IMPACT UNIT
WIRKUNG UND EVALUATION IN DER WISSENSCHAFTSKOMMUNIKATION

HOW-TO-REIHE

Wisskomm evaluieren

3. Erhebungsdesign einer Evaluation planen

Die How-To-Reihe *Wisskomm evaluieren* der *Impact Unit* liefert Hinweise, Beispiele und Lektüretipps für die Planung und Umsetzung aussagekräftiger Evaluationen in der Wissenschaftskommunikation. Sie richtet sich an Praktiker*innen in diesem Feld, die einen Einstieg in das Thema Evaluation suchen. Diese How-Tos sollen vor allem als Orientierungshilfe verstanden werden und weniger als strenges Regelwerk. Auch wenn sie als Reihe konzipiert ist, können Leser*innen jederzeit direkt den Teil der How-To-Reihe hinzuziehen, der ihnen akut weiterhelfen kann.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

wissenschaft : im dialog

IMPACT UNIT

Übersicht

HOW-TO-REIHE

Mit dem Projekt *Impact Unit - Evaluation und Wirkung in der Wissenschaftskommunikation* möchte *Wissenschaft im Dialog* zu einer stärkeren Wirkungsorientierung sowie aussagekräftigen Evaluationspraxis in der Wissenschaftskommunikation beitragen und eine Grundlage für fundierte Diskussionen des Feldes legen. Hierfür beobachtet und analysiert sie die aktuelle Evaluationspraxis, entwickelt Evaluationstools und Hilfsmittel für Praktiker*innen und unterstützt den Austausch zwischen Praxis, Forschung und Förderung.

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 0150862 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor*innen.

Impressum

Wissenschaft im Dialog gGmbH
Projekt Impact Unit – Wirkung und Evaluation in der Wissenschaftskommunikation
Charlottenstraße 80
10117 Berlin

Tel.: 030/206 22 95-0
E-Mail: info@w-i-d.de

Konzeption und Redaktion
Ricarda Ziegler, Projektleitung Impact Unit
Imke Hedder, Projektmanagement Impact Unit

Weitere Informationen und Tools finden Sie auf
www.impactunit.de

Stand
April 2021

How-To: Erhebungsdesign einer Evaluation planen

Das Erhebungsdesign beschreibt den Aufbau der Evaluation, d. h. die Strategie, nach der die benötigten Daten erhoben werden, um das festgelegte Erkenntnisinteresse zu verfolgen. Mit der Bestimmung des Erhebungsdesigns werden folgende Fragen geklärt:

1. Wann und wie häufig werden Daten erhoben?
2. Werden zusätzliche Vergleichswerte (zur Einordnung der Daten) benötigt?
3. Werden alle verfügbaren Daten oder nur eine Auswahl untersucht?

1. Wann und wie häufig werden Daten erhoben?

Worum geht's? Die Entscheidung darüber, ob eine einmalige Erhebung von Daten ausreicht oder mehrere Datenerhebungen nötig sind, ergibt sich aus der Frage, ob **eine Entwicklung oder eine „Momentaufnahme"** festgehalten werden soll. Eine Momentaufnahme wäre beispielsweise sinnvoll, wenn deskriptive Fragestellungen verfolgt werden, etwa wenn herausgefunden werden soll, ob die Teilnehmenden mit einer Veranstaltung zufrieden sind. Eine Entwicklung sollte festgehalten werden, wenn Veränderungen durch das Projekt von Interesse sind. Dies ist der Fall, wenn beispielsweise Wirkungsziele evaluiert werden. Solche Entwicklungen lassen sich im Gegensatz zu Momentaufnahmen allerdings nicht in einer einmaligen Erhebung festhalten.

Drei häufige Typen für Erhebungsdesigns sind:

- die einmalige Erhebung, die häufig im Anschluss an die Wissenschaftskommunikation durchgeführt wird
- der Vorher-Nachher-Vergleich, also eine Erhebung vor und nach der Wissenschaftskommunikation, um mögliche Wirkungen zu untersuchen
- die Langzeiterhebung, bei der mehrfache Erhebungen nach der Wissenschaftskommunikation in zeitlichen Abständen durchgeführt werden, um die Langfristigkeit von Wirkungen oder später einsetzende Wirkungen zu untersuchen

Mit der Frage nach der Anzahl der notwendigen Datenerhebungen geht auch die Frage nach dem richtigen **Zeitpunkt der jeweiligen Erhebung** einher. Diese hängt vorrangig davon ab, wann die gewünschten Informationen verfügbar sind (Beispiel: Können sich die Personen zu diesem Zeitpunkt bereits eine abschließende, reflektierte Meinung zum behandelten Thema gebildet haben?) und zuverlässig erscheinen (Beispiel: Können sich die Personen zu diesem Zeitpunkt noch gut an ihre spontane Reaktion erinnern?). Im Fall von Befragungen sollte auch bedacht werden, wann Befragte am ehesten dafür bereit wären.

Warum? In diesem fundamentalen Schritt entscheidet sich, **ob die praktischen Umstände der Evaluation der Evaluationsfrage gerecht werden** können. Wird eine einmalige Erhebung durchgeführt, obwohl eigentlich mehrfache Erhebungen nötig wären, dann kann die Evaluationsfrage auf Basis der vorliegenden Ergebnisse nicht beantwortet werden. Gleichzeitig muss durchdacht werden, ob sich überhaupt zu mehreren Zeitpunkten sinnvolle Gelegenheiten zur Datenerhebung ergeben und welchen zeitlichen und finanziellen Aufwand das geplante Erhebungsdesign erfordern würde. Damit ist auch die **Ressourcenplanung** für diese Entscheidung zentral.

Wie? Jedes Erhebungsdesign bringt bestimmte Bedingungen mit:

- ☐ Eine einmalige Erhebung sollte eingesetzt werden, wenn durch die Daten **deskriptive Fragestellungen** beantwortet werden.

- ☐ Ein Vorher-Nachher-Vergleich sollte eingesetzt werden, wenn die **Wirkung** des Projekts betrachtet werden soll.

→ Bedingung: Das Projekt ermöglicht eine **Vorab-Datenerhebung** – davon ausgeschlossen wären etwa Aktionen, die Personen überraschen sollen

- ☐ Eine Langzeiterhebung sollte eingesetzt werden, wenn **Entwicklungen im Fokus** stehen, die **nicht direkt im Anschluss an das Projekt deutlich werden** (häufig der Fall bei gewünschten Verhaltensänderungen) und/oder wenn die **Nachhaltigkeit einer Wirkung** untersucht werden soll.

→ Bedingung: Der Rahmen des Projekts lässt Erhebungen über einen längeren Zeitraum zu.

Sollte die verfolgte Fragestellung mehrere Datenerhebungen zu unterschiedlichen Zeitpunkten verlangen, aber die Rahmenbedingungen dafür nicht erfüllt sein, empfiehlt sich eine Anpassung der Fragestellung. In einzelnen Fällen kann der Einbezug zusätzlicher Vergleichswerte dabei helfen, Vermutungen über Wirkungen anzustellen, etwa durch Kontrollgruppen (siehe Punkt 2). Allerdings ist das keine zuverlässige „Kompensation" für einen Vorher-Nachher-Vergleich, weil diese externen Vergleichspunkte selten eins-zu-eins mit den Umständen und den erreichten Personen des eigenen Projekts vergleichbar sind.

Praxisbeispiel: Mehrfache Erhebungen

Im Kooperationsprojekt ***Wissenschaft debattieren!*** von *Wissenschaft im Dialog* und dem „ZIRN" (heute „ZIRIUS") der Universität Stuttgart werden sieben verschiedene Dialog- und Beteiligungsformate auf ihre Wirksamkeit untersucht und miteinander verglichen. Es wird ein Methodenmix gewählt. Hierzu gehört auch eine standardisierte Befragung, die Informationen der Teilnehmenden mittels eines Fragebogens einheitlich und zählbar erfragt, sodass Vergleiche zwischen den Teilnehmenden und zwischen den verschiedenen Erhebungszeitpunkten möglich sind. Eine Mehrfacherhebung wird dabei für jene Formate gewählt, **die besonders viele Teilnehmende aufweisen** und damit eine quantitative Auswertung auch bei sinkender Bereitschaft zur Evaluationsteilnahme mit jeder Erhebung gewährleisten.

Zwei bis drei Wochen vor der Veranstaltung wird den Teilnehmenden ein Fragebogen zugeschickt, der die Motivation zur Teilnahme, das Interesse am behandelten Thema in der Veranstaltung sowie an Wissenschaft allgemein erfragt. Zudem werden die Einstellungen und Bewertungen der Teilnehmenden zu Themen im Bereich Wissenschaft und Technik erfasst, zu Bürgerbeteiligungsformaten sowie (sofern das Veranstaltungsthema dies anbietet) bestimmte an das Thema angelehnte Verhaltensweisen (z. B. umweltfreundliches Alltagshandeln) und zuletzt Personenangaben.

Im Anschluss an die Veranstaltung können neue Informationen gesammelt werden, so etwa die Bewertung der Veranstaltung und die Nutzung der dort bereitgestellten Informationsangebote. Daneben werden die gleichen Fragen hinsichtlich des Interesses, der Einstellungen und Bewertungen der Themenbereiche sowie des Verhaltens gestellt, um einen direkten Vergleich zur Vorbefragung zu ermöglichen.

In einer letzten Erhebung **einige Monate später** werden die Fragenblöcke zu Interesse, Einstellung, Bewertung und Verhalten nochmals wiederholt, um mögliche „Nachwirkungen" zu erfassen. Es werden aber auch bewusst rückblickende Fragen auf die Veranstaltung gestellt, die mehr Reflexion erfordern: etwa, ob das Veranstaltungsformat im Nachhinein zweckdienlich für die Meinungsbildung der Teilnehmenden war.

Außerdem wird der Transfer des Erlebnisses in den Alltag erfragt, zum Beispiel mögliche Änderungen in der Mediennutzung seit der Veranstaltung, sowie anschließende Gespräche über die Veranstaltung.

Auch wenn die einzelnen Erhebungen zu großen Teilen den Zweck erfüllen, die Entwicklungen von Interessen, Einstellungen und Ähnlichem zu erfassen, **wird mit jedem Fragebogen auch die Möglichkeit ergriffen, weitere Informationen einzuholen**: Die erste Runde bietet die Chance, eine unvoreingenommene Meinung einzuholen; die zweite Runde bietet sich an, um spontane Reaktionen einzuholen, die zu einem späteren Zeitpunkt möglicherweise in Vergessenheit geraten; in der dritten Runde können Fragen gestellt werden, die zeitliche Distanz zur Veranstaltung verlangen. Hier geht es zum Praxisbeispiel.

2. Werden zusätzliche Vergleichswerte (zur Einordnung der Daten) benötigt?

Worum geht's? Im Fall von Mehrfacherhebungen wird eine Datenquelle (z. B. eine befragte Person) zu verschiedenen Zeitpunkten mit sich selbst verglichen. Daneben können aber auch **außenstehende Personen und Datenquellen zum Vergleich** herangezogen werden. Eine Option ist das Hinzuziehen einer Kontrollgruppe: Personen, die gar nicht an dem Projekt teilgenommen haben, aber dieselben Fragen gestellt bekommen. Eine weitere Alternative ist der externe Gruppenvergleich: Hier werden Daten herangezogen, die im Rahmen eines anderen Projekts unter einer ähnlichen Fragestellung generiert wurden.

Warum? Kontrollgruppen und externe Gruppenvergleiche können **Vergleichspunkte für die eigenen Ergebnisse** liefern, insbesondere bei der Einordnung, ob die eigenen Ergebnisse als ungewöhnlich oder üblich interpretiert werden sollten. Vergleichspunkte können die Aussagekraft der **Ergebnisse entweder absichern** oder als Warnung gelten, falls sie den Aussagen der eigenen **Ergebnissen widersprechen**. In jedem Fall sind diese Informationen hilfreich für die Interpretation der Daten und die Reflexion der Evaluation insgesamt.

Wie? Beim Einbezug zusätzlicher Vergleichswerte sollte Folgendes beachtet werden:

- ☐ **Strukturelle Unterschiede** zwischen den eigenen Daten und den Daten der zusätzlichen Vergleichswerte werden bestmöglichst vermieden.
  - ☐ Die Zuordnung von Personen zur Kontrollgruppe oder zur teilnehmenden Gruppe erfolgt idealerweise **zufällig** und ist nicht durch Faktoren wie beispielsweise Motivation, Zugänglichkeit, Alter o. Ä. bestimmt.
  - ☐ Beim Hinzuziehen eines externen Gruppenvergleichs wird ein kritischer Blick auf die **Zusammensetzung der Gruppe** geworfen, um ihre Merkmalsverteilung mit der selbst untersuchten Gruppe zu vergleichen. Mögliche Unterschiede, z. B. in den soziodemografischen Angaben, werden bei der Interpretation beachtet.
- ☐ Beim externen Gruppenvergleich werden die Aktivitäten des fremden Projekts sowie die **Fragestellungen und Erhebungsbedingungen sorgfältig mit dem eigenen Projekt verglichen**, mögliche Unterschiede transparent kommuniziert und in der Interpretation beachtet.
- ☐ Es wird besprochen, wie die Einteilung von Kontrollgruppe und teilnehmender Gruppe **die Evaluation beeinträchtigen könnte**. Für manche Projekte lässt sich so ein Kontrollgruppen-Design leichter umsetzen als für andere, wie das folgende Fallbeispiel veranschaulicht.

Praxisbeispiel: Fallstricke eines Kontrollgruppen-Designs

Im Rahmen eines Dissertationsprojekts wird ein Chemielabor für Schüler*innen durch mehrere Evaluationsphasen in einem Zeitraum von fünf Jahren begleitet. Während einige Klassen das Chemielabor testen, lernen die Kontrollklassen die Inhalte im Schulunterricht. Um sicherzugehen, dass die Unterschiede zwischen den Gruppen wirklich durch das Chemielabor hervorgerufen werden (und nicht mit grundsätzlichen Leistungsunterschieden zwischen den Klassen zusammenhängen), soll in einer Evaluationsrunde die Einteilung der Gruppen zufällig erfolgen: Per Los werden die Schüler*innen einer Klasse entweder dem Frontalunterricht oder dem Chemielabor zugeteilt. Die Folge: **Schüler*innen, die in der Kontrollgruppe landen und am klassischen Frontalunterricht teilnehmen müssen, beteiligen sich nur widerwillig am Unterricht. Protest kommt auch von Lehrkräften und Eltern**, denn erstere sind nicht zufrieden damit, halbe Klassen unterrichten zu müssen und letztere äußern ihre Sorge, dass ihre Kinder in der Kontrollgruppe Nachteile erfahren. Da unter diesen Umständen die Motivation der Teilnehmenden als unzureichend eingeordnet wird, um zu zuverlässigen Ergebnissen zu kommen, wird das Experiment abgebrochen.

Auch wenn ein Kontrollgruppen-Design aus statistischen Gründen befürwortet wird, lassen sich solche nicht-intendierte Wirkungen nicht immer vermeiden und sollten bedacht werden. Wichtig ist, dass Komplikationen transparent kommuniziert und die folgerichtigen Konsequenzen gezogen werden, wie in diesem Praxisbeispiel.

Wo kann ich mich weiter informieren?

- Für eine ausführliche Erklärung verschiedener Untersuchungsdesigns, inklusive Beispielen: Nicola Döring & Jürgen Bortz (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften (Kapitel 7)

Praxisbeispiele: Gruppenvergleiche heranziehen

Das Wiener Kindermuseum ZOOM setzt das Science-Awareness-Projekt **Draw a Scientist – Wer macht Wissenschaft?** um. Ziel des Projekts ist es, die Vorstellung von Grundschulkindern über Wissenschaftler*innen zu erfassen, indem sie zunächst Bilder von Forschenden und ihrer Arbeit malen. Anschließend werden die Kinder dazu angeregt, ihre Vorstellung zu reflektieren und werden über Forschende und ihre Arbeit abseits stereotyper Bilder aufgeklärt.

Das Projekt wird gemeinsam mit der türkischen Partnerorganisation Çocuk Istanbul konzipiert, beide Projekte werden getrennt voneinander wissenschaftlich begleitet. **Durch die gemeinsame Vorbereitung wird die Vergleichbarkeit der Projektergebnisse sichergestellt:** So wurde vorab entschieden, dass beide Projekte die gleichen Erhebungsmethoden nutzen, namentlich die Zeichenreflexionsbögen zur Auswertung der Zeichnungen und die Feedbackbögen für die Kinder und ihre Begleitpersonen. **Es wird allerdings auch deutlich gemacht, an welchen Stellen keine zuverlässigen Vergleichswerte bestehen.** Hierbei wird offen kommuniziert, an welchen Stellen sich die Evaluationsmethoden unterscheiden (in Wien finden qualitative Erhebungen statt, für die es kein Pendant in der Partnerevaluation gibt) und dass aufgrund der unterschiedlichen Sprachen nur quantitative Daten der geschlossenen Fragen aus den Feedbackbögen verglichen werden. Hier geht es zum Praxisbeispiel.

Auch im Forschungsprojekt ***Lernmotivation im Technikunterricht*** der Universität Stuttgart werden weitere Vergleichspunkte herangezogen. Ziel des Projekts ist es, ein Techniklabor als Ergänzung des Gymnasialunterrichts zu entwickeln und zu evaluieren. Hierfür werden neben qualitativen Methoden auch quantitative Befragungen der Schüler*innen in mehreren Erhebungen durchgeführt und zusätzlich Kontrollgruppen untersucht. **Als Kontrollgruppen gelten Klassen der gleichen Schule, die allerdings noch keinen Unterricht im Techniklabor hatten, aber auch Schüler*innen weiterer Schulen ohne ein Techniklabor.** Bei der ersten Erhebung aller Schüler*innen werden strukturelle Unterschiede zwischen Kontrollgruppen und der am Techniklabor teilnehmenden Gruppe untersucht und offen thematisiert; beispielsweise, dass im Fall der Schüler*innen von der teilnehmenden Schule geringfügig häufiger eines der Elternteile einen technischen Beruf ausübt als an einer der Kontrollschulen. Erhoben wurden u. a. die Einstellung zu, das Interesse an und der Umgang mit Technik. **Um die Vergleichbarkeit zu ähnlichen Projekten zu schaffen, werden bereits verwendete Fragen aus vorangegangenen Evaluationsstudien verwandter Modellprojekte zur Motivation des Techniknachwuchs eingesetzt und des Weiteren Fragen des Nachwuchsbarometers Technikwissenschaften herangezogen.** Die Zweitverwertung dieser Instrumente hat auch den Vorteil, dass sie bereits in der Praxis getestet wurden. Hier geht es zum Praxisbeispiel.

3. Werden alle verfügbaren Daten oder nur eine Auswahl untersucht?

Worum geht's? Neben der zeitlichen Abfolge der Erhebung und möglichen externen Vergleichspunkten stellt sich noch die Frage nach dem Umfang der selbst erhobenen Daten: **Werden alle verfügbaren Datenquellen (auch Grundgesamtheit genannt) einbezogen, handelt es sich um eine Vollerhebung. Wird eine Stichprobe von dieser Grundgesamtheit evaluiert, spricht man von einer Teilerhebung**. Eine Vollerhebung ist von Vorteil, weil jede Quelle und jede Perspektive in die Evaluation einfließt und in der abschließenden Reflexion des Projekts berücksichtigt werden kann. Bei Projekten mit einer überschaubaren Größe ist es in jedem Fall empfohlen, alle Datenquellen einzubeziehen. Sollte die Zahl der Medienberichte, der Teilnehmenden, der dokumentierten Nutzer*innen o. ä. Quellen, die evaluiert werden sollen, zu umfangreich für eine Vollerhebung sein, kann eine Stichprobe bestimmt werden. Diese Lösung ist auch sinnvoll, wenn das Personal in der Erhebungsphase begrenzt ist (z. B., wenn sämtliche Museumsbesucher*innen eines Tages von wenigen Evaluierenden vor Ort befragt werden müssten). **Im Fall einer Teilerhebung muss bei der Auswertung und Interpretation der Evaluation stets reflektiert werden, ob diese Stichprobe ein gutes Abbild der Grundgesamtheit darstellt** oder ob es dazu gekommen sein könnte, dass bestimmte Perspektiven über- oder unterrepräsentiert sind. (Hier spricht man auch von der „Repräsentativität" einer Stichprobe.) Das Risiko für so eine Verzerrung kann statistisch ermittelt werden und hängt auch eng mit der Art und Weise zusammen, wie die Stichprobe zusammengestellt wurde. Hierbei spricht man von der **Stichprobenziehung**.
Die gröbste und wichtigste Unterscheidung ist jene zwischen ***zufälligen* und *nicht zufälligen* Ziehungen**. Eine zufällige Stichprobenziehung wäre beispielsweise das Lotterieprinzip, bei dem jede Quelle im Lostopf landet und die gleiche Chance hat, Teil der Stichprobe zu werden. Eine nicht zufällige Variante wäre etwa eine willkürliche Auswahl, bei der einfach die am leichtesten zugänglichen Quellen herangezogen werden, ohne allen die gleiche Chance zu geben. Damit sind nicht zufällige Varianten statistisch gesehen anfälliger dafür, bestimmte Perspektiven zu überrepräsentieren, als zufällige Varianten.

Kurzbeispiel: Zufällige und nicht zufällige Stichprobenziehung

Im Rahmen einer Veranstaltung sollen kurze Interviews mit Besucher*innen stattfinden. Vor dem Ausgang stehen die Evaluierenden bereit, um vereinzelte Personen abzufangen und zu befragen. Hierbei kann es vorkommen, dass **Evaluierende intuitiv diejenigen abfangen, die sie anlächeln, die ihnen selbst sympathisch vorkommen oder die bei ihnen den Eindruck erwecken, als wären sie offen für eine Teilnahme an der Evaluation. Hierbei handelt es sich um ein nicht zufälliges Verfahren,** bei dem bestimmte Merkmale der Person über ihren Einbezug in die Stichprobe entscheiden. Die Folge einer solchen „Gelegenheitsstichprobe" kann sein, dass nur solche Personen an der Evaluation teilnehmen, die zufrieden mit der Veranstaltung waren, da die unzufriedenen Besucher*innen den Evaluierenden entgangen sind. Damit würde die Evaluation ein verzerrtes Bild der Bewertung ihrer Veranstaltung widerspiegeln. **Für ein zufälliges Verfahren müssten die Evaluierenden hierbei nach einem konsequenten Schema vorgehen: Zum Beispiel wird jede fünfte Person gefragt**, ob sie an der Evaluation teilnehmen will – unabhängig davon, ob sie fröhlich wirkt oder danach aussieht, als wäre sie an einem Gespräch interessiert. Dieses „Losverfahren" würde angewendet werden, bis alle Besucher*innen die Veranstaltung verlassen haben.

Neben der Stichprobenziehung spielt natürlich auch die **Größe der Stichprobe** eine Rolle dabei, inwieweit das Evaluationsergebnis insgesamt durch einzelne Befragte verzerrt werden kann. Eine besonders schlechte Bewertung unter zehn ansonsten moderat ausfallenden Feedbackbögen fällt stärker ins Gewicht als eine schlechte Bewertung im Rahmen von dreißig moderat ausfallenden Feedbackbögen. Insbesondere wenn Durchschnittswerte berechnet werden, ist daher die Größe der Stichprobe zu beachten. Die „angemessene" Stichprobengröße, die für fundierte Ergebnisse bei quantitativen Auswertungen und statistischen Berechnungen empfohlen wird, hängt immer von der **Art der Berechnung** ab (z. B. ob und wie viele Gruppen innerhalb der Stichprobe verglichen werden sollen, etwa auf Basis ihres Alters, Bildungshintergrunds o. ä.) und richtet sich nach der **Größe der Grundgesamtheit**. Online lassen sich nähere Erläuterungen zur Berechnung dieser Idealgröße sowie automatische Stichprobengrößen-Rechner finden. Als Faustregel empfiehlt sich, dass Ergebnisse zu Vergleichen zwischen Gruppen mit Vorsicht genossen werden sollten, wenn sie auf weniger als dreißig Quellen pro Gruppe basieren.

Kurzbeispiel: Gruppenvergleiche anstellen

Im Rahmen eines Wissenschaftsfestivals sollen Wirkungen bei den Besucher*innen evaluiert werden. Deshalb werden Personen gesucht, die an einer Vorher-Nachher-Befragung teilnehmen möchten. Die Evaluierenden wollen auch herausfinden, ob das Familienprogramm sowohl für Erwachsene als auch für ihre Kinder interessant und lehrreich ist. Um überzeugende Aussagen über Unterschiede zwischen den beiden Gruppen treffen zu können (z. B. über mögliche Unterschiede hinsichtlich der Lerneffekte bei Kindern und Erwachsenen) sollten am besten a) mindestens dreißig Kinder und dreißig Erwachsene evaluiert werden und b) diese Personen über ein zufälliges Verfahren „rekrutiert" werden (siehe hierzu die vorherige Beispielbox).

Warum? Die Entscheidung für eine Vollerhebung oder eine Teilerhebung ist wichtig für die **Ressourcenplanung und die Reflexion der Evaluation**. Eine Vollerhebung mag in der Erhebungsphase und der Auswertung mehr Zeit und Personal verlangen, auf der anderen Seite kann auch die Planung und Umsetzung einer aufwändigen Stichprobenziehung Zeit kosten. Ist eine Stichprobe notwendig, dann sollte eine bewusste Auseinandersetzung mit den Konsequenzen der möglichen Varianten zur Stichprobenziehung erfolgen, damit mögliche Einschränkungen bei der Interpretation ihrer Ergebnisse bekannt sind. Das ist wiederum wichtig für die angemessene **Verwertung und Kommunikation der Evaluationsergebnisse**, wie etwa bei der Formulierung von **Handlungsempfehlungen**.

Wie? Bei der Entscheidung, welche Daten einbezogen werden, gilt es folgende Punkte zu beachten:

- ☐ Wenn die Zahl der **Datenquellen überschaubar** ist, diese verfügbar sind und die zeitlichen und finanziellen Ressourcen es zulassen, wird eine Vollerhebung durchgeführt.
- ☐ Wenn die **Zahl der Quellen zu umfassend** ist, wird idealerweise eine zufällige Stichprobenziehung gewählt.

→ Bedingung: Alle Datenquellen haben **die gleiche Chance, in der Stichprobe zu landen**.

- ☐ Wenn **keine zufällige Stichprobenziehung** möglich ist, wird eine nicht zufällige Stichprobenziehung vorgenommen.

→ Folge: Es wird reflektiert, wie die Ergebnisse verwertet werden sollen und **ob es gewünscht ist, dass die Stichprobe repräsentativ ist, also ein gutes Abbild der Grundgesamtheit darstellt**. Ist dies der Fall, gibt es weitere, mehrstufige Varianten der Stichprobenziehung, die allerdings mit Aufwand und ggf. Kosten verbunden sind.

- ☐ Um überzeugende Aussagen zu gewährleisten, wird eine **angemessene Stichprobengröße** angestrebt, die **im sinnvollen Verhältnis zur Grundgesamtheit** steht. Im Mindesten werden bei geplanten Gruppenvergleichen dreißig Quellen pro Gruppe anvisiert.

→ Ist dies nicht möglich, müssen die Ergebnisse mit Vorsicht genossen und mögliche Verzerrungen im Hinterkopf behalten werden.

Wo kann ich mich weiter informieren?

- Ein Methodenbuch, das umfangreiche Informationen zur Datenauswahl liefert: Schnell, Hill & Esser (2018): Methoden der empirischen Sozialforschung (ab S. 239)
- Für weitere Informationen zu Repräsentativität und Methoden der Stichprobenziehung: Schnapp & Block (2019): Auswahl von Untersuchungsobjekten
- Für eine Kurzvorstellung weiterer Verfahren zur Stichprobenziehung: Hochschule Luzern (2021): Stichprobenziehung
- Für nähere Erklärungen und Tipps zur Ermittlung einer Stichprobe im Rahmen groß angelegter quantitativer Befragungen, inklusive Stichprobenrechner zur Bestimmung der angemessenen Größe: Survey Monkey (2021): Repräsentative Stichprobe berechnen: Formeln, Beispiele und Tipps