

33. Kongress der Deutschen Gesellschaft für Soziologie  
Universität Kassel, 9.-13.Okt. 2006  
Sektion Methoden der empirischen Sozialforschung

## Wissenschaftstheoretische Anforderungen an empirische Forschung und die Problematik ihrer Beachtung in der Evaluation

Oder: Wie sich die Evaluationsforschung um das Evaluieren  
drückt

*Helmut Kromrey*

### Vorbemerkungen

Wenn von Evaluation – in welcher Bedeutung des Begriffs auch immer – gesprochen wird, ist damit eine bewertende Aussage über einen Sachverhalt gemeint. Kommt dabei als Erkenntnisinstrument (auch) empirische Forschung ins Spiel, lautet das Ziel von Evaluation: empirisch gestützte Gewinnung von Bewertungen mit intersubjektivem Geltungsanspruch. Zugespitzt auf das Thema dieser Veranstaltung („Die empirische Erfassung von Qualität“): Es geht um die forschungsgestützte Gewinnung von *Qualitätsaussagen*, d.h. um empirisch begründete normative Aussagen, um „Werturteile“. Für die „herrschende Meinung“ in Methodologie und Wissenschaftstheorie (der analytisch-nomologischen Richtung, auf deren Argumentation ich mich im Folgenden beziehe) manövriert sich die Forschung damit in ein offensichtliches Dilemma: Werturteile sind empirisch nicht begründbar.

Umso mehr überrascht es, dass dies – abgesehen von ganz seltenen Ausnahmen<sup>1</sup> – bei Evaluationen trotz ihrer quantitativ zunehmenden Bedeutung als Problem anscheinend nicht erkannt, zumindest jedenfalls nicht thematisiert wird. Die skizzierte Problematik wirft ein ganzes Spektrum von Fragen auf, die natürlich in einem einzigen Vortrag nicht abgehandelt werden können. Ich beschränke mich hier auf die Wertproblematik im Rahmen von *Evaluation durch Forschung* (ungenauer: „Evaluationsforschung“).

Im Rahmen analytisch-nomologischer Wissenschaftstheorie ist – wie bereits angedeutet – der Anspruch, „empirisch gestützte Bewertungen mit intersubjektivem Geltungsanspruch“ zu gewinnen, methodologisch nicht ohne Weiteres legitimierbar:

Die unmittelbare empirische Begründung von Bewertungen durch Forschung ist nicht möglich; auch aus korrekten *empirischen* Beschreibungen und Analysen sind *normative* Aussagen nicht ableitbar.

Die Geltungsbegründung empirischer Analysen folgt einer anderen Logik als die Geltungsbegründung normativer Aussagen. Für die ersteren gibt es in der Wissenschaftstheorie klare Regeln, für die letzteren nicht.

Möglich ist es lediglich, die Forschung möglichst genau auf den Zweck „Bewertung“ auszurichten, indem wenn schon nicht direkt bewertende, so doch bewertungsrelevante Informationen gesammelt und systematisiert werden.<sup>2</sup>

Soll also unmittelbar „durch Forschung“ evaluiert werden, müssen Strategien gefunden werden, mit deren Hilfe erreichbar wird, dass die *empirischen* Daten einen quasi *normativen* Charakter erhalten, so dass sie „für sich selbst sprechen“ können. Eine explizite Geltungsbegründung daraus abgeleiteter Wertaussagen durch die Forschung wäre dann nicht mehr notwendig. Solche Strategien gibt es in der Tat; und drei von ihnen sollen hier skizziert werden.

---

1 Beispielsweise Christian Lüders wie auch Wolfgang Beywl in Flick 2006.

2 Dies geschieht z.B. in der Hochschulevaluation und/oder bei Akkreditierungen nach dem bekannten mehrstufigen Modell des peer review. Die Forschung hat hier nur die Funktion des Informationszulieferers; das Evaluieren (d.h. das Fällen der Werturteile) geschieht durch Experten oder durch ein dazu legitimiertes Gremium oder durch Aushandeln zwischen den beteiligten Parteien.

## Exkurs: Die Argumentationslogik analytisch-nomologischer, „wertfreier“ Forschung

Für *wissenschaftstheoretische* Argumentationen kann das von Hempel und Oppenheim konzipierte Schema wissenschaftlicher Erklärung (Hempel/Oppenheim 1948) als verbindliches Gerüst gelten:

- |                     |                                                                                                              |
|---------------------|--------------------------------------------------------------------------------------------------------------|
| <i>Explanans:</i>   | (1) Es gilt (mindestens) ein nomologisches Gesetz<br>z.B.: „Wenn A und B, dann C“                            |
|                     | (2) Die in der Wenn-Komponente genannten Randbedingungen sind empirisch erfüllt (z.B.: „A und B liegen vor“) |
| <hr/>               |                                                                                                              |
| <i>Explanandum:</i> | (3) Singulärer Satz, der den zu erklärenden Sachverhalt beschreibt (z.B. „C liegt vor“).                     |

*Gegeben* ist das zu erklärende „singuläre Ereignis“ (3), *gesucht* ist das „Explanans“ (1 und 2). Bei dieser Art von Erklärung muss (3) deduktiv-logisch aus (1) und (2) folgen, wobei (2) aus der Wenn-Komponente und (3) aus der Dann-Komponente des nomologischen Gesetzes abgeleitet wird. In der *alltäglichen empirischen Forschung*, die sich auf die analytisch-nomologische Wissenschaftstheorie beruft, scheint das H-O-Schema allerdings wenig praktische Bedeutung zu haben. Dieser Eindruck täuscht jedoch.<sup>3</sup>

Zum einen gilt die *Logik der Erklärung* auch für die Konstruktion des Forschungsdesigns zum empirischen *Test von Theorien und Hypothesen*. Lediglich das Erkenntnisinteresse ist ein anderes: Es sind nicht singuläre Ereignisse (3) „zu erklären“, sondern es sind nomologische Hypothesen (1) auf ihre empirische Geltung „zu prüfen“. Hierfür dürfte die skizzierte Logik auch von Forschungspraktikern leicht einzusehen sein. In anderen Forschungszusammenhängen dagegen werden seit Poppers Fokussierung der Methodologie auf die Logik des Testens wissenschaftstheoretische Grundlagen eher stiefmütterlich behandelt.

Bei genauerem Hinsehen können die drei Komponenten des H-O-Schemas aber z.B. auch im deskriptiven Survey-Modell – also bei empirischen Forschungen zum Zwecke deskriptiver Diagnose sozialer Problemfelder – wiedergefunden werden. Als methodologisches Gerüst ist in diesem Design nach Formulierung einer präzisen Fragestellung ein deskriptives Modell des zu untersuchenden Gegenstands auszuarbeiten. Dieses wird idealtypischerweise auf der Basis „empirisch bewährter“ Theorien (1) entwickelt, im Realfall faktischer Forschung ergänzt um Hypothesen möglichst hoher Plausibilität. Dieses Modell hat eine forschungsleitende Funktion,

<sup>3</sup> Ausführlicher dazu Kromrey 2006, S. 87 ff.

dient sozusagen als „Wegweiser“ im Forschungsprozess, ist also Erkenntnis**basis** und nicht Gegenstand der Überprüfung. Da jedoch die erhobenen Daten der empirischen „Diagnose“ sozialer Probleme dienen sollen, müssen sie „Erklärungswert“ haben und sowohl die Problemdimensionen differenziert beschreiben (3) als auch die relevanten Randbedingungen (2) für das Auftreten der Probleme erfassen.

Zu bedenken ist, dass das H-O-Erklärungsschema ebenso wie die Übernahme seiner Struktur für andere Erkenntniszwecke als *erkenntnistheoretische* Basis den erkenntnistheoretischen Realismus impliziert: Unterstellt wird auf der *Gegenstandsseite* eine „real existierende“ Welt, gekennzeichnet durch Merkmale wie Ordnung, Struktur und Tatsachenautonomie, die Geltung von Regelmäßigkeiten bzw. Gesetzmäßigkeiten und Kausalität. Unterstellt wird zudem auf der *Seite des erkennenden Subjekts* die prinzipielle, wenn auch möglicherweise unvollständige und teilweise fehlerbehaftete Erkennbarkeit dieser Realität durch Wahrnehmungssinne sowie unterstützende Instrumente.

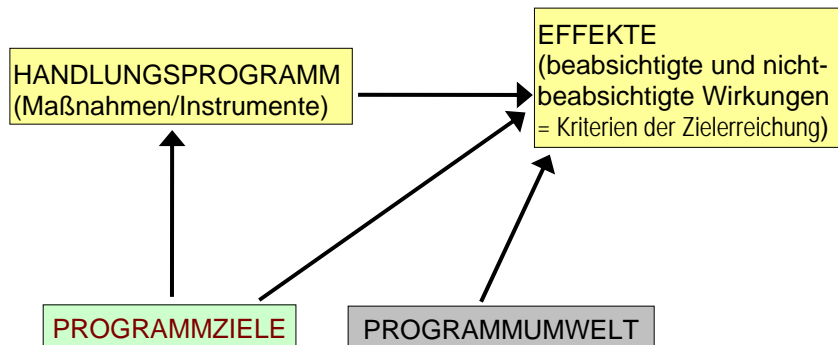
Das *Ziel* empirischer Wissenschaft ist hier die Erkenntnis der „wahren“ Strukturen und Gesetzmäßigkeiten der Realität sowie ihre Dokumentation in Theorien. Erreicht werden soll dies durch eine *Strategie* des „kontrollierten Ratens“ über das Aufstellen erkenntnisleitender ex-ante-Hypothesen und deren Konfrontation mit der (*objektiven*) Realität, abgebildet in (*subjektiven*) Wahrnehmungsdaten. Die darin implizierte Subjektivität wird kontrolliert durch strenge methodologische Regeln („Objektivierung“ der Verfahren, intersubjektive Nachprüfbarkeit).

Eine besondere Bedeutung kommt hier der Trennung deskriptiver (und damit „objektivierbarer“) von normativen Aussagen zu, deren subjektiver Charakter methodologisch nicht aufhebbar und deren intersubjektive Geltung daher mit empirischen Mitteln nicht begründbar ist. Die normativen Elemente der Erkenntnisgewinnung werden daher aufgeteilt in solche, die die *normative Basis* der Forschung bilden (der Forschung *vorgelagerte wissenschaftsimmanente* Werte), und die übrigen, d.h. die *nichtwissenschaftlichen* Interessen und Werte, die aus dem wissenschaftlichen Begründungskontext *ausgelagert* und in den Entstehungs- und Verwertungskontext des Forschungsprojekts verwiesen werden.

In genau dieser pragmatischen Strategie der „herkömmlichen empirischen Forschung“ zur Handhabung des Werturteilsproblems wird auch in der Evaluationsforschung die Lösung des Bewertungsdilemmas gesucht, wenn auch in unterschiedlichen Varianten.

## „Wertneutrale Evaluation“ im Design der Programmevaluation: der methodologische Idealtypus

So ist die skizzierte Forschungslogik einer auf Wertfreiheit verpflichteten Wissenschaft auf die Evaluationsforschung bruchlos anwendbar, sofern es sich bei ihrem Gegenstand um ein ausgearbeitetes „Programm“ handelt, das in Form explizit ausformulierter Programmziele die für eine „empirische Bewertung“ notwendige normative Basis bereits in die Evaluation mitbringt. Das Design der dieses Evaluationstyps berücksichtigt drei Dimensionen des zu bewertenden Gegenstands – Ziele, Maßnahmen, Effekte – sowie die programmexterne Umwelt als Quelle möglicher „Störvariablen“:



Man kann leicht erkennen, dass auch dieses Konzept die *Hempel-Oppenbeimsche* Logik verwendet.

Zum einen hat schon das „Programm“ als „technologische Aussage“ formal die gleiche Struktur wie eine „Erklärung“: Programmziele als angestrebte künftige Situation (Punkt 3), „Maßnahmen“ als Eingriffe in die gegenwärtigen „Randbedingungen“ (Punkt 2) sowie die theoretische Basis für die Art und Weise des Eingreifens (Punkt 1 im H-O-Schema).

Aber auch das Design der Evaluation orientiert sich an diesen Komponenten:

- Sowohl die existierenden Randbedingungen (2) als auch der Ist-Zustand der Zielvariablen (3) sind *vor* Programmbeginn –  $t_0$  – empirisch zu beschreiben.
- Während der Programmlaufzeit sind die Veränderungen der Randbedingungen (2) zu erfassen („monitoring“ sowohl der Maßnahmen als auch der Veränderungen in der Programmumwelt).
- Schließlich ist sicherzustellen, dass der Zustand der Zielvariablen (3) nach Programmdurchführung –  $t_1$  – wiederum empirisch beschrieben wird, so dass Art und Ausmaß der Veränderungen feststellbar sind.

Der Gegenstand, den es zu bewerten gilt, ist natürlich *als Gegenstand* nicht wertneutral oder zweckfrei. Ganz im Gegenteil: Das Programm soll etwas erreichen. Damit wird auch das Konzept „Evaluation als Programmwirkungsforschung“ mit dem Wertproblem konfrontiert. Für die Forschung ist es aber dadurch gelöst, dass es in den „Entstehungskontext“ verlagert wurde (Programmziele als normative Basis), wodurch die eigentliche „Evaluation“ einen deskriptiven (und somit „wertneutralen“) Charakter erhält: Der Bewertungsprozess reduziert sich damit auf einen Vergleich der vom Programm gesetzten Sollwerte (Zielerreichungskriterien) mit den gemessenen (und den Maßnahmen zurechenbaren) Effekten im Wirkungsfeld des Programms. Solche Aussagen lassen sich in vollem Umfang auf die Logik der Geltungs begründung empirischer Faktenbehauptungen stützen.

Dass die Realisierung dieser Aufgaben riesige Schwierigkeiten bereitet, tut der überzeugenden *Logik* des Modells keinen Abbruch.

Anders fällt das Urteil aus, wenn es um die praktische Bedeutung für den „alltäglichen“ Evaluationsbedarf geht. Die methodologischen Schwierigkeiten sind nämlich so riesig, dass diese überzeugende Logik nur unter sehr einschränkenden, ganz selten erfüllbaren Bedingungen praktisch einsetzbar ist. Überwiegend muss die Programmevaluationsforschung mit mehr oder weniger guten Annäherungen an das idealtypische Modell zufrieden geben. Es liegt also nahe, nach komplett andersartigen Ersatzlösungen zu suchen.

Eine solche Ersatzlösung folgt aus der Überlegung, „Qualität“ nicht erst anhand von „Effekten“, also als *Folge* der Eigenschaften des zu bewertenden Gegenstands oder Sachverhalts zu interpretieren, sondern *unmittelbar* zu messen.

## Evaluation als Qualitätsmessung: das methodologische Problemkind

Wenn es nämlich gelänge, am zu evaluierenden Gegenstand Qualitätsmerkmale zu bestimmen und präzise zu definieren, hätten wir einen direkten Weg, die Evaluation methodologisch zu „objektivieren“, d.h. am zu bewertenden „Objekt“ festzumachen. Wir könnten uns damit zugleich den schwierigen Umweg über die Messung von Outcome-Variablen und die methodisch äußerst schwierige Zurechnung ihrer Veränderungen als Zielerreichungs-Kriterien ersparen. Die implizite Annahme bei dieser Überlegung ist: Wenn die Qualität des zu bewertenden Sachverhalts hoch ist, dann werden auch seine Wirkungen positiv sein; bzw.: dann werden auch die mit ihm verknüpften Ziele erreicht werden. In diesem Fall hätte die Forschung lediglich die – forschungsmethodisch zur Alltagsroutine

zählende – Aufgabe zu erfüllen, „Qualität“ durch einen Satz qualitätsrelevanter Merkmale auszdifferenzieren und durch geeignete Indikatoren so zu operationalisieren, dass an ihnen situationsunabhängige „Qualitätsmesswerte“ abgelesen werden können.

Auch bei dieser Strategie wird die normative Basis der Evaluation in den Entstehungskontext ausgelagert, in dem der normative Begriff „Qualität“ von dazu legitimierter Seite festzulegen ist.

*Methodologisch* ist die o.g. Aufgabe allerdings gar nicht so simpel, wie es auf den ersten Blick erscheinen mag. Die „Sozialindikatorenbewegung“ in den 1970er Jahren hat sich damit intensiv auseinandergesetzt und eine Liste von Anforderungen an Indikatorensysteme formuliert,<sup>4</sup> die auch in unserem Fall Geltung beanspruchen kann.

*Grundlegender* ist jedoch eine andere Problematik, die in dieser Argumentation häufig übersehen wird: Neben dem Zutreffen der o.g. impliziten Annahme eines direkten Zusammenhangs zwischen Qualitätsmerkmalen und Zielerreichung (was empirisch geprüft werden kann) muss nämlich eine weitere, eine erkenntnistheoretische (und damit *axiomatische*) Voraussetzung als erfüllt angesehen werden: Qualität muss als direkte Eigenschaft des Objekts verstanden werden können (wie etwa Größe, Gewicht, Farbe usw.); bzw. methodologisch: Das Konstrukt „Qualität“ ist so zu definieren, dass seine Dimensionen als Merkmale des Gegenstands erscheinen. Schon eine oberflächliche semantische Analyse lässt erkennen, dass „Qualität“ eben nicht als direktes Merkmal(sbündel) des zu bewertenden Objekts zu verstehen ist, sondern als ein *relationales* Merkmal: als Eignung, Brauchbarkeit, Güte in Bezug auf bestimmte Ziele und Zwecke sowie auf bestimmte Nutzer- und Klientengruppen.

Damit (und wenn wir die methodologischen Anforderungen der Sozialindikatorenbewegung einbeziehen) haben wir aber wieder das Anforderungsniveau des Modells der Programmevaluation erreicht – ohne allerdings über deren Problemlösung zu verfügen, nämlich: vom Programm vordefinierte Ziele und Maßnahmen. Anders als zu Beginn postuliert, bleibt – wenn sich das Konstrukt „Qualität“ nicht aus dem Objekt selbst herleiten lässt – damit die Wert-Entscheidungs-Frage unbeantwortet. Die *normative Basis* für die Bewertung durch Qualitätsindikatoren setzt (und das heißt: die eigentliche Evaluation betreibt) diejenige Instanz, die festlegt, was als „Qualität“ gelten soll und welche Qualitätskriterien und –standards anzulegen sind. Und wenn diese Instanz Nebenziele verfolgt wie: a) das ganze Evaluationsverfahren solle möglichst ohne großen Ressourceneinsatz zu bewerkstelligen sein und/oder b) die Resultate dürften nicht noch interpretationsbedürftig, sondern sollten selbsterklärend sein (d.h.: höherer

---

4 s. z.B. Werner 1975.

Messwert = „mehr Qualität“), dann wird leicht in den Rang von Qualitätsindikatoren das erhoben, worüber Daten zur Verfügung stehen (d.h.: Qualität ist, was leicht messbar ist). Und dann werden vor allem quantitativ messbare Merkmale herangezogen (d.h.: Qualität wird operationalisiert durch Quantität).

Angesichts dieser erneuten Problematik verwundert es nicht, wenn die Forschung sich aus diesen Dilemmata zu befreien versucht, indem Sie Evaluation auf das reduziert, was sie unbestritten kann: Befragungen durchführen.

### Evaluation durch Befragung: das erkenntnistheoretische Problemkind

Statt aufwändige, methodisch kontrollierte Evaluation durch Programmforschung zu betreiben (deren Anwendungsvoraussetzungen selten erfüllbar sind) oder Qualitätsindikatoren zu messen (deren Gültigkeit fragwürdig ist), wird die Bewertung von Maßnahmen (Sachverhalten, Dienstleistungen) per „Betroffenenbefragung“ ermittelt. Die Adressaten und Nutzer, die Kunden und Klienten sind – so wird argumentiert – die von den zu evaluierenden Leistungen ganz konkret „Betroffenen“ und daher in der Lage, aus eigener Erfahrung auch deren Qualität sachverständig und zuverlässig zu beurteilen. Befragungen erscheinen erheblich weniger anspruchsvoll – sowohl hinsichtlich des Aufwands der Durchführung als auch hinsichtlich der Strategie der Objektivierung: Sind die erbrachten Dienstleistungen „schlecht“, so werden auch die Beurteilungen auf einer vorgegebenen Skala negativ ausfallen und umgekehrt. Befragt man eine hinreichend große Zahl von „Betroffenen“ und berechnet pro Skala statistische Kennziffern (etwa Mittelwerte oder Prozentanteile), dann kommen – so die weitere Argumentation – individuelle Abweichungen der einzelnen Urteilenden darin nicht mehr zur Geltung.

Sofern dies zuträfe,<sup>5</sup> wäre die Evaluation per Befragung der Königsweg zur Lösung aller Probleme der Evaluationsforschung – auch der Werturteilsproblematik, denn die Bewertungen nehmen hier die „per Betroffenheit dazu Legitimierten“ vor. Die Forschung selbst bliebe neutral; denn sie erhebt, systematisiert und analysiert lediglich.

Zwar sind „Messungen“ per Befragung nicht so problemlos wie dies dem Laien häufig erscheint (s. Kromrey 2006, S. 257 ff.). Doch sofern systematische

---

<sup>5</sup> Leider ist dieser Optimismus bei Lehrevaluationen nicht gerechtfertigt, wie differenzierte statistische Analysen von Daten aus Veranstaltungsbefragungen belegen (s. z.B. Kromrey 1994, 1995).



Verzerrungen vermieden werden können, lassen sich in der Tat bei hinreichend großer Befragtenzahl und bei repräsentativer Datenbasis individuelle Unterschiede – wie im obigen Zitat behauptet – „herausmitteln“. Einzulösen sind hierfür lediglich durch das Erhebungsinstrument und in der Befragungssituation einige formale methodologische Voraussetzungen – die allerdings nicht ohne weiteres als erfüllt gelten können:

- der „Gegenstand“ (das Objekt) der Beurteilung ist eindeutig definiert,
- das zu messende „Merkmal“ ist eindeutig definiert und operationalisiert,
- eine „Mess-Skala“ existiert und ist eindeutig definiert,
- die Befragten sind in der Lage, den „Gegenstand“ intersubjektiv übereinstimmend zu identifizieren, das zu messende „Merkmal“ intersubjektiv übereinstimmend zu erkennen und die „Mess-Skala“ in intersubjektiv übereinstimmender Weise anzuwenden.

Im Falle der Erhebung von Evaluationen wird die Situation zusätzlich dadurch schwieriger, dass es sich bei den zu messenden Merkmalen um die oben genannten „qualitätsrelevanten Merkmale“ (oder „Qualitätskriterien“) handelt, durch die der Begriff „Qualität“ operationalisiert wird. Und die hierauf anzuwendende Mess-Skala ist die Bezugsgröße, auf der das „Ausmaß“ von Qualität angebbar ist (also der „Qualitätsstandard“). Damit sind wir aber auch bei der *indirekten* Qualitätsmessung per Befragung wieder mit dem gleichen Problem konfrontiert wie beim Ansatz der *direkten* Messung von Objektqualität durch Indikatoren.

Das statistische „Ausmitteln“ von Messungenauigkeiten setzt bekanntlich die Existenz eines „wahren Wertes“ voraus, von dem die einzelnen Messwerte lediglich „zufällig“ abweichen. Bezogen auf die Beurteilungsvariation zwischen den einzelnen Befragten heißt dies: Um auf diese Weise zu einem gültigen Qualitätsmaß zu kommen, muss die Annahme gerechtfertigt sein, dass es einen „wahren“ Qualitätswert für den zu beurteilenden Sachverhalt gibt, um den die einzelnen Antworten „zufällig“ streuen. Diese Annahme wäre aber nur dann haltbar, wenn eines der beiden folgenden erkenntnistheoretischen Axiome zuträfe:

Alternative 1: Qualität ist ein „objektives“ Merkmal eines Sachverhalts,<sup>6</sup> dessen Ausprägung durch abbildende subjektive Wahrnehmung ohne systematische Verzerrung „gemessen“ werden kann, so dass bei hinreichend großer Zahl von Messungen der Erwartungswert dem „*wahren objektiven Wert*“ entspricht. Oder:

Alternative 2: Qualität ist ein „intersubjektiv gültiges“ Konzept, über das alle Menschen in gleicher Weise verfügen. Anders formuliert: Alle Menschen bewerten nach gleichen Kriterien und Standards in gleicher Weise; in konkreten Situationen

---

<sup>6</sup> Dies ist die identische Voraussetzung, die auch im Konzept der direkten Qualitätsmessung durch Indikatoren erfüllt sein muss.

auftretende Unterschiede zwischen Bewertern sind als Zufallsvariation anzusehen, so dass bei hinreichend großer Zahl von Messungen der Erwartungswert dem „wahren subjektiven Wert“ entspricht.

Grundlage für die erste Alternative ist die im erkenntnistheoretischen Realismus (vom frühen Empirismus bis zum Gründer des Wiener Kreises, Moritz Schlick) vertretene Überzeugung von der Möglichkeit abbildender Wahrnehmung der Realität: Das Wahrgenommene steht in einem genauen Entsprechungsverhältnis zum Wirklichen. In diesem Fall wäre „Qualität“ durch standardisierte Befragung „objektiv“ messbar.

Grundlage für die zweite Alternative wäre der erkenntnistheoretische Idealismus, am kompromisslosesten konzipiert in Platons „Ideenlehre“: Hinter der sinnlich wahrnehmbaren Welt stehen (als das „in Wahrheit Seiende“) die „Ideen“, die zwar der direkten Wahrnehmung nicht zugänglich, aber der unsterblichen Seele des Menschen von Anfang an mitgegeben sind. Erkenntnis besteht nach dieser Vorstellung im *Wiedererkennen* der allgemeingültigen Konzepte (der „Ideen“) in den empirischen (Einzel-)„Erscheinungen“. Sofern „Qualität“ der Status einer solchen „Idee“ zukäme (analog zu Gerechtigkeit, Gleichheit, Heldentum, Liebe etc.), wäre sie durch standardisierte Befragung „intersubjektiv“ messbar.

## Fazit

Die Versuche, das Wertproblem der Evaluationsforschung dadurch zu entschärfen, dass man die für das „wissenschaftliche Evaluieren“ erforderliche Wertbasis aus dem Begründungskontext empirischer Forschung hinausverlagert, erscheinen mir wenig überzeugend. Aus meiner Sicht bieten sich in dieser Situation zwei Alternativen für eine „wissenschaftliche Evaluation“ an.

Die eine bestünde darin, die Evaluation als einen „Sonderfall“ aus dem Aufgabengebiet einer wertneutral verfahrenen empirischen Forschung auszuheben und ihr die zusätzliche Aufgabe der (nach wissenschaftlicher Methodologie verfahrenen, intersubjektiv nachprüfbarer) Ableitung von Wertaussagen zuzuschreiben. Überlegungen in dieser Richtung werden von Christian Lüders (2006) angestellt. Eine überzeugende Methodologie ist allerdings derzeit nicht erkennbar.

Die andere Alternative wäre – und diese halte ich für die angemessenere Variante –, Evaluieren und Forsuchen klar zu trennen. Der Forschung ist die Aufgabe zuzuschreiben, alle für die Bewertung von Programmen, Maßnahmen etc. relevanten Informationen unter Einsatz des bewährten empirischen Instrumentariums zu erheben, zu analysieren und für Bewertungs- und Entscheidungsprozesse aufzubereiten. Die Funktion des Evaluierens sowie der Ableitung

möglicher Konsequenzen für das Evaluationsobjekt sollte dagegen einem dafür explizit legitimierten Gremium zugewiesen werden. Dass dieses Modell realisierbar ist und die Akzeptanz von Evaluation erhöht, zeigt das bereits angesprochene Modell der mehrstufigen Hochschulevaluation, beispielsweise in der vom Verbund Norddeutscher Universitäten praktizierten Variante: Selbstbeschreibung / Selbstevaluation – peer review – Auswertende Konferenz (Nordverbund 2004).

## Literatur:

- Beywl, Wolfgang (2006), »Evaluationsmodelle und qualitative Methoden«, in: Uwe Flick (Hg.), *Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen*, Reinbek bei Hamburg, S. 92-116
- Hempel, Carl G.; Oppenheim, Paul C. (1948): »Studies in the Logic of Explanation«, in: *Philosophy and Science*, Vol. 15, S. 135-175
- Kromrey, Helmut (1994), »Wie erkennt man "gute Lehre"? Was studentische Vorlesungsbefragungen (nicht) aussagen«, in: *Empirische Pädagogik*, 1994/2, S. 153-168
- ders. (1995), »Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern«. In: Peter Ph. Mohler. (Hg.): *Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung*, Münster, S. 105-128
- ders. (2006), *Empirische Sozialforschung. Modelle und Methoden der standardisierten Datenerhebung und Datenauswertung*, 11. Aufl., Stuttgart, utb 1040
- Lüders, Christian (2006), »Qualitative Evaluationsforschung – Was heißt hier Forschung?«, in: Uwe Flick (Hg.), *Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen*, Reinbek bei Hamburg, S. 33-62
- Verbund Norddeutscher Universitäten (Hg.) (2004), *10 Jahre Evaluation von Studium und Lehre*. Verbund-Materialien Band 16, Hamburg
- Werner, Rudolf (1975), *Soziale Indikatoren und politische Planung. Einführung in Anwendungen der Makrosoziologie*, Reinbek: rororo

**Helmut Kromrey**, Prof. Dr. rer.pol., Universitätsprofessor i.R. (Soziologie und Empirische Sozialforschung) an der Freien Universität Berlin und Adjunct Professor of Sociology der Graduate School of Management der Universität Educatis (Altdorf/Schweiz), Arbeitsschwerpunkte: Methoden/Methodologie der empirischen Sozialforschung, Evaluation. URL: [www.hkromrey.de](http://www.hkromrey.de)

Evelyn Funk, Dr. Wolfgang Meyer & Linh Nguyen

## Wissenschaftlichkeit und Partizipation

*Vortrag anlässlich der Frühjahrstagung 2013 des AK Methoden in der Gesellschaft für Evaluation (DeGEval), 26. April 2013, Eschborn*

### Wissenschaftlichkeit – 3 Argumentationsstränge

- A: Forschung darf nicht werten.

Lösung: „[...] Evaluieren und Forschen klar zu trennen“ (Kromrey 2007: 122)

- B: Evaluation muss werten.

„Valuation is the ‚what‘ of evaluation, and research design is the ‚how‘ of evaluation.“ (Crane 1988: 471)

- C: Wissenschaft wertet auch.

„Research under a paradigm must be a particularly effective way of inducing paradigm change.“ (Kuhn 2001: 52).

### Partizipation – Definitionen

“Participatory evaluation implies that, when doing an evaluation, researchers, facilitators, or professional evaluators collaborate in some way with individuals, groups, or communities who have a decided stake in the program, development project, or other entity being evaluated.“ (Cousins/Whitmore 1998: 5)

“(…) any evaluation in which there is a significant degree of collaboration or cooperation between evaluators and stakeholders in planning and/or conducting the evaluation.“ (Cousins, Donohue and Bloom 1996: 210)

“(…) any evaluation that involves program staff or participants actively in decision making and other activities related to the planning and implementation of evaluation studies“ (King 2005: 241)

## Literatur

COUSINS, J. BRADLEY AND ELIZABETH WHITMORE (1998): Framing Participatory Evaluation. In: *New Directions for Evaluation* 80, pp. 5-23.

COUSINS, J. BRADLEY, JOHN J. DONOHUE AND GORDON A. BLOOM (1996): Collaborative Evaluation in North America: Evaluators' Self-reported Opinions, Practices and Consequences. In: *American Journal of Evaluation* 17, pp. 207-226.

CRANE, JOHN A. (1988): Evaluation as Scientific Research. In: *Evaluation Review* 12 (5), pp. 467-482.

CULLEN, ANNE E., CHRIS L. S. CORYN AND JIM RUGH (2011): The Politics and Consequences of Including Stakeholders in International Development Evaluation. In: *American Journal of Evaluation* 32 (3), pp. 345-361.

DAIGNEAULT, PIERRE-MARC AND STEVE JACOB (2009): Toward Accurate Measurement of Participation: Rethinking the Conceptualization and Operationalization of Participatory Evaluation. In: *American Journal of Evaluation* 30, pp. 330-348.

KROMREY, HELMUT (2007) : Wissenschaftstheoretische Anforderungen an empirische Forschung und die Problematik ihrer Beachtung in der Evaluation. Oder: Wie sich die Evaluationsforschung um das Evaluieren drückt. In: *Zeitschrift für Evaluation* 1/2007, S. 113-123.

KUHN, THOMAS S. (2001): *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt: Suhrkamp Verlag.

## *Evaluationsstandards*

- ✓ DEGEVAL: Standards für Evaluation  
<http://www.degeval.de/degeval-standards>
- ✓ UNITED NATIONS EVALUATION GROUP: Standards for Evaluation in the UN System:  
[http://uneval.org/documentdownload?doc\\_id=22&file\\_id=561](http://uneval.org/documentdownload?doc_id=22&file_id=561)
- ✓ OECD DAC: Quality Standards for Development Evaluation:  
<http://www.oecd.org/development/evaluation/dcdndep/36596604.pdf>
- ✓ JCSEE: Program Evaluation Standards Statements:  
<http://www.jcsee.org/program-evaluation-standards/program-evaluation-standards-statements>
- ✓ INDEPENDENT EVALUATION GROUP: Sourcebook for Evaluating Global and Regional Partnership Programs, Indicative Principles and Standards:  
<http://siteresources.worldbank.org/EXTGLOREGPARPROG/Resources/sourcebook.pdf>
- ✓ SOCIÉTÉ FRANÇAISE DE L'ÉVALUATION: Charta zur Evaluation öffentlicher Politiken und öffentlicher Programme: <http://www.sfe-asso.fr/intranet/ckfinder/userfiles/files/charte-allemand.pdf>
- ✓ AUSTRALASIAN EVALUATION SOCIETY: Guidelines for the Ethical conduct of Evaluations:  
<http://www.aes.asn.au/images/stories/files/About/Documents%20-%20ongoing/AES%20Guidlines10.pdf>



# Wissenschaftlichkeit und Partizipation

*Vortrag zum Frühjahrstreffen des AK Methoden  
26. April 2013, Eschborn*

CEval -Referierende:

**Evelyn Funk**

**Dr. Wolfgang Meyer**

**Linh Nguyen**

Gliederung

Einleitung

Recherche

Wissenschaftlichk.

Partizipation

Schluss

- ✓ Einleitung
- ✓ Literaturrecherche
- ✓ Wissenschaftlichkeit in der Evaluation
- ✓ Partizipation in der Evaluation
- ✓ Schlussfolgerung

## Für was brauchen wir eigentlich Methoden?

- ✓ Methoden geben das „Wie“ vor
- ✓ Methoden sichern die Qualität
- ✓ Der richtige Methodeneinsatz ist professionell

These:

Professionelle Evaluation braucht Methoden

zur Sicherstellung ihrer Wissenschaftlichkeit  
und

zur Sicherstellung angemessener Beteiligung





## Übersicht zu Standardsystemen

- Gliederung
- Einleitung**
- Recherche
- Wissenschaftlichk.
- Partizipation
- Schluss

Standards	Wissenschaftlichkeit
<a href="#"><u>DAC-Standards</u></a>	<p><b>1.7 Quality control</b></p> <p><b>2.9 Selection of approach and methodology</b></p> <p><b>3.9 Validity and reliability of information sources</b></p> <p><b>3.13 Acknowledgement of changes and limitations of the evaluation</b></p>
<a href="#"><u>UNEG-Standards</u></a>	<p><b>3.7 Evaluation Methods should be sufficient rigorous</b></p>
<a href="#"><u>JCSEE-Standards</u></a>	<p><b>A2 Valid Information</b></p> <p><b>A3 Reliable Information</b></p>
<a href="#"><u>DeGEval Standards</u></a>	<p><b>G 5 Valide und reliable Informationen</b></p> <p><b>G 7 Analyse qualitativer und quantitativer Informationen</b></p>
<a href="#"><u>AES Guidelines</u></a>	<p><b>B13 Be sufficiently rigorous</b></p>
<a href="#"><u>Société française de l'Evaluation</u></a>	<p><b>Prinzip der Kompetenz</b></p>



## Übersicht zu Standardsystemen

Standards	Partizipation
<a href="#"><u>DAC-Standards</u></a>	1.4 Partnership approach 1.6 Capacity development 2.5 Stakeholder involvement 3.15 Incorporation of stakeholders' comments
<a href="#"><u>UNEG-Standards</u></a>	3.11 Stakeholders should be consulted
<a href="#"><u>JCSEE-Standards</u></a>	U2 Attention to Stakeholders U6 Meaningful Processes and Products
<a href="#"><u>DeGEval Standards</u></a>	N 1 Identifizierung der Beteiligten und Betroffenen N 8 Nutzung und Nutzen der Evaluation
<a href="#"><u>AES Guidelines</u></a>	-
<a href="#"><u>Société française de l'Evaluation</u></a>	Prinzip der Vielfalt

## Zentrale Fragestellung

Wie werden die methodischen Probleme der Sicherstellung von Wissenschaftlichkeit und Partizipation in der Evaluationsgemeinschaft thematisiert?

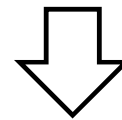
Welche Probleme werden angesprochen, welche Lösungen vorgeschlagen?

## Recherche nationaler und internationaler Fachzeitschriften:

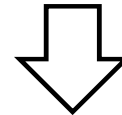
- ✓ American Journal of Evaluation
- ✓ Evaluation: The International Journal of Theory, Research and Practice
- ✓ Evaluation and Program Planning
- ✓ Evaluation Review: A Journal of Applied Social Research
- ✓ Journal of Programme Evaluation
- ✓ New Directions for Evaluation
- ✓ Practical Assessment, Research and Evaluation (online-Journal)
- ✓ Zeitschrift für Evaluation

## Selektionsprozess

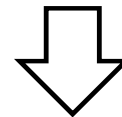
1. Stichwort-/Schneeballsuche in Zeitschriften nach „standards“, „scientific“, „participation“/ „participatory“ etc.



2. Identifikation 11 relevanter Artikel zu „Wissenschaftlichkeit in Evaluation“ und 20 relevanter Artikel zu „Partizipation in Evaluation“



3. Analyse der ausgewählten 31 Artikel



4. Ergebnis: 3 ausschlaggebende Artikel zu Wissenschaftlichkeit  
8 ausschlaggebende Artikel zu Partizipation

## Wissenschaftlichkeit in Evaluation

- ✓ Wenig Diskussion in der nationalen und internationalen Evaluationsgemeinschaft
  
- ✓ Essentielle Literatur
  - Kromrey 2007
  - Crane 1988

## 3 Argumentationsstränge:

- ✓ A: Forschung darf nicht werten.
- ✓ B: Evaluation muss werten.
- ✓ C: Wissenschaft wertet auch.

A: Forschung darf nicht werten.

„[...] Anspruch, ‚empirisch gestützte Bewertungen mit intersubjektivem Geltungsanspruch‘ zu gewinnen, [ist] methodologisch nicht ohne Weiteres legitimierbar.“ (Kromrey 2007: 114)



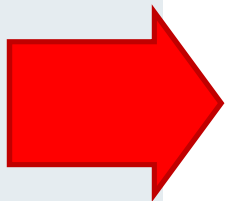
## A: Lösungsansatz

- ✓ Trennung von Evaluieren und Forschen
- ✓ Forschung hat beratende, begleitende Funktion.

B: Evaluation muss werten.

„Valuation is the ‚what‘ of evaluation, and research design is the ‚how‘ of evaluation.“

(Crane 1988: 471)



**Wertung als zentraler Bestandteil von  
Evaluation**

B: Evaluation muss werten.

„Selection of ‚valuation‘ as core construct requires the removal of some positivistic residuals in evaluators‘ thinking that lead them to shun valuations as subjective, personal, biased, and untestable.“

(Crane 1988: 480)

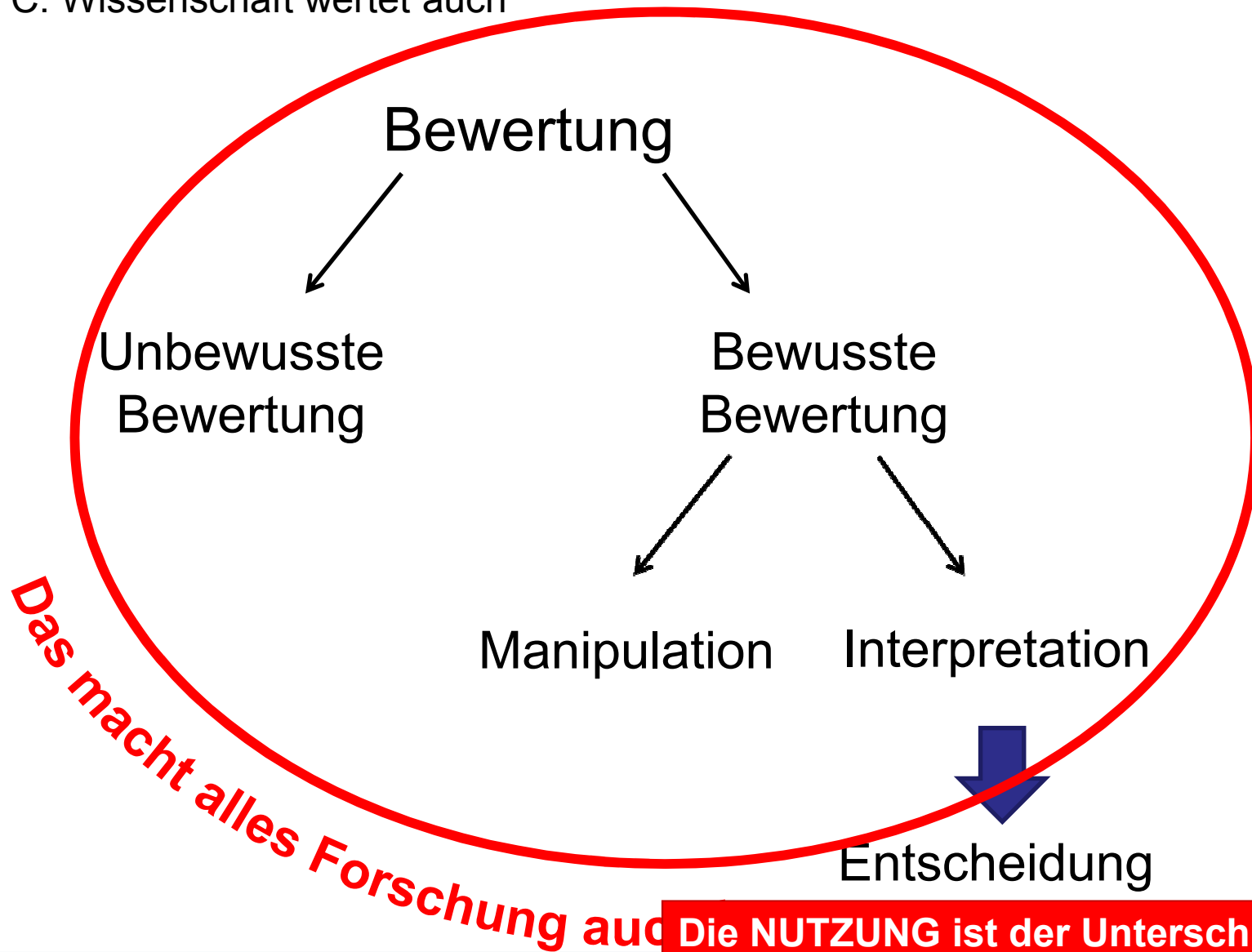
## B: Lösungsansatz

Kriterien zur Erfüllung eines wissenschaftlichen Konzeptes

- ✓ Sparsamkeit
- ✓ Allgemeingültigkeit
- ✓ Kohärenz
- ✓ Eindeutig zuzuordnen
- ✓ Logische Definition des Problemfeldes
- ✓ Möglichkeit kumulativer Forschung

C: Wissenschaft wertet auch

- Gliederung
- Einleitung
- Recherche
- Wissenschaftlichk.
- Partizipation
- Schluss



C: Wissenschaft wertet auch.

- ✓ Auswahl der Themen und Methoden von Paradigmen abhängig (Kuhn)
- ✓ Verlauf des Forschungsprozesses von Entscheidungen der Beteiligten abhängig
- ✓ Ergebnisse vom Verfahren abhängig
- ✓ Interpretation von willkürlich festgelegten Kriterien (z.B. Signifikanz) abhängig

## Fazit Wissenschaftlichkeit

- ✓ Es wird immer bewertet!
- ✓ Frage nach Kontrolle des Bewertungsprozess und methodische Sicherstellung der Wissenschaftlichkeit in Evaluation
- ✓ Bisher keine Diskussion in Evaluationsgemeinschaft!



- Gliederung
- Einleitung
- Recherche
- Wissenschaftlichk.
- Partizipation**
- Schluss

Stärken

Schwächen

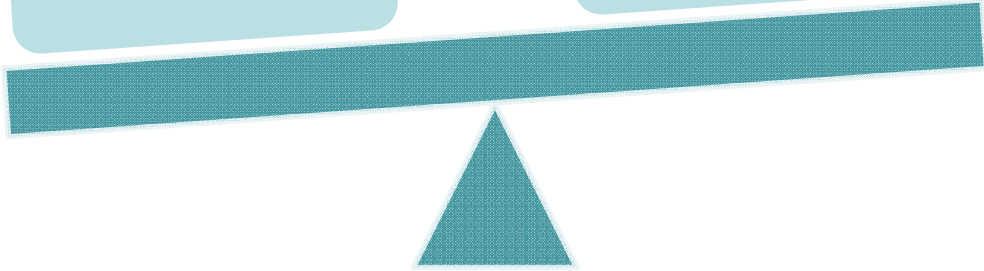
Pragmatisch

Politisch

Epistemologisch

Durchführbarkeit

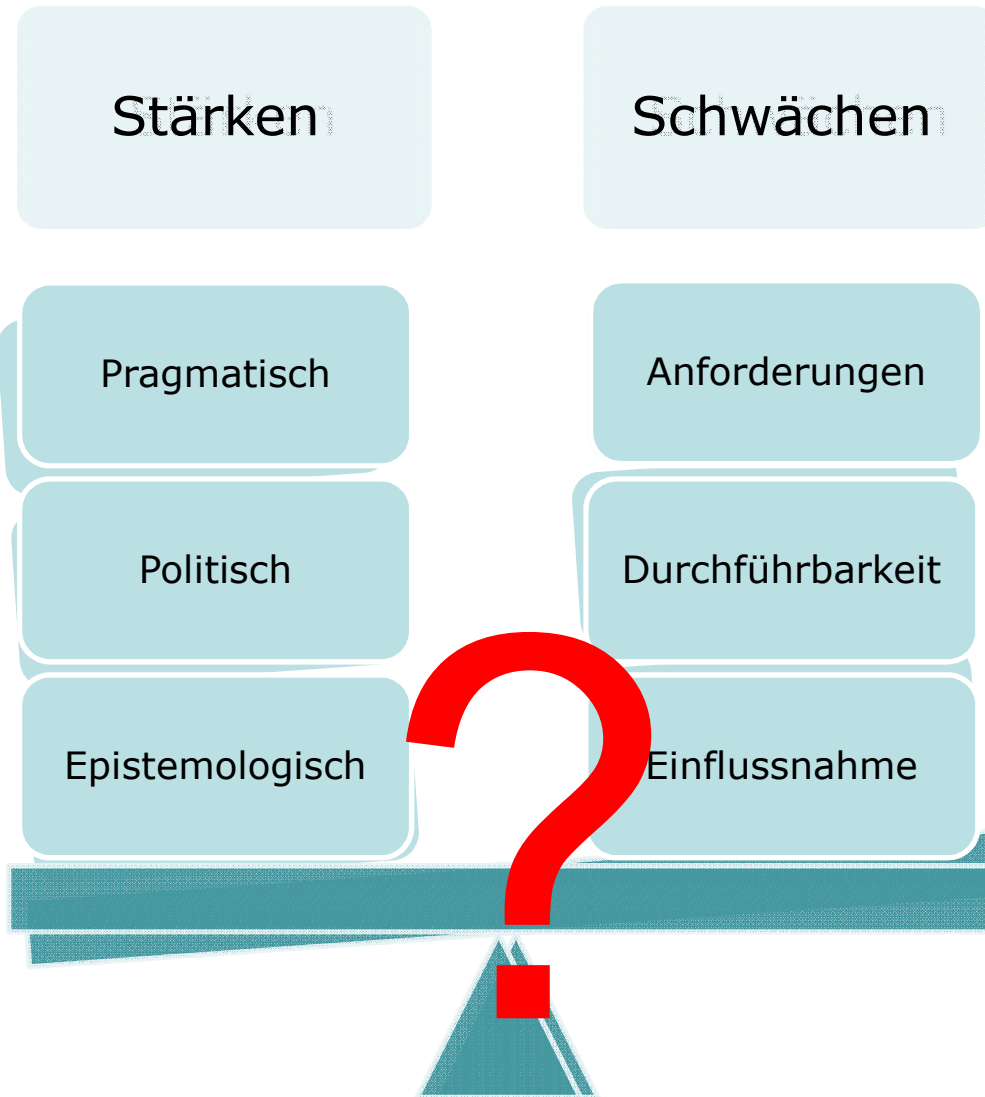
Einflussnahme







- Gliederung
- Einleitung
- Recherche
- Wissenschaftlichk.
- Partizipation**
- Schluss



## Definitionen

*Wer partizipiert woran in welcher Intensität?*

„Participatory evaluation implies that, when doing an evaluation, researchers, facilitators, or professional evaluators collaborate in some way with individuals, groups, or communities who have a decided stake in the program, development project, or other entity being evaluated.“ (Cousins/Whitmore 1998)

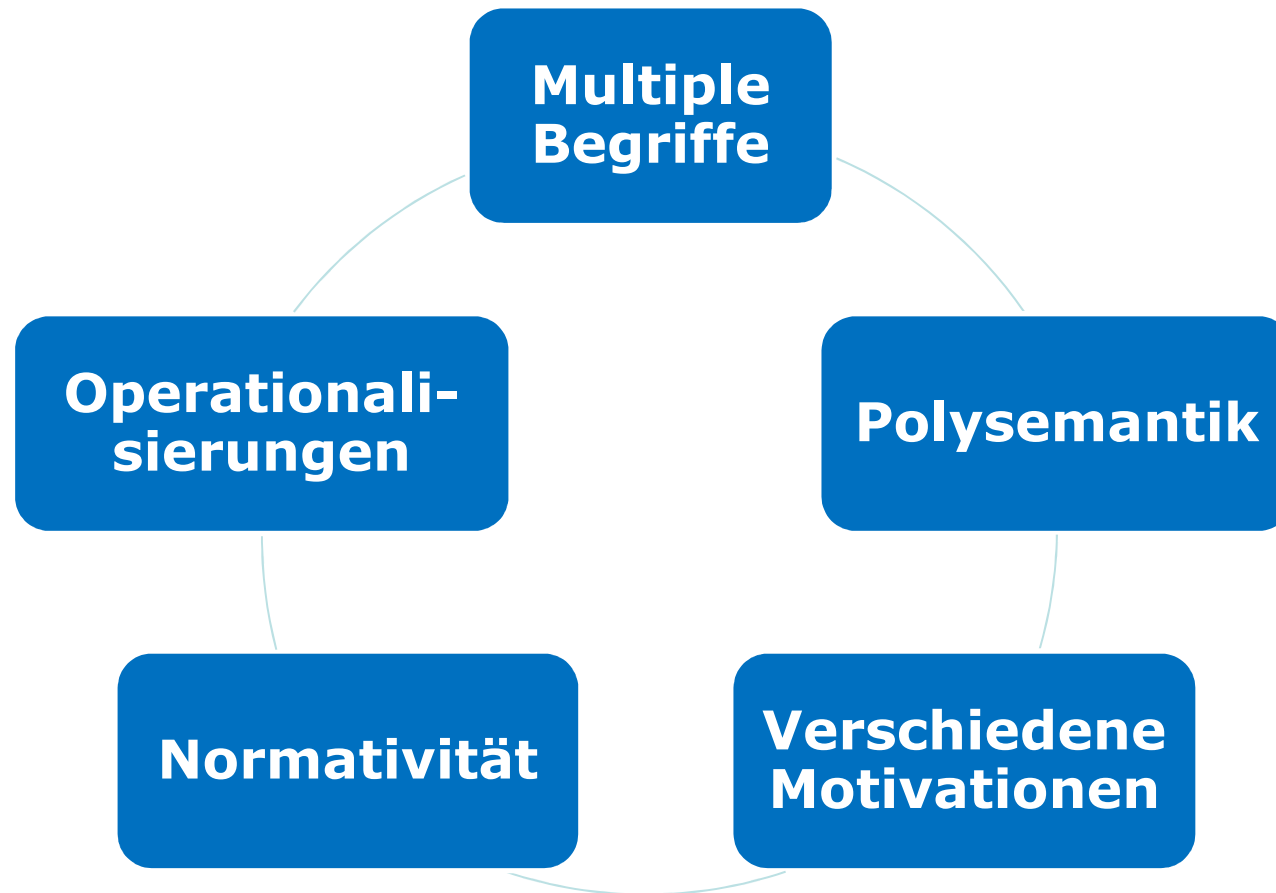
## Definitionen

*Wer partizipiert woran in welcher Intensität?*

„Participatory evaluation implies that, when **doing an evaluation**, researchers, facilitators, or professional evaluators collaborate **in some way** with **individuals, groups, or communities who have a decided stake** in the program, development project, or other entity being evaluated.“

(Cousins/Whitmore 1998)

## Diagnose: Konzeptuelle Unschärfe



Daigneault/Jacob 2009

## Participatory Evaluation Measurement Instrument (PEMI)

### 1. Ausmaß der Involviertheit

- Vier Entscheidungspunkte:
- 1) Evaluationsfragen
  - 2) Datensammlung und Auswertung
  - 3) Bewertung und Empfehlungen
  - 4) Reporting

Daigneault/Jacob 2009

## Participatory Evaluation Measurement Instrument (PEMI)

### 1. Ausmaß der Involviertheit

<b>Anzahl der Aufgaben</b>	<b>Involviertheit</b>	
0	Keine Involviertheit	.00
1	Eingeschränkte Involviertheit	.25
2	Moderate Involviertheit	.50
3	Starke Involviertheit	.75
4	Komplette Involviertheit	1.00

Daigneault/Jacob 2009

## Participatory Evaluation Measurement Instrument (PEMI)

1. Ausmaß der Involviertheit
2. Diversität der Partizipierenden

→ Vier Stakeholdertypen:

- 1) Entscheidungsträger
- 2) Durchführungsorganisationen und -personal
- 3) Zielgruppen
- 4) Zivilgesellschaft

Daigneault/Jacob 2009

## Participatory Evaluation Measurement Instrument (PEMI)

1. Ausmaß der Involviertheit
2. Diversität der Partizipierenden

<b>Anzahl der Stakeholder-Typen</b>	<b>Diversität</b>	
0	Keine Diversität	.00
1	Eingeschränkte Diversität	.25
2	Moderate Diversität	.50
3	Starke Diversität	.75
4	Komplette Diversität	1.00

Daigneault/Jacob 2009





## Participatory Evaluation Measurement Instrument (PEMI)

1. Ausmaß der Involviertheit
2. Diversität der Partizipierenden
3. Kontrolle des Evaluationsprozesses

---

### **Kontrolle des Prozesses**

---

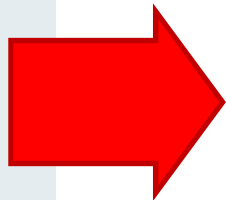
Exklusive Kontrolle der Evaluatoren/Auftraggeber	.00
Eingeschränkte Kontrolle der Teilnehmenden	.25
Geteilte Kontrolle zwischen Evaluatoren/Auftraggebern und Teilnehmenden	.50
Starke Kontrolle der Teilnehmenden	.75
Exklusive Kontrolle der Teilnehmenden	1.00

---

Daigneault/Jacob 2009

*Wer partizipiert woran in welcher Intensität?*

- ✓ (relativ) elaborierte Vorstellung eines Idealzustandes



**Aber: wenig Ideen darüber, wie dieser durch Anwendung verschiedener Methoden erreicht werden kann.**

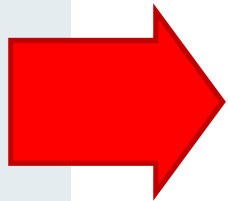
- ✓ Cousins, Donohue and Bloom 1996
- ✓ Cullen, Coryn and Rugh 2011

## Schlussfolgerungen

- ✓ Partizipation und Wissenschaftlichkeit werden in der Evaluation als **selbstverständlich** vorausgesetzt
- ✓ Die Notwendigkeit, beide Ansprüche **methodisch** abzusichern wird (noch) nicht gesehen
- ✓ Über **Wissenschaftlichkeit** wird in der Evaluationsgemeinschaft praktisch nicht diskutiert
- ✓ Über **Partizipation wird** hinsichtlich der Nutzung, nicht jedoch hinsichtlich deren Sicherstellung diskutiert

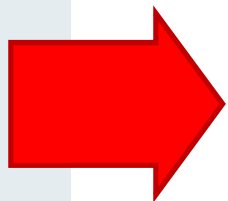
## Zwei zentrale methodische Aufgaben

- ✓ Sicherstellung **wertneutraler** Interpretationen und Schlussfolgerungen



**Ziel: Faire und rationale Entscheidungen**

- ✓ Sicherstellung **ausgewogener** Beteiligung und **ausreichender** Einbindung



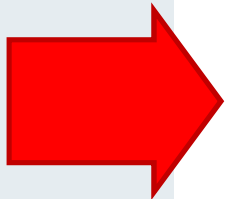
**Ziel: Akzeptanz und Ergebnisqualität**

## Zwei professionelle Ansprüche

- ✓ Festlegung von **Grenzwerten** („Mindeststandards“)

Wie viel **Wissenschaftlichkeit** ist notwendig um von einer professionellen Evaluation zu reden?

Wie viel **Beteiligung an Evaluation** ist notwendig um von einer partizipativen Evaluation zu reden?



**Ziel: Abgrenzung vom Alltagsvorgehen durch Methodik**

## Praktische Zielsetzungen

- ✓ Typologie der **Anforderungen an Wissenschaftlichkeit** für verschiedene Anwendungsfälle

(z.B. Monitoring vs. Evaluation)

- ✓ Typologie der **Anforderungen an Partizipation** für verschiedene Anwendungsfälle

(z.B. Pragmatic vs. Transformative Participatory Evaluation)

## Ausblick

- ✓ Aufarbeitung anderer Fachzeitschriften mit Evaluationsbezug
- ✓ Aufarbeitung der Literatur aus der Partizipationsforschung und der Wissenschaftslehre
- ✓ Aufarbeitung der Methodenliteratur
- ✓ Erstellung von Kriterien- und Methodenkatalogen
- ✓ Analyse der Evaluationspraxis
- ✓ Erstellung von Anforderungsprofilen



# Wissenschaftliche Ansprüche und Evaluierungspraxis – zwei Königskinder?

Ein Praxisbeispiel aus der deutschen  
Entwicklungszusammenarbeit

Frühjahrstagung des AK Methoden der DeGEval 2013

Sabine Dinges, Nadja Kabierski-Chakrabarti

Stabsstelle Monitoring und Evaluierung, GIZ





## Zum Titel...

Es waren zwei Königskinder,  
die hatten einander so lieb,  
sie konnten beisammen nicht kommen,  
das Wasser war viel zu tief.

→ Können **wissenschaftl. Ansprüche** und **Evaluierungspraxis**  
zueinander finden?



## Gliederung

1. **Ausgangsüberlegungen** zum Spannungsfeld zwischen wissenschaftl. Anforderungen und Praxis aus org. Perspektive
2. Institutionelle Verankerung von **Evaluierung in der GIZ**
3. Spannungsfelder bei dezentralen (**internen**) **Evaluierungen** und Lösungsansätze
4. Spannungsfelder bei zentralen (**unabhängigen**) **Evaluierungen** und Lösungsansätze
5. **Diskussionsfragen**



## Ausgangsüberlegungen (1/3)

1. Evaluierung ist in einen organisationalen Kontext eingebettet.
2. In diesem Kontext treffen unterschiedliche **Handlungslogiken** (Rationalitäten) aufeinander; unterschiedl. **Hierarchien von Wissensformen** (Erfahrungswissen vs. Evidenz)
3. Stakeholder haben unterschiedl. **Interessen**, Agenden, Werte
4. Sie agieren innerhalb institutioneller **Anreizsysteme** / Zwänge
5. und mit **beschränkten Ressourcen**.



## Ausgangsüberlegungen (2/3)

6. Evaluierungen erfüllen (für unterschiedl. Stakeholder) **versch. Funktionen** – nicht nur die Generierung von Evidenz, sondern bspw. Legitimierung, symbolische Funktion etc.
7. **Wissenschaftliche Ansprüche** sind nicht allen Stakeholdern vertraut
8. Die meisten der hier genannten organisationalen Faktoren sind **nicht oder schwer auflösbar / veränderbar**; sie haben aber...



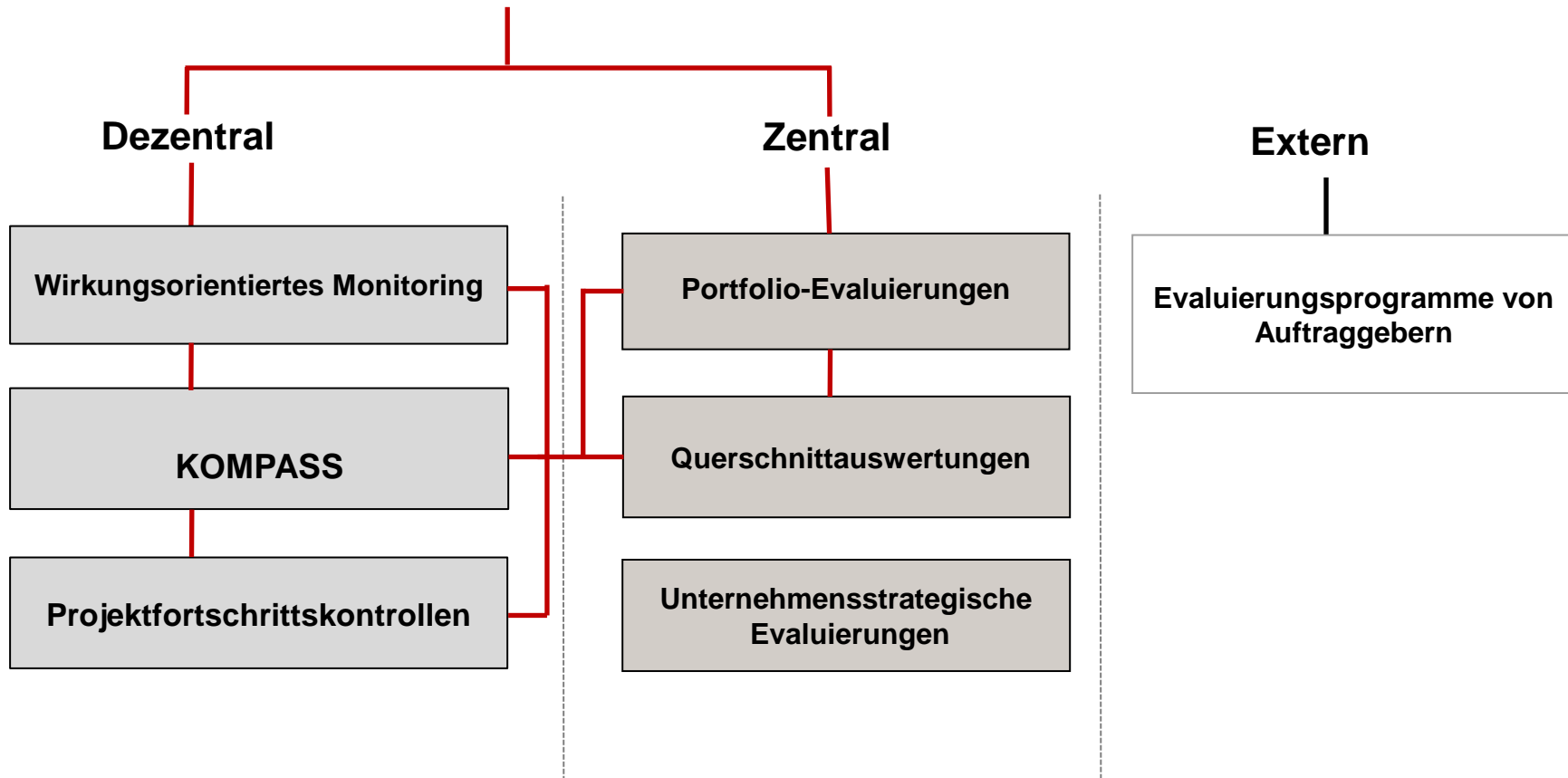
## Ausgangsüberlegungen (3/3)

### ... Konsequenzen für Evaluierung

- Die methodische Qualität von Evaluierungen ist (teilweise unberechtigt, teilweise berechtigt?) oftmals **nicht prioritär**
- Folgen für Evaluierungspraxis (z.B. beschränkte **Ressourcen**, **Gutachter-Auswahl** etc.)
- **Geringe methodische Qualität**



# M&E Instrumente der GIZ





## Spannungsfelder bei dezentralen (internen) Evaluierungen...

- Dezentrale Evaluierungen werden von den **AV** (Projektleiter vor Ort) **gesteuert**
- Interessen des AV: Vorbereitung einer Folge-Maßnahme (**planerischer Aspekt**); **Angst** vor schlechter Bewertung ?; Erfüllung von Vorgaben (intern und des Auftraggebers) ...
- Kein positives / negatives Anreizsystem für method. Qualität
- **Geringer Einfluss der Stabsstelle M&E** (= „Methoden-Anwalt“)
- Achilles-Verse: **Gutachter Auswahl, GA-Kompetenz**



## ... und Lösungsansätze

- **Beratungsangebot** durch die Stabsstelle (Anleitungen, Standardgliederung, [Help-m-e@giz.de](mailto:Help-m-e@giz.de))
- **Vorgaben**, bspw. 4-Augen-Prinzip bei **GA-Auswahl**
- **Fortbildungen** zu M&E für AV und Mitarbeiter
- **Gutachter-Pool** für Monitoring (→ Datenbasis)
- **Weitere Überlegungen** zur methodischen Qualitätssicherung:  
Deskstudy? Begleitung vor Ort? Einführung und Kommentierung  
IR? Aber: Gefahr der Überfrachtung





## Ausgangssituation bei zentralen (unabhängigen) Evaluierungen...

- Die Stabsstelle beauftragt und steuert Evaluierungen
- Mehr zeitliche, finanzielle und personelle Ressourcen
- Umfangreiches Qualitätssicherungssystem - standardisierter, dokumentierter Prozess und Formatvorlagen
  - Für Datenerhebung und Analyse
  - Für Formulierung von Empfehlungen
- Bsp. für Maßnahmen zur Annäherung an einige Gütekriterien: ...



## Qualität in der Datenerhebung und Analyse bei zentralen (unabhängigen) Evaluierungen...

- Vorlage eines Inception Reports welcher das anvisierte methodische Vorgehen darstellt.
- Annotierung gibt Qualitätsansprüche vor.
- Interne Validierung durch:
  - Methoden- und Datentriangulation
  - Gemischte Gutachterteams : Fachsektorexperte und Methodenexperte, internationale und nationale Fachkraft



## Qualität in der Datenerhebung und Analyse bei zentralen (unabhängigen) Evaluierungen...

- Int. Validierung durch:
- Peer Reading durch Fachsektor-KollegIn und MethodenkollegIn, teilw. durch Projektpersonal.
- Kalibrierung innerhalb der Stabsstelle
- Abnahme und Vertragsergänzung



## Qualität in der Datenerhebung und Analyse bei zentralen (unabhängigen) Evaluierungen...

- Transparenz, ggfs. Reproduzierbarkeit durch:
- Explizite Dokumentation von Methoden und Ergebnissen der qualitativen oder quantitativen Datensammlung und -Auswertung (Interviewleitfäden, Fragebögen, Transkriptionen, Kodierungen, Protokolle, etc.)



## Qualität in der Datenerhebung und Analyse bei zentralen (unabhängigen) Evaluierungen...

- Ext. Validität – Generalisierbarkeit
- Insbesondere interessant für Pilotvorhaben
- Klärung der Frage, ob die Ergebnisse kontextabhängig sind.
- Querschnittauswertungen: Vorschaltung von Metaevaluierungen vor Reviews und Synthesen zur Bewertung der methodischen Qualität der Einzelevaluierungen
- Ausbau von Hochschulkooperationen



## Qualität der Empfehlungen bei zentralen (unabhängigen) Evaluierungen...

- Trennung von Beschreibung und Bewertung und Empfehlung
- Vorgaben in der annotierten Berichtsgliederung = Referenzrahmen für die Formulierung von plausiblen, rationalen, evidenzbasierten und intersubjektiv nachvollziehbaren Bewertungen und Empfehlungen
- Transparent machen welche Forschungsergebnisse in die Bewertungen und Empfehlungen einfließen, wie diese gewichtet sind und warum.



## Qualität der Empfehlungen bei zentralen (unabhängigen) Evaluierungen...

- Interne Validierung durch Peer Review durch mindestens eine/n Fach- und eine/n MethodenkollegIn.
- Validierung der Verwertbarkeit durch eine Management Response



## Herausforderungen bei zentralen (unabhängigen) Evaluierungen...

- ...Gutachterangebot (fachliche versus methodische Kompetenz)
- Finanzielle und zeitliche Restriktionen
- „Methodenfantasielosigkeit“
- Datenbasis
- Vorstellungen der (internen) Auftraggeber, wenig Bewusstsein über methodische Ansprüche und Machbarkeit





## Herausforderungen bei zentralen (unabhängigen) Evaluierungen...

- Qualitätssicherung vs. politische Einflussnahme
  - Bspw. Beurteilung der Angemessenheit von Methoden oder Indikatoren im Rahmen der Qualitätssicherung;
  - Subjektive Einschätzung oder klare Indikation?
- Metaevaluierungen: Was sind die Mindeststandards, die die wissenschaftliche Güte garantieren, die den Ausschluss der einzelnen Studie aus der Ergebnissynthese rechtfertigen?



## Diskussionsfragen

- Wie können **institutionelle Anreize** (positiv / negativ) für methodische Qualität geschaffen werden?
- Wie viel methodische Qualität „**muss sein**“ für die jeweils verschiedenen Evaluierungszwecke? Was sind dann die **Ressourcenimplikationen**?
- Verständnis von Evaluierung als **institutionelle Praxis** oder **Forschung**?