Zur (Nicht-)Angemessenheit von Signifikanztests in der Evaluation*[*]

Wolfgang Beywl[1], Köln und Aarau, 31. März 2010

Evaluationen stützen häufig die Bewertung eines Programms darauf ab, welchen Unterschied es macht, sei es gegenüber der Ausgangssituation davor, oder gegenüber einer Vergleichsgruppe ohne Programmteilnahme. Hierfür werden Aussagen zur Signifikanz von datenbasierten Ergebnissen gemacht. Das Wort Signifikanz hat zwei ganz verschiedene Konnotationen: einmal statistisch, einmal praktisch. Diese beiden Terme werden zunächst geklärt. Irrtümlicherweise werden oft für beide Sachverhalte statistische Signifikanztests berechnet und in Evaluationsbeeichten ausgewiesen. Dieser Beitrag klärt, unter welchen Umständen solche Testverfahren der schließenden Statistik sinnvoll und zulässig sind und formuliert hierfür schrittweise anwendbare Regeln, deren Einhaltung zu genaueren und glaubwürdigeren Evaluationen beiträgt. Abschließend wird kurz angesprochen, wie praktische Signifikanz oder Bedeutsamkeit bestimmt werden kann.

In Evaluationen* geht es häufig darum, den Beitrag zu bestimmen, den eine bestimmte Intervention* zur Zielerreichung* oder Wirksamkeit* eines Programms* leistet, um seinen Wert* zu beurteilen, seine Übertragfähigkeit* an andere Orte bzw. in andere Kontexte* zu bewerten usw.

Ein typischer Fall: Um den Wert eines Programms zu bestimmen, soll festgestellt werden, ob und in welchem Maße es seine Ziele* erreicht und die gewünschten Resultate* erbringt. Gemessen werden sollen insbesondere Outcomes*, das heißt z. B. Lernzielerreichungen bei den Zielgruppen* des Programms. Es wird nach einem Erhebungsdesign*[2] gesucht, mit dem belegt werden kann, ob Zuwächse der Outcomes auf die Interventionen rückführbar sind. Hierzu misst man die Outcomes zu unterschiedlichen Zeiten bei derselben Personengruppe (z. B. vor und nach einer Intervention) oder zum gleichen Zeitpunkt bei mehreren Personengruppen (also z. B. bei einer Experimentalgruppe* mit und einer Vergleichsgruppe* ohne Intervention). Die Herausforderung liegt dann darin, auf Basis der gewonnenen Daten zu bestimmen, ob die gemessenen Unterschiede (oder Nicht-Unterschiede) eine (wie stark) positive oder negative Bewertung* des Programms rechtfertigen.

---

* Alle mit Stern gekennzeichnete Terme sind im "Glossar wirkungsorientierte Evaluation" (Beywl/Niestroj 2009) definiert, das insgesamt 369 Begriffe inklusive der vollständigen Literaturverweise enthält. Ein Glossarauszug mit allen markierten Termen findet sich im Anhang zu diesem Text.

1 Für diesen Text habe ich mir bei Fachleuten mit vertiefter statistischer Expertise Rat geholt, um Sicherheit zu den Grundannahmen zu gewinnen. Das heißt nicht, dass ich deren Fachpositionen zutreffend wiedergebe; sie haben den Text nicht "abgenommen". Evaluatorinnen und Evaluatoren, die Statistik auf Basisniveau anwenden, haben mir mit vielen Nachfragen geholfen, die Mehrebenen-Argumentation besser nachvollziehbar zu machen, ein Stück zu didaktisieren. Allen bin ich sehr zu Dank verpflichtet und bitte um Nachsicht, wenn die folgende alphabetische Aufzählung nicht vollständig ist: Lars Balzer, Hanne Bestvater, Dieter Brauns, Thomas Gautschi, Jan Hense, Susanne Mäder, Benjamin Merle, Uwe Neugebauer, Melanie Niestroj, Reinhard Oppermann, Dirk Scheffler, Berthold Schobert, Dörte Schott, Werner Stangl, Sandy Taut, Elitsa Uzunova, Peter Michael Vock.

2 Zu den Erhebungsdesigns in der Evaluation siehe Balzer (2005).

Die zwei «Signifikanzen» in der Evaluation

In der Evaluation gibt es zwei unterschiedliche Bedeutungen von "Signifikanz", die unabhängig voneinander sind. Dass die beiden Sachverhalte mit dem gleichen Wort bezeichnet werden, ist vielfach Ausgangspunkt für Missverständnisse und Kunstfehler.

I) «**Statistische Signifikanz**» als Term der Inferenzstatistik. Die Inferenzstatistik wird gelegentlich gegenüber der deskriptiven Statistik als "fortgeschrittener" oder "wertvoller" wahrgenommen, da sie nicht nur die erhobenen Daten* beschreibt und zu Parametern verdichtet (z. B. Mittelwerte), sondern auch Schlüsse auf deren Signifikanz ermöglicht, also die Wahrscheinlichkeit, mit der bei kleineren Erhebungsmengen*[3]/ gemessene/für sie berechnete Parameter induktiv auf größere Mengen oder gar auf die Gesamtheit der existierenden Merkmalsträger (Grundgesamtheiten* oder Populationen) verallgemeinert werden können. Es geht also um Generalisierung. Mit relativ geringem Aufwand (Erhebungsmenge, welche die statistisch erforderliche Mindestgröße aufweist, z. B. 3.500 Elemente), kann mit einer angebbaren Wahrscheinlichkeit auf Verteilungen in einer Gesamtheit von z. B. 60 Millionen Elementen geschlossen werden (dies ermöglicht z. B. kostengünstige Wahl- oder anderweitige Umfrageforschung). Signifikant sind in diesem Denken Ergebnisse, die wahrscheinlichkeitsstatistisch belegt sind.

II) «**Praktische Signifikanz**» [«*significance (of a programme*)]*» als Synonym für Relevanz oder Bedeutsamkeit* eines evaluierten Programms. Diese ergibt sich aus der Bewertung, also der Interpretation* der erhobenen Daten und der darin hergestellten Relation zu den gesetzten Kriterien*. Sind z. B. die Unterschiede zwischen den gemessenen Mittelwerten von Teilgruppen so groß, dass das Programm mehr oder weniger positiv (oder negativ) zu bewerten ist? Für diese Aufgabe sind andere Verfahren als die der schließenden Statistik erforderlich, nämlich systematische Bewertungsverfahren, wie die von Scriven (1981) vorgeschlagene «Weight-and-Sum-Methodology» oder die «Multiattribute Evaluation» nach Edwards/Newman (1982). Die Feststellung der Bedeutsamkeit geschieht, in Analogie etwa zur Feststellung von Gültigkeit* (Validität) von Ergebnissen*, nicht mittels Statistik sondern mittels logischem Schließen, oft abgestützt auf theoretisch begründeten Annahmen, z. B. Modellierung im Rahmen «logischer Modelle»* (z. B. Wyatt Knowlton/Philips 2009) oder verschiedener «Validierungsverfahren» (vgl. z. B. Lamnek, 2005).

Eine solche bewertende Schlussfolgerung* wird häufig mit Hilfe von Signifikanztests aus der «schließenden Statistik»[4] begründet: Es wird z. B. ausgewiesen, auf welchem Signifikanzniveau Gruppen von Personen - bezüglich der bei ihnen gemessenen Outcomes - ähnlich oder verschieden sind, bezogen auf statistische Parameter von Verteilungsdaten wie Mittelwerte oder Korrelationskoeffizienten.

3 Um Missverständnisse zu vermeiden, wird in diesem Text allgemein von Erhebungsmengen gesprochen, die wie folgt definiert sind: "Menge von Erhebungseinheiten, die nach einem ausgewiesenen Auswahlverfahren gebildet ist und bei der Daten erhoben werden." (Beywl/Niestroj 2009, S. 34) Die [repräsentative] Stichprobe* ist ein später in Text beschriebener Sonderfall der Erhebungsmenge, welche die Grundgesamtheit "repräsentiert". Stichproben im so verstandenen Sinne sind demnach immer repräsentativ.

4 So wird die z. B. von z. B. Kromrey (2006, S. 393) genannt; alternative Benennungen sind "induktive Statistik" (z. B. Assenmacher 2009) oder "Inferenzstatistik" (Nachtigall 2006); teils wird auch von "Teststatistik" gesprochen (z. B. Field 2009, S. 26-29; S. 285-308) oder vom "Testen von Hypothesen" (z. B. Bortz/Döring 2006, S.494).

Während bei Zusammenhangsmaßen bereits die Größe z. B. des Korrelationskoeffizienten oder die «Effektstärke» (z. B. *r* größer .6) einen ersten (dabei meist nicht allein ausreichenden) Hinweis auf die Bedeutsamkeit des Ergebnisses (nicht aber auf dessen Verallgemeinerbarkeit; s. u.) gibt[5], besteht z. B. bei Mittelwertsunterschieden Unsicherheit darüber, ob/in welchem Maße der gemessene absolute Unterschied bedeutsam ist (also z. B. Klassen-Mittelwert in einem standardisierten Mathematiktest auf einer Notenskala 6 bis 1 (1=sehr gut) zu Beginn des Jahres 2.8, am Ende 2.5). Um diese Unsicherheit zu mindern, werden oft Signifikanztests gerechnet und Signifikanzniveaus in den Evaluationsberichten* ausgewiesen.

Auch wenn Signifikanztests gelegentlich sinnvoll und zulässig sind, so sind sie doch meist ein falscher und irreführender Weg, um ein Programm systematisch zu bewerten:

I. Signifikanztests sind ausschließlich für bestimmte Absichten sinnvoll (nämlich, wenn *Teilg*ruppen einer größeren Gesamtheit/von größeren Gesamtheiten betrachtet werden). Für eine solche Verallgemeinerungsabsicht sind Signifikanztests ausschließlich dann zulässig, wenn die Teilgruppen repräsentativ sind für die Gesamtheit(en), auf die geschlossen werden soll. Beide Voraussetzungen *fehlen* in vielen Evaluationen.[6] Das nachfolgende Regelwerk unter I. legt dar, unter welchen Voraussetzungen Signifikanztest angemessen und zulässig sind und wann nicht.

II. Auch dann, wenn Signifikanztests angemessen und zulässig sind, muss in der Evaluation ein weiterer Schritt getan werden, um die (Nicht-)Relevanz eines Mittelwertunterschiedes festzustellen: Beispiel: Wenn bei großen repräsentativen Erhebungsmengen ein kleiner, dabei statistisch hochsignifikanter Mittelwertsunterschied gemessen wurde, ist zusätzlich ein Urteil erforderlich, ob dieser den Einsatz des Programms rechtfertigt (im Sinne z. B. einer Kosten-Outcome- oder Kosten-Wirkungsanalyse* oder -abschätzung). Dies erfordert die Angabe von Kriterienpunkten* bzw. Kriterienzonen*, z. B. Zahlenwerten (wie etwa: ein Unterschied von 0.3 Notenpunkten oder mehr führt zu der Bewertung, dass die eingeführte Zusatz-Unterrichtsstunde in Mathematik beibehalten werden sollte. Andernfalls sollte nach kostengünstigeren Lösungen gesucht werden). Das Festlegen solcher Kriterienpunkte ist ein unverzichtbarer Schritt, um die Bedeutsamkeit z. B. eines Gruppenunterschiedes beurteilen zu können (siehe unter «II Wege zur Bewertung der Bedeutsamkeit»).

---

5 "Bedeutsam" heißt hier, dass es sich um einen starken Zusammenhang der beiden Variablen handelt; ob dieser auch relevant für das Programm und schließlich ausschlaggebend für Schlussfolgerungen oder Empfehlungsoptionen ist, entscheidet sich mit dem logischen Modell des Evaluationsgegenstandes*, dem Evaluationszweck* und anderen übergeordneten Aspekten.

6 Diese Behauptung resultiert für aus meinen Erfahrungen mit ca. 200 selbst durchgeführten oder im Rahmen von Qualifikationsarbeiten oder Beratungen begleiteten Evaluationen, unter denen es lediglich eine Handvoll mit repräsentativen Stichproben gab. Sie müsste empirisch erhärtet werden. Wenn in der Literatur die Bedeutung der schließenden Statistik für die Sozialwissenschaften wie folgt begründet wird, "Wie in den Kapiteln über die Inferenzstatistik mehrfach betont, haben wir es in den Sozialwissenschaften bei empirischen Untersuchungen in der Regel nur mit Stichproben zu tun, von denen wir auf die uns eigentlich interessierenden Grundgesamtheiten schließen wollen" (Kriz 1973, S. 219), wird oft der kurz darauf folgende Satz übersehen: "Die Voraussetzungen für eine Zufallsstichprobe müssen dann natürlich erfüllt sein." In der Evaluation handelt es sich meistens um angefallene Stichproben oder Vollerhebungen. Im ersten Fall ist ein Schließen auf eine Grundgesamtheit meist nicht zulässig, da die Stichproben nicht repräsentativ für die Grundgesamtheit sind, und im zweiten Fall ist das Schließen auf eine Grundgesamtheit überflüssig (vgl. Regel I zur Anwendung von Signifikanztests, S. 2).

Statistische Signifikanz im Sinne von Übertragbarkeit* oder Generalisierbarkeit* alleine sagt nichts aus über praktische Signifikanz oder Bedeutsamkeit eines Programms, und umgekehrt: Statistische Signifikanz kann z. B. sowohl für relevante wie für irrelevante Korrelationen zutreffen. Für eine bestimmte Erhebungsmenge (z. B. eine Schulklasse) festgestellte praktische Signifikanz darf erst nach einer zusätzlichen statistischen Prüfung auf eine größere Gesamtheit übertragen oder verallgemeinert werden; hier können Signifikanztests zum Zuge kommen.

Zu I: Regeln zur Anwendung von statistischen Signifikanztests in der Evaluation

Nachfolgend sind Regeln formuliert, welche bei der Entscheidung für Anwendung oder Nicht-Anwendung von Signifikanztests beachtet werden sollen. Dabei werden (1) zunächst Konstellationen genannt, in denen solche Tests überflüssig sind. Dann (2) solche, in denen sie meist nicht zulässig sind. Dann wird auf Konstellationen eingegangen, in denen sie zulässig sind (3 und 4); letzteres erfolgt unter Bezug auf die vorangegangenen Regeln. Dieser Argumentationsgang von der Überflüssigkeit/Unzulässigkeit zur (eingeschränkten) Sinnhaftigkeit/Zulässigkeit soll unterstreichen, dass Signifikanztests in der Evaluation die Ausnahme von der Regel sind.

1 Überflüssig sind Signifikanztests (wie alle anderen Verfahren der Inferenzstatistik) bei Vollerhebungen (n der Erhebungsmenge = N der Grundgesamtheit), da die Menge, bei welcher die Daten erhoben werden (alle Fälle), identisch ist mit der Gesamtheit der Fälle, über die Aussagen gemacht werden sollen. Hier sind ausschließlich die Verfahren der deskriptiven Statistik sinnvoll. Diese Aussage mag manchem banal erscheinen oder tautologisch, denn schon vom Namen her geht es ja der Inferenzstatistik um das Schließen oder die statistische Induktion auf ein größeres Ganzes. In der Forschungspraxis finden wir jedoch nicht selten Signifikanz-Maße z. B. für einen Mittelwertsvergleich von zwei Teilgruppen (z. B. n= 40 und n=60), deren Elementezahl addiert der Grundgesamtheit (N=100) entspricht. Diese sind überflüssig, da sie keine sinnvolle Zusatzinformation geben, die die Glaubwürdigkeit der Evaluation erhöhen oder sie "wertvoller" machen. Sie sind nicht nur Zeit- und Platzverschwendung sondern können darüber hinaus in die Irre führen, da sie mit «praktischer Signifikanz» verwechselt werden. Darüber hinaus verkomplizieren sie unnötigerweise die Ergebnisdarstellungen: Statistische Laien verstehen sie weder noch können sie diese kontrollieren. Sie sind in solchen Fällen nichts anderes als ein Bluff von (scheinbaren) Experten.

2 Meist nicht zulässig sind Signifikanztests bei angefallenen Erhebungsmengen. "Angefallen" meint, dass das Zustandekommen der Erhebungsmenge nicht derart kontrolliert erfolgte, dass sie sicher ein verkleinertes (=repräsentatives) Abbild der Grundgesamtheit darstellt.

   Beispiele für angefallene Erhebungsmengen:

   (a) Selbstselektion oder andere «willkürliche» Auswahlverfahren*, also z. B. Auswahl immer der folgenden Person, zu der nach einem abgeschlossenen Straßeninterview als nächstes Augenkontakt hergestellt werden kann (Convenience- oder Bequemlichkeits-Auswahl), und die dann auch bereit ist zu einem Interview (Selbstselektion).

   (b) (Stark) unvollständig [undercoverage] oder (stark) überzogen [overcoverage] realisierte Erhebungsmenge einer Zufallsauswahl (Diekmann 2007, S. 376-380). Beispielsweise bei einem Rücklauf von 30% oder 150 % von einer angezielten Zufallsstichprobe.

   (c) «Gematchte» Gruppen, bei denen die Experimentalgruppe angefallen ist (bspw. Jugendliche haben sich aus eigenem Antrieb für eine Ausbildung entschieden, womit die Ausbildungsgruppe unkontrolliert zustande gekommen ist) und die Vergleichsgruppe dieser Experimentalgruppe dann angepasst wird. Zwar ist die Vergleichsgruppe (durch Bestimmung statistischer Zwillinge → Matching*) in Bezug auf die Experimentalgruppe absichtsvoll und kontrolliert gebildet; keine der beiden Gruppen für sich *und auch nicht beide zusammen, repräsentieren* jedoch eine Grundgesamtheit, auf die geschlossen werden könnte.[7]

7 Dies ist keinesfalls ein Argument gegen Vergleichgruppendesigns, denn sie ermöglichen ein gewisses Maß der Kontrolle bewirkenden Faktoren (unabhängiger Variablen). Wann immer sinnvoll

3 In Abweichung von Regel 2 sind Signifikanztests in Fällen von (a), (b) und (c) zulässig wenn die folgenden beiden Bedingungen additiv gegeben sind:

A Die Grundgesamtheiten, auf die geschlossen werden soll, sind beschrieben (einschließlich Unsicherheiten der Beschreibung); dies setzt voraus, dass die Zahl der Elemente in der Grundgesamtheit bekannt ist. Für unbegrenzte und zahlenmäßig nicht bekannte Grundgesamtheiten (z. B. alle in einem Jahr auf der Welt an Festnetztelefonen gesprochenen Wörter) ist ausschließlich eine echte Zufallsauswahl* der Weg, um Repräsentativität zu erreichen.

und

B Eine theoretisch fundierte und datenbasierte Begründung ist gegeben, weshalb die vorliegenden (angefallenen) Erhebungsmengen einer Zufallsstichprobe nahekommen (auch Grenzen sind angegeben). Hierzu muss ein theoretisches Wirkmodell der relevanten Variablen vorliegen, also nicht nur, was sind die relevanten bewirkenden (unabhängigen) und was sind die bewirkten (abhängigen) Variablen, sondern auch: Was sind die wichtigsten beeinflussenden Drittvariablen (insb. Moderatorvariablen*, welche den Einfluss einer unabhängigen auf eine abhängige Variable beeinflussen, sei es verstärkend, abschwächend oder gar umkehrend). Es reicht keinesfalls zu überprüfen, ob in der angefallenen Erhebungsmenge anteilig gleich viele Männer und Frauen, Alte und Junge, Inländer und Ausländer sind, sondern diese Überprüfung muss für alle relevanten Drittvariablen vorgenommen sein und eine genügende Ähnlichkeit der Verteilungen der Werte dieser Variablen in Erhebungsmenge und Grundgesamtheit ergeben.[8]

Es wird aus arbeitsökonomischen Gründen für die Anwendung der Regel 3 das folgende schrittweise Vorgehen empfohlen:

i) Grobe Abschätzung, ob A und B beide gegeben sind

ii) Erstes Rechnen von Signifikanztests ´auf Probe´

iii) Für signifikante Ergebnisse: datenbasierte, theoretisch angeleitete Prüfung, ob A und B gegeben sind

iv) Aufnahme der Ergebnisse zu ii) in die Evaluationsergebnisse*/Berichterstattung* bei Wiedergabe der unter iii) geprüften Voraussetzungen

4 Zulässig (und oft auch sinnvoll) sind Signifikanztests (z. B. Mittelwertvergleiche) bspw. zur Überprüfung von Unterschieds- oder Zusammenhangshypothesen (Bortz 2005, S. 135-239) bei

- zufallsgezogenen (Art des Zufallsziehungsverfahrens ist angegeben) und *dadurch* repräsentativen Erhebungsmengen (Stichproben)

  oder

- Erhebungsmengen, die auf theorie-/forschungsergebnis-geleiteten quotierten Auswahlverfahren beruhen und *dadurch* repräsentativ sind.
  (Bedingung 3.B ist gegeben und auch Bedingung 3.A ist gegeben, da die Verteilung der relevanten Quotierungsmerkmale in der Grundgesamtheit für die Durchführung einer solchen Quotierung bekannt sein muss, um gezielt zu quotieren. Hingegen sind Erhebungsmengen, die durch eine willkürliche Quotierung gebildet wurden, ein Unterfall von 2.b.)

---

und durchführbar sollen diese gewählt werden, doch muss in den unter Regeln 1 und 2 genannten Fällen auf Signifikanztests von Gruppenunterschieden verzichtet werden.

8 Der Umkehrschluss, dass Quotierung etwa nach Merkmalen Geschlecht oder Alter ausschliesslich dann angeraten ist, wenn dadurch Repräsentativität der Erhebungsmenge angestrebt und herstellbar ist, ist falsch: eine wie immer grobe Quotierung ist allemal besser als eine willkürliche Bildung der Erhebungsmenge.

und

wenn die Erhebungsmenge die in den inferenzstatistischen Lehrbüchern geforderten Dateneigenschaften aufweist: angemessene Stichprobengröße in Relation zur Grundgesamtheit, ähnliche Varianzen zu vergleichender Verteilungen, annähernde Normalverteilung in den zu vergleichenden Gruppen usw. (wenn die ersten zwei Bedingungen erfüllt sind, aber keine Normalverteilung vorliegt, ist die Anwendbarkeit nonparametrischer Tests zu prüfen, vgl. Field 2009, S. 539-583).
Zu den für die einzelnen Tests je zu prüfenden Voraussetzungen, z. B. bzgl. Skalenniveaus (nominal, ordinal, intervall, metrisch) finden sich in der Standardliteratur Hinweise auf je anzuwendende Tests (z. B. Bortz 2005, S. 213-235; Entscheidungsbaum für angemessene Tests in Field 2009, S. 781). Auch für kleine repräsentative Erhebungsmengen gibt es spezielle statistische Verfahren (Prein/Kluge/Kelle 1994), für die ebenfalls zu prüfen ist, ob die erforderlichen (Verteilungs- und anderen) Bedingungen für ihre Anwendbarkeit gegeben sind.

Alle diese Testverfahren werden in der Evaluation eher selten sinnvoll und zulässig sein. Wenn die Voraussetzungen für sie vorliegen, sollten sie auf jeden Fall eingesetzt werden, denn man kann wertvolle zusätzliche Informationen* gewinnen und den Wert der Evaluation erhöhen.

Zu II) Hinweise für eine Bewertung der praktischen Signifikanz (Bedeutsamkeit)

In der Evaluation muss man vor der Datenerhebung oder spätestens vor dem Vergleich der Gruppenergebnisse Kriterienpunkte und -zonen festlegen, bei deren Übertreffen/Erreichen man von einem bedeutsamen Unterschied spricht. Als methodischer Zugang zu einer solchen «antizipatorischen Ergebnisnutzung»* eignet sich z. B. die mock-evaluation nach Patton (2008, S. 472-473).

Hierbei müssen in den für den Vergleich relevanten Kriteriendimensionen operationalisierte Kriterien mit Kriterienpunkten gesetzt werden, und zwar vorab (vor der Messung, spätestens vor der Auswertung); z. B.: Wann ist ein Mittelwertsunterschied so groß, dass die Wirkungseinschätzung* zu einer Bewertung des Programms als «befriedigend» oder besser führt (Erfolgspunkt*, evtl. Erfolgsspanne*). *Wie* man zu diesen Kriterienpunkten kommt, ist ein anderes Thema.

Achtung: Wenn man auf diese Weise feststellt, dass ein Unterschied/ein Zusammenhang bedeutsam ist, gilt dies ausschließlich für die untersuchte Erhebungsmenge (eventuell für den Vergleich dieser mit einer Vergleichsgruppe). Verallgemeinerung auf größere bzw. andere Fälle/Mengen ist nicht zulässig. Hierzu bedürfte es eines Signifikanztests im Sinne der statistischen Signifikanz, der zulässig ist, wenn die Erhebungsmenge die statistisch geforderten Voraussetzungen erfüllt.

Literatur

Assenmacher, Walter (2009): Induktive Statistik. 2., überarb. Aufl. Berlin: Springer

Averch, Harvey A. (2004): "Using Expert Judgment". In: Wholey, Joseph S./Hatry, Harry/Newcomer, Kathryn (Hg.): The handbook of practical program evaluation. San Francisco: Jossey-Bass, S. 292-309.

Balzer, Lars (2005): Wie werden Evaluationsprojekte erfolgreich? Eine empirische Studie und ein integrierender theoretischer Ansatz zum Evaluationsprozess. Landau: Verlag Empirische Pädagogik.

Beywl, Wolfgang/Niestroj, Melanie (2009): Das A-B-C der wirkungsorientierten Evaluation. Glossar - Deutsch Englisch - der wirkungsorientierten Evaluation. 2., vollständig überarbeitete Auflage. Köln: Univation.

Bortz, Jürgen (2005): Statistik für Human- und Sozialwissenschaftler. Berlin: Springer.

Davidson, Jane E. (2005): Evaluation methodology basics. The nuts and bolts of sound evaluation. Thousand Oaks: Sage.

Diekmann, Andreas (2007): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. 17. Auflage. Reinbek: Rowohlt.

Edwards, Ward/Newman, J. Robert (1982): Multiattribute evaluation. Beverly Hills: Sage Publication. 3rd edition. London: Sage.

Field, Andy (2009): Discovering statistics using SPSS (and sex and drugs and rock 'n' roll). 3rd edition. London: Sage.

Henry, Gary T. (2002): "Choosing criteria to judge program success: A values inquiry". In: Evaluation, Jg. 8, No. 2, S. 182-204.

Kriz, Jürgen (1973): Statistik in den Sozialwissenschaften. Einführung und kritische Diskussion. Reinbek: Rowohlt.

Kromrey, Helmut (2006): Empirische Sozialforschung. Modelle und Methoden der standardisierten Datenerhebung und Datenauswertung. 11., überarbeitete Auflage. Stuttgart: Lucius & Lucius.

Lamnek, Siegfried (2005): Qualitative Sozialforschung. Lehrbuch. 4., vollständig überarbeitete Auflage. Weinheim: Beltz.

Patton, Michael Q. (2008): Utilization-focused evaluation. 4th edition. Thousand Oaks: Sage.Prein, Gerald/Kluge, Susann/Kelle, Udo (1994): Strategien zur Sicherung von Repräsentativität und Stichprobenvalidität bei kleinen Samples:

Scriven, Michael (1981): "The Weight and Sum Methodology". In: American Journal of Evaluation, Jg. 2, No. 1, S. 85-90

Scriven, Michael (2007): Key Evaluation Checklist KEC. Michigan: Western Michigan University.

Wyatt Knowlton, Lisa W./Phillips, Cynthia C. (2009): The logic model guidebook. Better strategies for great results. Los Angeles: Sage.

Vedung, Evert (2007): "Merit criteria and performance standards"(Hg.): *Evaluation research methods.* London: Sage, S. 183-200.

Anhang: Glossarbegriffe in diesem Text

Begriff Univation	Begriff engl. [Original]	Beschreibung
antizipatorische Ergebnisnutzung	mock evaluation; simulation of use	Simulation einer späteren [[Nutzung]] von [[Evaluationsergebnissen]]. Insbesondere für umfangreiche Evaluationsvorhaben sollte dies versucht werden, da Überlegungen zur erwarteten Nutzung die Ausgestaltung einer [[Evaluation]] sinnvoll beeinflussen können. (siehe auch [[simulierte Daten-Interpretations-Sitzung]]) (Patton 2008, S. 472-473; Wottawa/Thierau 2003, S. 89)
Auswahlverfahren	sampling strategy	Art und Weise (z. B. [[Zufallsauswahl]] oder „absichtsvoll“), nach der die [[Auswahl]] von [[Erhebungseinheiten]] erfolgt, bei denen [[Daten]] erhoben werden. (Cochran 2007; Patton 2002, S. 230-246)
Bedeutsamkeit (eines Programms)	significance (of a programme)	Aktuelle und künftige Wichtigkeit, Sichtbarkeit sowie (potenzieller) [[Impact]] eines [[Programms]] für die (Welt-)Gesellschaft. Sie zeigt sich insbesondere an der Langlebigkeit und der [[Übertragfähigkeit]] des Programms. Eine generische Kriteriendimension ([[Kriteriendimension, generische]]) für die [[Bewertung]] von Programmen. (Scriven 2007)
Berichterstattung (der Evaluation)	reporting (of the evaluation)	Aufgabe, die schwerpunktmäßig in der Phase der [[Ergebnisvermittlung]] einer [[Evaluation]] zu bearbeiten ist. Mit Hilfe von Medien, wie schriftlichen Berichten ([[Bericht, schriftlicher]]), Plakaten, Flyern oder Folien, allein stehend oder kombiniert mit mündlichen Präsentationen, werden die [[Evaluationsergebnisse]] den [[Adressierten]] vermittelt, was auch interaktiv erfolgen kann. (siehe auch [[Berichterstattungsformat]]) (Torres/Preskill/Piontek 2005; Alkin/Christie/Rose 2006)
Bewertung	judgement	Neben der [[Beschreibung]] eine Hauptaufgabe der [[Evaluation]]. Sie bringt die erzeugten beschreibenden [[Informationen]] in Relation zu ausgewiesenen [[Kriterien]], um schließlich zusammenfassende Urteile über [[Güte]] und/oder [[Tauglichkeit]] eines [[Evaluationsgegenstands]] zu treffen. (Mathison 2005, S. 214)
Daten	data	Menge aller Merkmalsmessungen. In geeigneter Form festgehaltene, archivierte und abrufbare symbolische Repräsentationen der bei [[Aussageeinheiten vorhandenen [[Merkmalsausprägungen. Daten sind durch empirische Erhebungsakte transformierte Merkmalsausprägungen und können damit, nach [[Auswertung mittels (quantitativer) Statistik und/oder qualitativer Verfahren für die [[Beschreibung und [[Bewertung im Rahmen einer [[Evaluation genutzt werden. (Bortz/Döring 2006, S. 2; Kromrey 2006, S. 228-234)
Erfolgspunkt	success criterion	[[Kriterienpunkt]] in einer für den [[Erfolg]] eines [[Programms]] relevanten [[Kriteriendimension]] (z. B. Innovationsgehalt, Reife, Stabilität, [[Zielerreichung]], [[Wirksamkeit]]), bei dessen Erreichen das Programm (in dieser Dimension) als erfolgreich bewertet wird (ggf. mit auf- und absteigenden Erfolgs-Gradierungen). Sollte frühzeitig durch relevante [[Beteiligte]] festgelegt werden. (siehe auch [[Erfolgsspanne]]; [[Minimal-Erfolgspunkt]], [[Optimal-Erfolgspunkt]]) (Henry 2002)
Erfolgsspanne	success zone	Eine positive Kriterienzone ([[Kriterienzone, positive]]), innerhalb derer ein [[Programm]] als [[Erfolg]] gewertet wird: Orientiert an den [[Zielen]] des Programms werden zwei [[Kriterienpunkte]] für ein operationalisiertes Kriterium ([[Kriterium, operationalisiertes]]) festgelegt. Erreicht das Programm den [[Minimal-Erfolgspunkt]] und überschreitet den [[Optimal-Erfolgspunkt]] nicht, so spricht man auf dieser Dimension von einem Programmerfolg. Wenn die Angabe quantifizierter Bewertungskriterien nicht möglich ist, kann dies durch eine genaue qualitative Beschreibung des erwünschten Zielzustandes ersetzt werden ([[Erfolgsbild]]). (Beywl/Schepp-Winter 1999, S. 68)
Ergebnisse (der Evaluation)	Evaluation findings	[[Evaluationsergebnisse]]
Erhebungsdesign	design of data collection	Teil des [[Datenerhebungsplans]], in dem festgelegt ist, bei wie vielen Gruppen (einer oder mehreren) von [[Erhebungseinheiten]] (oft sind dies Personen) zu welchen Untersuchungszeitpunkten [[Daten]] erhoben werden sollen. Dazu gehören auch Aussagen zur [[Auswahl]] der Erhebungseinheiten. (Fraenkel/Wallen 2008, S. 261-275; Balzer 2006, S. 218-232)
Erhebungsmenge	sample	Menge von [[Erhebungseinheiten]], die nach einem ausgewiesenen [[Auswahlverfahren]] gebildet ist und bei der [[Daten]] erhoben werden. (siehe auch [[Stichprobe]]) (Fraenkel/Wallen 2008, S. 90)

Evaluation	evaluation	Wissenschaftliche Dienstleistung, die insbesondere öffentlich verantwortete und/oder finanzierte [[Evaluationsgegenstände]] (Politiken, [[Programme]], Projekte, [[Maßnahmen]]…) systematisch, transparent und auf [[Daten]] gestützt beschreibt ([[Beschreibung]]) und ausgewogen bewertet ([[Bewertung]]), so dass [[Stakeholder]] ([[Auftraggebende]] etc.) die erzeugten [[Evaluationsergebnisse]] für vorgesehene [[Evaluationszwecke]] wie [[Rechenschaftslegung]], [[Entscheidungsfindung]] oder [[Verbesserung]] nutzen. (Beywl/Widmer 2009, S. 16; Balzer 2005)
Evaluationsbericht	evaluation report	Gegliederte, auf [[Adressierte]] zugeschnittene Wiedergabe der [[Evaluationsergebnisse]] in Wort, Bild oder audiovisueller Form. Um Nachvollziehbarkeit der [[Berichterstattung]] zu gewährleisten, in der Regel verbunden mit einer [[Beschreibung]] des [[Evaluationsgegenstands]] sowie des Vorgehens der [[Evaluation]]. Das Vorgehen soll auch beschrieben werden, um die [[Güte]] und [[Tauglichkeit]] der Evaluation überprüfbar zu machen. (siehe auch [[Evaluationsstandards]]) Unterschieden werden können an Berichtsformen insbesondere schriftliche Berichte ([[Bericht, schriftlicher]]), mündliche Berichte sowie solche, die sich audiovisueller Medien bedienen (alle Kombinationen sind möglich). (siehe auch [[Berichterstattungsformat]]) (Beywl/Faust 1999; Torres/Preskill/Piontek 2005)
Evaluationsergebnisse	evaluation findings	Die durch eine konkrete [[Evaluation]] bereitgestellten [[Beschreibungen]] und [[Bewertungen]] (auch [[Schlussfolgerungen]], und ggf. [[Empfehlungen]]). Sie werden im Rahmen von [[Rückmeldungen]] bzw. der [[Berichterstattung]] der Evaluation an [[Adressierte]] weitergegeben. Es gehört zur Aufgabe der [[Evaluierenden]], die [[Nutzung]] der Evaluationsergebnisse im Sinne der [[Evaluationszwecke]] anzustoßen und zu unterstützen. Es ist zu klären, in welchem Maße den Evaluierenden ein [[Redaktionsrecht]] im Zusammenhang mit der Veröffentlichung der Evaluationsergebnisse eingeräumt wird. Die Evaluationsergebnisse sind durch sprachliche Konvention von den [[Resultaten]] des [[Programms]] zu unterscheiden. (Beywl/Kehr/Mäder/Niestroj 2007, S. 63-70)
Evaluationsgegenstand	evaluand; evaluation object	Das, zu dem eine [[Evaluation]] [[Beschreibungen]] und [[Bewertungen]] erstellt. Oft werden Evaluationsgegenstände als [[Programme]] aufgefasst. Die theoretischen und methodischen Ansätze der [[Programmevaluation]] unterscheiden sich von denen der [[Personalevaluation]] und der [[Produktevaluation]]. (Stake 2004, S. 3-4)
Evaluationszweck	purpose of evaluations	Als Aussagesatz formulierte Bestimmung, was die [[Evaluation]] in Bezug auf den [[Evaluationsgegenstand]] und seine veränderbaren [[Bedingungen]] bewirken soll. Der Evaluationszweck wird von [[Auftraggebenden]] und/oder anderen [[Stakeholdern]] festgelegt und bezeichnet, was diese mit der [[Evaluation]] und den erzeugten [[Evaluationsergebnissen]] erreichen wollen. Der Evaluationszweck bestimmt die Richtung der Evaluation: Jeder Evaluationsschritt ist so anzulegen, dass er dem gewählten Evaluationszweck dient. Evaluationszwecke sind: [[Optimierung]]/[[Verbesserung]], [[Entscheidungsfindung]], [[Rechenschaftslegung]], [[Wissensmanagement]] und [[Wissensgenerierung]]. Der Terminus Zweck (der Evaluation) wird von [[Zielen]] abgegrenzt, die im Bereich des [[Evaluationsgegenstands]] zu finden sind (z. B. [[Detailziele]] eines [[Programms]]). (DeGEval 2002, S. 13)
Experimentalgruppe	treatment group	Die Personengruppe, die einer Experimentalhandlung unterzogen wird, indem sie bspw. an einem zu evaluierenden [[Programm]] teilnimmt. Die Experimentalgruppe wird im Rahmen einer [[experimentaldesigngesteuerten Evaluation]] mit einer [[Kontrollgruppe]] und im Rahmen einer [[quasi-experimentaldesigngesteuerten Evaluation]] mit einer [[Vergleichsgruppe]] verglichen, um Hinweise auf die [[Wirksamkeit]] des Programms zu gewinnen. (Bortz/Döring 2006, S. 113)
Generalisierbarkeit (von Evaluationsergebnissen)	generalisability (of evaluation findings)[ecological generalisability]	Sonderfall von [[Übertragbarkeit]], der dann vorliegt, wenn [[Evaluationsergebnisse]] auch für andere (identische oder sehr ähnliche) [[Evaluationsgegenstände]] in sehr ähnlichen [[Kontexten I]] oder zu späteren Zeitpunkten gelten. Dies setzt in der Regel stark standardisierte [[Programme]] sowie stabile Kontextbedingungen voraus (wie sie z. B. im Lernlabor gegeben sind). (Fraenkel/Wallen 2008, S. 104)

Grundgesamtheit	population	Gesamtheit aller Elemente, welche in empirischen Untersuchungen einbezogen werden. Es kann sich um die Gesamtheit der [[Aussageeinheiten]] handeln, um die der [[Erhebungseinheiten]] sowie um die der [[Auswahleinheiten]]. Grundgesamtheiten müssen zeitlich und räumlich bestimmt sein (z. B. alle angemeldeten Schüler und Schülerinnen in 4. Klassen von öffentlichen und als Ersatzschulen anerkannten privaten Schulen in Deutschland zum Stichtag 1. Januar 2010). Der Umfang der Grundgesamtheit kann bekannt oder unbekannt sein (z. B. alle sich illegal am 1. Juni 2011 in Österreich aufhaltenden Ausländer und Ausländerinnen), die einzelnen Elemente können identifiziert (z. B. über eine Liste oder einen Datensatz) oder nicht identifiziert sein. In der Statistik verwendet man den Grossbuchstaben „N“ als Abkürzung für die Grundgesamtheit. Wenn keine „Vollerhebung“ erfolgt, sondern Teilmengen einer Grundgesamtheit betrachtet werden, bezeichnet man diese als [[Auswahlmengen]]. Bei Erfüllung strenger Anforderungen („verkleinertes Abbild“) können bei Teilmengen gewonnene Ergebnisse auf die Grundgesamtheit übertragen werden (oder es kann sogar, bei vorhandener statistischer Repräsentativität, mit Hilfe der Inferenzstatistik auf Verteilungen in der Grundgesamtheit geschlossen werden). (Schnel/Hill/Esser 2008, S. 265-267)
Gültigkeit	validity	Eigenschaft zur Beschreibung und Bewertung der Güte von [[Datenerhebungsmethoden]] und [[Daten]]. Für die [[Genauigkeit]] der Datengrundlage einer [[Evaluation]] ist die interne und externe Gültigkeit zentral. Auch die Gültigkeit von [[Datenerhebungsinstrumenten]] kann beurteilt werden (Inhaltsvalidität; Kriteriumsvalidität; Konstruktvalidität). Die zur Gewinnung von Daten eingesetzten Instrumente sollen die Merkmale oder Verhaltensweisen, die sie zu messen vorgeben, auch tatsächlich erfassen. (Synonym: Validität) (Fraenkel/Wallen 2008 S. 146-182; Kromrey 2006, S. 200-205)
Informationen	information	Sprachlich gefasste und damit explizierte [[Evaluationsergebnisse]], die durch [[Interpretation]] von [[Daten]] entstehen. Die zweite Hauptphase im [[Evaluationsprozess]], die [[Informationsgewinnung]], dient deren systematischer und nachvollziehbarer Erzeugung. In dem Maße, in dem sich die [[Adressierten]] der Evaluationsergebnisse diese Informationen aktiv aneignen, sich mit den durch die [[Evaluierenden]] bereit gestellten [[Schlussfolgerungen]], [[Bewertungen]] und [[Empfehlungen]] auseinandersetzen, wird bei ihnen Wissen aktiviert, das für verschiedenste [[Evaluationszwecke]] einsetzbar ist ([[Wissensmanagement]]). (Seiler/Reinmann-Rothmeier 2004, S. 13)
Interpretation	interpretation	Auslegung, Erklärung der sozialen Bedeutung, des Zustandekommens eines, oder des Zusammenhangs mehrerer [[Merkmale]] eines [[Evaluationsgegenstands]]. Interpretation erzeugt [[Evaluationsergebnisse]] durch Ordnen, Verdichten oder Verknüpfen von [[Daten]] sowie deren Transformation in [[Informationen]], also [[Beschreibungen]], [[Schlussfolgerungen]], [[Bewertungen]] und ggf. [[Empfehlungen]]. (Bude 2004; Denzin 1998)
Interventionen (eines Programms)	programme interventions	[[Aktivitäten]] der in einem [[Programm]] Tätigen, welche direkt auf die Auslösung von [[Outputs]], [[Outcomes]] oder [[Impacts]] gerichtet sind. In detailliert ausgefüllten [[logischen Modellen]] (bspw. dem [[Programmbaum]]) ist auszuweisen, zur Erreichung welcher [[Resultate]] welche spezifischen Interventionen eingesetzt werden. Die [[Programmtheorie]] enthält darüber hinaus theoretisch begründete oder empirisch belegte Erklärungen dazu, weshalb von spezifischen Interventionen erwartet wird, dass sie bestimmte Resultate auslösen. Im [[Monitoring]] sollte zwischen [[Kennzahlen]] zu Interventionen und solchen zu anderen Aktivitäten unterschieden werden. (Synonym: Treatment) (Chen/Rossi 1992)
Kontext	context	Rahmenbedingung eines [[Programms]] ([[Kontext I]]) bzw. einer [[Evaluation]] ([[Kontext II]]). Wenn der Begriff Kontext in diesem Glossar allein steht, ist Kontext I gemeint.
Kosten-Wirkungs-Analyse	cost-effectiveness-analysis	Überprüfung, wie kostengünstig (Kosten im weitesten Sinne) ein [[Programm]] [[Outcomes]] und [[Impacts]] hervorbringt, also [[Wirtschaftlichkeit]] aufweist. (siehe auch [[programmkosten-nutzengesteuerte Evaluation]]) (McDavid/Hawthorn 2006, S. 252-263)
Kriterienpunkt	criterion point [standard]	Wendepunkt im Verlauf der Ausprägungen eines operationalisierten Kriteriums ([[Kriterium, operationalisiertes]]), an dem die [[Bewertung]] des [[Evaluationsgegenstands]] umschlägt (von negativ auf positiv oder umgekehrt). (siehe auch [[Erfolgspunkt]], [[K. O.-Kriterienpunkt]], [[Trumpf-Kriterienpunkt]]) (Fournier 1995, S. 16)

Kriterienzone	citerion zone	Festgelegter Bereich in Hinblick auf die Ausprägung eines operationalisierten Kriteriums ([[Kriterium, operationalisiertes]]), der zwischen zwei [[Kriterienpunkten]] liegt. Liegen [[Kriteriendaten]] vor, die sich in diesem Bereich befinden, fällt die [[Bewertung]] eines [[Evaluationsgegenstands]] positiv aus, sofern es sich um eine positive Kriterienzone ([[Kriterienzone, positive]]) handelt. Liegen die gemessenen Werte außerhalb dieser Zone, erfolgt eine negative Bewertung. (Entsprechend anders verhält es sich mit einer negativen Kriterienzone.) (Davidson 2005b)
Kriterium	criterion	Gesichtpunkt, auf welchen bei der [[Bewertung]] eines [[Evaluationsgegenstands]] explizit Bezug genommen wird. Etymologisch hergeleitet aus altgr. „krites“ (der Richter). Kriterium (altgr. Kriterion) meint Beurteilungsgesichtpunkt für den Richter, die Bezugsmaßstäbe, an denen bewertet wird. Nach Konkretisierungsgrad zu unterscheiden sind allgemeinere [[Kriteriendimensionen]] von operationalisierten Kriterien ([[Kriterien, operationalisierte]]). (Scriven 1959)
logisches Modell	logic model	Visuelle und/oder schriftliche Veranschaulichung der (Ablauf-)Logik eines [[Programms]]. Dies geschieht meist in Form einer Tabelle mit den klassischen Elementen als Spalten ([[Inputs]]/[[Ressourcen]], [[Aktivitäten]], [[Outputs]], [[Outcomes]], [[Impacts]]) und verdeutlicht die angenommenen Verbindungen zwischen diesen [[Programmelementen]]. Anders als bei einer [[Programmtheorie]] ist keine forschungsbasierte Begründung einer kausalen Verbindung zwischen den Elementen des Modells erforderlich. Das logische Modell dient Programmverantwortlichen sowie [[Evaluierenden]] als Strukturierungshilfe und Kommunikationsgrundlage über das Programm, zur Erstellung eines Plans zum [[Monitoring]] sowie zur Fokussierung einer [[Evaluation]]. Verbreitung fand das logische Modell in den USA durch die W. K. Kellogg Foundation. Auf einem logischen Modell basierte Evaluationen beantworten zumeist die Frage nach der [[Zielerreichung]]. (siehe auch [[Programmbaum]]). Obwohl die (vereinfachende) grafischen Datellung dies nahelegt, sind die in logischen Modellen abgebildeten Programme selten linear, sondern komplex und schleifenartig verknüpft ([[Kaskadenprogramm]], [[eingebettetes Programm]]). (Frechtling 2007; Wyatt Knowlton/Phillips 2009; WKKF 2001)
Matching	matching	Konstruktion einer [[Vergleichsgruppe]] oder [[Kontrollgruppe]] durch [[Auswahl]] von [[Aussageeinheiten]], die einer Experimentalhandlung ([[Programm]], [[Intervention]]) nicht unterworfen wurden und für die in der [[Experimentalgruppe]] Pendants („Zwillinge“) mit ähnlichen Merkmalsprofilen existieren. Matching-Verfahren werden u. a. im Rahmen von [[quasi-experimentaldesigngesteuerter Evaluation]] eingesetzt. (Rossi/Lipsey/Freeman 2004, S. 275-279)
Moderatorvariable	moderator variable	[[Variable]], welche die Stärke der Beziehung einer unabhängigen auf eine abhängige Variable beeinflusst, oft im [[Kontext I]] des [[Evaluationsgegenstands]] verortbar. (Bortz/Döring 2006, S. 3)
Outcomes (eines Programms)	outcomes (of a programme)	[[Programmelement]]: [[Intendierte Resultate]] eines [[Programms]] bei [[Zielgruppen]], wie z. B. Veränderungen bzw. Stabilisierungen im Wissen, den Einstellungen, sozialen Werten ([[Wert, sozialer]]) oder dem Können ([[Outcomes I]]), im Verhalten ([[Outcomes II]]) oder in der Lebenslage/dem Status der Zielpersonen ([[Outcomes III]]). Welche Outcomes angestrebt werden, soll in den [[Zielen]] des Programms festgehalten sein ([[Outcome-Ziele]]). (siehe auch [[logisches Modell]], [[Programmbaum]]) (Patton 2008, S. 240; Plantz/Greenway/Hendricks 1997)
Praktische Signifikanz	significance	[[Bedeutsamkeit]]
Programm	programme	Beschriebene und durchgeführte, intentional aufeinander bezogene Bündel von [[Aktivitäten]], [[Interventionen]], [[Maßnahmen]], Projekten oder Teilprogrammen. Ein Programm besteht aus einer Folge von auf ausgewiesene [[Ziele]] hin ausgerichteten [[Interventionen]]. Es wird auf der Basis von verfügbaren [[Ressourcen]] ([[Inputs]]) durchgeführt und ist darauf gerichtet, vermittels bereitgestellter Leistungen ([[Outputs]]) bestimmte Veränderungen/Stabilisierungen bei bezeichneten [[Zielgruppen]] ([[Outcomes]]) oder in sozialen Systemen ([[Impacts]]) auszulösen. [[Evaluationsgegenstand]] können sowohl das [[Konzept]] des Programms, als auch seine Umsetzung (Aktivitäten bzw. Interventionen) und seine [[Resultate]] sein. Je nach [[Evaluationsfeld]] hat das Wort „Programm“ eine andere Bedeutung – hier ist es ein Fachbegriff der Evaluationssprache. Programme unterscheiden sich u. a. in ihrer Größe, gemessen in eingesetzten oder umgesetzten Finanzmitteln, der Anzahl von [[Stakeholdern]] oder in ihrem Komplexitätsgrad. (siehe auch [[Kaskadenprogramm]], [[eingebettetes Programm]]) (Donaldson 2007)

Resultate (eines Programms)	results (of a programme)	[[Programmelement]]: Durch [[Aktivitäten]] bzw. [[Interventionen]] eines [[Programms]] bereitgestellte Leistungen oder Produkte ([[Outputs]]) und ausgelöste Veränderungen/Stabilisierungen bei [[Zielgruppen]] ([[Outcomes]]) oder bei Organisationen und anderen sozialen Aggregaten ([[Impacts]]). (siehe auch [[Wirkungen]], [[Effekte]], [[intendierte Resultate]], [[nicht-intendierte Resultate]], [[unerwünschte Resultate]]) (Kusek/Rist 2004; Beywl 2006a, S. 36-37)
Schlussfolgerungen (einer Evaluation)	conclusions (of an evaluation)	Verdichtende Art von beschreibenden [[Evaluationsergebnissen]] oder [[Informationen]] ([[Beschreibung]]), die eine [[Evaluation]] zur Verfügung stellt. Schlussfolgerungen formulieren meist auf die [[Evaluationsfragestellungen]] bezogene zusammenfassende Beschreibungen und [[Interpretationen]], die auf die erhobenen [[Daten]] rückführbar sind. Eine Evaluation muss Schlussfolgerungen und [[Bewertungen]], kann darüber hinaus [[Empfehlungen]] bereitstellen. (Fournier 1995)
Stichprobe	sample	[[Erhebungsmenge]], die vorzugsweise durch [[Zufallsauswahl]] gebildet wird („probabilistische Stichprobe“). (Bortz/Döring 2006, S. 396-479)
Tauglichkeit (eines Programms)	worth (of a programme)	Im Gegensatz zur [[Güte]] der [[Wert]] eines [[Evaluationsgegenstands]] im Sinne des Gebrauchswerts für anzugebende [[Zielgruppen]] in bestimmten Situationen und zu angegebenen Zeitpunkten. Wenn die Tauglichkeit eines [[Programms]] durch eine [[Evaluation]] belegt ist, heißt dies nicht automatisch, dass dasselbe Programm auch in anderen Verwendungssituationen funktionieren wird. Insbesondere [[Kontext I]], [[Struktur]] und [[Incomes]] müssen auf hinreichende Ähnlichkeit geprüft werden. Eine [[Übertragbarkeit]] der [[Evaluationsergebnisse]], z. B. auf andere Kontexte oder auf die Gesamtheit des Bundesgebietes, ist als eigenständiger Schritt erforderlich, da die Tauglichkeit je nach Kontext usw. dramatisch variieren kann. Tauglichkeit ist eine generische Kriteriendimension ([[Kriteriendimension, generische]]) für die [[Bewertung]] von Programmen. (Scriven 1991a, S. 383-383; Beywl 1988, S. 149)
Übertragbarkeit (von Evaluationsergebnissen)	transferability (of evaluation findings)	Gütemerkmal von [[Evaluationsergebnissen]], das zum Ausdruck bringt, in welchem Maße die in einer konkreten [[Evaluation]] gewonnenen Ergebnisse auch für [[Evaluationsgegenstände]] in anderen [[Kontexten I]] oder andere (ähnliche) Evaluationsgegenstände gelten. (In der Sozialforschung spricht man – enger gefasst – von externer Validität oder [[Generalisierbarkeit]].) (Guba/Lincoln 1989, S. 241-242)
Übertragfähigkeit (von Programmen)	portability (of programmes)	Eine [[Kriteriendimension]] für [[Programme]], zeitlich und logisch der Kriteriendimension [[Nachhaltigkeit]] vorgeschaltet, die angibt, ob und in welchem Umfang ein zu einem Zeitpunkt/an einem Ort/in einem [[Kontext I]] erfolgreiches Programm an anderen Orten/zu anderen Zeitpunkten/unter anderen Kontexten vergleichbar gute [[Resultate]] zu erreichen im Stande ist. Kriterien aus der Software-Evaluation können die Konkretisierung dieser Dimension anregen: z. B. Anpassungsfähigkeit an andere [[Bedingungen]], Fähigkeit zur Koexistenz mit anderen Programmen. (King 2008, S. 139; Gediga/Hamborg/Düntsch 2001)
Vergleichsgruppe	comparison group	Personengruppe, die im Rahmen einer [[quasi-experimentaldesigngesteuerten Evaluation]] mit der [[Experimentalgruppe]] verglichen wird, woraus sich Aussagen zur [[Wirksamkeit]] einer Experimentalhandlung ([[Programm]], [[Maßnahme]], [[Intervention]]) ableiten lassen sollen. Sie wird aus „statistischen Zwillingen“ der Mitglieder der Experimentalgruppe zusammengesetzt ([[Matching]]). Idealerweise unterscheidet sie sich von der Experimentalgruppe nur durch den Umstand, dass sie der Experimentalhandlung nicht unterzogen wird. Zu diesem Zweck müssen viele Einflussgrößen für beide Situationen gleich gehalten werden. (Fraenkel/Wallen 2008, S. 602)
Wert (eines Evaluationgegenstandes)	value (of an evaluand)	Summe der Eigenschaften eines [[Evaluationsgegenstands]], die zu der [[Bewertung]] führt, dass er mehr oder weniger gut oder schlecht ist. Dabei kann die über Zeit und Raum relativ stabile Wertdimension der [[Güte]] von der weniger stabilen der [[Tauglichkeit]] unterschieden werden, die zwischen den [[Stakeholdern]] meist stark umstritten ist. Eine dritte Wertdimension ist die [[Bedeutsamkeit]]. Die Bestimmung des Werts eines Evaluationsgegenstands, die [[Bewertung]], ist die zentrale und unverzichtbare Aufgabe jeder [[Evaluation]]. (Scriven 2007; Stake/Schwandt 2006)

Wirksamkeit (eines Programms)	(caused) effectiveness (of a programme)	Grad, zu dem ein [[Programm]] bestimmte [[Wirkungen]] auslöst, die in seinen [[Zielen]] als anzustrebend vorgegeben sind. Es ist erforderlich, die Wirksamkeit eines Programms auf Basis einer [[Wirkungsmodellierung]] zu plausibilisieren und/oder empirisch nachzuweisen ([[Wirkungsnachweis, empirischer]]). Voraussetzung für den Nachweis sind theoretisch und durch Forschung begründete [[Konzepte]] (z. B. lerntheoretisch basierte Curricula) in Kombination mit strengen [[Erhebungsdesigns]] (z. B. randomisierte Erhebungsdesigns mit [[Kontrollgruppen]]). Eine Bestimmung der Wirksamkeit setzt die Messung der [[Zielerreichung]] voraus. (Synonym: Effektivität) (Davidson 2005a)
Wirkungseinschätzung	effects assessment	Systematisches Verfahren, in dem Expertinnen und Experten oder [[Stakeholder]] im Umfeld eines [[Programms]] dessen [[Wirksamkeit]] auf Basis ihres Erfahrungswissens einschätzen. Streng genommen handelt es sich um subjektive Meinungen über [[Wirkungen]], die durch fachliche Vorlieben, persönliche Interessen, affektive Bezüge zum Programm oder andere Faktoren beeinflusst sein können, ohne dass dies gemäß [[Evaluationsplan]] kontrolliert wird. Wirkungseinschätzung ist somit die schwächste, manchmal jedoch einzig realisierbare Form der [[Wirkungsfeststellung]]. (Synonym: Wirkungsabschätzung) (Beywl 2006a)
Ziele (eines Programms)	programme goals; aims; objectives	In der Zukunft liegende, erwünschte Zustände, die durch ein [[Programm]] ausgelöst werden sollen. Es handelt sich um einfache bis hoch komplexe gedankliche Vorwegnahmen künftiger Situationen. Wenn sich die Agierenden diese Ziele zu Eigen machen (Selbstverpflichtung), können sie individuelles und kollektives Handeln darauf ausrichten und abstützen. Vielfach sind Ziele implizit, werden stillschweigend als geltend unterstellt ([[stillschweigendes Wissen]]). In der [[Evaluation]] ist es jedoch wünschenswert, dass für die [[Evaluationsgegenstände]] Ziele expliziert, d. h. verschriftlicht sind. Als Konkretionsstufen von Programmzielen werden [[Leitziele]], [[Mittlerziele]] und [[Detailziele]] (auch Praxis- oder [[Handlungsziele]]) unterschieden, die zusammengefügt und aufeinander abgestimmt ein [[Zielsystem]] bilden. Vielfach werden [[Kriterien]] zur Bewertung von Programmen aus Zielen abgeleitet. Eine Alternative hierzu stellt die [[zielfreie Evaluation]] dar. (Patton 2008, S. 231-243; Weiss 1998, S. 51-55; Reischmann 2006, S. 180-196)
Zielerreichung	goal attainment; goal achievement	Grad, zu dem ein [[Evaluationsgegenstand]] ([[Programm]], [[Maßnahme]] usw.) seine gesetzten [[Ziele erreicht. Zielerreichung ist eine verbreitete [[Kriteriendimension]], wobei ein Programm umso besser bewertet wird, je genauer und vollständiger es seine Ziele erreicht. Oft wird – fälschlicherweise – mit der Prüfung der Zielerreichung auch der [[Wirkungsnachweis]] als erbracht angesehen. Die Zielerreichungsüberprüfung beinhaltet jedoch nicht den Nachweis, dass die in den Zielen als erwünscht charakterisierten und festgestellten [[Resultate]] tatsächlich ursächlich durch [[Interventionen]] des Programms ausgelöst wurden. Die Zielerreichung kann ganz oder teilweise auf andere Faktoren (z. B. Veränderungen von [[Kontext I]]) zurückgehen. Die Messung der Zielerreichung ist jedoch notwendige Voraussetzung für die Überprüfung der [[Wirksamkeit]] des Programms. (Provus 1971)
Zielgruppe (eines Programms)	target group (of a programme)	Die Personengesamtheit, an die sich ein [[Programm]] richtet, d. h. bei der [[Outcomes]] ausgelöst werden sollen. (Synonyme: Leistungsempfangende, Teilnahmekunden) (Lazenbatt 2002, S. 263)
Zufallsauswahl	randomisation	[[Auswahl]] nach dem Zufallsprinzip, d. h. jedes Element der [[Grundgesamtheit]] hat statistisch die gleiche Chance, in die [[Erhebungsmenge]] aufgenommen zu werden. (Field 2009, S. 17-18)

Bezugsquellen des Glossars auf http://www.univation.org/index.php?class=Calimero_Webpage&id=9015