin, dass die Evaluatoren/-innen sich für die Konzeption des Programms verantwortlich machen, was nicht ihre Aufgabe ist, weil sie v.a. für dessen theoriegesteuerte Beschreibung und Bewertung verantwortlich sind. Darüber hinaus ist Vorsicht davor geboten, möglicherweise eine falsche Theorie aufzustellen, welche die Evaluation in die Irre leiten kann. Dies kann insbesondere dann passieren, wenn die Evaluatoren/-innen die Sichtweisen der Programmakteure unkritisch übernehmen oder, da ihnen ein Referenzrahmen, sei es aus geprüften wissenschaftlichen Theorien oder aus empirisch abgesicherten Programmtheorien, fehlt (vgl. Friedman 2001, S. 167ff). Leider muss davon ausgegangen werden, dass nur den wenigsten Programmen im Bereich der Sozialwissenschaften, bzw. Sozialpolitik eine fundierte Theorie zu Grunde liegt, so dass in den meisten Fällen die oben genannten Schwächen zum Tragen kommen können.
Werteberücksichtigung	Wertedistanziert (Bei der Formulierung der Programmtheorie sollten zwar die Stakeholder einbezogen werden, ein Umgang mit evtl. Wertekonflikten wird jedoch nicht ausführlich thematisiert).
Wichtige Quellen	Glaser & Strauss (1967), Weiss (1995), Bickman (1990), Chen (1990), Rogers (2000).

8.7 Spannungsthemen-gesteuerte Evaluation

Synonyme	Responsive (Fallstudien-) Evaluation.
Charakterisierung Modell	Responsive Evaluationen „antworten“ (am.: *respond*) auf Informationsanliegen der verschiedenen am Programm Beteiligten und Betroffenen. Sie erstellen eine dichte, vertiefte Beschreibung eines bestimmten Programms bzw. seiner lokalen Umsetzung („Fall“). Diese Beschreibung ist so organisiert, dass die Beteiligten relevante Informationen erhalten, diese vor dem Hintergrund ihrer jeweiligen Werte und Interessen beurteilen und Schlussfolgerungen für die künftige Programmplanung und -umsetzung ziehen können.
Steuerungsfaktoren	Die Abgrenzung des Gegenstandes erfolgt Schritt für Schritt im Evaluationsprozess. Wertgeladene „Spannungsthemen“ (*issues*) sind dabei die zentralen Steuerungsfaktoren. Sie kennzeichnen von den Beteiligten als problematisch, konfliktreich und ungelöst wahrgenommene Programmbestandteile.
Hauptzwecke	Es geht primär darum, ein Programm intensiv zu schildern und seine Wirkungsweise zu erhellen. Diese Schilderung ist so angelegt, dass Daten, Informationen und Interpretationen durch die Programmbeteiligten für eine intensive Auseinandersetzung mit dem Programm genutzt werden können. Vorrangig zielt die responsive Evaluation auf Programmverbesserung; sie kann ihre Berichte auch so ausrichten, dass sie Grundsatzentscheidungen über Programme ermöglichen.
Quellen der Fragestellungen	Die Konkretisierung der Fragestellungen erfolgt unter der Maßgabe, welche Informationen für die vorgesehenen Adressaten der Evaluation am interessantesten sind. Damit „antwortet“ der responsive Ansatz beständig auf die evaluativen Bedarfe der verschiedenen Adressaten und Adressatinnen, die sich im Verlauf der Evaluation auch verändern können. Dies erfolgt vielfach durch Gespräche, im Rahmen informell geprägter Treffen und bei anderen Feldkontakten der Evaluatoren/-innen. Dabei sollen auch die intendierten Nutznießer/-innen oder Zielgruppen eines Programms einbezogen werden; dies kann auch „stellvertretend“ durch das Evaluationsteam erfolgen. Der Kontext und der Prozess der Umsetzung des Programms werden intensiv beschrieben, auch dann, wenn der Fokus der Evaluation auf Wirkungen gelegt wird.
Typischerweise eingesetzte Methoden	Aus diesem Grundsatz der Multiperspektivität folgt, möglichst verschiedenartige und mehrere Methoden und Datenquellen parallel einzusetzen (mit einer Präferenz für qualitative Daten). Dies soll auch zur Gültigkeit der gewonnenen Daten beitragen. Für die responsiven Evaluatoren/-innen dient diese Triangulation weniger dazu, vorgefasste Interpretationen zu bestätigen, sondern dazu, zusätzliche Interpretationen aufzufinden. Es dominieren in der responsiven Fallstudie qualitative Zugänge: Beobachtung, dichte Beschreibung des Kontextes, narrative Verfahren, Interviews und Auswertung von Dokumenten. Charakteristisch für die Datenerhebung ist die Konzentration auf natürliche Situationen, die in einer Alltagssituation nachvollziehbar machen, wie das Programm abläuft und welche insbesondere kurzfristigen Wirkungen es hat. Zu diesem Zweck wurde eine spezifische Fallstudien-Methode entwickelt (Stake 1995).

Stärken	Chancen des Fallstudienansatzes liegen besonders darin, dass sie eine vertiefte Grundlage für einen erweiterten Diskurs über Armuts- und Reichtumspolitik unterstützen. Bei bundesweiten Programmen können responsive Fallstudien genutzt werden, um einzelne vertiefte Evaluationen durchzuführen, welche die Schnittstelle zwischen Programmprozess und Programmwirkungen erhellen. Die responsive Evaluation eignet sich besonders zur Unterstützung der Entwicklung und Förderung neuer Programme, da sie ihren Betrachtungsfokus flexibel und zeitnah an die oft schnellen Veränderungen in der Programmdurchführung anpassen kann. Bei unscharfem und – z.B. wegen starker Wertspannungen zwischen den Beteiligten – nicht abgeschlossenem Zielklärungsprozess kann der Fallstudienansatz zur Verdichtung von Programmkonzeptionen beitragen. Der responsive Ansatz eignet sich besonders zur Evaluation lokaler Programme oder lokaler Umsetzungen von regionalen bzw. bundesweiten Programmen. Experimentierende (Heiner 1998), an einer überschaubaren Zahl von Standorten stattfindende Modellprogramme können als „Fälle“ evaluiert werden.
Schwächen	Bei explizit wirkungsorientiert beauftragten Evaluationen kann es schwierig sein, den Fokus allein auf Wirkungen zu setzen. Ein typischer Fall ist, dass die Programmumsetzer es im Unterschied zum Auftraggeber für zentral erachten, den Prozess des Programms, z.B. seine Umsetzung von der Bundes- über die Landes- auf die lokale Ebene zu betrachten. Breit angelegte Programme auf Landes- oder Bundesebene lassen sich nur bedingt responsiv evaluieren, da zu viele Beteiligte und Betroffene in die Abstimmungsprozesse einzubeziehen wären (einen Ausweg bietet die nutzungsorientierte Evaluation; vgl. Kap. 2.3.3.2). Als begrenzter Ausweg ist es möglich, durch theoretische Vorannahmen gesteuerte Stichproben, z.B. aus einer Gesamtheit von mehreren hundert lokalen Umsetzungen, zu ziehen. Meist reichen jedoch die theoretischen Grundlagen speziell über die relevanten Beeinflussungsfaktoren von Programm-Prozessen und -wirkungen nicht, um eine Stichprobe auszuwählen, die eine Übertragbarkeit auf die Grundgesamtheit (Gesamtzahl stattfindender Umsetzungen) ermöglicht. Oft werden solche Fälle ausgewählt, von denen am meisten gelernt werden kann, wobei es weniger um Übertragbarkeit von Ergebnissen geht. Die methodische Offenheit – die sich im Fehlen von Verfahrensanweisungen, operational beschreibendem Vorgehen, systematischen Forschungen über den Fallstudienansatz ausdrückt – kann leicht zu einer nicht-adäquaten Anwendung führen.
Wertberücksichtigung	Die für die Fallstudie verantwortlichen Evaluatoren/-innen selbst fällen keine zusammenfassenden Urteile über das Programm, sondern überlassen dies den Adressaten/-innen der Berichte und Präsentationen. Dabei legen sie die Untersuchungen so an und verfassen die schriftlichen Berichte so, dass die Vielfalt und Differenz der Wertperspektiven möglichst erhalten bleibt. Das Programm soll so beschrieben werden, dass die unterschiedlichen Sichtweisen, Erfahrungen und Bewertungskriterien der Programmbeteiligten aus den Berichten deutlich werden. Die Programmevaluation kann auch darin gipfeln, kontrastierende Befunde und gegensätzliche Schlussfolgerungen zu liefern; die Bewertung wird den Beteiligten überlassen. Fairness der Evaluation bedeutet hier, Perspektiven auch dann transparent zu machen, wenn sie über geringe Artikulationsfähigkeit und -macht verfügen. Eine Parteinahme für arme und sozial desintegrierte Bevölkerungsgruppen im Sinne einer pointierten Privilegierung von Schlussfolgerungen und Empfehlungen ginge jedoch über die werterelativistische Grundposition des responsiven Ansatzes hinaus. Eine große Chance von responsiven Fallstudien liegt in ihren konzeptionellen Nutzungsmöglichkeiten als erfahrungsbasierte Grundlage für die Meinungsbildung über Themen von Armut und Reichtum.
Wichtige Quellen	Stake (1995), Greene/ABMGS (2001).

8.8 Dialoggesteuerte Evaluation

Synonyme	Konstruktivistische Evaluation; Fourth Generation Evaluation.
Charakterisierung Modell	Basierend auf der konstruktivistischen Erkenntnistheorie: Erkenntnis wird immer als individuelle Deutung eines einzelnen Menschen auf der Grundlage seiner Weltsicht und Kenntnisse angesehen. Eine intersubjektive Einigung auf bestimmte Deutungen ist möglich, ebenso eine Weiterentwicklung von Deutungen durch dialogischen Austausch und Anreicherung mit Informationen. Die universell gültige Beurteilung einer Deutung als „richtig“ oder „wahr“ ist jedoch ausgeschlossen. Demnach sind in der Dialoggesteuerten Evaluation alle Beteiligten und Betroffenen und auch die Evaluatoren/-innen „Erkenntnisinstrumente“, die zu einem geteilten, informierten Verständnis durch Auseinandersetzung untereinander und mit Informationen beitragen. Die Evaluation soll so gesteuert werden, dass sich möglichst alle Stakeholder einbringen, indem die Aufmerksamkeit der Stakeholder gewonnen, sie informiert und befähigt werden, ihre Umwelt (die im Bereich der Programmziele liegt) zu gestalten. Die Stakeholder sollen die Evaluation gemäß ihrer Interessenlage mit steuern, die Erarbeitung neuer Deutungen soll sich an ihren Bedürfnissen orientieren. Die Evaluatoren/-innen haben den Auftrag, die Stakeholder während des gesamten Evaluationsprozesses regelmäßig zu informieren, sie in alle Entscheidungen einzubeziehen und ihnen – besonders bei der Interpretation von Daten **–** beratend zur Seite zu stehen. Dialoggesteuerte Evaluation hat zwei Phasen, die sich überschneiden oder auch parallel verlaufen können. In der Entdeckungs-Phase beschreiben die Evaluatoren/-innen, was im Programm vor sich geht. Das kann verkürzt werden, wenn es bereits eine ausgearbeitete Darstellung mit einem hohen Gehalt an verarbeiteten Deutungen zum Programm gibt. Vorsicht ist geboten, wenn diese Deutungen einer einzigen „positivistischen Quelle“ entstammen. Die Integrations-Phase soll die vorliegende oder entwickelte Darstellung durch neue Informationen und Deutungen aktualisieren. Die neue Deutung soll passen, funktionieren, Erklärungshilfen bieten und selbst auch wieder offen für Veränderung sein. Der Dialog zunächst innerhalb der einzelnen Gruppen, dann zwischen ihnen, wird fortlaufend rückgekoppelt. Es soll soviel Konsens wie möglich hergestellt werden. Themen, zu denen kein Konsens erreicht werden kann, dürfen nicht übergangen werden, sondern sollen im Rahmen eines organisierten und evtl. durch die Evaluatoren/-innen moderierten Forums weiter verhandelt werden. Zur Klärung offener Fragen sollen die Evaluatoren/-innen zusätzliche hilfreiche Informationen/Daten sammeln. Durch dauerhaft ungelöste Gegensätze ergeben sich dabei ggf. immer wieder neue Ansatzpunkte für eine Evaluation. Das (immer als vorläufig zu verstehende) Ergebnis der Evaluation besteht eher in einer gründlichen, von den Beteiligten geteilten Beschreibung des Gegenstandes denn in quantitativen Messungen und Statistiken. Die Evaluatoren/-innen weisen darauf hin, dass es keine einzig richtige und beständige Deutung gibt. Die eigenständige Weiterführung der Evaluation durch die Beteiligten und Betroffenen wird angeregt (s.a. Kap. 2.3.4). Im Evaluationsbericht sollen nicht nur die Ergebnisse öffentlich, sondern auch die Vorgehensweisen der Evaluation transparent gemacht werden, damit diese auf ihre Qualität geprüft werden können. Guba und Lincoln haben eigene Qualitätskriterien für die Datenerhebung und Informationsgewinnung entwikkelt, die herkömmliche Gütekriterien für Untersuchungen ersetzen.
Steuerungsfaktoren	Die im Dialog geklärten Annahmen, Anliegen und Spannungsthemen der Stakeholder (*claims, concerns and issues)*. Die erkenntnistheoretische Setzung (es gibt keine ultimativen Ergebnisse, Antworten oder Interpretationen) erfordert eine gemeinschaftliche, kommunikative und immer wieder für Revi-

	sion offene schrittweise Entwicklung des Evaluationsablaufes.
Hauptzwecke	Die Vielfalt der Deutungen bzgl. des Programms unter den Betroffenen und Beteiligten soll deutlich gemacht werden und eine dialogische Auseinandersetzung darüber initiiert werden.
Quellen der Fragestellungen	Fragestellungen werden mit den Beteiligten und Betroffenen den Steuerungsfaktoren entsprechend gemeinsam entwickelt und im Verlauf der Evaluation evtl. verändert oder spezifiziert. Dabei können sich die Fragestellungen auf alle Aspekte des Programms beziehen, sie müssen es aber nicht.
Typischerweise eingesetzte Methoden	Eine ausgewogene Mischung aus quantitativen und besonders qualitativen Methoden (bspw. formulieren einer Programmtheorie, Diskussionen, ...), die jedoch alle dem Zweck der Sammlung von verschiedenen Deutungen zum Evaluationsgegenstand bzw. deren Gegenüberstellung/Weiterentwicklung dienen. Außerdem sollen sie die Kommunikation zwischen allen Beteiligten und Betroffenen unterstützen. Zusätzliche Daten, um die Deutungen der Stakeholder mit Daten zu konfrontieren, werden durch Interviews, Literaturrecherche oder Nutzung anderer, auch nicht-reaktiver Informationsquellen erhoben. Es kommen alle Methoden in Frage, mit einem Vorzug für qualitative dialogische Methoden.
Stärken	Der Evaluationsprozess und die Ergebnisse sind völlig transparent, das begünstigt die Akzeptanz und Nutzung der Ergebnisse. Alle Betroffenen und Beteiligten werden einbezogen. Selbst wenn kein ausreichender Konsens zur Orientierung von Datenerhebungen gefunden werden sollte, wird erwartet, dass allein der Verständigungsprozess positive Auswirkungen auf die Teilnehmenden und das Programm hat. Dadurch, dass Stakeholder als „Instrumente“ aktiv werden, müssen ggf. keine oder weniger eigenständige Instrumente entwickelt werden. Verschiedene Perspektiven auf und verschiedene Quellen von Informationen sind sichergestellt. Die Evaluation identifiziert keine Verantwortlichen für Erfolg oder Misserfolg eines Programms und ist so weniger bedrohlich.
Schwächen	Eine Dialoggesteuerte Evaluation erfordert, dass Auftraggeber/-innen, Stakeholder und Evaluatoren/-innen darin übereinstimmen, dass eine solche Evaluation praktikabel ist und dass sie in diesem Prozess zusammenarbeiten wollen. Alle müssen akzeptieren, dass das Design der Studie und die Fragestellungen sich erst in deren Verlauf herausbilden werden und dass Ergebnisse möglicherweise widersprüchlich, in jedem Fall wenig eindeutig sein werden. Ob ein Konsens gefunden wird, ist nicht vorauszusagen. Darüber hinaus, sind die Ergebnisse nicht langlebig, sondern der Prozess muss idealerweise immer weiter geführt werden. Ein gefundener Konsens ist nicht übertragbar auf andere Settings und Programme. Vielfach wird es politischen Entscheidern/-innen oder Programmmanagern/-innen schwer fallen, sich auf ein derart offenes, „unsicheres“ und nicht abschließbares Vorhaben einzulassen. Es bedarf einer festen Gruppe verschiedenster, engagierter Betroffener und Beteiligter, die zu kontinuierlicher Mitarbeit auch bei Austragung von Divergenzen bereit ist. Alle müssen offen sein für den Prozess von Entdeckung und Integration, was viel Engagement in der Reflexion auch der eigenen Standpunkte voraussetzt. Diese Forderung erscheint fast utopisch. Eine Gefahr für die Evaluation sind unzureichend informierte und uninteressierte Stakeholder, die „schlechte Datenquellen“ darstellen. Diese einzubeziehen und zu motivieren bzw. auszubilden kann die Evaluation leicht überfordern.
Werteberücksichtigung	Werterelativistisch. Alle Werte der Stakeholder werden als gleich berechtigt und wichtig angesehen und sollen gleichermaßen in einen Wertekonsens einfließen. Jedoch ist jeder Konsens immer wieder offen für neue Werte und wird nicht höher als die einzelnen Werte der Stakeholder geachtet.

Wichtige Quellen	Guba / Lincoln (1981, 1989).

8.9 Entscheidungsgesteuerte Evaluation

Synonyme	Decision/Accountability-Oriented Studies (Entscheidungsgesteuerte/Rechenschaftslegungsorientierte Evaluation).
Charakterisierung Modell	Die Evaluation soll präzise so geplant und terminiert werden, dass sie für im Voraus bestimmte Entscheidungssituationen vor, während oder nach der Programmdurchführung rechtzeitig die erforderlichen Informationen bereitstellt. Die Ergebnisse und Zwischenergebnisse sollen auf die zentralen Fragestellungen der Beteiligten an Entscheidungen zum Programm empirisch abgesicherte Antworten bereitstellen. Dies soll ermöglichen, dass Entscheidungen – sei es in Bezug auf die Verbesserung des Programms (formative Evaluationsleistung) oder auf die Einstellung, Fortführung oder Erweiterung/Verbreitung des Programms (summative Evaluation) informierter getroffen werden können. Die Entwicklung des Entscheidungsgesteuerten Ansatzes resultierte aus der vielfach belegten Erfahrung, dass Evaluationen, die unter Missachtung von Entscheidungssituationen geplant und durchgeführt werden, sehr häufig nicht genutzt werden. Das Modell stellt das erste in einer Reihe von Ansätzen dar, die Nutzung von Evaluationsergebnissen systematisch zu erhöhen. Mögliche Entscheidungssituationen, die durch bereitgestellte Informationen „unterfüttert“ werden sollen, sind z.B. die Feststellung der Bedarfe der Zielgruppen, die Formulierung der Programmziele, die Auswahl verschiedener Dienstleister für die Programmdurchführung, die Budgetierung und Ausstattung eines Programms mit Ressourcen etc. Eine andere typische Entscheidungssituation besteht darin, ob als „Pilote“ umgesetzte Prototypen eines Programms in ein Regelprogramm auf Bundesebene überführt werden sollen, oder für welche erprobten Prototypen dies geschehen soll. Auch die Rechenschaftslegung „War es angesichts der Programmresultate gerechtfertigt, die öffentlichen Mittel einzusetzen?“ ist eine typische Aufgabe für die Entscheidungsgesteuerte Evaluation. Besonders wichtig können Informationen über Alternativen, über Entwicklungen bei Unterlassung einer programmförmigen Bearbeitung des sozialen Problems oder Abschätzungen über Nettokosten oder Nettoerlöse einer Programmdurchführung sein. Auch die Prüfung von Programmplänen auf Konsistenz, Passung bzw. Entwicklung von Mitarbeiterqualifikationen, Kosten-Effektivität oder langfristige Outcomes können zentrale Leistungen der Evaluation sein. Die Evaluation kann sich somit entweder auf bestimmte Programmphasen oder die Frage der Programmverbreitung beschränken oder über den gesamten Programmverlauf hin stattfinden. Wird sie seit Programmbeginn durchgeführt, liegen frühzeitig Grundlagendokumente und Daten für die Erstellung einer grundsätzlichen Entscheidungsvorlage oder eines Rechenschaftsberichtes vor, die für den abschließenden Entscheidungszeitpunkt nicht mehr gesondert erhoben, sondern mit Blick auf zentrale Lücken, die aus Sicht der Entscheidenden bestehen, ergänzt werden müssen. Kennzeichnend für den Ansatz ist, dass die absehbaren Tagesordnungspunkte von Entscheidungssituationen ausschlaggebend sind dafür, welche Fragestellungen, welches Untersuchungsdesign und welche Erhebungsmethoden im Rahmen der jeweiligen Evaluation eingesetzt werden. Stufflebeam stellt in Vorausschau der im Programmverlauf zu erwartenden Entscheidungen das sog. CIPP-Modell vor, das verschiedenen Klassen von Entscheidungssituationen im Programmverlauf eine „Evaluationsart“ zuordnet, zu deren Beschreibung er die essentiellen Evaluationsfragestellungen und die zu deren Beantwortung angemessenen Methoden der Datenerhebung zusammenstellt. Das jeweilige Design muss jedoch von den Evaluierenden selbst

	entwickelt werden. Evaluatoren/-innen arbeiten bei einer Entscheidungsgesteuerten Evaluation mit den Adressaten/-innen ihrer Informationen zusammen, also den Beteiligten und Betroffenen (bzw. Repräsentanten/-innen der verschiedenen Gruppen), um herauszufinden, worin ihre Unsicherheiten und worin ihre abweichenden oder gegensätzlichen Einschätzungen der empirischen Realität bestehen. Hieraus leiten die Evaluatoren/-innen die Fragestellungen ab, die empirisch bearbeitet werden. Es sollten ausdrücklich nicht nur diejenigen Personen in die Klärung der Evaluationsfragestellungen einbezogen werden, die tatsächliche Entscheidungsträger sind, sondern es sollen Fragen aller Beteiligten aufgegriffen werden. Dabei soll aber auch darauf geachtet werden, Informationsinteressen von Entscheidenden mit in die Entwicklung des Evaluationsdesigns einzubeziehen, die relativ weit vom Programm entfernt sind, aber besonders längerfristig erhebliche Einwirkungsmöglichkeiten haben (z.B. Parlamentarier, Haushaltspolitiker, Stiftungsräte von Förderstiftungen). Die vergleichsweise breite Sicht auf die Informationsinteressen vieler Beteiligtengruppen soll nicht verhindern, dass die Evaluation auf herausragende Entscheidungssituationen zuarbeitet, die in der Regel durch Programmfinanziers, -träger und -verantwortliche beherrscht werden. Besonders ihnen muss die Evaluation hilfreiche Informationen zur Verfügung stellen.
Steuerungs-faktoren	Die Unsicherheiten, untereinander abweichenden und konträren Realitätssichten der Entscheidenden und weiterer Beteiligter, Tagesordnungspunkte vorauszusehender Entscheidungssituationen und (z.B. auch gesetzliche) Erfordernisse der Programmrechenschaftslegung.
Hauptzwecke	Evaluationsergebnisse sollen v. a. dazu genutzt werden, Entscheidungen über die Verbesserung oder Grundsatzentscheidungen zu Programmen auf gesicherter empirischer Datengrundlage zu treffen.
Quellen der Fragestellungen	Fragestellungen stammen von den Beteiligten, insbesondere von denjenigen, die Entscheidungen zu treffen haben.
Typischerweise eingesetzte Methoden	Methoden der Daten- und Informationssammlung sind nicht festgelegt, es können alle Methoden, wie z.B. schriftliche Befragungen oder Interviews, Beobachtungen sowie die verschiedensten Designs eingesetzt werden. Experimentelle Versuchspläne werden wegen ihrer Inflexibilität, auf sich verändernde Entscheidungssituationen zu reagieren, tendenziell ausgeschlossen.
Stärken	Durch den systematischen Einbezug der Informationsinteressen der Beteiligten – insbesondere der Entscheidenden **–** wird es wahrscheinlicher, dass Evaluation akzeptiert und ihre Ergebnisse tatsächlich genutzt werden, insbesondere dann, wenn zu den zentralen Fragestellungen der Beteiligten genau zum richtigen Zeitpunkt die erforderlichen Daten und Informationen vorliegen. Evaluation kann so kontinuierlich und systematisch für die Verbesserung des Programms in jeder seiner Phasen nutzbar gemacht werden. Für eine Rechenschaftslegung kann auf vorhandene Daten und Dokumente zurückgegriffen werden. Durch die Fokussierung auf notwendige Informationen und den Zeitpunkt, zu dem sie bereitstehen müssen, erbringt die Evaluation idealerweise nur nutzbare Leistungen, so dass sie selbst eine hohe Kostenwirksamkeit erreichen kann **–** ein in Zeiten knappster Finanzmittel wichtiges Merkmal dieses Modells. Das Modell signalisiert, dass – im Rahmen des Möglichen – auf veränderte Informationsinteressen der Entscheidenden flexibel durch Änderungen am Evaluationsdesign reagiert werden kann. Die Stakeholder und Entscheidenden werden in der Zusammenarbeit mit den Evaluatoren/-innen unterstützt, ihre Entscheidungssituationen samt der erforderlichen Informationsgrundlage systematisch und rechtzeitig vorzubereiten und erlangen so mehr Klarheit über das Programm und dessen Steuerung.
Schwächen	Die Standardkritik an diesem Evaluationsmodell lautet, dass Entscheidungssituationen selten klar vorauszusehen und planbar sind. Außerdem wird angenommen, dass Entscheidungen weniger rational auf Grundlage von fundierten Informationen und vielmehr beeinflusst durch Routinen, Machtgefüge

	oder andere außerhalb der Sache liegenden Einflüsse getroffen werden. Abgesehen davon, dass diese Einwände gegen jedwedes Evaluationsmodell erhoben werden können, das praktische Nutzung seiner Ergebnisse anstrebt, kann erwidert werden, dass durch das Entscheidungsgesteuerte Modell ein – wenn auch oft geringer – Beitrag zu einer Kultur rationalerer Entscheidungen geleistet wird. Eine weitere Begrenzung des Ansatzes liegt in dem Erfordernis einer kontinuierlichen Ansprechbarkeit von Stakeholdern und ihrer Bereitschaft, konstruktiv und zielgerichtet an der Klärung von Fragestellungen für Entscheidungssituationen mitzuarbeiten. Zwar ist die zeitliche Belastung und auch die Bereitschaft zur Konfliktaustragung für die Beteiligten deutlich geringer als bei partizipativen Evaluationsansätzen, doch besteht ihrerseits ein großes Vertrauen in die Unparteilichkeit der Evaluatoren/-innen, welches eine unverzichtbare Voraussetzung darstellt. Ist dieses nicht vorhanden, sind die Interessenskonflikte – insbesondere die unausgesprochenen – zu groß oder gibt es andere Gründe für taktische Manöver bis hin zu gezielten Fehlinformationen und Zurückhaltung von Wissen über Entscheidungssituationen, droht die Evaluation ins Leere zu laufen. Wenn die Evaluatoren/-innen dies nicht bemerken oder darauf nicht angemessen reagieren, werden sie selbst zum Spielball von Machtauseinandersetzungen und drohen ihre Glaubwürdigkeit vollends zu verlieren. Verfügen sie aus wirtschaftlichen Gründen über keine Option, einen Abbruch der Evaluation ins Spiel zu bringen, droht Nutzlosigkeit der Evaluation, womit sie ihren zentralen selbst gesetzten Zweck nicht erreicht. Wenn die verschiedenen Stakeholder-Gruppen nicht angemessen einbezogen werden (können), liegt eine Gefahr darin, dass hauptsächlich „oberste Entscheider/-innen“ Evaluationsfragestellungen bestimmen und/oder mit Informationen versorgt werden und damit die Evaluation parteilich wird. Mögliche Grenzen für neue/abgeänderte Fragestellungen sollten vorab bedacht und kommuniziert werden, damit die Flexibilität der Evaluation nicht überfordert wird, was im schlechten Fall zu immer wieder abgebrochenen Erhebungen und damit zu inakzeptablen zahlreichen Blindleistungen führen würde.
Werteberücksichtigung	Wertepriorisierend; dabei „gemäßigt“ demokratisch, da in der Tendenz die Werte derjenigen mehr Gewicht erhalten, die an zentralen Entscheidungen beteiligt sind. Die Anforderungen an faires und ausgewogenes Handeln der Evaluatoren/-innen sind hier besonders hoch, da einflussschwächere Beteiligte lediglich mittelbar die Agenda der Evaluation mit bestimmen.
Wichtige Quellen	Cronbach (1963), Stufflebeam (1966, 1967), Alkin (1969), Webster (1975, 1995).

8.10 Nutzungsgesteuerte Evaluation

Synonyme	---
Charakterisierung Modell	Dieses Modell legt Planung, Durchführung und Ergebnisvermittlung von Evaluationen so an, dass ihre Nutzung durch die vorgesehenen Nutzer/-innen optimiert wird. Unter Nutzung wird sowohl die Verwendung der Ergebnisse (die insbesondere in Form von Berichten und Präsentationen vermittelt werden), als auch die des Evaluationsprozesses selbst verstanden. Der Merksatz dieses Evaluationsmodells lautet: Eine Evaluation, deren Ergebnisse durch die Praxis nicht genutzt werden, ist Verschwendung! Es gilt das Primat der Nützlichkeit für die Praxis vor der Nützlichkeit für die Wissenschaft. Die Nutzbarkeit einer Evaluation hängt wesentlich davon ab, ob das Evaluationsteam die primär vorgesehenen Nutzer (*primary intended users*) dafür gewinnen kann, die Bedingungen und Voraussetzungen der Nutzung im Vorhinein zu klären. Nutzung ist eine Handlung, die durch konkrete Personen ausgeführt wird. Diese müssen bekannt sein, und auch die von ihnen beabsichtigten Verwendungen der Evaluation und ihrer Ergebnisse. Es ist wichtig, dass die Gültigkeit der Ergebnisse einer Evaluation auf methodisch angemessenem Niveau gesichert ist. Es ist allerdings noch wichtiger, dass die vorgesehenen Nutzer, die ja in Bezug auf die empirische Sozialforschung meist Laien sind, den Ergebnissen ebenfalls Gültigkeit zumessen. Wenn die Beteiligten und Betroffenen die Ergebnisse der Evaluation nicht nachvollziehen und als gültig akzeptieren können, dann werden sie sich kaum die Mühe machen, diese Ergebnisse zu interpretieren, geschweige denn, darauf eine Entscheidung zu gründen. Face Validity bezeichnet die unmittelbare Einsichtigkeit von Daten und Darstellungen und ist eine zentrale Grundlage für die Akzeptanz der Evaluationsergebnisse. Es versteht sich, dass solche augenscheinliche Gültigkeit mit einigen herkömmlichen Standards des Untersuchens oder des Testens in Konflikt steht. Die Verständlichkeit zunächst der Daten, dann der Interpretationen, der Umsetzung zu Schlussfolgerungen und schließlich der Empfehlungen ist ebenso zentral wie Praktikabilität. Diese ist dann gegeben, wenn die Empfehlungen im zeitlichen, finanziellen, kulturellen und politischen Rahmen des Programms sowie der tragenden Organisationen und Personen umgesetzt werden können.
Steuerungsfaktoren	Steuerungsfaktoren sind die intendierten Nutzungen der intendierten Hauptnutzer/-innen. D.h. es wird in der Anfangsphase einer Evaluation möglichst präzise festgelegt, wer die Evaluation, ihre Ergebnisse und ggf. auch ihren Prozess wann für welche Entscheidungen oder Planungen nutzen will.
Hauptzwecke	Entscheidungsfindung, Programmverbesserung oder Erweiterung der Wissensgrundlage in einem bestimmten Feld, je nach Interessen des Auftraggeber/-innen/ oder Hauptnutzer/innen.
Quellen der Fragestellungen	Diese gehen hervor aus den im Dialog mit den vorgesehenen Nutzern/-innen identifizierten Informationsinteressen. Hierzu werden verschiedene Verfahren moderierter Sitzungen, schriftlicher Gruppen-Planungsverfahren u.ä. eingesetzt. In der Regel werden zwischen den Fragestellungen Prioritäten gesetzt als Grundlage, um die Evaluationsressourcen entsprechend aufzuteilen. Fragestellungen können im Ablauf der Evaluation an Relevanz gewinnen oder verlieren oder es können auch neue Fragestellungen hinzukommen.
Typischerweise eingesetzte Methoden	Als Evaluationsmethoden kommen im nutzungsfokussierten Ansatz grundsätzlich alle Verfahren in Betracht, dabei besonders diejenigen, die für die vorgesehenen Nutzer/-innen nachvollziehbar und plausibel darstellbar sind. Diese beiden Eigenschaften werden im Zweifel über technische Eleganz

	und Komplexität gesetzt. Fortgeschrittene quantitative Erhebungsdesigns sowie statistische Auswertungsmethoden kommen dann zum Einsatz, wenn sie relevante Informationsbedarfe der intendierten Nutzer befriedigen und wenn sie so dargestellt und ihre Ergebnisse so aufbereitet werden können, dass sie die Rezeptionsfähigkeiten und -motivationen der intendierten Nutzer/-innen treffen. Die nutzungsfokussierte Evaluation ist insofern flexibel und pragmatisch in Bezug auf die Methodenentscheidung. Dabei ist die Breite qualitativer Methodenoptionen besonders ausgeprägt
Stärken	Das Modell hebt darauf ab, die Einwirkung der Evaluation zu maximieren. Dabei wird unter Einwirkung die aktive Nutzung von Prozessen und Ergebnissen der Evaluation durch relevante Beteiligte verstanden. Da gemäß diesem Modell Evaluationen durchgängig entlang der präzisierten und gegengeprüften Nutzungserwartung der Adressatengruppen ausgerichtet werden, ist die Wahrscheinlichkeit der tatsächlichen Nutzung hoch. Wegen der engen Abstimmung mit den Beteiligten, die an strategischen Punkten der Evaluation wie Festlegung der Fragestellungen oder Interpretation der Ergebnisse stattfindet, können in Organisationen/Politikfeldern Widerstände gegen Evaluation gemindert werden. Evaluation unterstützt ein zielgerichtetes Wissensmanagement und Lernen in Organisationen.
Schwächen	Das Modell sieht vor, dass substantielle Zeit- und materielle Ressourcen in die nutzungsorientierte *Steuerung* von Evaluationen investiert werden. Für umfangreiche Datenerhebungen auf der Basis aufwändig entwickelter und getesteter Instrumente werden damit die Ressourcen eingeschränkt. Eine hohe Wirkung auf die Verbesserung von Programmen oder die Vorbereitung von Entscheidungen wird erkauft durch eine fokussierte, oft qualitativ/deskriptiv geprägte Datengrundlage. Die aktiv Beteiligten müssen bereit sein, erhebliche zeitliche und andere Ressourcen in die aktive Begleitung der Evaluation zu investieren. Der Evaluationsprozess ist anfällig gegenüber Versuchen von bestimmten Nutzergruppen, Druck auf die Anlage des Evaluationsprozesses oder die Darstellung von Ergebnissen auszuüben. Personen oder Gruppen, deren Interessen von Ablauf und Ausgang der Evaluation stark betroffen sind, können insbesondere versuchen, die Evaluation zu verzögern oder den Gegenstandsbereich so einzuschränken, dass anderen Gruppen wichtige Fragestellungen nicht untersucht werden. Nutzungsfokussierte Evaluatoren/-innen müssen über ein breites Kompetenzprofil verfügen; sie müssen starke Kommunikatoren/-innen sein und Abstimmungsprozesse beteiligtenorientiert und gleichzeitig effizient führen können. Außerdem müssen sie über ein breites Methodenrepertoire in quantitativen und qualitativen Verfahren verfügen. Solche Kombinationen sind selten oder erfordern die Zusammenstellung eines Evaluationsteams.
Wertberücksichtigung	In alle Phasen der Evaluation – beginnend bei der Festlegung des Evaluationszweckes bis hin zur Darstellungsform von Ergebnissen – sind wertebasierte Entscheidungen zu treffen. Ganz besonders gilt dies für die Interpretation von Daten, die je nach Wertposition, Menschenbild, theoretischen Annahmen der Interpretierenden zu gänzlich verschiedenen Erklärungen/Ursachenzuschreibungen und auch Schlussfolgerungen/Empfehlungen führen können. Die anzulegenden Wertgrundlagen werden von Evaluatoren/-innen mit den relevanten Nutzern/-innen geklärt und bei Differenzen wird nach möglichst großen Überschneidungsbereichen gesucht; die Wertposition der Auftraggeber/-innen findet – im Ausmaß variierend nach Partizipationsgrad der Evaluation **–** besondere Berücksichtigung. Angestrebt wird, dass Wertunterschiede transparent werden und eine wertebasierte Diskussion über zu ziehende Schlussfolgerungen oder Empfehlungen zwischen den Beteiligten und Betroffenen stattfindet.
Wichtige Quellen	Patton (1997); Beywl/Joas (2000).

8.11 Stakeholder-Interessen-gesteuerte Evaluation

Synonyme	Deliberative Democratic Evaluation.
Charakterisierung Modell	Durch die Berücksichtigung demokratischer Prinzipien im gesamten Evaluationsprozess soll die Evaluation gerade dort zu verteidigbaren, ausgewogenen (entgegen voreingenommenen, parteiischen) Ergebnissen und Schlussfolgerungen kommen, wo es gegensätzliche Wert- und Interessen-Positionen gibt. Zentral geschieht dies durch den gleichberechtigten Einbezug der Betroffenen und Beteiligten und aller Adressatengruppen in alle Phasen des Evaluationsprozesses. House und Howe sehen *deliberative democratic evaluation* nicht als isoliertes Evaluationsmodell an. Wenn andere Evaluationen folgenden drei Anforderungen genügen, können auch diese als *deliberative democratic* bezeichnet werden: Aufnahme/Einbezug, Dialog/Auseinandersetzung, Überlegung/Abwägung. Der *Einbezug* der relevanten Interessen aller verschiedenen Stakeholder-Gruppen schließt Voreingenommenheit/Beeinflussung von Prozess und Ergebnis der Evaluation aus. Um möglichst viele Gruppen einbeziehen zu können, müssen ggf. Vertreter für solche Gruppen bestimmt werden, die sich nicht selbst vertreten können. *Der Dialog/die Auseinandersetzung* mit und unter den Stakeholdern ist notwendig, um diejenigen Interessen zu identifizieren, die sie tatsächlich zu bestimmten Äußerungen oder Handlungen bewegen. Oft sind nicht alle dieser vorantreibenden Interessen den Handelnden bewusst, noch häufiger machen sie diese nicht transparent, z.B. um ihre Durchsetzungschancen zur erhöhen. Die Aufdeckung und Kommunikation dieser Interessen ist ein aufwändiger Teil der Evaluation. Durch *Überlegung/Abwägung* kommt die Evaluation zu ihrem eigentlichen Ergebnis, indem Daten, Informationen und die Interessen (auch Werte) miteinander in Verbindung gebracht werden. Der Prozess der Abwägung ist durch die Evaluatoren/-innen systematisch und vorausschauend zu strukturieren. Um zu Schlussfolgerungen zu gelangen, werden zunächst vorläufige Befunde und Stellungnahmen gemeinsam zusammen getragen, die alle Perspektiven repräsentieren müssen. Diese vorläufigen Befunde werden diskutiert. Die Formulierung von Schlussfolgerungen ist letztlich Aufgabe des Evaluators/der Evaluatorin, die diese jedoch aus dem vorherigen Dialog heraus bildet. Die Evaluatoren/-innen sollen nicht wertend sein, aber Prozesse anstoßen und begleiten, die in einem ausgewogenen diskursiven Prozess zu einem gewerteten Ergebnis kommen, das wiederum Entscheidungen vorbereiten kann. Das heißt, der professionelle Kanon zur Erreichung von Validität in der Datenerhebung bleibt unberührt. Die Evaluatoren/-innen müssen alle Stakeholder-Gruppen identifizieren und sicher gehen, dass zumindest ein Repräsentant jeder Gruppe in die Evaluation aktiv einbezogen wird. Sollte es dabei große Unterschiede in der Gewichtigkeit der Interessen und der Macht der Beteiligten geben, muss sie sicherstellen, dass es Instrumente gibt, dieses Ungleichgewicht zu moderieren (z.B. durch eine Ombudslösung). Auch sollten die Evaluatoren/-innen ein Auge drauf haben, dass das Engagement der Stakeholder-Vertreter möglichst authentisch und angemessen intensiv ist.
Steuerungsfaktoren	Die transparent und einem demokratischen Abwägungsprozess zugänglich gemachten Interessen aller Beteiligten und Betroffenen.
Hauptzwecke	Demokratische Partizipation soll genutzt werden, um schließlich zu sicheren und ausgewogenen Schlussfolgerungen und zu einer Einschätzung des Programms zu kommen und ist auch ein Wert an sich.

Quellen der Fragestellungen	Die Fragestellungen werden im gleichberechtigten Dialog möglichst aller erarbeitet, zumindest aber in Absprache mit wichtigen Beteiligten- und Betroffenengruppen. Kernfragestellungen beziehen sich dabei immer auf die Güte und Verwendbarkeit des Programms in Hinblick auf die Interessen der Betroffenen und Beteiligten.
Typischerweise eingesetzte Methoden	Es werden über den ganzen Prozess der Evaluation hinweg solche Methoden eingesetzt, die Inklusion, Dialog und Abwägung erlauben und fördern, z.B.: Diskussionen, Umfragen, Debatten. Aus der deutschen Tradition kommen hierfür insbesondere die Moderationsmethode oder die Planungszelle in Frage, aus der angelsächsischen z.B. die *nominal group technique* und oder *open-space*-Arrangements.
Stärken	Der Ansatz sieht Evaluation als eine Möglichkeit zur Demokratisierung von lokaler und nationaler Gesellschaft beizutragen und stellt einen Versuch dar, Evaluationen gerecht zu machen, indem eine demokratische Steuerung des Evaluationsprozesses eingeführt wird (Partizipation). Dabei spielen sowohl Elemente der direkten wie der repräsentativen Demokratie eine Rolle (letzteres bei der Vertretung, der Artikulations- und Durchsetzungsschwachen). Verzerrungen durch z.B. Vorgaben der Auftraggeber/-innen oder Einflussnahme der hauptamtlichen Programmmitarbeitenden kann damit entgegengewirkt werden. Durch die Partizipation wird die Akzeptanz der Evaluation und eine Umsetzung und Nutzung der Evaluationsergebnisse gefördert. Das *committment* der Beteiligten mit dem Evaluationsprozess wird gefördert Eine Besonderheit des Ansatzes ist es, dass den Evaluatoren/-innen das Recht vorbehalten ist, Stellungnahmen und Interessen, die sie als undemokratisch oder unethisch empfinden, auszugrenzen, so dass der Prozess der Evaluation im Rahmen eines geteilten demokratischen Grundverständnisses stattfindet.
Schwächen	Es kann den Evaluatoren/-innen ggf. schwer fallen, die tatsächlichen Informationsinteressen der Stakeholder zu identifizieren und auf das zu evaluierende Programm zu beziehen. Auch bedeutet es für sie eine sehr hohe Anforderung, durch den intensiven Dialog hindurch ihre Unvoreingenommenheit zu bewahren und nicht für eine Gruppe Partei zu ergreifen (in den deutschen Ausbildungen zur Mediation werden hier methodisches Wissen und eine professionelle Ethik vermittelt). Je komplexer die Situation des Programms ist, desto schwieriger und komplexer wird auch der Dialog. Er kann so sehr viel Zeit und Raum in der Evaluation einnehmen, zu Verzögerungen und gar zur Nicht-Bearbeitung vorgesehener Datenerhebungen führen. Es muss darüber hinaus mindestens eine repräsentative Gruppe von über die Dauer der Evaluation engagierten Beteiligten und Betroffenen gefunden werden. Durchführbar ist eine demokratisch ausbalancierte Evaluation nur, wenn die Auftraggeber/-innen bereit sind, einen großen Teil ihrer Macht zu teilen und vorläufige Ergebnisse zu relativ frühen Zeitpunkten an eine vergleichsweise breite Öffentlichkeit weiterzugeben, wobei sie die Verwendung der Informationen nicht kontrollieren können. Mit einer *deliberative democratic evaluation* entsteht aus einem institutionell gebundenen Programm ein offenes System. Der Ansatz erscheint als eine idealtypische Vorstellung von Evaluation, wie sie zumindest in vollem Umfang in der Realität – besonders bei bundesweiten Programmen **–** kaum durchführbar sein wird. Auch die Übertragbarkeit auf andere Modelle muss bezweifelt werden. Die Autoren erkennen diese Schwierigkeit an und plädieren für eine Durchführung „so gut wie möglich“.

Werteberück-sichtigung	Wertepriorisierend. Interessen und Werte werden als zentral, dabei als veränderbar oder einander annäherbar angesehen. Dies soll im Rahmen eines rationalen, demokratisch kontrollierten Prozess der Überlegung/Abwägung geschehen, so dass Werte und Interessen in allen Phasen der Evaluation ausschlaggebend sind. Es wird eindeutig gefordert, dass nicht unbesehen eine der Wertepositionen – sei es auch die der Benachteiligten – übernommen wird.
Wichtige Quellen	House & Howe (1998, 2000), Whitmore (1998).

8.12 Selbstorganisationsgesteuerte Evaluation

Synonyme	Empowerment Evaluation, Inclusive evaluation.
Charakterisierung Modell	*Empowerment Evaluation* bezeichnet die Anwendung von Evaluations-Konzepten, -techniken und -ergebnissen mit dem Zweck, Programmverbesserung und Selbstorganisation in einem partizipativen Prozess zu fördern. Die Programmverantwortlichen, -mitarbeitenden und -nutzer/-innen einschließlich der Auftraggeber/-innen führen gemeinsam ihre eigene Evaluation durch, wobei die externen Evaluatoren/-innen wesentlich als Evaluations-Berater/-innen oder Evaluations-Trainer/-innen fungieren oder als Dienstleister/-innen bei der praktischen Durchführung der Evaluation, je nach den programmintern vorhandenen Potentialen und Ressourcen. Selbstorganisationsgesteuerte Evaluation wird häufig in drei Schritten durchgeführt: 1. Programm-Betroffene und -beteiligte formulieren die Mission oder Vision, die sie mit dem Programm verfolgen, alternativ kann der Fokus auch auf angezielten Resultaten liegen. 2. Die wichtigsten Programmaktivitäten werden identifiziert und ihre Durchführung von den Teilnehmenden systematisch beschrieben und bewertet, die Bewertung wird diskutiert. 3. Ziele und Strategien für eine zukünftige Verbesserung des Programms werden auf dieser Grundlage entwickelt, eine Dokumentationsweise für das weitere Vorgehen wird vereinbart. Nach Abschluss dieser Phasen soll sich der Prozess idealerweise zirkulär verstetigen, da aktuell gewonnene Ergebnisse hieraus immer nur für begrenzte Zeit gültig sind. Möglichst viele Betroffene und Beteiligte sollen in den Prozess der Evaluation einbezogen werden und sich dafür verantwortlich fühlen. Die Interpretationen erfolgen aus den Perspektiven der Beteiligten und Betroffenen, die sich auf der Nachvollziehbarkeit ihrer Erhebungen und Schlussfolgerungen verpflichten. Die Unparteilichkeit der Schlussergebnisse soll dadurch sichergestellt werden, dass in deren Diskussion alle Beteiligten und Betroffenen einbezogen werden. Dabei wird festgehalten, welche Ergebnisse auf Konsens treffen und welche nicht. Minderheitenmeinungen können ebenfalls formuliert und schriftlich festgehalten werden, je nach Vereinbarung, die zwischen den Beteiligten zu Beginn der Evaluation getroffen wurde. Das Konzept der *Empowerment Evaluation* steht dem der Selbstevaluation in Deutschland nahe, die als eine Methode der Qualitätsentwicklung im Bereich der sozialen Arbeit entwickelt wurde. Auch dort werden Untersuchungsprozesse parallel zu Praxisverbesserungsprozessen betrieben, allerdings nehmen an Selbstevaluation gemäß der Konzepte von Heiner (1988ff) und von v. Spiegel (1993ff) in der Regel ausschließlich die Fachkräfte (*program staff*) teil.
Steuerungsfaktoren	Die länger- und mittelfristigen Ziele, Bedürfnisse und Motive der Stakeholder.
Hauptzwecke	Der übergeordnete Zweck ist die Stärkung der Selbstbestimmung (*Empowerment*). Konkret sollen nicht nur Güte und Verwendbarkeit des Programms empirisch bestimmt werden, sondern es sollen auch die Programmbeteiligten und Betroffenen zu einer systematischen und andauernden Selbstevaluation befähigt werden, die sie kontinuierlich bei ihrer Arbeit unterstützt (Qualifizierungszweck ähnlich dem in der deutschen Selbstevaluation).
Quellen der Fragestellungen	Die Gemeinschaft von Praktiker- und Nutzergruppen des Programms im lokalen Setting entwickelt (ggf. zusammen mit oder moderiert durch externe Evaluatoren/-innen) die Fragestellungen, die im Verlauf der Evaluation in fortwährender Interaktion verändert und weiterentwickelt werden können.
Typischerweise eingesetzte Methoden	Es können verschiedenste einfach zu beherrschende qualitative oder quantitative Methoden eingesetzt werden.

Stärken	Die Beteiligten und Betroffenen werden zu einer selbstständigen Weiterführung der Evaluation befähigt. Idealerweise entsteht eine Atmosphäre gemeinsamen Lernens und gemeinsamer Verantwortung. Durch den starken Einbezug in den Prozess der Evaluation kann soziale Benachteiligung im Bereich der Handlungskompetenzen ausgeglichen werden. Die Interessen der benachteiligten Stakeholder können durch die Diskussion von Evaluationsergebnissen einen Eingang in die öffentliche Diskussion finden. Widerstände gegen die Evaluation und die Nutzung ihrer Ergebnisse werden durch den Einbezug/die Partizipation der Stakeholder-Gruppen abgebaut. Die Evaluation ist sehr kostengünstig durchzuführen. Oft reichen wenige Leistungstage einer externen Evaluationsberatung aus. Den Stundeneinsätzen der Fachkräfte stehen sowohl Qualifizierungseffekte als auch unmittelbare Qualitätsverbesserungen des Programms gegenüber.
Schwächen	Es muss eine hoch rezeptive Gruppe von Betroffenen und Beteiligten gefunden werden, die zum langfristigen Mitarbeiten in der (Selbst-) Evaluation bereit ist. Die Teilnehmenden müssen Risiken bei der Weiterentwicklung des Programms eingehen können und gleichzeitig die Verantwortung für die von ihnen eingeleiteten Untersuchungen und ggf. Verbesserungen übernehmen, was bei quer zu Hierarchien zusammengesetzten Gruppen zu erheblichen Spannungen führen kann. Eine starke Unterstützung der externen Evaluatoren/-innen zu Gunsten der Benachteiligten mag die Glaubwürdigkeit der Evaluation beinträchtigen, Wertkonflikte werden eher virulent, was ggf. einen zusätzlichen Mediationsprozess erfordert (ggf. Rollenkonflikt für die externen Evaluatoren/-innen). Besonders dann, wenn die Beteiligten aus einer einzigen Organisation (z.B. von einem Qualifizierungsträger für Benachteiligte) stammen, mag den Evaluationsergebnissen geringe Glaubwürdigkeit zugeschrieben werden. Dies kann dadurch aufgefangen werden, dass die Evaluationsberater/-innen darauf drängen, auch Externe (in diesem Fall z.B. Vertreter/-innen von Praktikumbetrieben oder potentielle Arbeitgeber/-innen) einzubeziehen.
Werteberücksichtigung	Wertepositioniert: *Empowerment Evaluation* hat eine eindeutige Werteorientierung: Sie wird eingesetzt, um Betroffenen dabei zu helfen, sich selbst zu helfen und ihre Programme zu verbessern. Es werden die Werte insbesondere der benachteiligten Gruppen aufgenommen, die „gestärkt“ werden sollen.)
Wichtige Quellen	Fetterman (1996; 2000).