WISSENSCHAFTLICHE DISKUSSIONSPAPIERE

Eva Quiring | Andreas Stöhr | Gunda Görmar

Berufsübergreifendes Konzept zur Evaluation von Ausbildungsordnungen

Bundesinstitut für Berufsbildung BIBB

Forschen
Beraten
Zukunft gestalten

WISSENSCHAFTLICHE DISKUSSIONSPAPIERE

Heft 172

Eva Quiring | Andreas Stöhr | Gunda Görmar

Berufsübergreifendes Konzept zur Evaluation von Ausbildungsordnungen

Bundesinstitut für Berufsbildung BIBB

Forschen
Beraten
Zukunft gestalten

Die WISSENSCHAFTLICHEN DISKUSSIONSPAPIERE des Bundesinstituts für Berufsbildung (BIBB) werden durch den Präsidenten herausgegeben. Sie erscheinen als Namensbeiträge ihrer Verfasser und geben deren Meinung und nicht unbedingt die des Herausgebers wieder. Sie sind urheberrechtlich geschützt. Ihre Veröffentlichung dient der Diskussion mit der Fachöffentlichkeit.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

© 2016 by Bundesinstitut für Berufsbildung, Bonn

Herausgeber: Bundesinstitut für Berufsbildung, Bonn
Umschlaggestaltung: CD Werbeagentur Troisdorf
Satz: W. Bertelsmann Verlag GmbH & Co. KG
Herstellung: Bundesinstitut für Berufsbildung, Bonn

Bestell-Nr.: 14.172

Bundesinstitut für Berufsbildung Arbeitsbereich 1.4 –
Publikationsmanagement/Bibliothek
Robert-Schuman-Platz 3
53175 Bonn

Internet: www.bibb.de
E-Mail: zentrale@bibb.de

ISBN 978-3-945981-29-0

CC-Lizenz
Der Inhalt dieses Werkes steht unter einer Creative-Commons-Lizenz (Lizenztyp: Namensnennung – Keine kommerzielle Nutzung – Keine Bearbeitung – 4.0 Deutschland).

Weitere Informationen finden Sie im Internet auf unserer Creative-Commons-Infoseite www.bibb.de/cc-lizenz.

Diese Netzpublikation wurde bei der Deutschen Nationalbibliothek angemeldet und archiviert: urn:ubn:de:0035-0593-1

Internet: www.bibb.de/veroeffentlichungen

Inhaltsverzeichnis

Abkürzungsverzeichnis . . . 5

Einleitung . . . 6

1 Definitionen von Evaluation . . . 7

2 Vorüberlegungen und theoretische Grundlagen . . . 9
2.1 Ziel und Zweck von Evaluationen in der Ordnungsarbeit . . . 9
2.2 Die Ausbildungsordnung als Evaluationsgegenstand . . . 11
2.2.1 Die Dimension „Ordnungsrahmen“ . . . 13
2.2.2 Die Dimension „Kontext“ . . . 14
2.2.3 Die Dimension „Weisungsgeber, Weisungsnehmer, Sozialpartner“ . . . 14
2.2.4 Die Dimension „Zielgruppen“ . . . 15
2.3 Die Forschungsfragen . . . 16
2.4 Akteure während der Evaluation von Ausbildungsordnungen . . . 18
2.5 Die Unabhängigkeit zum Evaluationsgegenstand . . . 20
2.5.1 Selbstevaluation . . . 21
2.5.2 Fremdevaluation . . . 23
2.5.3 Rahmenbedingungen für Evaluationen . . . 23
2.6 Methodische Aspekte von Evaluationen . . . 23
2.6.1 Die Triangulation von Methoden . . . 24
2.6.2 Die Triangulation von Daten . . . 25
2.7 Berichtswesen . . . 26
2.8 Mögliche Evaluationsansätze . . . 26
2.8.1 Der managementorientierte Ansatz . . . 27
2.8.2 Der zielorientierte Ansatz . . . 28
2.8.3 Der partizipative Ansatz . . . 29
2.8.4 Der konsumentenorientierte und der expertenorientierte Ansatz . . . 30

3 Leitlinien für die Evaluation von Ordnungsmitteln . . . 31
3.1 Evaluationsstandards . . . 31
3.2 Die Zusammensetzung des Evaluationsteams . . . 32
3.3 Perspektivenvielfalt bei der Analyse des Evaluationsgegenstands . . . 33
3.4 Die Funktionen einer Evaluation im Ordnungsbereich . . . 34
3.5 Planung und Durchführung der Evaluation . . . 34

4 Arbeitshilfen zur Umsetzung der Evaluationsleitlinien . . . 35

Literatur . . . 36

Anhang Arbeitshilfen 39
Arbeitshilfe: Qualitätskriterien, Gütekriterien und Untersuchungsdesign 40
Arbeitshilfe: Stichprobenauswahl 49
Arbeitshilfe: Erhebung quantitativer Daten 61
Arbeitshilfe: Skalenniveaus und Auswertung quantitativer Daten 73
Arbeitshilfe: Erhebung qualitativer Daten 105

Abstract 118

Abbildungen

Grafik 1: Ablauf einer Evaluation 7
Grafik 2: Kennzeichen wissenschaftlich durchgeführter Evaluationen 8
Grafik 3: Ziel/Zweck der Evaluation 10
Grafik 4: Evaluationszyklus 11
Grafik 5: Wirkungsmodell 13
Grafik 6: Kernelemente des Ordnungsrahmens 13
Grafik 7: Organisationsmodell 19
Grafik 8: Unabhängigkeitsgrad zum Evaluationsgegenstand 21
Grafik 9: Aspekte der Evaluationsmethodik 24
Grafik 10: Evaluationskonzepte 25
Grafik 11: Evaluationsansätze 26
Grafik 12: Stärken/Schwächen des managementorientierten Ansatzes 27
Grafik 13: Arbeitsschritte einer zielorientierten Evaluation 28
Grafik 14: Stärken/Schwächen des zielorientierten Ansatzes 29
Grafik 15: Stärken/Schwächen des partizipativen Ansatzes 30
Grafik 16: Leitlinien für die Evaluation von Ordnungsmitteln 31
Grafik 17: Dimensionen der Methodenkompetenz 33
Grafik 18: Funktionen der Evaluation 34

Tabellen

Tab. 1: Typische Forschungsfragen 17

Abkürzungsverzeichnis

AEA	American Evaluation Association
BIBB	Bundesinstitut für Berufsbildung
BMBF	Bundesministerium für Bildung und Forschung
BMWi	Bundesministerium für Wirtschaft und Technologie
CEval	Centrum für Evaluation
DeGEval	Deutsche Gesellschaft für Evaluation e. V.
DGB	Deutscher Gewerkschaftsbund
GAP	gestreckte Abschlussprüfung
KMK	Ständige Konferenz der Kultusminister der Länder
SEVAL	Schweizerische Evaluationsgesellschaft

Einleitung

Die Evaluation von Ausbildungsordnungen ist nicht nur ein wichtiger Bestandteil der Qualitätssicherung beruflicher Bildung, sondern spielt auch bei der Modernisierung von Berufen eine bedeutende Rolle. Herangehensweisen, die bei der Evaluation von Programmen, Projekten, Prozessen oder Organisationen häufig gewählt werden, lassen sich nicht automatisch auf die Evaluation von Ausbildungsordnungen übertragen. Deshalb wurde ein berufsübergreifendes Konzept zur Evaluation von Ausbildungsordnungen entwickelt, das speziell auf diesen Evaluationsgegenstand angepasst ist.

Dieses berufsübergreifende Konzept soll einen Beitrag dazu leisten, die Qualität von Evaluationen im Nachgang zur Ordnungsarbeit zu optimieren. Um dieses umfangreiche Thema adäquat beschreiben und analysieren zu können und um angemessene Empfehlungen für eine Weiterentwicklung von Evaluationen im Ordnungsbereich abgeben zu können, wurde das Konzept in insgesamt drei Abschnitte unterteilt.

Im ersten Teil des Konzepts (Kapitel 1 und 2) werden **Vorüberlegungen** angestellt und theoretische Grundlagen gelegt, um die Evaluationspraxis des Ordnungsbereichs innerhalb der Evaluationslandschaft verorten zu können. Die zentralen Fragen lauten u. a.: *Warum wird evaluiert? Wer bzw. was wird evaluiert? Welche Aspekte eines Gegenstands stehen im Mittelpunkt der Betrachtung? Wo ist die Evaluation angesiedelt? Wie wird evaluiert? Wann wird evaluiert?*

Der zweite Teil des Konzepts (Kapitel 3) skizziert darauf aufbauend für zukünftige Evaluationen im Nachgang zur Ordnungsarbeit sogenannte **Evaluationsleitlinien**. Diese Leitlinien umreißen sowohl für die Evaluatorinnen und Evaluatoren als auch für die an der Weisung beteiligten Akteure einen Handlungsrahmen. Darin werden Fragen hinsichtlich der zu berücksichtigenden *Evaluationsstandards*, der *Zusammensetzung des Evaluationsteams*, der *Sicherstellung einer Perspektivenvielfalt bei der Analyse des Evaluationsgegenstands*, der *Funktionen einer Evaluation* sowie schließlich hinsichtlich ihrer empfohlenen *Planung und Durchführung* beantwortet.

Im dritten Teil des Konzepts (Kapitel 4) finden sich Kurzbeschreibungen der sogenannten **Arbeitshilfen**. Die ausführlichen Arbeitshilfen im Anhang zum Konzept decken ein breites Feld ab und sollen Anregungen zur Behandlung unterschiedlicher Problemstellungen liefern, aber auch dabei helfen, Übersicht über formale Abläufe zu erlangen.

1 Definitionen von Evaluation

Der Begriff Evaluation wird in der Literatur nicht einheitlich definiert. Zudem wird der Begriff Stockmann (2010: 9) zufolge geradezu inflationär verwendet, und in vielen Kontexten benutzt, in denen er zumindest von seiner wissenschaftlichen Bedeutung her wenig sinnstiftend scheint. So wird beispielsweise eine einfache fragebogengestützte Seminarauswertung oder eine spontane mündliche Abfrage von Workshop-Teilnehmern und Teilnehmerinnen zu ihrer Zufriedenheit häufig als Evaluation bezeichnet, obwohl es sich schlicht um eine Veranstaltungsauswertung handelt, die meist nicht die Kennzeichen einer wissenschaftlichen Evaluation (vgl. Grafik 2) aufweisen kann.

Im wissenschaftlich geführten Diskurs wird Evaluation als Teil angewandter Sozialforschung gesehen. Mit der Definition von Rossi/Freeman/Hofmann (1988: 3) lassen sich einige andere Definitionen recht gut bündeln: danach ist Evaluation als *eine systematische Anwendung sozialwissenschaftlicher Forschungsmethoden zur Beurteilung der Konzeption, Ausgestaltung, Umsetzung und des Nutzens sozialer Interventionsprogramme zu verstehen*. Es gelten dabei die in der Wissenschaft grundlegenden Regeln für das Sammeln valider und reliabler Daten, vgl. Stockmann (2007: 28)[1]. Dabei folgt Evaluation immer einem bestimmten Ablauf, egal welchem Zweck die Erkenntnisse aus der Evaluation letztlich dienen sollen (vgl. Grafik 1).

Grafik 1
Ablauf einer Evaluation[2]

Im Unterschied zur Grundlagenforschung ist Evaluation als angewandte Sozialforschung *nie zweckfrei*, da an ihrem Ende immer eine datenbasierte sowie zielgerichtete Handlungsempfehlung steht.

Evaluation hat zudem den Anspruch *nützlich* zu sein, ihre Ergebnisse sollen fundierte Entscheidungen überhaupt erst ermöglichen. Um diesem Anspruch gerecht werden zu können, muss sie gemäß ihrer Wortbedeutung (valor [lat.]: Geltung, Wert; evaluation [engl.]: Bewertung, Beurteilung; Evaluation [altdeutscher Sprachgebrauch]: Schätzung, Wertschätzung) Bewertungen vornehmen. Diese bewertenden Stellungnahmen müssen sich jedoch auf die mit wissenschaftlichen Methoden gewonnenen Daten stützen. Außerdem müssen die jeweils herangezogenen *Bewertungskriterien* präzise festgelegt und transparent gemacht werden.

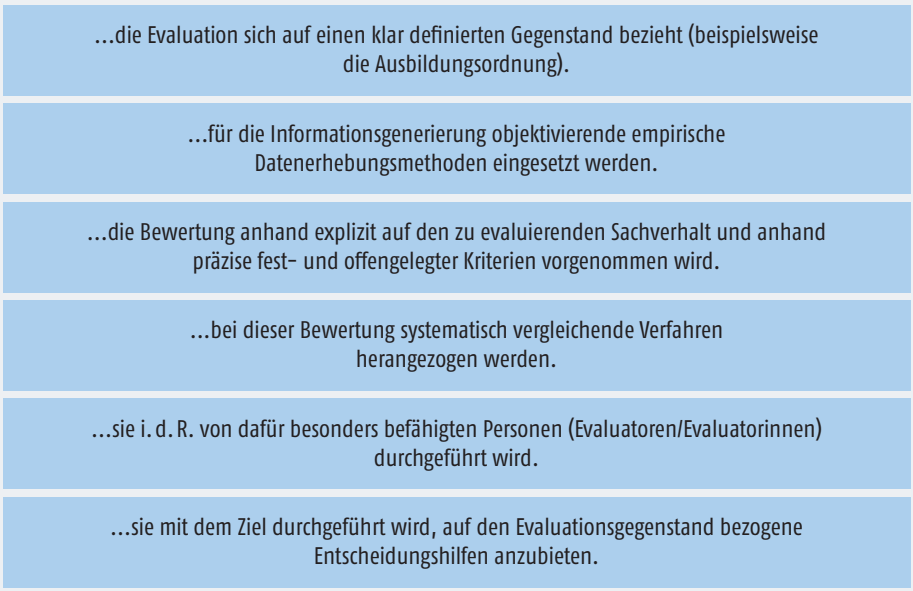
Um die Abgrenzung zum teilweise unwissenschaftlich gebrauchten Begriff der Evaluation zu verdeutlichen, sollte klargestellt werden, dass zudem nicht jede Form der *Bewertung* auch eine Evaluation ist. An dieser Stelle kann das eingangs erwähnte Beispiel der Seminar- oder Workshop-Auswertung fortgeführt werden: Nur weil der Seminarleiter oder die Seminarleiterin die

[1] Stockmann führt in diesem Zusammenhang folgende Autoren an: Clemens (2000), Rossi/Freeman/Hofmann (1988), Kromrey (1995), Wottawa/Thierau (1998), Bortz/Döring (2002).
[2] Eigene Darstellung in Anlehnung an die Erläuterungen von Stockmann (2007: 26).

Daten aus den Fragebögen oder den mündlichen Rückmeldungen als ‚gar nicht so schlecht' ansieht bzw. bewertet, wurde noch keine Evaluation nach wissenschaftlichen Kriterien durchgeführt. Dafür hätte er/sie u. a. seine/ihre impliziten und gegebenenfalls kaum reflektierten Bewertungskriterien zumindest offenlegen müssen.

Weiterhin sollte Evaluation nicht verwechselt werden mit Umfragen- oder Meinungsforschung, Gutachten, Prüfungen, ökonomischen Effizienzmessungen oder Erfolgskontrollen, vgl. Stockmann (2010: 66), da bei all diesen datengenerierenden Verfahren nicht alle Kennzeichen gegeben sind, die eine wissenschaftlich durchgeführte Evaluation auszeichnen. Zu diesen Kennzeichen zählt im Einzelnen, dass…

Grafik 2
Kennzeichen wissenschaftlich durchgeführter Evaluationen[3]



[3] In Anlehnung an Stockmann (2010: 66). Diese Kennzeichen finden sich zudem in den Definitionen von Bortz/Döring (2002), Cronbach (1963), Kromrey (2001), Patton (1997), Rossi (1999), Scriven (1974), Suchman (1967), Weiss (1998), Wottawa/Thierau (1998) etc. wieder, die zusammengefasst in Balzer (2005: 16) nachzulesen sind.

2 Vorüberlegungen und theoretische Grundlagen

Evaluationen lassen sich anhand verschiedener Gestaltungsdimensionen klassifizieren. Mögliche Differenzierungsmerkmale sind beispielsweise:

- Die **Ziele** und der **Zweck**, die mit der Evaluation verknüpft sind. (*Warum wird evaluiert?*) Diese beiden Merkmale werden im nachfolgenden Kapitel 2.1 näher erläutert.
- Der **Gegenstand**, der evaluiert werden soll. (*Wer/was wird evaluiert? Welche Aspekte eines Gegenstands stehen im Mittelpunkt der Betrachtung?*) Dieses Differenzierungsmerkmal wird gemeinsam mit den damit in Verbindung stehenden Forschungsfragen in den Kapiteln 2.2 und 2.3 behandelt.
- Die **Akteure**, welche die Evaluation durchführen bzw. bei ihrer Durchführung mitwirken. (*Wo ist die Evaluation angesiedelt? Wer evaluiert?*) Diese Fragen sind Gegenstand der Kapitel 2.4 und 2.5.
- Der **Zeitpunkt**, zu dem die Evaluation ansetzt. (*Wann wird evaluiert?*) Dieses Merkmal wird in Kapitel 2.6 näher erläutert.
- Die **Methodologie** sowie das **Design**, welche(s) der Evaluation zugrunde liegt. (*Wie wird evaluiert? Wie werden die Ergebnisse dargestellt?*) Dieses Differenzierungs-merkmal wird in Kapitel 2.6 bis 2.8 aus unterschiedlichen Blickwinkeln betrachtet.

2.1 Ziel und Zweck von Evaluationen in der Ordnungsarbeit

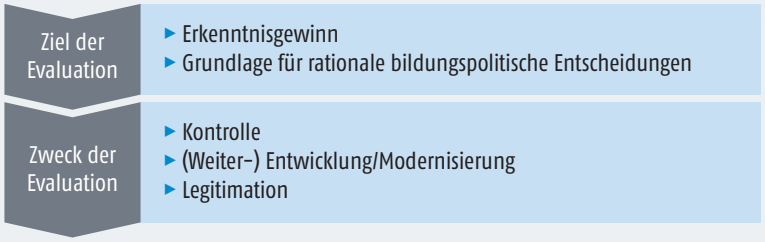
Die Ziele für eine Evaluation im Nachgang zur Ordnungsarbeit sind teilweise sehr heterogen.

Im Falle der Evaluation von Neuordnungen soll i. d. R. geklärt werden, ob und inwieweit die Ziele und Intentionen der Neuordnung in die Praxis der beruflichen Ausbildung von Betrieb und Berufsschule sowie in den Prüfungen umgesetzt werden. Gegebenenfalls weiterführend soll sie aufzeigen, wie Qualifikationen verwertet werden und inwieweit die Ausbildungsordnung grundsätzlich anforderungsgerecht gestaltet ist. Daneben gibt es auch spezifischere Gründe für eine Evaluation, wie beispielsweise technische Neuerungen, die im aktuell gültigen Berufsbild noch nicht ausreichend berücksichtigt werden. Ziel solcher Evaluationen ist es, folglich Klarheit darüber zu gewinnen, ob die jeweilige Ausbildungsordnung gegebenenfalls anzupassen bzw. zu ändern ist. Ähnlich gestaltet sich dies auch bei Evaluationen von Erprobungsverordnungen. Sie zielen u. a. darauf ab, eventuell vorhandene Schwachstellen des aktuellen Ausbildungsprofils aufzudecken, die im Zuge der Überführung in die Regelverordnung dann noch rechtzeitig korrigiert werden können. Speziell bei der Evaluation von Erprobungsverordnungen ist die Bestandsaufnahme ein wichtiges Forschungsziel: Es wird nach der Anzahl der Ausbildungsverhältnisse gefragt, die seit dem Erlass geschlossen und auch wieder gelöst wurden, nach der Art der Ausbildungsbetriebe sowie nach demografischen Daten der Auszubildenden etc. Ziel ist es daher auch, erste umfassende und aussagekräftige Daten zu diesem neuen Beruf zu erheben. Bei der Evaluation von Teilbereichen einer Ausbildungsordnung, wie beispielsweise den Prüfungsregelungen, steht die Frage im Vordergrund, ob die verwendeten Prüfungskonzepte (noch) zeitgemäß sind. Sind neue Prüfungskonzepte eingeführt, so ist ein Evaluationsziel u. a. die Messung ihrer Akzeptanz bei den Betroffenen.[4]

[4] In diesem Zusammenhang sind auch die Weisungen zur Evaluation der Gestreckten Abschluss- bzw. Gesellenprüfung (GAP) zu sehen. Dabei wurden im Zeitraum von Anfang 2003 bis Ende 2007 rund 20 Berufe hinsichtlich der zum damaligen Zeitpunkt neu eingeführten GAP evaluiert.

Die Ziele von Evaluationen im Nachgang zur Ordnungsarbeit lassen sich folglich unter den beiden Schlagwörtern *Erkenntnisgewinn* und *Grundlage zur Entscheidungsfindung* zusammenfassen. In den Zielformulierungen ist zudem, wenn auch eher indirekt, der Zweck einer Evaluation enthalten. Dabei sind grob drei Zweck-Arten zu unterscheiden (vgl. Grafik 3):

Grafik 3
Ziel/Zweck der Evaluation[5]



Kontrolle: Die gewonnenen Erkenntnisse sollen darüber Aufschluss geben, ob das mit der aktuell gültigen Ausbildungsordnung verfolgte Ziel erreicht werden kann.

(Weiter-)Entwicklung/Modernisierung: Die gewonnenen Erkenntnisse sollen dazu dienen, die Ausbildungsordnung – falls notwendig – weiterentwickeln bzw. modernisieren zu können. Darüber hinaus kann der damit verbundene Innovationsgedanke ganz allgemein Ideen und Anregungen für das duale System liefern. Die erzielten Erkenntnisse müssen sich nicht auf die evaluierte Ausbildungsordnung beschränken.

Legitimation: Die gewonnenen Erkenntnisse sollen eine Entscheidungsgrundlage liefern, mithilfe derer der Erhalt oder auch eine notwendige Weiterentwicklung bzw. Modifikation der Ausbildungsordnung begründet werden kann.

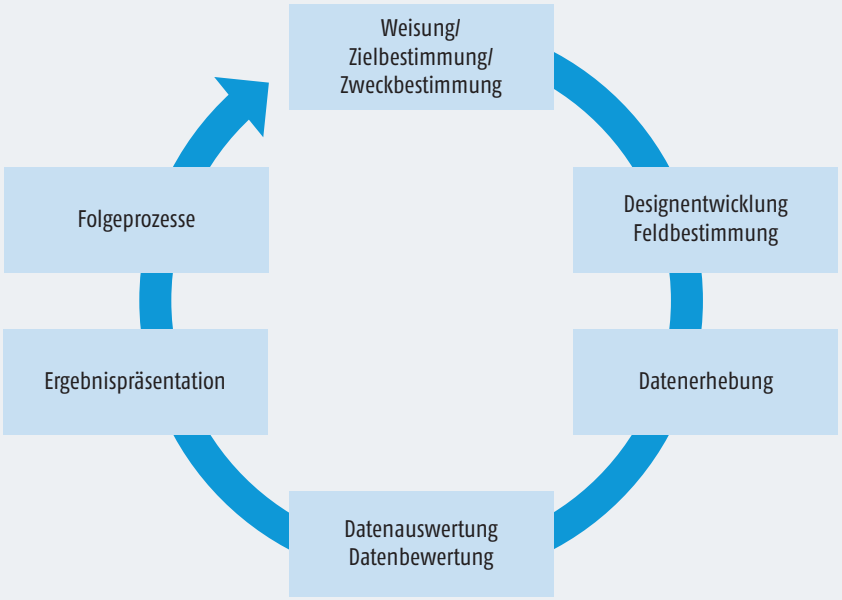
Gemäß der *Standards für Evaluation* der Deutschen Gesellschaft für Evaluation e. V. (DeGEval) ist der Evaluationszweck stets zu klären: „Es soll deutlich bestimmt sein, welche Zwecke mit der Evaluation verfolgt werden, so dass die Beteiligten und Betroffenen Position dazu beziehen können und das Evaluationsteam einen klaren Arbeitsauftrag verfolgen kann“ (DeGEval 2008a).

Die Evaluation mit ihren Zielen und dem Zweck kann u. a. als Zyklus dargestellt werden. Sind die Ergebnisse präsentiert, folgt ein neuer Prozess bzw. Zyklus, der andere Ziele mit einem an-

[5] Eigene Darstellung in Anlehnung an das Modell „Ziele von Evaluation“ von Stockmann (2007: 37).

deren Zweck verfolgt, wie beispielsweise ein sich an die Evaluation anschließendes Neuordnungsverfahren (vgl. Grafik 4):

Grafik 4
Evaluationszyklus



2.2 Die Ausbildungsordnung als Evaluationsgegenstand

Unmittelbar mit der Zielklärung verbunden ist die präzise Erfassung des Evaluationsgegenstands. Dieser Evaluationsgegenstand, von Stockmann (2010: 67) sowie Wottawa/Thierau (2003: 59) auch als *Objekt der Bewertung* bezeichnet, ist die Ausbildungsordnung oder die Erprobungsverordnung. Daneben können aber auch Teilbereiche einer Ausbildungsordnung wie beispielsweise Prüfungsanforderungen und Zusatzqualifikationen, Fortbildungsordnungen oder Ausbildungsbausteine Gegenstand von Evaluationen sein. Die folgenden Ausführungen konzentrieren sich jedoch ausschließlich auf die Evaluation von Ausbildungs- und Erprobungsverordnungen.

Exkurs

Die meisten Autoren und Autorinnen, die sich in der Fachliteratur theoretisch mit dem Evaluationsgegenstand auseinandersetzen, konzentrieren sich auf *Programme* oder *Projekte*. Was unter Programmen und Projekten genau zu verstehen ist, wird im nachfolgenden Infokasten 1 kurz dargestellt.

Infokasten 1

Instrumentell betrachtet handelt es sich bei *Programmen* und *Projekten* um Maßnahmenbündel zur Erreichung spezifizierter Ziele. Mithilfe konzeptionell zugeschnittener, aufeinander abgestimmter Aktivitäten bzw. Interventionen sollen diese Ziele unter Einsatz von Ressourcen (Mittel, Personal) zu den gewünschten Resultaten bei den Zielgruppen führen. Mit den Programmen bzw. Projekten sollen Innovationen innerhalb sozialer Systeme eingeleitet werden.

Vgl. Stockmann (2010: 68) sowie Beywl (2006: 115)

Eine Ausbildungsordnung ist jedoch weder ein Programm noch ein Projekt, weshalb es teilweise schwerfällt, die dazu in der Fachliteratur getroffenen Aussagen auf den Evaluationsgegenstand Ausbildungsordnung anzuwenden. Bei einer Ausbildungsordnung handelt es sich vielmehr um eine *Rechtsverordnung*, die ihre Legitimation aus dem Berufsbildungsgesetz (BBiG) bezieht. Zwar finden sich einige Charakteristika, die ein Programm oder ein Projekt ausmachen, auch in einer Ausbildungsordnung wieder – so können einer Ausbildungsordnung beispielsweise *konkrete Ziele entnommen werden*, die in §1, Abs. 3 BBiG (Zielsetzung der Berufsausbildung) festgeschrieben sind – jedoch würde eine Forschungslogik, welche eine Ausbildungsordnung als Programm oder Projekt begreift, schnell an ihre Grenzen stoßen: Dies wäre dann zum Beispiel der Fall, wenn nach weiteren Charakteristika eines Programms, wie etwa einem *eigenen Budget*, *stabilen Finanzmittelzuweisungen* etc. gefragt würde, vgl. Stockmann (2010: 68).

Wie ein Evaluationsgegenstand grundsätzlich umfassend begriffen werden kann, führt Kromrey (2006: 115–116) vor, indem er eine modellhafte Strukturierung des Objektbereichs vornimmt: „In gedanklichen Vorleistungen ist der Gegenstand der Untersuchung so zu durchleuchten, so in seine vielfältigen Facetten zu zerlegen und zu ordnen, dass daraus ein problemangemessenes Forschungsdesign entwickelt und begründet werden kann." Im Rahmen einer deskriptiven Untersuchung wird somit die empirische Struktur des Realitätsausschnitts in *Dimensionen* erfasst, die für die aktuelle Fragestellung besonders bedeutsam erscheinen.

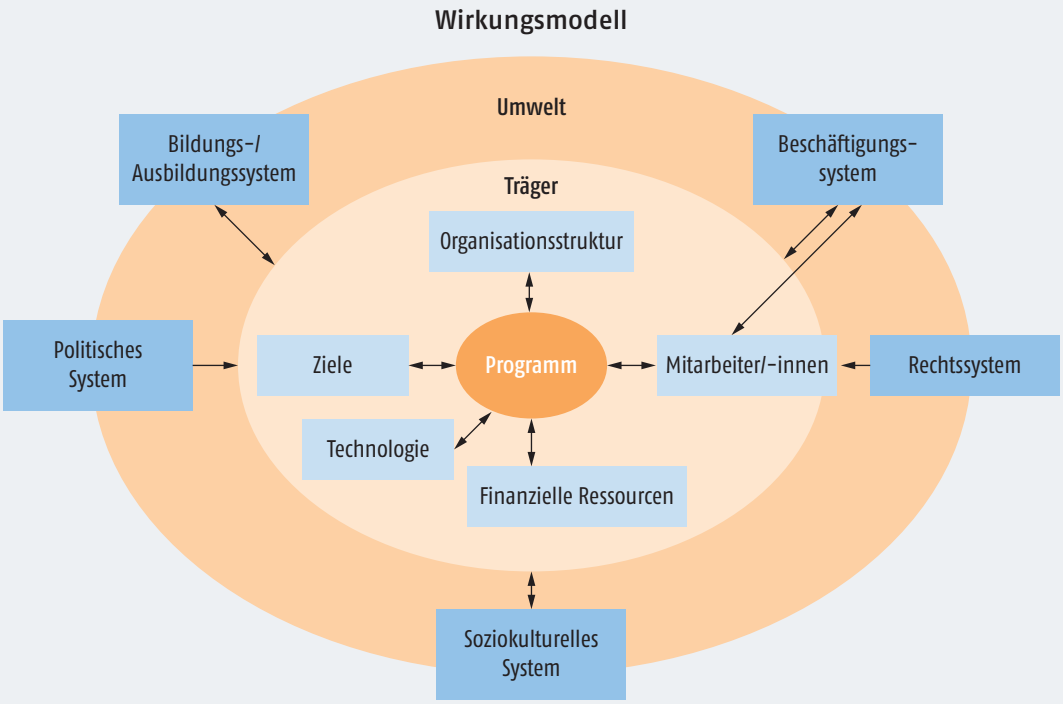
Beywl/Univation (2004) geht im Vergleich dazu etwas pragmatischer zu Werke und ‚fasst' den Evaluationsgegenstand zunächst ganz allgemein, indem er ihn

- hinsichtlich der Interventionen, Träger und Zielgruppen beschreibt und anschließend
- empirisch entlang der gegebenen Evaluationsfragestellungen untersucht.

Behielte man die Logik Kromreys bei und beließe es bei diesen drei von Beywl/Univation genannten Dimensionen (Interventionen, Träger, Zielgruppen), würde jedoch grundsätzlich eine entscheidende Realitätskomponente fehlen: die Umwelt. Stockmann (2007: 52) misst im Rahmen seines Wirkungsmodells dieser Dimension Bedeutung bei und bildet die Umwelt durch verschiedene Subsysteme ab (vgl. Grafik 5).

Diese vier identifizierten Dimensionen (Interventionen, Träger, Zielgruppe und Umwelt), müssen nun in ihrer terminologischen Logik dem Evaluationsgegenstand Ausbildungsordnung angepasst werden, da sich auch die Terminologie in der Evaluationsfachliteratur vorwiegend auf Programme und Projekte bezieht. Statt von Interventionen, sollte eher vom *Ordnungsrahmen* gesprochen werden, da eine Ausbildungsordnung nicht aus intervenierenden Elementen besteht, sondern vielmehr als ein ordnender Rahmen aufgefasst werden kann, für den klare Standards formuliert sind. Da es keinen offiziellen Träger einer Ausbildungsordnung gibt, sollte zudem statt vom Träger vom *Weisungsgeber* (in diesem Falle das zuständige Fachministerium), vom *Weisungsnehmer* (dem BIBB) sowie von den *Sozialpartnern* die Rede sein. Schließlich sollte statt Umwelt der Begriff *Kontext* verwendet werden, da Umwelt erfahrungsgemäß bei vielen Beteiligten die Aufmerksamkeit allein auf ökologische Aspekte lenkt und eben nicht auf die von Stockmann intendierten Subsysteme.

Grafik 5
Wirkungsmodell[6]



2.2.1 Die Dimension „Ordnungsrahmen“

Um die Dimension „Ordnungsrahmen“ vollständig erfassen zu können, ist zunächst die Auflistung der *Kernelemente* einer Ausbildungsordnung erforderlich (vgl. Grafik 6):

Grafik 6
Kernelemente des Ordnungsrahmens[7]

6 Eigene Darstellung in Anlehnung an Stockmann (2010: 93).
7 Eigene Darstellung in Anlehnung an die Ausführungen in BIBB (2006: 11–12).

Diese sieben Kernelemente stecken den Ordnungsrahmen ab, innerhalb dessen ein Beruf ausgebildet wird. Sie sind der Dreh- und Angelpunkt einer jeden Evaluation einer Ausbildungsordnung und werden dabei in ihrer *praktischen Umsetzung* untersucht.

Die Art und Weise der praktischen Umsetzung hängt zum einen von Faktoren ab, die unmittelbar durch die Ausbildungsordnung (Paragrafenteil und Ausbildungsrahmenplan mit sachlicher und zeitlicher Gliederung) bestimmt werden. Zum anderen wird die praktische Umsetzung aber auch durch andere (externe) Faktoren beeinflusst. Um dieses Spektrum der Einflussfaktoren erfassen zu können, ist die Betrachtung der Dimension „Kontext“ erforderlich.

2.2.2 Die Dimension „Kontext“

Der Kontext bzw. die Rahmenbedingungen, in denen eine Ausbildungsordnung betrachtet werden muss, sind je nach Beruf sehr unterschiedlich. Ein Beruf mit vielen Ausbildungsverhältnissen wie zum Beispiel Kaufmann und Kauffrau im Einzelhandel muss mit gänzlich anderen Rahmenbedingungen umgehen als beispielsweise der Beruf Geigenbauer/Geigenbauerin mit relativ wenigen Auszubildenden und gänzlich anderen Branchenverhältnissen.

Zum Kontext, in dem die Ausbildungsordnung zu sehen ist, zählen daher zum Beispiel neben der Branchenstruktur, den Marktbedingungen sowie der Vorbildung der Auszubildenden auch die Umstände, unter denen die vorangegangene (Neu-)Ordnung stattgefunden hat. Darüber hinaus können auch bildungs- und arbeitsmarktpolitische Aspekte dem Kontext zugeschrieben werden, vgl. Paulini (1995: 40–41).

Bei der Auflistung der typischen Forschungsfragen (vgl. Kapitel 2.3), die der Weisungsgeber oder das Projektteam beantwortet haben möchte, fällt auf, dass vereinzelt Bereiche angesprochen werden, die nicht ausschließlich durch die Ausbildungsordnung beeinflusst werden können. So beispielsweise bei der Forschungsfrage „Wird der Beruf von den Betrieben nachgefragt?“. Als Hypothese könnte daher aufgestellt werden: „Die neugeordnete Ausbildungsordnung ist gelungen und erzeugt dadurch eine größere Nachfrage als vor ihrer Neuordnung.“ Sicherlich kann eine ‚gelungene‘ Ausbildungsordnung dazu beitragen, dass der Ausbildungsberuf attraktiv und modern gestaltet ist und somit eine gewisse Nachfrage hervorruft. In diesem Falle wäre die Ausbildungsordnung als *unabhängige Variable* anzusehen und die Nachfrage als *abhängige Variable,* vgl. Sedlmeier/Renkewitz (2008: 128). Ein konjunktureller Abschwung, demografische Einflüsse oder eine relative Unbeliebtheit des Berufs können jedoch trotz gelungener Ausbildungsordnung dazu führen, dass ein Beruf seltener nachgefragt wird. Dann verliert die Ausbildungsordnung ihren Status als unabhängige Variable und andere Faktoren treten an ihre Stelle. Dies sollte ebenfalls bei der Dimension Kontext im Blick behalten werden.

Der Kontext mit den von Stockmann benannten Subsystemen muss im Rahmen einer Evaluation daher immer mit einbezogen werden und ist als ein Teil des Evaluationsgegenstandes anzuerkennen; nicht zuletzt weil die praktische Umsetzung einer Ausbildungsordnung stets kontextabhängig erfolgt.

2.2.3 Die Dimension „Weisungsgeber, Weisungsnehmer, Sozialpartner“

Zur vollständigen Erfassung des Evaluationsgegenstandes gehört es weiterhin, die Institutionen bzw. deren Vertretungen zu identifizieren, die für die ursprüngliche (Neu-)Ordnung des zu evaluierenden Berufes verantwortlich waren. Diese sind in der Regel Vertreterinnen und Vertreter:

- der zuständigen Fachministerien,
- des Bundesministeriums für Bildung und Forschung,
- des BIBB (Berufeverantwortliche),
- der Arbeitnehmer und Arbeitnehmerinnen,

- der Arbeitgeber,
- der Länder,
- von Fachverbänden.

Sie alle sind auf strategischer und operativer Ebene mit verantwortlich für die Ausbildungsordnung und damit wichtige *Informationsträger* sowie *Multiplikatoren*. Im Rahmen der Evaluation einer Ausbildungsordnung sollten diese daher als wichtige Probanden, von denen Daten zu erheben sind bzw. die wichtige Unterlagen zur Verfügung stellen können, im Sinne der Daten-Triangulation (vgl. Kapitel 2.6.2) wahrgenommen werden.[8]

2.2.4 Die Dimension „Zielgruppen"

Die Betrachtung des Evaluationsgegenstandes Ausbildungsordnung bliebe unvollständig, würde die Dimension „Zielgruppen" nicht beachtet werden. Zu dieser Dimension gehören zum einen die durch die Ausbildungsordnung unmittelbar angesprochenen Personengruppen. Zu ihnen zählen:

- Auszubildende,
- Ausbildungsbetriebe und Ausbilderinnen und Ausbilder,
- Berufsschullehrerinnen und Berufsschullehrer,
- Prüferinnen und Prüfer sowie die
- Zuständigen Stellen.

Zum anderen können je nach Forschungsfragestellung (vgl. Kapitel 2.3) auch noch weitere Zielgruppen in den Fokus einer Evaluation genommen werden. Mögliche Zielgruppen wären in diesen Fällen:

- potenzielle Auszubildende,
- an der Ausbildung Interessierte, die jedoch keinen Ausbildungsplatz erhalten,
- fertig ausgebildete Berufseinsteigerinnen und Berufseinsteiger,
- potenzielle Ausbildungsbetriebe, die derzeit den entsprechenden Beruf aber nicht ausbilden,
- Auszubildende ähnlicher oder verwandter Berufe,
- Ausbildungsbetriebe ähnlicher oder verwandter Berufe,
- Fachverbände.

Für alle aufgeführten Zielgruppen gilt, dass sie wichtige *Informationsträger* sind und – je nach den Forschungsfragestellungen – identifiziert und befragt werden sollten. Zudem sind sie als *Multiplikatoren* zu betrachten, da es in erheblichem Maße von ihnen abhängt, wie eine Ausbildungsordnung praktisch umgesetzt wird.

Es bleibt festzuhalten, dass die umfassende Betrachtung des Evaluationsgegenstands die Berücksichtigung der eben erläuterten vier Dimensionen erfordert, die eng miteinander verbunden sind. Diese Form der Betrachtung fordern auch die *Standards für Evaluation* der DeGEval unter den beiden Aspekten *Nützlichkeit* und *Genauigkeit* (vgl. Infokasten 2).

Die Betrachtung aller vier Dimensionen kann zu Schwierigkeiten führen: Die Evaluation scheint ‚auszuufern'. Es sollte daher bei der Konzeptionierung einer Evaluation unbedingt überlegt werden, ob und gegebenenfalls wie der Evaluationsgegenstand eingegrenzt werden kann ohne dabei eine der vier Dimensionen unberücksichtigt zu lassen. Maßgeblich ist in diesem Zusammenhang zunächst das Forschungsinteresse, das in der Weisung festgehalten ist.

[8] Diese Dimension sollte nicht verwechselt werden mit den Akteuren während der Evaluation. Diese werden in Kapitel 2.4 gesondert behandelt.

Infokasten 2

Beschreibung des Evaluationsgegenstandes

„Der Evaluationsgegenstand soll klar und genau beschrieben und dokumentiert werden, so dass er eindeutig identifiziert werden kann."

Kontextanalyse

„Der Kontext des Evaluationsgegenstandes soll ausreichend detailliert untersucht und analysiert werden."

Identifizierung der Beteiligten und Betroffenen

„Die am Evaluationsgegenstand beteiligten oder von ihm betroffenen Personen bzw. Personengruppen sollen identifiziert werden, damit deren Interessen geklärt und so weit wie möglich bei der Anlage der Evaluation berücksichtigt werden können."

Auswahl und Umfang der Informationen

„Auswahl und Umfang der erfassten Informationen sollen die Behandlung der zu untersuchenden Fragestellungen zum Evaluationsgegenstand ermöglichen und gleichzeitig den Informationsbedarf des Auftraggebers und anderer Adressaten und Adressatinnen berücksichtigen."

DeGEval (2008a)

2.3 Die Forschungsfragen

Eng mit dem Ziel der Evaluation sowie dem Evaluationsgegenstand verbunden sind die Forschungsfragen, die in der Regel entweder den ministeriellen Weisungen oder den Projektbeschreibungen entnommen werden können.

Zentrale Forschungsfragen, die bisher bei Evaluationen von Ausbildungsordnungen von Interesse waren, sind beispielhaft in Tabelle 1 in der Spalte *Typische Forschungsfragen* in zusammengefasster Form dargestellt. Aus diesen Fragen kann ein *Erkenntnisinteresse* abgeleitet werden (vgl. die gleichnamige Spalte).

Bei manchen der o. g. Forschungsfragen fällt auf, dass diese nicht eindeutig gestellt wurden. Was beispielsweise eine ‚stimmige' oder ‚passgenaue' Ausbildungsordnung ausmacht, ist nicht einfach zu klären. Dies verweist auf ein grundsätzliches Problem, welches vielen Evaluationen gemein ist: sowohl die Definitionen als auch zum Teil die Bewertungskriterien sind häufig unklar.

Um eine Ausbildungsordnung bewerten zu können, muss spätestens zu Beginn der Evaluation geklärt sein, welche Erwartungen (zum Beispiel des weisungsgebenden Ministeriums) an eine Ausbildungsordnung gestellt werden, was diese leisten muss und was genau ihre ‚Passgenauigkeit' ausmacht, was also unter einem bestimmten Begriff verstanden wird.

Tabelle 1
Typische Forschungsfragen

Erkenntnisinteresse	Typische beispielhafte Forschungsfragen bzw. Forschungsimpulse[9]
Umsetzbarkeit der Ausbildungsordnung in der betrieblichen Praxis	Wie werden die Ordnungsmittel im Betrieb umgesetzt (Transfer, Formen betrieblicher Ausbildung, hemmende und fördernde Faktoren)?
Praktikabilität von Prüfungen und Aussagekraft von Prüfungen	Wie gestaltet sich der Prüfungsablauf? Bilden die Prüfungsaufgaben die berufliche Handlungsfähigkeit ab? Wie gestalten sich die praktischen Prüfungsaufgaben? Wie hoch ist die Bestehen-Quote?
Qualifizierungsanforderungen der Betriebe und Niveau der Inhalte	Ist das Berufsbild auf einem sachgerechten und für die Betriebe adäquaten Niveau formuliert? Stimmen die Ausbildungsinhalte mit den Qualifikationsanforderungen im Betrieb überein?
Passgenauigkeit der Ausbildungsordnung	Ist die Berufsstruktur insgesamt stimmig? In welchem Umfang entspricht das Ausbildungsberufsbild den aktuellen Qualifikationsanforderungen der ausbildenden Betriebe?
Branchenanalyse und Ausbildungsmarktanalyse	Wie viele Ausbildungsverhältnisse bestehen derzeit und wie viele Ausbildungsverträge wurden gelöst? Wie ist die Entwicklung der Ausbildungsvertragsverhältnisse einschließlich ihrer regionalen Verteilung? Wie hoch ist der künftige Bedarf an ausgebildeten Fachkräften?
Zukünftige Qualifikationsanforderungen	In welchem Umfang entspricht das Ausbildungsberufsbild den zukünftigen Qualifikationsanforderungen der ausbildenden Betriebe?
Abgrenzung zu anderen Berufen	Wie verhält sich der Zuschnitt des Berufsprofils im Hinblick auf die Abgrenzung zum Beruf XY?
Umsetzbarkeit des Rahmenlehrplans	Gibt es Probleme bei der Umsetzung des neuen Rahmenlehrplans? Wie gelingt die Umsetzung der Inhalte des Rahmenlehrplans? Bilden die Inhalte des Rahmenlehrplans die berufliche Handlungsfähigkeit ab?
Akzeptanz und Bewährung der Ausbildungsordnung	Wird die Berufsstruktur von der Praxis angenommen?
Verwertbarkeit der Qualifikation auf dem Arbeitsmarkt	Wird der Beruf von Betrieben nachgefragt?
Fortbildungs-, Aufstiegs-, Entwicklungsmöglichkeiten	Ist der Zuschnitt der Ausbildungsordnung im Hinblick auf Fortbildungs- und Aufstiegsmöglichkeiten förderlich?
Einsatzbereiche nach Abschluss der Ausbildung	In welchen Bereichen werden die ausgebildeten Fachleute eingesetzt?
Verbleib der Absolventen und Absolventinnen	Wo verbleiben erfolgreiche Absolventen und Absolventinnen der Berufsausbildung? Wie viele ehemalige Auszubildende werden vom Ausbildungsbetrieb in ein Beschäftigungsverhältnis übernommen?
Verwertbarkeit der Ausbildungsordnung in angrenzenden Branchen	Gibt es Einsatzmöglichkeiten für das Berufsbild auch in sogenannten „Randbereichen"?

[9] Die oben aufgeführten Forschungsfragen wurden zwölf Evaluationsprojekten entnommen, die im BIBB durchgeführt worden sind. In diese Projekte waren insgesamt 22 Berufe einbezogen.

(Fortsetzung Tab. 1)

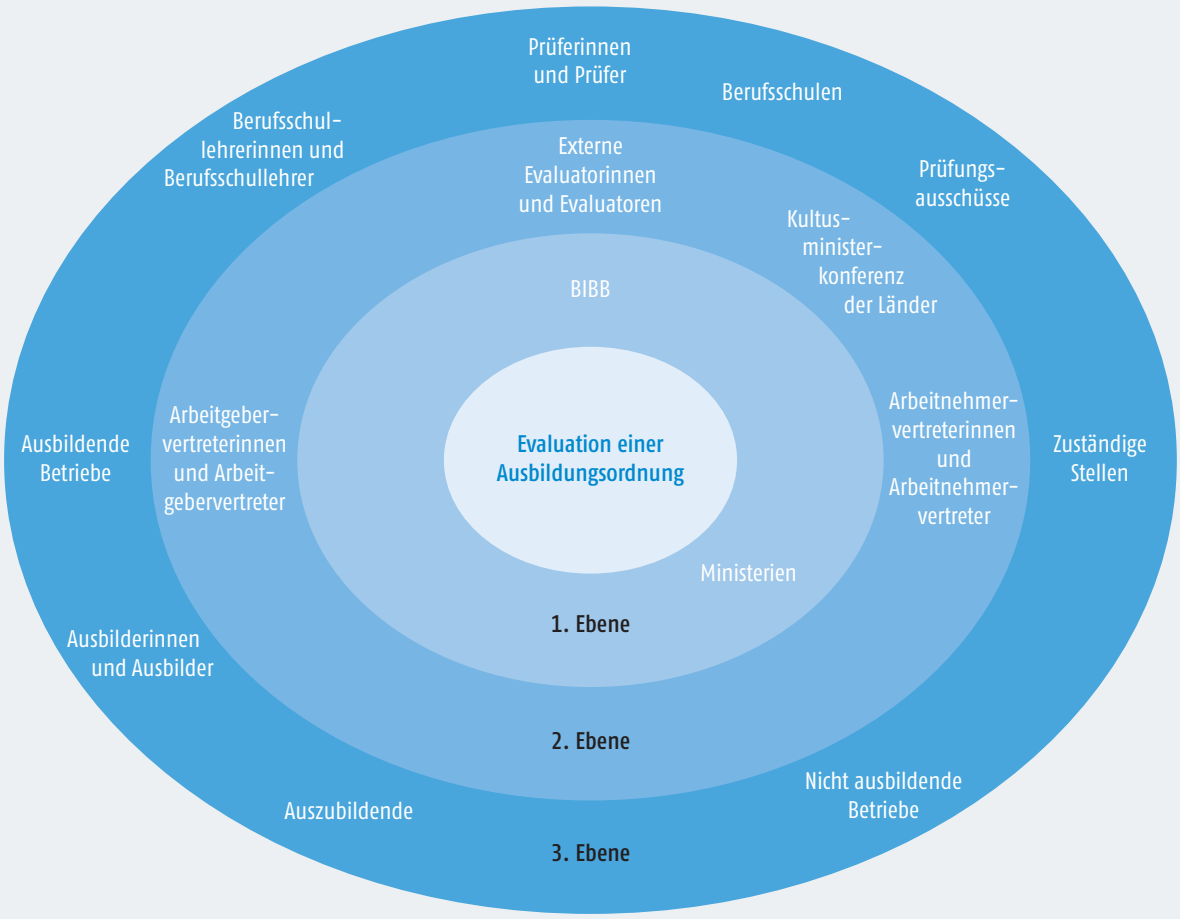
Erkenntnisinteresse	Typische beispielhafte Forschungsfragen bzw. Forschungsimpulse[9]
Empfehlungen/Arbeitshilfen für mögliche Neuordnung	Welche Änderungen in der Ausbildungsordnung sind hinsichtlich der Qualifikationsanforderungen im Betrieb erforderlich? Gibt es Verbesserungsvorschläge?
Entscheidung zur Überführung der Erprobungsverordnung in eine Regelverordnung	Kann die Erprobungsverordnung in eine unbefristete Ausbildungsordnung überführt werden?
Berufliche Handlungskompetenz	Wie ist der Aussagewert der Abschlussprüfung in Bezug auf die von den Betrieben erwartete berufliche Handlungskompetenz?
Umsetzbarkeit überbetrieblicher Ausbildung/ Verbundausbildung	Wie gelingt die Umsetzung der Lernziele der Ausbildungsordnung in der betrieblichen Praxis? Wie gelingt die Umsetzung der Lernziele der Ausbildungsordnung in der überbetrieblichen Ausbildung bzw. Verbundausbildung?
Identifikation von Fortbildungsthemen	Welche Fortbildungsthemen sind nach der Ausbildung von Bedeutung?
Stellenwert von Qualifikationen	Welchen Stellenwert haben die erworbenen Qualifikationen (z. B. im Verhältnis zum Servicegedanken)?
Angebot (Bedarf) an Zusatzqualifikationen	Welche Zusatzqualifikationen werden bereits während der Ausbildung angeboten?
Branchenveränderungen	Wie wirken sich die veränderten Anforderungen an die Unternehmen auf die Berufe in qualitativer (aber auch quantitativer) Hinsicht aus?
Anforderungen an die Ausbilderinnen und Ausbilder	Welche Qualifikationen braucht das Ausbildungspersonal?
Berufsspezifische Fragestellungen	Diverse, z. B.: Welche Konsequenzen ergeben sich aufgrund der sportpraktischen und der kaufmännisch-administrativen Tätigkeiten für die künftige Qualifikationsstruktur des zu überarbeitenden Berufs? Welche Relevanz haben die Strukturmerkmale Antrag, Vertrag und Leistung für die Beschäftigung von Fachkräften und für die Ausbildung?

2.4 Akteure während der Evaluation von Ausbildungsordnungen

Die Analyse des Evaluationsgegenstands (vgl. Kapitel 2.2) ergab, dass die Betrachtung der Dimensionen „Weisungsgeber, Weisungsnehmer, Sozialpartner“ und „Zielgruppen“ wichtig ist und u. a. von diesen Personengruppen auch Daten erhoben werden sollten.

Manche Vertreterinnen und Vertreter dieser Gruppen waren unmittelbar an der Ausgestaltung der jeweiligen zu evaluierenden Ausbildungsordnung beteiligt, beispielsweise als Mitglied eines begleitenden Beirats. Diese Überschneidung kann Verwirrung stiften, weshalb an dieser Stelle ein *Organisationsmodell* eingeführt wird, das aufzeigen soll, welche Akteure sich grundsätzlich an der *Ausgestaltung der Evaluation einer Ausbildungsordnung* beteiligen können (vgl. Grafik 7):

Grafik 7
Organisationsmodell



Dieses Vorgehen, alle denkbaren Akteure, die bei der Planung und Durchführung einer Evaluation beteiligt sein können, in einem Modell abzubilden, bezeichnet Stockmann (2010: 162) auch als Stakeholder- bzw. *Akteursanalyse*. Sie dient u. a. dazu im Vorfeld abzuwägen, wessen Beteiligung bei der Ausgestaltung einer Evaluation aufgrund von Bestimmungen unumgänglich ist sowie darüber hinaus möglich und sinnvoll sein könnte.

► Akteure auf der ersten Ebene des Organisationsmodells

Das BIBB ist Weisungsnehmer; ihm obliegt die Verantwortung für die Planung und Durchführung der Evaluation einer Ausbildungsordnung. Als Weisungsgeber treten die zuständigen Ministerien in Erscheinung. Im Organisationsmodell sind daher auf der 1. Ebene als Akteure eingezeichnet:

- Berufe-Expertinnen und Berufe-Experten des BIBB,
- Evaluationsexpertinnen und Evaluationsexperten des BIBB,
- Vertreterinnen und Vertreter der zuständigen Ministerien.

Im BIBB werden die staatlich anerkannten dualen Ausbildungsberufe geordnet und evaluiert. Daher kann es vorkommen, dass eine Mitarbeiterin oder ein Mitarbeiter des BIBB jene Berufe evaluiert, an deren Ordnung sie oder er in der Vergangenheit selbst beteiligt war. Diese Situation bzw. Personalunion entspricht einer *Selbstevaluation* und erfordert ein besonders methodisches Vorgehen, welches in Kapitel 2.5.1. beschrieben wird.

► Akteure auf der zweiten Ebene des Organisationsmodells

Meist wird die Evaluation von einem Beirat begleitet. Solchen Projektbeiräten gehören i. d. R. Vertreterinnen und Vertreter

- der Kultusministerkonferenz der Länder und
- der Arbeitnehmer- sowie Arbeitgeberorganisationen an.

Grundsätzlich hat ein Beirat ausschließlich eine beratende Funktion. Die Projektverantwortung liegt auf der 1. Ebene des Organisationsmodells.

Auf der 2. Ebene dieses Modells sind zudem externe Evaluatorinnen und Evaluatoren genannt. Da Teile einer Evaluation an externe Dienstleister vergeben werden können, haben diese auch Einfluss auf die Durchführung einer Evaluation und sollten daher auch als Akteure wahrgenommen werden.

► Akteure auf der dritten Ebene des Organisationsmodells

Der äußere Kreis im Organisationsmodell umschließt jene Akteure, die in Kapitel 2.2.4 auch als mögliche *Zielgruppen* vorgestellt wurden. Auch sie können an der Ausgestaltung einer Evaluation aktiv beteiligt werden und müssen nicht zwangsläufig „nur“ als Probanden wahrgenommen werden. Diese Form der Beteiligung von Zielgruppen an der Ausgestaltung von Evaluationen wird in der Fachliteratur auch als *partizipativer Evaluationsansatz* bezeichnet. Näheres hierzu in Kapitel 2.8.3.

Festzuhalten bleibt, dass die Akteure der 2. und 3. Ebene des Organisationsmodells bei der Ausgestaltung und Durchführung der Evaluation unterschiedlich eingebunden werden können. Anzuraten ist auf jeden Fall, sich das vorhandene Know-how der 2. Ebene zunutze zu machen; jedoch sollte die Beteiligung mit Augenmaß vorgenommen werden, damit während der gesamten Evaluation für alle Beteiligten klar zu erkennen ist, dass das BIBB der projektverantwortliche Akteur ist.

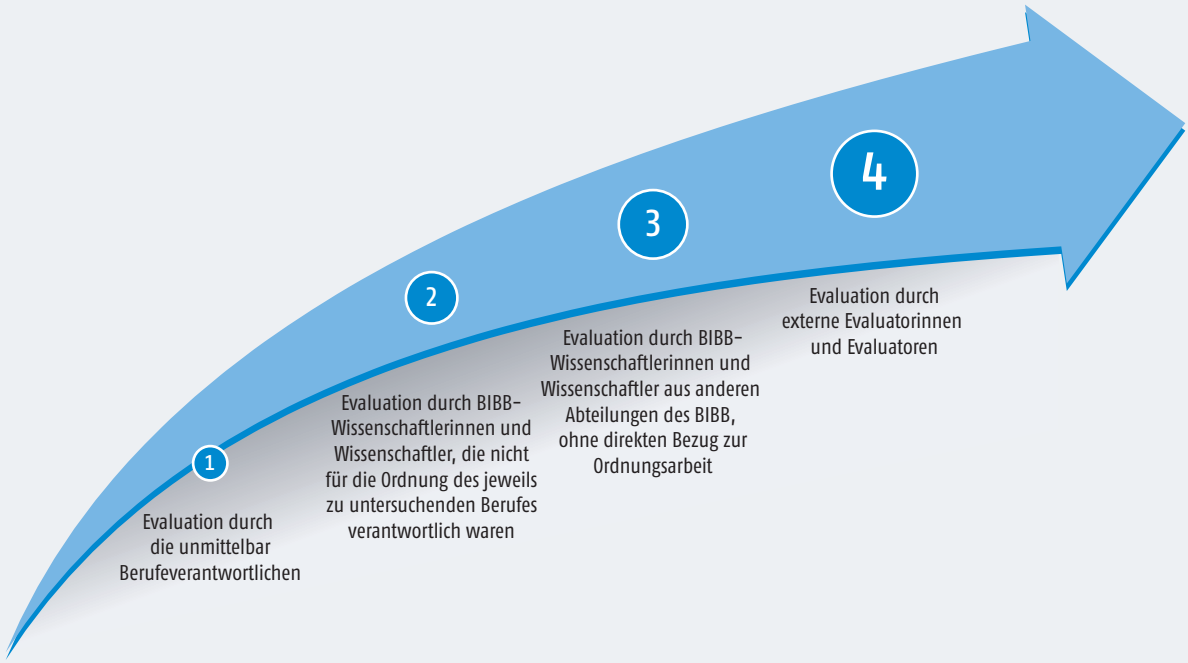
Eine Beteiligung der 3. Ebene an der Planung einer Evaluation ist demgegenüber nur schwierig umzusetzen. So wirft nicht nur die Entscheidung, welche Vertreterinnen und Vertreter mit welcher Begründung ausgewählt werden, häufig große Schwierigkeiten auf. Daher sollten eher die Akteure der 2. Ebene des Modells als *repräsentative* Vertreterinnen und Vertreterinnen der 3. Ebene angesehen und von einer Einbindung der 3. Ebene in die Ausgestaltung einer Evaluation eher abgesehen werden. Nicht zuletzt hängt die Form der Akteurs-Einbindung in die Planung und Durchführung der Evaluation jedoch auch in entscheidendem Maße von dem gewählten *Evaluationsansatz* ab (weitere Ausführungen zu möglichen Evaluationsansätzen folgen in Kapitel 2.8)

2.5 Die Unabhängigkeit zum Evaluationsgegenstand

Wie im vorangegangenen Kapitel kurz angesprochen, können Evaluationen zum Teil von Personen durchgeführt werden, die eng mit dem zu evaluierenden Beruf verbunden sind, da sie an seiner vorangegangenen (Neu-) Ordnung beteiligt waren. In anderen Fällen kann beispielsweise durch personelle Wechsel diese Nähe nicht gegeben sein. Kategorisieren ließe sich diese Bandbreite an Settings, indem der Einfluss der Evaluierenden auf die zuvor durchgeführte Ordnung des Berufes gemessen wird. Je höher demnach der *Mitwirkungsgrad am ursprünglichen Ordnungsverfahren* war, desto niedriger wäre der *Unabhängigkeitsgrad im Rahmen der dazugehörigen Evaluation*. Dabei wird unter Unabhängigkeit die u. a. von Weiss (1998: 38–39) erläuterte *autonomy* verstanden.

Grafik 8 veranschaulicht die möglichen Unabhängigkeitsgrade der Evaluierenden zum Evaluationsgegenstand. Diese Grafik zeigt, dass den höchsten Unabhängigkeitsgrad jene Evaluation aufweist, die ausschließlich durch Externe durchgeführt wird. Diese Form der Fremdevaluation steht in Konkurrenz zur Selbstevaluation. Beide Pole sind in der Evaluations-Community nicht unumstritten und bergen sowohl Vor- als auch Nachteile[10]. Daher folgt in den drei nachfolgenden Unterkapiteln eine kurze Skizzierung beider Evaluationsformen, ein Abgleich mit der Evaluationspraxis des BIBB sowie eine Betrachtung einzuhaltender Rahmenbedingungen sowohl bei Fremd- als auch Selbstevaluationen.

Grafik 8
Unabhängigkeitsgrad zum Evaluationsgegenstand



Exkurs

Manche Autorinnen und Autoren, zu denen u. a. auch Weiss (1998: 37–39), Fitzpatrick et. al. (2004: 23–24) sowie Stockmann (2007: 61–62) zählen, unterscheiden zudem *interne* und *externe* Evaluationen. Dabei würde Position 1 auf dem oben abgebildeten Pfeil genauso zur *internen Evaluation* zählen wie Position 2, da auch hierbei eine Evaluation abteilungsintern durchgeführt wird. Eine *externe Evaluation* wäre demnach nur bei den Positionen 3 und 4 gegeben, wenn die Untersuchung von Personen durchgeführt wird, die nicht der Ordnungsabteilung des BIBB angehören.

2.5.1 Selbstevaluation

Selbstevaluationen, wie in Infokasten 3 beschrieben, werden häufig kritisiert. Diese Kritik bezieht sich laut Müller-Kohlenberg (2004: 71) meist auf die Gefahr eines *Interessenkonflikts*, in dem sich die Evaluierenden befinden können, welche unter Umständen über eine geringere *Methodenkompetenz verfügen oder „betriebsblind" sein könnten*.

[10] Eine entsprechende Gegenüberstellung beider Evaluationsformen findet sich u. a. bei Stockmann (2007: 61) sowie Stockmann (2010: 81).

Selbstevaluationen bieten auf der anderen Seite aber auch Chancen. Diese Chancen beziehen sich meist auf die *hohe Sachkenntnis* der Selbstevaluierenden. Sie kennen den im Mittelpunkt stehenden Untersuchungsgegenstand wie keine andere Person. Eine weitere Chance stellt zudem der *Lerneffekt* dar, den eine Evaluation für die Beteiligten mit sich bringen kann. Müller-Kohlenberg bezeichnet diesen möglichen Lerneffekt auch als *Prozessnutzen,* also das Lernen durch die unmittelbare Durchführung der Evaluation selbst.

Infokasten 3

Selbstevaluation wird gemäß der Definition nach Müller-Kohlenberg (2004: 71) als ein systematisches, datenbasiertes Verfahren der Beschreibung und Bewertung verstanden, bei dem die praxisgestaltenden Akteure identisch sind mit den evaluierenden Akteuren. Diese Akteure überprüfen systematisch ihre eigene Tätigkeit. Aktiv Mitwirkende sind gleichzeitig praxis- und evaluationsverantwortlich.

Die aufgeführten Chancen sind nach Ansicht Müller-Kohlenbergs gleichzeitig auch ein Ziel von Selbstevaluationen: die möglichst *unmittelbare Veränderung und Optimierung der Praxis* sowie die *Weiterqualifizierung der Beteiligten*. Diese Veränderung und Optimierung der Praxis wurde bereits in Kapitel 2.1 als *Zweck der Evaluation* bestimmt. Die selbst abgeleiteten Handlungsempfehlungen können unter bestimmten Voraussetzungen, die in Kapitel 2.5.3 noch näher erläutert werden, zu einer größeren Bereitschaft der Beteiligten führen, den Evaluationsgegenstand sinnvoll weiterzuentwickeln. Zudem wird Selbstevaluationen ein gewisser Schulungseffekt zugeschrieben. So kann beispielsweise der offene Umgang mit Interessenskonflikten, der Austausch über Schwierigkeiten und die gemeinsame Suche nach Lösungen aber auch die intensive Auseinandersetzung mit dem Untersuchungsgegenstand dazu führen, dass sich die Beteiligten fachlich weiterqualifizieren.

In der Praxis können die meisten Evaluationen von Ausbildungsordnungen nicht eindeutig den Positionen 1 bis 4 der Grafik 8 zugeordnet werden. Zum Teil werden Daten selbst erhoben und ausgewertet, zum Teil wird die Datenerhebung und Datenauswertung auch an Externe vergeben. Da die Berichtshoheit immer beim BIBB verbleibt, kann eine Tendenz zu Selbstevaluationen festgestellt werden. Einschränkend ist jedoch anzumerken, dass das BIBB selbst keinen direkten Einfluss auf die unmittelbare Umsetzung der Evaluationsergebnisse hat, da diese Entscheidung beim Weisungsgeber verbleibt. Somit ist ein wichtiges Charakteristikum von Selbstevaluationen bei der Evaluation von Ausbildungsordnungen nicht gegeben.

Es bleibt festzuhalten, dass mit der Kritik sowie den benannten Chancen, die gegenüber Selbstevaluationen geäußert werden, vorausschauend umgegangen werden sollte. Müller-Kohlenberg (2004: 79) schlägt bezüglich der möglichen Interessenskonflikte vor, diese offen und tolerant zu behandeln, damit sie das Selbstevaluationsverfahren und seine Ergebnisse möglichst wenig beeinträchtigen.

Der Prozessnutzen, der als Stärke von Selbstevaluationen angesehen wird, spielt im Handlungsfeld der Ordnungsarbeit eine wichtige Rolle. Die persönliche Inaugenscheinnahme des Untersuchungsgegenstandes, die Gewinnung grundlegender Erkenntnisse und die darüber hinaus stattfindende Weiterqualifizierung sind speziell für die fortlaufende Modernisierung der Ordnungsarbeit von großer Bedeutung.

2.5.2 Fremdevaluation

Im Vergleich zur Selbstevaluation wird bei einer Fremdevaluation der Evaluationsgegenstand von ausschließlich externen Evaluierenden untersucht[11]. Die Fremdevaluation, die geprägt ist durch hohe Unabhängigkeit (vgl. Infokasten 4), weist jedoch ebenso wie die Selbstevaluation sowohl Stärken als auch Schwächen auf. Nach Stockmann (2007: 61–62) zählen zu ihren Stärken neben der *Unabhängigkeit*, eine *große Glaubwürdigkeit* sowie ein *professionelles Evaluationswissen*. Zudem wird externen Evaluierenden häufig unterstellt, dass sie in der Lage seien, für ‚frischen Wind' zu sorgen, wodurch speziell eine konkrete Umsetzung der Handlungsempfehlungen positiv beeinflusst werden kann.

Zu den Schwächen der Fremdevaluation zählt eine *geringere Sachkenntnis* der externen Evaluatoren, zum Beispiel im unmittelbaren Vergleich mit den Berufe-Verantwortlichen. Zudem besteht gelegentlich der Eindruck, dass manche extern Evaluierende eben nicht unabhängig sind und über ein professionelles Evaluationswissen verfügen. Häufig wird davon ausgegangen, Fremdevaluationen seien hinsichtlich ihrer Stärken den Selbstevaluationen klar überlegen und daher immer vorzuziehen. Die Praxis bestätigt diese einfache Schlussfolgerung jedoch nicht grundsätzlich. Es ist festzuhalten, dass die (vermeintlichen) Stärken einer externen Evaluation im Einzelfall gründlich zu prüfen sind.

Infokasten 4

Laut Definition von Beywl/Univation (2004) ist unter einer Fremdevaluation eine Untersuchung zu verstehen, bei der die Evaluierenden gegenüber dem Fach- und Wissensgebiet, zu dem der Evaluationsgegenstand gehört, „fremd" sind. Sie verfügen somit über eine geringere Fach- und Feldkompetenz im Evaluationsfeld. Da sie den Geltungsansprüchen des jeweiligen Evaluationsfeldes weniger verpflichtet sind, fällt es ihnen oft leichter, eine unabhängige Position zu wahren und neue Perspektiven der Beschreibung und Bewertung einzubringen.

2.5.3 Rahmenbedingungen für Evaluationen

Sowohl für Fremd- als auch für Selbstevaluationen gibt es Rahmenbedingungen, die eingehalten werden müssen, damit eine Evaluation erfolgreich verlaufen kann. Da sich die Rahmenbedingungen bei den beiden Herangehensweisen teilweise unterscheiden, wurden von der Gesellschaft für Evaluation DeGEval für beide Herangehensweisen spezifische *Standards für Evaluation* festgeschrieben. Die Funktion dieser Evaluationsstandards besteht u. a. darin, Orientierung bei der Planung und Durchführung von Evaluationen zu geben, vgl. DeGEval (2008a).[12]

2.6 Methodische Aspekte von Evaluationen

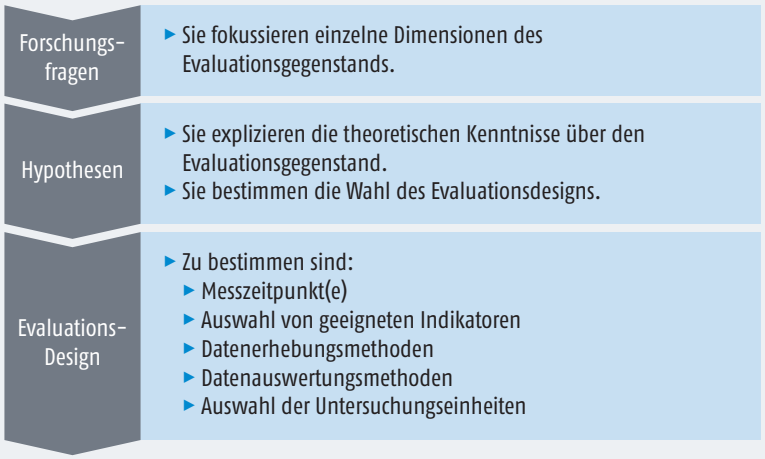
„Empirische Sozialforschung ist im Regelfall dadurch charakterisiert, dass es Patentrezepte, die für alle Fälle gültig sind, nicht gibt. Von einigen Ausnahmen abgesehen (etwa regelmäßig wiederkehrende Routineuntersuchungen in der Markt- und Meinungsforschung) ist der Forscher gezwungen, in gründlicher Auseinandersetzung mit dem Untersuchungsgegenstand immer wieder neu ein für die jeweilige Fragestellung geeignetes Untersuchungsdesign zu entwickeln, d. h. einen speziellen Untersuchungsplan zu entwerfen" (Kromrey 2009: 65).

[11] Stockmann (2007: 61–62) unterscheidet zudem interne und externe Evaluationen, wobei gemäß seiner Kategorisierung die Selbstevaluation genauso zur *internen Evaluation* bzw. „In-house"- Evaluation gezählt wird wie Evaluationen durch institutsinterne Stellen, wie beispielsweise eine institutseigene Evaluations-Stabsstelle. Eine *externe Evaluation* wäre demnach nur dann gegeben, wenn die Untersuchung von Personen durchgeführt wird, die nicht dem Institut angehören.

[12] Besonders lesenswert sind die Erläuterungen zu den *Standards für Evaluation*, die alle 25 Einzelstandards relativ ausführlich behandeln. Das Dokument steht als kostenloses Download unter folgendem Link zur Verfügung: www.alt.degeval.de/calimero/tools/proxy.php?id=19074.

Demnach ist es auch für die Evaluation von Ausbildungsordnungen erforderlich, die Aspekte *Forschungsfragen, Hypothesen* sowie *Evaluationsdesign*[13] jeweils genauer zu betrachten (vgl. Grafik 9) und es gibt folglich diverse Möglichkeiten, wie ein Design entworfen werden kann: „We have learned that the choice of methods (and measures and instruments and data) depends much more on the type of question being asked than on the qualities of any particular method" Chelimsky (1995) zitiert nach Stockmann (2007: 46).

Grafik 9
Aspekte der Evaluationsmethodik



Grundsätzlich erfolgt eine Annäherung an das Design über die Eingrenzung des Evaluationsgegenstands sowie die Formulierung von Fragestellungen, die Einordnung der Problemstellung in vorhandene Kenntnisse sowie die Bildung von Hypothesen. Die folgenden Fragen können dabei hilfreich sein:

- *Welche theoretischen Kenntnisse sind über den Untersuchungsgegenstand sowie über Beziehungen zwischen den Dimensionen vorhanden?*
- *Welche Vermutungen können oder müssen zusätzlich formuliert werden? Vgl.* Kromrey (2009: 71)

Sind keine ausreichenden Vorkenntnisse vorhanden, so sind explorative Vorstudien anzustellen.

2.6.1 Die Triangulation von Methoden

In den meisten Evaluationen von Ausbildungsordnungen kommen sowohl qualitative als auch quantitative Methoden zum Einsatz. Mit dieser sogenannten *Methoden-Triangulation*, also der Anwendung mehrerer Methoden zur Erhebung und Auswertung von Daten, wird darauf reagiert, dass jede Methode Stärken und Schwächen hat. Durch die Triangulation wird zu einer Kombination von sowohl qualitativen als auch quantitativen Methoden ermutigt, mit der ein Maximum an Erkenntnissen hinsichtlich des zu untersuchenden Gegenstands erzielt werden kann. Wichtig dabei ist, neben einer *präzisen Anwendung der* jeweiligen *Methode*, ein *methodenkritischer Auswahlprozess*.

[13] Gemäß Meyer (2007: 163) gibt es keine allgemeinverbindliche Darstellungsform für ein Evaluationsdesign, weshalb in Grafik 10 diese Darstellungsform gesetzt wurde.

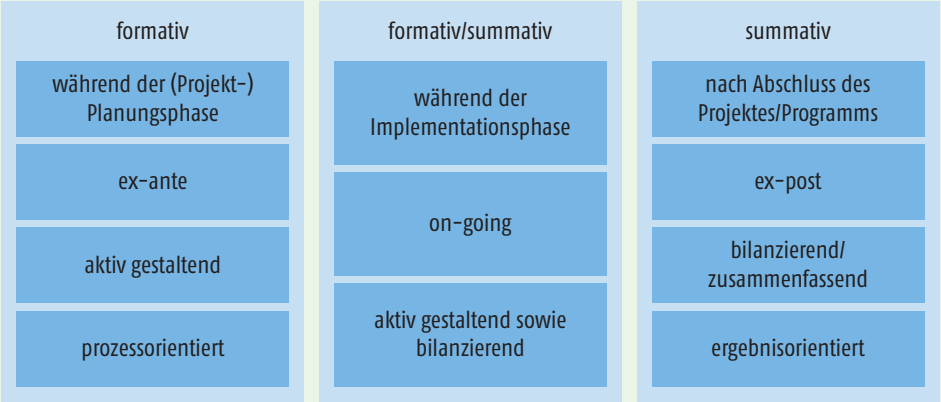
2.6.2 Die Triangulation von Daten

Bei der sogenannten *Daten-Triangulation* werden unterschiedliche *Datenquellen* genutzt. Dabei findet eine gezielte sowie systematische Auswahl und Einbeziehung von Personen/Untersuchungsgruppen, Zeitpunkten und lokalen Settings (z. B. eine Fallstudie in einem speziellen Betrieb oder die Wahl bestimmter Kammerbezirke) statt. Bei der Evaluation von Ausbildungsordnungen geschieht dies beispielsweise indem neben den Statistiken des Datenzentrums der Bundesagentur für Arbeit oder von DESTATIS (Sekundärdaten) auch Primärdaten erhoben werden in Betrieben, Berufsschulen, bei zuständigen Stellen oder auch bei Verbänden. Es ist im Zuge einer Evaluationskonzeptionierung notwendig, die Gründe der Methodenwahl kritisch zu reflektieren. Dieser methodenkritische Auswahlprozess sollte im Sinne sowohl einer Methoden- als auch Daten-Triangulation stattfinden, da qualitative und quantitative Methoden sowie unterschiedliche Untersuchungsgruppen oft Informationen liefern, die bestehende Zusammenhänge erklären oder fehlerhafte Interpretationen von Befunden korrigieren können, vgl. Kelle (2008: 261)

Exkurs

Die Festlegung der Messzeitpunkte darf nicht verwechselt werden mit der grundsätzlichen Terminierung einer Evaluation. Dabei werden die folgenden drei Perspektiven unterschieden (vgl. Grafik 10):

Grafik 10
Evaluationskonzepte[14]



Es ist entscheidend, ob das Projekt oder Programm als abgeschlossen angesehen werden darf oder als in der Implementationsphase befindlich. Wird davon ausgegangen, dass sich zum Beispiel eine Erprobungsverordnung noch in der Implementationsphase und eine seit Jahren umgesetzte Ausbildungsordnung bereits in der ex-post-Phase befinden, so ist eine eindeutige Zuordnung möglich. Die Evaluation einer Ausbildungsordnung soll zum Beispiel in Erfahrung bringen, ob die Ausbildungspraxis mit der derzeitigen Ausbildungsordnung zufrieden ist. Sie ist daher eindeutig ergebnisorientiert.

Würde eine Ausbildungsordnung hingegen ebenfalls als noch veränderbar angesehen werden, beispielsweise durch ein sich anschließendes Neuordnungsverfahren, so wäre ihre Evaluation eher als on-going Evaluation zu werten, bei der zeitgleich sowohl eine formative als auch summative Perspektive eingenommen wird.

Hinsichtlich des Messzeitpunktes wird bei der Evaluation von Ausbildungsordnungen in der Regel eine sogenannte *Querschnittbetrachtung* vorgenommen, d. h. die befragten Personen können nur zu einem einzigen Zeitpunkt befragt werden. Prozessbeobachtungen bzw. *Längsschnittbe-*

[14] Eigene Darstellung in Anlehnung an Balzer (2005: 196) sowie Stockmann (2007: 34).

trachtungen erfolgen in diesem Zusammenhang kaum, da diese erheblichen personellen und finanziellen Aufwand erfordern würden.

2.7 Berichtswesen

Das Verfassen von Berichten ist eine Kernaufgabe der Evaluierenden. Evaluatoren sollten sich immer vergegenwärtigen, dass die Nützlichkeit dessen, was sie schreiben, von großer Bedeutung ist. Cronbach[15] weist in diesem Zusammenhang darauf hin: „(…) that one important role of the evaluator is to illuminate, not to dictate, the decision. Helping clients to understand the complexity of issues, not to give simple answers to narrow questions, is a role of evaluation."

Der Bericht ist zudem das zentrale Dokument, das auch nach dem Ende der Evaluation bestehen bleibt. Auf ihn wird im Zuge von sich anschließenden Verfahren zurückgegriffen. Daher sollte er unbedingt auch den durch die DeGEval festgeschriebenen Nützlichkeitsstandards entsprechen.

2.8 Mögliche Evaluationsansätze

Wie in Kapitel 2.4 dargestellt wurde, können die Akteure der ersten, zweiten und dritten Ebene des Organisationsmodells grundsätzlich bei der Ausgestaltung und Durchführung einer Evaluation unterschiedlich eingebunden werden. In welcher Form diese Einbindung stattfindet, hängt

Grafik 11
Evaluationsansätze[16]

Managementorientiert	▸ Ziel: Entscheidungsträger/-innen sollen mit Informationen über den Programm- bzw. Projektverlauf sowie Inputs, Outputs, Outcomes und deren Verhältnis zueinander versorgt werden. ▸ Vertreter dieses Ansatzes: D. Stufflebeam, M. Alkin, M. Patton
Zielorientiert	▸ Ziel: Überprüfung von spezifizierten Programm- bzw. Projekt-Zielen hinsichtlich ihres Erreichungsgrades. ▸ Vertreter dieses Ansatzes: R.W. Tyler
Partizipativ	▸ Ziel: unmittelbare Einbeziehung der Programm- bzw. Projekt-Beteiligten bei der Planung und Durchführung der Evaluation. ▸ Vertreter dieses Ansatzes: R. Stake, E. Guba, Y. Lincoln, M. Patton, D. Fetterman, D. Mertens
Expertenorientiert	▸ Ziel: Beurteilung von Programmen bzw. Projekten auf Basis professioneller Expertisen (Review- oder Akkreditierungsverfahren). ▸ Vertreter dieses Ansatzes: E.W. Eisner
Konsumentenorientiert	▸ Ziel: Information potenzieller Konsumenten/Konsumentinnen über bestimmte Qualitätsaspekte von Programmen bzw. Produkten. ▸ Vertreter dieses Ansatzes: M. Scriven

u. a. von dem gewählten Evaluationsansatz ab. Wie beispielsweise Balzer (2005: 23–66) und Stockmann/Meyer (2010: 101) ausführen, gibt es eine Vielfalt grundlegender Evaluationsan-

[15] Zitiert nach Fitzpatrick (2004: 96).
[16] In Anlehnung an Stockmann (2007: 48–49). Er hat die Ansätze in Anlehnung an Fitzpatrick (2004) übersetzt.

sätze. Für diese Arbeit soll auf die Systematisierung der Ansätze nach Fitzpatrick/Sanders/Worthen (2004) zurückgegriffen werden.[17] Nach Ansicht Fitzpatricks (2004: 68 ff.) gibt es insgesamt fünf verschiedene Evaluationsansätze, denen sich Evaluationsmodelle unterschiedlicher Autoren und Autorinnen zuordnen lassen[18] (vgl. Grafik 11). Diese fünf Ansätze gehen jeweils von einem anderen Ziel/Zweckgedanken aus und sehen zum Teil auch eine unterschiedliche Stakeholder-Einbindung vor. Im Folgenden werden diese fünf Ansätze einzeln vorgestellt und in einem Theorie-Praxis-Vergleich den bisherigen Evaluationen von Ausbildungsordnungen gegenübergestellt.

2.8.1 Der managementorientierte Ansatz

Das Ziel des managementorientierten Ansatzes ist es, Entscheidungsträgerinnen und Entscheidungsträger mit Informationen über den Programm- bzw. Projektverlauf sowie über Inputs, Outputs, Outcomes und deren Verhältnis zueinander zu versorgen, vgl. Stockmann (2007: 48). "The decision maker is the audience to whom a management-oriented evaluation is directed, and the decision maker's concerns, informational needs, and criteria for effectiveness guide the direction of the study" Fitzpatrick (2004: 88). Nach Ansicht von Fitzpatrick (2004: 95–97) und Stockmann (2007: 48–49) können die *Stärken* und *Schwächen* des managementorientierten Ansatzes wie folgt zusammengefasst werden (vgl. Grafik 12).

Grafik 12
Stärken/Schwächen des managementorientierten Ansatzes[19]

Stärken des managementorientierten Ansatzes
▸ Klare Fokussierung auf Informationsbedürfnisse von Entscheidungsträgern.
▸ Zeitgerechte Bereitstellung von Informationen.
▸ Nützlichkeitsaspekte der Ergebnisse stehen im Vordergrund.
▸ Evaluation aller Programmkomponenten zu verschiedenen Entwicklungsphasen.
▸ Programmentwicklung soll durch die Ergebnisse möglich sein.
Stärken des managementorientierten Ansatzes
▸ Gefahr des Tunnelblicks: Die Evaluationsfragestellung wird durch die Interessen der Entscheidungsträgerinnen und Entscheidungsträger dominiert.
▸ Gefahr der Vereinnahmung der Evaluatorinnen und Evaluatoren durch den Auftraggeber: die Interessen der Stakeholder rücken in den Hintergrund.
▸ Implizite Annahme, dass Programmentscheidungen jeweils im Voraus eindeutig definiert werden können und damit klare Entscheidungsalternativen spezifiziert werden können.

Evaluationen von Ausbildungsordnungen entsprechen meist dem managementorientierten Ansatz. Eine *Fokussierung auf die Informationsbedürfnisse der Entscheidungsträgerinnen und Entscheidungsträger* (Weisungsgeber), sowie eigenes Forschungsinteresse sind häufig an vorge-

[17] Die Systematisierung von Fitzpatrick/Sanders/Worthen wurde gewählt, da sie sich durch gute Nachvollziehbarkeit auszeichnet und für die Kategorisierung von Evaluationen von Ausbildungsordnungen geeigneter erscheint als beispielsweise das Baummodell von Alkin/Christie (2004) zitiert nach Stockmann/Meyer (2010: 113).
[18] Grundsätzlich sollte diese Kategorisierung von Modellen nicht als absolut betrachtet werden. Vielmehr gibt es auch Evaluationsmodelle, die Versatzteile mehrerer Ansätze miteinander kombinieren, vgl. Fitzpatrick (2004: 68).
[19] Eigene Darstellung in Anlehnung an Stockmann (2007: 48–49).

gebenen Forschungsfragestellungen (aus Weisungen) ausgerichtet. Nachteilig bei dieser Vorgehensweise kann eine *Verengung allein auf die Interessen der Entscheidungsträgerinnen* und *Entscheidungsträger* sein[20].

Exkurs

Die hier und in den nachfolgenden Unterkapiteln aufgeführten Stärken und Schwächen der jeweiligen Ansätze wurden von Fitzpatrick (2004) und Stockmann (2007) festgehalten. Sie beziehen sich auf Evaluationsmodelle, die eben diesen Ansätzen zugeordnet werden können. Diese Stärken- Schwächen-Zuweisung sollte nicht als starr angesehen werden. Vielmehr soll die Diskussion der benannten Stärken und Schwächen den eigenen Blick schärfen. Darüber hinaus kann die Kenntnis über mögliche Stärken und Schwächen bei der Entwicklung eines eigenen Modells bzw. eines Designs hilfreich sein und im Folgeprozess die Evaluierenden darin unterstützen, möglichen Schwierigkeiten und Herausforderungen aktiv begegnen zu können.

2.8.2 Der zielorientierte Ansatz

Der zielorientierte Evaluationsansatz bezieht sich explizit auf die Beantwortung der Frage nach der Erreichung spezifizierter Ziele eines Programms oder Projekts[21]. Das Vorgehen im Rahmen einer solchen zielorientierten Evaluation ist geprägt durch Arbeitsschritte (vgl. Grafik 13), die Ralph W. Tyler (1935, 1950) beschrieben hat, der als prominentester Vertreter des zielorientierten Ansatzes gilt und selbst aus der Curriculumforschung stammt[22].

Grafik 13
Arbeitsschritte einer zielorientierten Evaluation[23]

1. Formulierung von allgemeinen (Bildungs-) Zielen
2. Klassifizierung dieser Ziele
3. Beschreibung dieser Ziele durch Verhaltenskategorien
4. Identifikation von Situationen, in denen Personen dieses Verhalten zeigen können
5. Auswahl sowie ggf. Entwicklung von Messmethoden
6. Datensammlung
7. Vergleich der Ergebnisse mit den zuvor definierten Zielen

In der Regel werden die Ergebnisse, die nach Bearbeitung aller sieben Arbeitsschritte im Raum stehen, zur Optimierung des Projekts genutzt, sodass eine bessere Zielerreichung möglich werden kann.

Zu den *Stärken* und *Schwächen* des Ansatzes zählen laut Fitzpatrick (2004: 82–84) und Stockmann (2007: 48–49) die folgenden Aspekte (vgl. Grafik 14):

[20] Fitzpatrick (2004: 96) erläutert diese Schwäche des managementorientierten Ansatzes wie folgt: "A potential weakness of this approach is the evaluator's occasional inability to respond to questions or issues that may be significant – even critical – but that clash with or at least do not match the concerns and questions of the decision maker who, essentially, controls the evaluation."
[21] Zur sprachlichen Vereinfachung wird nachfolgend ausschließlich von Projekten als Synonym für alle erdenklichen Evaluationsgegenstände gesprochen.
[22] Für Tyler waren Curricula notwendig, „(...) um spezifische Lernziele zu organisieren und den Unterricht zu planen. Die Untersuchung der operationalisierten Lernziele sollte die Bewertung der Qualität des amerikanischen Erziehungssystems unterstützen" Balzer (2005: 29).
[23] Eigene Darstellung in Anlehnung an Balzer (2005: 29) sowie Fitzpatrick (2004: 72).

Grafik 14
Stärken/Schwächen des zielorientierten Ansatzes[24]

Stärken des zielorientierten Ansatzes

- Einfach und leistungsfähig.
- Leicht legitimierbar.
- Bringt Projektverantwortliche dazu, Projektziele zu spezifizieren und zu reflektieren.

Schwächen des zielorientierten Ansatzes

- Projektziele sind oft verschwommen.
- Kluft zwischen offiziellen und tatsächlichen Zielen.
- Zielveränderungen.
- Unterschiedliche Akteure mit unterschiedlichen Zielen.
- Fehlende Standards zur Beurteilung der Relevanz von beobachteten Diskrepanzen zwischen Soll und Ist.
- Auslassen von Informationen über den Wert eines Projekts/Programms, die nicht in dessen Zielsetzung wiedergegeben werden.

Fitzpatrick und Stockmann untermauern mit den folgenden Fragen die Schwächen dieses Ansatzes sehr deutlich: "Who really determines the goals and objectives? Do they include all important outcomes? Have all those affected by the program agreed on these particular goals or objectives? Who has determined that a particular criterion level is more defensible than alternatives? On what evidence? These and other questions must be addressed if an objectives-oriented approach is to be defensible."

2.8.3 Der partizipative Ansatz

Das Ziel des partizipativen Evaluationsansatzes ist es, die Programm- bzw. Projekt-Beteiligten und Betroffenen bei der Planung und Durchführung der Evaluation unmittelbar mit einzubeziehen. In welchem Ausmaß diese Einbeziehung vorgenommen wird, hängt in erster Linie vom Partizipationsverständnis ab, das in den Modellen, welche diesem Ansatz zugeordnet werden, recht unterschiedlich gehandhabt wird. Egon G. Guba und Yvonne Lincoln zählen beispielsweise zu den prominentesten Vertretern und Vertreterinnen dieses Ansatzes.[25] Die ihrem Modell zugrunde liegende, als konstruktivistisch bezeichnete Erkenntnistheorie, „(…) bedeutet, dass Erkenntnis immer auf der Basis der Erfahrungen und Interpretationen eines einzelnen Menschen gebildet wird und damit subjektiv ist. Zur Mehrung von Erkenntnissen ist ein Dialog mit anderen erforderlich, ein endgültiges ‚richtig' oder ‚falsch' ist nicht erzielbar. Der Idee der responsiven Evaluation folgend sind es darüber hinaus die Bedürfnisse der Beteiligten und Betroffenen, die den Evaluationsprozess steuern" Balzer (2005: 51).

Aus diesen Überlegungen heraus entwickelten Guba und Lincoln ein Modell, das ein übergreifendes Verständnis und eine besondere Bezugnahme auf die menschlichen, politischen, sozialen, kulturellen und kontextuellen Elemente des Evaluationsprozesses herzustellen versucht.

[24] Eigene Darstellung in Anlehnung an Stockmann (2007: 48–49).
[25] Neben Guba und Lincoln (1981, 1989) zählen Robert Stake (1978), mit seinem Modell der „Responsiven Evaluation", David Fetterman (1994, 200, 2004), mit seinem „Empowerment Ansatz", aber auch Michael Bamberger et. al. (2006), mit ihrem „Real World Evaluation"-Modell und Melvin M. Mark (2000) mit ihrem „Betterment-Driven" Modell, zu den prominenten Vertretern bzw. Vertreterinnen des partizipativen Evaluationsansatzes. Alle Autoren und Autorinnen sind jeweils zitiert nach Stockmann/Meyer (2010: 107–141).

Dies geschieht, indem die verschiedenen Betroffenen und Beteiligten mit ihren unterschiedlichen Interessen von den Evaluierenden in einen demokratischen Aushandlungsprozess einbezogen werden, an dessen Ende die Einigung über die Interpretation der erhobenen Daten zu einer besonderen Relevanz der Resultate führt. Auf diese Weise avancieren die Evaluierenden zu Moderierenden eines offenen Prozesses, vgl. Stockmann (2007: 45).

Ein weit weniger ‚radikales' Modell findet sich bei Michael Bamberger (2007: 113 ff. und 373–376), bei dem sich die Partizipation ausschließlich auf die Einbindung der Projektverantwortlichen bezieht und nicht auch auf die Zielgruppen. Dieses „Real World Evaluation"-Modell legt zwar großen Wert auf die Analyse des Umfelds und der Bedürfnisse aller Stakeholder, konzentriert sich jedoch hinsichtlich der Beteiligung am Evaluationsprozess ausschließlich auf den Auftraggeber.

Zu den *Stärken und Schwächen*, die mit diesem partizipativen Vorgehen verbunden sind, zählen laut Fitzpatrick (2004: 146–149) und Stockmann (2007: 48–49) (vgl. Grafik 15):

Grafik 15
Stärken/Schwächen des partizipativen Ansatzes[26]

Stärken des partizipativen Ansatzes

- Klare Fokussierung auf die Bedürfnisse derjenigen, die von der Evaluation profitieren sollen.
- Vielfalt der verschiedenen Interessen und Perspektiven werden bewusst erfasst.
- Der Kontext des Evaluationsgegenstandes steht im Mittelpunkt.
- Kapazitätsbildung bei den Evaluationsbeteiligten (Stakeholdern).

Schwächen des partizipativen Ansatzes

- In der Praxis schwer umzusetzen wegen der Gefahr einer vereinfachten unreflektierten Anwendung.
- Die Objektivität der Evaluatoren/Evaluatorinnen ist kritisch zu sehen.
- Übertragung der Bewertungskomponenten vom Evaluator auf die Evaluationsbeteiligten bedeutet Abkehr vom eigentlichen Konzept der Evaluation.
- Sehr zeit- und kostenintensiv.

2.8.4 Der konsumentenorientierte und der expertenorientierte Ansatz

Ziel des konsumentenorientierten Ansatzes ist es, potenzielle Konsumenten und Konsumentinnen über verschiedene Qualitätsaspekte von Produkten oder Dienstleistungen zu informieren. Dafür werden u. a. Evaluationschecklisten verwendet. Ein recht prominentes Beispiel sind die Tests der Stiftung Warentest, denen dieser Evaluationsansatz zugrunde liegt.

Hinter dem expertenorientierten Ansatz verbergen sich in der Regel Review- oder Akkreditierungsverfahren. Bewertungen des Evaluationsgegenstands werden dabei ausschließlich durch Expertinnen und Experten vorgenommen, die sich an veröffentlichten Standards orientieren.

Die genaue Beschreibung beider Ansätze sowie deren Vor- und Nachteile sind bei Fitzpatrick (2004: 100–128), Stockmann (2007: 48–49) sowie Stockmann/Meyer (2010: 131–138) nachzulesen.

[26] Vgl. Stockmann (2007: 48–49).

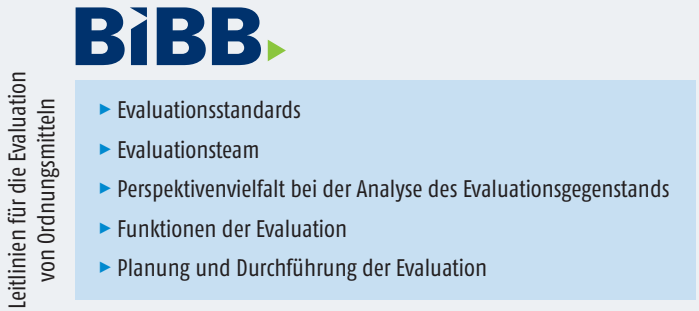
3 Leitlinien für die Evaluation von Ordnungsmitteln

Die Evaluation von Ausbildungsordnungen unterliegt Besonderheiten, welche einem „klassischen Methodenansatz“ nicht immer gerecht werden können. So sind zum Beispiel die Validierung eines Fragebogens oder ein Pre-Test nicht immer möglich, da es oft nur einen einzigen, nicht wiederholbaren Messzeitpunkt (zum Beispiel im Rahmen einer Abschlussprüfung) gibt. Daher muss im Bereich der Evaluation von Ausbildungsordnungen ein Methodenmix eingesetzt werden, der solchen möglichen Besonderheiten gerecht wird.

Nachfolgend sollen daher Leitlinien für die Evaluation von Ordnungsmitteln vorgestellt werden. Diese Leitlinien können grundsätzlich dem *managementorientierten Evaluationsansatz* zugeordnet werden. Zudem greifen diese Leitlinien Aspekte auf, die sich im *zielorientierten Ansatz* wiederfinden lassen.

Die Leitlinien zeichnen sich durch die in der nachfolgenden Grafik 16 dargestellten fünf Eckpunkte aus. In den folgenden Kapiteln werden diese Eckpunkte näher erläutert.

Grafik 16
Leitlinien für die Evaluation von Ordnungsmitteln



3.1 Evaluationsstandards

Neben den Maßgaben empirischer Sozialforschung – hinterlegt in der BIBB-Institutsanweisung „Regeln zur Sicherung guter wissenschaftlicher Praxis“ – sind für Evaluationen im Ordnungsbereich des BIBB die *Standards für Evaluation* der DeGEval – Gesellschaft für Evaluation e. V. (2014) relevant.

Evaluation wird im Rahmen dieser beiden Vorgaben als *anwendungsorientierte Sozialforschung* verstanden, die sich sozialwissenschaftlicher Forschungsmethoden bedient. Ihr liegen folgende Kriterien zugrunde, die zeitgleich grundlegende Standards von Wissenschaftlichkeit sind:

- Systematik
- Nachvollziehbarkeit des Erkenntnisweges (Transparenz)
- Methodische Kontrolle
- Kritische Reflexion

Ergänzt werden diese Kriterien um folgende Standards:

- Nützlichkeit (utility)
- Durchführbarkeit (feasibility)
- Genauigkeit (accuracy)
- Korrektheit/Fairness (priority).

Diese Standards basieren auf den ursprünglich 1986 durch die American Evaluation Association (AEA) verfassten *Program Evaluation Standards* und wurden von der DeGEval sowie der befreundeten Schweizerischen Evaluationsgesellschaft (SEVAL) fast vollständig übernommen. Das BIBB, als Mitglied der DeGEval, fühlt sich zur Einhaltung dieser Standards verpflichtet.

Für den gesamten Prozess der Planung und Durchführung einer Evaluation bieten die o. g. Kriterien und Standards nicht nur Orientierung, sondern geben gleichzeitig auch einen Handlungsspielraum vor. „Evaluationen müssen nützlich sein, d. h. dem Informationsbedarf der vorgesehenen Nutzer/-innen dienen. Sie müssen durchführbar sein, d. h. realistisch, durchdacht, diplomatisch und kostenbewusst geplant und ausgeführt werden. Evaluationen sollen respektvoll und fair mit den betroffenen Personen und Gruppen umgehen und wissenschaftlichen Ansprüchen genügen“ (Stockmann 2007: 6).

Die vier o. g. Evaluations-Standards gliedern sich in insgesamt 25 einzelne Standards, die auf den Internetseiten der DeGEval unter http://www.degeval.de/de/degeval-standards/standards/ (Stand 1/2014) nachzulesen sind. Sie beschreiben den Evaluationsprozess von der Entscheidung über die Durchführung bis hin zur personellen Ausstattung einer Evaluation.

Als wichtige Ergänzung sind zudem die *Empfehlungen zur Anwendung der Standards im Handlungsfeld der ‚Selbstevaluation‘* der DeGEval (2004) zu berücksichtigen. Sie gehen zusätzlich zu den vier Evaluationsstandards speziell auf die Besonderheiten ein, die im Rahmen einer Selbstevaluation zu berücksichtigen sind (vgl. Kapitel 2.5.1).

Das Vorhaben, die eigene durchzuführende Evaluation mit allen 25 Einzel-Standards in Einklang zu bringen, stellt eine große Herausforderung für die Evaluatorinnen und Evaluatoren dar.

Bereits 1999 hielten Beywl/Widmer, die die Evaluationsstandards der American Evaluation Association ins Deutsche übersetzten, diesbezüglich fest, dass die mit den Standards geforderten Anforderungen sehr anspruchsvoll sind und teilweise auch miteinander konkurrieren. „Eine unter realen Bedingungen stattfindende Evaluation vermag demzufolge kaum alle Standards gleichzeitig in vollem Umfang zu erfüllen“ (Sanders 1999: 8).

Vor diesem Hintergrund kann auch für Evaluationen im Ordnungsbereich festgehalten werden, dass diese Standards grundsätzlich berücksichtigt werden müssen und entsprechend der Rahmenbedingungen gewissenhaft abzuwägen sind. Gegebenenfalls kann es durch die (zum Beispiel personellen oder zeitlichen) Umstände einer Untersuchung zu erzwungenen Abweichungen kommen. Dann müssen diese Abweichungen in einem Evaluationsbericht explizit benannt werden.

3.2 Die Zusammensetzung des Evaluationsteams

Ein Evaluationsteam sollte möglichst ausgewogen besetzt sein. Es sollte neben einer Berufe-Expertin bzw. einem Berufe-Experten – passend zum untersuchten Beruf – mindestens mit einer Evaluationsexpertin bzw. einem Evaluationsexperten besetzt sein. Dieses „Tandemmodell“ soll dazu beitragen, dass (wichtige) Entscheidungen im Laufe der Evaluation gemeinsam diskutiert und getroffen werden, wodurch u. a. auch mögliche Interessenskonflikte, alternative methodische Vorgehensweisen etc. gemeinsam reflektiert werden können. Im Falle der Vergabe von Eva-

luationsaufgaben an externe Evaluierende kann darüber nachgedacht werden, das Tandem aus einem/einer internen Berufe-Experten/-Expertin und einem/einer externen Evaluations-Experten/-Expertin zusammenzusetzen.

Welche *Evaluationskenntnisse* eine Evaluationsexpertin oder ein Evaluationsexperte mitbringen sollte, ist u. a. nachzulesen in den *„Empfehlungen für die Aus- und Weiterbildung in der Evaluation. Anforderungsprofile an Evaluatorinnen und Evaluatoren“* der DeGEval (2008b). In diesen Empfehlungen werden sog. Kompetenzfelder definiert, die neben Kenntnissen zu Theorie und Geschichte der Evaluation, Methodenkompetenzen, Organisations- und Feldkenntnissen auch die Sozial- und Selbstkompetenzen umfassen. Nachfolgend soll, im Hinblickauf die Datenerhebung und Datenauswertung, der *Methodenkompetenz* die größte Aufmerksamkeit gewidmet werden. Sie wird den o. g. Empfehlungen folgend durch fünf Dimensionen bestimmt (vgl. Grafik 17).

Grafik 17
Dimensionen der Methodenkompetenz[27]

Grundzüge empirischer Sozialforschung u. Untersuchungsdesign	▸ Wissenschaftliche Grundlagen ▸ Entwicklung und Operationalisierung von Fragestellungen ▸ Planung empirischer Untersuchungen ▸ Auswahl von Messverfahren
Datenerhebung	▸ Grundlagen über Erhebungsformen ▸ Entwicklung von Erhebungsinstrumenten
statistische Kenntnisse	▸ Univariate Häufigkeitsauszählungen ▸ Kreuztabellierung ▸ Varianzanalyse ▸ Verfahren zur Messung von Zusammenhängen, Signifikanztests
Datenverarbeitung Datenaufbereitung Dateninterpretation	▸ Anwendungskenntnisse in relevanten Softwarepaketen und quantitativen sowie qualitativen Datenanalysen ▸ Kodierung und Rekodierung ▸ Dateninterpretation und Reporting
Kenntnisse der Projektorganisation	▸ Zeitplanung, Durchführungsplanung und -kontrolle ▸ Kostenplanung und -kontrolle

Über die *Organisations- und Feldkenntnisse*, die ebenfalls als ein Kompetenzfeld von der DeGEval festgelegt wurden, müssen im Falle der Evaluation von Ausbildungsordnungen vor allem die Berufe-Expertinnen und Berufe-Experten verfügen.

3.3 Perspektivenvielfalt bei der Analyse des Evaluationsgegenstands

Im Rahmen von Evaluationen im Zusammenhang mit der Ordnungsarbeit sind gemäß den Ausführungen in Kapitel 2.2 die vier Gegenstandsdimensionen *Ordnungsrahmen, Kontext, Weisungsgeber-Weisungsnehmer, Sozialpartner* und *Zielgruppen* zu berücksichtigen. Diese umfassende

[27] Eigene Darstellung in Anlehnung an DeGEval (2008b: 19).

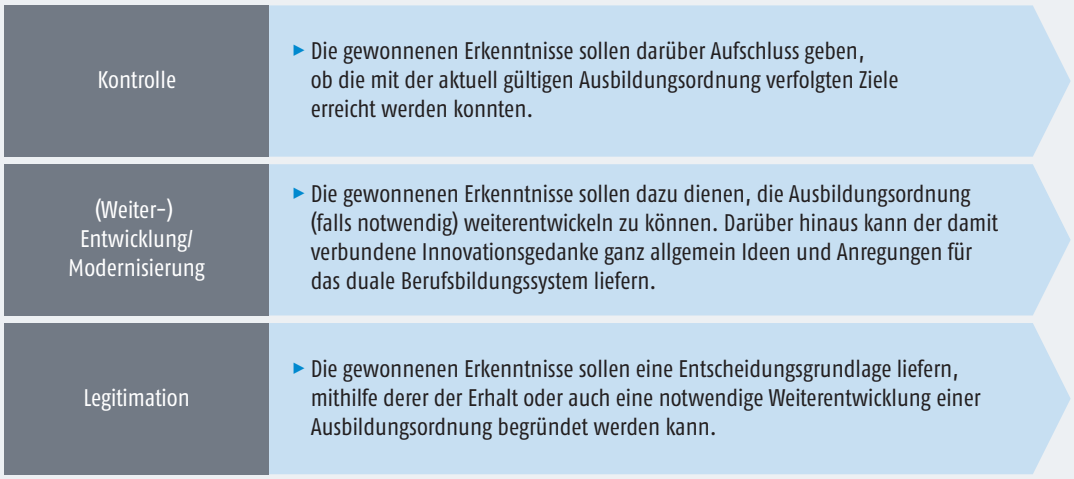
Betrachtung soll sicherstellen, dass die Analyse des Evaluationsgegenstands nicht zu fokussiert verläuft.

Mit der Perspektivenvielfalt ist darüber hinaus auch die Perspektivenvielfalt bei der Datenerhebung und -auswertung gemeint. Im Sinne einer *Methodentriangulation* sollen die Stärken und Möglichkeiten sowohl der quantitativen als auch der qualitativen Methoden genutzt werden. Im Sinne einer *Datentriangulation* sollen Daten von Informationsträgern erhoben werden, die jeweils unterschiedlichen Gruppen angehören.

3.4 Die Funktionen einer Evaluation im Ordnungsbereich

Evaluationen im Ordnungsbereich sollen zum Zweck der Kontrolle, der (Weiter-) Entwicklung sowie der Legitimation durchgeführt werden (vgl. Grafik 18). Daher ist es wichtig, dass die Evaluatorinnen und Evaluatoren die erhobenen Daten im Abschlussbericht nicht nur *darstellen*, sondern auch *interpretieren* und demzufolge *Bewertungen vornehmen*, aus denen sich Handlungsempfehlungen ableiten lassen. Auf welche Bewertungskriterien in diesem Zusammenhang zurückgegriffen wird, ist offenzulegen. Auf welche Weise diese Bewertungskriterien festgelegt wurden, beispielsweise in Absprache mit dem Weisungsgeber, ist im abschließenden Evaluationsbericht zu erwähnen.

Grafik 18
Funktionen der Evaluation



3.5 Planung und Durchführung der Evaluation

Gemäß Kromrey (2006: 65) gibt es kein Patentrezept für ein Untersuchungsdesign, das immer als angemessen erachtet werden kann. Vielmehr muss für jeden Fall neu entschieden werden, wie die Evaluation zu gestalten ist. Dennoch können für die Evaluation einer Ausbildungsordnung *zentrale Arbeitsschritte* identifiziert werden, die (fast) immer wiederkehren (zum Beispiel die Entwicklung eines geeigneten Evaluationsdesigns oder einer realistischen Zeitplanung). Bei näherer Betrachtung dieser zentralen Arbeitsschritte, die im Laufe einer Evaluation zu berücksichtigten sind, wird deutlich, dass diese unterschiedlich komplex sind und zum Teil umfangreiche Erläuterungen erfordern. Dazu wurden einige Arbeitshilfen entwickelt, die den Evaluationsprozess unterstützen sollen und im nachfolgenden Kapitel vorgestellt werden. Diese Arbeitshilfen greifen einige ausgewählte Themenfelder kurz auf, stellen aber keinen Ersatz für die jeweils notwendige Fachliteratur dar.

4 Arbeitshilfen zur Umsetzung der Evaluationsleitlinien

► Arbeitshilfe: Qualitätskriterien, Gütekriterien und Untersuchungsdesign

Evaluationen mit einem Multimethoden-Ansatz zeichnen sich dadurch aus, dass sie sich sowohl qualitativer als auch quantitativer Erhebungs- und Auswertungsmethoden bedienen. Welche Unterschiede es dabei hinsichtlich der Forschungslogik sowie der Gütekriterien zu berücksichtigen gilt, erläutert diese Arbeitshilfe.

► Arbeitshilfe: Stichprobenauswahl

Sowohl bei quantitativen als auch bei qualitativen Methoden müssen gezielt Quellen bzw. Datenträger ausgewählt werden, von denen Daten erhoben werden können. Diese Arbeitshilfe geht gezielt darauf ein, was es im Rahmen dieser Auswahlprozesse zu berücksichtigen gilt, wie dieser Auswahlprozess gestaltet werden kann und welche Stichprobengrößen im Falle von Teilerhebungen gewählt werden sollten.

► Arbeitshilfe: Erhebung quantitativer Daten

Die Erhebung rein quantitativer Daten ist neben der Erhebung qualitativer Daten zentraler Bestandteil fast jeder Evaluation. Welche Methoden es gibt und was bei ihrem Einsatz zu berücksichtigen ist, soll diese Arbeitshilfe erörtern.

► Arbeitshilfe: Skalenniveaus und Auswertung quantitativer Daten

Beim Einsatz quantitativer Methoden ist die Wahl des Skalenniveaus mit entscheidend dafür, welche Auswertungsverfahren überhaupt angewendet werden dürfen. Die Arbeitshilfe möchte Aspekte beleuchten, die es bei der Auswertung und Bewertung von quantitativ erhobenen Daten zu berücksichtigen gilt.

► Arbeitshilfe: Erhebung qualitativer Daten

Die Erhebung qualitativer Daten unterscheidet sich teilweise erheblich von der quantitativen Datenerhebung. Welche qualitativen Erhebungsmethoden im Rahmen von AUSBILDUNGSORDNUNG-Evaluationen hilfreich sein können und was bei ihrer Anwendung zu beachten ist, soll diese Arbeitshilfe darstellen.

Literatur

Balzer, Lars: Wie werden Evaluationen erfolgreich? Landau 2005.

Bamberger, Michael; Rugh, Jim; Mabry, Linda: Real World Evaluation. Working under budget, time, data and political constraints. Thousand Oaks 2006.

Beywl, Wolfgang: Evaluationsmodelle und qualitative Methoden. In: Flick, Uwe: Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg 2006, S. 92–116.

Brandt, Tasso: Sozialer Kontext der Evaluation. In: Stockmann, Reinhard: Handbuch zur Evaluation. Eine praktische Handlungsanleitung. Münster 2007, S. 164–194.

Bundesinstitut für Berufsbildung (Hrsg.): Ausbildungsordnungen und wie sie entstehen. Bonn 2011.

Deutsche Gesellschaft für Evaluation (Hrsg.): Empfehlungen für Aus- und Weiterbildung in der Evaluation. Anforderungsprofile für Evaluatorinnen und Evaluatoren. Mainz 2008.

Deutsche Gesellschaft für Evaluation (Hrsg.): Empfehlungen zur Anwendung der Standards für Evaluation im Handlungsfeld der Selbstevaluation. Alfter 2004.

Deutsche Gesellschaft für Evaluation (Hrsg.): Standards für Evaluation. Mainz 2014.

Fitzpatrick, Jody L.; Sanders, James R.; Worthen, Blaine R.: Program Evaluation. Alternative approaches and practical guidelines. White Plains 1997.

Fitzpatrick, Jody L.; Sanders, James R.; Worthen, Blaine R.: Program Evaluation. Alternative approaches and practical guidelines. 3rd edition. Boston 2004.

Kelle, Udo: Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte. Wiesbaden 2008.

Krämer, Heike: Evaluation Mediengestalter/Mediengestalterin für Digial- und Printmedien. Ergebnisse und Ausblick. Bielefeld 2004.

Kromrey, Helmut: Empirische Sozialforschung. Stuttgart 2006.

Meyer, Wolfgang: Datenerhebung: Befragungen – Beobachtungen – Nicht-reaktive Verfahren. In: Stockmann, Reinhard: Handbuch zur Evaluation. Eine praktische Handlungsanleitung. Münster 2007, S. 223–277.

Meyer, Wolfgang: Informationssammlung und -berwertung. In: Stockmann, Reinhard; Meyer, Wolfgang: Evaluation. Eine Einführung. Opladen 2010, S. 191–234.

Müller-Kohlenberg, Hildegard: Verfahren, Chancen und Grenzen der Selbstevaluation – Wieviel Professionalität kann bei Selbstevaluationen erwartet werden? In: Ermert, Karl: Evaluation in der Kulturförderung. Wolfenbüttel 2004, S. 70–80

OECD (Hrsg.): Glossar entwicklungspoltischer Schlüsselbegriffe aus den Bereichen Evaluierung und ergebnisorientiertes Management. Paris 2009.

Paulini, Hannelore; Drochner, Ilse; Borch, Hans: Kriterienkatalog für die Evaluierung von Ausbildungsordnungen. BWP, Heft 24. Bonn 1995, S. 37–42.

Rossi, Peter H.; Lipsey, Mark W.; Freeman, Howard E.: Evaluation. A systematic approach. Thousand Oaks 2004.

Sanders, James R. (Hrsg.); Beywl, Wolfgang; Widmer, Thomas: Handbuch der Evaluationsstandards. Die Standards des „Joint Committee on Standards for Educational Evaluation“. Opladen 1999.

Sedlmeier, Peter; Renkewitz, Frank: Forschungsmethoden und Statistik in der Psychologie. München 2008.

Stockmann, Reinhard: Einführung in die Evaluation. In: Stockmann, Reinhard: Handbuch zur Evaluation. Eine praktische Handlungsanleitung. Münster 2007, S. 40–70.
Stockmann, Reinhard: Evaluation und Qualitätsentwicklung. Eine Grundlage für wirkungsorientiertes Qualitätsmanagement. Münster 2006.
Stockmann, Reinhard: Evaluationsprozess. In: Stockmann, Reinhard; Meyer, Wolfgang: Evaluation. Eine Einführung. Opladen 2010, S. 159–189.
Stockmann, Reinhard: Wissensbasierte Evaluation. In: Stockmann, Reinhard; Meyer, Wolfgang: Evaluation. Eine Einführung. Opladen 2010, S. 55–100.
Stockmann, Reinhard; Meyer, Wolfgang: Evaluation. Eine Einführung. Opladen 2010.
Taylor-Powell, Ellen; Hermann, Carol: Collecting Evaluation Data: Surveys. Madison, Wisconsin 2000.
Univation – Institut für Evaluation (Hrsg.): Beywl, Wolfgang; Niestroj, Melanie: Das A-B-C der wirkungsorientierten Evaluation. Köln 2009.
Weiss, Carol H.: Evaluation. Methods for studying programs and policies.Upper Saddle River, New Jersey 1998.
Widmer, Thomas: Meta-Evaluation. Kriterien zur Bewertung von Evaluationen. Bern 1996.
Wottawa, Heinrich; Thierau, Heike: Lehrbuch Evaluation. Bern 2003.

Anhang Arbeitshilfen

Arbeitshilfe: Qualitätskriterien, Gütekriterien und Untersuchungsdesign

Evaluationen mit einem Multimethoden-Ansatz zeichnen sich dadurch aus, dass sie sich meist sowohl qualitativer als auch quantitativer Erhebungs- und Auswertungsmethoden bedienen. Welche Unterschiede dabei hinsichtlich der Güte- bzw. Qualitätskriterien sowie des Untersuchungsdesigns zu berücksichtigen sind, erläutert diese ARBEITSHILFE.

In der Vergangenheit äußerten Vertreter/-innen der quantitativen Sozialforschung zum Teil heftige Kritik und Vorwürfe gegen qualitative Methoden. Diese hätten in den Wissenschaften nichts zu suchen, da sie subjektiv und willkürlich, insgesamt also unwissenschaftlich und darüber hinaus auch noch teuer und zeitraubend seien. Besondere Kritik betraf die relativ kleinen Fallzahlen, mit denen qualitative Untersuchungen in der Regel auskommen. An qualitativen Methoden orientierte Wissenschaftlerinnen und Wissenschaftler hielten dem entgegen, dass viele Untersuchungsgegenstände mit quantitativen Methoden überhaupt nicht zu erfassen seien. Wichtige Fragestellungen und Beobachtungen würden ohne qualitative Methoden einfach unterbleiben, bestimmte Zusammenhänge wären niemals entdeckt worden.

Das grundlegende Problem dieser Diskussionen bestand darin, dass zwei unterschiedliche Untersuchungsansätze direkt miteinander verglichen wurden und häufig die aus der quantitativen Forschung bekannten Gütekriterien an qualitative Vorgehensweisen angelegt wurden.

Jede wissenschaftliche Untersuchung muss bestimmten Standards genügen, je nach eingesetzten Untersuchungsmethoden. Qualitative Untersuchungsmethoden erfordern bestimmte Qualitätskriterien, zum Beispiel im Hinblick auf die Interpretation erhobener Daten oder Informationen. Für quantitative Methoden sind sogenannte Gütekriterien als Qualitätsstandards festgelegt. Beide werden nachfolgend erläutert.

Qualitätskriterien qualitativer Sozialforschung

Wegen der zum Teil recht kontrovers geführten Diskussion um die Beurteilung qualitativer Forschungsarbeiten wurden in den vergangenen Jahren Kriterien entwickelt, die zur Gütebeurteilung von qualitativer Forschung[28] angewandt werden können. HERR/ANDERSEN (2005) sowie STEINKE (1999) haben die wesentlichen Kriterien herausgearbeitet und zur Bewertung qualitativer Forschung vorgeschlagen.

Intersubjektive Nachvollziehbarkeit/Prozessvalidität

Sie wird durch die Dokumentation des Forschungsprozesses erreicht. Konkret sollte im Evaluationsbericht das Vorverständnis, die Erhebungsmethodik und der Erhebungskontext, die Daten sowie die Auswertungsmethodik dokumentiert werden. Hierbei soll beschrieben werden, wie die Evaluierenden zu ihren Ergebnissen gefunden haben. Gleichzeitig sollen die subjektive Position der Forschenden und ihre Auseinandersetzung mit Entscheidungen und Problemen im Forschungsprozess beleuchtet werden.

Indikatoren für die Beurteilung der Güte sind die Transparenz in der Dokumentation des Forschungsprozesses mit dem Ziel, intersubjektive Nachvollziehbarkeit zu ermöglichen sowie die Angemessenheit der Methodenwahl in Bezug auf die Ziele der Forschungsarbeit. Hier spielen

[28] Hierunter fallen ebenfalls Aktionsforschungsprozesse, zu denen AO-Evaluationen gezählt werden können.

unterschiedliche Aspekte eine Rolle wie bspw. die Frage danach, wie forschungsleitende Annahmen begründet werden oder ob das Thema aus unterschiedlichen Blickwinkeln (z. B. durch Methodentriangulation) betrachtet wurde. Auch der Aspekt, wie die Beziehung zu den Akteuren des Aktionsforschungsprozesses gestaltet wurde, ist hier interessant.

Prüfung der Gegenstandsangemessenheit

Vor Umsetzung des Forschungsprozesses ist auf der Basis von Bewertungskriterien, die festzulegen sind, zu beurteilen, inwieweit das geplante Vorgehen tatsächlich gegenstandsangemessen ist. Hierbei stehen die Fragen im Vordergrund, ob der Evaluationsgegenstand überhaupt einen qualitativen Forschungsprozess erforderlich macht und ob die ausgewählte Erhebungs- und Analysemethodik dem Evaluationsgegenstand gegenüber angemessen ist. Eine Umsetzung dieser Forderung findet sich ebenfalls in der Erläuterung und Begründung der Methodenwahl im Evaluationsbericht.

Empirische Verankerung der Theoriebildung und Theorieprüfung/Ergebnisvalidität

Sie ist besonders dann als Gütekriterium anzulegen, wenn durch die qualitative Forschungsarbeit neue Theorien entwickelt oder angewendet werden sollen, aber auch wenn es um die Diskussion von Ergebnissen geht. Hierbei ist zum einen entscheidend, ob hinreichende Belege für den theoretischen Ansatz im empirischen Material zu finden sind. Außerdem ist hier der Umgang mit von der Theorie abweichenden negativen Fällen, Situationen und Ereignissen von entscheidender Bedeutung.

Ein wichtiges Verfahren zur Überprüfung der empirischen Verankerung ist die kollegiale Validierung der Forschungsergebnisse. Konkret bedeutet dies, dass beispielsweise im Rahmen von Teamsitzungen Ergebnisse gemeinsam diskutiert werden.

Limitation

Die sogenannte Limitation beschreibt die Grenzen des Geltungsbereichs einer qualitativen Forschungsarbeit. Hierbei gilt es, im Evaluationsbericht begründete Aussagen zur Generalisierung (im Sinne einer ‚Verlängerung', nicht einer statistischen Generalisierbarkeit!) der Ergebnisse zu machen und Grenzen und Einschränkungen aufzuzeigen. Um eine begründete Aussage zur möglichen Verallgemeinerung der Ergebnisse zu treffen, ist die Beschreibung der Kontexte der erhobenen Daten und der für das Untersuchungsphänomen relevanten Bedingungen notwendig.

Als Techniken, die zur Herausarbeitung des Geltungsbereichs einer qualitativen Forschungsarbeit herangezogen werden, gelten z. B. die Bildung von Idealtypen, der Fallvergleich und die Fallkontrastierung.

Reflektierte Subjektivität/Dialogvalidität/demokratische Validität

Mit der Reflektierten Subjektivität als qualitativem Gütekriterium ist gemeint, dass sich Evaluatorinnen und Evaluatoren immer ihrer Rolle als Teil des Forschungsprozesses bewusst sind und diese auch reflektieren. Diese Reflexion bezieht sich auf den Einstieg ins Forschungsfeld, auf den gesamten Forschungsprozess, auf die möglichen biografischen Beziehungen der evaluierenden Personen zum Forschungsthema wie z. B. die Beteiligung am vorangegangenen Ordnungsverfahren des nun zu evaluierenden Berufes, sowie auf die Beziehung zwischen Evaluierenden und Informanten.

Eng mit der Subjektivität, die es zu reflektieren gilt, verbunden ist die Dialogvalidität (kollegiale Validierung). Hierunter wird eine Absicherung der Forschungsergebnisse durch Peerreview-Verfahren verstanden. Dies kann durch die Arbeit in Forschungsteams, den intensiven

Austausch über die Ergebnisse oder zumindest durch den kritischen und reflexiven Dialog mit anderen, mit dem Forschungsfeld vertrauten, aber auch nicht so sehr vertrauten Personen, denen ein ‚fremder Blick' noch möglich ist, erreicht werden.

Mit der demokratischen Validität wird die Partizipation der Beteiligten in den Mittelpunkt der Gütebeurteilung des Aktionsforschungsprozesses gestellt. In der Literatur wird dies auch als ökologische oder Kontextvalidität bezeichnet.

Kohärenz

Kohärenz bedeutet, dass die Ergebnisse qualitativer Forschungsarbeiten und besonders der durch sie generierten Theorien hinsichtlich ihrer Widersprüche in den Daten und der Interpretation der Daten bearbeitet werden. Dies bedeutet nicht, dass diese Widersprüche zugedeckt oder beschönigt werden sollen, sondern dass nicht gelöste Fragen und Widersprüche offengelegt und dokumentiert werden.

Relevanz

Hiermit ist die Alltagstauglichkeit und Praxisnähe (i. S. v. einer Anwendungsorientierung) einer qualitativen Forschungsarbeit gemeint. Dies bezieht sich sowohl auf die Praxisrelevanz der Ergebnisse als auch auf die Ausgestaltung eines alltagstauglichen Untersuchungsdesigns oder eine direkte Verbindung zwischen Forschung und Praxis.

Gütekriterien quantitativer Sozialforschung

In der quantitativen Sozialforschung werden mindestens die drei Gütekriterien Objektivität, Reliabilität und Validität gefordert, wenn eine Untersuchung wissenschaftlichen Qualitätsansprüchen genügen soll. Hinzukommen können noch eine Reihe von Nebenkriterien, wie zum Beispiel Utilität und Effizienz, die sich mit Aspekten wie Nützlichkeit und Wirtschaftlichkeit von Untersuchungen beschäftigen. Für die wissenschaftliche Qualität sind aber Objektivität, Reliabilität und Validität ausschlaggebend.

Diese Trias von Testgütekriterien gehen auf die Tradition der klassischen Testtheorie zurück. Nach Leonhart (2004) gibt es eine logische Beziehung zwischen den drei Gütekriterien. Ein nicht objektiver Test kann nicht reliabel sein, und wenn er nicht reliabel ist, kann er auch nicht valide sein.

Objektivität

Eine Untersuchung bzw. ein Untersuchungsinstrument ist dann objektiv, wenn verschiedene Untersuchende bei denselben Personen oder denselben Untersuchungsgegenständen unabhängig voneinander die gleichen Resultate erzielen. Die Untersuchungsergebnisse müssen also reproduzierbar sein.

> Die Objektivität eines Tests gibt an, in welchem Ausmaß die Testergebnisse vom Testanwender unabhängig sind. Bortz/Döring (1995).
>
> Mit Objektivität ist gemeint, inwieweit das Testergebnis unabhängig ist von jeglichen Einflüssen außerhalb der getesteten Person, also dem Versuchsleiter, der Art der Auswertung, den situativen Bedingungen, der Zufallsauswahl der Testitems usw. Es ist ersichtlich, dass es sehr viele verschiedene Arten von Objektivität bei Tests zu unterscheiden gilt. Leonhart (2004)

Wie Bortz und Döring (1995) weiter feststellen, ist die Objektivität meist ein unproblematisches Gütekriterium, wenn Durchführung, Auswertung und Ergebnisinterpretation vorab standardisiert festgelegt worden sind.

Beispiele für Objektivität im Rahmen der Evaluation: Es sollte keine Rolle spielen, ob z. B. Lehrlinge durch Mitarbeiter und Mitarbeiterinnen des BIBB oder BIBB-externe Evaluatoren und Evaluatorinnen befragt werden. Die Objektivität könnte hingegen durchaus eingeschränkt werden, wenn beispielsweise Fragebogen von einer strengen Lehrkraft der Berufsschule oder einer netten jungen Studentin, die als wissenschaftliche Hilfskraft die Evaluation unterstützt, überreicht werden. Um dieses Problem zu vermeiden, bieten u. a. Online-Befragungen ein hohes Maß an Objektivität.

Auswertungsobjektivität sollte gegeben sein, wenn die Daten elektronisch, beispielsweise mit SPSS – und angemessenen Methoden – analysiert werden.

Die Feststellung bzw. Messung der Objektivität ist zusammenfassend betrachtet nicht im gleichen Maße möglich wie beispielsweise bei der Reliabilität. Evaluatoren und Evaluatorinnen sollten sich jedoch der Wichtigkeit des Gütekriteriums bewusst sein und Maßnahmen zur Sicherstellung der Objektivität dazu im Bericht darlegen.

Reliabilität

Die Reliabilität (oder Zuverlässigkeit) gibt an, wie genau ein Test misst. Dabei gilt es, den möglichen Fehleranteil an einer Messung so gering wie möglich zu halten.

> Die Reliabilität eines Tests kennzeichnet den Grad der Genauigkeit, mit dem das geprüfte Merkmal gemessen wird. Bortz/Döring (1995).
>
> Als Messgenauigkeit wird dabei nicht die Zahl der Dezimalstellen der Messwerte bezeichnet, sondern die Zuverlässigkeit, mit der bei einer wiederholten Messung unter gleichen Bedingungen dasselbe Messergebnis herauskommt. Leonhart (2004)

Um die Zuverlässigkeit eines Tests oder eines Testverfahrens bestimmen zu können, gibt es verschiedene mögliche Vorgehensweisen. Hier drei der bekanntesten:

Testwiederholung: Derselbe Test wird nach einigen Monaten[29] mit denselben Personen wiederholt und dann werden die Ergebnisse miteinander verglichen.

Testhalbierung: Derselbe Test wird zwei miteinander vergleichbaren Personengruppen zur einmaligen Beurteilung vorgelegt.

Parallel-Test: Zwei ähnliche Tests, die das gleiche Merkmal messen, werden an ein- und dieselbe Personengruppe verteilt. Dann werden die Ergebnisse miteinander verglichen.

Für die Evaluationspraxis eignet sich am besten die Berechnung der *internen Konsistenz – Cronbach's Alpha*. Da es meist keine Möglichkeit für ausführliche Pre-Tests gibt und Messwiederholungen eher die Ausnahme bilden, bleibt die Berechnung von Cronbach's Alpha die einzige Alternative. Der Wertebereich für die Reliabilität, z. B. Cronbach's Alpha wird nach Fisseni 1997 (zitiert nach Bühner: 2004) wie folgt definiert:

< 0,8 niedrig
0,8 bis 0,9 mittel
> 0,9 hoch.

[29] In der Psychologie etwa vier bis sechs Wochen, bei längeren Zeiträumen kann es auch bei stabilen Merkmalen zu leichten Veränderungen kommen.

Zusätzlich sollten die *korrigierten Trennschärfen* der Items berechnet werden. Diese geben an, wie gut ein einzelnes Item mit den anderen Items einer Skala korreliert. Diese Korrelation ist nach Bühner wie folgt zu bewerten:

< 0,3 niedrig
0,3 bis 0,5 mittel
> 0,5 hoch.

SPSS gibt im Output auch an, wie sich Cronbach's Alpha verändern würde, wenn einzelne Items entfernt werden.

Grundsätzlich steigt mit der Anzahl der Items auch der Grad der Genauigkeit, vorausgesetzt, dass es sich um reliable Items handelt. Beispielsweise kann die Zufriedenheit anhand eines Items erfragt werden, genauere Informationen werden jedoch mit mehreren Items zu erzielen sein.

Validität

Eine Untersuchung bzw. ein Untersuchungsinstrument (z. B. ein Fragebogen) ist dann valide, wenn Items abgefragt werden, die in ihrer Summe Hinweise auf den zu untersuchenden Sachverhalt bzw. die Forschungsfrage liefern.

Es gibt mehrere Arten von Validität. Die wichtigste davon ist die Konstruktvalidität. Aus einem Konstrukt, z. B. „berufliche Handlungsfähigkeit", müssen sich Hypothesen ableiten lassen, die mittels eines oder mehrerer Messinstrumente überprüft werden können. Dabei muss überlegt werden, welche Hypothesen sich eindeutig aus dem Konstrukt ableiten lassen oder welche Hypothesen bereits bekannt sind und für einen speziellen Fall. z. B. einen neuen Ausbildungsberuf, noch einmal überprüft werden sollen.

Beispielsweise ist eine Fragenbatterie, die ausschließlich auf erlerntes Wissen in Berufsschule und Betrieb fokussiert ist, nicht valide, wenn es eigentlich um die Frage nach erlernten Fachkompetenzen geht.

Ein Beispiel ist die prognostische Validität, wenn das Kriterium erst in der Zukunft erhoben werden kann. Z. B. ist ein Eignungstest für Berufsanfänger/-innen dann valide, wenn er den tatsächlichen Berufserfolg vorhersagen kann.

Erst wenn klar ist, was das zugrunde liegende Konstrukt ist und welche Hypothesen und damit verbundene Fragen sich davon ableiten lassen, können Überlegungen zum Forschungsdesign angestellt werden.

Aufgrund der problematischen Messung der Validität und der hohen Komplexität im sozialwissenschaftlichen Bereich sind auch die Anforderungen an die Validität im Vergleich zur Reliabilität niedriger(siehe Bühner, 2004):

Werte <0,4 sind niedrig
0,4 bis 0,6 mittel
ab 0,6 hoch.

> Die Validität eines Tests gibt an, wie gut der Test in der Lage ist, genau das zu messen, was er zu messen vorgibt.
>
> Bortz/Döring (1995).

Untersuchungsdesign

Nach Erarbeitung geeigneter Hypothesen (und entsprechender Fragen) müssen Überlegungen dahingehend angestellt werden, welches Untersuchungsdesign (auch Forschungsdesign genannt) für die Beantwortung der Fragen am besten geeignet ist. Das Untersuchungsdesign legt fest, auf welche Weise die erstellten Hypothesen und dazugehörenden Fragestellungen untersucht werden sollen. Es legt also fest, welcher Gegenstand (z. B. Prüfungsanforderungen), zu welcher Zeit (z. B. nach der Abschlussprüfung), auf welche Weise (z. B. mittels Fragebogen) unter Heranziehung welcher Informationsquellen (z. B. Auszubildende) empirisch überprüft werden soll.

Grundsätzlich werden drei Designs unterschieden:

Das experimentelle Design

Dieser Ansatz entspricht dem klassischen (meist naturwissenschaftlichen) Experiment. Eine Hypothese wird geprüft, indem eine unabhängige Variable – nach Ausschaltung aller möglichen Störvariablen – systematisch manipuliert und der dabei auftretende Effekt registriert bzw. gemessen wird. Bei den (seltenen) Experimenten im sozialwissenschaftlichen Bereich ist die Randomisierung der Untersuchungsteilnehmer und Untersuchungsteilnehmerinnen unabdingbare Voraussetzung.

Praktisch bedeutet dies, dass eine in Bezug auf ein bestimmtes Merkmal homogene Stichprobe in eine Versuchsgruppe und eine Kontrollgruppe aufgeteilt und in die Untersuchung einbezogen wird. In der Versuchsgruppe wird ein sogenanntes Treatment eingesetzt (z. B. ein Medikament oder ein Lehrgang), welches in der Kontrollgruppe nicht genutzt wird. Dann werden die beiden Gruppen auf Veränderungen (z. B. verringertes Schmerzempfinden oder Fremdsprachenkenntnisse) untersucht und die Ergebnisse miteinander verglichen. Wichtig ist dabei, dass alle anderen Möglichkeiten, die eine Veränderung verursachen könnten, ausgeschlossen worden sind, sodass nur das Treatment als veränderndes Element infrage kommt.

Der Vorteil dieses Designs besteht darin, dass eindeutige kausale Aussagen möglich sind. Die zufällig ausgewählten Personen oder Objekte im Sinne einer Stichprobe stehen stellvertretend für eine Grundgesamtheit von Personen oder Objekten mit genau definierten Eigenschaften. Die gemessenen Effekte, die in der Stichprobe aufgetreten sind, dürfen auf die Grundgesamtheit übertragen werden (Repräsentativität).

Diese Art von experimentellem Design wird auch als „Laborexperiment“ bezeichnet. Es kommt im Rahmen von Evaluationen von Ausbildungsordnungen nicht vor, da keine zufällige Verteilung auf eine Versuchs- und eine Kontrollgruppe möglich ist.

Das quasi-experimentelle Design

Hierbei wird wie bei dem eben beschriebenen „klassischen“ Experiment verfahren. Allerdings ist (im sozialwissenschaftlichen Bereich) eine Randomisierung der Stichprobe also eine zufällige Verteilung auf eine Versuchs- und eine Kontrollgruppe, bei diesem Design keine Voraussetzung und in der Regel auch nicht mehr möglich, weil andere Personen oder Institutionen bereits eine Vorauswahl getroffen haben. Die zu untersuchende Stichprobe, deren Festlegung in der Regel erst nach der Durchführung einer Maßnahme (z. B. Berufsausbildung) erfolgt, wird vielmehr nach bestimmten Merkmalen (z. B. ein Prüfungsjahrgang, Alter oder Geschlecht der Auszubildenden o. ä.) ausgewählt.

Das Problem dieses Designs besteht darin, dass die nichtzufällige Zuordnung der Untersuchungsteilnehmer und Untersuchungsteilnehmerinnen auf eine Versuchs- und eine Kontroll-

gruppe Einflussfaktoren in Kauf nimmt, die nicht auf das Treatment zurückzuführen sind. Eindeutige kausale Aussagen sind bei diesem Design daher nur bedingt möglich.

Das nicht-experimentelle Design

Dieses Design verzichtet gänzlich auf eine Kontrollgruppe. Abhängige und unabhängige Variablen werden gemessen, wobei deren Beziehungen oder Wechselwirkungen meist unbekannt sind. Störvariablen können oft nicht kontrolliert oder ausgeschaltet werden. Im Wesentlichen sind „nur“ korrelative Aussagen möglich.

Es kann i. d. R. keine eindeutige Aussage darüber getroffen werden, ob das gewählte Treatment (z. B. eine bestimmte Prüfungsstruktur) wirklich die einzige Ursache für eine gemessene Veränderung (z. B. höhere Motivation der Auszubildenden) war oder nicht. Andere mögliche Einflussfaktoren (z. B. Ausbilderverhalten) müssen daher zunächst erkannt und dann in die Ergebnisinterpretation mit einbezogen werden.

Nicht-experimentelle Designs stellen bisher das gängige Vorgehen bei der Evaluation von Ausbildungsordnungen dar, da die Bildung einer Kontrollgruppe i. d. R. nur schwer möglich ist oder ein quasi-experimentelles Design aus Kosten- und Zeitgründen bereits von Beginn an nicht zur Debatte steht.

Untersuchungs- bzw. Messzeitpunkte

Zum Forschungsdesign gehört auch die Klärung der Frage, wann und wie oft eine Untersuchung durchgeführt werden soll. In der Regel wird zwischen Querschnittuntersuchungen (eine Untersuchung wird zu einem bestimmten Zeitpunkt einmalig durchgeführt) und Längsschnittuntersuchungen[30] (eine bestimmte Untersuchung wird mehrmals, meist in Jahresabständen, durchgeführt) unterschieden.

Art der Messung

Unterschieden werden Einmalmessungen, Vorher-Nachher-Messungen sowie Messungen in regelmäßigen zeitlichen Abständen. Diese Messungen können mit und ohne Kontrollgruppen, die randomisiert oder nicht-randomisiert sind, ausgeführt werden.

Zusammenfassung

Insgesamt betrachtet ergeben sich vielfältige Möglichkeiten eine Evaluation zu gestalten, wie in der Tabelle dargestellt.

	Messung vor dem Treatment	Messung während des Treatments	Messung nach dem Treatment
Experimentelles Design			
Quasi-experimentelles Design			
Nicht-experimentelles Design			

[30] Zum Beispiel Panel-Untersuchungen, bei denen eine festgelegte Personengruppe über einen bestimmten Zeitraum hinweg untersucht wird, oder Kohorten-Studien, bei denen bestimmte Jahrgänge im Hinblick auf ein bestimmtes Merkmal oder mehrere bestimmte Merkmale längerfristig (meist über Jahre hinweg) untersucht werden.

„Die Aufgabe eines Forschungsdesigns ist es (…) eine exakte Bewertung über einen vermuteten Ursache-Wirkungszusammenhang zu ermöglichen. Zur Durchführung dieser Aufgabe gibt es aber keinen allgemeinverbindlichen „Königsweg“.

Manche Autoren folgen einem Ablaufschema und planen die verschiedenen Schritte einer Evaluation in ihrer zeitlichen Reihenfolge. Andere orientieren sich an dem Informationsbedarf und den daraus folgenden Erhebungsmethoden (…). Wieder andere rücken schließlich den Dialog mit den Beteiligtengruppen in den Vordergrund und leiten daraus die Aufgabe zur Planung der (…) Evaluation ab.

Selbst die Kernelemente eines Forschungsdesigns – die Kontrolle von Störeinflüssen und die Isolierung von Ursachen-Wirkungszusammenhängen – sind nicht unumstritten. Ihnen wird der Anspruch einer partizipativen Erschließung des Sinns von Kausalzusammenhängen oder auch die Vermittlung systemischen Denkens zum Verstehen komplexer Modelle als Zielvorstellung entgegengesetzt.“

Meyer in Stockmann (2007).

Literatur

Bortz, Jürgen; Döring, Nicola: Forschungsmethoden und Evaluation für Sozialwissenschaftler. Berlin und Heidelberg 1995.

Bühner, Markus: Einführung in die Test- und Fragebogenkonstruktion. München 2004.

Flick, Uwe: Qualitative Forschung. Theorie, Methoden, Anwendung in Psychologie und Sozialwissenschaften. Reinbek 1995.

Glaser, Barney G.; Strauss, Anselm L.: The Discovery of Grounded Theory. Hawthorne, New York 1967.

Herr, Kathryn; Anderson, Gary L.: The Action Research Dissertation. A Guide for Students and Faculty. Thousand Oaks, California 2005.

Kelle, Udo; Erzberger, Christian: Stärken und Probleme qualitativer Evaluationsstudien – ein empirisches Beispiel aus der Jugendhilfeforschung. In: Flick, Uwe (Hrsg.): Qualitative Evaluationsforschung – Konzepte, Methoden, Umsetzungen. Reinbek 2006, S. 284–300.

Kelle, Udo; Erzberger, Christian: Qualitative und quantitative Methoden – Kein Gegensatz. In: Flick, Uwe; Kardorff, Ernst von; Steinke, Ines (Hrsg.): Qualitative Forschung. Ein Handbuch. Reinbek 2005, S. 299–309.

Leonhart, Rainer: Lehrbuch Statistik. Einstieg und Vertiefung. Bern 2004.

Mayring, Philipp: Qualitative Inhaltsanalyse. Grundlagen und Techniken. Weinheim 1983.

Meyer, Wolfgang. Evaluationsdesigns. In: Stockmann, Reinhard. Handbuch zur Evaluation. Eine praktische Handlungsanleitung. Münster 2007, S. 143–163.

Steinke, Ines: Kriterien qualitativer Forschung. Ansätze zur Bewertung qualitativ-empirischer Sozialforschung. Weinheim und München 1999.

Stockmann, Reinhard: Handbuch zur Evaluation. Eine praktische Handlungsanleitung. Münster 2007.

Stockmann, Reinhard; Meyer, Wolfgang: Evaluation. Eine Einführung. Opladen 2010.

Arbeitshilfe: Stichprobenauswahl

Sowohl bei quantitativen als auch bei qualitativen Methoden müssen gezielt aus Quellen bzw. Datenträgern Fälle i. S. v. Erhebungseinheiten ausgewählt und entsprechende Daten erhoben werden. Diese Arbeitshilfe geht gezielt darauf ein, was es im Rahmen dieser Auswahlprozesse zu berücksichtigen gilt, wie dieser Auswahlprozess gestaltet werden kann und welche Stichprobengrößen im Falle von Teilerhebungen gewählt werden sollten (standardisierte Erhebungen) bzw. welche Prinzipien bei Fallauswahlen in qualitativen Studien berücksichtigt werden müssen.

1 Quantitative Methoden

Alle Methoden, die der zahlenmäßigen Darstellung empirischer Sachverhalte dienen, werden den quantitativen Methoden zugeordnet. Dabei steht *das Messen zählbarer Eigenschaften* bzw. *Merkmale* im Fokus. Diese zählbaren Eigenschaften werden standardisiert abgebildet durch Befragungen (z. B. mittels geschlossener Fragen im Fragebogen), durch nicht-reaktive Verfahren (z. B. Inhaltsanalyse in Form einer Kategorienbildung, Sekundäranalyse) sowie durch Beobachtungen (z. B. Messung von Reaktionszeiten).[31]

Stichprobenauswahl im Rahmen der Evaluation von Ausbildungsordnungen

Die Auswahl von Datenträgern, von denen jeweils solche zählbaren Merkmale erhoben werden sollen, ist die Aufgabe der Stichprobenauswahl. Grundsätzlich kann natürlich anstelle einer Stichprobenauswahl eine Vollerhebung vorgenommen werden.

Bei der Evaluation von Ausbildungsordnungen werden häufig Stichproben gezogen. Hierbei stellt sich grundsätzlich die Frage nach dem *Umfang* und der *regionalen Verteilung* der Stichprobe. Dabei ist zu bedenken, dass gegebenenfalls mehrere Gruppen, wie z. B. Ausbilderinnen und Ausbilder oder Lehrkräfte, in die Untersuchung mit einbezogen werden müssen, um Fragen hinreichend beantworten zu können.

Demnach ist zunächst zu überlegen, welchen Umfang (*Grundgesamtheit*) diese Gruppen de facto haben und wie groß die daraus zu ziehenden Stichproben sein sollen. Darüber hinaus ist zu klären, ob die Stichproben z. B. hinsichtlich ihrer regionalen Verteilung oder der Betriebsgröße so gewählt werden können, dass sie ein möglichst genaues Abbild der jeweiligen Grundgesamtheit sein können und somit (soweit als möglich) *repräsentativ* sind.

Regionale Verteilung

Die Regionale Verteilung spielt in den Lehrbüchern über Erhebungsmethoden in der Regel eher eine untergeordnete Rolle. Im Rahmen der Evaluation von Ausbildungsordnungen kann diese jedoch von großer Bedeutung sein. Eine Ausbildungsordnung kann in einer ausgewählten Region z. B. aufgrund der vorhandenen Infrastruktur überaus „erfolgreich“ sein, d. h. eine hohe Zahl an Auszubildenden aufweisen, während sie in einer anderen Region derzeit keine Rolle spielt, weshalb auf Grundlage dieser Ausbildungsordnung dort niemand oder nur wenige ausgebildet werden. Diese regionale Verteilung ist daher in der Wahl der Stichprobengrößen unbedingt zu berücksichtigen.

[31] Eine Übersicht der gängigen quantitativen sowie qualitativen Datenerhebungsverfahren findet sich bei Stockmann/Meyer 2010, S. 210–212.

Grundgesamtheit

„Unter Grundgesamtheit ist diejenige Menge von Individuen, Fällen, Ereignissen zu verstehen, auf die sich die Aussagen der Untersuchung beziehen sollen und die im Hinblick auf die Fragestellung und die Operationalisierung vorher eindeutig abgegrenzt werden muss." Kromrey (2006: 269). Nach Bortz/Döring beziehen sich Stichprobenuntersuchungen auf Populationen, die in begrenzten Zeiträumen real existieren, und sind daher über diese hinaus nicht generalisierbar. Das trifft insbesondere auf human- oder sozialwissenschaftliche Fragestellungen zu, die sich mit der Zeit verändern. Zur Definition der Population oder Grundgesamtheit sollten möglichst operationale, leicht erhebbare Merkmale verwendet werden, vgl. Bortz/Döring (2002: 399).

Für die Evaluation von Ausbildungsordnungen bedeutet dies, dass vor Untersuchungsbeginn geklärt werden muss, von welchen Personengruppen Daten erhoben werden sollen. Relevante Akteure des Evaluationsgegenstands wurden den beiden Dimensionen „Weisungsgeber, Weisungsnehmer, Sozialpartner" und „Zielgruppen" zugeordnet. Diese könnten z. B. Ausbilderinnen und Ausbilder (in Betrieben und bei Bildungsträgern), Mitarbeiterinnen und Mitarbeiter von zuständigen Stellen, Lehrkräfte an Berufsschulen, Auszubildende (aus unterschiedlichen Ausbildungsjahrgängen, in unterschiedlichen Fachrichtungen oder Schwerpunkten) und Vertreterinnen und Vertreter von Fachverbänden sein.

Zur Bestimmung einer Grundgesamtheit gilt es herauszufinden, wie groß die jeweilige Grundgesamtheit tatsächlich ist und wie sich diese regional verteilt.

Um das zu klären, gibt es diverse Informationsquellen und Datenpools, die herangezogen werden können:

1. im BIBB vorhandene Daten,
2. Daten der zuständigen Stellen,
3. Daten der KMK (z. B. Liste aller Berufsschulen, die den zu untersuchenden Beruf ausbilden),
4. Daten von Fachverbänden (z. B. Adressdateien von Ausbildungsfirmen),
5. Daten bei der Bundesagentur für Arbeit (z. B. Adressen aus der Betriebsdatei),
6. Daten von Fachverbänden (Berufsverbänden),
7. Daten, deren Zugang Fachbeiräte oder Projektbeiräte ermöglichen.

Es sind folglich viele Wege gangbar, um zu den benötigten Informationen über die jeweiligen Grundgesamtheiten zu gelangen. Eine Klärung hinsichtlich der Grundgesamtheiten (Personengruppen), die von Interesse sind, sollte so früh wie möglich erfolgen, da es meist viel Zeit in Anspruch nimmt, an die gewünschten Datensätze zu gelangen.

Vollerhebung versus Stichprobe

In der Regel ist die Größe der jeweiligen Grundgesamtheit so umfangreich, dass aus Zeit- und Kostengründen eine Stichprobenziehung erforderlich ist. Bei „kleineren" Berufen hingegen kann auch eine *Vollerhebung* der jeweiligen Grundgesamtheiten erfolgen, bei der von *allen* Personen, Betrieben, Berufsschulen etc., die sich in den relevanten Grundgesamtheiten befinden, Daten erhoben werden. Ob und wann eine Vollerhebung unwirtschaftlich wird, kann nicht eindeutig bestimmt werden. Die Entscheidung für oder gegen eine Vollerhebung hängt häufig von den zur Verfügung stehenden Personalressourcen sowie von den vorhandenen finanziellen Mitteln ab.

Um ein vollständiges Bild der Verteilung von Merkmalen wie z. B. die Zufriedenheit mit dem gewählten Beruf zu erhalten, müssten alle Personen befragt werden, die diesen Beruf erlernt haben. Da dies aufgrund des hohen Aufwands in der Regel nicht möglich ist, wird nur ein bestimmter Teil (die Stichprobe) aus der Menge aller betroffenen Personen (aus der Grundgesamtheit) befragt. Für Bortz (1999: 86) muss die Frage beantwortet werden, wie gewährleistet werden

kann, dass eine Stichprobe eine Grundgesamtheit möglichst genau repräsentiert. Er unterscheidet „globale Repräsentativität“ (in Bezug auf alle Merkmale) und „spezifische Repräsentativität“ (in Bezug auf bestimmte Merkmale). Hinsichtlich der Menge der zu erfassenden Merkmale sieht er einen Vorteil von Stichproben gegenüber Vollerhebungen. Bortz geht davon aus, dass aufgrund der vergleichsweise geringen Anzahl von Untersuchungsteilnehmern eine größere Anzahl von Merkmalen sorgfältiger und kontrollierter erhoben werden kann, vgl. Bortz/Döring (2002: 398).

Ergebnisse aus Stichproben dienen der Schätzung der Grundgesamtheit. Die Abweichungen zwischen den Ergebnissen, z. B. die Mittelwerte aus der Befragung der Stichprobe und den Ergebnissen, die sich ergeben hätten, wenn alle Personen dieser Grundgesamtheit befragt worden wären, werden durch den *Standardfehler* (des Mittelwertes) ausgedrückt. Dieser ist definiert als Standardabweichung der Mittelwerte von gleich großen Zufallsstichproben einer Population, vgl. Bortz (1999: 89)[32].

Grundsätzlich gilt, dass sich die festgestellten Merkmale der Stichprobe mit wachsendem Stichprobenumfang immer mehr den tatsächlichen Merkmalen der Grundgesamtheit annähern. Je größer also der Stichprobenumfang ist, desto kleiner wird der Standardfehler und desto genauer wird der unbekannte Populationsparameter geschätzt, vgl. Bortz (1999: 89). Dies spricht natürlich für eine möglichst große Stichprobe. Aus der Praxis ist aber bekannt, dass bei gut gewählter Stichprobe die Zahl der Befragten nicht unbedingt groß sein muss. Darüber hinaus lässt sich ab einem bestimmten Punkt der Zugewinn an Informationsgenauigkeit nur noch unter großem Aufwand steigern.[33]

Stichprobenumfang

Die Wahl des richtigen Stichprobenumfangs hängt auch von den statistischen Auswertungsverfahren ab, mithilfe derer die erhobenen Daten ausgewertet werden sollen. Der Stichprobenumfang muss ein anderer sein, wenn z. B. Mittelwerte aus den erhobenen Daten verglichen werden sollen im Vergleich zu einem Mittelwert, der mit einem Erfahrungswert verglichen werden soll usw.

Als „Faustregel“ kann die nachfolgende Formel von Meyer (2007: 233) zur Ermittlung des benötigten Stichprobenumfangs eingesetzt werden:

$$n = \frac{N}{1 + d^2 (N - 1)}$$

n = Umfang der Stichprobe

N = Umfang der Grundgesamtheit

d = Tolerierter Stichprobenfehler (Irrtumswahrscheinlichkeit, bei 5 % ist d = 0,05)

Für ein konkretes Beispiel mit einer Populationsgröße von N = 1.724 und einer Irrtumswahrscheinlichkeit von 5 % ergibt sich also folgende Stichprobengröße:

$$n = \frac{1.724}{1 + 0{,}5^2 * (1.724 - 1)} = 324{,}82$$

[32] Anders ausgedrückt: Der Standardfehler gibt die theoretische Streubreite der Gruppenmittelwerte an, die sich ergeben würde, wenn alle möglichen Stichproben (und nicht nur eine) aus der Grundgesamtheit gezogen würden.

[33] Der Standardfehler nimmt proportional zur Quadratwurzel des Stichprobenumfanges ab. Um den Standardfehler zu halbieren, muss man den Stichprobenumfang also vervierfachen.

Die errechnete Stichprobengröße liegt bei 324,82, also bei 325 Personen.

Beispielhaft sind bei Meyer einige Stichprobengrößen zu Grundgesamtheiten angegeben:

Grundgesamtheit (N)	Stichprobenumfang (n)	Grundgesamtheit (N)	Stichprobenumfang (n)
10	10	10.000	385
50	45	20.000	392
100	80	50.000	397
500	222	100.000	398
1.000	286	1.000.000	400
5.000	370	10.000.000	400

Quelle: Meyer (2007: 233)

Eine weitere Methode zur Berechnung von Stichprobengrößen bietet sich über bekannte oder geschätzte Anteile an. Im ungünstigsten Falle ist keine Einschätzung über die gesuchten Parameter bekannt, weshalb von einem Verhältnis 50/50 ausgegangen werden muss; p wäre dann 0,5. Ist hingegen Vorwissen über die Verteilung des Merkmals in der Grundgesamtheit vorhanden, kann ein anderes Verhältnis gewählt werden.

Bei dem konkreten Beispiel kann diese Formel wie folgt aussehen:

$$n \geq \frac{N}{1 + \frac{(N-1)\,\varepsilon^2}{z^2\,P\,Q}}$$

n = wird gesucht

N = 1.724 Auszubildende

ε = 0,05 (also Fehlertoleranz von 5 %)

z = 1,96 (also Sicherheitswahrscheinlichkeit von 95 %)

P = 0,5 (geschätzt auf 50/50)

Q = 1 − 0,5

$$n \geq \frac{1.724}{1 + \frac{(1.724-1) * 0{,}05^2}{1{,}96^2 * 0{,}5 * 0{,}5}} = 314{,}31$$

n ≥ 314,31 (Die Stichprobe muss mind. 314 Auszubildende umfassen.)

Eine hilfreiche Liste mit empfohlenen Stichprobengrößen findet sich auch auf Seite 10 des Arbeitspapiers „Sampling“ von Taylor-Powell (1998):

http://learningstore.uwex.edu/assets/pdfs/G3658-3.PDF

Stichproben aus N ≤ 30 gelten in der Regel als zu klein, um aussagefähige Rückschlüsse auf eine Grundgesamtheit zu erlauben. In diesen Fällen darf nicht von einer Normalverteilung ausgegangen werden.

Stichprobenarten

Die Stichprobenarten stellen Techniken dar, mit deren Hilfe Stichproben aus einer Grundgesamtheit gezogen werden können. Grundsätzlich werden *zufallsgesteuerte Auswahlverfahren* und *bewusste Auswahlverfahren* unterschieden. Nachfolgend werden exemplarisch einige Möglichkeiten der Stichprobenziehung kurz erläutert. Weitere Auswahlverfahren finden sich beispielsweise in Kromrey (2006: 279–315).

Die Entscheidung für ein bestimmtes Auswahlverfahren hängt in entscheidendem Maße von der Zusammensetzung und Heterogenität der Grundgesamtheit ab.

Zufallsstichprobe

Wenn über die Verteilung der Merkmale innerhalb der Grundgesamtheit nichts bekannt ist, aber alle Elemente der Grundgesamteit bekannt sind und der Erhebung prinzipiell zur Verfügung stehen, sollte eine Zufallsstichprobe gezogen werden, vgl. Bortz (1999: 86). Jedes Element, das heißt z. B. jeder Ausbildungsbetrieb, muss dann die gleiche Wahrscheinlichkeit haben, in die Stichprobe zu gelangen. So kann beispielsweise aus einer vorliegenden Liste aller Ausbildungsbetriebe, die z. B. den gleichen Beruf ausbilden, jeder 15. Betrieb ausgewählt werden (systematische Zufallsstichprobe). Sind anstelle der Betriebe jedoch einzelne Personen Elemente der Stichprobe, müssten aus einer Liste, z. B. von allen Auszubildenden oder Lehrkräften, in gleicher Weise Personen in die Stichprobe aufgenommen werden.

Klumpenstichprobe

Bedingung für dieses Verfahren ist, dass die Grundgesamtheit einfach, sozusagen „natürlich“ in mehrere Klumpen (engl. cluster) unterteilt werden kann. So können z. B. Auszubildende in einem Beruf ausgewählt werden, die in ausgewählten Kammerbezirken eingetragen oder in bestimmten Berufsschulzentren zusammengefasst sind. Nach Auswahl spezieller Klumpen werden Daten von allen Elementen des Klumpens, z. B. alle Ausbilderinnen und Ausbilder in einem bestimmten Kammerbezirk sowie alle Auszubildenden dort, im Sinne einer Vollerhebung ermittelt. Nach Bortz (1999: 88) ist die Bezeichnung „Klumpenstichprobe“ nur dann gerechtfertigt, wenn mehrere zufällig ausgewählte Klumpen vollständig befragt werden.

Ein Effekt der Klumpenstichprobe ist, dass die Streuung des untersuchten Merkmals größer sein wird als bei der Anwendung der Zufallsstichprobe.

Wird aus dem Klumpen lediglich eine Zufallsstichprobe gezogen, so spricht man von einem mehrstufigen Auswahlverfahren. Dieses ist insbesondere bei aufwändigen Befragungen wie Interviews zu empfehlen.

Geschichtete Stichprobe

Bei der geschichteten Zufallsstichprobe wird die Grundgesamtheit in mehrere weitgehend homogene Gruppierungen (z. B. Altersgruppen der Auszubildenden, Schulabschluss der Auszubildenden), welche als Schichten bezeichnet werden, unterteilt. Danach wird separat aus jeder Gruppierung eine einfache Zufallsstichprobe gezogen. Diese Stichproben werden dann beim Schluss auf die Grundgesamtheit entsprechend den Umfängen der einzelnen Schichten, die bekannt sein müssen, gewichtet. Nach Bortz (1999: 88) kann diese Methode angewendet werden, wenn bekannt ist, welche Determinanten die Verteilung des untersuchungsrelevanten Merkmals beeinflussen und wie sich diese in der Grundgesamtheit verteilen.

Ein Effekt der geschichteten Stichprobe ist, dass die Streuung des untersuchten Merkmals kleiner sein wird als bei der Anwendung der Zufallsstichprobe. Die Streuung der Merkmale beim Vergleich der einzelnen Schichten untereinander wird hingegen größer. Da die Werte der

Grundgesamtheit nicht nur mit zunehmendem Stichprobenumfang, sondern auch mit abnehmender Streuung der untersuchten Merkmalsausprägungen geringer wird, ist die Schichtung eine Möglichkeit, die Verlässlichkeit von Verallgemeinerungen zu erhöhen. Daneben bietet die Schichtung auch die Chance, den notwendigen Stichprobenumfang so klein wie möglich zu halten.

Bezogen auf die Evaluation von Ausbildungsordungen könnten beispielsweise folgende Determinanten Einfluss auf die zu untersuchenden Variablen haben:

Zielgruppe Auszubildende: Geschlecht, Bildungabschlüsse, Alter bzw. Fortschritt in der Ausbildung, Migrationshintergrund.

Zielgruppe Ausbildungsbetriebe/Ausbilderinnen/Ausbilder: Branche, regionale Verteilung, Größe der Betriebe.

Allen drei eben skizzierten Formen der zufallsgesteuerten Stichprobenziehung ist gemeinsam, dass Zufallsauswahlen über die Merkmalsträger (z. B. Personen, Betriebe) der jeweiligen Stichprobe entscheiden. Dem gegenüber stehen die rein bewussten Auswahlverfahren sowie Mischformen der Stichprobenziehung. Als Mischform ist insbesondere die Quotenstichprobe zu nennen.

Als Beispiel besonders heterogener Gruppen können die Ausbildungsbetriebe genannt werden, während in der Gruppe der Auszubildenden und evtl. auch in der Gruppe der Ausbilderinnen und Ausbilder von einer geringeren Heterogenität auszugehen ist.

Es sollte deutlich geworden sein, dass eine Zufallsstichprobe nicht immer die beste Stichprobe für die Klärung von quantitativen Fragestellungen ist. Bestimmte soziale Phänomene sind nur über geschichtete oder über Klumpenstichproben sinnvoll statistisch modellierbar. Hierzu ein Beispiel:

In einem Forschungsprojekt, in dem in Bezug auf Einkommensunterschiede in verschiedenen Anstellungsverhältnissen von Absolventen und Absolventinnen der Büroberufe statistisch valide Varianzmaße berechnet werden sollen, kann in der Regel. nicht mit einfachen Zufallsstichproben gerechnet werden. Denn die Varianzmaße der Einkommen von Berufstätigen in Büroberufen, die über Bund-Länder-Tarife angestellt sind, unterscheiden sich von denen derjenigen, die in privatwirtschaftlichen Anstellungsverhältnissen beschäftigt sind. Eine per se deutlich unterschiedliche Einkommens-Range der Berufstätigen in Bund-Länder-Tarif-Verhältnissen und der Berufstätigen im privatwirtschaftlichen Sektor kann durch die Zufallsstichprobe nicht berücksichtigt werden. Um die Einkommensvarianz statistisch solide zu berechnen, wird für die erste Gruppe eine kleinere (Zufalls-)Stichprobe benötigt als für die zweite Gruppe, sodass bei einem Vergleich der Varianzmaße mit anderen Sampleverfahren operiert werden muss.

Quotenstichprobe

Die Quotenstichprobe gehört mehr zu den bewussten als zu den zufallsgesteuerten Auswahlverfahren. Zur Bestimmung der Quoten haben die Evaluatoren und Evaluatorinnen relativ freie Hand. Letztlich entscheidet deren Erfahrung mit dem Untersuchungsgegenstand über eine geeignete Einbeziehung von Personen bzw. Merkmalsträgern. Die Zusammensetzung der Stichprobe erfolgt hinsichtlich ausgewählter Merkmale und wird über vorgegebene Quoten realisiert. Dies kann z. B. ein prozentualer Anteil an Auszubildenden in einem Beruf je Kammerbezirk sein, wobei beispielweise die Stärke von Ausbildungsjahrgängen berücksichtigt wird.

Theoretische Stichprobe

Um eine Theorie prüfen zu können, werden für bestimmte Forschungsfragen besonders typische oder untypische Elemente (z. B. Personen) ausgewählt. Diese Art der Stichprobenziehung wird primär in der qualitativen Forschung angewandt.

Grundsätzlich zu beachten ist, dass *nicht zufällig ausgewählte* Stichproben in der Regel für *inferenzstatistische Auswertungen* ungeeignet sind. So wird bereits mit der Stichprobenauswahl festgelegt, ob die Daten einer Evaluation nur auf deskriptiver Ebene oder auch auf inferenzstatistischer Ebene bearbeitet werden dürfen. Weitere Ausführungen hierzu könnender Arbeitshilfe Erhebung quantitativer Daten sowie der Arbeitshilfe Skalenniveaus und Auswertung quantitativer Daten entnommen werden.

Repräsentativität

Die wichtigste Frage, die sich im Zusammenhang mit der Stichprobenauswahl stellt, ist die nach der Repräsentativität der gewählten Stichprobe. Die Annahme, dass mit steigendem Stichprobenumfang die Genauigkeit einer Erhebung/Messung steigt, gilt nur für repräsentative Stichproben.

Grundsätzlich gilt: Ist die Stichprobe ein möglichst genaues Abbild der Grundgesamtheit, dann ist sie auch repräsentativ. Dieser Regelsatz wirft aber ein Problem auf: Um wissen zu können, ob die in der Stichprobe gemessenen Merkmalsausprägungen repräsentativ für eine Grundgesamtheit sind, müssen determinierende Merkmalsausprägungen der Grundgesamtheit bekannt sein. Dieser Hinweis ist insofern wichtig, als es obligatorisch geworden ist, nach jeder Befragung reflexartig nach der Repräsentativität der Ergebnisse zu fragen. Sofern sich sowohl die Datenerhebung als auch die Datenauswertung an den Vorgaben guter wissenschaftlicher Praxis orientiert, und sofern die Stichprobenziehung wie oben beschrieben korrekt durchgeführt wurde, darf in der Regel davon ausgegangen werden, dass die erhobenen Daten repräsentativ sind.

Nach Abschluss der Rücklauf-Phase empfiehlt sich eine Non-Responder-Analyse. Insbesondere bei vorhandenen Daten zu determinierenden Merkmalsausprägungen sollte überprüft werden, welche „Gruppe“ sich tatsächlich an der Erhebung beteiligt hat. Wurde eine geschichtete Stichprobe erhoben, können so Rückschlüsse auf den Erfolg der Stichprobenrekrutierung gezogen werden und ggf. eine Nachfassaktion gestartet werden.

Bei der Interpretation von Untersuchungsergebnissen ist die Freiwilligkeit der Teilnahme somit nicht unerheblich. Bortz/Döring (2002: 75–77) charakterisieren freiwillige Untersuchungsteilnehmerinnen und Untersuchungsteilnehmer wie folgt: höhere Bildung, besserer Notendurchschnitt, tendenziell höhere Intelligenz, Freiwillige sind geselliger und brauchen mehr soziale Anerkennung, häufiger Frauen.

Auch wenn sich die Erkenntnisse aus wissenschaftlichen Experimenten nicht unmittelbar auf Evaluationen umlegen lassen, sollte dieser Aspekt zumindest berücksichtigt und in die Planung mit einbezogen werden. Bei Evaluationen hängt die Bereitschaft zur Beantwortung von Fragebogen bzw. die Teilnahme an Interviews häufig vom Grad der (zugesicherten) Anonymität ab und davon, welche Konsequenzen bzw. welcher Nutzen für die Betroffenen aufgrund der Befragung zu erwarten sind.

Es empfiehlt sich also, an die Bereitschaft zur Teilnahme an Befragungen zu appellieren, bei den Einleitungen der verwendeten Fragebogen die Wichtigkeit der Teilnahme zu betonen, sowie Anonymität und Sicherheit zu vermitteln. Gute wissenschaftliche Praxis zeichnet sich in diesem Zusammenhang aber auch dadurch aus, dass die eingesetzten Erhebungsinstrumente fachge-

recht konstruiert, die zu befragenden Personenkreise adäquat ausgewählt wurden sowie die Datenerfassung und Datenauswertung verlässlich erfolgt sind.

Stichprobenplan

Der Stichprobenplan fasst alle wichtigen Eckdaten zusammen, die für die Durchführung der Erhebung wichtig sind. Folgende Fragen sollte der Stichprobenplan beantworten können:

- Welche Merkmalsträger (z. B. Auszubildende, Betriebe, Kammerbezirke) sind für die Untersuchung wichtig?
- Wie groß sind die jeweiligen Grundgesamtheiten dieser Merkmalsträger?
- Welches Auswahlverfahren soll zur Ziehung der Stichprobe herangezogen werden?
- Sind Besonderheiten bzw. Rahmenbedingungen zu berücksichtigen, die mithilfe eines speziellen Auswahlverfahrens berücksichtigt werden können?
- Welchen Umfang soll die jeweils zu ziehende Stichprobe haben?
- Wie groß muss die Gruppe der Merkmalsträger sein, die angeschrieben bzw. angesprochen wird, damit bei einem als relativ normal zu bezeichnenden Rücklauf schließlich die Stichprobengröße erreicht werden kann? (Ist genügend Puffer eingebaut?)
- Bei welcher kritischen (Rücklauf-) Größe muss eine zweite Erhebungswelle oder eine Nachfassaktion gestartet werden?

Literatur

Bortz, Jürgen: Statistik für Human- und Sozialwissenschaftler. Heidelberg 2005.
Bortz, Jürgen; Döring, Nicola: Forschungsmethoden und Evaluation. Berlin 2002.
Bortz, Jürgen: Statistik für Sozialwissenschaftler. Berlin 1999.
Köhler, Wolfgang; Schachtel, Gabriel; Voleske, Peter: Biostatistik. Heidelberg 2007.
Kromrey, Helmut: Empirische Sozialforschung. Stuttgart 2006.
Meyer, Wolfgang (2007). Datenerhebung: Befragungen – Beobachtungen – Nicht-reaktive Verfahren. In Stockmann, Reinhard, Handbuch zur Evaluation. Eine praktische Handlungsanleitung. Münster 2007, S. 223–277.
Stockmann, Reinhard; Meyer, Wolfgang: Evaluation. Eine Einführung. Opladen 2010.

2 Qualitative Methoden

Zu den qualitativen Methoden zählen solche, die der Erhebung nicht standardisierter (nicht numerischer) Daten dienen. Diese Daten werden meistens durch Befragungen erhoben. Neben den sogenannten Interview- oder Gesprächsverfahren kommen auch Beobachtungen (z. B. in Form von Betriebsbesichtigungen) und sogenannte nicht-reaktive Verfahren (z. B. durch Analysen sogenannter „natürlicher Gespräche“ oder sogenannter „Verhaltensspuren“) zum Einsatz.

In der Regel werden wegen des hohen zeitlichen Aufwands, den qualitative Methoden verursachen, nur geringe Fallzahlen[34] in die Untersuchung mit einbezogen. Dies ist aber nur eine „technische“ Argumentation. Vielmehr muss bei qualitativen Studien ein ganz anderes Ziel bedacht werden, das im Folgenden näher erläutert wird.

Vorüberlegungen: Das Ziel qualitativer Studien

Das Ziel qualitativer Studien ist keine statistische Repräsentativität, sondern die *phänomenologische* – d. h. umfassende und vielschichtige – *Repräsentation* komplexer sozialer Wirklichkeiten,

[34] In der qualitativen Forschung werden Fallzahlen als „Fallauswahlen“ bzw. „qualitative Samples“ bezeichnet – in Abgrenzung zu Stichproben.

die hermeneutisch rekonstruiert (Erkenntnisprinzip des verstehenden Nachvollzugs) und als *Muster* umfassend dargestellt werden. Ausführliche Erläuterungen hierzu finden sich u. a. bei Kruse (2010, Oktober: 9 ff.). Muster stellen dabei sinnstrukturelle Konsistenzen im Hinblick auf Gemeinsamkeiten und Unterschiede in den untersuchten sozialen Phänomenen dar. Sinnstrukturelle Konsistenzen sind hier als *qualitative Regelmäßigkeiten* gemeint, nicht im Sinne von Häufigkeitsverteilungen, wie in der quantitativen Forschung.

Bildlich bedeutet dies: Muster stellen spezifische Sinnfiguren oder Gesamtgestalten dar, wie bei einem Mosaik: Ein Mosaik – als soziales Phänomen – ist aus unterschiedlichen Teilphänomenen aufgebaut (Sinnbausteinen), die zueinander in ihrer strukturellen Ordnung ein spezifisches Muster bzw. Motiv bilden. Qualitative Forschung hat in dieser Hinsicht sozusagen die Aufgabe, in einem großen Mosaik – oder in verschiedenen Mosaiken, die miteinander verglichen werden – Gemeinsamkeiten und Unterschiede in den jeweiligen sinnstrukturellen Kompositionen aufzudecken.

Mögliche Formen der Fallauswahl

Um diese phänomenologische Repräsentation in qualitativen Studien auf der Ebene der untersuchten Erhebungs- bzw. Falleinheiten zu erreichen, wird aber eine spezifische Anlage der Fallauswahl, des *qualitativen Samples*, notwendig. Das Grundprinzip qualitativer Samples sind *kontrastierende* bzw. *komparative* Fallauswahlen. Die Kontrastierungsdimensionen in Hinblick auf das qualitative Sample können dabei sehr unterschiedliche sein (s. u.). Ziel dessen ist, die *Heterogenität des Feldes*, vgl. Kelle/Kluge (1999: 38 ff.); Merkens (2003); Kruse (2010, Oktober: 81 ff.) zu berücksichtigen. Dies wird jedoch eben nicht über statistische Verfahren der Samplebildung zu erreichen versucht, sondern durch eine *bewusste Fallauswahl*, die in methodischer Anlehnung an das „*theoretical sampling*“ von Glaser/Strauss (im Überblick Strauss/Corbin 1996: 148 ff.) nach dem Prinzip der *maximalen* bzw. *minimalen strukturellen Variation* operieren. Siehe weiterführend Kleining (1982); Kelle/Kluge (1999: 44 ff.).

Hierzu ist jedoch in einem ersten Schritt erforderlich, die „Grundgesamtheit“ des zu befragenden Feldes, also das empirische Feld, aus dem Interviewpersonen bzw. die Erhebungseinheiten falltypologisch befragt werden sollen, von außen her einzugrenzen, um so eine Kontrastierung nach innen besser zu ermöglichen, vgl. Merkens (2003); Helfferich (2005); Kelle/Kluge (1999: 38 ff.). Die Grundgesamtheit wird dabei nur über theoretische Überlegungen in einem qualitativen Sinne bestimmt (welche Erhebungseinheiten gehören prinzipiell dazu, welche nicht) und nicht wie in der quantitativen Forschung über bestimmte quantitative Bestimmungsverfahren, da es ja um eine qualitative Repräsentation spezifischer „Falltypen“ geht, nicht um eine Gewährleistung der Stichprobe in statistischer Hinsicht.

Das Prinzip der maximalen und minimalen strukturellen Variation

Das Prinzip der *maximalen strukturellen Variation* verfolgt dabei die Logik, dass gemeinsame Muster, die in ganz unterschiedlichen Fall-Lagerungen rekonstruiert werden konnten, die Tragweite der getroffenen Aussagen über die befragten Fälle hinweg abstrahieren können, jedoch eben nicht im induktiv-statistischen Sinne.

Das Prinzip der *minimalen strukturellen Variation* soll in dieser Hinsicht im Prinzip einen Gegenhorizont bilden können: Es soll untersucht werden, inwiefern bei ganz ähnlichen Fällen in Bezug auf spezifische Teilphänomene auch sinnstrukturelle Unterschiede herausgearbeitet werden können.

Beide Strategien werden in der Regel in einem Forschungsprojekt parallel verfolgt: Werden nun die jeweiligen Ergebnisse in Bezug auf die sinnstrukturellen Muster, die auf der Basis der beiden Samplingstrategien gewonnen werden konnten, sozusagen übereinander geblendet, kann

ein umfassendes und ganzheitliches Gesamtbild des untersuchten sozialen Phänomens generiert werden.

So können beispielsweise für die maximale strukturelle Variation im Rahmen der Evaluation von Ausbildungsordnungen viele unterschiedliche Fälle gebildet bzw. Personen befragt werden: Auszubildende, Berufsanfängerinnen und Berufsanfänger, Ausbilderinnen und Ausbilder, Lehrkräfte, Mitarbeiterinnen und Mitarbeiter aus Betrieben und Kammern sowie Weiterbildungskräfte usw. Und das nicht nur in einer Region sondern in verschiedenen Regionen; zudem werden Personen mit unterschiedlichem Berufsalter befragt usw.

Für eine minimale strukturelle Variation würden nur Ausbilderinnen und Ausbilder, nur Auszubildende oder nur Fachkräfte aus Berieben befragt, und hierbei jeweils viele Fälle erhoben, die sich stark ähneln, z. B. Auszubildende nur aus einer spezifischen Kohorte, um Gemeinsamkeiten und Unterschiede in diesem eng umgrenzten sozialen Feld umfassend und genau herausarbeiten zu können.

Die beiden Beispiele verdeutlichen nochmals, dass in der Forschungspraxis in der Regel. beide Strategien zugleich (partiell) verfolgt werden.

Rekrutierungsstrategien

Diese allgemeinen Überlegungen zeigen, dass in qualitativen Forschungsprojekten in der Regel andere Rekrutierungsstrategien als in standardisierten Projekten erforderlich sind, um so die bewusste Kontrastierung von Erhebungseinheiten, d. h. Fällen, zu erreichen:

Die am häufigsten verwendeten Rekrutierungsstrategien sind dabei die Einbeziehung von Gutachtern, z. B. andere Forscherinnen und Forscher in dem Untersuchungsfeld, Kennerinnen und Kenner der „Szene“, Multiplikatorinnen und Multiplikatoren (Fachkräfte, die in dem Untersuchungsfeld in anderen beruflichen Funktionen tätig sind, z. B. Ausbilderinnen oder Ausbilder, Weiterbildungsfachkräfte) oder Gatekeeper, die Zugangsmöglichkeiten verschaffen (Fachkräfte, die in dem Untersuchungsfeld eine zentrale berufliche Position innehaben). Es wird ein kontrolliertes und reflektiertes Schneeballprinzip („Kennen Sie jemanden, der jemand kennt, der für mich interessant ist, zu befragen?“) initiiert, um so zu einer Erweiterung des Kontaktfeldes zu gelangen, die Annoncierung bzw. die Werbung über Infoflyer oder das direkte Aufsuchen im Feld („pick-up“).

Alle Verfahren weisen dabei in ihrer inneren Logik Verzerrungsmöglichkeiten auf, sodass in der Forschungspraxis verschiedene Rekrutierungsstrategien gleichzeitig gewählt werden sollten.

Sampleverfahren

Das Prinzip der minimalen bzw. maximalen strukturellen Variation kann grundsätzlich über zwei verschiedene Sampleverfahren erreicht werden:

Theoretisch begründete Vorabfestlegung des Samples

Das erste Sampleverfahren ist die theoretische Vorabfestlegung von Fällen anhand verschiedener Merkmalsausprägungen. Hier wird nach dem Prinzip der maximalen strukturellen Variation eine Spanne von extrem unterschiedlichen Erhebungseinheiten, d. h. Fällen, aufgebaut. Die Merkmalskategorien werden von vornherein begründet, um dann passende Interviewpersonen zu suchen, vgl. Kelle/Kluge (1999: 46 ff.). Solche Varianzmerkmale können z. B. standarddemografische Aspekte sein (Alter, Geschlecht, soziale Herkunft, Bildungsniveau etc.) oder forschungsthematisch spezifische Aspekte, die im Verlauf der Forschung deutlich werden. So könnte z. B. in der Evaluation von Ausbildungsordnungen angenommen werden, dass *Lernstile*

oder *Lernpräferenzen* der Auszubildenden ein zentrales thematisches Merkmal sind, oder die *spezifischen Branchenstrukturen* eine hohe Bedeutung aufweisen.

Die Gefahr bei diesem Verfahren ist, dass man verschiedenen *Kategorienfehlern* aufsitzt, d. h., dass man z. B. annimmt, dass die Kategorie Geschlecht Unterschiede produziert – und man so im Laufe der Forschung bestehende Stereotype reproduziert.

Zudem weist die eventuell gerade erst im Forschungsprozess deutlich gewordene Verfolgung spezifischer thematischer Aspekte auf eine zweite grundsätzliche Samplingstrategie hin, die des theoretical samplings.

Begründung im Verlauf des Datenerhebungsprozesses – das „theoretical sampling" der Grounded Theory

Das zweite Sampleverfahren ist das durch Glaser/Strauss, vgl. Strauss/Corbin (1996: 148 ff.) begründete Verfahren des „theoretical sampling", vgl. Kelle/Kluge (1999: 44 ff.) im Zusammenhang des Forschungsprogramms der „Grounded Theory", vgl. Strauss/Corbin (1996: 3 ff.). Hier erfolgt die Begründung einer Fallauswahl mit maximal bzw. minimal strukturell variierenden Fällen erst im Verlauf der Feldforschungsphase bzw. des Datenerhebungsprozesses: Ausgehend von der Analyse eines ersten Interviews bzw. einer ersten Erhebungseinheit wird nach weiteren Interviewfällen bzw. Erhebungseinheiten gesucht, die sich von den vorherigen falltypisch unterscheiden. Die Varianzmerkmale werden also erst im Forschungsprozess festgelegt: So könnte z. B. in der Evaluation von Ausbildungsordnungen bei der bisherigen Fallauswahl nicht daran gedacht worden sein, dass *Lernstile*, *Lernpräferenzen* und *Branchenstrukturen* eine hohe thematische, d. h. sinnstrukturelle Bedeutung haben für das Untersuchungsfeld. Allerdings zeigt sich dies nun durch die ersten Interviews mit verschiedenen Personen im Untersuchungsfeld, sodass diese spezifischen forschungsthematischen Aspekte weiter verfolgt werden, und nun gezielt ähnliche oder sehr unterschiedliche Interviewpersonen gesucht werden, die zu diesen Themen einen direkten Feldbezug haben, um die Bedeutung dieser thematischen Dimensionen weiter zu klären.

Fallauswahl im Rahmen der Evaluation von Ausbildungsordnungen

Die Fallauswahl im spezifischen Feld der Evaluation von Ausbildungsordnungen erfolgt in der Regel. nach dem Prinzip der maximalen, strukturellen Variation. Auch hier gilt, dass im Vergleich zur quantitativen Forschungslogik, bei der beispielsweise isolierte Merkmale gemessen werden und durch soziale Sachverhalte bestätigt oder eben nicht bestätigt werden können, die Erfassung der Komplexität sowie das grundsätzliche „verstehen wollen" sozialer Phänomene im Vordergrund steht.

Die Untersuchungseinheiten, z. B. Auszubildende, Ausbilderinnen und Ausbilder, Lehrkräfte aber auch Betriebe, werden im Rahmen eines qualitativen Ansatzes, der sich zudem an die Grounded Theory anlehnt, zusätzlich eher *sukzessive* ausgewählt, entlang eines spiralförmig-dynamischen Erkenntnisprozesses.

Für die Auswahl von Untersuchungseinheiten im Rahmen der Evaluation von Ausbildungsordnungen bedeutet diese sukzessive Vorgehensweise, dass z. B. in einem bestimmten Kammerbezirk Interviews mit Vertreterinnen und Vertretern der zuständigen Stellen, Ausbilderinnen und Ausbildern, Lehrkräften und ggf. auch mit Auszubildenden selbst geführt werden. Ein ähnliches Setting sollte dann z. B. in einer anderen Region gewählt werden und anschließend auch in Regionen mit anderen infrastrukturellen Rahmenbedingungen (Prinzip der maximalen strukturellen Variation). Das Hauptanliegen dieser Erhebungen besteht darin, möglichst viele unterschiedliche Informationen zu sammeln, die den Erkenntnisprozess anreichern. Können ab einem bestimmten Zeitpunkt keine neuen Informationen mehr gewonnen werden, kann dieser Prozess

beendet und evtl. ein anderer Fokus gewählt werden, Prinzip der Sättigung, vgl. Strauss/Corbin (1996: 148 ff.).

Häufig wird die Zusammensetzung der Untersuchungseinheit im Rahmen eines qualitativen Vorgehens auch als *Fallstudie* bezeichnet.

Literatur

Helfferich, Cornelia: Qualität qualitativer Daten – Manual zur Durchführung qualitativer Einzelinterviews. Wiesbaden 2005.

Kelle, Udo; Kluge, Susann: Vom Einzelfall zum Typus. Opladen 1999.

Kleining, Gerhard: Umriss zu einer Methodologie qualitativer Sozialforschung. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 34 (1982) 2, S. 224–253.

Kruse, Jan: Einführung in die Qualitative Interviewforschung (Reader). Online-Publikation. Freiburg 2011.

Merkens, Hans: Auswahlverfahren, Sampling, Fallkonstruktion. In: Flick, Uwe: Qualitative Forschung. Ein Handbuch. Reinbek 2003, S. 286 ff.

Strauss, Anselm; Corbin, Juliet: Grounded Theory: Grundlagen qualitativer Sozialforschung. Weinheim 1996.

Arbeitshilfe: Erhebung quantitativer Daten

Die Erhebung rein quantitativer Daten ist neben der Erhebung qualitativer Daten zentraler Bestandteil fast jeder Evaluation. Welche Methoden es gibt und was bei ihrem Einsatz zu berücksichtigen ist, soll diese Arbeitshilfe erörtern.

Vorüberlegungen zur Herkunft der Daten

Nach Bortz/Döring (2002: 374) können drei Datenquellen unterschieden werden: Die Nutzung vorhandener Daten, die Datenbeschaffung durch Dritte und die eigene Datenbeschaffung.

Nutzung vorhandener Daten/Sekundärdaten

Um eine Evaluation effizient zu gestalten, sollte nach Möglichkeit auf bereits bestehende Daten zurückgegriffen werden. So werden in der Regel bei der Evaluation von Ausbildungsordnungen vorhandene Daten des statistischen Bundesamtes, der DIHK, der KMK u. ä. m. genutzt. Sie liefern Auskünfte über die Vertragslösungsquoten, das Geschlechterverhältnis innerhalb eines Ausbildungsberufes usw. Zur Kategorie *Sekundärdaten* zählen weiterhin Dokumente, die im Rahmen der Evaluation untersucht werden können, wie beispielsweise Sitzungsprotokolle vorausgegangener Ordnungsverfahren, Statistiken von Verbänden oder auch Publikationen über den Beruf.

Datenbeschaffung durch Dritte/Primärdaten

In der Regel können Evaluationen von Ausbildungsordnungen jedoch nicht ausschließlich mittels einer Analyse von Sekundärdaten durchgeführt werden. Es bedarf ebenso einer Erhebung von neuen Daten, sogenannter Primärdaten. Bortz/Döring gehen davon aus, dass insbesondere beim Anspruch auf Repräsentativität von schwer zu erreichenden Populationen oder bei besonders großen Grundgesamtheiten die Vergabe der Datenerhebung an kommerzielle Anbieter sinnvoll sein kann. Die Einbindung Dritter birgt unter Umständen aber auch Risiken in sich: So weisen nur wenige Evaluatorinnen und Evaluatoren oder Marktforscherinnen und Marktforscher fundierte Kenntnisse des Ausbildungsordnungsgeschäftes auf, sind selbst keine Berufe-Experten/-Expertinnen und haben folglich nur bedingt Erfahrung mit dem Evaluationsgegenstand.

Eigene Datenbeschaffung/Primärdaten

Andererseits können die Primärdaten auch durch das BIBB selbst erhoben werden. Bei ausreichenden personellen und finanziellen Kapazitäten sollte immer über den vollständigen Verbleib der Primärdatenerhebung im BIBB nachgedacht werden.

In beiden Fällen der Primärdatenerhebung ist eine gründliche Planung von großer Bedeutung.

Grundlagen zur Primärdatenerhebung

Im Unterschied zur *qualitativen* Datenerhebung, die durch eine nicht standardisierte oder nur teil-standardisierte Herangehensweise geprägt ist (offene Fragen), zeichnet sich die *quantitative* Datenerhebung durch starke Strukturierungen (geschlossene Fragen) aus.

Die Daten werden mittels **vollstandardisierter Verfahren** *schriftlich* (Fragebogen per Post, online oder classroom), *mündlich* (Interviews Face-to-Face oder per Telefon/CATI[35]) oder *beobachtet* (Zählungen oder Begehungen) erhoben.

Werden quantitative Methoden eingesetzt, so müssen sie in der Form gestaltet sein, dass sie die **Gütekriterien** (Objektivität, Reliabilität und Validität) erfüllen.

Ziel vollstandardisierter Verfahren ist es, eine numerische Darstellung von Sachverhalten zu erlangen, um damit statistische Berechnungen bzw. Auswertungen zu ermöglichen. Welche Möglichkeiten es dabei gibt, erläutert die Arbeitshilfe: Skalenniveaus und Auswertung quantitativer Daten.

Speziell für die statistische Verifizierung oder Falsifizierung von Annahmen bzw. Hypothesen, also für die Inferenzstatistik, ist es äußerst wichtig

a. im Rahmen der Konstruktion (beispielsweise eines Fragebogens) zu klären, was durch die Daten ausgesagt werden soll, d. h. welche statistischen Berechnungen möglich sein sollen und
b. in Erfahrung zu bringen, wie viele Personen bzw. Einheiten befragt werden können (Größe der Grundgesamtheit) und wie viele befragt werden müssen (Größe der Stichprobe).

Die gewünschte *statistische Auswertung* gibt vor, welches Skalenniveau bei den einzelnen Antwortmöglichkeiten zu wählen ist und wie viele Personen oder Einheiten mindestens befragt werden müssen.

Entscheidend für die Erstellung der Erhebungsinstrumente sind die forschungsleitenden Fragestellungen. Grob kategorisiert können im Rahmen von Evaluationen folgende Fragestellungen auftauchen, die über eine reine Beschreibung/Deskription von Sachverhalten hinausgehen:

- **Zusammenhänge:** Gibt es Zusammenhänge zwischen interessierenden Variablen? (z. B. Dauer der Berufserfahrung und Einstellung zu bestimmten Fragen)
- **Unterschiede:** Gibt es Unterschiede zwischen interessierenden Gruppen (unabhängige Variable z. B. Bundesland, Branche, Wahlqualifikation) in definierten abhängigen Variablen (z. B. Zufriedenheit, Bestehensquote, Qualifikation von Ausbildern/Ausbilderinnen)?
- **Verläufe:** Gibt es Veränderungen im zeitlichen Verlauf? (z. B. Bestehensquote in einem vorher definierten Zeitraum)
- **Übereinstimmungen:** Wie übereinstimmend wird ein bestimmter Sachverhalt von mehreren Beurteilern eingeschätzt? (z. B. Einschätzungen der Befragten zum Praxisbezug von Prüfungsaufgaben)
- **Modelltestungen:** Wie gut eignen sich bestimmte Prädiktoren zur Vorhersage eines definierten Sachverhalts? (z. B. Aussagewert von Abschlussprüfung oder Schulnoten in Bezug auf die von den Betrieben erwartete Handlungskompetenz)

Die Voraussetzungen zur Beantwortung solcher Fragestellungen müssen schon bei der Indikatorenentwicklung, ganz entscheidend jedoch bei der Zusammenstellung und Auswahl der Erhebungsinstrumente geschaffen werden. Eine optimale Stichprobengröße soll dabei die statistische Aussagekraft stärken und eine ökonomische Vorgehensweise sichern.

Zum Thema „standardisierte Befragung“ existiert eine Fülle an Fachliteratur, die auf den letzten Seiten dieser Arbeitshilfe auszugsweise als Literaturtipps aufgeführt ist.

Viele Lehrbücher widmen sich dem Thema Fragebogenkonstruktion sowie dem Aspekt der Item- bzw. Fragenformulierung. Da der letztgenannte Aspekt nicht nur für die Erhebung quanti-

[35] Computer Assisted Telephone Interview (CATI).

tativer Daten von Bedeutung ist, sondern genauso auch für qualitative Verfahren, finden sich nachfolgend einige ausgewählte zentrale Hinweise.

Schriftliche Befragung

Postalisch

Der Fragebogen wird den Probanden per Post oder per E-Mail zugeschickt. Sie beantworten den Fragebogen ohne die Anwesenheit einer „Aufsichtsperson".

Die **Vorteile** einer schriftlichen Befragung per Post oder E-Mail lassen sich wie folgt zusammenfassen:

- Viele Personen können zeitgleich befragt werden.
- Sie sind leichter erreichbar als per Telefon oder durch einen persönlichen Besuch.
- Die Beantwortung der Fragen kann zeitlich vom Probanden selbst festgelegt werden.
- Es entsteht kein Einfluss durch den Interviewer bzw. die Interviewerin (Interviewer-Bias).
- Der durch die Befragten wahrgenommene Anonymitäts-Grad ist relativ hoch.
- Gerade wenn der Fragebogen per E-Mail zugeschickt werden kann, sind die Kosten zu Beginn gering. Die Kosten der Dateneingabe nach erfolgreicher Zurücksendung ausgefüllter Fragebogen sind jedoch nicht außer Acht zu lassen. Je nach Stichprobengröße kann die Dateneingabe durchaus teuer werden.

Im Vergleich dazu lassen sich auch **Nachteile** von schriftlichen Befragungen per Post oder E-Mail festhalten:

- Die Erhebungssituation ist kaum kontrollierbar. Ein Teil der Befragten tendiert dazu, den Fragebogen nicht oder nur unvollständig auszufüllen.
- Es ist nicht nachzuvollziehen, in welcher Situation der/die Proband/-in den Fragebogen ausgefüllt hat: in relativer Ruhe/Unruhe in der Straßenbahn auf dem Weg zur Arbeit, im Falle von Prüflingen beispielsweise direkt im Anschluss an eine anstrengende schriftliche Prüfung, oder kurz vor Feierabend, mit den Gedanken bereits auf dem Nachhauseweg.
- Die Rücklaufquote von schriftlichen Befragungen liegt nicht selten unter 50 %.

Classroom

Hierbei handelt es sich um eine schriftliche Befragung einer in einem Raum anwesenden Gruppe unter „Aufsicht". Classroom-Befragungen werden im Rahmen von Ausbildungsordnungsevaluationen (z. B. im Anschluss an schriftliche Prüfungen) durchgeführt. Die „Aufsichtsperson" ist in diesen Fällen für die Ausgabe und das Einsammeln der Fragebögen zuständig und steht für Rückfragen der Probanden zur Verfügung. Dies setzt jedoch voraus, dass die Aufsichtsperson, beispielsweise mit einem separaten Anschreiben, ausreichend über die Erhebung informiert ist.

Der große Unterschied zur o. g. postalischen Befragung besteht darin, dass die Classroom-Befragung auf die Nachteile der postalischen Befragung reagieren kann: Die Erhebungssituation ist relativ kontrollierbar; die Rücklaufquote ist in der Regel hoch. Durch die Anwesenheit einer Aufsichtsperson sowie die empfundene „Verpflichtung" zur Teilnahme ist jedoch auch die Gefahr gegeben, dass die Probanden nicht gänzlich authentisch antworten.

Online

Hierbei wird der Fragebogen digital auf einer Online-Plattform zur Verfügung gestellt. Der Zugang zum Fragebogen kann durch ein Passwort geschützt sein, das nur an die Probanden vergeben wird, die an der Befragung teilnehmen.

Die Vergabe eines Passworts setzt voraus, dass die Probanden bekannt sind und beispielsweise per E-Mail kontaktiert werden können und mit dieser E-Mail das Passwort erhalten. Dies hat den großen Vorteil, dass dadurch die genaue Stichprobengröße bekannt ist. Wird der Online-Fragebogen nicht passwortgeschützt und beispielsweise auf einem Portal eingestellt, das die Zielgruppe häufig nutzt, kann den jeweiligen Fragebogen prinzipiell jede Person ausfüllen, wodurch die angesprochene Stichprobengröße nicht mehr kontrollierbar ist.

Online-Befragungen unterscheiden sich erheblich von postalischen oder Classroom-Befragungen: Während Papier sprichwörtlich geduldig ist, besitzt das Internet eine Ausstrahlung von Unruhe und Ungeduld.

Die folgenden Hinweise sollten bei der Erstellung eines Online-Fragebogens unbedingt berücksichtigt werden.[36]

- Eine klare und verständliche Sprache ist noch entscheidender als üblicherweise.
- Vermeiden Sie unbedingt lange Sätze. Bildschirmseiten werden in der Regel eher ‚gescannt' als sorgfältig gelesen. Während man es hinnimmt, auf einem gedruckten Stück Papier einen Satz auch zweimal zu lesen, wird das im Internet kaum akzeptiert.
- Eine Online-Befragung sollte nicht länger als 10–15 Minuten in Anspruch nehmen.
- Programmieren Sie auf jeden Fall einen Fortschrittsanzeiger (Balken mit %-Angaben) auf jeder Seite ein, sodass der/die Proband/-in immer verfolgen kann, wie weit die Befragung fortgeschritten ist. Eine realistische Fortschrittsanzeige ist geeignet, den Durchhaltewillen zu stärken.
- Beachten Sie beim Layout des Fragebogens die Eigenarten unterschiedlicher Bildschirmformate sowie die unterschiedlichen Darstellungsarten je nach gewähltem Browser (Explorer, Firefox, Google, Apple, etc.). Text, der über die gesamte Bildschirmbreite läuft, wird in der Regel vom User nicht akzeptiert.
- Ein professionelles Design wird von den Online-Usern vorausgesetzt, sie haben einen geschulten Blick für Dilettantismus. Minimalismus wird bevorzugt, weshalb größtenteils auf farbige Schriften oder überflüssige Grafikelemente verzichtet werden sollte.
- Berücksichtigen Sie, dass bei einem Online-Fragebogen das Vor- und Zurückblättern vergleichsweise schwerfälliger geschieht als bei Papierfragebogen. Rückverweise auf vorangegangene Fragen sollten daher besser nicht erfolgen.
- Ein großer Unterschied zwischen Online- und Papierfragebogen besteht darin, dass das Papier zwischendrin zur Seite gelegt werden kann und der Fragebogen zu einem späteren Zeitpunkt fertig ausgefüllt wird. Die Frustrationstoleranz ist höher. Die „Rückkehrfunktion" ist durch eine entsprechende Programmierung auch bei Online-Befragungen möglich, wird jedoch nur selten von den Probanden genutzt. Die Frustrationstoleranz bei Online-Befragungen ist erheblich niedriger.

Die **Vorteile eines Online-Fragebogens** liegen im Vergleich zu klassischen paper pencil-Befragungen in den folgenden Bereichen:

- Filterfragen können einfach eingebaut werden.
- Es können Kontrollen für unlogisches oder inkonsistentes Antwortverhalten programmiert werden.
- Der postalische Versand der Fragebogen entfällt. Gleichzeitig ist jedoch der Aufwand für eine konsistente Programmierung des Online-Fragebogens nicht zu unterschätzen.
- Einmal programmiert, stellt der Umfang der befragten Personen (vorausgesetzt aktuelle E-Mail-Adressen liegen vor bzw. die Probanden nutzen häufig das Portal, auf dem der Link eingestellt ist) nicht den kostenentscheidenden Faktor dar.

[36] Diese Hinweise sind größtenteils aus Kuckartz (2009: 34–37) entnommen.

- Die Daten werden automatisch erfasst, es bedarf in der Folge keiner aufwendigen manuellen Datenerfassung.
- Durch die direkte automatisierte Datenerfassung sinkt die Fehlerquote durch Übertragung auf 0 %.

Delphi-Methode

Nach Bortz/Döring (2002: 261) handelt es sich bei der Delphi-Methode um eine spezielle Form der schriftlichen Befragung. Sie ist eine hochstrukturierte Gruppenkommunikation, deren Ziel es ist, aus Einzelbeiträgen der an der Delphi-Studie beteiligten Personen konsensorientierte Lösungen für komplexe Probleme zu erarbeiten. Sie findet besonders dann Anwendung, wenn es um die Einschätzung und Vorhersage von Sachverhalten geht, die nicht direkt abgebildet werden können, da sie nicht aktuell präsent oder real existent sind (z. B. weil sie in der Zukunft oder Vergangenheit liegen). Bei den meisten Delphi-Studien liegt der Zeithorizont in der Zukunft (10 oder mehr Jahre).

Die Grundidee der Delphi-Methode besteht darin, über mehrere Befragungswellen Expertenmeinungen zu definierten Problemstellungen einzuholen und diese Erkenntnisse z. B. für einen Blick in die Zukunft oder als Entscheidungshilfen zu nutzen.

Bis zum heutigen Tage gibt es keine einheitliche Definition der Delphi-Methode. Als Merkmale der Delphi-Methode führen Häder/Häder (1994) folgende Punkte an:

- Verwendung eines formalisierten Fragebogens, der sich an Expertinnen und Experten richtet.
- Vergleich von Einzelantworten zur statistisch ermittelten Gruppenantwort.
- Mehrfache Wiederholung der Befragung.
- Die Expertinnen und Experten bleiben untereinander anonym.

Die Vorgehensweise bei einer Delphi-Methode lässt sich grob wie folgt beschreiben, vgl. Glinz (2005), wobei es grundsätzlich zwei Formen der Delphi-Methode gibt, die paper-pencil Methode und die Delphi-Konferenz Häder/Häder (1994 u.1998):

In einem *ersten Forschungsschritt* werden Experten und Expertinnen damit beauftragt, eine Schätzung abzugeben.

In einem *zweiten Schritt* wird versucht, die Meinungen dieser Expertinnen und Experten zu relativieren. Hierfür gibt es unterschiedliche Herangehensweisen, die nachfolgend kurz skizziert werden:

a. In Variante eins werden einzelnen Schätzerinnen und Schätzern die Ergebnisse der Expertinnen und Experten zur Verfügung gestellt. Nach Einsichtnahme in die zur Verfügung gestellten Schätzergebnisse werden die Experten und Expertinnen erneut beauftragt eine Schätzung abzugeben.
b. In Variante zwei werden die Ergebnisse der Expertinnen und Experten an eine Diskussionsleitung gegeben, welche die einzelnen Ergebnisse miteinander vergleicht. Bei großen Abweichungen von der allgemeinen Einschätzung muss sich der oder die jeweils Betroffene vor den anderen Experten und Expertinnen und der Diskussionsleitung rechtfertigen.
c. In Variante drei folgt nach jeder Schätzung eine gemeinsame Diskussion zwischen den Expertinnen und Experten, wobei Abweichungen wie bei Variante zwei durch den jeweiligen Experten oder die jeweilige Expertin zu rechtfertigen sind.

Dieser zweite Abschnitt wird so lange wiederholt, bis die Ergebnisse bzw. deren Abweichungen untereinander innerhalb eines vorgegebenen Rahmens bleiben. Das Mittel der letzten Schätzrunde bildet dann das letztendliche Ergebnis.

Häder/Häder (1994) gehen von einer zunehmenden Spezialisierung der Experten und Expertinnen und von einer zunehmenden Komplexität von Entscheidungen aus. Einen wesentlichen Einsatzbereich für die Delphi-Methode sehen sie (unter Berufung auf mehrere Autorinnen und Autoren) im Bildungswesen bei der Evaluation von Bildungsinhalten.

Des Weiteren wird diese Methode für die Bestimmung von Entwicklungsprognosen eingesetzt. Gewonnene Beschreibungen zukünftiger Entwicklungen können als Grundlage für die Bestimmung von Zukunftsszenarien genutzt werden. Die sorgfältige Auswahl der Experten und Expertinnen ist eine wichtige Voraussetzung zur Durchführung. Aufgrund des hohen Kosten- und Zeitaufwands wird sie meist nur in großen Projekten eingesetzt.

Mündliche Befragung

Survey (Face-to-Face-Interviews)

Unter einem Survey wird eine mündliche Befragung durch einen Interviewer oder eine Interviewerin mithilfe eines voll-standardisierten Fragebogens (in Papierform oder digital auf dem Laptop des Interviewers oder der Interviewerin) verstanden.

CATI (Computer assisted/gestützte Telefon Interviews)

CATI bezeichnet eine telefonische Befragung durch einen Interviewer oder eine Interviewerin mithilfe eines voll-standardisierten computergestützten Fragebogens. CATI wird in der Regel von einem Telefonstudio aus durchgeführt, welches neben der eigentlichen Computerunterstützung der Befragung – Steuerung des Interviews, sofortige Eingabe der Daten etc. – auch die automatische Wahl von Telefonnummern und Zuordnung von Befragten zu Interviewern erlaubt.

► Survey und CATI im Vergleich

Der Kontakt ausschließlich übers Telefon wirkt auf manche Befragte unpersönlich. Das direkte Gespräch vermittelt hingegen Wertschätzung und Wichtigkeit. Bei Face-to-Face-Befragungen lässt sich über nonverbale Kommunikationselemente (z. B. Interviewer-Auftreten und nicht zuletzt der persönliche Besuch an sich) die Wichtigkeit des oder der Befragten für die Untersuchung unterstreichen, vgl. Jahr/Edinger (2008: 29).

Die CATI-Befragung weist im Gegensatz dazu eine hohe zeitliche Flexibilität auf.

Bei beiden Erhebungsmethoden ist die Nutzung von Filterfragen sehr gut möglich. Die Filterführung ist dann optimal einzusetzen, wenn der Fragebogen digitalisiert ist und die Führung automatisch erfolgt. Eine automatische Filterführung entlastet den Interviewer bzw. die Interviewerin enorm.

Hinsichtlich der Dauer sollte für beide Befragungsarten der Richtwert 20–30 Minuten betragen, wobei die Bereitschaft auch länger für ein Interview zur Verfügung zu stehen, bei Surveys höher ist als bei CATI-Befragungen. Je nach Befragungsgruppe wird ein Telefonat, das länger als 20 Minuten dauern soll, auch äußerst kritisch gesehen.

► Fehlerquellen bei mündlichen Befragungen

Ergebnisse, die durch Befragungen gewonnen werden, können nur die Aussagen der Personen widerspiegeln, sozusagen die „subjektive Wahrheit“ der Befragten. Damit kann zwar ein Rück-

schluss auf das Verhalten einer Person gezogen werden, objektive Daten können damit jedoch nicht ermittelt werden. Bortz/Döring (2002: 232) nennen diesbezüglich drei Fehlerquellen:

- Antworttendenzen,
- Selbstdarstellung,
- soziale Erwünschtheit.

Nach Bühner (2004: 56) spielt die Motivation der Befragten bei der Behandlung von Items eine wesentliche Rolle. Diese kann durch eine klare und einfache Formulierung der Items sowie eine vertretbare Länge des Erhebungsinstruments gesteigert werden. Auch die Reihenfolge der Items spielt eine Rolle bei der Beantwortung. Ja-sage- (Zustimmungs-) oder Nein-sage-Tendenz gehören zu den häufigsten Erscheinungen bei Antwort-Tendenzen.

Während diese beiden Tendenzen Antworten im Extrembereich nach sich ziehen, kann sich ein Teil der Befragten nicht festlegen und wählt die mittlere Kategorie, um eine differenzierte Urteilsabgabe zu vermeiden, vgl. Bortz/Döring (2002: 236).

Die beiden Autoren führen eine Reihe von Antwortverfälschungen im Rahmen von Interviews an, die durch wissenschaftliche Studien belegt sind, vgl. Bortz/Döring (2002: 251):

- Das Bemühen, dem Interviewer/der Interviewerin gefallen zu wollen
- Wissen um die Teilnahme an einer Befragung hat Auswirkung auf das Ergebnis
- Geringe Bereitschaft zur Selbstenthüllung
- Spezifische Motive zur Selbstdarstellung und Streben nach Konsistenz
- Antizipation möglicher negativer Konsequenzen nach bestimmten Antworten
- Konkrete Vermutungen über den Auftraggeber bzw. dessen Untersuchungsziele.

Die Wichtigkeit einer sorgfältigen Auswahl und Schulung der Interviewer/-innen wird anhand dieser Beispiele deutlich.

Soziale Erwünschtheit wird von Bortz/Döring als Sonderform der Selbstdarstellung beschrieben. Durch die Furcht vor sozialer Verurteilung neigen Personen zu konformem Verhalten und orientieren sich an verbreiteten Normen und Erwartungen.

Aus Evaluationsstudien ist bekannt, dass beispielsweise Patientinnen und Patienten selten kritische Urteile über Ärzte und Ärztinnen abgeben, was aus einer Form der Abhängigkeit resultieren kann. Ähnlich verhält es sich bei Befragungen von Firmenpersonal. Die Gewährung der Anonymität ist besonders bei kleinen Stichproben kaum möglich, sodass vieles nicht mitgeteilt bzw. beschönigt dargestellt wird. Eine positive Darstellung der eigenen Person ist hingegen in Bewerbungssituationen nachvollziehbar. „Objektive“ Ergebnisse lassen sich somit anhand von Fragebögen oder Interviews nur bedingt erheben, eine – zumindest in einigen Fällen – geeignetere Methode ist daher die Beobachtung.

Verwendung standardisierte (Test-)Verfahren

Auch wenn sie bei der Evaluation von Ausbildungsordnungen in der Regel nicht zur Anwendung kommen, sollten aus Gründen der Vollständigkeit die sog. standardisierten (Test-)Verfahren erwähnt werden. Sie spielen eine besondere Rolle, wenn Persönlichkeitsmerkmale erhoben werden sollen, wie zum Beispiel Intelligenz, Einstellungen, Depressivität oder Konzentrationsfähigkeit. Der Vorteil des Einsatzes solcher (oft käuflicher) Instrumente liegt in der gesicherten Validität und Reliabilität der Instrumente sowie in der Vergleichbarkeit der Ergebnisse mit der sogenannten „Normalbevölkerung“ auf Basis von Normwerten. So gut wie alle Instrumente weisen metrisches Skalenniveau auf, wodurch eine Vielzahl von Berechnungsmöglichkeiten besteht.

Beobachtung

Unter Berufung auf Laatz (1993) beschreiben Bortz/Döring (2002: 262) „beobachten“ als Sammeln von Erfahrungen in einem nicht kommunikativen Prozess, der stärker zielgerichtet und methodisch kontrolliert ist als Alltagsbeobachtungen. Bezeichnend dafür ist die Verwendung von Instrumenten, die die Selbstreflektiertheit, Systematik und Kontrolliertheit während der Beobachtung gewährleisten. Im Gegensatz zur Beobachtung, die bei einer qualitativen Herangehensweise offen angelegt werden sollte (vgl. Arbeitshilfe: Erhebung qualitativer Daten), empfehlen die Autoren vor dem Hintergrund einer quantitativen Herangehensweise einen genauen Beobachtungsplan, der vorschreibt, was für die Beobachtung wesentlich und unwesentlich ist, wie das Beobachtete gedeutet werden darf, wann sie stattfindet und wie sie protokolliert wird.

Als Formen der Beobachtung werden die folgenden Formen genannt:

- teilnehmend – verdeckt: z. B. Evaluierende besuchen unangekündigt und für die Auszubildenden nicht klar zu erkennen eine Lehrwerkstatt,
- teilnehmend – offen: z. B. Evaluierende beobachten Auszubildende im Rahmen einer angekündigten Betriebsbesichtigung,
- nicht teilnehmend – verdeckt: z. B. Evaluierende beobachten für die Auszubildenden unsichtbar (z. B. hinter einer Scheibe) den Verlauf der praktischen Prüfung.

Bei offenen Beobachtungen muss damit gerechnet werden, dass die zu Beobachtenden über den Zweck und das Ziel der Beobachtung Bescheid wissen und sich so sozial erwünscht verhalten (z. B. Tragen von Kopfbedeckungen in Küchen bei einer angekündigten Hygienekontrolle). Soll die Beeinflussung der Probanden durch die Beobachtenden vermieden werden, sind nonreaktive Beobachtungen anzustreben, wobei die Einhaltung ethischer Grundsätze gewahrt werden sollte. Ähnlich wie beim Interview sollten Beobachterinnen und Beobachter einem Training unterzogen werden. Um die Reliabilität der Beobachtungen zu überprüfen, können mehrere Beobachterinnen oder Beobachter eingesetzt und ihre Ergebnisse auf Übereinstimmung geprüft werden.

Zählung

Eine quantitative Erfassung vorab bestimmter Merkmale erfolgt, zum Beispiel entlang einer Checkliste.

Begehung

Die quantitative Erfassung bestimmter Merkmale eines sozialen Raums (z. B. einer Lehrwerkstätte) wird durchgeführt.

Erhöhung der Ausschöpfungsquote

Um den Erfolg einer schriftlichen Befragung zu erhöhen, müssen verschiedene Faktoren berücksichtigt werden.

► Durchführung eines Pretests

Der Pretest dient der Verbesserung von Untersuchungsinstrumenten *vor* Durchführung der eigentlichen Erhebung. Er wird erforderlich, um Fragenfehler und nicht trennscharfe Items zu reduzieren (Item-Analyse) und das gesamte Instrument zu präzisieren. Ein gut getesteter Fragebogen kann ein Garant dafür sein, dass weniger Probanden/Probandinnen den Test zwischendrin abbrechen. Zudem bietet speziell ein Online-Fragebogen dem/der Probanden/Probandin in der Regel keine Möglichkeiten ‚mal eben‘ schriftliche Randbemerkungen an der passenden Stelle abzugeben. Daher muss der Fragebogen präzise sein.

Bei Online-Befragungen ist ein Pretest zudem aufgrund technischer Aspekte unbedingt durchzuführen. Wichtige Checks, die durchgeführt werden sollten, sind dabei[37]:

- Korrekte visuelle Darstellung des Fragebogens
- Einwandfreies Funktionieren der eingesetzten Antwortformate
- Die Filterführung
- Die Übermittlung der Daten
- Die Probandenverwaltung
- Der Export der Antwortdaten
- Der Import der Daten in das Analyseprogramm, z. B. Excel oder SPSS.

▸ Ansprechende Gestaltung des Anschreibens an die Befragten

Die Befragten sollen durch eine ansprechende *Gestaltung* und den *Inhalt* des Anschreibens zur Teilnahme an der Befragung überzeugt werden[38]. Bei Online-Befragungen erfolgt das Anschreiben über die E-Mail, mit der auch das Passwort zur Befragung sowie der direkte Link zum Fragebogen verschickt wird.

Es ist die Aufgabe der Befragenden, den Befragten im Anschreiben zu demonstrieren, dass sich die Teilnahme für sie oder Mitglieder ihrer peer-group entweder lohnt oder ihnen zumindest nicht schadet. Es ist grundsätzlich auch zu überlegen, *wer* angeschrieben wird. Beispielsweise sollten Ausbilderinnen und Ausbilder, Prüferinnen und Prüfer oder Berufsschullehrerinnen und Berufsschullehrer ein anderes Anschreiben erhalten als die Auszubildenden selbst.

Auszubildende können eher dazu motiviert werden an einer Befragung teilzunehmen, wenn das Anschreiben

a. ihren Lesegewohnheiten entspricht und
b. ihnen in einfachen Worten vermittelt wird, welche Vorteile die Untersuchungsergebnisse mit sich bringen können.

Folgende Aspekte sollten im Anschreiben berücksichtigt werden:

- Ein Bild kann als „eye catcher“ dienen, weil es die Aufmerksamkeit weckt.
- Klare Teilnahme-Appelle, wie z. B. „Mitmachen + mit Fakten überzeugen!“ motivieren zum Mitmachen.
- Auf persönliche relevante Motive der Teilnehmerinnen und Teilnehmer soll Bezug genommen werden.
- Hinweise auf mögliche Einfluss- und Gestaltungsmöglichkeiten sollen gegeben werden.
- Logos beteiligter Akteure (BIBB, Ministerien, ggf. Sozialpartner) sollen gezielt ein als Verweis auf „anerkannte Autoritäten“ eingesetzt werden.
- Mittels eines Logos soll deutlich gemacht werden, wer die Untersuchung durchführt.
- Wie viel Zeit die Beantwortung aller Fragen voraussichtlich in Anspruch nehmen wird, muss unbedingt angegeben werden.
- Es muss klar erkennbar sein, wo der ausgefüllte Fragebogen abzugeben ist bzw. an welche Adresse dieser geschickt werden soll.
- Anonymität muss explizit zugesichert werden, darüber hinaus muss plausibel dargestellt werden, was mit den Daten geschieht. Zudem ist klarzustellen, dass Befragte ihren Namen nicht angeben müssen und den Fragebogen ohne Absender zurücksenden sollen.

[37] Vgl. Kuckartz (2009: 47–50).
[38] In manchen Fällen ist es sinnvoll – beispielsweise bei der Befragung von Auszubildenden – alle Informationen auf der ersten Seite des Fragebogens zu platzieren und ganz auf ein zusätzliches Anschreiben zu verzichten.

► Ein Fragebogen, der die Befragten zeitlich und inhaltlich nicht überfordert

Es ist dringend darauf zu achten, dass der Fragebogen nicht die Geduld der Befragten überbeansprucht und die Befragten auch nicht inhaltlich überfordert.

► Nachfassaktionen: Versand eines ersten Erinnerungsschreibens/einer Mail ca. zwei Wochen nach Beginn der Befragung

Mindestens eine Nachfassaktion ist notwendig, da damit der Fragebogenrücklauf um etwa 10 Prozent erhöht werden kann. Auch der Versand eines zweiten Erinnerungsschreibens oder einer Mail, etwa drei Wochen nach Beginn der Befragung, kann nochmals zu einer Steigerung des Rücklaufs von ungefähr 10 Prozent beitragen. Jede weitere Nachfassaktion bringt dann nur noch ca. 3–5 Prozent.

Die Nutzung von Fach- oder Verbandszeitschriften oder von Newsletter sollte ebenfalls in Betracht gezogen, um auf die Befragung aufmerksam zu machen.

Der Umgang mit Missings[39]

Durch gezielte Nachfassaktionen kann der Rücklauf, wie eben beschrieben, zum Teil erhöht werden. Dennoch werden in der Regel immer einige Probanden nicht antworten, sprich den Fragebogen nicht online ausfüllen, nicht per Post zurückschicken oder kein Interesse an einer telefonischen Befragung haben. Sie gelten als sog. **Unit Nonresponse**.

Darüber hinaus gibt es sog. **Item Nonresponse** innerhalb eines ausgefüllten Fragebogens. Hierbei handelt es sich um Fragen, die von dem Probanden bspw.

- bewusst ausgelassen wurden,
- übersehen wurden,
- nicht beantwortbar waren,
- uneindeutig waren.

Diese fehlenden Daten können unter Umständen den Datensatz entscheidend verzerren. Es gibt mehrere Möglichkeiten in SPSS mit diesen Item Nonresponse umzugehen:

Fallausschluss Test für Test (nicht zwingend zu empfehlen)

Probanden mit fehlenden Daten werden von den Berechnungen in SPSS einfach ausgeschlossen. Aus dieser Vorgehensweise erwachsen jedoch neue Probleme: die bei SPSS als Standardeinstellung aktivierte Option sorgt dafür, dass in verschiedenen Berechnungen stets unterschiedliche Probanden des Datensatzes eingebunden werden.

Listenweiser Fallausschluss (nur bei Datensätzen mit geringer Fehlquote < 5 % zu empfehlen)

Hierbei wird der Datensatz radikal auf alle Fälle, die vollständig vorliegen, reduziert. Beide Fallausschlüsse verringern die Größe des Datensatzes, und es werden wertvolle Informationen einfach verschenkt.

Fälle gewichten (zu empfehlen)

Mit der Gewichtung eines Datensatzes kann erreicht werden, dass das Profil des Datensatzes einem gewünschten Profil wie beispielsweise der zugrunde liegenden Grundgesamtheit angenähert wird.

[39] Die nachfolgenden Informationen sind vorwiegend den Schulungsunterlagen der Firma NETQUES entnommen, die im Juni 2009 im BIBB eine Fortbildung zum Thema Stichprobenanalyse angeboten hatte.

Fehlender Wert durch Mittelwert ersetzen (zu empfehlen)

Aus den bestehenden Daten einer Variablen werden fehlende Werte durch Mittelwert- oder Medianbildung ersetzt. (Da Mittelwerte sehr stark durch Ausreißer beeinflusst werden, bietet es sich an, den Median aufgrund der Robustheit zu verwenden.)

Regression

Hier werden die fehlenden Werte im Datensatz mit den durch die Regressionsgerade vorhergesagten Werten ersetzt. Bei der Mittelwertersetzung (siehe oben) werden die Zusammenhänge unterschätzt, bei der Regression überschätzt.

Multiple Imputationen (fundierte Kenntnisse des Verfahrens erforderlich, äußerst zeitaufwendig)

Hierbei wird jeder fehlende Wert durch mehrere zufällig oder geschätzte Werte ergänzt.

► Abschließende Anmerkung

Im Falle von Evaluationen ist die Ersetzung einzelner Werte nicht nur methodisch, sondern auch *inhaltlich* zu prüfen. Sind viele missing values bei einem Item zu verzeichnen, ist die Sinnhaftigkeit bzw. Interpretierbarkeit nach vielfacher Imputation fraglich. Die Gründe für das Auslassen des Items liefern möglicherweise mehr Information als „künstlich" produzierte Antworten. Bei Ersetzung der missing values durch Mittelwerte können einfache Häufigkeitsdiagramme nicht mehr erstellt werden.

Weist eine Person sehr viele fehlende Werte auf, ist möglicherweise der Ausschluss dieser Person sinnvoller.

Literatur

Bortz, Jürgen: Statistik für Human- und Sozialwissenschaftler. Heidelberg 2005.

Bühner, Markus: Einführung in die Test- und Fragebogenkonstruktion. München 2004.

Glinz, Martin: Delphi-Methode. Augsburg 2005. URL: http://glossar.hs-augsburg.de/Delphi-Methode (Stand 07.01.2013).

Kromrey, Helmut: Empirische Sozialforschung. Stuttgart 2006.

Kuckartz, Udo u. a.: Evaluation online. Internetgestützte Befragung in der Praxis. Wiesbaden 2009.

Porst, Rolf: Fragebogen. Ein Arbeitsbuch. Wiesbaden 2009.

Häder, Michael; Rexroth, Margrit: Erfassung kognitiver Aspekte des Antwortverhaltens in einer Delphi-Studie. ZUMA-Arbeitsbericht 98/06. Mannheim 1998.

Häder, Michael; Häder, Sabine: Die Grundlagen der Delphi-Methode: ein Literaturbericht. Mannheim 1994. URL (PID): http://nbn-resolving.de/urn:nbn:de:0168-ssoar-70116.

Fragebogenkonstruktion

Deml, Sonja: Wie erstelle ich einen Fragebogen für eine empirische Forschungsarbeit? Norderstedt 2001.

Kallus, Wolfgang: Erstellung von Fragebogen. Wien 2010.

Kirchhoff, Sabine u. a.: Der Fragebogen. Datenbasis, Konstruktion und Auswertung. Wiesbaden 2008.

Konrad, Klaus: Mündliche und schriftliche Befragung. Bamberg 1999.

Porst, Rolf: Fragebogen: Ein Arbeitsbuch. Wiesbaden 2009.

Raab-Steiner, Elisabeth; Benesch, Michael: Der Fragebogen. Von der Forschungsidee bis zur SPSS/PASW Auswertung. Stuttgart 2010.

Online-Befragung

Batinic, Bernad; Bosnjak, Michael: Fragebogenuntersuchung im Internet. Göttingen 2000.
Kuckartz, Udo u. a.: Evaluation online : internetgestützte Befragung in der Praxis. Wiesbaden 2009.

CATI-Befragung

Balli, Christel; Krekel, Elisabeth M.; Sauter, Edgar: Qualitätsentwicklung in der Weiterbildung. Zum Stand der Anwendung von Qualitätssicherungs- und Qualitätsmanagementsystemen bei Weiterbildungsanbietern. Bonn 2002.
Berger, Klaus: Evaluierung der Bund-Länder-Ausbildungsplatzprogramme Ost – Erwerbssituation der Programmabsolventinnen und -absolventen ein halbes Jahr nach Ausbildungsabschluss. Ergebnisse einer computergestützten Telefonbefragung von Absolventinnen und Absolventen des Jahres 2005 im Ausbildungsplatzprogramm Ost und einer betrieblichen Vergleichsgruppe. Bonn 2006.
Buchwald, Christina: Studie zur Bildungslandschaft in Aschersleben. Eine Untersuchung zur Integration einer weiterführenden Schule in freier Trägerschaft in die Bildungslandschaft der Stadt Aschersleben. Forschungsberichte aus dem zsh 06–2. Halle 2006.
Buchwald, Christina: Telefoninterview ist nicht gleich Telefoninterview. Berlin 2005.
Buchwald, Christina: Telefonische Bevölkerungsbefragungen im Vergleich zu telefonischen Interviews mit politischen Eliten. Ausschöpfung und Kontakthäufigkeit. Baden-Baden 2008.
Buchwald, Christina: Zur Leistungsfähigkeit telefonischer Befragungen. Das Methodenprojekt des SFB 580 – zwischen Methodenentwicklung und Dienstleistung. Heft 4. Halle 2002.
Buchwald, Christina; Lukanow, Katja: Das Telefoninterview – Instrument der Zukunft. Halle 2006.
Buchwald, Christina; Lukanow, Katja: Qualitätsmanagement und Qualitätssicherung. 7. Wissenschaftliche Tagung des Statistischen Bundesamtes. GESIS-Tagungsberichte. Bonn 2007.
Falk, Martin: Erfolg von personalwirtschaftlichen Maßnahmen zur Überwindung des IT-Fachkräftemangels. Mering 2003.
Falk, Martin; Bellmann, Lutz; Velling, Johannes: IKT-Fachkräftemangel und Anpassungsreaktionen der Unternehmen. Nürnberg 2002.
Häder, Michael; Häder, Sabine: Telefonbefragung über das Mobilfunknetz: Konzept, Design und Umsetzung einer Strategie zur Datenerhebung. Wiesbaden 2009.
Hall, Anja: Die BIBB/BAuA-Erwerbstätigenbefragung 2006: Methodik und Frageprogramm im Vergleich zur BIBB/IAB-Erhebung 1998. Bonn 2009.
Martens, Bernd; Ritter, Thomas: Eliten am Telefon: Neue Formen von Experteninterviews in der Praxis. Baden-Baden 2008.
Wiener, Bettina; Lukanow, Katja: Medienstudie Halle 2005. Eine Untersuchung zur IT-, Medien- und Kommunikationsbranche der Stadt Halle. Halle 2005.
Zopf, Susanne; Tiemann, Michael: BIBB/BAuA-Employment Survey 2005/06. Bonn 2010.

Arbeitshilfe: Skalenniveaus und Auswertung quantitativer Daten

Beim Einsatz quantitativer Methoden ist die Wahl des Skalenniveaus mit entscheidend dafür, welche Auswertungsverfahren überhaupt angewendet werden dürfen. Die Arbeitshilfe möchte daher zunächst kurz auf die möglichen Skalenniveaus eingehen und darauf aufbauend Aspekte und Verfahren beleuchten, die es bei der Auswertung und Bewertung von quantitativ erhobenen Daten zu berücksichtigen gilt.

Inhalt

Wozu dient Statistik? . . . 74
Messung und Skalen . . . 74
Nominalskala . . . 74
Ordinalskala (Rangskala) . . . 75
Intervallskala . . . 75
Verhältnisskala (Ratio(nal)skala) . . . 77
Berechnungsmöglichkeiten bei den einzelnen Skalenniveaus . . . 78
Die Normalverteilung und ihre Bedeutung für die Statistik . . . 79
Deskriptive Statistik . . . 81
Ausgewählte Lageparameter . . . 81
Ausgewählte Streuungsmaße . . . 82
Bivariate Analyse . . . 86
Inferenzstatistik . . . 87
Übersicht der inferenzstatistischen Verfahren . . . 87
Darstellung einzelner Signifikanztests . . . 89
Testverfahren zur Ermittlung von signifikanten Unterschieden (für Variable mit metrischem Skalenniveau) . . . 96
Nichtparametrische Tests (verteilungsfreie Tests) für Unterschiede . . . 100
Zusammenfassung . . . 103
Literatur . . . 104

Wozu dient Statistik?

- Zur **Reduktion:** Die Informationsflut der erhobenen Daten soll kanalisiert werden mit dem Ziel, Übersichtlichkeit herzustellen, beispielsweise mittels Häufigkeitsauszählungen.
- Zur **Exploration:** Es sollen Strukturen und Zusammenhänge der erhobenen Daten erkundet werden, beispielsweise mittels Kreuztabellen, Faktorenanalysen, Datamining, etc.
- Zur **Extrapolation:** Die Erkenntnisse aus den Daten einer Stichprobe sollen auf nicht untersuchte Einheiten mittels inferenzstatistischer Methoden übertragen werden, wozu beispielsweise die lineare Regression zählt.

Messung und Skalen

Daten sind das Ergebnis von Messvorgängen. „Messen" bedeutet im Rahmen der quantitativen Forschung, dass Eigenschaften von Objekten nach bestimmten Regeln in Zahlen ausgedrückt werden. Im Wesentlichen bestimmt die jeweilige Art einer Eigenschaft, wie gut ihre Ausprägung gemessen werden kann, d. h. wie gut sie sich in Zahlen ausdrücken lässt.

Eine *Eigenschaft* eines Objektes – in unserem Falle beispielsweise eines Ausbildungsbetriebs – kann u. a. *die Anzahl der aktuellen Ausbildungsverhältnisse* sein. Diese Eigenschaft lässt sich sehr gut in einer Zahl ausdrücken. Die *Qualität der Ausbildung* dieses Betriebs, als eine weitere Eigenschaft des Objektes, ist hingegen viel schwieriger direkt in Zahlen auszudrücken. Zur Operationalisierung von Qualität, sprich der Benennung geeigneter Indikatoren zu Messung von Qualität, bedarf es mehrerer Arbeitsschritte. Letztlich können aber auch hierfür Indikatoren gefunden werden, denen eine Zahl zugeschrieben werden kann.

Die „Messlatte", auf der die Ausprägungen einer Eigenschaft aufgetragen werden, heißt Skala. Je nachdem, in welcher Art und Weise eine Eigenschaft eines Objektes in Zahlen ausgedrückt/gemessen werden kann, lassen sich Skalen mit unterschiedlichen Skalenniveaus unterscheiden. Im Folgenden werden die vier möglichen Skalenniveaus kurz vorgestellt.

Nominalskala

Die Nominalskala stellt die einfachste Form des Messens dar. Hier bedeuten unterschiedliche Zahlen nichts anderes als unterschiedliche Merkmalsausprägungen. Diese Zahlen stehen jedoch **nicht** für ein „mehr oder weniger" oder „größer oder kleiner", sie haben **keine mathematische Bedeutung**. Vielmehr kann lediglich mittels einer Häufigkeitsauszählung ermittelt werden, wie häufig die einzelnen Merkmalsausprägungen genannt wurden.

Beispiele für Nominalskalen:

Objekt: Geschlecht	
mögliche Merkmalsausprägungen	„Skala"
männlich	*1*
weiblich	*2*

Objekt: Bundesland	
mögliche Merkmalsausprägungen	„Skala"
Baden-Württemberg	*1*
Bayern	*2*
...	*X*
Sachsen-Anhalt	*15*
Thüringen	*16*

Die Zuordnung von Zahlen zu Ausprägungen (z. B. 1 = männlich; 2 = weiblich) erfolgt willkürlich. Nominalskalierte Variablen können nach Anzahl ihrer möglichen Ausprägungen in dichotome (2 Ausprägungen wie z. B. Geschlecht m/w, oder ja/nein) oder polytome Variablen (z. B. Familienstand, Religionsbekenntnis, Postleitzahl, Bundesland) eingeteilt werden.

Ordinalskala (Rangskala)

Die Ordinalskala erlaubt die Aufstellung einer Rangfolge mithilfe von Rangwerten, d. h. ordinalen Zahlen. Zur reinen Häufigkeitsauszählung, wie bei der Nominalskala, kommt folglich ein **ordnender Vergleich** hinzu, es sind Aussagen über „besser – schlechter"-Relationen möglich. Als klassisches Beispiel für rangskalierte Daten können Schulnoten genannt werden. Die Merkmalsausprägungen treten in vergleichbaren, diskreten Kategorien auf und lassen sich z. B. nach Qualität, Stärke oder Intensität anordnen. Die Zuordnung von Zahlen zu Ausprägungen erfolgt nach einer Rangfolge. Die Rangwerte selbst sagen jedoch nichts über die Abstände oder über die Beziehungen zwischen den Merkmalsausprägungen aus. Es sind folglich keine Aussagen hinsichtlich der Fragen „wie viel mehr?" oder „wie viel besser?" möglich (vgl. Infokasten 1).

Infokasten 1

„Aus der olympischen Medaillenvergabe weiß man zwar, dass Gold besser als Silber und Silber besser als Bronze ist. Dabei kann aber durchaus (zum Beispiel 100 m Lauf) der Unterschied Gold (10,1 sec) zu Silber (10,2 sec) klein sein, während zwischen Silber und Bronze (10,7 sec) ein großer Abstand klafft (0,5 sec). Diese Information geht verloren, nur Rangplätze zählen." KÖHLER/SCHACHTEL/VOLESKE (2007).

Beispiel für ordinalskalierte Variablen:

Objekt: Einschätzung des Verlaufs der Prüfung „Die Abschlussprüfung 2010 verlief für unsere Auszubildenden …"		Anderes Beispiel: Wie sehr stimmen Sie der folgenden Aussage zu:	
mögliche Merkmalsausprägungen	**Skalierung**	**Nach meiner Ausbildungszeit stehen die Chancen auf einen Arbeitsplatz sehr gut.**	
sehr gut	1	stimmt genau	1
gut	2	stimmt eher	2
weniger gut	3	stimmt eher nicht	3
schlecht	4	stimmt gar nicht	4

Weitere Beispiele für ordinalskalierte Variablen: Alter in Kategorien, höchste abgeschlossene Schul-/Ausbildung, Ratings, einzelne Items in Fragebogen wie z. B. die visuelle Analogskala. Bei dieser Form der Einschätzung werden nur die Endpunkte einer Skala verbal benannt und die Distanz dazwischen wird mit einer Linie dargestellt:

sehr zufrieden _____ gar nicht zufrieden

Intervallskala

Bei einer Intervallskala lassen sich zusätzlich zu den Eigenschaften der Ordinalskala die Abstände zwischen den verschiedenen Merkmalsausprägungen exakt bestimmen. Allerdings existiert **kein natürlicher Nullpunkt** für die Skala. Der Messwert 0 wird also willkürlich festgelegt und besagt nicht, dass ein Merkmal nicht vorhanden ist. Ein Beispiel für eine Intervallskala ist die Celsius-Temperaturskala: Der Temperaturunterschied zwischen 5 °C und 10 °C ist genauso

groß wie derjenige zwischen 20 °C und 25 °C. Allerdings bedeutet 0 °C nicht, dass keine Temperatur vorhanden ist und der Messwert 0 könnte auch irgendeiner anderen Temperatur zugeordnet werden. Aufgrund dieser Beliebigkeit des Nullpunktes ist es falsch zu behaupten, 20 °C seien doppelt so warm wie 10 °C.

Beispiel für eine Intervallskala:

Objekt	Betriebsergebnis in EURO	Mögliche Frage dazu: „Welches Betriebsergebnis erzielte Ihr Betrieb im Jahr 2010?“
Betrieb C	5.200.000,–	
Betrieb F	2.500.000,–	
Betrieb B	276.000,–	
Betrieb A	–276.000,–	
Betrieb D	–948.000,–	

Bei intervallskalierten Daten ist es möglich, die Abstände zwischen den Ausprägungen zu vergleichen. Die Intervallskala ist deshalb auch nicht mehr diskret sondern kontinuierlich. Die **Berechnung von Verhältnissen** (halb so viel, doppelt so viel) ist aber auf diesem Niveau noch **nicht erlaubt.**

Weitere Beispiele für intervallskalierte Variablen: Datum nach christlicher Zeitrechnung, Normwerte psychologischer Testverfahren (IQ, Z, z, C, T).

Likert-Skalierung als „Sonderfall“: Der nachfolgend dargestellten Skala, die häufig in standardisierten Fragebogen verwendet wird, wird teilweise Intervallskalenniveau zugesprochen, auch wenn sie lediglich das Niveau einer Ordinalskala aufweist.

Objekt: Bewertung der GAP		Mögliche Frage dazu: „Wie bewerten sie die Neuregelung der Prüfung mittels der GAP?“
mögliche Merkmalsausprägungen	**Skalierung**	
Lehne ich stark ab	1	
Finde ich nicht sehr gelungen	2	
Ich sehe Vor- und Nachteile	3	
Befürworte ich größtenteils	4	
Befürworte ich vorbehaltlos	5	

Weshalb dieser sog. Rating-Skala Intervallskalenniveau zugesprochen wird, hat hauptsächlich pragmatische Gründe. Zum einen stehen für die Auswertung von intervallskalierten Daten mehr und aussagekräftigere statistische Verfahren zur Verfügung. Zum anderen gelangt man in der Forschung mit Rating-Skalen oftmals auch dann zu sinnvollen Ergebnissen, die sich in der Praxis bewähren, wenn man das (höhere) Intervallskalenniveau unterstellt, vgl. Sedlmeier/Renkewitz (2008: 66).

Metrisches Niveau darf bei diesen Merkmalsausprägungen insbesondere dann angenommen werden, wenn mehrere Items eines standardisierten Verfahrens zu einer „Skala“ zusammenge-

fasst wurden, deren Reliabilität und Validität nachgewiesen wurde. Bespiele dafür gibt es aus der Psychologie bei der Messung von Merkmalen wie Intelligenz, Persönlichkeitsmerkmalen, Leistungsmerkmalen und klinischen Merkmalen wie z. B. Depression oder Ängstlichkeit. Nach einer Testung von vielen Items an einer repräsentativen Eichstichprobe wird anhand der Trennschärfen eine Auswahl der geeigneten Items getroffen oder es werden im Rahmen einer Faktorenanalyse latente Dimensionen aus dem Itempool extrahiert. Bortz/Döring (2002: 222) beschreiben das als einfachste Konstruktion von Skalen und gehen davon aus, dass die Kategorien der Rating-Skala äquidistant sind, wobei der mittlere Skalenwert nicht immer eindeutig zu interpretieren ist. Allerdings sind sie auch der Meinung, dass in der Praxis Likert-Skalen sehr häufig eingesetzt werden und beinahe jede Ansammlung von 5-stufigen Items als Likert-Skala bezeichnet wird, ohne dass der Nachweis für die Angemessenheit dieser Bezeichnung durch eine Item-Analyse geliefert wird, vgl. Bortz/Döring, (2002).

Im Rahmen von Evaluationen sind Likert-Skalen, die im Rahmen einer Pilot-Testung entwickelt werden, eine durchaus praktikable Erhebungsmethode. Bei entsprechender Reliabilität können durchaus Mittelwerte errechnet und interpretiert und somit auf einfache Weise Verläufe oder Gruppenunterschiede beschrieben und auf Signifikanz getestet werden. Denkbar wäre der Einsatz z. B. bei der Messung der Zufriedenheit.

Verhältnisskala (Ratio(nal)skala)

Bei einer Verhältnisskala existiert ein sinnvoll definierbarer **echter Nullpunkt**. Die Verhältnisse zwischen den Messwerten, welche die Merkmalsausprägungen abbilden, dürfen berechnet werden. Nicht nur die Differenz, sondern auch der Quotient aus zwei Messwerten darf verwendet werden. Nur bei diesem Skalenniveau sind also **alle mathematischen Rechenoperationen** sinnvoll und **erlaubt**.

Beispiele für Verhältnisskalen:

Objekt: Weiterbildungskosten	Skala: Ausgaben in EURO	Mögliche Frage dazu: Wie hoch waren die Kosten in Ihrem Betrieb für Weiterbildungen im Jahr 2010 insgesamt?
Betrieb C	172.370	
Betrieb D	50.850	
Betrieb E	27.500	
Betrieb B	14.962	
Betrieb A	5.600	
Betrieb X	...	

Das Merkmal Weiterbildungskosten mit Ausprägungen in Euro genügt einer Verhältnisskala. Der Nullpunkt ist nicht willkürlich definierbar, der natürliche Nullpunkt liegt bei 0,– EUR, da es keine „Negativausgaben" gibt. Daher sagt auch der Quotient etwas über das Verhältnis zweier Messwerte aus: 60.000,– EUR Weiterbildungskosten ist zweimal so viel wie 30.000,– EUR Weiterbildungskosten. Der Quotient ist 2.

Das Betriebsergebnis (in EUR gemessen) erfüllt dagegen nicht die Anforderungen einer Verhältnisskala, es existiert kein natürlicher Nullpunkt. 0,– EUR Betriebsergebnis ist ein willkürlich gesetzter Nullpunkt. Eine Aussage wie „das Betriebsergebnis von Betrieb B mit 276.000,– EUR ist doppelt so groß wie das von Betrieb A" (negatives Betriebsergebnis von minus 276.000,– EUR) ist nicht zulässig.

Weitere Beispiele für Merkmale, die mit der Verhältnisskala gemessen werden können, sind: Altersangabe in Jahren, Körpergröße, Gewicht, Einkommen, Beschäftigtenzahlen, Prüferstunden, Betriebszugehörigkeit in Jahren, Ausbildungsdauer etc.

Berechnungsmöglichkeiten bei den einzelnen Skalenniveaus

Tabelle 1
Möglichkeiten der einzelnen Skalenniveaus

Skalenniveau	Mögliche Aussagen	Erlaubte Operationen	Mögliche Lagemaße
Nominalskala	Gleichheit, Ungleichheit (Äquivalenz-Aussagen)	= ≠	Modus
Ordinalskala	siehe oben sowie: Größer-kleiner-Relationen (Ordnungs-Aussagen)	siehe oben sowie < >	siehe oben sowie: Median
Intervallskala	siehe oben sowie: Gleichheit von Differenzen (Distanz-Aussagen)	siehe oben sowie: - +	siehe oben sowie: Arithmetisches Mittel
Verhältnisskala	siehe oben sowie: Gleichheit von Verhältnissen (Verhältnis-Aussagen)	siehe oben sowie: × ÷	siehe oben sowie: Geometrisches Mittel

Die Skalenniveaus der einzelnen Fragen oder Messungen müssen, soweit wie möglich, bereits zu Beginn einer Untersuchung, spätestens aber bei der Erstellung eines Befragungs- oder Messinstruments, festgelegt werden. Die Festlegung des Skalenniveaus sollte gut überlegt werden, da bereits vor der eigentlichen Datenerhebung bestimmt wird, welche Auswertungsmethoden nach der Erhebung überhaupt erlaubt und möglich sein werden, vgl. Köhler/Schachtel/Voleske (2007):

Tabelle 2
Mögliche Auswertungsmethoden und Tests der einzelnen Skalenniveaus

Skalenniveau	Statistisch (erlaubte) Auswertungsmethoden	Maßzahlen und Tests
Nominalskala	Häufigkeiten	*Modalwert* *C, D, χ^2-Test*
Ordinalskala	siehe oben sowie: Rangplätze	*Interquartilsabstand* *$r_{Spearman}$, Z, U-und W-Test*
Intervallskala	siehe oben sowie: Messwerte	*Varianz s^2/ Standardabweichung s* *$r_{Pearson}$, $\bar{x}$, t-Test* *Regressionsanalyse* *Varianzanalyse*
Verhältnisskala	siehe oben sowie: Messwerte	*G, V*

Die Normalverteilung und ihre Bedeutung für die Statistik

Häufig ist aufgrund einer relativ großen Grundgesamtheit die Ziehung einer Stichprobe erforderlich (nähere Informationen hierzu finden Sie in der Arbeitshilfe Stichprobenauswahl). Ist die Stichprobe ordnungsgemäß gezogen worden, so nähern sich die ermittelten Werte mit zunehmender Stichprobengröße der sogenannten Normalverteilung an. Die Normalverteilung von Zahlenwerten ergibt sich dabei aus einer mathematischen Gesetzmäßigkeit, dem **zentralen Grenzwertsatz**[40].

Infokasten 2

Zentrales Grenzwerttheorem:

„Die Verteilung von Mittelwerten aus Stichproben des Umfanges n, die einer beliebig verteilten Grundgesamtheit entnommen werden, ist normal, vorausgesetzt, n ist genügend groß. Bortz/Döring (2002: 414).

Voraussetzung für die Gültigkeit dieses Satzes sind: Mittelwert und Varianz der Grundgesamtheit müssen endlich sein und n sollte ≥30 sein, vgl. Bortz/Döring (2002: 414).

Nach Leonhart (2004: 116) kann die Schätzung von Populationsparametern über zwei Vorgehensweisen präzisiert werden: Über die Erhöhung des Stichprobenumfangs oder über die Mittelung von vielen Stichprobenmittelwerten.

► Was ist die Bedeutung der Normalverteilung?

1. Die Normalverteilung wird verwendet, um die Gültigkeit von Aussagen einzuschätzen.
2. Die Normalverteilung wird für die Anwendung vieler statistischer Verfahren vorausgesetzt.
3. Viele andere Wahrscheinlichkeitsverteilungen können durch die Normalverteilung angenähert werden.

Die Normalverteilung (vgl. Abbildung 1) ist zwar ein Idealfall[41], hat aber eine Reihe interessanter und nützlicher Eigenschaften, die auch auf viele andere Verteilungen (meist unsymmetrische) übertragen werden dürfen. Viele Merkmale und Ereignisse wie z. B. die Zeit, die Lehrlinge zur Beantwortung der schriftlichen Fragen in der Zwischenprüfung eines Jahrgangs benötigen, sind zufallsverteilt (und besitzen eine begrenzte Varianz), sie lassen sich daher in einer statistischen Verteilung darstellen.

Es gilt, dass rund 68 Prozent der Messwerte innerhalb einer Standardabweichung um den Mittelwert (auf der Fläche unter der Kurve) liegen.

95,5 Prozent der Messwerte liegen innerhalb von zwei Standardabweichungen um den Mittelwert (auf der Fläche unter der Kurve).

99,7 Prozent der Messwerte liegen innerhalb von drei Standardabweichungen um den Mittelwert (auf der Fläche unter der Kurve).

Wenn im Rahmen einer Messung davon ausgegangen werden darf, dass ein zu messendes Merkmal in der Grundgesamtheit in etwa nor-

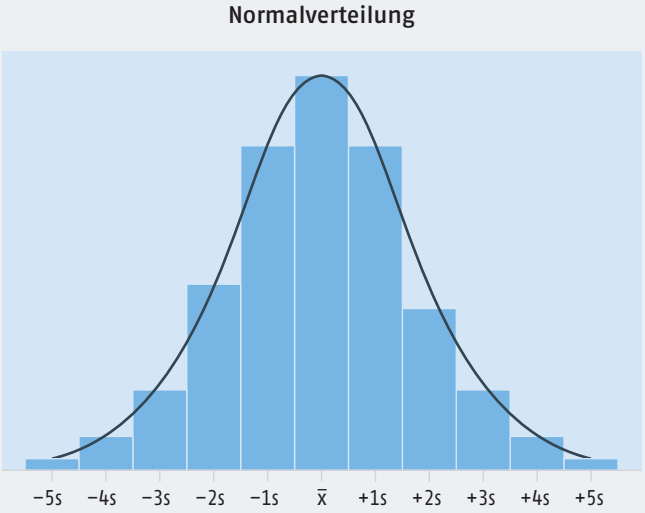
Infokasten 3

„Besonders oft werden wir die zweite Eigenschaft noch anwenden, dass die Fläche über dem Intervall $(\bar{x} - 2s;\ \bar{x} + 2s)$ etwa 95,5 % der Gesamtfläche unter der Kurve ausmacht. Da diese Fläche der Anzahl Beobachtungen entspricht, liegen also 95,5 % der beobachteten Werte im Bereich zwischen $\bar{x} - 2s$ und $\bar{x} + 2s$."

Köhler, Schachtel, Voleske (2007)

[40] Vgl. Pafnuti Lwowitsch Tschebyschow, 1846 und Alexander Michailowitsch Ljapunow, 1906.
[41] Sie ist eigentlich ein Sonderfall der Wahrscheinlichkeitsverteilung und kommt in ihrer „Idealform" tatsächlich nur selten vor.

Abbildung 1
Die Normalverteilung



malverteilt ist, dann darf auch eine entsprechende Häufigkeitsverteilung erwartet werden. Aus dieser Häufigkeitsverteilung darf dann die tatsächliche Verteilung des Merkmals geschätzt werden.

Die Normalverteilung, die auch als Gauß'sche-Verteilung[42] oder Glockenkurve bezeichnet wird, stellt eine **Wahrscheinlichkeitsverteilung** dar. Sie gibt an, mit welcher Wahrscheinlichkeit zufällig erhobene Einzelergebnisse[43] auftreten werden. (Sie steht im Gegensatz zur Häufigkeitsverteilung, bei der angegeben wird, wie häufig ein Wert aufgetreten ist.)

Bei einer Wahrscheinlichkeitsverteilung gilt, dass Ereignisse, deren Wahrscheinlichkeiten gering sind, selten auftreten und Ereignisse, deren Wahrscheinlichkeiten hoch sind, häufig auftreten. Die Häufigkeit des Auftretens von Ereignissen oder Merkmalen in der Stichprobe ist also davon – und nur davon – abhängig, wie häufig diese Ereignisse oder Merkmale in der Grundgesamtheit tatsächlich vorhanden bzw. verteilt sind.

► Eigenschaften der Normalverteilung

- Die Normalverteilung ergibt eine symmetrische (glockenförmige) Kurve.
- Das arithmetische Mittel, der Modalwert und der Median fallen in einem (dem höchsten) Wert der Verteilung zusammen, sind also identisch. Sie teilen die Verteilung in zwei gleich große Hälften.
- Die Fläche unter der Kurve entspricht 100 Prozent der Messwerte.
- Die Normalverteilung nähert sich der x-Achse, ohne sie jemals zu erreichen.
- Es existieren unendlich viele verschiedene Normalverteilungen, die jeweils durch zwei Größen eindeutig bestimmt werden können: durch den Mittelwert sowie durch die Standardabweichung.

[42] Johann Carl Friedrich Gauß, deutscher Mathematiker, Astronom, Geodät und Physiker.
[43] Zufallsergebnisse ergeben sich aus Zufallsstichproben. Alle Personen oder Ereignisse oder Merkmale der Grundgesamtheit haben die gleiche Wahrscheinlichkeit in die Zufallsstichprobe mit aufgenommen zu werden.

Deskriptive Statistik

„Unter deskriptiver Statistik versteht man eine Gruppe statistischer Methoden zur Beschreibung von Daten anhand statistischer Kennwerte, Graphiken, Diagramme und/oder Tabellen“ Leonhart (2004: 17).

Mittels der deskriptiven Statistik werden erhobene Daten bereinigt und geordnet, durch Lageparameter und Streuungsmaße beschrieben und schließlich auch grafisch dargestellt. Die deskriptive Statistik gibt also **Auskunft** darüber, **wie die einzelnen Daten** liegen bzw. **verteilt sind,** aber nicht, warum diese so liegen bzw. verteilt sind.

Zu klären ist, warum die Daten so verteilt sind und welche Schlüsse daraus gezogen werden dürfen. Nur wenige statistische Modelle (lineare Strukturgleichungsmodelle) können diesen Kausalitätsnachweis erbringen. Insbesondere bei Evaluationsergebnissen kann das „Warum“ für ein Ergebnis nicht statistisch ermittelt, sondern eher durch nachvollziehbare Interpretationen und Rückschlüsse aus den Daten formuliert werden.

Im Vergleich zur deskriptiven Statistik ermöglicht die Inferenz- (schließende) Statistik objektive Entscheidungen über die Brauchbarkeit von Hypothesen, vgl. Bortz (1999: 1). Auf Basis von Stichprobendaten können allgemeine, begründete Aussagen über die dahinterliegende Population bzw. Grundgesamtheit gemacht werden, indem Unterschiede oder Zusammenhänge auf Signifikanz getestet werden, vgl. Nachtigall/Wirtz (2006: 100).

Was grundsätzlich aus jedem Datensatz herausgelesen werden kann, sind die sogenannte Lageparameter und Streuungsmaße. Zur Beantwortung vieler Fragen, die sich aus Untersuchungen im Ordnungsbereich ergeben, reichte bisher häufig bereits die Angabe dieser Maßzahlen aus, um die gegebenen Forschungsfragen beantworten zu können.

Wichtig ist in diesem Zusammenhang, dass je nach Skalenniveau unterschiedliche Lageparameter und Streuungsmaße verwendet werden dürfen (vgl. Tabelle 1 auf S. 7).

Ausgewählte Lageparameter

► Modus bzw. Modalwert (D):

Das ist der *Wert, der am häufigsten als Messwert auftritt* und in der Regel das Dichtezentrum (unimodale Verteilung) bzw. bei mehreren verschiedenen Modalwerten die Dichtezentren (bimodale Verteilung) einer Datenreihe bildet. Er muss nicht berechnet werden; man kann ihn in einer Häufigkeitstabelle oder einer grafischen Darstellung einfach ablesen.

Beispiel für den Modalwert:

Im Rahmen einer bildungspolitischen Maßnahme sollten insbesondere „benachteiligte Bevölkerungsgruppen“ erreicht werden. Als Modalwert konnte die Gruppe „mit Migrationshintergrund“ ermittelt werden, diese Gruppe wurde am häufigsten durch die Maßnahme angesprochen.

Benachteiligte Bevölkerungsgruppe	N
Personen mit Migrationshintergrund	44
Personen mit Behinderungen	36
Sonstige benachteiligte Personen	41

► Median (Z):

Der Median (als Zentralwert) halbiert eine geordnete Datenreihe derart, dass 50 Prozent der Messwerte oberhalb und 50 Prozent der Messwerte unterhalb des Zentralwertes liegen. Zwischen einer geraden und ungeraden Reihe von Messwerten muss unterschieden werden.

Beispiel für die Darstellung des Median bei einer ungeraden Anzahl von Werten:

	Nach 1 Monat	Nach 2 Monaten	Nach 3 Monaten	Nach 4 Monaten
Anzahl der Ausbildungsabbrecher und Ausbildungsabbrecherinnen nach X Monaten	2	1	3	1

1. Schritt zur Berechnung des Medians: Werte sortieren

Monate	1	1	2	3	3	3	4
Position	1.	2.	3.	4.	5.	6.	7.

2. Schritt zur Berechnung des Medians:

Position des mittleren Wertes bestimmen:

$$\frac{n+1}{2} = \frac{7+1}{2} = 4$$

3. Schritt zur Berechnung des Medians: Mittleren Wert ablesen an der Position 4: 3 Monate. Ergebnis: Der Median liegt bei 3 Monaten.

4. Schritt Interpretationsbeispiel des errechneten Medians: 50 Prozent der Ausbildungsabbrecherinnen und Ausbildungsabbrecher haben ihre Ausbildung bis zum 3. Monat abgebrochen und die andere Hälfte ab dem 3. Monat.

► Arithmetisches Mittel (AM oder $\bar{x}$):

Die Summe aller Messwerte geteilt durch die Anzahl aller Messwerte, gibt den *Durchschnitt* ($\bar{x}$) an. Extrem hohe oder niedrige Einzelwerte verzerren das AM. Im Einzelfall ist zu überlegen, ob solche Extremwerte aus den Daten ausgeklammert werden können, um doch noch einen interpretierbaren Durchschnittswert zu erhalten.

Beispiel für die Darstellung des arithmetischen Mittels:

Durchschnittliche Weiterbildungskosten der befragten mittelständischen Betriebe in der Branche XY im Jahr 2009: 54.256,40 EUR.[44]

Ausgewählte Streuungsmaße

Mit den Streuungsmaßen lassen sich Informationen gewinnen, wie weit um die zentrale Tendenz herum die Werte streuen. Als klassisches Beispiel kann hier der Mittelwert genannt wer-

[44] Um im Evaluationsbericht eine korrekte Darstellung des AM zu gewährleisten, sollte immer neben dem Wert des AM auch die Stichprobengröße sowie der Wert der errechneten Standardabweichung benannt werden. Wie man die Standardabweichung berechnet, wird später im Text noch beschrieben.

den, der niemals ohne Angabe der Standardabweichung und Angabe der Größe der Stichprobe kommuniziert werden sollte. Je weiter sich die Werte vom Mittelwert entfernen, desto größer ist die Streuung und desto flacher die Verteilung.

Zu beachten ist, dass Streuungsmaße nicht für nominalskalierte Variablen berechnet werden können.

▸ Variationsbreite/Spannweite/engl. range (R):

Sie gibt die Streubreite der Werte vom niedrigsten bis zum höchsten Wert an. Die Variationsbreite (*R*) gibt also den Bereich an, in dem alle Messwerte liegen:

$$R = x\,max - x\,min$$

Durch die Spannweite kann deutlich werden, ob bspw. eine Gruppe A (z. B. Auszubildende im 2. Lehrjahr) in ihren Aussagen bezogen auf die Zufriedenheit mit der Ausbildung homogener ist als eine andere Gruppe B (Auszubildende im 3. Lehrjahr). Dann wäre das R von Gruppe A nämlich kleiner als das R von Gruppe B.

Die Spannweite ist jedoch sehr anfällig für Ausreißer.

▸ Varianz (s^2):

Sie ist ein Maß dafür, wie weit die einzelnen Messwerte im Durchschnitt vom arithmetischen Mittel entfernt liegen. Sie ist die Summe der quadrierten Abweichungen der einzelnen Messwerte vom Mittelwert. Leonhart (2004: 45) unterscheidet zur Berechnung der Varianz zwei Formeln: Eine für die Varianz der Population und nachstehende Formel zur Berechnung der Varianz einer Stichprobe:

$$s_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1}$$

▸ Standardabweichung/engl. standard deviation (s):

Die eben vorgestellte Varianz kommt jedoch hinsichtlich ihrer Aussagekraft erst richtig zum Tragen, wenn aus ihr die Wurzel gezogen wird, um die Standardabweichung zu berechnen. Sie ist das Maß für die Streuung der Werte um ihren Mittelwert:

$$s = \sqrt{s^2}$$

Der Unterschied zur Spannweite besteht darin, dass nicht nur die beiden Maximalwerte berücksichtigt werden (der Werte vom niedrigsten und vom höchsten Wert), sondern alle Werte. Es gilt: je höher die Varianz und Standardabweichung ausfallen, desto weiter streuen die Werte um den Mittelwert.

Ist die ermittelte Standardabweichung bei Gruppe A größer als bei Gruppe B, so bedeutet dies, dass die ermittelten Werte in der Gruppe A (bspw. durchschnittliche Zeit der Auszubildenden beim Kundeneinsatz vor Ort) heterogener sind als in der Gruppe B.

▸ Standardfehler/mittlerer Fehler des Mittelwertes ($\sigma_{\bar{x}}$):

Um den wahren Mittelwert (μ) einer Grundgesamtheit zu erfahren, müsste man alle Mitglieder (oder Bestandteile/Ereignisse/Merkmale) dieser Grundgesamtheit erfassen. Da dies in der Regel nicht möglich ist, wird eine Stichprobe aus der jeweiligen Grundgesamtheit gezogen.

Der Mittelwert, der sich aus der Untersuchung dieser Stichprobe ergibt ($\bar{x}$), weicht vom tatsächlichen Mittelwert der Grundgesamtheit (μ) ab. Der Mittelwert einer Stichprobe ist also ein Schätzwert für den wahren Mittelwert der Grundgesamtheit.

Die Standardabweichung einer Kennwerteverteilung (z. B. vieler Mittelwerte) aus einer Population nennt man „Standardfehler $(s_{\bar{x}})$ des Mittelwertes". Dabei gilt grundsätzlich: Je größer die Stichprobe, desto kleiner die Abweichung vom tatsächlichen Mittelwert. Leider gilt auch: Um den Standardfehler zu halbieren, muss der Stichprobenumfang vervierfacht werden.

Die Berechnung des Standardfehlers des Mittelwertes gibt an, wie groß die Streuung des Mittelwertes der Stichprobe ($\bar{x}$) um den wahren Mittelwert der Grundgesamtheit (μ) ist.

Bei der Ermittlung der Lageparameter und der Streuungsmaße einer Stichprobe sollte auch immer der Standardfehler des Mittelwertes erhoben werden:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

Hinweis: Die Formel ist zitiert nach Leonhart (2004: 117).

Es gilt: je größer der Standardfehler ist, desto unsicherer ist die Schätzung des Mittelwerts. Das heißt, die Genauigkeit der Schätzung hängt entscheidend von der Streuung der Variablenwerte ab; je kleiner die Streuung der Messwerte, desto genauer die Schätzung, je größer die Streuung desto unpräziser.

Beispiel zur Interpretation von Lage- und Streuungsmaßen:

Im Rahmen der Untersuchung zur gestreckten Abschlussprüfung in den Produktions- und Laborberufen der Chemischen Industrie wurden die Betriebe standardisiert gefragt:

„Wie hat sich durch die Einführung der Gestreckten Abschlussprüfung die bisherige Möglichkeit Ihres Betriebes, die Vermittlung von Ausbildungsinhalten zeitlich flexibel gestalten zu können, geändert?"

Ist deutlich verbessert worden (1)	Ist verbessert worden (2)	hat sich nicht verändert (3)	Ist verloren gegangen (4)	Ist deutlich verloren gegangen (5)

Aus den Antworten der Betriebe ließen sich die nachfolgenden Daten mittels SPSS erschließen: **Statistiken**

Hat sich die zeitliche Flexibilität geändert?

N	Gültig	186
	Fehlend	8
Mittelwert		3,35
Standardfehler des Mittelwertes		,061
Median		3,00
Modus		3
Standardabweichung		,827
Varianz		,684

Hat sich die zeitliche Flexibilität geändert?

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	deutlich verbessert worden	5	2,6	2,7	2,7
	verbessert worden	15	7,7	8,1	10,8
	nicht geändert	102	52,6	54,8	65,6
	verloren gegangen	42	21,6	22,6	88,2
	deutlich verloren gegangen	22	11,3	11,8	100,0
	Gesamt	186	95,9	100,0	
Fehlend	999	8	4,1		
Gesamt		194	100,0		

Wichtiger Hinweis vorab: Dieses Beispiel behandelt die ordinalskalierten Variablen so, als handele es sich um intervallskalierte Variablen. Dieses Verfahren ist zwar gängige Praxis, es ist jedoch immer mit Vorsicht zu behandeln. Auch muss spätestens im Evaluationsbericht ein Hinweis auf diese (im Grunde nicht zulässige) Vorgehensweise gegeben werden.

Welche Informationen lassen sich nun aus diesen Daten ablesen?

Der Mittelwert (hier bei fünfstufiger Antwortskala) beträgt 3,35. Die Antwort „hat sich nicht geändert" war mit dem Wert 3 codiert, dieser Wert wurde am häufigsten gewählt. Deshalb liegt der Modalwert auf der 3 und auch der Median liegt erwartungsgemäß bei 3.

Für die meisten befragten Betriebe (54,8 %) hat sich also offensichtlich keine Änderung bei der zeitlichen Flexibilität (bei der Vermittlung von Ausbildungsinhalten) ergeben.

Der wahre Mittelwert der Grundgesamtheit (alle in den Produktions- und Laborberufen ausbildenden Betriebe) liegt (mit einer 95 % Wahrscheinlichkeit) in einem Intervall von 3,23 bis 3,47, weil der Standardfehler des Mittelwertes bei dieser Stichprobe 0,061 beträgt.

$$\mu_{1,2} = \bar{x} \pm z * \sigma_{\bar{x}}$$

Hinweis: Formel nach Bortz (1999: 101)

34,4 Prozent der befragten Betriebe geben an, dass zeitliche Flexibilität bei der Vermittlung von Ausbildungsinhalten „verloren gegangen" (22,6 %) oder sogar „deutlich verloren gegangen" (11,8 %) ist. Für 65,6 Prozent der befragten Betriebe hat sich, im Hinblick auf zeitliche Flexibilität, nichts geändert oder es hat sich sogar etwas verbessert. Etwa ein Drittel aller Ausbildungsbetriebe scheint also Probleme mit der zeitlichen Flexibilität zu haben[45].

68 Prozent aller Wertungen liegen im Bereich von 2,52 bis 4,18.

95,5 Prozent aller Wertungen liegen im Bereich von 1,70 bis 5,0.

Hintergrundinformation: In der Praxis wird die Standardabweichung meist nur berichtet, aber selten für sich interpretiert. Ein Profi wird sich das AM und die Standardabweichung anschauen und seine Schlüsse ziehen. Wenn der Wertebereich so klein ist wie hier (1 bis 5), kann

[45] Warum dies so ist, kann natürlich mit diesen Daten nicht geklärt werden und bedarf weiterer, insbesondere qualitativer Untersuchungen.

die Standardabweichung im Übrigen auch gar nicht so groß werden. „Gefährlich“ sind Standardabweichungen eher bei Werten, wo es große Ausreißer geben kann.

► Variationskoeffizient/engl. coefficient of variation (*V*):

Er dient dem Vergleich mehrerer verschieden großer Mittelwerte hinsichtlich der Streubreite der Daten und drückt das Verhältnis der Standardabweichung zum Mittelwert in Prozent aus:

$$V = \frac{s_x}{x} * 100$$

Hinweis: Formel nach Leonhart (2004: 51).

Mittels des Variationskoeffizienten lassen sich verschiedene Verteilungen miteinander vergleichen, selbst wenn die Verteilungen nicht die gleiche Skala/Maßeinheit aufweisen. Der größere Prozentwert zeigt auch die größere Streuungsbreite an.

Beispiel:

Der Variationskoeffizient ergibt sich aus den oben aufgeführten Daten (Produktions- und Laborberufe der Chemischen Industrie).

$$\mathrm{V} = \frac{0{,}827}{3{,}35} * 100 = 24{,}69$$

Das bedeutet, dass die Streuung 24,69 Prozent des Mittelwertes ausmacht. Die Streuung ist hier also relativ gering. Dieser Wert könnte jetzt mit dem Variationskoeffizienten einer anderen Gruppe verglichen werden, beispielsweise mit einer Gruppe, bei der auch die Gestreckte Abschlussprüfung eingeführt wurde und der eine vergleichbare Frage wie im Beispiel oben gestellt wurde.

Bivariate Analyse

Im Gegensatz zu den vorangegangenen Ausführungen zu univariaten Analysen, bei denen die Beschreibung von Lage und Streuung der Variablen im Vordergrund stand, sollen nun *die Beziehungen zwischen zwei Variablen* näher betrachtet werden.

► Kreuztabelle

In einer Kreuztabelle werden die Häufigkeitsverteilungen von zwei Variablen dargestellt und analysiert. Sie kann für alle Skalenniveaus verwendet werden, wobei die metrisch skalierten Daten vorab gruppiert werden sollten (z. B. Bildung von Altersgruppen statt jedes Alter einzeln aufzuführen), man spricht dann von kategorialen Variablen.

Die Anzahl der Kategorien sollte nicht zu groß gewählt werden, um die Kreuztabelle noch übersichtlich zu halten. Alle Kategorien sollten die gleiche Breite aufweisen und eine eindeutige Zuordnung zu einer Kategorie (Exklusivitätskriterium) ermöglichen und alle Messwerte umfassen (Exhaustivitätskriterium), vgl. Bortz/Döring (2002: 139). Als klassisches Beispiel kann die Gruppierung von Altersangaben in Jahren genannt werden: Kategorie 1: 21 bis 30 Jahre, Kategorie 2: 31 bis 40 Jahre, Kategorie 3: 41 bis 50 Jahre, usw.).

Beispiel:

Ausbildungsverhältnisse für Fachangestellte für Bürokommunikation (öffentlicher Dienst)

	Geschlecht		
Jahr	weiblich	männlich	gesamt
1992	1.255	94	1.349
1993	2.088	170	2.258
1994	2.363	206	2.569
1995	2.327	243	2.570
1996	2.380	277	2.657
1997	2.905	402	3.307
gesamt	**13.318**	**1.392**	**14.710**
% gesamt	90,5 %	9,5 %	100 %

Quelle: Wissenschaftliches Diskussionspapier, Heft 37, Evaluation der Büroberufe

Inferenzstatistik

Mithilfe der Inferenzstatistik kann aufgrund von Daten, die in einer Stichprobe erhoben worden sind, auf die Verteilung von Merkmalen in einer Grundgesamtheit geschlossen werden, aus den beobachteten Messwerten (der Stichprobe) können also Schätzwerte für die Grundgesamtheit abgeleitet werden.

Mit Methoden (Tests) der Inferenzstatistik können Populationsparameter geschätzt und Hypothesen getestet werden.

Mit den Daten, die sich im Rahmen der deskriptiven Statistik gewinnen lassen und den unten in der Tabelle genannten sieben inferenzstatistischen Verfahren lassen sich nahezu alle Fragen, die im Zusammenhang mit der Evaluation von Ausbildungsordnungen auftreten können, beantworten.

Übersicht der inferenzstatistischen Verfahren

Skalenniveau	Anzahl der Stichproben*	Anzahl der Messzeitpunkte	Testverfahren	Voraussetzungen / Besonderheiten
Nominalskala	≥ 2	1	Chi-Quadrat-Test Binomialtest[46]	minimale Voraussetzungen: nominalskalierte Daten
Ordinalskala	2	1	Mann-Whitney U-Test (bei unabhängigen Stichproben)	auch für „kleine" Gruppen mit N ≤ 20 geeignetes Testverfahren
Ordinalskala	1	2 (Messzeitpunkte od. 2 Variablen)	Wilcoxon-Test für Paardifferenzen (bei abhängigen Stichproben)	auch für „kleine" Gruppen mit N ≤ 20 geeignetes Testverfahren
Intervallskala	1	1	T-Test (für eine Stichprobe)	Vergleich einer empirischen Stichprobe mit einem Referenz-Mittelwert

[46] Der Binominaltest überprüft, ob der Anteil in einer Stichprobe (z. B. Frauen) sich signifikant vom Anteil einer Grundgesamtheit (oder einer anderen Stichprobe) unterscheidet.

Skalenniveau	Anzahl der Stichproben*	Anzahl der Messzeit-punkte	Testverfahren	Voraussetzungen / Besonderheiten
Intervallskala	2	1	T-Test (für unabhängige Stichproben)	Varianzhomogenität als Voraussetzung, für Gruppen mit N ≥ 30
Intervallskala	1	2	T-Test (für abhängige Stichproben)	Normalverteilung der Differenzen, für Gruppen mit N ≥ 30
Intervallskala	≥ 2	1	Einfaktorielle Varianzanalyse (ANOVA)	unabhängige Stichproben, homogene Varianzen, für Gruppen N ≥ 30

Quelle: NETQUES – Beratung für HR Management, Wuppertal

► *) Abhängige und unabhängige Stichproben

Stichproben werden als „abhängig" bezeichnet, wenn dem Wert der einen Variablen genau ein Wert einer anderen Variablen zugeordnet werden kann. Vereinfacht ausgedrückt können drei Arten abhängiger Stichproben unterschieden werden:

Prä-post-Vergleich: Personen werden zu zwei Zeitpunkten in Bezug auf dasselbe Merkmal befragt, wobei eine exakte Zuordnung der Messwertpaare möglich sein muss. Bei Evaluationsfragestellungen kann in solchen Fällen die Anonymität der Befragung nicht immer gewährleistet werden.

Matched Samples: Zum Vergleich von zwei Gruppen, beispielsweise Interventions- und Kontrollgruppe, werden „künstliche Zwillinge" verglichen. Beim Vorgang der Parallelisierung werden nach vorher definierten Kriterien vergleichbare Personen (z. B. in Bezug auf Alter, Geschlecht, Bildung) herangezogen, um die Wirksamkeit einer Maßnahme zu überprüfen.

„echte Zwillinge": Damit sind klassische Zwillings- oder Geschwisterstudien sowie Studien zu Ehepaaren gemeint.

Natürlich gibt es noch eine ganze Reihe weiterer Verfahren, die aber im Rahmen der vorliegenden Arbeitshilfe nicht weiter ausgeführt werden können. Diesbezüglich wird auf die umfangreiche Methodenliteratur verwiesen.

► Signifikanz

In den meisten statistischen Berechnungen taucht häufig der Begriff „Signifikanz" (Sig. oder ähnlich abgekürzt) auf. Was aber bedeutet es, dass ein Ergebnis signifikant, also überzufällig ist? Dazu muss folgende Überlegung angestellt werden: Die Ergebnisse aus einer Stichprobe stellen eine Schätzung dar: wie sich die Ergebnisse verteilen würden, wenn alle Personen (oder Merkmale) einer Grundgesamtheit (zum Beispiel alle Ausbilderinnen und Ausbilder im Beruf Mechatroniker und Mechatronikerin) befragt bzw. untersucht worden wären.

Voraussetzung für eine möglichst genaue Schätzung der Population ist eine repräsentative Stichprobe (weitere Informationen hierzu finden Sie in der Arbeitshilfe Stichprobenauswahl). Anhand der so erhobenen Stichprobenparameter können Rückschlüsse auf die zugrunde liegende Population gezogen werden. Signifikanztests können jedoch keine 100-prozentig gesicherten Ergebnisse liefern, sondern nur mit einer vorher definierten Wahrscheinlichkeit Aussagen treffen. D. h., ein signifikantes Ergebnis ist ein Ergebnis, das überzufällig entstanden ist. In den Sozialwissenschaften hat sich eine Irrtumswahrscheinlichkeit von 5 Prozent etabliert, so-

dass Aussagen über Unterschiede oder Zusammenhänge mit einer Wahrscheinlichkeit von 95 Prozent getroffen werden können.

Die Signifikanz eines statistischen Tests kann auf zwei Arten bestimmt werden: durch den p-value (probability value, häufig als sig. bezeichnet) oder über den empirisch errechneten Wert des jeweiligen Signifikanztests (z. B. t oder Chi^2).

Wird die Signifikanz in einem sogenannten „**p-Wert**“ ausgedrückt, gilt die Regel:

Ist der **p-Wert kleiner als oder gleich wie das gewählte Signifikanzniveau** ($\leq\alpha$), d. h. 0,05 oder 0,01 oder 0,001, dann ist das Ergebnis nicht mit dem Zufall vereinbar und es darf von einem statistisch signifikanten Ergebnis ausgegangen werden. Das heißt die Alternativhypothese[47] H_1 ist richtig; es gibt also einen Unterschied/Zusammenhang zwischen den gefundenen Werten.

Ist der **p-Wert größer als das gewählte Signifikanzniveau** ($>\alpha$), also größer als 0,05 oder 0,01 oder 0,001, dann ist die Nullhypothese[48] H_0 richtig, es gibt also keinen statistisch bedeutsamen Unterschied/Zusammenhang zwischen den Werten.

Der p-Wert gibt an, wie wahrscheinlich das empirisch gefundene Ergebnis (bei gültiger H_0) ist. Ein Ergebnis darf erst als „signifikant“ klassifiziert werden, wenn die Wahrscheinlichkeit $\leq$5 % (bzw. 1 % oder 0,1 %) ist.

Wird die Signifikanz in einem sogenannten „**t-Wert**“ ausgedrückt, gilt die Regel:

Ist der **t-Wert kleiner als der sogenannte „kritische Wert“** (der einer t-Verteilungstabelle entnommen werden kann bzw. von SPSS automatisch errechnet wird, siehe Infokasten 6), dann ist das Ergebnis ein Zufallsergebnis und die Nullhypothese H_0 ist richtig, es gibt also keinen bedeutsamen Unterschied/Zusammenhang zwischen den Werten. Ist der **t-Wert größer als der sogenannte „kritische Wert“**, dann ist das Ergebnis kein Zufallsergebnis mehr und die Alternativhypothese H_1 ist richtig, es gibt also einen statistisch bedeutsamen Unterschied/Zusammenhang zwischen den Werten.

Gleiches gilt für die Chi^2-Verteilung, die entsprechenden Tabellen sind in allen gängigen Statistikbüchern (dort meist im Anhang) zu finden.

Merke: Zu beachten ist, dass ein signifikantes Ergebnis noch keine Auskunft über die Bedeutsamkeit des Befundes gibt! „Die inhaltliche Bedeutsamkeit ist eine Frage der absoluten Differenz von Stichprobenmittelwert und Mittelwert der Grundgesamtheit und nicht eine Frage, die durch den Signifikanztest beantwortet wird.“ Kuckartz (2010: 145).

Um die Bedeutsamkeit eines ermittelten Zusammenhangs einordnen zu können, bedient man sich der Effektgröße bzw. Effektstärke (engl. effect size).

Hinweis: Möglichkeiten zur Berechnung von Effektgrößen (für die gängigsten statistischen Tests) bieten zum Beispiel die Internetseiten verschiedener Universitäten an. Dort können online Effektgrößen berechnet werden.

Darstellung einzelner Signifikanztests

Nachfolgend werden einige der gebräuchlichsten Signifikanztests vorgestellt. Sie dienen dazu, Hypothesen zu überprüfen und sollten entsprechend dem Skalenniveau der Daten ausgewählt werden. Das Ergebnis eines Signifikanztests ist folglich eine Wahrscheinlichkeitsaussage über

[47] Die Alternativhypothese stellte eine Aussage dar, die im Mittelpunkt des Interesses steht. Mit ihr möchte man ein bestimmtes Phänomen erklären und Zusammenhänge offenlegen.
[48] Die Nullhypothese ist im Vergleich zur Alternativhypothese eine formale Gegenhypothese. Mit ihr wird behauptet, dass die zur Alternativhypothese komplementäre Aussage richtig ist.

ein Stichprobenergebnis im Lichte der zuvor aufgestellten Hypothesen, vgl. Sedlmeier (2008: 366).

► Chi-Quadrat-Tests (χ^2-Tests) für nominalskalierte Daten

Bei nominalskalierten Daten ist die Berechnung von Lage- und Streuungsparametern nur sehr eingeschränkt möglich. Es kann aber anhand von Häufigkeiten z. B. festgestellt werden, in welchen Bundesländern die befragten Personen leben, oder ob Metallbauer und Metallbauerinnen insgesamt mit ihrem gewählten Beruf zufriedener sind als Feinwerkmechaniker und Feinwerkmechanikerinnen, oder ob es in einem Bundesland mehr Metallbauer und Metallbauerinnen oder mehr Feinwerkmechaniker und Feinwerkmechanikerinnen gibt. Aus solchen Daten können natürlich wieder Rückschlüsse auf andere erhobene Daten gezogen werden.

Aus beliebigen nominalskalierten Daten einer Stichprobe dürfen ebenfalls Aussagen zur Grundgesamtheit (Population) gemacht werden.

Jene Signifikanztests, welche zur Analyse von Häufigkeiten nominalskalierter Daten angewendet werden, werden als χ^2-Tests (Chi-Quadrat-Tests) bezeichnet.

Beim χ^2-Test bezieht sich die Nullhypothese auf die erwartete Häufigkeitsverteilung in der Grundgesamtheit und die Abweichung von derselben (Residuum). Gibt es zwischen der beobachteten und der erwarteten Häufigkeitsverteilung eine Abweichung, deren Wahrscheinlichkeit – bei Gültigkeit der Nullhypothese – $\leq \alpha$ ist, so ist das Ergebnis signifikant und die Alternativhypothese wird angenommen.

Grundsätzlich kann mit dem χ^2-Test jede Annahme über die Verteilung einer (nominalen) Häufigkeit getestet werden. In der Regel wird man aber von einer Gleichverteilung des getesteten Merkmals ausgehen.

Die χ^2-Verteilung ist eine Stichprobenverteilung (so wie die t-Verteilung) und ihre Form hängt von der Zahl der Freiheitsgrade ab. Diese Verteilung hat nur positive Werte. Jedem χ^2-Wert ist ein p-Wert zugeordnet (der aus entsprechenden Tabellen abgelesen werden kann).

► χ^2-Test für eine Variable

Mit dem χ^2-Test für eine Variable kann geprüft werden, ob die in einer Stichprobe beobachtete Häufigkeitsverteilung der untersuchten Variable signifikant von einer Häufigkeitsverteilung abweicht, die in der Grundgesamtheit vermutet wird, vgl. Sedlmeier/Renkewitz (2008: 552).

Beispiel: Eine Annahme könnte sein, dass die Betriebe, in denen Metallbauer und Metallbauerinnen ausgebildet werden, gleichmäßig über alle 11 befragten Bundesländer verteilt sind (H_0).

187 ausbildende Metallbaubetriebe aus 11 Bundesländern haben beantwortete Fragebogen an das BIBB zurückgesendet.

Bundesland (k)	Beobachtetes N	Erwartete Häufigkeit $\frac{N}{k}$	Residuum
Baden-Württemberg	7	17,0	-10,0
Bayern	68	17,0	51,0
Brandenburg	11	17,0	-6,0
Bremen	7	17,0	-10,0
Hessen	9	17,0	-8,0
Niedersachsen	18	17,0	1,0

Bundesland (k)	Beobachtetes N	Erwartete Häufigkeit $\frac{N}{k}$	Residuum
Nordrhein-Westfalen	37	17,0	20,0
Rheinland-Pfalz	22	17,0	5,0
Sachsen-Anhalt	4	17,0	-13,0
Schleswig-Holstein	2	17,0	-15,0
Thüringen	2	17,0	-15,0
Gesamt (N)	**187**		

Statistik für Test	
	Bundesland
Chi-Quadrat	232,118[a]
df	10
Asymptotische Signifikanz	,000

a: Bei 0 Zellen (,0 %) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 17,0.

Aus einer χ^2-Tabelle (siehe Anhang in den bekannten Statistik-Lehrbüchern) wird der χ^2-Wert für df[49] (Freiheitsgrade) = 10 abgelesen. Dieser Tabellenwert beträgt 18,31. Der in der Untersuchung ermittelte χ^2-Wert beträgt 232,12 und liegt damit deutlich über dem χ^2-Wert aus der Tabelle. Es gilt: χ^2(Untersuchung) > χ^2(Tabelle), damit ist eine signifikante Abweichung gegeben und die Alternativhypothese H_1 (es gibt eine signifikante Abweichung zwischen der beobachteten und der erwarteten Häufigkeitsverteilung) wird angenommen.

Wie aufgrund der vorliegenden Daten zu erwarten war, verteilen sich die ausbildenden Metallbaubetriebe aus der Stichprobe nicht gleichmäßig über die Bundesländer.

Dieses Ergebnis wird auch dadurch angezeigt, dass $p < 0{,}001$ beträgt, d. h., dass das Ergebnis hoch signifikant ist (hier wurde der χ^2-Wert von SPSS in einen p-Wert umgerechnet).

► χ^2-Test für zwei Variablen – Testung auf signifikante Zusammenhänge zwischen zwei Merkmalen

Mit diesem Test wird in der Regel geprüft, ob zwischen zwei untersuchten Merkmalen ein Zusammenhang besteht. Die Nullhypothese H_0 besagt hier, dass zwischen zwei Variablen in der Population kein Zusammenhang besteht. Die Ergebnisinterpretation folgt auch beim χ^2-Test für zwei Variablen der gleichen Logik, wie der χ^2-Test für eine Variable.

Bei einem signifikanten Ergebnis für einen χ^2-Test kann auch geklärt werden, wie stark dieser Zusammenhang ist. Dazu muss der Phi-Koeffizient, der Kontingenzkoeffizient C oder Cramers V berechnet werden.

[49] Der Freiheitsgrad df wird aus der Anzahl der möglichen Kategorien berechnet: df = k–1. Formel nach Sedlmeier (2008: 555).

► Zusammenhangsmaß für nominalskalierte Variablen: der Kontingenzkoeffizient C

Sofern der χ^2-Test ein signifikantes Ergebnis geliefert hat, können zwei Variablen, die Nominalskalenniveau aufweisen, mithilfe des Kontingenzkoeffizienten C (nach Pearson) hinsichtlich der **Stärke ihres Zusammenhangs** hin untersucht werden. Dabei wird eine Variable als r (row) und die zweite Variable als c (column) bezeichnet[50]. Für C gilt immer: $C \geq 0$ und < 1. Der maximale Wert, den C annehmen kann, hängt von der Spalten- und Zeilenzahl der Kontingenztafel ab. So kann z. B. in einer 3(r) *x* 3(c) Kontingenztafel der Wert von C maximal 0,82 sein.

$$c_{max} = \sqrt{\frac{R-1}{R}}$$

Hinweis: Die Formel dazu ist zitiert nach Leonhart (2004: 226).

Der Anteil des empirisch errechneten Zusammenhangs C geteilt durch den maximalen Zusammenhang (C_{max}) gibt den korrigierten Zusammenhang C_{corr} an. Dieser kann Werte zwischen 0 und 1 annehmen und entsprechend interpretiert werden: 0,2 entspricht einem schwachen, 0,5 einem mittleren und 0,7 einem starken Zusammenhang.

Je größer C_{corr} ausfällt, desto stärker der Zusammenhang, vgl. Kuckartz (2010: 256).

► Zusammenhangsmaß für ordinalskalierte Variablen: Spearman-Korrelation

Die Rangkorrelation nach Spearman basiert auf der Vergabe von Rangplätzen und darf auch auf Ordinalskalenniveau gerechnet werden. Ein linearer Zusammenhang muss nicht bestehen, es genügt auch Monotonie (nur monoton steigende oder fallende Werte bzw. Kurven).

Beispiel:

Im Rahmen einer Bildungsmaßnahme wurden Workshops zur Kompetenzberatung angeboten. Um herauszufinden, für welche Klientel der Workshop besonders hohe Bedeutung hatte, wurde diese Einschätzung (1 = sehr geringe Bedeutung, 5 = sehr hohe Bedeutung, Ordinalskalenniveau) mit der allgemeinen Selbstwirksamkeitserwartung (Intervallskalenniveau) korreliert. Aufgrund einer ordinalskalierten Variable wurde die Spearman-Korrelation angewendet.

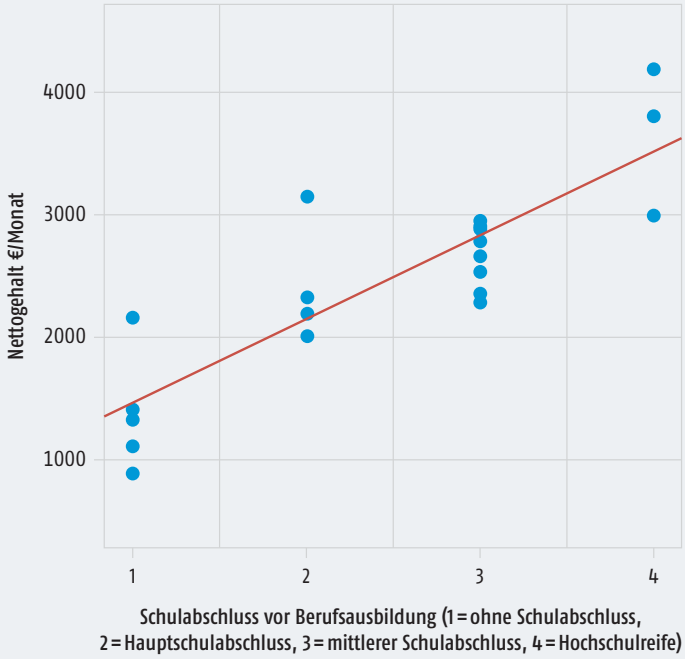
Correlations

			1. Welche Bedeutung messen Sie dem Workshop bei?	Allgemeine Selbstwirksamkeit
Spearman's rho	1. Welche Bedeutung messen Sie dem Workshop bei?	Correlation Coefficient	1,000	,288**
		Sig. (2-tailed)	-	,001
		N	140	140
	Allgemeine Selbstwirksamkeit	Correlation Coefficient	,288**	1,000
		Sig. (2-tailed)	,001	-
		N	140	144

**: Correlation is significant at the 0.01 level (2-tailed).

[50] Als Erstes werden die beobachteten Häufigkeiten in einer Kontingenztafel mit r x c Feldern notiert, danach werden die Randverteilungen berechnet. Zur weiteren Berechnung von C muss zunächst χ^2 (Chi²) berechnet werden. χ^2 wird dann in die Formel für C eingetragen.

Abbildung 2



Es zeigte sich ein *signifikanter Zusammenhang* zwischen den beiden Variablen ($p = 0{,}001$). Mit einer höheren Selbstwirksamkeitserwartung gehen signifikant höhere Werte für die Bedeutung des Workshops einher. Das heißt, Personen, die stärker an eigene Fähigkeiten glauben, haben von diesem Workshop mehr profitiert. Aus Ergebnissen wie diesem können Empfehlungen zur Gestaltung der künftigen Workshops abgeleitet werden. Die *Stärke des Zusammenhangs* kann bei einem Korrelationskoeffizienten von $r_{Spearman} = 0{,}288$ allerdings nur als gering eingeschätzt werden.

► Zusammenhangsmaß für metrische Variablen: Korrelationskoeffizient nach Pearson r

Die zugrunde liegenden Daten müssen mindestens auf Intervallskalenniveau (metrisch) sein und sollten normalverteilt sein, damit $r_{Pearson}$ gebildet werden darf.

Der Korrelationskoeffizient $r_{Pearson}$ weist die *Stärke des Zusammenhangs* von Variablen aus, erklärt die wahrscheinliche Ursache für diesen Zusammenhang aber nicht.

Wenn die Daten aus zwei Variablen grafisch dargestellt werden, dann ergibt sich eine sogenannte Punktwolke (Scatterplot). Durch diese Punktwolke wird eine Gerade gelegt. Bei starker Korrelation liegen fast alle Datenpunkte nahe an der Geraden bei schwachem Zusammenhang liegen die meisten der Datenpunkte von der Geraden entfernt.

Der den Zusammenhang beschreibende Korrelationskoeffizient r nimmt immer Werte zwischen –1 und +1 an. Das Vorzeichen von r ergibt sich aus der Steigung der Geraden und bestimmt die Lage der Geraden und der Punktwolke im Koordinatensystem.

Bei $r = +1$ liegt ein maximaler starker gleichberechtigter Zusammenhang vor, steigen die Werte der Variablen x, steigen auch die Werte der Variablen y. Alle empirischen Beobachtungen liegen im Achsenkreuz auf einer ansteigenden Geraden. Bei $r = -1$ liegt ein maximal starker gegenläufiger Zusammenhang vor (steigt x, fällt y). Alle Punkte liegen auf einer fallenden Geraden. Bei $r = 0$ liegt kein statistischer Zusammenhang vor. Alle Werte, die dazwischen liegen, können entsprechend interpretiert werden. Es wird jeweils Linearität zwischen x und y angenommen.

Beispiel für einen positiven Zusammenhang (positives r): Zusammenhang zwischen Schulabschluss und späterem Nettoeinkommen. Mit der Höhe des Schulabschlusses nimmt auch die Höhe des späteren monatlichen Nettoeinkommens zu.

Die Erstellung eines Scatterplots empfiehlt sich in jedem Fall, da grafische Darstellungen einen raschen Überblick über die Daten ermöglichen und Ausreißer erkannt werden können.

Der Wert von r (zwischen –1 und +1) ist immer mit Vorsicht zu interpretieren. Nicht immer sind alle Faktoren, die zum Ergebnis einer Korrelation beigetragen haben, bekannt. Daher empfiehlt es sich, nicht einfach von einer hohen oder niedrigen Korrelation von zwei Variablen zu sprechen, sondern das Ergebnis möglichst umsichtig zu interpretieren, das heißt, alle möglichen anderen Variablen in die Überlegungen und die Erklärungsversuche mit einzubeziehen und ggf. vergleichbare Werte aus anderen Untersuchungen – mit gleichen oder ähnlichen Untersuchungsgegenständen – in die Ergebnisinterpretation mit einzubeziehen.

Als Faustregel gilt ein Korrelationskoeffizient

Wert	Interpretation
r bis 0,2	als sehr geringe Korrelation
r bis 0,5	als geringe Korrelation
r bis 0,7	als mittlere Korrelation
r bis 0,9	als hohe Korrelation
r über 0,9	als sehr hohe Korrelation

Beispiel:

Im Rahmen der Evaluation der Gestreckten Gesellenprüfung (GAP) wurden Betriebe, die Metallbauer und Metallbauerinnen ausbilden, sowohl nach ihrer Betriebsgröße, als auch nach der Motivation (im Zusammenhang mit der Gestreckten Gesellenprüfung) der Ausbilderinnen und Ausbilder im Betrieb befragt. Daraufhin wurde die Hypothese aufgestellt, dass es durch die Gestreckte Gesellenprüfung zu einem Motivationsschub bei den Ausbilderinnen und Ausbildern gekommen sein könnte. Als H_1 wurde formuliert, dass der Motivationsschub umso höher ausfällt, je größer der Betrieb ist (weil große Betriebe in der Regel mehr Personal, Räume und Material für die Umsetzung der Gestreckten Gesellenprüfung bereitstellen können).

Dieser mögliche Zusammenhang wurde mittels einer $r_{Pearson}$ untersucht. SPSS[51] hat folgendes Ergebnis aufgezeichnet:

Korrelationen			
		Größe des Ausbildungsbetriebes	Auswirkung GAP auf Motivation der Ausbildungsverantwortlichen
Größe des Ausbildungsbetriebes	Korrelation nach Pearson	1	,014
	Signifikanz (2-seitig)		,850
	N	187	187

[51] Ursprünglich Statistical Package for the Social Sciences, ist SPSS heute nur noch ein Firmenname. SPSS gehört heute zu IBM.

Korrelationen			
		Größe des Ausbildungsbetriebes	Auswirkung GAP auf Motivation der Ausbildungsverantwortlichen
Auswirkung der GAP auf Motivation der Ausbildungsverantwortlichen	Korrelation nach Pearson	,014	1
	Signifikanz (2-seitig)	,850	
	N	187	187

Es zeigt sich *kein Zusammenhang* (r = 0,014) zwischen beiden Variablen, darüber hinaus *ist das Ergebnis nicht signifikant*. Dies bedeutet, dass das vorliegende Ergebnis absolut zufällig ist und die Betriebsgröße keinen Einfluss auf die Motivation (im Zusammenhang mit der Gestreckten Gesellenprüfung) der Ausbildungsverantwortlichen hat.

► Bestimmtheitsmaß B (Determinationskoeffizient)

Für intervallskalierte Daten kann die *Stärke des Zusammenhangs* auch durch das Bestimmtheitsmaß B dargestellt werden. B lässt Rückschlüsse darauf zu, welcher Anteil der Veränderung des einen Merkmals aus den Veränderungen des anderen Merkmals erklärt werden kann. B nimmt immer einen positiven Wert an, liegt also zwischen 0 und 1 oder zwischen 0 und 100 Prozent.

Formel für das Bestimmtheitsmaß nach Clauss/Finze/Partzsch (2002: 65):

$B = r^2$

B beschreibt nach Leonhart (2004: 195) „den Anteil der gemeinsamen Varianz beider Merkmale an der Gesamtvarianz von 1."

Für das oben genannte Bespiel ergibt sich somit eine gemeinsame erklärte Varianz der beiden Variablen von $0{,}014^2$ und somit von 0,02 Prozent (oder 0,000196).

An diesem Beispiel wird deutlich, dass es nur Sinn macht B zu berechnen, wenn das Ergebnis signifikant ist, was in diesem Falle nicht gegeben war.

Sowohl r als auch B lassen nur Aussagen über den Grad des Zusammenhangs von Variablen zu. Der Schluss auf die möglichen Ursachen dieses Zusammenhangs ist in der Regel Spekulation und wird weder von r noch von B unterstützt.

Sogenannte „Scheinkorrelationen", also der Versuch, aufgrund einer Korrelation einen Kausalzusammenhang zwischen Variablen herzustellen ohne die sogenannte(n) intervenierende Variable(n) zu berücksichtigen, führen bei der Ergebnisinterpretation schnell zu falschen Schlüssen (vgl. Infokasten 4).

Infokasten 4

„Ein bekanntes Beispiel in der Statistik ist die Korrelation zwischen der Zahl der Kindergeburten und der Zahl der Storchenpaare in verschiedenen Regionen. Obwohl es eine Korrelation zwischen der Zahl der Geburten und der Zahl der Storchenpaare gibt, gibt es keinen kausalen Zusammenhang. Tatsächlich gibt es aber einen kausalen Zusammenhang zu einer dritten (intervenierenden) Variablen: der Ländlichkeit der Region. Je ländlicher eine Region ist, desto höher ist die Zahl der Kindergeburten und desto größer ist die Zahl der Storchenpaare. Dies führt zu der Korrelation zwischen der Zahl der Kindergeburten und der Zahl der Storchenpaare."

Quelle: http://de.wikipedia.org/wiki/Scheinkorrelation (Stand 07.01.2013).

► Regressionsrechnung

Die Signifikanz zeigt den *überzufälligen Zusammenhang* zwischen zwei Variablen auf.

Die Korrelation zeigt die *Stärke des Zusammenhangs* zwischen zwei Variablen auf.

Die Regression zeigt die *Art dieses Zusammenhangs* auf.

Mittels einer Regressionsanalyse können bei stochastischem Zusammenhang zwischen zwei Variablen Werte vorhergesagt werden, vgl. Leonhart (2004: 229). Die Regressionsrechnung (die Intervallskalenniveau voraussetzt) unterscheidet zwischen abhängigen und unabhängigen Variablen, welche vorher inhaltlich oder theoretisch begründet werden müssen. Die Regressionsanalyse erlaubt es, die Art der Zusammenhänge zu modellieren, vgl. Rudolf/Müller (2004: 31).

Einfache Lineare Regression

Aus einer unabhängigen Variablen soll die abhängige Variable geschätzt werden; dies geschieht, bei einfacher linearer Regression, mittels einer Regressionsgeraden mit der allgemeinen Funktion $y=a+bx$. Sind a und b bekannt, ist auch die Lage der Regressionsgeraden bekannt und (bei linearem Zusammenhang) zu jedem gegebenen Wert x kann ein Wert y geschätzt werden.

Multiple lineare Regression

Aus mehreren unabhängigen Variablen soll eine abhängige Variable geschätzt oder deren Einfluss auf diese Variable erklärt werden. Nach Rudolf/Müller (2004: 41) ist es möglich, „(…) aus einer großen Anzahl von Prädiktorvariablen diejenigen auszuwählen, die zur Vorhersage der Kriteriumsvariablen optimal geeignet sind.“ Bei der Anwendung dieses Verfahrens sollten die Voraussetzungen dafür geprüft werden. Hinweise dazu finden sich bei Rudolf/Müller (2004).

Testverfahren zur Ermittlung von signifikanten Unterschieden (für Variable mit metrischem Skalenniveau)

Nach der Beschreibung von Zusammenhangsmaßen für Daten mit unterschiedlichem Skalenniveau, erfolgt nun eine Darstellung der wichtigsten Signifikanztests **zur Beurteilung der Signifikanz von Unterschieden**. Auch hier werden die Verfahren getrennt nach Skalenniveau dargestellt, beginnend mit den Verfahren für metrische Daten.

Infokasten 5

Grundsätzliches zu Tabellen:

Unterschiedlichen Signifikanztests liegen unterschiedliche Verteilungen zugrunde.

Die sich daraus ergebenden Verteilungstabellen (zum Beispiel z-Tabelle, t-Tabelle usw.) werden, aus Platzgründen und Gründen der grafischen Darstellung, in der Regel nur mit ***positiven*** Werten aufgeführt. Ist ein negatives Ergebnis vorhanden, wird dieses in der Tabelle wie das entsprechende positive Ergebnis abgelesen.

Eine Ausnahme bildet die F-Verteilung, die sich nur im positiven Bereich bewegt.

▸ One-Sample – t-Test

Beim t-Test mit einer Stichprobe kann zum Beispiel die Frage geklärt werden, ob der in einer Stichprobe gefundene Mittelwert $\bar{x}$ signifikant von einem theoretischen Mittelwert μ_T abweicht bzw. ob diese Abweichung zufällig ist oder nicht. Die Grundgesamtheit muss dabei normalverteilt und das Merkmal intervallskaliert[52] sein. Der t-Test kann auch für Stichproben mit $N < 30$ (Standard $N \geq 30$) verwendet werden.

(fiktives) Beispiel:

Es ist aus einer Studie bekannt, dass die Teilnehmerinnen und Teilnehmer der Fortbildung zum geprüften Medienfachwirt bzw. zur geprüften Medienfachwirtin im Bundesdurchschnitt 32 Jahre alt sind. Es wird vermutet, dass in einem bestimmten Bundesland die Teilnehmerinnen und Teilnehmer älter sind. Folgende Hypothesen wurden formuliert:

H_0: Die Teilnehmer und Teilnehmerinnen aus Bremen unterscheiden sich im Alter nicht vom Bundesdurchschnitt von 32 Jahren,

H_1: Die Teilnehmerinnen und Teilnehmer aus Bremen unterscheiden sich im Alter signifikant vom Bundesdurchschnitt. ($\alpha = 5\,\%$).

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Alter	57	35,70	9,710	1,286

One-Sample Test

	Test Value = 32					
					95 % Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Alter	2,878	56	,006	3,702	1,13	6,28

Es zeigt sich ein höchst signifikanter Unterschied im Alter der Teilnehmerinnen und Teilnehmer zwischen Bremen und dem Bundesdurchschnitt (Population) ($p = 0{,}006$). Die Teilnehmerinnen und Teilnehmer aus Bremen sind im Durchschnitt 35,7 Jahre alt, die Standardabweichung liegt bei 9,71 Jahren. Der empirisch gefundene t-Wert von 2,88 liegt über dem kritischen Wert der t-Verteilung (bei $df = 56$) von 2,0. Somit wird die Alternativhypothese angenommen und H_0 verworfen.

[52] Bei sogenannten Ratingskalen (z. B. „sehr einfach" bis „sehr schwer" oder „stimme voll zu" bis „lehne strikt ab" usw.) – die auch in Fragebogen des BIBB häufig eingesetzt werden – gibt es eine lange andauernde Grundsatzdiskussion in der Fachwelt, ob Ratingskalen das Intervallskalenniveau erfüllen oder letztlich doch nur Ordinalskalen sind, die bis zum heutigen Tage anhält. Die Ergebnisse von Berechnungen aus Ratingskalen auf Intervallskalenniveau könnten aber ungenauer sein als auf Ordinalskalenniveau, daher sollte (bis zur Klärung des Problems) mit der Interpretation der Ergebnisse, insbesondere wenn diese grenzwertig sind, vorsichtig umgegangen werden. Will man „auf der sicheren Seite" sein, sollte man einen Test für ordinalskalierte Daten (sogenannte Rangtests) anwenden.

► t-Test für abhängige (gepaarte)[53] Stichproben

Mit dem t-Test kann auch die Frage geklärt werden, ob die Mittelwerte zweier abhängiger Stichproben (aus zwei Populationen) signifikant verschieden sind.

Beispiel:

Auszubildende in den Produktions- und Laborberufen der Chemischen Industrie wurden befragt, wie schwierig sie den schriftlichen und den praktischen Teil der Gestreckten Abschlussprüfung (Teil 2) fanden:

Wie schwierig fanden Sie den *schriftlichen Teil* der Gestreckten Abschlussprüfung Teil 2?			
sehr einfach	eher einfach	eher schwer	sehr schwer

Wie schwierig fanden Sie den *praktischen Teil* der Gestreckten Abschlussprüfung Teil 2?			
sehr einfach	eher einfach	eher schwer	sehr schwer

H_1: Es gibt einen Unterschied bei der Einschätzung des Schwierigkeitsgrades der schriftlichen und der praktischen Prüfungsteile.

H_0: Es gibt keinen Unterschied bei der Einschätzung des Schwierigkeitsgrades der schriftlichen und der praktischen Prüfungsteile. ($\alpha = 5\,\%$)

Das Ergebnis des t-Tests für abhängige (gepaarte) Stichproben:

Statistik bei gepaarten Stichproben				
	Mittelwert	N	Standardabweichung	Standardfehler des Mittelwertes
schriftlicher Teil wie schwierig?	2,75	853	,806	,028
praktischer Teil wie schwierig?	2,14	853	,884	,030

Test bei gepaarten Stichproben								
	Gepaarte Differenzen					T	df	Sig. (2-seitig)
	Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	95 % Konfidenzintervall der Differenz				
				Untere	Obere			
schriftlicher Teil wie schwierig? praktischer Teil wie schwierig?	,613	1,036	,035	,544	,683	17,286	852	,000

[53] „Abhängig" bzw. „gepaart" bedeutet in diesem Zusammenhang: von einer Person wurden zwei Variablen erhoben.

Es zeigt sich ein höchst signifikantes Ergebnis ($p<0{,}001$), somit wurde der Schwierigkeitsgrad für die schriftliche Prüfung ($AM = 2{,}75$; $SD = 0{,}81$) signifikant schwerer eingeschätzt als der praktische Teil ($AM = 2{,}14$; $SD = 0{,}88$).

Auch hier kann eine Interpretation des Testergebnisses des empirischen t-Werts vorgenommen werden. Allerdings kann aufgrund der hohen Freiheitsgrade von $df = 825$ (errechnet aus Anzahl der Messwertpaare – 1) kein kritischer t-Wert abgelesen werden. Da sich aber die t-Verteilung an die Standardnormalverteilung annähert, kann der kritische z-Wert von ± 1,96 zum Vergleich herangezogen werden. Der t-Wert von 17,286 liegt weit über 1,96 und somit kann das Ergebnis als signifikant klassifiziert werden.

Damit wird die Nullhypothese (H_0; Es gibt keinen Unterschied) verworfen und die Alternativhypothese (H_1; Es gibt einen Unterschied) angenommen.

Infokasten 6

Grundsätzlich gilt:

Alle empirisch errechneten t-Werte, die größer als der kritische t-Wert sind (zu entnehmen aus einer t-Tabelle oder von SPSS automatisch berücksichtigt), dürfen als signifikant betrachtet werden, da $p < \alpha$ (je größer der t-Wert, umso kleiner der p-Wert) und die Alternativhypothese wird angenommen.

Alle t-Werte, die kleiner als der kritische t-Wert sind, müssen als nicht signifikant betrachtet werden, da $p > \alpha$ (je kleiner der t-Wert, umso größer der p-Wert) und die Nullhypothese wird nicht verworfen.

Bei Freiheitsgraden > 120 wird anstelle des kritischen t-Werts der z-Wert von 1,96 zum Vergleich herangezogen.

Für das vorliegende signifikante Ergebnis wurde die Effektgröße über eine online-Eingabemaske (der Universität des Saarlandes) berechnet.

Es zeigte sich ein relativ großer Effekt[54] von 0,72 (Konfidenzintervall: 0,65 bis 0,80). Damit kann gesagt werden, dass der schriftliche Teil viel schwieriger eingeschätzt wurde als der praktische Teil.

Der t-Test für abhängige Stichproben eignet sich insbesondere für die Messung signifikanter Unterschiede bei prä-post-Messungen, da es sich hierbei um ein sehr trennscharfes Verfahren handelt.

► t-Test für unabhängige Stichproben

Mittelwertsunterschiede unabhängiger Stichproben können ebenso mittels t-Test auf Signifikanz getestet werden. Als Beispiel könnten die Lehrlinge der chemischen Industrie herangezogen werden. Es könnte untersucht werden, ob es signifikante Unterschiede in der Einschätzung der Schwierigkeit des schriftlichen Teils zwischen Lehrlingen im Labor- und im Produktionsbereich gibt. Eine weitere Möglichkeit wäre die Untersuchung von Unterschieden zwischen männlichen und weiblichen Lehrlingen in der Einschätzung der Schwierigkeit.

Voraussetzungen: metrisches Skalenniveau (im Falle der Einschätzung der Schwierigkeit durch die Lehrlinge umstritten), Normalverteilung der Daten in beiden Gruppen und homogene Varianzen. Die Varianzhomogenität wird durch SPSS beim unabhängigen t-Test automatisch mitberechnet. Zur Interpretation des t-Tests unabhängig sollte also das Ergebnis des Levene-Tests nicht signifikant sein.

[54] Klassifizierung der Effektgröße: ab 0,2=klein, ab 0,5=mittel, ab 0,8=groß, vgl. Leohart (2004: 398).

► Varianzanalyse (*ANOVA*[55])

Mittelwertunterschiede von mehr als zwei voneinander unabhängigen Stichproben können mittels Varianzanalyse auf Signifikanz getestet werden. Das ist sozusagen eine Erweiterung des t-Tests auf mehr als zwei Gruppen. Die Nullhypothese in der Varianzanalyse besagt, dass verschiedene untersuchte Populationen (zum Beispiel Auszubildende in unterschiedlichen Berufen) den gleichen Mittelwert haben. Die Alternativhypothese bei der Varianzanalyse formuliert lediglich das Gegenteil (es gibt Mittelwertunterschiede).

Voraussetzungen sind, dass die gefundenen Werte normalverteilt sind, dass die untersuchten Populationen gleiche Varianz haben (Varianzhomogenität) und dass die untersuchten Stichproben unabhängig sind.

► F-Test (Untertest zur Varianzanalyse)

„Der Vergleich der Varianzschätzung zwischen Gruppen mit der Varianzschätzung innerhalb von Gruppen wird in der Varianzanalyse anhand des Verhältnisses beider Größen durchgeführt. Dieses Verhältnis der beiden Varianzschätzungen wird auch als F-Wert bezeichnet (das F steht für R. A. Fisher, den Statistiker, der die Varianzanalyse entwickelte)“ (Sedlmeier/Renkewitz 2008).

Nach der Interpretation des Levene-Tests (sollte wie beim t-Test nicht signifikant sein) erfolgt die Interpretation der ANOVA anhand des p-Werts oder alternativ über den F-Wert und den kritischen Wert in der F-Verteilung.

► post-hoc-Tests (Untertest zur Varianzanalyse)

Da die ANOVA eine sogenannte „Omnibushypothese“ prüft, ist nach wie vor unklar, zwischen welchen Gruppen (hier Auszubildende in unterschiedlichen Berufen) Unterschiede bestehen. Um diese Frage beantworten zu können muss bei einem signifikanten Ergebnis der ANOVA ein sogenannter post-hoc-Test (von denen es mehrere gibt und die von SPSS angeboten werden) durchgeführt werden. Tukey oder Scheffe liefern Subsets, in denen jene Gruppen zusammengefasst sind, zwischen denen keine signifikanten Unterschiede bestehen.

Nichtparametrische Tests (verteilungsfreie Tests) für Unterschiede

Der t-Test und der F-Test verlangen sowohl Normalverteilung der Daten als auch Intervallskalenniveau. Sogenannte Rangtests können auch bei ordinalskalierten Daten angewendet werden und verlangen keine Normalverteilung, sie werden deshalb auch als „verteilungsfreie“ oder „nicht-parametrische“ Tests bezeichnet. Sie eignen sich beispielsweise sehr gut für Daten mit Ausreißern oder heterogenen Varianzen.

► U-Test (Mann-Whitney-Test) für zwei unabhängige Stichproben

Dieser Test basiert auf der Vergabe von Rangzahlen für erhobene Datenwerte. Er hat geringere Voraussetzungen als der t-Test und dient dem Vergleich zweier unabhängiger Stichproben.

Die aus einer Untersuchung erhaltenen Daten (Zahlenwerte) werden in eine gemeinsame Rangreihe gebracht. Daraus ergeben sich für die einzelnen Daten Rangplätze. Nun wird untersucht, wie viele Rangplatzüberschreitungen oder Rangplatzunterschreitungen (bezogen auf die mittleren Ränge) zwischen den ausgewählten Stichproben bestehen. Aus den Rangplätzen der Überschreitungen und der Unterschreitungen werden die Rangsummen gebildet.

[55] Analysis Of VAriance = Varianzanalyse.

Ist die Nullhypothese (es gibt keinen Unterschied) gültig, dann sind die mittleren Rangplätze nahezu gleich groß. Sind die mittleren Rangplätze „auf den ersten Blick“ deutlich unterschiedlich, darf davon ausgegangen werden, dass die Alternativhypothese (es gibt einen Unterschied) gültig ist.

Die Interpretation des Ergebnisses ist wie bei den zuvor präsentierten Testverfahren auf zwei Arten möglich, über den empirisch ermittelten Wert und den p-value.

Bei kleinen Stichproben ($N \leq 20$) kann der sogenannte U-Wert aus einer Tabelle direkt abgelesen werden. Bei größeren Stichproben muss der U-Wert in einen z-Wert transformiert werden (das macht SPSS automatisch) und dann kann das Ergebnis aus einer z-Tabelle abgelesen werden. Die Signifikanz des Ergebnisses kann anschließend mit einer Standardnormalverteilungstabelle geprüft werden (auch das macht SPSS automatisch).

Beispiel:

Auszubildende in den Produktions- und Laborberufen der Chemischen Industrie wurden befragt, wie schwierig sie den praktischen Teil der Gestreckten Abschlussprüfung (Teil 2) fanden:

Wie schwierig fanden Sie den *praktischen Teil* der Gestreckten Abschlussprüfung Teil 2?			
sehr einfach (1)	eher einfach (2)	eher schwer (3)	sehr schwer (4)

Folgende Hypothesen wurden formuliert:

H_0: Es gibt keinen Unterschied in der Einschätzung der Schwierigkeit des praktischen Teils zwischen Chemikanten/Chemikantinnen und Chemielaboranten/Chemielaborantinnen

H_1: Chemikanten/Chemikantinnen und Chemielaboranten/Chemielaborantinnen schätzen den Schwierigkeitsgrad des praktischen Teils unterschiedlich ein ($\alpha = 5\,\%$)

Das Ergebnis (nach Ausfiltern der Fälle „keine Angabe“) war:

Ränge				
	AZUBI als	N	Mittlerer Rang	Rangsumme
praktischer Teil wie schwierig?	Chemikant/Chemikantin	322	297,66	95845,00
	Chemielaborant/Chemielaborantin	316	341,76	107996,00
	Gesamt	638		

Statistik für Test [a]	
	praktischer Teil wie schwierig?
Mann-Whitney-U	43842,000
Wilcoxon-W	95845,000
Z	-3,516
Asymptotische Signifikanz (2-seitig)	,000

a. Gruppenvariable: AZUBI als

Es zeigte sich ein höchst signifikantes Ergebnis ($p < 0{,}001$), somit darf die Nullhypothese (es gibt keinen Unterschied zwischen Chemikanten/Chemikantinnen und Chemielaboranten/Chemielaborantinnen im Hinblick auf die Einschätzung des Schwierigkeitsgrades der praktischen Prüfungen) verworfen werden und die Alternativhypothese angenommen werden. Es gibt also bei der Einschätzung des Schwierigkeitsgrades der praktischen Prüfungen einen Unterschied zwischen den befragten Chemikanten und Chemielaboranten.

Alternativ zum p-Wert kann das Ergebnis auch anhand des z-Wertes interpretiert werden. Da die Stichproben der Chemikanten/Chemikantinnen und Chemielaboranten/Chemielaborantinnen größer als $N = 20$ waren, wurden die berechneten U-Werte in z-Werte überführt. Der z-Wert liegt hier, wie oben der Tabelle zu entnehmen ist, bei $Z = -3{,}516$ und somit weit über dem kritischen Wert von $\pm 1{,}96$ (dieser Wert aus der Standardnormalverteilung gilt bei allen zweiseitigen Testungen bei einer Irrtumswahrscheinlichkeit von 5 %)

Da die Chemielaboranten/Chemielaborantinnen die höheren mittleren Rangplätze (Mittlerer Rang = 341,76) aufweisen als die Chemikanten/Chemikantinnen (Mittlerer Rang = 297,66) kann gesagt werden, dass die Chemielaboranten und Chemielaborantinnen den Schwierigkeitsgrad der praktischen Prüfung signifikant höher eingeschätzt haben als die Chemikanten und Chemikantinnen.

► Wilcoxon-Test für zwei abhängige Stichproben

Der Test dient dem Vergleich zweier abhängiger Stichproben, wenn die Daten Ordinalskalenniveau aufweisen oder den Anforderungen metrischer Verfahren (Normalverteilung der Differenzen) nicht genügen.

Stichproben sind dann abhängig, wenn einer Person genau zwei Messwerte zugeordnet werden können. Das können Messungen des gleichen Merkmals zu zwei unterschiedlichen Zeitpunkten sein oder zu einem Zeitpunkt wurden zwei Variablen erhoben, die miteinander verglichen werden sollen.

(fiktives) Beispiel:

Vorgesetzte wurden zum Entlassungsrisiko ihrer Beschäftigten befragt, die eine Fortbildung zum Bankfachwirt/zur Bankfachwirtin machen. Die Einschätzung der „Fortbildungs-Wirksamkeit“ wurden mittels eines Fragebogens (BDI) zu zwei Zeitpunkten erhoben: zu Beginn der Fortbildung ihrer Beschäftigten und zum Abschluss der Fortbildung.

H_0: Es gibt keinen Unterschied hinsichtlich des Entlassungsrisikos zwischen den beiden Messzeitpunkten.

H_1: Bei Abschluss der Fortbildung ist das Entlassungsrisiko niedriger als zu Beginn der Fortbildung ($\alpha = 5\,\%$).

Ergebnis des Wilcoxon-Tests:

Ranks		N	Mean Rank	Sum of Ranks
Abschluss BDI – Beginn: Summe BDI	Negative Ranks	642	476,411215	305856
	Positive Ranks	239	345,878661	82665
	Ties	43		
	Total	924		
a	E: Summe BDI < A: Summe BDI			
b	E: Summe BDI > A: Summe BDI			
c	E: Summe BDI = A: Summe BDI			

In 642 von 924 Fällen (negative Ranks) waren die Werte zum Abschluss der Fortbildung niedriger als zu Beginn, die Fortbildung kann sozusagen als erfolgreich eingeschätzt werden. In 239 Fällen waren die Werte des Entlassungsrisikos zum Zeitpunkt des Abschlusses höher als bei der Aufnahme (positive Ranks) und in 43 Fällen wurde zu beiden Zeitpunkten der exakt gleiche Wert erhoben (Ties = Rangplatzbindungen).

Test Statistics(b)	
	E: Summe BDI – A: Summe BDI
Z	-14,778
Asymp. Sig. (2-tailed)	<0,000
a	Based on positive ranks.
b	Wilcoxon Signed Ranks Test

Das Ergebnis des Wilcoxon Tests ist höchst signifikant ($p<0{,}001$).

Auch hier kann zur Interpretation des Ergebnisses der z-Wert herangezogen werden. Der empirisch ermittelte Wert von 14,778 liegt weit über dem kritischen Wert von 1,96.

Somit kann von signifikant niedrigeren Entlassungsrisikowerten nach der Fortbildung ausgegangen werden.

Zusammenfassung

Bei der Erstellung von Erhebungsinstrumenten ist die Berücksichtigung des Skalenniveaus wichtig.

Primär nominalskalierte Daten erlauben nur wenige statistische Testverfahren, die über eine geringe Trennschärfe verfügen. Das heißt, Unterschiede müssen sehr groß ausgeprägt sein, damit sie zu einem signifikanten Ergebnis führen können.

Nonparametrische Tests für ordinale Daten liefern etwas mehr Information, deren Ergebnisse sind jedoch schwer kommunizierbar. Beispielsweise können aus mittleren Rangplätzen, wie sie beim U-Test als Ergebnis berichtet werden, nur Tendenzen in eine Richtung abgelesen werden. Und auch wenn ein Ergebnis signifikant ist, kann die Bedeutsamkeit (Effektgröße) nicht berechnet werden. Daten mit wenig Varianz (wie z. B. Items mit Ausprägungen von „1“ bis „4“), wie sie auch in den Beispielen angeführt sind, schwächen die Aussagekraft dieser Verfahren, weil zu viele Rangplatzbindungen bestehen. Gleiches gilt für die Spearman-Korrelation.

Die aussagekräftigsten Ergebnisse lassen sich mit metrischen Daten und Testverfahren erzielen. Aufgrund leicht verständlicher Mittelwerte können diese Ergebnisse gut kommuniziert werden. Die Möglichkeit der Berechnung von Effektgrößen ermöglicht zusätzlich eine Aussage über die Bedeutsamkeit der Ergebnisse.

Bei der sorgfältigen Konstruktion eines Erhebungsinstruments mit inkludiertem Pre-Test kann die Reliabilität von Item-Gruppen überprüft und das Instrument ggf. optimiert werden.

Literatur

Bortz, Jürgen: Statistik für Human- und Sozialwissenschaftler. Heidelberg 2005.
Bortz, Jürgen; Döring, Nicola: Forschungsmethoden und Evaluation. Berlin 2002.
Clauss, Günter; Finze, Falk-Rüdiger; Partzsch, Lothar: Statistik. Für Soziologen, Pädagogen, Psychologen und Mediziner. Grundlagen. Frankfurt 2002.
Köhler, Wolfgang; Schachtel, Gabriel; Voleske, Peter: Biostatistik. Heidelberg 2007.
Kromrey, Helmut: Empirische Sozialforschung. Opladen 2002.
Leonhart, Rainer: Lehrbuch Statistik. Einstieg und Vertiefung. Bern 2008.
Ludwig-Mayerhofer, Wolfgang: ILMES – Internet-Lexikon der Methoden der empirischen Sozialforschung. URL: http://wlm.userweb.mwn.de/ilmes.htm.
Nachtigall, Christof; Wirtz, Markus: Wahrscheinlichkeitsrechnung und Inferenzstatistik. Statistische Methoden für Psychologen Teil 2. Weinheim 2006.
NETQUES – Daten & Diagnostik: Inferenzstatistische Verfahren (Schulungsunterlagen). Wuppertal 2009.
Rudolf, Matthias; Müller, Johannes: Multivariate Verfahren. Göttingen 2012.
Sedlmeier, Peter; Renkewitz, Frank: Forschungsmethoden und Statistik in der Psychologie. München 2008
Voss, Werner; Khlavna, Veronika; Schöneck, Nadine M.: Einführung in die Datenanalyse und Datenmanagement mit SPSS (Schulungsunterlagen). Bochum 2006.

Arbeitshilfe: Erhebung qualitativer Daten

Die Erhebung qualitativer Daten unterscheidet sich erheblich von der quantitativen Datenerhebung. In dieser Arbeitshilfe soll ein Überblick über die Möglichkeiten qualitativer Datenerhebung gegeben werden. Im Fokus stehen dabei neben Interviewtechniken eine weitere Form kommunikativer Erhebung, die Gruppendiskussion sowie die Beobachtung. Welche dieser qualitativen Erhebungsmethoden im Rahmen der Evaluation von Ausbildungsordnungen hilfreich sein können und was bei ihrer Anwendung zu beachten ist, soll in der Arbeitshilfe ebenfalls erläutert werden.

Vorüberlegungen

Grundsätzlich gilt, dass es nicht „*die*“ Erhebungsmethode gibt. Vielmehr bestimmen der Forschungsgegenstand sowie die jeweilige Phase innerhalb des Forschungsprozesses bzw. der bereits erlangte Erkenntnisgrad die Wahl der Forschungsmethode.[56]

Im Vorfeld der Datenerhebung sind drei zentrale Fragen zu klären, vgl. Kruse (2010, Oktober: 59 ff.):

- Von *wem* sollen Daten erhoben werden? (vgl. Arbeitshilfe Stichprobenauswahl)
- *Wie* sollen die Daten erhoben werden?
- *Wie* sollen die erhobenen Daten dokumentiert werden?

Qualitative Daten können beispielsweise mittels Interviews (siehe Helfferich 2005), Gruppendiskussionsverfahren (siehe Loos/Schäffer 2005), Beobachtungen (siehe Leggewie 1995) oder durch visuelle Erhebungen (siehe Raab 2008) generiert werden. Damit diese Daten jedoch auch ausgewertet werden können, müssen sie in angemessener Art und Weise dokumentiert sein. Möglich ist dies bspw. mittels Tonbandaufzeichnung und/oder Transkription (siehe hierzu auch die Online-Plattform: www.audiotranskription.de), Beobachtungsprotokoll, Fotos, schriftlich oder per Tonband/Video dokumentierte Diskussionsrunde etc.

Im Folgenden werden die gängigsten qualitativen Erhebungsmethoden vorgestellt, wobei der Schwerpunkt auf Interviewverfahren liegt, da diese im Rahmen der Evaluation von Ausbildungsordnungen am häufigsten eingesetzt werden. Im zweiten Teil dieser Arbeitshilfe werden noch zwei weitere Methoden erläutert, die Gruppendiskussion sowie die Beobachtung, da auch sie in manchen Evaluationen sinnvoll eingesetzt werden können.

Interviews

Ein Interview kann sowohl face-to-face, per Telefon oder per Internet erfolgen. Häufig hängt das gewählte Medium von den zur Verfügung stehenden finanziellen Mitteln sowie der zur Verfügung stehenden Zeit aller Beteiligter ab. Zu empfehlen ist in der Regel das face-to-face-Interview in einem Umfeld, das dem oder der Befragten vertraut ist. Findet das Interview bspw. im Büro des oder der Befragten selbst statt, so kann ggf. direkt auf benötigte Unterlagen zugegriffen werden. Andererseits sollte auch darauf geachtet werden, dass das Interview *möglichst* an einem ungestörten Ort durchgeführt wird – gerade auch in Hinblick auf die Aufnahmequalität (die Betriebskantine zur Mittagszeit wäre bspw. ein denkbar ungünstiger Ort). Aber auch das Büro

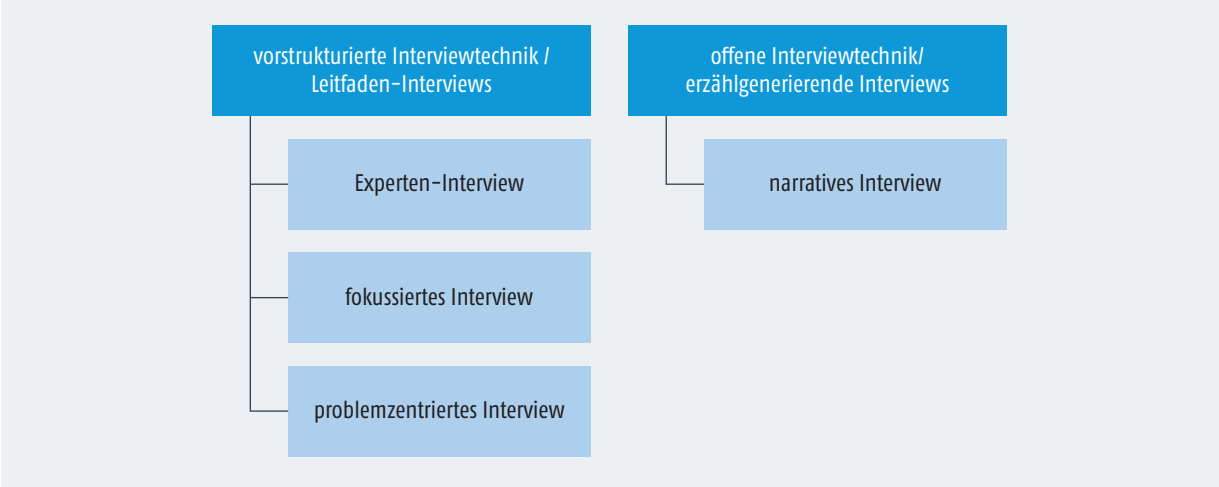
[56] Zur *Gegenstandsangemessenheit* der Methodenwahl siehe auch Lamnek (1995a); des Weiteren Helfferich (2005: 19 ff.); Kruse (2010, Oktober: 53 ff.).

kann als Gesprächsraum ungünstig sein, wenn bspw. durch Telefonanrufe etc. das Interview unterbrochen wird. Bei der Ortswahl gilt allerdings: Vorschlagsrecht hat der Proband oder die Probandin.

Grundsätzlich stellt ein Interview eine komplexe kommunikative Situation dar. Es lässt sich als asymmetrische Kommunikationsform beschreiben, die aber sowohl von dem/der Interviewten als auch von dem Interviewer oder der Interviewerin gemeinsam hergestellt und unterhalten wird, da beide nicht umhin kommen in der Situation herauszufinden, was die jeweils andere Person eigentlich will, was seine oder ihre tatsächlichen Interessen sind, wie er oder sie die Situation sieht, wie er oder sie das Gegenüber einschätzt usw., vgl. Helfferich (2005: 19 ff.); Kruse (2010, Oktober: 95 ff.).

In Abgrenzung zu standardisierten Befragungstechniken in der quantitativen Forschung besteht das Grundprinzip nichtstandardisierter, d. h. qualitativer Interviewführung darin, *so wenig direktiv wie möglich* zu verfahren, d. h. dem oder der Interviewten den weiten Raum zu geben, die eigenen Relevanzen entwickeln und formulieren zu können, vgl. Bohnsack (2000: 20 ff.). Der Grad der Direktivität hängt jedoch vom jeweiligen Forschungsinteresse ab, vgl. Honer (2006: 97) sowie Helfferich (2005: 100 ff.).

Nichtstandardisierte Interviews lassen sich somit nach dem Grad der Strukturierung unterscheiden:



Bei vorstrukturierten Interviews sollte grundsätzlich darauf geachtet werden, dass die vorgegebene Struktur v. a. das Relevanzsystem des Interviewers bzw. der Interviewerin abbildet und dieses explizit dokumentiert wird. Was der oder die Befragte jenseits dieses Relevanzsystems sagt, sollte vom Interviewer bzw. von der Interviewerin ‚bis auf Weiteres' als wichtig angesehen und behandelt werden. Ob speziell diese Aussagen für den Forschungsgegenstand von Bedeutung sind, sollte nicht im Interview selbst entschieden werden, sondern erst in der Auswertung der Daten nach dem Interview, vgl. Kruse (2010, Oktober: 130 ff.).

Leitfadeninterviews

Leitfadeninterview ist ein Oberbegriff für verschiedene Interviewverfahren, die mittels eines Gesprächsleitfadens realisiert werden: Durch den Einsatz eines Interviewleitfadens wird das Interview teilstrukturiert. Der Leitfaden ist dabei der von dem Forscher oder der Forscherin vorbereitete, leitende „rote Faden", der sich nach seinen oder ihren Vorstellungen durch das gesamte Interview zieht. Er besteht aus Fragen, die einerseits sicherstellen, dass bestimmte Themenbereiche angesprochen werden, die andererseits aber so offen formuliert sind, dass narrative Poten-

ziale des Befragten dadurch genutzt werden können. Der Leitfaden sollte aus diesem Grunde nicht zu umfangreich sein, vgl. Helfferich (2005: 158 ff.).

Im Vergleich zum narrativen Interview (s. u.) kann beim Leitfadeninterview sehr viel leichter sichergestellt werden, dass die interessierenden Aspekte auch alle angesprochen werden. Insofern wird eine Vergleichbarkeit mit anderen Interviews besser und direkter möglich, wenn diesen der gleiche Leitfaden zugrunde lag. Damit ist ein Leitfadeninterview immer dann das Mittel der Wahl, wenn konkrete Aussagen zu einem spezifischen Themenfeld erhoben und diese mit anderen Interviews verglichen werden sollen. Damit stehen jedoch alle Leitfadeninterviews im Spannungsfeld von Strukturierung versus Offenheit, vgl. Kurse (2010, Oktober: 65 ff.).

► Kommunikation im Leitfadeninterview

Kennzeichnend für Leitfadeninterviews sind trotz der grundsätzlichen Strukturierung der Interviewkommunikation die *offen formulierten Fragen*, auf die der oder die Befragte in selbst gewählten Formulierungen frei antworten soll: Es obliegt dabei den Befragten selbst, zu bestimmen, wie umfangreich und detailliert er oder sie auf die jeweilige Fragestellung eingeht. Die Fragen selbst sollten von der Interviewerin oder dem Interviewer so formuliert sein, dass die oder der Befragte zu einer detaillierten Schilderung des Sachverhalts in eigenen Worten animiert wird. Die Aufgabe des Interviewers bzw. der Interviewerin besteht dann darin, während der Erzählung, auf Themen oder wichtige Aspekte zu achten, die von den Befragten eingebracht werden, denn die Interviewerin oder der Interviewer hat hier die Möglichkeit, unabhängig vom Leitfaden, spontan und vertiefend nachzufragen. Somit gibt der Leitfaden den Ablauf der „Frage-Antwort-Sequenz“ nicht vor, sondern eröffnet vielmehr je nach Gesprächsverlauf notwendige Spielräume in der Abfolge der Fragen, für die Fragenformulierung oder auch für Verständnisfragen, vgl. Krieger (2010: 1) sowie Helfferich (2005:100 ff.). Dies fordert von den Interviewenden ein hohes Maß an intellektueller und kommunikativer Kompetenz und verlangt zudem hohe Sensibilität, Reflexivität, Überblick und eine permanente Vermittlung zwischen dem Interviewverlauf und dem Leitfaden, vgl. Helfferich (2005: 138 ff.). Unterstützt wird sie oder er dabei vom Leitfaden in der Funktion einer „Checkliste“, vgl. Helfferich (2005: 158 ff.). Der Leitfaden hilft sicherzustellen, dass alle Fragen vollständig und hinreichend spezifisch in der Befragung behandelt werden. Außerdem ermöglicht er die Vergleichbarkeit der einzelnen Interviews. Es sollte jedoch darauf geachtet werden, dass es zu keiner „Leitfadenbürokratie“ kommt, vgl. Kruse (2010, Oktober: 64 ff.), also zu keiner zu starken Orientierung am Leitfaden – im Sinne eines *Abfragens*, und nicht mehr im Sinne eines *Erfragens* – was den eigentlichen Gesprächsablauf ignoriert.

Narratives Interview

Das narrative Interview geht maßgeblich auf die Forschungsarbeiten von Fritz Schütze in der Biografieforschung zurück. Im Gegensatz zu den vorstrukturierten Interviewtechniken ist das narrative Interview durch eine offene Interviewtechnik gekennzeichnet, die klar zum Ziel hat, im Sinne eines erzählgenerierenden Interviews, die Befragten möglichst ungestört *erzählen* zu lassen. Die Interviewenden verzichten u. a. bewusst auf eine Strukturierung bspw. in Form eines Interviewleitfadens. Dadurch erhalten die Befragten den Raum, ihre Relevanzsysteme, Deutungsmuster und Sichtweisen zu verbalisieren, vgl. Kruse (2010, Oktober: 53). Durch das monologische „erzählen-lassen“ gelingt es den Befragten eher, sich in eine zurückliegende Handlungs- oder Erlebnissituation zurückzuversetzen. Dadurch können Erlebtes oder auch Erfahrungen der Befragten mit bestimmten Ereignissen aus der Vergangenheit und deren Verarbeitung wieder sichtbar werden, die für Forschende von großem Interesse sein können.

► Kommunikation im narrativen Interview

Narrative Interviews können nur „face to face“ durchgeführt werden, da es in der Regel nur in diesem Rahmen möglich ist eine Atmosphäre zu erzeugen, welche die Befragten zum Erzählen veranlasst. Dies setzt die Kompetenz der Befragenden voraus, ihre Rolle als aufmerksame, interessierte, insbesondere aber als stumme Zuhörer ernst zu nehmen und das Interview nicht durch Zwischenfragen zu unterbrechen; vgl. Friebertshäuser/Prengel (1997: 449). Weiterhin ist es wichtig, dass die Befragten davon ausgehen können, dass die Inhalte der Darstellung den Befragenden noch nicht bekannt sind. Außerdem sollte darauf geachtet werden, dass sich das Thema für eine narrative Darstellung eignet und hinreichend eingegrenzt ist.

Die Interviewtechnik sieht vor, die Befragten mit einer eindeutig narrativen Ausgangsfrage dazu zu animieren, möglichst viel zu erzählen. Diese so zustande kommende *Stegreiferzählung* enthält optimaler Weise kaum Argumentationen sondern fast ausschließlich Erzählungen. Der Ablauf eines narrativen Interviews gliedert sich in der Regel in folgenden vier Phasen:

1) Erklärungsphase: Die narrative Ausgangsfrage wird gestellt und die oder der Befragte wird dazu ermutigt mit der Stegreiferzählung zu beginnen.

2) Erzählphase: In dieser Phase sollte die Interviewerin oder der Interviewer der oder dem Befragten Aufmerksamkeit signalisieren, gleichzeitig aber strikt zurückhaltend agieren und die Erzählungen auf keinen Fall durch Nachfragen unterbrechen.

3) Nachfragephase: Erst wenn die Erzählphase definitiv abgeschlossen ist, sollte in die Nachfragephase gewechselt werden, um unklar gebliebene Fragen, nicht nachvollziehbare Erzählstellen oder auch Widersprüchlichkeiten zu klären. Die Nachfragen sollten dabei erzählgenerierend angelegt sein, also die Befragten zu genaueren Ausführungen anregen.

4) Bilanzierungsphase: Zum Abschluss sollte das Interview bilanziert werden. Dabei geht es hauptsächlich um Nachfragen zu den *Eigentheorien* der oder des Befragten, die im Verlaufe des Interviews aufgestellt wurden. Zudem können Unklarheiten geklärt werden.

Grundsätzlich steht und fällt das narrative Interview jedoch mit der Kompetenz und dem Willen der Befragten, offen von sich zu erzählen; dabei kann die monologische Erzählsituation durch das geänderte Rollenverhalten zudem zur Verunsicherung auf beiden Seiten beitragen. Bei Interviews mit Personen aus anderen Kulturkreisen muss beachtet werden, dass sich diese Methode unter Umständen problematisch gestaltet, da hier zum Teil auch andere Erzählmuster vorliegen. Nicht zuletzt ist die Auswertung der Interviews teilweise sehr intensiv und zeitaufwendig, da die Erzählphase von wenigen Minuten bis hin zu mehreren Stunden dauern kann, vgl. Kruse (2010, Oktober: 54).

Das Experten-Interview

Das Hauptmotiv für die Durchführung eines Experten-Interviews stellt das sachliche Interesse an einem bestimmten Forschungsgegenstand dar. Die Experten und Expertinnen stellen fundiertes Wissen als „sachorientierte Gutachter“ zur Verfügung, welches im Rahmen des Interviews zudem konstruktiv erläutert werden soll, vgl. Miek (2005: 13) sowie Meuser/Nagel (1994). Solche Interviews sind im Grunde genommen *Gespräche*.

Diese Interviewform funktioniert nur dann, wenn die Befragten im Gegenüber auch fachlich kompetente Gesprächspartner/-innen erkennen, die Interviewenden also über hohe Ko-Expertise verfügen. Kenntnisse gängiger Fachausdrücke sowie von Grundlagen des jeweiligen Fachgebiets sind daher Voraussetzung für ein gelungenes Interview bzw. Gespräch.

► Wer aber ist eigentlich Expertin/Experte?

Expertin oder Experte ist, wer dieses Mandat bekommt, also von anderen Menschen als solcher oder solche benannt wird, vgl. Bogner/Littig/Menz (2005: 39 ff.). Dazu zählen insbesondere Personen, die aufgrund ihrer langjährigen Erfahrung bzw. Tätigkeit über bereichsspezifisches Können und Wissen verfügen. Hierzu zählen z. B. umfangreiches Erfahrungs- bzw. Entwicklungswissen oder spezielles Betriebs-, Prozess- oder Kontextwissen, welches von den Experten und Expertinnen durch ihre Stellung in einer Institution, einer Organisation oder in einem Betrieb erworben wurde.

Obwohl es sich beim Expertenwissen um ein spezialisiertes Sonderwissen handelt, ist es den betroffenen Personen nicht notwendigerweise im Modus des „diskursiven Bewusstseins" direkt verfügbar. Diskursiv verfügbar bzw. klar und deutlich präsent sind erinnerte Entscheidungsverläufe und offizielle Entscheidungskriterien, nicht aber die fundierende Logik des Entscheidens und der Routinen des Expertenhandelns, vgl. Bogner/Littig/Menz (2005: 33 ff.) sowie Meuser/Nagel (2006: 58).

Das Ziel des Interviews ist in erster Linie, komplexes Fach- und Erfahrungswissen zu erfragen oder zu rekonstruieren. In zweiter Linie können auch die Abfrage von Meinungen, Einschätzungen, Alltagstheorien oder Stellungnahmen im Fokus des Interviews stehen. Dabei werden die Befragten als Repräsentanten einer bestimmten Gruppe gesehen. Im klassischen Experten-Interview treten die Befragten als biografische Personen in den Hintergrund, stattdessen interessieren die in einen Funktionskontext eingebundenen Akteure. Vgl. Meuser/Nagel (2006: 57).

Exkurs:

Allerdings wird diese Aufteilung von Rolle und Person auch kritisiert, sodass Bogner und Menz mit ihrem theoriegenerierenden Experten-Interview (siehe in Bogner/Littig/Menz 2005: 33 ff.) ein komplexeres Verfahren vorstellen, dass die Generierung von Experten-Wissen gerade auch vor dem Hintergrund einer komplexen Rollensozialisation zu rekonstruieren vermag.

► Kommunikation im Experten-Interview

Ein Interviewleitfaden stellt das Grundgerüst eines jeden Experten-Interviews dar, wenngleich er auch flexibel und nicht im Sinne eines standardisierten Ablaufschemas gehandhabt wird. Der Leitfaden unterstützt zudem die Interviewenden darin sicherzustellen, dass alle Fragen vollständig und hinreichend spezifisch in der Befragung behandelt werden.

Sinnvoll ist es, solche Interviews mit einer offenen Frage einzuleiten, welche den Befragten die Möglichkeit einräumt, ihre Erläuterungen zu einem bestimmten Sachverhalt eigenstrukturiert darzustellen, vgl. Kruse (2010, Oktober: 57).

Im weiteren Dialog mit den Experten und Expertinnen werden stark themenfokussierte Fragen gestellt, die in der Terminologie der Experten und Expertinnen beantwortet werden sollten, vgl. Miek (2005: 4). Dabei ist die Rolle der Kommunikanten wechselseitig aktiv, der Interviewer oder die Interviewerin gibt allerdings meistens die konkrete Gesprächsstruktur vor.

Die Herausforderung für die Interviewenden liegt darin, ein gelungenes Maß zwischen strukturiertem und offenem Interview zu wählen. Die Leitfragen sollen die Experten dazu anregen, ihr fundiertes Wissen preiszugeben. Dabei sollen die Experten und Expertinnen aber nicht zu sehr in eine durch die Interviewenden vordefinierte Richtung getrieben werden.

Die oben beschriebene flexible Handhabung eines Interviewleitfadens ist daher äußerst wichtig. Für den Fall, dass der gewünschte Detaillierungsgrad durch die bisherigen Aussagen nicht erreicht worden ist, sollten die jeweiligen Experten gebeten werden, auf bestimmte Aspekte noch näher einzugehen.

Fokussiertes Interview[57]

Beim fokussierten Interview findet eine klare Fokussierung auf einen vorab bestimmten Gesprächsgegenstand bzw. Gesprächsanreiz statt. Diese Form des Interviews wurde in den 1940er-Jahren maßgeblich durch Robert Merton und Patricia Kendall entwickelt, die dem Feld der Kommunikations- und Medienforschung (insbesondere auch der Propagandaanalyse) zugeordnet werden können. Ihr Interesse galt insbesondere den Wirkungen von medialen Kommunikationsprozessen und Mediendokumentationen, vgl. Kruse (2010, Oktober: 56).

Aufgabe der Evaluierenden ist es, im Vorfeld des Interviews den Gesprächsgegenstand bzw. Gesprächsanreiz genau zu analysieren. Dabei bestimmen sie die jeweils relevanten, objektiven Bestandteile des Gegenstands und die möglichen subjektiven Interpretationen der interviewten Personen, um diese später miteinander vergleichen zu können, vgl. Flick (2005: 118).

► Kommunikation im fokussierten Interview

Die Kommunikation im fokussierten Interview ist geprägt durch offen formulierte Fragen und eine nicht direktive (nicht-beeinflussende) zurückhaltende Gesprächsführung. Der Interviewleitfaden ist ähnlich dem Experten-Interview flexibel zu handhaben. Die Interviewführung ist insgesamt geprägt durch maximale Offenheit bei der Fragestellung und minimale Lenkung wie dies z. B. erforderlich ist bei der Überleitung auf ein anderes Thema. Merton/Kendall haben vier Qualitätskriterien herausgearbeitet, die bei fokussierten Interviews besondere Beachtung finden sollten, vgl. Hopf (2005: 354):

Reichweite: Das Spektrum der im Interview angeschnittenen Problemstellungen darf nicht zu eng sein. Das heißt, die Befragten müssen eine maximale Chance haben, auf die gewählte Stimulus-Situation (z. B. ein Film, Bild, Text etc.) zu reagieren.

Spezifität: Die im Interview aufgeworfenen Themen sollen in spezifizierter Form abgehandelt werden. Das heißt, die Befragten sollen konkret werden in ihren Äußerungen zum Gegenstand.

Tiefe: Im Interview soll die Tiefendimension angemessen repräsentiert werden. Das heißt, die Befragten sollen bei der Darstellung der affektiven, kognitiven und wertebezogenen Bedeutung, die bestimmte Situationen für sie haben, unterstützt werden.

Personaler Kontext: Der persönliche Kontext, in dem die analysierten Deutungen und Reaktionen stehen, muss in ausreichendem Maße erfasst werden.

Problemzentriertes Interview

Das Problemzentrierte Interview (PZI) wurde in den 1980er-Jahren v. a. durch Andreas Witzel bekannt und wird in der Regel in Kombination von mehreren qualitativen Datenerhebungs- und Auswertungsmethoden (Leitfadeninterview, Fallanalyse, Biografische Methode, Gruppendiskussion und Inhaltsanalyse) eingesetzt, vgl. Witzel (2000: Absatz 4) sowie Flick (2005: 135). So dient bspw. ein Kurzfragebogen zur Ermittlung von demografischen Daten (Alter, Beruf etc.) und entlastet den Erzählstrang der Befragten von Unterbrechungen. Auch können die Informationen aus dem Kurzfragebogen den Gesprächseinstieg in das sich anschließende leitfadengestützte Interview erleichtern. Der Leitfaden hat für die Interviewenden eher die Funktion einer Gedächtnisstütze bzw. eines Orientierungsrahmens. Durch diesen soll gewährleistet werden, dass im Rahmen der Vergleichbarkeit aller Interviews die gleichen inhaltlichen Aspekte angesprochen werden. Die Reihenfolge der Fragen richtet sich dabei nach dem Gesprächsverlauf und ist somit nicht starr vorgegeben.

[57] Das von Niethammer (2005: 595 ff.) als Fachinterview bezeichnete Interview, lässt sich dem fokussierten Interview zuordnen.

► Kommunikation im problemzentrierten Interview

Beim PZI handelt es sich um eine offene, teilstrukturierte Befragung. „Offen" bezieht sich dabei auf die Möglichkeit der Befragten, sich frei zu äußern und das wiederzugeben, was ihnen besonders relevant erscheint. „Teilstrukturiert" bezieht sich auf die Vorgehensweise der Befragung, welche sich auf einen Interviewleitfaden stützt, mit dem immer wiederkehrend stark strukturierende Vorgaben gesetzt werden. Von diesem Leitfaden darf im Gespräch situativ abgewichen werden, vgl. Witzel (2000). Witzel sieht nämlich gerade das systematische Wechselspiel von Induktivität (Offenheit) und Deduktivität (Strukturierungen) als zentrales Merkmal des PZI an. Das Erzählprinzip steht somit nur zum Teil im Vordergrund, denn der Interviewer oder die Interviewerin lenkt das Gespräch immer wieder auf den Untersuchungsgegenstand bzw. die vorliegende Problemstellung zurück.

► Zentrale Grundpositionen des problemzentrierten Interviews

Die drei zentralen Grundpositionen des PZI sind gemäß Witzel (2000: Absatz 4) die Folgenden:

Problemzentrierung: Die Problemstellung wird von den Forschenden gesetzt und steht im Fokus. Die Befragenden verfügen über Vorwissen zum Problembereich und nutzen dieses, um die Ausführungen der Befragten besser nachvollziehen und interpretieren zu können. Anders als beim narrativen Interview können die Evaluatoren hierzu auch durch gezieltes Nachfragen die subjektive Sichtweise der Befragten vertiefend erfassen.

Gegenstandsorientierung: Diese betont die Flexibilität des PZI. Dahinter steht der Anspruch, dass sich die Methode am Gegenstand orientiert und nicht umgekehrt. So besteht u. a. die Möglichkeit, die Gesprächstechniken flexibel einzusetzen, d. h., den Befragenden obliegt es, je nach sprachlicher Ausdrucks- und Reflexionsfähigkeit der Befragten, stärker erzählgenerierende oder dialogische Techniken einzusetzen oder z. B. eine Diskussion oder eher eine Narration zuzulassen.

Prozessorientierung: Diese erlaubt ein gezieltes Nachfragen im Gesprächsverlauf. Das Interview wird prozesshaft auf die subjektive Sichtweise der Befragten zugespitzt.

Gruppendiskussion

Die Gruppendiskussion stellt eine eigenständige qualitative Methode der rekonstruktiven Sozialforschung dar, vgl. Kruse, (2010, Oktober: 58). Allgemein wird die Gruppendiskussion nach Lamnek definiert als ein Gespräch einer Gruppe von Untersuchungspersonen zu einem bestimmten Thema unter Laborbedingungen, vgl. Lamnek (1995: 131).

Das Ziel einer Gruppendiskussion ist es, interne Interaktions-, Diskurs- und Gruppenprozesse, die zur Bildung von *kollektiven Orientierungsmustern* und sozialen, konstruierten Bedeutungsmustern beitragen, zu einem bestimmten Thema zu erheben, vgl. Kruse (2010, Oktober: 276 ff.). Um dies erreichen zu können, steht seitens der Moderation die *Initiierung eines selbstläufigen Diskurses* in der Gruppe im Mittelpunkt, vgl. Bohnsack (2000: 123 ff.) sowie Loos/Schäffer (2005). Dies setzt jedoch wiederum voraus, dass eine *Realgruppe,* vgl. Loos/Schäffer (2005) vorliegt. Das heißt, dass die diskutierende Gruppe als soziale Gruppe auch außerhalb der Forschungssituation bestehen muss, und nicht nur für diese künstlich zusammengestellt wurde.

Im Vergleich dazu gibt es auch die Form der Gruppendiskussion, bei der eine künstlich zusammengestellte Gruppe als Proband dient. Diese Tradition des Gruppendiskussionsverfahrens wird üblicherweise als *focus group* bezeichnet, vgl. Bohnsack (2000: 123 ff.) sowie Loos/Schäfer

(2005). Hier steht sehr viel weniger die Initiierung eines selbstläufigen Diskurses im Mittelpunkt als die Informationsgewinnung.[58]

Eine Gruppendiskussion kann sich sowohl auf verschiedene Themen beziehen (Familie, Arbeit, Freizeit) als auch auf Meinungen, Einstellungen, auf bestimmte Verhaltensweisen, Normen, Werte einzelner Gruppenmitglieder oder der ganzen Gruppe. Die Methode sollte besser nicht verwendet werden, wenn die Forschung zum Ziel hat, mehr über Handlungspraxen, subjektive Intentionen („was war die Motivation?“) oder Biografien herausfinden zu wollen.

Die Größe der Gruppe sollte zwischen fünf (Gruppendiskussion) und zwölf Teilnehmenden (focus group) liegen. Die Gruppenzusammensetzung sollte je nach der spezifischen Gruppendiskussionsmethode und dem Forschungsgegenstand gewählt werden. Unterschieden werden hier:

- homogene und heterogene Gruppen,
- natürliche Gruppe und Realgruppe (z. B. Auszubildende eines Betriebs),
- künstliche Gruppe (z. B. Auszubildende unterschiedlicher Betriebe).

Voraussetzung für eine gute Gruppendiskussion ist jedoch grundsätzlich, dass alle Teilnehmenden ähnlich stark vom Thema der Fragestellung betroffen sind.

► Kommunikation mit der Gruppendiskussion

Der idealtypische Diskussionsverlauf in der Gruppendiskussion teilt sich in vier Phasen.

Eröffnungsphase:

Zunächst stellt sich die Diskussionsleitung vor, benennt den Diskussionsgegenstand und gibt erste Informationen zum Projekt. Dabei gilt die Devise „weniger ist mehr“, um mögliche Richtungsangaben für die sich anschließende Diskussion zu vermeiden, vgl. Kruse (2010, Oktober: 285 ff.). Sofern sich die Gruppe untereinander nicht kennt (im Falle von focus groups), ist jetzt Zeit für eine kurze Vorstellungsrunde der Teilnehmenden.

Einstiegsphase:

Es wird ein Grundreiz (Diskussionsstimulus) in Form einer offenen oder geschlossenen Frage, eines Statements oder eines kurzen Films in die Gruppe hineingegeben. Wichtig dabei ist, dass die Moderatorin oder der Moderator immer die ganze Gruppe anspricht und die direkte Aufforderung an eine einzelne Person vermeidet. Durch eine offene Eröffnungsfrage zu Beginn der Gruppendiskussion sollen die Teilnehmenden animiert werden, eigene, für sie persönlich relevante Erfahrungen bzw. Beschreibungen einzubringen, mit dem Ziel, einen *selbstlaufenden Diskurs* ins Rollen zu bringen.

Erhöhte Aktionsphase:

Läuft die Diskussion, so sollte die Haltung der Moderatorin oder des Moderators eher zurückhaltend sein, d. h. sie oder er sollte sich während der Diskussionsphase mit Kommentaren und Zwischenfragen zurückhalten, damit die Diskussion nicht stoppt oder beeinflusst wird. Nur wenn die Diskussion ins Stocken gerät, sollten dezente Interventionen durch die Moderatorin oder den Moderator erfolgen, um den Gesprächsfluss wieder herzustellen, vgl. Kruse (2010, Oktober: 286 ff.). Auch obliegt es der Moderation, die Diskussion durch die Einbringung weiterer gezielter Stimuli zu strukturieren. Dabei sollte dem Prinzip der Offenheit gefolgt und geschlossene, direkte oder auch suggestive Fragen vermieden werden, vgl. Kruse, (2010, Oktober: 287 ff.)

[58] Spöttl (2005: 611 ff.) wählt für seine Methode Experten-Facharbeiter-Workshop (EFW) beispielsweise eine focus group.

Auslaufphase:

Wenn sich die Diskussion dem Ende zuneigt, können durch gezielte Fragestellungen eventuell noch nicht erörterte Aspekte diskutiert werden. Auch bietet diese Phase die Gelegenheit für immanente Nachfragen, um Verständnisfragen zu stellen oder auf eventuelle Brüche in der Argumentation innerhalb der Gruppe einzugehen. Aus Gründen der Vergleichbarkeit und auch damit keine forschungsrelevanten Fragestellungen ausgelassen werden, ist es ratsam, im Vorfeld der Gruppendiskussion eine Liste anzufertigen, auf der alle Themen aufgeführt sind, die bearbeitet werden sollen.

▸ Vorteile/Nachteile der Gruppendiskussion

Unter ökonomischen Gesichtspunkten wird der Aufwand an Zeit, Personal und Geld erheblich reduziert, indem eine Gruppe von Menschen gleichzeitig befragt wird. Nicht zu unterschätzen ist dabei allerdings der womöglich hohe Aufwand bei der Vereinbarung eines Termins, an dem auch tatsächlich alle Mitglieder einer Gruppe teilnehmen können, wie auch der Aufwand für die Durchführung, Transkription und Auswertung der Diskussion.

Der Gruppendynamik und der Diskussion unter den Teilnehmenden wird bei der Gruppendiskussion – anders als bei der focus group – eine hohe Bedeutung beigemessen: Blumer (1969/1973: 123) zitiert nach Flick (2005: 170)verdeutlicht „ (…) solch eine Gruppe, die gemeinsam ihren Lebensbereich diskutiert und ihn intensiv prüft, wenn ihre Mitglieder sich widersprechen, wird mehr dazu beitragen, die den Lebensbereich verdeckenden Schleier zu lüften, als jedes andere Forschungsmittel, das ich kenne“. Es finden Positionierungen statt, die in dieser Form in Einzelinterviews nicht zwingend auftreten würden.

Der Vorteil dieser Gruppendynamik kann aber gleichzeitig auch ein Nachteil dieser Methode sein. Denn die Dynamik, die durch die einzelnen Gruppenmitglieder initiiert wird, verhindert unter Umständen gleichzeitig die Meinungsäußerungen bestimmter Personen und kann zu einer Verzerrung des Gesamtbildes durch „Schweiger“ und „Vielredner“ führen. Wie sich letztendlich die Diskussion entwickelt und welche Wendungen sie nimmt, ist kaum vorhersagbar und stellt besondere Anforderungen an die Moderation. Moderierende können Entscheidungen über den Einsatz weiterer Stimuli folglich nur aus der Situation heraus treffen, vgl. Flick (2005: 177).

Ein weiterer Vorteil dieser Methode ist ihre „Ideenproduktion“, die dann entsteht, wenn sich die Gruppenmitglieder gegenseitig zu neuen Ideen anregen. Nicht zuletzt entspricht der Prozess der Meinungsbildung, der unter dem sozialen Einfluss der Gruppe innerhalb der Diskussion entsteht, am ehesten dem Prozess der Meinungsbildung im wahren Leben.

Beobachtung

Nach Leggewie (1995: 193) ist die Beobachtung bzw. die „qualitative Feldforschung (…) immer dann eine Methode der Wahl, wenn sozial-räumlich überschaubare Einheiten menschlichen Zusammenlebens (…) ganzheitlich erfasst werden sollen.“ Wenn also der situative Handlungskontext erschlossen werden soll. Die Beobachtung ist darüber hinaus aber auch immer dann Methode der Wahl, wenn verbale Daten nicht ausreichen, wenn Selbstinszenierungen hinterfragt werden sollen, wenn eine Befragungs- oder auch Laborsituation die Datenerhebung negativ beeinträchtigen würde und/oder wenn *eigene* Erlebnisdaten gefragt sind, vgl. Stegmaier (2009).

Die teilnehmende Beobachtung, als die am häufigsten verwendete Methode, ist dabei eine Feldstrategie, die gleichzeitig Dokumentenanalyse, Interviews, direkte Teilnahme und Beobachtung sowie Introspektion kombiniert, vgl. Flick (2005: 206). Im Rahmen der Evaluation von Ausbildungsordnungen kann diese Methode beispielsweise bei Betriebsbesichtigungen oder der Teilnahme an praktischen Prüfungen zum Einsatz kommen.

Beobachten ist grundsätzlich ein Phänomen des alltäglichen Handelns, beispielsweise wenn aus dem Fenster heraus spielende Kinder beobachtet werden. Die wissenschaftliche Beobachtung unterscheidet sich jedoch erheblich von der alltäglichen Beobachtung. Bei der wissenschaftlichen Beobachtung wird versucht, klassische Beobachtungsfehler zu vermeiden bzw. zu kontrollieren. So wird die Wahrnehmung der Kinder erst dann zu einer wissenschaftlichen Beobachtung, wenn sie drei Kriterien folgt, vgl. Schöne (2003):

- Absicht : der Forscher oder die Forscherin verfolgt ein bestimmtes Ziel
- Selektion: es erfolgt eine Auswahl bestimmter Aspekte der Wahrnehmung
- Auswertung: das Beobachtete wird beschrieben und interpretiert.

„Teilnehmende Beobachtung will ihren Untersuchungsgegenstand von innen heraus verstehen. Mit ihrer Hilfe können subjektive Sichtweisen, die Abläufe sozialer Prozesse oder die kulturellen und sozialen Regeln, die diese Prozesse prägen, verstanden werden. Das der teilnehmenden Beobachtung zugrunde liegende Erkenntnisprinzip heißt *Verstehen*. Ausgangspunkt teilnehmender Beobachtung ist in der Regel der Einzelfall, von dem aus zu allgemeinen oder vergleichenden Aussagen geschritten wird. Erst wird der einzelne Fall rekonstruiert, dann werden die Analysen und Ergebnisse anderer Fälle zum Vergleich herangezogen und schließlich wird daraus eine Typologie entwickelt. Was als Einzelfall verstanden wird, ist vom theoretischen Standpunkt abhängig, von dem aus der Fall untersucht wird: Es können Subjekte und ihre Sichtweisen sein, Interaktionen oder soziale und kulturelle Kontexte“ Schöne (2003: 6).

Im Falle der Beobachtung ist ein spiralförmiger Forschungsprozess dringend zu empfehlen. Dabei folgt auf einen ersten beobachteten Fall zunächst eine Auswertung, ehe wieder ins Feld zurückgegangen wird und die nächste Beobachtung erfolgt. Der Ablauf von Beobachtungen kann daher, in Anlehnung an Schröer (1997), Spradley (1980) und Stegmaier (2009), grob wie folgt skizziert werden:

1. Klärung des Feldzugangs („Türöffner“ finden, Erlaubnis einholen),
2. Auswahl des Settings sowie des Beobachtungsgegenstands,
3. *Exploratives* Beobachten (offen wahrnehmen, viel „einsammeln“, viel fragen, Memos direkt nach (auf keinen Fall während!) der Beobachtung verfassen, dabei Beobachtungen klar von Vermutungen/Interpretationen trennen),
4. Auswertung des ersten Beobachtungsfalls im Team! Offene Fragen notieren,
5. Erneuter Gang ins Feld, die Beobachtung erfolgt zunehmend *fokussierter,*
6. Auswertung der neuen Daten und Vergleich des ersten Falls mit dem zweiten Fall,
7. Erneuter Gang ins Feld; die Beobachtung erfolgt zunehmend *selektiver,*
8. Auswertung der neuen Daten und Vergleich mit den vorherigen Fällen,
9. Datensättigung erreicht; Ausstieg aus dem Feld.

Grundsätzlich lassen sich wissenschaftliche Beobachtungsverfahren nach fünf Dimensionen klassifizieren, vgl. Flick (2005: 200):

Verdeckte vs. offene Beobachtung:

Inwieweit wissen die Beobachteten, dass sie beobachtet werden und inwieweit soll die Rolle der Beobachter und die Forschungsfrage aufgedeckt werden?

Nicht teilnehmende vs. teilnehmende Beobachtung:

Inwieweit werden Beobachtende zum Teil des aktiv untersuchten Feldes, d. h. nehmen selbst am Feldgeschehen teil?

Systematische vs. unsystematische Beobachtung:

Wird ein mehr oder minder standardisiertes Beobachtungsschema verwendet oder wird das Feld eher ‚offen‘ beobachtet?

Beobachtung natürlicher vs. künstlicher Situationen:

Wird im interessierenden Feld beobachtet, d. h. ist der Beobachtungsgegenstand in komplexe, soziale Situationen eingebunden, oder findet die Interaktion isoliert in einem speziellen Raum statt, der eine bessere Beobachtbarkeit ermöglicht?

Selbst- vs. Fremdbeobachtung:

Meist werden andere Menschen beobachtet. Welcher Stellenwert wird dabei der reflektierenden Selbstbeobachtung der Forschenden beigemessen?

Die teilnehmende Beobachtung selbst erfolgt durch das Eintauchen des Forschers bzw. der Forscherin in das untersuchte Feld. Er oder sie nimmt am Alltags- oder auch am Arbeitsleben der im Fokus stehenden Personen oder Gruppen (offen oder verdeckt) teil. Dies eröffnet die Möglichkeit, die Konstitution der sozialen Wirklichkeit der Beobachteten aus der Außenperspektive zu analysieren, vgl. Flick (2005: 205).

Dabei sind die Beobachtungen der Forschenden die Grundlage für Interpretationen. Es erfordert eine große Sorgfalt bei der Interpretation der sozialen Bedeutung von Beobachtetem, um sicherzustellen, dass auch die vorhandenen Sinnstrukturen angemessen erfasst worden sind.

► Vorteile/Nachteile der teilnehmenden Beobachtung

Durch die teilnehmende Beobachtung entsteht meist ein umfassenderes Bild eines sozialen Phänomens (Inhalte, Zustände, Prozesse, Strukturen). Durch die persönliche Teilnahme nimmt das „Verstehen“ einer Situation zu.

Ein Nachteil ist der hohe Zeit- und Personalaufwand und die damit einhergehende Kostenintensität dieser „Feldstrategie“. Das besondere Problem der teilnehmenden Beobachtung ist jedoch der ständige Balanceakt zwischen Subjektivität und Intersubjektivität (relativierte Beobachtungsposition), unter Umständen schwindet dabei die notwendige Distanz zum Feld, was in der Ethnografie als das Problem des „going nativ“ bezeichnet wird, vgl. Girtler (2001).

Literatur

Blumer, Herbert: Der methodologische Standort des Symbolischen Interaktionismus. In: Arbeitsgruppe Bielefelder Soziologen (Hrsg.). Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit, Reinbek 1973, S. 80–146.

Bohnsack, Ralf: Rekonstruktive Sozialforschung. Opladen 2000.

Flick, Uwe: Handbuch Qualitative Sozialforschung. Grundlagen, Konzepte, Methoden und Anwendungen. Weinheim 1995a.

Flick, Uwe: Qualitative Sozialforschung. Eine Einführung. Reinbek 2005, S. 118–124.

Friebertshäuser, Barbara; Prengel, Annedore: Forschungsmethoden in der Erziehungswissenschaft. Weinheim 1997, S. 387 ff.

Friebertshäuser Barbara; Prengel, Annedore: Handbuch qualitativer Forschungsmethoden in der Erziehungswissenschaft. Weinheim,1997.

Girtler, Roland: Methoden der Feldforschung. Wien 2001.

Gold, Raymond: Roles in sociological field Observation. In: McCall, George J.; Somons, Jerry L. (Hrsg): Issues in Participant Observation. 1969, S. 30–39). Reading, Mass.: Addison-Wesley.

Krieger, Claus. (ohne Jahreszahl): Leitfaden-Interviews. In: Miethling, Wolf-Dietrich (Hrsg.): Qualitative Forschungsmethoden in der Sportpädagogik. Schorndorf 2008, S. 45–63.

Kruse, Jan: Einführung in die qualiative Interviewforschung. Freiburg 2010, S. 53–59.

Lamnek, Siegfried: Qualitative Sozialforschung. Band 1: Methodologie. Weinheim1995a.

Lamnek, Siegfried: Qualitative Sozialforschung. Band 2: Methoden und Techniken. Weinheim 1995b.
Lamnek, Siegfried: Gruppendiskussion. Theorie und Praxis. Weinheim 2005.
Leggewie, Heiner: Beobachtungsverfahren. In: Flick, Uwe u. a.: Handbuch Qualitative Sozialforschung. Grundlagen, Konzepte und Anwendungen. Weinheim 1995, S. 189–208.
Mayer, Horst: Interview und schriftliche Befragung. Entwicklung, Durchführung und Auswertung. Oldenburg 2002.
Mieg, Harald. A.: Experteninterviews in den Umwelt- und Planungswissenschaften. Eine Einführung und Anleitung. Zürich 2005.
Niethammer, Manuela: Fachinterview. In: Rauner, Felix (Hrsg.): Handbuch Berufsbildungsforschung. Bielefeld 2005.
Raab, Jürgen: Visuelle Wissenssoziologie. Theoretische Konzeption und materiale Analysen. Konstanz 2008.
Schöne, Helmar (2003): Die teilnehmende Beobachtung als Datenerhebungsmethode in der Politikwissenschaft. Methodologische Reflexion und Werkstattbericht. Forum Qualitative Sozialforschung [On-line Journal], 4(2). URL: www.qualitative-research.net/fqs-texte/2-03/2-03schoene-d.htm (Stand 01/2016).
Schröer, Norbert: Wissenssoziologische Hermeneutik. In: Hitzler, Ronald; Honer, Anne (Hrsg.): Sozialwissenschaftliche Hermeneutik. Opladen 1997.
Spöttl, Georg: Experten-Facharbeiter-Workshop. In: Rauner, Felix. (Hrsg.): Handbuch Berufsbildungsforschung. Bielefeld 2005.
Spradley, James P.: Participant Observation. New York 1980.

Literatur zu Erhebungsmethoden

Experteninterview:

Meuser, Michael; Nagel, Ulrike: Expertenwissen und Experteninterview. In: Hitzler, Ronald (Hrsg.): Expertenwissen. Die institutionalisierte Kompetenz zur Konstruktion von Wirklichkeit. Opladen 1994, S. 180–192.
Bogner, Alexander; Littig, Beate Menz, Wolfgang.(Hrsg.): Das Experteninterview. Theorie, Methode, Anwendung. Wiesbaden 2005.

Fokussiertes Interview:

Merton, Robert K.: The focused interview and focus groups. In: Public Opionion Quarterly, Vol. 51 (Issue 4). London 1987, S. 550–566.

Problemzentriertes Interview:

Witzel, Andreas (2000): Das problemzentrierte Interview. Forum Qualitative Sozialforschung [On-line Journal], 1(1), Art. 22. URL: http://nbn-resolving.de/urn:nbn:de:0114-fqs0001228. (Stand 01/2016).

Leitfadeninterview:

Helfferich, Cornelia: Qualität qualitativer Daten. Manual zur Durchführung qualitativer Einzelinterviews. Wiesbaden 2005.
Kruse, Jan: Einführung in die qualitative Interviewforschung. Freiburg 2015.

Narratives Interview:

Schütze, Fritz: Biografieforschung und narratives Interview. In: Neue Praxis, Heft 3/1983. Lahnstein 1983.
Glinka, Hans-Jürgen: Das narrative Interview. Weinheim 1998.
Küsters, Ivonne: Narrative Interviews. Grundlagen und Anwendung. Wiesbaden 2009.

Gruppendiskussionsverfahren:

Bohnsack, Ralf: Rekonstruktive Sozialforschung. Opladen 2000, S. 123–142.
Loos, Peter; Schäffer, Burkhard: Das Gruppendiskussionsverfahren. Wiesbaden 2005.
Lamnek, Siegfried: Gruppendiskussion. Theorie und Praxis. Weinheim 2005.

Beobachtung:

Schöne, Helmar (2003): Die teilnehmende Beobachtung als Datenerhebungsmethode in der Politikwissenschaft. Methodologische Reflexion und Werkstattbericht. Forum Qualitative Sozialforschung [On-line Journal], 4(2). URL: www.qualitative-research.net/index.php/fqs/article/view/720/1559 (Stand 01/2016).

Abstract

Die Evaluation von Ausbildungsordnungen ist nicht nur ein wichtiger Bestandteil der Qualitätssicherung beruflicher Bildung, sondern spielt auch bei der Modernisierung von Berufen eine bedeutende Rolle. Herangehensweisen, die bei der Evaluation von Programmen, Projekten, Prozessen oder Organisationen häufig gewählt werden, lassen sich nicht automatisch auf die Evaluation von Ausbildungsordnungen übertragen. Deshalb wurde ein berufsübergreifendes Konzept zur Evaluation von Ausbildungsordnungen entwickelt, das speziell an diesen Evaluationsgegenstand angepasst ist.

The evaluation of training regulations is not only an important part of the quality assurance in vocational education and training, but also plays an important role in the modernisation of occupations. Approaches, which are often chosen to evaluate programs, projects, processes or organisations, may not be suitable to evaluate training regulations. Therefore, an inter-professional concept, especially adjusted to this subject, has been developed.

Bundesinstitut für Berufsbildung
Robert-Schuman-Platz 3
53175 Bonn

Telefon: (0228) 107-0
Telefax: (0228) 107 2976/77

Internet: www.bibb.de
E-Mail: zentrale@bibb.de

Bundesinstitut für Berufsbildung
BIBB

▶ Forschen
▶ Beraten
▶ Zukunft gestalten