Die Gegenstandsangemessenheit empirischer Datenerhebungsmethoden im Kontext von Lehrevaluationen an Hochschulen

Inaugural-Dissertation

zur

Erlangung der Doktorwürde

des

Fachbereichs Erziehungswissenschaften

der Philipps-Universität Marburg/Lahn

vorgelegt von

Claus Stefer

aus Hanau am Main

Marburg/Lahn 2013

Vom Fachbereich Erziehungswissenschaften der Philipps-Universität Marburg
als Dissertation angenommen am: 20.11.2013

Abschluss der mündlichen Prüfung am: 26.11.2013

Betreuer: Prof. Dr. Udo Kuckartz
Zweitgutachterin: Prof. Dr. Inka Bormann

Inhaltsverzeichnis

1. Vorwort

„Wie soll ich meine Daten erheben?“ Mit dieser Frage ist jede Person, die ein empirisches Forschungsvorhaben angeht, irgendwann konfrontiert. Die Antworten können dabei sehr unterschiedlich ausfallen, denn in der Geschichte der empirischen Sozialforschung hat sich eine Vielzahl von Datenerhebungsmethoden herausgebildet, die zur Lösung unterschiedlichster forschungspraktischer Probleme und Fragen, aber auch aus verschiedenen wissenschaftstheoretischen Standpunkten heraus entstanden sind. Forschenden stellt sich folglich stets die Aufgabe, Verfahren zu wählen, die den Anforderungen der eigenen Untersuchung am besten entsprechen, was häufig auch eine Modifikation oder Kombination von Datenerhebungsmethoden erforderlich machen kann. Von der Wahl geeigneter Datenerhebungsmethoden hängt maßgeblich ab, ob bzw. wie gut die erzeugten Daten zur Bearbeitung der Forschungsfrage geeignet sind. In diesem Sinne gilt die Wahl geeigneter Methoden auch als ein wesentliches Gütekriterium empirischer Untersuchungen. Auf der Suche nach einem Maßstab für die Eignung eines Verfahrens begegnet den Forschenden in der Regel der Begriff der *Gegenstandsangemessenheit*, mit dem umschrieben wird, dass die Tauglichkeit einer Datenerhebungsmethode anhand der Anforderungen von Forschungsfrage, -gegenstand sowie weiterer beeinflussender Faktoren bestimmt werden soll. Steinke charakterisiert die Gegenstandsangemessenheit als Kennzeichen qualitativer Forschung (Steinke 1999, S. 38 ff), Flick et al. (Flick/von Kardorff/Steinke 2000, S. 327) und Helfferich (Helfferich 2009, S. 46) benennen sie als wichtigen Indikator und zentrales Kriterium zur Auswahl von Methoden, Brüsemeister bezeichnet sie als Zentralbegriff qualitativen Forschens (Brüsemeister 2008, S. 28). Doch bei genauerer Betrachtung zeigt sich an dieser Stelle eine Forschungslücke: Zwar finden sich im Kontext einzelner Methodenbeschreibungen oder auch in einigen Lehrbüchern konkretere Hinweise auf Möglichkeiten zur Überprüfung der Angemessenheit,

auch finden sich in Form von fachlichen Standards vereinzelt mehr oder weniger explizite Verweise auf entsprechende Möglichkeiten, allerdings stehen diese Einzelaspekte bislang unverbunden nebeneinander. Eine übergreifende und auch in der Praxis nutzbare Definition des Konstruktes der Gegenstandsangemessenheit oder gar leitende Kriterien zur Bestimmung angemessener Methoden existieren bislang nicht.

An dieser Stelle setzt die vorliegende Arbeit an und geht der Forschungsfrage nach: *Wie lässt sich im Rahmen von Lehrevaluation an Hochschulen die Gegenstandsangemessenheit von Datenerhebungsmethoden bestimmen?* Sie befasst sich mit einer Untersuchung des Konstruktes der Gegenstandsangemessenheit von Datenerhebungsmethoden, um Faktoren herauszuarbeiten, welche die Einschätzung der Eignung der fraglichen Methoden fundiert ermöglichen. Die Untersuchung erfolgt anhand des Beispiels von Lehrevaluationen an Hochschulen, weil sich dieser Gegenstand durch eine relativ klare Struktur auszeichnet, was die Identifikation von Akteuren und weiteren Einflussgrößen ermöglicht. Zudem verfügt die Evaluationsforschung, der dieser Gegenstand zuzurechnen ist, über eine breite und sehr fundierte eigene Wissensbasis. An diese Basis kann angeknüpft werden, um weitere relevante Aspekte zu identifizieren. Auf diese Weise wird auch eine Verallgemeinerung möglich, welche die Grundlage für eine Übertragung der Erkenntnisse auf andere Gegenstände ist. Die Arbeit ist hiermit so ausgerichtet, dass sie sowohl einen Beitrag zur methodologischen Diskussion als auch zur Lösung des Problems der Auswahl von Datenerhebungsmethoden in der Praxis, und zwar auch über den reinen Evaluationskontext hinaus, darstellen will.

Entsprechend des Erkenntnisinteresses der Arbeit werden in Kapitel 2 zunächst die Felder der Evaluation im Allgemeinen und der Hochschulevaluation im Besonderen dargestellt. Aufbauend auf diesen Grundlagen wird in Kapitel 3 die Frage der Gegenstandsangemessenheit von Datenerhebungsmethoden in Evaluationen aus theoretisch-methodologischer Perspektive bearbeitet, und es wird untersucht, welche Implikationen sich für die praktische Auswahl von Erhebungsverfahren ergeben. Ein zweiter erforderlicher Ausgangspunkt zur Beantwortung der Forschungsfrage ist eine empirische Untersuchung der Sichtweisen von Stakeholdern einer Evaluation, die in Kapitel 4 dargestellt

wird. Dabei wird in den Blick genommen, welche Vorstellungen von Gegenstandsangemessenheit sich auf Seiten der Stakeholder erkennen lassen, welche Prioritäten sie bezüglich der Kriterien setzen und welche Bedürfnisse verschiedene Stakeholdergruppen äußern. Diese Ergebnisse werden abschließend zu stakeholderbezogenen Kriterien ausgearbeitet. In Kapitel 5 schließlich werden die Erkenntnisse aus den Kapiteln 3 und 4 zusammengeführt und zu einem in der Planungspraxis anwendbaren Kriterienkatalog zur Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden konzentriert.

Während der Arbeit an dieser Dissertation habe ich viel Hilfe erfahren, für die ich mich herzlich bedanken möchte:

Prof. Dr. Udo Kuckartz, dem Betreuer dieser Arbeit, danke ich für die vielfältige Unterstützung, die er mir während der gesamten Arbeit an diesem Text hat zuteil werden lassen. Besonders danke ich ihm für die Möglichkeiten, die er mir zur Erarbeitung des komplexen Themas geboten hat sowie für die Beratung in methodologischen wie in organisatorischen Fragen. Die immer sehr vertrauensvolle Zusammenarbeit war ein wichtiger Faktor für das Gelingen des Vorhabens.

Meinen Interviewpartner/-innen danke ich für die Bereitschaft, mir ihre Sicht auf das Thema zu offenbaren – trotz voller Terminpläne und fehlenden persönlichen Nutzens.

Ursula Lapczynski danke ich für ihre Unterstützung bei Teilen der Transkription, die für mich eine sehr große Hilfe war. Rita Stefer danke ich für das stets schnelle und genaue Lektorat des Textes sowie für die große Bereitschaft zu geduldigen Rückmeldungen.

Schließlich danke ich besonders meiner Frau Antonia. Nicht nur organisatorische Hilfe und Entlastung, sondern gerade die vielen stets kritischen Gespräche und Ideen zu inhaltlichen Fragen haben einen unschätzbaren und entscheidenden Anteil am erfolgreichen Abschluss der Arbeit.

2. Evaluation und Lehrevaluation an Hochschulen: Ein Überblick

Der Begriff *Evaluation* hat Eingang in den alltäglichen Sprachgebrauch gefunden, besonders im Umfeld geistes-, sozialwissenschaftlicher und verwandter Theorie und Praxis. Mit großer Selbstverständlichkeit wird er in den unterschiedlichsten Kontexten verwendet, in der Politik ebenso wie z. B. im Gesundheitswesen oder in der sozialen Arbeit. Was genau mit dem Begriff bezeichnet wird, bleibt dabei allerdings oft unklar. Häufig werden damit Tätigkeiten bezeichnet, deren Ziele sich in einem Spektrum zwischen (positiver) Weiterentwicklung und (negativer) Rationalisierung bzw. Einsparung eines Evaluationsgegenstandes, z.B. eines Programms[1] oder einer Maßnahme, bewegen. Die jeweilige Bedeutung ist sehr stark vom Kontext der Begriffsverwendung abhängig und dabei so unterschiedlich wie das Interesse an Evaluationen und ihren angenommenen oder erhofften Wirkungen und Folgen.

In diesem Kapitel wird zunächst eine grundlegende Eingrenzung des Begriffs Evaluation vorgenommen, bevor verschiedene Aspekte des Bezugsrahmens vorgestellt werden. Daran anschließend werden Entwicklung und Hintergründe des Begriffs sowie Ziele und Probleme von Evaluationen im Sektor Hochschule dargestellt.

2.1. Evaluation

Der Begriff Evaluation bedeutet wörtlich *Bewertung* (Dudenredaktion 2005). Im Sinne dieser sicher allgemeinsten Definition wird der Begriff oft in der Alltagssprache ver-

[1] „Programme sind beschriebene und durchgeführte, intentional aufeinander bezogene Bündel von Interventionen, Maßnahmen, Projekten oder Teilprogrammen. Sie bestehen aus einer Folge von auf ausgewiesene Ziele hin ausgerichteten Aktivitäten. Sie werden auf der Basis von verfügbaren Ressourcen durchgeführt und sind darauf gerichtet, vermittels bereitgestellter Leistungen (outputs) bestimmte outcomes bei bezeichneten Zielgruppen oder im sozialen System zu erreichende impacts auszulösen.“ (Beywl, Eval-Wiki)

standen und eingesetzt. Die Verwendung dieses (vermeintlichen) Fachbegriffs bringt dabei allerdings nicht unbedingt erläuternde Klarheit mit sich, wie es eigentlich von den Sprechern intendiert ist, sondern lässt eine Fülle von Fragen offen. So bedeutet Bewertung zunächst nämlich nichts weiter, als dass irgendetwas von irgendjemandem nach irgendwelchen Kriterien in irgendeiner Weise bewertet wird (Kromrey 2005, S. 31). Der Begriff steht nicht per se für eine bestimmte Vorgehensweise oder gar für festgelegte Strategien oder die Einhaltung von Qualitätsmerkmalen: Der tägliche Blick aus dem Fenster mit der darauf folgenden Feststellung, ob das Wetter gut oder schlecht ist, ist in diesem Sinne bereits ein Urteil infolge eines Bewertungsprozesses, also eine Evaluation. Das eingesetzte Verfahren und seine Kriterien mögen bewährt sein, sind aber keineswegs transparent oder nachvollziehbar. Bei dieser Vorgehensweise können unterschiedliche Bewerter/-innen bei der Beurteilung des gleichen Gegenstandes durchaus zu unterschiedlichen Ergebnissen kommen, weil sie verschiedene Bewertungskriterien anlegen. Die jeweiligen Hintergründe und Standpunkte bedingen hier die Unterschiede. Dieses Phänomen ist vermutlich jedem vertraut, sei es von der Suche nach dem „Wählerwillen“ im Nachgang von Wahlen oder von der Beurteilung einer Unterrichtseinheit durch die Teilnehmenden.

Diese Art von mehr oder weniger willkürlicher Bewertung ist nicht für jeden Bereich tauglich. Denn spätestens dann, wenn die Bewertung mit Folgen für andere einhergeht, beispielsweise im sozialpolitischen Bereich, werden berechtigte Fragen nach der Grundlage der Bewertung auftauchen. Soll eine Bewertung nachvollziehbar sein, ist es notwendig, sie nach dokumentier- und nachvollziehbaren Verfahrensweisen und Kriterien durchzuführen und diese transparent zu machen. So wird es möglich, eine Bewertung sachlich zu diskutieren.

Seit den 1930er Jahren und ausgehend von den USA lässt sich die Bestrebung feststellen, den Prozess der Evaluation zu professionalisieren, um dem Problem der Nicht-Wissenschaftlichkeit bzw. der Willkür der Bewertung zu begegnen (vgl. Gollwitzer/ Jäger 2007, S. 8 ff., Hense 2006, S. 31 ff. oder Beywl 1988). Ein entscheidender Entwicklungsschritt in diesem Bestreben war die Integration empirischer Forschungsmethoden.

Denn auch an empirische Sozialforschung wird der Anspruch gerichtet, transparent und nachvollziehbar zu sein. Diese Verbindung zwischen Evaluation und empirischen Forschungsmethoden hat sich auch in einem eigenen Begriff niedergeschlagen, der auf Edward Suchman zurückgeht: *Evaluationsforschung* (Suchman 1967, S. 7).

2.1.1. Evaluation und Evaluationsforschung

Evaluations*forschung*[2] unterscheidet sich von Evaluation durch die systematische Anwendung empirischer Forschungsmethoden (Suchman 1967, S. 7, Gollwitzer/Jäger 2007, S. 6, Meyer/Höhns 2002, S. 2 ff., Bortz/Döring 2005, S. 102). Die Bewertung eines Gegenstandes wird in der Evaluationsforschung also aufgrund transparenter und nachprüfbarer Verfahren vorgenommen und ebenso offengelegt wie die Bewertungskriterien und sonstige Faktoren, die Einfluss auf die Bewertung haben. Auch so können, wie Beywl ausführt, sich widersprechende Bewertungen zu Stande kommen, je nachdem, welche Werteposition der/die Beurteilende einnimmt (Beywl 2006, S. 95). Allerdings werden die Gründe offengelegt und können somit Gegenstand einer Diskussion werden. An Verfahren steht dabei das gesamte Instrumentarium empirischer Sozialforschung zur Verfügung. Neben den quantitativen Datenerhebungs- und Auswertungsverfahren, die oftmals mit Evaluation assoziiert werden, haben zwischenzeitlich auch zahlreiche Varianten qualitativer Verfahren Einzug in die Evaluationsforschung gehalten.

Bewertungen, die im Sinne der Evaluationsforschung zu Stande kommen, müssen nicht unbedingt treffender oder genauer sein also solche, die ohne empirische Forschungsmethoden entstanden sind. So kann eine rein erfahrungsbasierte Wettervorhersage ebenso exakt – oder unter Umständen sogar genauer – sein als eine, die mit Hilfe wissenschaftlicher Erkenntnisse und aufwändiger Computersimulation erstellt wurde. Dennoch ist der Einsatz von Methoden empirischer Sozialforschung unverzichtbar, wenn eine Evaluation mehr sein soll als eine subjektive und nur schwer nachvollziehbare in-

[2] Stefan Hirschauer schlägt unter dem Hinweis auf Begriffe wie „Arbeitsmarktforschung“ oder „Familienforschung“ vor, Forschung unter dem Einsatz empirischer Forschungsmethoden als *evaluierende Forschung* zu bezeichnen und den Begriff Evaluationsforschung für Forschung *über* Evaluationen zu verwenden (Hirschauer 2006, S. 406). Damit befindet er sich in großer Nähe zu Suchman, der in seiner Begriffskreation *evaluative research* dem Wort *evaluative* die Funktion eines Adjektives zuweist (Suchman 1967, S. 7).

dividuelle Einschätzung. Zum einen ist Nachvollziehbarkeit mit Blick auf die Wirkung der Evaluation bedeutsam, da ihre Ergebnisse in aller Regel nicht ausschließlich einer eingeweihten Fachöffentlichkeit vorgestellt werden, sondern oftmals eine breitere Öffentlichkeit und ggf. auch fachliche Laien betreffen, die nicht unbedingt mit den zugrunde liegenden Verfahren und Kriterien vertraut ist. So wird sie zu einem gewichtigen Aspekt bei der Geltungsbegründung. Zum anderen dient die Nachvollziehbarkeit zur Dokumentation der Qualität der Evaluation. So lässt sich beispielsweise überprüfen, ob die gewählte Vorgehensweise ethischen Standards genügt oder ob die abschließend getroffenen Bewertungen auch tatsächlich eine Grundlage in den erhobenen Daten haben. Zusammengefasst lässt sich festhalten: Der Einsatz empirischer Methoden im Rahmen von Evaluationen leistet einen entscheidenden Beitrag zur Absicherung der Ergebnisse, erleichtert die Beweisführung und bildet die Grundlage einer sachlichen Betrachtung und Diskussion der Evaluation und ihrer Ergebnisse[3].

2.1.2. Grundlagen- und Evaluationsforschung

Das Bestreben, durch den Einsatz dokumentierter und theoretisch abgesicherter Methoden nachprüfbare und gut begründete Ergebnisse hervorzubringen, ist natürlich nicht nur Kennzeichen von Evaluationsforschung, sondern liegt jeder ernsthaften wissenschaftlichen Forschung zugrunde. Somit stellt sich die Frage, worin die Besonderheiten von Evaluationsforschung liegen und weshalb sie mitunter – nicht nur im Kreise professionell Evaluierender – als eigene Disziplin angesehen wird.

Evaluationsforschung kann als eine besondere Form der angewandten Sozialforschung angesehen werden (vgl. etwa Stockmann 2006a, S. 16, Kromrey 2003). Als angewandte Sozialforschung wird empirische Forschung bezeichnet, die auf die unmittelbare Lösung praktischer, etwa gesellschaftlich-politischer Probleme hin angelegt ist (Habermehl 1992, S. 9). Angewandte Sozialforschung ist also Forschung, die konkret zweckgebunden durchgeführt wird und zu deren ausdrücklichem Ziel es gehört, Ergebnisse zu produzieren, die in der Praxis direkt nutzbar und somit handlungsrelevant sind. Von der

[3] Der Einsatz empirischer Methoden im Rahmen von Evaluationen wird auch in den *Standards für Evaluation* der DeGEval gefordert, siehe Kapitel 2.1.4.

angewandten Sozialforschung ist die Grundlagenforschung zu unterscheiden, die eben nicht unmittelbar nutzenbezogen ausgerichtet ist, sondern das Ziel verfolgt, Wissen und Erkenntnisse – wenn möglich *Wahrheit* – über Forschungsgegenstände zu erzeugen. Die Ergebnisse von Grundlagenforschung unterliegen nicht dem Anspruch praktischer Verwendbarkeit, sondern dienen der Vermehrung des wissenschaftlich gesicherten Wissensbestandes und sind etwa oftmals wichtige Beiträge zur Theoriebildung/-testung oder Grundlage für weitere Forschung. Anders ausgedrückt: „Grundlagenforschung formuliert Hintergrundwissen, dessen funktionaler Wert nicht unmittelbar erkennbar sein muss." (Bortz/Döring 2005, S. 103)i. Diese grundlegende Unterscheidung zwischen angewandter Sozialforschung und Grundlagenforschung gilt auch für die Unterscheidung zwischen Evaluationsforschung und Grundlagenforschung. Darüber hinaus allerdings ist Evaluationsforschung durch weitere spezifische Merkmale gekennzeichnet, die sie nicht nur von der Grundlagenforschung absetzt, sondern sie auch zu einem Spezialfall der angewandten Sozialforschung macht:

Unterscheidungsmerkmale zwischen Grundlagen- und Evaluationsforschung

Ziele: Die unterschiedlichen Zielsetzungen von Evaluations- und Grundlagenforschung wurden bereits dargestellt. Patton fasst sie folgendermaßen zusammen: „Research [i. S. v. Grundlagenforschung, Anm. d. A.] aims to produce knowledge and truth. Useful evaluation supports action." (Patton 1997) Die Unterscheidung auf der Ebene der Untersuchungsziele ist die Grundlage, aus der sich viele der nachfolgend benannten Unterschiede ergeben bzw. auf die sie sich zurückführen lassen.

Fragestellungen und Untersuchungsgegenstände: In der Grundlagenforschung werden die Fragestellungen und damit einhergehende Hypothesen in der Regel durch die Forschenden und somit aus der Wissenschaft heraus bestimmt. Evaluationsforschung ist hingegen oftmals Auftragsforschung, d. h. dass die Fragestellungen von Auftraggebern vorgegeben werden. Aber auch Evaluationen, die nicht im Auftrag durchgeführt werden, dienen meist außerwissenschaftlichen Erkenntnisinteressen und Verwertungskontexten (vgl. Balzer 2005, S. 18, Stockmann 2006a, S. 16).

Rollen: Evaluatoren sind oftmals in ein ihre Rolle bestimmendes System eingebunden und diesem gegenüber verpflichtet. Grundlagenforscher sind hingegen prinzipiell freier und nur sich selbst, der wissenschaftlichen Öffentlichkeit und dem Erkenntnisgewinn verpflichtet – so zumindest der idealtypische Entwurf (Balzer 2005, S. 18/19).

Urteilskriterien und Bewertung: Eines der zentralen Kennzeichen, wenn nicht sogar die Kernleistung von Evaluationsforschung ist, wie bereits erwähnt, die Tatsache, dass sie stets mit Bewertung verbunden ist (Stockmann 2006a, S. 16). Während Grundlagenforschung den aktuellen Status quo möglichst wertneutral untersucht, beschreibt, zu erklären versucht und eventuell auch zukünftiges Geschehen aus ihm ableitet (Balzer 2005, S. 19), urteilt Evaluation über den untersuchten Gegenstand. Neben dem Vergleich von Ist- und Sollzustand oder der Überprüfung vorher festgelegter Erfolgskriterien (Balzer 2005, S. 19) wird auch das Verfahren des Benchmarkings eingesetzt, in dem Evaluationsgegenstände (z.B. Schulsysteme) hinsichtlich ihrer Zielerreichung miteinander verglichen werden. Der Aspekt der unbedingten Bewertung unterscheidet Evaluationsforschung nicht nur von Grundlagenforschung, sondern auch von anderen Formen der angewandten Sozialforschung.

Datenerhebungs- und Datenauswertungsmethoden: In der Regel unterscheiden sich die eingesetzten Datenerhebungs- und -auswertungsmethoden zwischen Grundlagen- und Evaluationsforschung nicht (Balzer 2005, S. 19, Stockmann 2006a, S. 17). Allerdings merkt Kromrey an, dass Evaluationsforschung die Komplexität des untersuchten Gegenstandes nicht in dem Maße reduzieren kann, wie es bei Grundlagenforschung etwa in Form von Laborexperimenten üblich ist. Realitätsferne Laborsituationen beispielsweise werden als ungeeignet erachtet (Kromrey 2003). Gerade in der Erfassung der Komplexität der untersuchten Gegenstände liegt häufig ein notwendiger Bestandteil von Evaluation, denn eine fundierte und angemessene Bewertung ist kaum möglich, wenn evtl. beeinflussende Faktoren ausgeblendet werden.

Zeit: Zeitvorgaben sind bei Evaluationen eine Rahmenbedingung mit besonderem Gewicht. Die Ergebnisse der Evaluation haben in aller Regel zu einem vorher definierten Zeitpunkt vorzuliegen. Zu spät vorgelegte Resultate sind oft nicht mehr verwertbar. Grundlagenforschern steht im Prinzip mehr Zeit zur Verfügung, die auch frei gewählt werden kann (Balzer 2005, S. 19/20). „Neue Erkenntnisse [...] lassen sich [in der Grundlagenforschung, Anm. d. A.] schließlich nicht mit dem Terminkalender erzwingen“ (Kromrey 2003).

Ergebniserwartung: Von Evaluationen wird ein Ergebnis erwartet, das bestimmten vorher formulierbaren Ansprüchen genügt – zumindest müssen Bewertungen eines möglichst exakt bestimmten Gegenstandes vorgenommen werden. Werden diese Ergebnisse nicht produziert, ist eine Evaluation gescheitert. Die Ergebnisse von Grundlagenforschung hingegen sind vorher oft nicht absehbar. Und auch Grundlagenforschung, die zu keinen der Intention der Forschung entsprechenden oder sogar zu ihr entgegenstehenden Ergebnissen kommt, kann u. U. als erfolgreich bezeichnet werden: „Deren informationsreiches Scheitern ist nicht selten der Startpunkt für grundlegende Erkenntnisse, die eine neue Forschungslinie begründen“ (Kromrey 2003). Schließlich geht es letzten Endes um die Vermehrung des wissenschaftlichen Wissensbestandes, in dem auch Scheitern von Forschung eine durchaus wichtige Erkenntnis sein kann.

Veröffentlichungen: Evaluationsergebnisse werden nur selten frei und in der Form publiziert, wie es bei Ergebnissen aus der Grundlagenforschung eher die Regel ist. Erkenntnisgewinn und -verbreitung sind schließlich zentrale Anliegen der Grundlagenforschung (Balzer 2005, S. 20). Evaluationsergebnisse hingegen werden typischerweise nur den Stakeholdern[4], manchmal sogar nur bestimmten Stakeholdergruppen zugänglich gemacht.

[4] Als Stakeholder einer Evaluation wird die Gesamtheit der an einer Evaluation beteiligten oder durch sie betroffenen Personen bezeichnet (Beywl, Widmer/Landert/Bachmann 2000). Als Beteiligte werden dabei jene Personen bezeichnet, die aktiv an der Evaluation mitwirken, sei es in auftraggebender, planender, durchführender oder datengebender Funktion. Die Gruppe der Betroffenen umfasst hingegen jene Personen, auf die die Ergebnisse der Evaluation Auswirkungen entfalten, ohne dass sie selbst aktiv in den Prozess eingebunden wären.

Empfänger von Ergebnissen: Beide Forschungsvarianten haben nur im Ausnahmefall ein großes Publikum, aber aus unterschiedlichen Gründen. Evaluationsforschung verallgemeinert in der Regel nicht, weshalb ihre Ergebnisse eben nur für eine eng umrissene Zielgruppe, meist die Stakeholder der Evaluation, von Bedeutung sind (Balzer 2005, S. 20).[5] Allerdings liegt in der Heterogenität der Stakeholder eine besondere Herausforderung: Da Evaluationsergebnisse nicht nur durch die Scientific Community aufgenommen und bewertet werden, müssen sie so aufbereitet werden, dass sie auch für ein interessiertes, aber weniger fachkundiges Publikum nachvollziehbar und verständlich sind (Bortz/Döring 2005, S. 103, Kelle/Erzberger 2006, S. 284). Die Ergebnisse von Grundlagenforschung indes sind mitunter so speziell und/oder komplex, dass sie außerhalb der Wissenschaft oder manchmal auch nur eines kleinen Kreises von spezialisierten Wissenschaftlern nicht auf unmittelbares Interesse stoßen.

Ergebnisnutzung: Während Ergebnisse aus Grundlagenforschung nicht mit dem direkten Zweck der Nutzung verbunden sind, ist die Verwendung von Ergebnissen aus Evaluationsprojekten gerade der dahinter liegende Sinn (Balzer 2005, S. 20). Kromrey nennt außerwissenschaftliche Erkenntnis- und Verwertungsinteressen als zentrale Definitions- und Legitimationskriterien und fordert, die unmittelbare Verwertbarkeit der erzielten Resultate während des gesamten Evaluationsprozesses nie aus dem Blick zu verlieren (Kromrey 2003). Anders formuliert:

> „An evaluation is supposed to be used in immediate or future debate and decision-making on the problem or intervention at hand. This ist not at all the case with fundamental research. Fundamental researchers have no specific application in mind. The basic researcher attempts to live according to the legendary, humorous toast at the Christmas party of the Cambridge economists: ‚Long live economics and may it never be of any use.' In evaluation, use ist more or less planned. Basic research may be useful, but its use is accidental and uplanned." (Vedung 2006, S. 120)

[5] Je nach Evaluationsgegenstand kann die Menge der Stakeholder durchaus sehr groß werden, etwa wenn staatlich geförderte Programme evaluiert werden und letztlich der Staat in Form seiner geldgebenden Bürger Empfänger der Ergebnisse ist. Solche Dimensionen dürften in der Summe der durchgeführten Evaluationen allerdings einen geringen Anteil ausmachen.

> Daher leitet sich die Forderung ab, dass die Ergebnisse von Evaluationsforschung möglichst fehler- und irrtumsfrei sein müssen. Grundlagenforschung hingegen „[...]darf sich irren, Hypothesen dürfen sich als falsch erweisen. [...] Denn informationsreiches Scheitern ist nicht selten der Startpunkt für grundlegende Erkenntnisse, die eine neue Forschungslinie begründen. Diesen Luxus darf sich angewandte Sozialwissenschaft nicht leisten." (Kromrey 2003)

Diese Darstellung ist durchaus idealisierend. Ob es Grundlagenforschung in der dargestellten Form, losgelöst von äußeren Zwängen, überhaupt gibt, wird kontrovers diskutiert (Bortz/Döring 2005, S. 99). Dennoch handelt es sich um zwei sich unterscheidende und auch in der Praxis unterscheidbare Richtungen[6] und um hilfreiche Verständniskategorien. Im direkten Vergleich beider Richtungen genießt allgemein die Grundlagenforschung das höhere Ansehen. Der Hauptgrund liegt darin, dass angewandte Sozialforschung – und somit auch Evaluationsforschung – in der Regel Kompromisse zwischen Wissenschaftlichkeit und Praxisbedürfnissen eingehen muss und so eine aus wissenschaftlicher Sicht ungewöhnliche Verschiebung der Rangordnungen entsteht: Vor allem bei Evaluationen müssen im Zweifelsfall grundlagenwissenschaftlich-methodische Ansprüche (etwa die Kontrolle störender Faktoren) hinter jene an die Verwendbarkeit der Ergebnisse zurücktreten (Balzer 2005, S. 21, Kromrey 2005). Und gerade dieser Ausgleich zwischen wissenschaftlichen und verwertungsinduzierten Ansprüchen ist eine der zentralen Herausforderungen für Evaluatoren, die so im Bereich der Grundlagenforschung nicht existiert. Um hier sicher agieren zu können, ist eine Fülle von Wissen, Fähigkeiten und Feingefühl notwendig, nicht zuletzt im Bereich der Forschungsmethoden:

> „Der Evaluator muss das Repertoire des Forschers beherrschen, und dieses gleichzeitig auf reale Situationen, die sich seiner Kontrolle entziehen, anwendbar machen. Rückt er die Wissenschaft zu sehr in den Vordergrund, droht die Gefahr, dass das Projekt am Widerstand des Feldes scheitert oder aber zwar exakte, aber praktisch völlig irrelevante oder ignorierte Ergebnisse produziert. Betont er zu stark die Notwendigkeiten des Feldes, könnte die Wissenschaftlichkeit auf der Strecke

[6] Die Deutsche Forschungsgemeinschaft (DFG) etwa fördert ausschließlich Forschungsvorhaben der Grundlagenforschung.

> bleiben, was Ergebnisse ebenfalls unbrauchbar macht. Gute Evaluationen sind daher ein Drahtseilakt, und Evaluation als Disziplin ist keine Forschung zweiter Klasse, sondern als angewandte Wissenschaft ein Teilgebiet sozialwissenschaftlicher Forschung, das weltweit seit vielen Jahren zu den am stärksten wachsenden Betätigungsfeldern von Akademikern zu zählen ist.“ (Balzer 2005, S. 21)

Über all diese Unterschiede hinweg bleibt der zentrale und sicher markanteste Unterschied jener, der die Ziele betrifft:

> „In conclusion, the basic difference between evaluation research and fundamental research is that the former is intended for use. Evaluation research should be useful and used, but that is not integrated into the ethos of fundamental research.“ (Vedung 2006, S. 135)

Vor dem dargestellten Hintergrund ist die Definition des Begriffs Evaluation, wie sie die *DeGEval Gesellschaft für Evaluation* vorgenommen hat, sehr präzise und bildet auch in dieser Arbeit die Grundlage des Verständnisses des Begriffs Evaluation:

> „Evaluation ist die systematische Untersuchung des Nutzens oder Wertes eines Gegenstandes. Solche Evaluationsgegenstände können z.B. Programme, Projekte, Produkte, Maßnahmen, Leistungen, Organisationen, Politik, Technologien oder Forschung sein. Die erzielten Ergebnisse, Schlussfolgerungen oder Empfehlungen müssen nachvollziehbar auf empirisch gewonnenen qualitativen und/oder quantiativen Daten beruhen.“ (DeGEval Gesellschaft für Evaluation 2008, S. 15)

2.1.3. Evaluationsmodelle

Trotz der rasch fortschreitenden Entwicklung der Disziplin Evaluation, die in den USA in den 1930er Jahren ihren Anfang genommen (Gollwitzer/Jäger 2007, S. 8/9, Hense 2008) und seitdem zu einem hohen Grad an Professionalisierung geführt hat, hat sich im wissenschaftlichen Sinne noch keine umfassende Evaluationstheorie entwickelt. Als Theorie wird dabei ein Modell verstanden, das zentrale Grundannahmen sowie beschreibende und erklärende Aussagen über einen Teil empirischer Realität (hier: Evaluation und ihre Wirkungsweisen) enthält, aus denen sich Hypothesen über Zusammenhänge sowie Regeln für deren Messung bilden lassen (Stockmann 2007a, S 40). Auf einer darunter liegenden Ebene ist jedoch eine Fülle unterschiedlicher Evaluationsmodelle[7]

[7] Alternativ, aber oftmals synonym wird häufig auch der Begriff *Evaluationsansätze* verwendet. Gelegentlich werden Evaluationsmodelle auch als Evaluationstheorien bezeichnet. Damit wird aber

entstanden, bei denen es sich um „theoretisch begründete, in Publikationen schriftlich dargelegte und durch praktische Evaluationserfahrungen gesättigte Konzepte dazu, wie Evaluationen geplant und durchgeführt werden sollen“(Beywl 2006, S. 92/93), handelt. Der Wert dieser Modelle liegt darin, dass sie ausformulierte Vorschläge zur praktischen Strukturierung des Evaluationsvorhabens sowie zu seiner Reflexion bereitstellen (Stockmann 2007a, S. 40). Dabei setzt jedes Modell eigene Schwerpunkte und nimmt bestimmte Aspekte des Evaluationsgegenstandes und/oder -feldes in besonderer Weise in den Fokus oder betrachtet den Gegenstand von einer spezifischen Position aus, um eine angemessene Bewertung zu ermöglichen. Die Reichweite der verschiedenen Modelle ist sehr unterschiedlich. Einige geben den Evaluatoren recht detaillierte schrittweise Anleitungen für die Durchführung an die Hand, andere hingegen betrachten hauptsächlich einen Teilaspekt (z.B. die Ergebnisnutzung) und blenden andere Bereiche (etwa die Datengewinnung) aus. Auch dies zeigt, dass die Modelle bisher noch nicht den Status einer Theorie erreicht haben.

Im Folgenden werden beispielhaft das CIPP-Modell, die Vier Ebenen, die Goal-free Evaluation, Nutzenorientierte Evaluation, Responsive Evaluation und Empowerment Evaluation kurz skizziert[8] (Balzer 2005, S. 29 ff., Hense 2008, Beywl):

Das CIPP-Modell, das auf Daniel Stufflebeam zurückgeht, beschreibt vier Aspekte eines Evaluationsgegenstandes, die je nach Rahmen einer Evaluation zu berücksichtigen sind: 1. den Kontext (*Context*, „C“), d. h. die Umgebung des Evaluationsgegenstandes, Bedarfslagen und Ziele, 2. den Input („I“), also Pläne, Konzepte und Personen, 3. den Prozess (*Process*, „P“) selbst, hier stehen Handlungen, Aktivitäten und Beziehungen im Vordergrund, sowie schließlich 4. das Produkt (*Product*, „P“), nämlich Effekte, intendierte und ungewollte Wirkungen und Zielerreichung. Ziel des Modells ist es, alle für die Bewertung relevanten Faktoren im Kontext des Evaluationsgegenstandes zu erfassen (Stufflebeam et al. 1971).

nicht, wie dargelegt, eine Theorie im wissenschaftlichen Sinne impliziert, sondern auch dieser Begriff wird dann synonym verwendet.

[8] Die Auswahl der hier aufgeführten Evaluationsmodelle aus der langen Liste der in Frage kommenden erfolgte mit dem Ziel, Beispiele für unterschiedliche Schwerpunktsetzungen von Modellen aufzuzeigen.

Die Vier Ebenen (The Four Levels), [9] ein durch Donald Kirkpatrick entwickeltes Modell, zielt hauptsächlich auf den Bereich der Evaluation von Trainingsmaßnahmen und Seminaren. Auch Kirkpatrick arbeitet, wie der Name des Modells ja unmittelbar verrät, mit einer Vierteilung, allerdings bezieht sich diese auf unterschiedliche Ebenen, denen Ergebnisse einer Evaluation zuzurechnen sind und somit auf die Art der erhaltenen Informationen: 1. die Ebene der Reaktion (unmittelbare Reaktionen der Teilnehmenden nach Abschluss einer Maßnahme, meist die Zufriedenheit der Teilnehmenden), 2. die Ebene des Lernerfolgs (Erkenntnisse über Art und Umfang erworbenen Wissens und erlernter Fähigkeiten, setzt die Definition von Lernzielen voraus), 3. die Ebene des Verhaltens (vornehmlich die Verhaltensänderung, also etwa der Wissenstransfer in Alltagssituationen) und abschließend 4. die Ebene der Endergebnisse (Erkenntnisse über endgültige Effekte der Maßnahme, z.B. Produktivitäts- oder Qualitätssteigerung). Zu jeder der Ebenen schlägt Kirkpatrick passende Methoden vor. Im Rahmen von Evaluationsprojekten sind die Ergebnisse aller vier Ebenen wichtig (Kirkpatrick 1987).

Goal-free Evaluation wurde durch Michael Scriven ausgearbeitet und verlangt, dass das Evaluationsteam den gewählten Gegenstand evaluiert, ohne dessen intendierte Ziele zu kennen. So soll gewährleistet werden, dass nicht nur die erwünschten, sondern die tatsächlichen – ggf. auch negativen – Wirkungen einer Maßnahme erfasst werden. Dieses Modell dient also dem Ziel, eine größere Aufgeschlossenheit der Evaluatoren zu gewährleisten. Zielfreiheit meint nicht, dass die Evaluation selbst oder der evaluierte Gegenstand zielfrei wären (Scriven 1973).

Nutzungsorientierte Evaluation (Utilization-focused Evaluation) wurde in erster Linie durch Michael Patton geprägt. Der zentrale Merksatz dieses Modells lautet „Eine Evaluation, deren Ergebnisse durch die Praxis nicht genutzt werden, ist Verschwendung!“ (Beywl). Es wird als wesentlich erachtet, dass nicht nur die Auftraggeber, sondern alle Stakeholder die Gültigkeit der Evaluation und ihrer

[9] Ob es sich bei den *Vier Ebenen* tatsächlich um ein Evaluationsmodell handelt, wird kontrovers betrachtet. In Balzers Typologie, auf die später Bezug genommen wird, werden sie als Modell geführt und dieser Standpunkt wird hier übernommen.

Ergebnisse anerkennen. Die Nützlichkeit für die Praxis wird der Nützlichkeit für die Wissenschaft vorgeordnet. Evaluatoren haben dabei auch die Aufgabe, die tatsächliche Verwendung der Evaluationsergebnisse zu unterstützen. Patton hat einen fünfstufigen Ablaufplan entwickelt, anhand dessen nutzungsorientierte Evaluationen verwirklicht werden sollen: 1. werden die Nutzergruppen identifiziert. 2. werden die Nutzer zum geplanten Schwerpunkt der Evaluation und zur Verwertung ihrer Ergebnisse verpflichtet. 3. werden angemessene Methoden, Designs und Messungen erarbeitet und verwendet. 4. werden die Daten ausgewertet und interpretiert, es werden Entscheidungen gefällt und 5. werden Entscheidungen über die weitere Nutzung getroffen (Patton 1997).

Responsive Evaluation, ein auf Robert Stake zurückgehendes Evaluationsmodell, rückt die Anliegen und Informationsbedürfnisse der Beteiligten und Betroffenen in den Vordergrund. Die Responsivität des Modells liegt darin, dass diese Bedürfnisse bei der Bewertung des Evaluationsgegenstandes berücksichtigt und die Stakeholder während des gesamten Evaluationsprozesses einbezogen werden. Dementsprechend wird auch kein vorher festgelegtes Evaluationsmodell verfolgt, sondern im Prozess wird auf die jeweiligen Erfordernisse reagiert (Stake 1975).

Empowerment Evaluation schließlich, ein durch David Fetterman erarbeitetes Modell, hat das Ziel, Programmteilnehmer dabei zu unterstützen, sich selbst und das betreffende Programm zu evaluieren, um so die Praxis zu verbessern, indem Personen vor Ort Entscheidungen treffen und den zukünftigen Weg selbst bestimmen. Dem Evaluationsteam kommt in diesem Modell eine beratende und keine durchführende Rolle zu, wodurch sich dieses Modell fundamental von vielen anderen unterscheidet. Damit geht dieses Modell eigene Wege, um eine Nutzung der Evaluationsergebnisse zu erreichen und es zielt gleichzeitig auf die Befähigung und Stärkung der Stakeholder, auf Hilfe zur Selbsthilfe ab (Fetterman 2001).

An dieser Stelle könnte eine sehr große Anzahl weiterer Modelle aufgeführt werden, die sich von den hier vorgestellten teilweise stark, manchmal auch nur in Nuancen un-

terscheiden.[10] Diese Vielfalt an Modellen birgt, wie Stockmann festhält, das Problem, dass sie potenzielle Anwender ratlos zurücklässt (Stockmann 2007a, S. 40): Die Funktion, als Orientierungsrahmen zu dienen, die Ausarbeitung von Evaluationsfragen zu unterstützen, ihre Umsetzung in konkrete Arbeitsschritte zu erleichtern und idealerweise Anleitungen für die Praxis zu geben, würde aufgrund der kaum zu durchblickenden Vielfalt unterschiedlicher Modelle weitgehend konterkariert.

Um die Anwender bei der Suche nach einem zur jeweiligen Fragestellung passenden Evaluationsmodell zu unterstützen, das zumindest als Vorlage für Anpassungen und Veränderungen genutzt werden kann, wurde und wird vielfach versucht, die Modelle nach bestimmten Gesichtspunkten in Typologien zu ordnen und so die Orientierung und den gezielten Zugriff auf passende Modelle zu erleichtern:

> „Vergleichende Darstellungen von Evaluationsmodellen (Modelltypologien) funktionieren wie Reiseführer, welche die Landschaft in sinnvoller Gliederung beschreiben und für die Reisenden zugänglich machen. Geschichte, Topographie, Sehenswürdigkeiten und praktische Tipps sollen zum Besuch motivieren und helfen, sich die Landschaft aktiv zu erschließen.“ (Beywl 2006, S. 93)

Zwar bringen die Kategorisierungen ihrerseits Probleme wie eine mitunter starke Simplifizierung oder auch die Implikation einer nicht unbedingt vorhandenen Trennschärfe zwischen verschiedenen Ansätzen mit sich. Auch können beinahe beliebig viele Kriterien zur Bildung der Modellkategorien herangezogen werden, z. B. die Einteilung nach den Verwendung findenden empirischen Methoden, nach den Spezifika von Evaluationsgegenständen oder dem Umgang mit Werten innerhalb der verschiedenen Modelle (vgl. Stockmann 2007a, S. 40). Dennoch spiegeln diese Typologien den Alltag von Evaluationen und seine Vielfalt wider und sind letztlich unverzichtbar für die Suche nach einem passenden Modell – vorausgesetzt, es lässt sich eine Typologie finden, die den benötigten Schwerpunktsetzungen genüge leistet.

Zwei aktuelle und im deutschsprachigen Raum populäre Typologien sind die von Lars Balzer und Wolfgang Beywl:

[10] Weitere populäre und häufig genannte Modelle sind neben den Aufgeführten etwa die Theoriebasierte Evaluation, die Partizipative Evaluation oder die Fourth Generation Evaluation.

Lars Balzer wählt für die Darstellung und Sortierung der unterschiedlichen Evaluationsmodelle und -ansätze ein Klassifikationsmodell, das sich an Alkins Klassifikation orientiert und die Ansätze nach ihrer Schwerpunktsetzung gruppiert (Balzer 2005, S. 26 ff., Alkin 2004):

1. Ansätze, die ihr Hauptaugenmerk auf die Methoden und Verfahren der Evaluation richten,
2. Ansätze, die ihren Schwerpunkt auf den Akt der Bewertung legen und
3. Ansätze, bei denen die Nutzung der Evaluationsergebnisse in den Mittelpunkt gestellt wird, die also den Nutzungsaspekt betonen. Schließlich werden
4. Evaluationsstandards- und richtlinien beschrieben, die den Versuch darstellen, ansatzunabhängig Anforderungen an die Durchführung und Bewertung von Evaluationsprojekten zu formulieren.

Der Vorteil dieser Klassifikation liegt, wie Balzer verdeutlicht, einerseits darin, dass sie die keineswegs geradlinig verlaufende Entwicklung der Disziplin Evaluation widerspiegelt (die erste Kategorie ist die historisch älteste, die dritte die jüngste). Andererseits sind die Kategorien klar gegeneinander abgrenzbar. Balzers Typologie ist darüber hinaus nicht nur gut nachvollziehbar, sondern stellt deutlich sichtbar die Verschiebung der Schwerpunktsetzung innerhalb der Disziplin dar. Vor allem aber repräsentiert sie die drei gegenwärtig identifizierbaren Hauptäste der Diskussion der wissenschaftsbasierten Praxis Evaluation: Methoden, Bewertung und Nutzung (Beywl 2006, S. 94).

Beywls Modelltypologie hingegen orientiert sich an einem anderen für die Evaluation konstitutiven Element, an ihrer Kernleistung: der Bewertung (Beywl 2006, S. 94 ff.). Evaluationsmodelle werden danach eingeteilt, wie der Akt des Bewertens innerhalb oder außerhalb der Evaluation verortet ist, wie er organisiert ist und wie mit den zugrunde liegenden Werten umgegangen wird. So entstehen vier Typen von Evaluationsmodellen.

Die erste und größte Gruppe ist die der wertedistanzierten Modelle, bei denen die Klärung und Priorisierung von Werten außerhalb der Evaluation angesiedelt wird. Mo-

delle dieses Typs versuchen, Werte zu neutralisieren und so ihren Einfluss auf den Evaluationsprozess zu minimieren. Davon unterscheiden sich prinzipiell die übrigen drei Modelltypen, innerhalb derer Werte und Bewertung als integraler Bestandteil von Evaluationen angesehen werden. Die Modelle dieser Gruppen unterscheiden sich durch die für die Evaluierenden vorgesehenen Handlungsweisen bei divergierenden Werten. Innerhalb der werterelativistischen Modelle (zweite Gruppe) werden Werte in allen Phasen der Evaluation berücksichtigt. Wertegemeinsamkeiten und v.a. -konflikte werden herausgearbeitet und bestimmen die Evaluationsplanung. Evaluierende nehmen dabei nicht Partei. Die dritte Gruppe der wertepriorisierenden Modelle umfasst solche Modelle, die soziale Werte explizit berücksichtigen und dafür Aushandlungsspielräume zur Verfügung stellen, solange durch diese Aushandlung andere Anforderungen an Evaluationen (etwa Nützlichkeit, Durchführbarkeit) nicht eingeschränkt werden. Die letzte Gruppe der wertepositionierten Modelle (vierte Gruppe, umfasst gegenwärtig lediglich die Empowerment Evaluation) räumt Werten eine privilegierte Position ein, indem die Klärung von Wertfragen in den Evaluationsprozess integriert wird, da die involvierten Personen die Evaluation selbst durchführen und von Experten dabei unterstützt werden. Somit sind die Werte der Beteiligten zentrales Element der Evaluationsarbeit und der Umgang mit ihnen liegt völlig in der Hand der Betroffenen.

Innerhalb dieser Typologie lassen sich Unterschiede in den präferierten Methoden erkennen. Entlang einer fiktiven Linie zwischen den Modellen, die zur ersten Gruppe gehören und jenem der vierten Gruppe lässt sich eine Zunahme der Bedeutung von qualitativen Methoden erkennen. Je stärker Werte berücksichtigt werden, desto größer wird die Bedeutung qualitativer Methoden, und zwar über die systematische Informationsgewinnung hinaus (Beywl 2006, S. 113). Außerdem ermöglicht sie den Zugang zu den Modellen nach einem der zentralsten Kriterien von Evaluationen: der Frage nach dem Wie und Warum der Bewertung.

Bereits diese beiden Typologien verdeutlichen, wie unterschiedlich der Blick auf Evaluationsmodelle sein kann – und es könnte hier eine große Anzahl weiterer Typologien

aufgezählt werden.[11] Jeder Typologie liegt dabei eine spezifische Sichtweise auf Evaluation zugrunde, jede betont einen bestimmten Aspekt der betrachteten Modelle, meist zu Lasten eines anderen oder auch mehrerer anderer. Nicht zuletzt aufgrund dieser Betonung von Elementen sind Typologien unverzichtbar, um sich in der Vielfalt der Evaluationsmodelle zurechtzufinden, sie unter verschiedenen Gesichtspunkten zu betrachten und um schließlich eine zur gegebenen Problemstellung passende Auswahl zu treffen.

2.1.4. Evaluationsstandards

Merkmal und Anspruch professionell durchgeführter Arbeiten – dies gilt auch für Evaluationen – ist ein möglichst hohes Qualitätsniveau des Produktes. Um diese Qualität zu gewährleisten, etablieren sich oftmals allgemein anerkannte Regeln, die als Gütekriterien eine möglichst objektivierte Einschätzung der Qualität geleisteter Arbeit ermöglichen sollen. Im Bereich der Evaluation als Sonderform der angewandten Sozialforschung bietet es sich zunächst an, auf die Gütekriterien empirischer Sozialforschung zurückzugreifen. Allerdings wirken Evaluationen oftmals über den Bereich hinaus, der durch die dort gängigen Gütekriterien abgedeckt wird. Die vielschichtigen Interessen der unterschiedlichen Stakeholdergruppen können hier ebenso als Beispiel für nicht abgedeckte Bereiche herangezogen werden wie intendierte und nicht intendierte Rückwirkungen in das Feld. Dementsprechend haben sich im Bereich der Evaluation eigene Qualitätsregeln, sog. Standards, etabliert:

> „Ein Standard ist ein Prinzip, auf das sich die in einem Fachgebiet tätigen Praktiker geeinigt haben, und dessen Beachtung dazu beiträgt, daß die Qualität und die Fairneß der jeweiligen beruflichen Tätigkeit – z.B. Evaluation – verbessert werden." (Sanders/Beywl 2006, S. 27)

Einen Grundkonsens über Standards für die Durchführung von Evaluationen stellen die 1994 in den USA erschienenen *Program Evaluation Standards* (Sanders/American

[11] Andere bekannte und populäre Typologien sind beispielsweise der *Evaluation Theory Tree* von Alkin, die Typologie von Stufflebeam, Madaus und Kellaghan oder die Typologie von Worthen und Sanders.

Association of School Administrators/Joint Committee on Standards for Educational Evaluation 1994) des *Joint Committee on Standards for Educational Evaluation* dar. Bei diesen Standards handelt es sich um eine grundlegend überarbeitete Fassung der im Jahre 1981 erschienenen *Standards for evaluation of educational programs projects and material*, sie haben also bereits eine längere Tradition. Da die Disziplin der professionellen Evaluation in den USA im internationalen Vergleich am weitesten entwickelt ist und somit in einer Vorreiterfunktion steht, haben diese Standards auch über die USA und den englischsprachigen Raum hinaus Verbreitung gefunden und Einfluss gewonnen. Seit 1999 liegt unter dem Titel *Handbuch der Evaluationsstandards. Die Standards des „Joint Committee on Standards for Educational Evaluation“* eine deutsche Übersetzung dieser Standards vor (Sanders/Beywl 2006). Diese bildet auch die Grundlage der 2002 erschienenen *Standards für Evaluation* der deutschsprachigen DeGEval (DeGEval Gesellschaft für Evaluation 2008, S. 2). Sowohl die Standards des *Joint Committee* als auch jene der DeGEval sind unabhängig von bestimmten Evaluationsansätzen. Ihr Anspruch ist es, in unterschiedlichsten Evaluationskontexten Anwendung finden zu können und gültig zu sein[12].

Die Standards für Evaluation der DeGEval

Die *DeGEval Gesellschaft für Evaluation* ist die deutschsprachige Fachgesellschaft für im Bereich der Evaluation tätige Personen und Organisationen mit unterschiedlichsten Hintergründen. Von selbständigen Evaluatoren über Politik- und Unternehmensberatungen, Stiftungen und Forschungseinrichtungen bis hin zu Hochschulen sind unterschiedlichste Gruppen aus Deutschland, Österreich und der Schweiz in der DeGEval vertreten. Somit haben die durch die DeGEval erarbeiteten und herausgegebenen *Standards für Evaluation* (DeGEval Gesellschaft für Evaluation 2008) besonderes Gewicht

[12] Speziell für den Bereich der Hochschulen im europäischen Raum wurden 2005 durch die European Association for Quality Assurance in Higher Education die *Standards und Leitlinien für die Qualitätssicherung im Europäischen Hochschulraum* veröffentlicht (European Association for Quality Assurance in Higher Education 2006). Diese Standards fokussieren zwar speziell den Sektor der Hochschulen, jedoch wurden bei ihrer Entwicklung die international am breitesten anerkannten und wichtigen Standards des Joint Committee nicht erkennbar berücksichtigt (Schmidt 2009, S. 168). Vielleicht ist es darauf zurückzuführen, dass sie in der Literatur zur Hochschulevaluation eine äußerst nachrangige Rolle spielen.

im deutschen Sprachraum. Die 25 Einzelstandards der DeGEval sind in vier Gruppen unterteilt: Standards zur Nützlichkeit, zur Durchführbarkeit, zur Fairness sowie zur Genauigkeit.[13] Diese vier Aspekte einer Evaluation sind es dementsprechend auch, die als bestimmend für die Qualität einer Evaluation angesehen werden[14]:

Nützlichkeit: „Die Nützlichkeitsstandards sollen sicher stellen, dass die Evaluation sich an den geklärten Evaluationszwecken sowie am Informationsbedarf der vorgesehenen Nutzer und Nutzerinnen ausrichtet." (DeGEval Gesellschaft für Evaluation 2008, S. 21) Die acht Nützlichkeitsstandards (bezeichnet durch das Kürzel *N*, gefolgt von einer Zahl zur exakten Benennung des Standards) beziehen sich auf Fragen der Identifikation der Beteiligten und Betroffenen (N1), auf die Klärung der Evaluationszwecke (N2), auf die Glaubwürdigkeit und Kompetenz des Evaluators/der Evaluatorin (N3), auf Auswahl und Umfang der Informationen (N4), auf die Transparenz von Werten (N5), die Vollständigkeit und Klarheit der Berichterstattung (N6), die Rechtzeitigkeit der Evaluation(N7) sowie auf Nutzung und Nutzen der Evaluation (N8).

Durchführbarkeit: „Die Durchführbarkeitsstandards sollen sicher stellen, dass eine Evaluation realistisch, gut durchdacht, diplomatisch und kostenbewusst geplant und ausgeführt wird." (DeGEval Gesellschaft für Evaluation 2008, S. 26) Zur Durchführbarkeit (Kürzel *D*) wurden drei Standards definiert: Angemessene Verfahren (D1), Diplomatisches Vorgehen (D2) und Effizienz von Evaluation (D3).

Fairness: „Die Fairnessstandards sollen sicher stellen, dass in einer Evaluation respektvoll und fair mit den betroffenen Personen und Gruppen umgegangen wird." (DeGEval Gesellschaft für Evaluation 2008, S. 28) Die fünf Fairnessstandards (Kürzel *F*) befassen sich mit den Themen Formale Vereinbarungen (F1), Schutz individueller Rechte (F2), Vollständige und faire Überprüfung (F3), unparteiischer

[13] Die Standards des Joint Committees sind gleichermaßen unterteilt. Die englischen Originalbezeichnungen lauten *utility*, *feasability*, *propriety* und *accuracy* (Sanders/American Association of School Administrators/Joint Committee on Standards for Educational Evaluation 1994).

[14] An dieser Stelle wird nur eine beispielhafte Auswahl der Standards vorgestellt. Eine digitale, vollständige Version der *Standards für Evaluation* ist über die Homepage der DeGEval (www.degeval.de) kostenfrei zu beziehen.

Durchführung und Berichterstattung (F4) und Offenlegung der Ergebnisse (F5).

Genauigkeit: „Die Genauigkeitsstandards sollen sicher stellen, dass eine Evaluation gültige Informationen und Ergebnisse zu dem jeweiligen Evaluationsgegenstand und den Evaluationsfragestellungen hervorbringt und vermittelt." (DeGEval Gesellschaft für Evaluation 2008, S. 31) Die neun Einzelstandards umfassende Gruppe der Genauigkeitsstandards (Kürzel *G*) fokussiert folgenden Bereiche: Beschreibung des Evaluationsgegenstandes (G1), Kontextanalyse (G2), Beschreibung von Zwecken und Vorgehen (G3), Angabe von Informationsquellen (G4), Valide und reliable Informationen (G5), Systematische Fehlerprüfung (G6), Analyse qualitativer und quantitativer Informationen (G7), Begründete Schlussfolgerungen (G8) und schließlich Meta-Evaluation (G9).

Bei den Standards für die Durchführung von Evaluationen handelt es sich um Maximalstandards. Das bedeutet, dass die Standards den Entwurf einer best practice, gewissermaßen einer idealen Evaluation darstellen, an der sich Evaluatoren orientieren und die sie anstreben sollen (Balzer 2005, S. 74, Stockmann 2007a, S. 67). Dabei ist es in der Praxis kaum möglich, tatsächlich alle Standards komplett zu erfüllen. Dies sieht das Konzept der Maximalstandards allerdings auch nicht zwingend vor[15] und unterscheidet sich dadurch vom Konzept der Minimalstandards, bei dem stets alle Standards zu erfüllen sind. Es kann sehr gute Gründe geben, von der beschriebenen best practice abzuweichen, diese müssen allerdings nachvollziehbar dargelegt werden können. Somit dienen die Standards für die Durchführung von Evaluationen nicht nur als Qualitätssicherungs-, sondern auch als Katalysationsinstanz (vgl. Balzer 2005, S. 76): Die Qualität von Evaluationen lässt sich anhand der Standards erkennen und muss diskutiert werden, es soll eine Evaluationskultur entstehen. Darüber hinaus geben sie konkrete Hilfen bei der Planung und Durchführung von Evaluationen und unterstützen

[15] Im Einführungsteil der deutschen Übersetzung findet sich folgender Abschnitt: „Das Joint Committee ist sich bewußt, daß nicht alle Standards in allen Evaluation gleichermaßen anwendbar sind. Es muß fachkundig entschieden werden, welche Standards für die jeweilige Situation anwendbar sind. Die Nutzer der Standards sollten die Sachdienlichkeit jedes Standards für einen bestimmten Kontext sorgfältig prüfen und daraufhin entscheiden, welchen Standards die größte Bedeutung beigemessen werden soll. Diese Entscheidungen sollten schriftlich festgehalten werden, damit später darauf Bezug genommen werden kann." (Sanders/Beywl 2006, S. 27)

dabei, hohe Qualität zu erreichen. Ziel der Standards ist es also, zu einer optimalen Anpassung der Evaluation an die spezifischen Gegebenheiten beizutragen und somit eine *best fitted practice* zu ermöglichen. Eine dogmatische Anwendung der Standards führt nicht zum Ziel, der konkrete Anwendungskontext muss immer beachtet werden (Balzer 2005, S. 78).

Die konzeptionelle Unabhängigkeit der Standards von bestimmten Evaluationsansätzen bringt den großen Vorteil mit sich, dass die Standards unter verschiedenen, auch sich verändernden Evaluationsbedingungen und -rahmen zur Anwendung kommen können. Allerdings sind sie notwendigerweise, um der Konzeption der Übertragbarkeit zu entsprechen, recht offen formuliert. Um dennoch die konkrete Anwendung zu erleichtern, enthalten sowohl die englischsprachige Originalausgabe (Sanders/American Association of School Administrators/Joint Committee on Standards for Educational Evaluation 1994) als auch die deutsche Übersetzung (Sanders/Beywl 2006) eine Fülle von Material, so etwa Fallbeispiele und Hinweise auf typische Fehlerquellen zu jedem der Standards. Auch zu den Standards der DeGEval ist ein eigener Band mit Erläuterungen, Definitionen, Anwendungsbereichen etc. verfügbar (DeGEval Gesellschaft für Evaluation 2008).

Trotz der unstrittigen Wichtigkeit der Idee der Standards und der vielfältigen Handreichungen muss Balzer feststellen, dass die *Standards für Evaluation* bei vielen in der Praxis Tätigen noch gar nicht bekannt sind, so dass viele Evaluationsprojekte ohne die explizite Beachtung der Standards realisiert werden (Balzer 2005, S. 77/78).

2.2. Lehrevaluation an deutschen Hochschulen

Evaluation der Lehre gehört an deutschen Hochschulen zunehmend zum Alltag von Studierenden und Lehrenden. Dabei tritt Evaluation in unterschiedlichen Formen auf. Das Spektrum reicht von Studierendenbefragungen über summative oder formative interne Evaluationen bis hin zu Akkreditierungsverfahren.[16] Unterschiedlichste Indikatoren können als Bewertungsgrundlage herangezogen werden. Oft erfolgt die Lehrevaluation

[16] Uwe Schmidt weist in seinem Aufsatz darauf hin, dass einem strengen Verständnis von Evaluation längst nicht alle der zum Einsatz kommenden Verfahren genügen (Schmidt 2009, S. 165).

dabei im Kontext des Qualitätsmanagements[17] eingebunden und soll einen Beitrag zur Feststellung, Erhöhung und/oder Sicherung der Güte universitärer Lehre leisten. In dieser Rolle ist Lehrevaluation naturgemäß einer Vielzahl von Interessen und Ansprüchen ausgesetzt, die je nach Stakeholdergruppe variieren können. Kromrey identifiziert drei generelle, zentrale Funktionen (Paradigmen), die hinter Evaluationen stehen können und die in Lehrevaluationen fast immer gemeinsam auftreten: Forschung, Kontrolle und Entwicklung (Kromrey 2006b, S. 238). Bei Lehrevaluationen wird das Forschungsparadigma bei dem Wunsch, empirisches Wissen über erfolgreiches Lehren und Lernen zu sammeln, sichtbar. Das Kontrollparadigma scheint in Form der Frage nach einer effektiven und effizienten Verwendung der Geldmittel auf und das Entwicklungsparadigma schließlich kommt durch das Ziel der Verbesserung der Qualität von Lehre und Studium zum Tragen. Stockmann ergänzt diese Liste um ein weiteres Paradigma, nämlich das der Legitimation (Stockmann 2006a, S. 20). Legtimierende Evaluationen sollen einen Nachweis über das Verhältnis von Input zu Output sowie über Wirkungen in einer zeitlichen Perspektive erbringen, beispielsweise den Wirkungsgrad einer Maßnahme feststellen. Jede Stakeholdergruppe einer Universität – etwa Studierende, Lehrende, Verwaltungsmitarbeiter/-innen, Geldgeber – wird diese vier Paradigmen unterschiedlich gewichten.

Lehrevaluation findet nicht selten vor einem gesetzlichen Hintergrund statt. Spätestens durch den Bologna-Prozess, ein politisches Verfahren zur Schaffung eines einheitlichen europäischen Hochschulwesens, kann bzgl. des Einsatzes von Evaluationen im Kontext vorgeschriebener (Re-)Akkreditierungsverfahren eine deutliche Häufigkeitszunahme verzeichnet werden. Davon unabhängig ist Evaluation seit der Novelle des Hochschulrahmengesetzes im Jahre 1998 gesetzlich verankert und hat Eingang in die Mehrzahl der Hochschulgesetze auf Länderebene gefunden (Ernst 2008, S. 16), Schmidt 2009, S. 163). Neben diesen gesetzlich verankerten Evaluationen, gewissermaßen am anderen Ende einer fiktiven Skala der Organisationshierarchie, existieren auch jene Formen von

[17] *Qualitätsmanagement* bezeichnet die Gesamtheit der Maßnahmen, die eine Institution zur Sicherung und Steigerung der Qualität ihrer Produkte, Dienstleistungen etc. ergreift (Hense 2008, Beywl).

Evaluation, die auf Betreiben der unmittelbar durch die Qualität der Lehre betroffenen oder für sie verantwortlichen Personen – etwa Studierende oder Veranstaltungsleiter/-innen –, die ein vorrangiges Interesse an der tatsächlichen inhaltlichen Qualität einer Veranstaltung haben, initiiert und durchgeführt werden. Zwischen diesen Polen entfaltet sich ein weites Spektrum unterschiedlichster Ebenen, von denen ausgehend Lehrevaluationen initiiert werden können.

Neben der durch den Bologna-Prozess angestossenen Umstrukturierung der Hochschule ist ein vielleicht noch gewichtigerer Hintergrund der Lehrevaluation eine generelle Wandlung der Steuerung im öffentlichen Verwaltungssektor, dem auch die Hochschulen letztlich angehören, die sog. *Neue Steuerung* bzw. das *Neue Steuerungsmodell*. Im Rahmen dieses Modells, das eng mit dem internationalen Trend des *New Public Management* verwandt ist, erfolgt die Steuerung über *Produkte*, also bei Hochschulen über Vorgaben bzgl. Forschung und Lehre, und nicht mehr wie vorher über Mittelzuweisungen: Die Inputsteuerung wird durch eine Outputsteuerung ersetzt.

Dies ist, im Überblick dargestellt, der Hintergrund, vor dem Uwe Schmidt folgendes feststellt:

> „Hochschulevaluation erfährt in Deutschland insgesamt eine breite Akzeptanz in dem Sinne, dass ihre Existenz nicht grundlegend in Frage gestellt wird." (Schmidt 2009, S. 169)

Dieser grundlegenden Akzeptanz stehen allerdings auch ein gewisser Argwohn und eine Skepsis gegenüber, die sich v. a. gegen die möglichen Folgen von Evaluationen richten. Kritiker sehen in Evaluationen hochschulpolitische Kontroll- und Machtinstrumente, Ökonomisierungs- und Standardisierungsinstrumente mit subtilen Machteffekten, die in eine Art Ausscheidungskampf münden. Ebenso könnten falsch angegangene Evaluationen Motivation zerstören und Anreizverzerrungen[18] bedingen (Ernst 2008, S. 16). Um diesem Skeptizismus und den unerwünschten negativen Folgen zu begegnen, fordert Stefanie Ernst ein hohes Maß an Gegenstandsangemessenheit.

[18] Der Begriff *Anreizverzerrung* beschreibt den Effekt, dass ein äußerlicher Anreiz, in diesem Fall die Evaluation bzw. ihre erwarteten Folgen, Einfluss auf Entscheidungen und Handlungen der Betroffenen, hier der Evaluierten, ausübt, also etwa dazu führt, dass möglichst viele Ressourcen zu Lasten anderer Bereiche in den evaluierten investiert werden, um eine gute Bewertung zu erhalten.

Nachfolgend werden die einzelnen zuvor skizzierten Themen näher betrachtet. Nach einer knappen Darstellung der geschichtlichen Entwicklung der Hochschulevaluation wird der Wechsel der Steuerungsparadigmen näher betrachtet, um schließlich die Ziele, aber auch die Probleme der Hochschulevaluation zu skizzieren.

2.2.1. Abriss der historischen Entwicklung

Wie die gesamte Disziplin der Evaluation kann auch die Hochschulevaluation in den USA auf eine längere Geschichte zurückblicken als in Deutschland und Europa (vgl. Gollwitzer/Jäger 2007, S. 8 ff.; Grüner 1993, S. 31 ff.; Rindermann 2009, S. 30 ff.). Während die Bewertung der Qualität von Lehrveranstaltungen in den USA vor dem Hintergrund des amerikanischen Hochschulsystems zu sehen ist, das von den Studierenden ein deutlich höheres Maß an finanzieller Eigenbeteiligung verlangt als das deutsche und wo Evaluationen somit nicht zuletzt die Funktion eines Qualitätsnachweises haben, stehen die Anfänge der Diskussion um die Qualität der Lehre an deutschen Hochschulen zunächst in einem anderen Zusammenhang: Sie sind in der Zeit der Studentenbewegung der 1968er Jahre zu finden (Rindermann 2009, S. 32). Der Begriff *Evaluation* tauchte damals noch nicht auf, sondern es wurde von *Vorlesungsrezensionen* gesprochen. Unter diesem Titel wurden in studentischen Blättern Besprechungen von Inhalt und Didaktik einzelner Veranstaltungen einzelner Dozenten durchgeführt[19] (Rindermann 2009, S. 32). Ziel dieser Rezensionen war es, einen Beitrag zur Demokratisierung und Umstrukturierung des Hochschulsystems zu leisten. Zwar etablierte sich diese Form der Bewertung der Qualität von Lehrveranstaltung nicht dauerhaft, sondern verschwand im Laufe der 1970er Jahre wieder, gleichwohl führte sie zumindest zu einer ersten Sensibilisierung der Wissenschaftler für Fragen der Qualität von Lehre, Studium und Hochschuldidaktik (vgl. Rindermann 2009, S. 31 ff.).

Während in den 1960er und 70er Jahren vor allem die Kritik an den Verhältnissen in der Hochschule die treibende Kraft hinter den Veranstaltungsrezensionen (bzw.

[19] Dieser Weg wurde in Anlehnung an Karl Marx beschritten, der gefordert hatte, die „versteinerten Verhältnisse dadurch zum Tanzen zu zwingen, dass man ihnen ihre eigne Melodie vorsingt“ (Karl Marx zitiert nach Rindermann 2009).

Evaluationen) war, hat die zweite Welle der Hochschulevaluation einen veränderten Hintergrund. Sie nahm ihren Anfang mit der Wiedervereinigung von West- und Ostdeutschland. In diesem Kontext stand die Frage im Raum, welche Hochschulen und Fächer weitergeführt werden sollten und in welcher Weise[20]. Evaluationen sollten helfen, diese Frage fundiert zu beantworten (Schmidt 2009, S. 163). Neben diesem politisch-organisatorisch motivierten Interessensstrang lassen sich noch zwei weitere identifizieren. Zum einen gibt es ein gesellschaftliches oder doch zumindest öffentliches Interesse an der Qualität der Hochschullehre, das etwa in Form der seit Beginn der 1990er Jahre regelmäßig in Magazinen wie Spiegel, Focus oder Stern erscheinenden Universitätsrankings Ausdruck findet, zum anderen haben die Studierenden einer Hochschule selbst ein vitales Interesse an qualitativ hochwertiger Lehre, das z.B. durch studentische Evaluationsinitiativen, etwa über Fachschaften oder in Form von Mitarbeit in universitären Gremien Ausdruck findet (vgl. Rindermann 2009, S. 33 ff.).

Die drei Stränge dieser zweiten Evaluationswelle sind letztlich vor den Anforderungen gesellschaftlicher Rechenschaftslegung, Qualitäts- und der Effektivitätssteigerung, aber selbstverständlich auch im Kontext einer zunehmend prekären Finanzsituation zu sehen. Auch die zunehmende Internationalisierung der Hochschulen und die Konkurrenz zwischen ihnen spielen ein Rolle (Rindermann 2009, S. 34 ff.). Hier wird auch der zentrale Unterschied zwischen den ersten Evaluationsbestrebungen während der Studentenbewegung und denen der zweiten Welle deutlich: Während erstere, ganz im Sinne ihrer Intention als Protest und Katalysator der Auseinandersetzung mit der Hochschule, vor allem von inneruniversitärem Interesse waren, haben letztere eine gesamtgesellschaftliche Reichweite.

Ein weiterer bedeutsamer Unterschied, der nicht nur das staatliche Interesse an Evaluation im Hochschulsektor unterstreicht, sondern auch ein Garant dafür ist, dass sie dauerhaft durchgeführt wurde bzw. wird, ist die Tatsache, dass Evaluation mittlerweile in vielen die Hochschule betreffenden Gesetzen verankert ist. Seit der Novelle des Hochschulrahmengesetzes im Jahre 1998 ist Evaluation dort gesetzlich verankert

[20] Dieser Umstand hat neben der ohnehin bereits bestehenden kritischen Haltung Anteil an der bis heute vorhandenen negativen Konnotation von Evaluation im Hochschulbereich.

(Schmidt 2009, S. 163). In der Folge ist Evaluation der Hochschullehre mittlerweile auch in zahlreichen Ländergesetzen festgeschrieben und damit verpflichtend geworden (Ernst 2008, S. 16). Außerdem erhöht sich durch die Akkreditierung von Bachelor- und Masterstudiengängen zunehmend die rechtliche Verbindlichkeit von Qualitätssicherungsverfahren (Schmidt 2009, S. 163).

Zwar unterscheiden sich die Hintergründe der beiden beschriebenen Phasen der Evaluation der Hochschullehre teilweise grundlegend, v.a. hinsichtlich ihres ideellen Hintergrundes, ihrer Reichweite sowie des Grades der Formalisierung, dennoch dienen beide Ansätze dem gleichen Zweck, der nicht nur für die Studierenden, sondern für die gesamte Gesellschaft von Belang ist: sie sollen dazu beitragen, die Hochschulausbildung so gut wie möglich zu machen. Die zweite Phase ist indes auch in einem anderen grundlegenden Kontext zu sehen, der das Thema Evaluation an Hochschulen nach vorne gebracht hat, allerdings nicht primär der Idee der Verbesserung der Lehre entstammt: dem Wechsel der Steuerungsparadigmen von Kommunen, Einrichtungen des Öffentlichen Dienstes und somit auch der Hochschulen.

2.2.2. Hintergrund: Steuerungsparadigmen

Reform der Verwaltungssteuerung: das Neue Steuerungsmodell und das klassisch-bürokratische Verwaltungsmodell

In Folge der in der Mitte der 1980er Jahre begonnen Diskussion um die Erneuerung und Modifikation der kommunalen Verwaltungen im Sinne des *New Public Management* legte im Jahr 1993 die *Kommunale Verwaltungsstelle für Verwaltungsvereinfachung* (KGSt, seit November 2005 *Kommunale Verwaltungsstelle für Verwaltungsmanagement*) die erste Skizze des *Neuen Steuerungsmodells* (NSM) vor (Kommunale Gemeinschaftsstelle für Verwaltungsmanagement (KGSt) 1993). Dabei handelt es sich um eine deutsche Variante des *New Public Management*-Ansatzes. Das Neue Steuerungsmodell ist ein Modell der Verwaltungswissenschaft, dessen zentrales Ziel die Flexibilisierung kommunaler Verwaltung ist, um sie an sich ändernde gesellschaftliche Rahmenbedingungen anzupassen. Wesentlich ist dabei die Abkehr von der Steuerung der

Verwaltung über den Input, also die zur Verfügung gestellten Ressourcen, hauptsächlich die Haushaltsmittel. Stattdessen erfolgt die Steuerung im Rahmen des NSM über den Output, d. h. über konkrete Zielsetzungen oder Produkte. Praktisch erfolgt diese Steuerung über Vereinbarungen, sog. *Kontrakte*, die zwischen der Verwaltungsleitung und den jeweiligen Verwaltungseinheiten geschlossen werden.

Hauptantrieb für die Umgestaltung der Verwaltungssteuerung waren durch sich ändernde wirtschaftliche Rahmenbedingungen und die Kosten der deutschen Einheit bedingte zunehmende finanzielle Probleme, die eine Veränderung der Verwaltungsstrukturen und ein effizienteres, ökonomischeres Handeln erforderlich machten. Außerdem wuchs die Komplexität der Aufgaben, mit deren Verrichtung kommunale Verwaltungen betraut sind. Gleichzeitig forderten die Bürger eine stärker kundenorientierte Verwaltung (vgl. etwa Bogumil 1998, S. 83 ff., Kracht 2006, S. 22, Riese 2007, S. 10). Darüber hinaus identifiziert Löffler weitere übergeordnete Entwicklungen, in deren Kontext die Veränderung der Verwaltungsorganisation zu sehen ist, bzw. die sie begünstigten. Dazu gehören vor allem der Wandel von einer Welt der Stabilität zu einer dynamischen, interdependenten und komplexen Welt, der Wertewandel von traditionalen Pflicht- und Akzeptanzwerten hin zu Werten der Selbstverwirklichung sowie die Änderung des relativen Preises von Informations- und Kommunikationstechnologie (Löffler 1998, S. 339). Das NSM stellt nun den Versuch dar, eine bürokratisch organisierte Verwaltung zu flexibilisieren, um diesen Anforderungen und Entwicklungen entsprechen zu können. Damit einher geht ein gänzlich neues Verständnis von Verwaltung:

> „Während sich die Bürokratie im Weberschen Sinne als möglichst stabiler, mechanisierter Apparat darstellte, etabliert sich heute ein Verständnis von der Verwaltung als dynamisch komplexes soziales Gebilde.“ (Schedler/Proeller 2006, S. 18)

Das Referenzmodell, vor dem das NSM zu sehen ist, ist das klassisch-bürokratische Modell von Staat und Verwaltung, das auf die Stein-Hardenbergische Reformen[21] zurückgeht (Naschold/Bogumil 2000, S. 84). Innerhalb dieses Modells ist die Verwaltung streng hierarchisch organisiert. Die Steuerung der Abläufe und Tätigkeiten erfolgt über

[21] Als Stein-Hardenbergische Reformen werden ein eine Reihe von Staats- und Verwaltungsreformen des preußischen Staates bezeichnet, die zwischen 1806 und 1822 durchgeführt wurden.

festgelegte Regeln (Verwaltungshandeln als rechtsstaatliche Regelanwendung). Ein strategisches Management, das die allgemeine Ausrichtung der Verwaltung überwacht und ggf. beeinflusst, gibt es nicht. Es existiert eine strenge funktionale Arbeitsteilung nach dem Verrichtungsprinzip, d. h. nach der Zuständigkeit für einen speziellen, definierten Aufgabenbereich. Kooperationen zwischen verschiedenen Arbeitsbereichen haben dabei oftmals mit Koordinations- und Schnittstellenproblemen zu kämpfen (Naschold/ Bogumil 2000, S. 87).

Dieses klassische Verwaltungsmodell, so das weitgehend übereinstimmende Urteil, hatte die Grenzen seiner Leistungsfähigkeit überschritten (Naschold/Bogumil 2000, S. 86): Die zunehmend komplexen Aufgaben konnten angesichts einer internationalen Wettbewerbsökonomie und der Qualitäts- und Demokratisierungsforderungen von Seiten der Bürger und Kunden mit den organisatorischen und personalwirtschaftlichen Potenzialen immer weniger bearbeitet werden. Über dieses Metaurteil hinaus benennt Stefan Kracht folgende Hauptkritikpunkte am bürokratischen Steuerungsmodell (Kracht 2006, S. 53 ff.):

- Steuerungsmängel, wie z.B. die wenig effiziente Mittelverwaltung der Kameralistik, etwa der Mittelverfall bei sparsamer Wirtschaft, oder das Auseinanderfallen von Erfüllung der Sachaufgabe und der Verantwortung dafür,
- Planungsdefizite, z.B. eine fehlende strategische Planung oder fehlende Anreize zur Leistungsverbesserung,
- Ausufernde Bürokratisierung als systemimmanenter Bestandteil des hierarchischen Steuerungsmodells sowie eine
- Fehlende Kosten-/Leistungstransparenz, da etwa durch bestimmte Maßnahmen entstehende Kosten nicht erfasst werden und Entscheidungsträger somit bei der Inputorientierung nicht wissen, welche Kosten durch ihre Entscheidungen entstehen, wodurch auch kein effizienter Anreiz zur effizienten Mittelverwendung besteht.

Das NSM soll diesen zentralen Problem- und Kritikpunkten durch die Einführung von Managementkonzepten, die der Privatwirtschaft entlehnt sind, begegnen. Verwaltungs-

handeln wird dabei als Dienstleistung verstanden. Unter der Bezeichnung Neues Steuerungsmodell wird demnach eine Fülle von Einzelmaßnahmen zusammengefasst. Folgende Elemente stehen dabei im Vordergrund (vgl. Kracht 2006, S. 86ff., Kommunale Gemeinschaftsstelle für Verwaltungsmanagement (KGSt) 1996, S. 69 ff.):

- Dezentralisierung zum Zwecke der Verschlankung und Entbürokratisierung, etwa:
  - Verlagerung der Verantwortlichkeiten von oben nach unten, Zusammenführung von Fach- und Ressourcenverantwortung,
  - Trennung von Politik (in der Funktion der strategischen Verantwortung, zuständig für das *Was*) und Verwaltung (operative Verantwortung, zuständig für das *Wie*),
  - dezentrale Budgetverantwortung.

- Contracting: Kontrakte sind verbindliche Leistungsvereinbarungen, die zwischen Verwaltungsverantwortlichen und Verwaltungseinheiten geschlossen werden. Das Schließen von Kontrakten soll den Steuerungsverlust der ehemaligen Verwaltungsspitze ausgleichen, indem die Anbindung der dezentralen Einheiten in den Gesamtorganismus gewährleistet wird. Das Contracting ist das zentrale Lenkungsinstrument des NSM.

- Outputsteuerung: Jeder Verwaltungsteil erbringt bestimmte, in den Kontrakten vereinbarte Leistungen. Diese Leistungen werden mit den eingesetzten Ressourcen verknüpft, beispielsweise über die dezentral verwalteten Budgets, um den Preis der Leistung erkennbar zu machen.

- Controlling: Das Controlling soll im Sinne eines Soll-Ist-Vergleichs die Erreichung der in den Kontrakten vereinbarten Leistungen überprüfen. Außerdem soll es bereits während eines Haushaltsjahres Kostentransparenz erzeugen (nicht wie vorher erst nach seinem Abschluss) und so entscheidungsrelevante Daten bereitstellen. Somit kommt dem Controlling eine zentrale Position innerhalb des NSM zu, Löffler bezeichnet es gar als Herzstück des NSM (Löffler 1998, S. 346).

KBR	NPM
Steuerung über Regeln	Steuerung durch Ziele/Ergebnisse (Performanz-Management inkl. Ergebnisbudgetierung und Plafondierung)
Funktionale Arbeitsteilung nach dem Verrichtungsprinzip mit prozesskettenbezogenen Schnittstellenproblemen in der Kooperation	Produktbezogene Organisation in Form einer Prozesskette
Ausgeprägte Hierarchisierung	Kontraktmanagement verselbständigter Ergebniseinheiten
Geringer Einsatz von Wettbewerbsinstrumenten	Aufgabenauslagerung und Bildung von (Quasi-) Märkten
Fehlen eines strategischen Managements	Kundenorientierung

Tabelle 2.1.: Vergleich der Steuerungsmodelle nach Naschold/Bogumil 2000, S. 87

- Kundenorientierung: Die gesamte Verwaltung richtet sich stärker auf die Interessen der Bürger, anstatt sich wie zuvor an den internen bürokratischen Abläufen zu orientieren. Der Bürger wird gewissermaßen als Kunde gesehen, dem gegenüber hochwertige Dienstleistungen erbracht werden sollen. In diesem Kontext erlangen auch Modelle zur Qualitätssicherung Bedeutung.

- Bildung von Märkten bzw. Quasimärkten: Durch die Einführung von Wettbewerb in das bisher quasi wettbewerbsfreie System sollen ökonomische Prinzipien gestärkt werden. Wettbewerber der öffentlichen Verwaltung sind sowohl private Unternehmen (beispielsweise im Sektor der Abfallwirtschaft) als auch andere Verwaltungsträger. Teilweise steht die Verwaltung auch mit fiktiven privaten Dienstleistern im Wettbewerb (Kracht 2006, S. 74).

Im direkten Vergleich stellen sich die wesentlichen Unterschiede zwischen dem klassisch-bürokratischen Regulierungsmodell (KBR) und einem New Public Management-Regulierungsmodell (NPM), somit also auch dem NSM, dar, wie in Tabelle 2.1 gegenübergestellt.

Neue Steuerung an Hochschulen

Während die Konzepte des New Public Management und später der Neuen Steuerung im Feld der Kommunalverwaltungen Mitte der 1980er Jahre Gegenstand der Diskussio-

nen wurden, sind sie im Bereich der Hochschulen etwa zehn Jahre später ins Blickfeld gerückt (Bogumil/Heinze 2009, S. 7). Hauptsächlicher Antrieb für die Bestrebung, die bestehenden Lenkungssysteme zu verändern, war die angespannte finanzielle Situation der Hochschulen (vgl. Serrano-Velarde 2008, Riese 2007, Kracht 2006). Dem entsprechend sollen die ergriffenen Maßnahmen v.a. dazu beitragen, Effizienz und Effektivität der Leistungserbringung zu steigern. Allerdings war das nicht die einzige Funktion. Auch die Steigerung der Qualität von Forschung und Lehre und ein verstärkter Zuschnitt auf die Bedürfnisse der Kunden, also einerseits der Studierenden, andererseits der späteren Arbeitgeber der Absolventen, waren Gründe für die Einführung des NSM.

Zwar wurde das NSM nicht für die Hochschulsteuerung konzipiert, dennoch liegt eine Übertragung des Modells von Kommunalverwaltungen auf andere Institutionen des öffentlichen Dienstes nahe, v.a. wenn Probleme und grundlegende Strukturen ähnlich sind. Den Ideen des NSM entsprechend setzen Wissenschaftsministerien und Hochschulen folglich Maßnahmen um wie die Stärkung der Hochschulautonomie (= Dezentralisierung), Globalbudgets[22] (= dezentrale Budgetverantwortung) oder Zielvereinbarungen (= Contracting)[23] (Bogumil/Heinze 2009, S. 7). Gleichwohl handelt es sich bei den bisher erfolgten Umsetzungen um kein in sich geschlossenes Konzept, sondern es werden in Hochschulen üblicherweise – so, wie im Bereich der Kommunalverwaltungen häufig auch – hauptsächlich Einzelelemente des Maßnahmenbündels NSM in unterschiedlichen Kombinationsvariationen auf das Verwaltungshandeln übertragen.

Folgende Elemente des NSM spielen für die Umsetzung an Hochschulen in der Praxis entscheidende Rollen (Kracht 2006, S. 118 ff.):

- Dezentralisierung, v. a. die Lösung von der Fremdsteuerung durch das Ministerium (innerhalb der Hochschule führt dies i. d. R. zu einer Stärkung der Hochschulleitung und eher zu einer Schwächung ihrer dezentralen Struktur),

[22] Als Globalbudget wird ein Verfahren der Mittelbereitstellung bezeichnet, bei dem nur ein pauschaler Betrag je Einheit bewilligt wird. Über dieses Budget kann die Verwaltungseinheit, in diesem Falle die Hochschule, weitestgehend frei entscheiden und ihn nach eigenen Regeln intern weiter verteilen.

[23] Bogumil und Grohs weisen darauf hin, dass bei der Einführung von NSM-Elementen an Hochschulen kaum auf die Erfahrungen zurückgegriffen wurde, die bei seiner Einführung in kommunalen Verwaltungen gewonnen werden konnten, obwohl durchaus robuste Erkenntnisse vorlägen (Bogumil/Grohs 2009, S. 139). Kracht weist darauf hin, dass bisher kein empirischer Erfolgsnachweis bzgl. des NSM geführt wurde (Kracht 2006, S 86).

- Zielvereinbarungen, zunächst auf der Ebene Hochschule – Ministerium, danach hochschulintern bis auf Fachbereichsebene,
- Outputsteuerung, also die Koppelung des Mittelflusses an konkrete Leistungen,
- Controlling, die Überwachung der Zielerreichung,
- Wettbewerb in Gestalt des Wettbewerbs um Studierende und als Wettbewerb um finanzielle Zuweisungen,
- Qualitätsmanagement, d. h. die Bestrebung, die Qualität weiter zu verbessern, v.a. durch Akkreditierungsverfahren, Evaluation von Lehrangeboten, u. U. mit Fokussierung der Kundensicht durch eine starke Beteiligung der Studierenden.

Trotz der mitunter recht unterschiedlichen Schwerpunktsetzungen lässt sich eine gewisse Modellvorstellung der modernisierten Hochschulsteuerung erkennen:

> „Im Kontext des 'Neuen Steuerungsmodells' gelten Zielvereinbarungen als das entscheidende Mittel, um die alte, der kameralistischen Verwaltungsführung entsprechende 'Inputsteuerung' der Hochschulen auf die sogenannte 'Outputsteuerung' umzustellen." (Stock 2004, S. 31)

Somit avanciert das Contracting zum zentralen Steuerungsinstrument (Kracht 2006, S. 21). Nun führt das bloße Schließen von Zielvereinbarungen natürlich noch nicht zum Ziel, sondern es bedarf des Controllings, um die Einhaltung der Zielvereinbarungen zu überprüfen. Ein wesentliches Mittel zur Überprüfung der Zielerreichung ist die Evaluation[24] im Sinne eines Soll-Ist-Vergleichs, die somit in der Funktion des Controllings neben dem Contracting das Herzstück des NSM bildet (Löffler 1998, S. 346).

An dieser Stelle wird eine der zentralen Fragen des NSM an Hochschulen sichtbar, nämlich ob sich die Leistungen, die eine Hochschule erbringt, ohne weiteres in überprüfbare Ziele übertragen lassen, ob ihre Dienstleistungen in steuerungstaugliche Produkte transformiert werden können (Kracht 2006, S. 78). Denn nur dann, wenn die Ziele klar erkennbar formuliert und so beschaffen sind, dass ihre Erreichung auch tatsächlich

[24] Neben der Evaluation kommt auch das Verfahren der Akkreditierung zum Einsatz, das die Einhaltung von Mindeststandards überprüft.

überprüft werden kann, können Evaluationen zu aussagekräftigen Ergebnissen kommen. Es wird deutlich, dass der Erfolg der Zielvereinbarungen (und damit auch der des NSM) maßgeblich von einer trennscharfen Zielbestimmung abhängt: nur wenn die Kontrakte tatsächlich überprüfbare Ziele enthalten, ist auch eine Zielkontrolle möglich (Kracht 2006, S. 78).

Gemeinsam mit den in Kapitel 2.2.1 (*Abriss der historischen Entwicklung*) dargelegten Zielen der Verbesserung der Lehre und der Rechenschaftslegung gegenüber der Gesellschaft hat die hier dargestellte Rolle von Evaluationen im Rahmen des Neuen Steuerungsmodells wesentlich dazu beigetragen, dass sich Evaluation als Verfahren an den Hochschulen etabliert hat.

2.2.3. Ziele

Mit Evaluationen werden, wie sich anhand des bereits Dargestellten erkennen lässt, je nach Kontext sehr unterschiedliche Zielsetzungen verbunden. Im Wesentlichen lassen sich die Ziele von Evaluationsaktivitäten nach drei verschiedenen Bereichen unterscheiden:

1. Hochschulinterne Qualitätsentwicklung und -sicherung,

2. Controlling und Rechenschaftslegung gegenüber Ministerien sowie

3. Rechenschaftslegung bzw. Qualitätsnachweis gegenüber gegenwärtig Studierenden, Studieninteressenten und der Gesellschaft.

Diese Bereiche schließen sich nicht gegenseitig aus, da die jeweiligen Erkenntnisinteressen nicht unbedingt grundsätzlich verschiedenartig sind. Letztlich geht es, gemäß der Definition der DeGEval, immer um die Bestimmung des Wertes eines Evaluationsgegenstandes. So ist die Bewertung der Qualität der Lehrveranstaltungen eines Fachbereichs eben nicht nur für die Lehrenden oder die mit der Sicherung der Lehrqualität Beauftragten von Interesse, sondern gleichzeitig eine für die nach außen gerichtete Rechenschaftslegung interessante Information. Die Unterscheidung der genannten Bereiche

ergibt sich demnach vielmehr aus der vorrangigen Nutzung der Evaluationsergebnisse, also den konkreten Zielen, zu deren Erreichung die Evaluationsergebnisse beitragen sollen.

Innerhalb eines jeden dieser drei Zielbereiche wird zur Beantwortung der spezifischen Fragen als Informationsbasis auf die gleichen Evaluationsgegenstände zurückgegriffen. An Hochschulen handelt es sich dabei in erster Linie um die drei folgenden:

1. Konkrete Lehrveranstaltungen,
2. Lehrgebiete, etwa Lehrstühle, Institute oder Fachbereiche,
3. Die Organisation, und zwar sowohl in der Bedeutung von *Institution* oder *Infrastruktur* als auch im Sinne von *Planung und praktischer Durchführung.*

Der Blickwinkel auf diese Informationsbasis ändert sich dabei in Abhängigkeit vom Standpunkt der Auftraggeber und damit der intendierten Nutzung der Evaluationsergebnisse: Für das Controlling eines Fachbereichs sind andere Fragestellungen erforderlich als für die Verbesserung einer einzelnen Lehrveranstaltung. So werden innerhalb eines jeden Zielbereichs spezifische Fragen gestellt, die sich allerdings, wie bereits erwähnt, nicht unbedingt ausschließen, sondern mitunter überschneiden. Somit steht jede Evaluation in einer bestimmten Funktion. Folgende Funktionen oder Erkenntnisinteressen (von Kromrey als *Paradigmen* bezeichnet) lassen sich dabei, wie in Kapitel 2.2 bereits angerissen, unterscheiden (vgl. Kromrey 2006b, Stockmann 2006b, Schmidt 2009):

Entwicklung bzw. Lern-/Dialogfunktion: Die sicher klassischste Funktion von Evaluationen an Hochschulen ist, wie in Kapitel 2.2.1 dargestellt, die Verbesserung und somit Weiterentwicklung einer Lehrveranstaltung[25] oder, allgemeiner gesprochen, des Evaluationsgegenstandes. Die Bewertungen werden hierbei als Ausgangspunkt der Identifikation von individuellen Stärken und Schwächen des Gegenstandes genutzt, um aus ihnen im Idealfall Ideen für eine Veränderung abzuleiten. Die Ergebnisse können gleichzeitig die Grundlage eines Austauschs zwischen

[25] Dieses Ziel ist vermutlich im gesamten Bildungssektor, auch außerhalb der Hochschule, eines der verbreitetsten.

den Stakeholdergruppen sein.[26] Dabei ist dieses Ziel der Entwicklung keineswegs auf einzelne Lehrveranstaltungen begrenzt, sondern kann ohne weiteres auf größere Einheiten wie etwa Curricula oder ganze Fachbereiche bezogen sein.

Kontrolle: Die Kontrollfunktion von Evaluationen ist jene, die durch die Implementation des Neuen Steuerungsmodells stark an Bedeutung gewonnen hat. Gleichzeitig ist sie wohl auch die am negativsten konnotierte. In dieser Funktion liefern Evaluationen Informationen über die Aufgabenerfüllung von Personen innerhalb des Evaluationsgegenstandes, über die Verwendung von Geld- und Sachmitteln und dergleichen mehr, kurz: Sie dienen letztlich der Überwachung der Aufgabenerfüllung und Sachorientierung. Sie leisten somit einen bedeutenden Beitrag zur Erzeugung von Transparenz, führen aber auch schnell zu Unbehagen auf Seiten der Kontrollierten. Innerhalb dieser Funktion sind Lehrberichte das entscheidende Element der Lehrevaluation.

Forschung bzw. Erkenntnisfunktion: Evaluation kann in soweit Forschungsfunktionen erfüllen, als dass sie grundlegendes empirisches Wissen über Bedingungen erfolgreicher Arbeit hervorbringen kann. Dabei kann es sich etwa um Bedingungen des Lehrens und Lernens, die Akzeptanz von Maßnahmen oder auch um die Rahmenbedingungen und ihre Veränderung handeln. Dieses Wissen wiederum kann für sich stehen und die gesicherte Wissensbasis verbreitern, aber beispielsweise auch zur Weiterentwicklung oder zur Entscheidungsfindung herangezogen werden.

Legitimation: In der Rolle eines legitimatorischen Instrumentes können Evaluationen etwa genutzt werden, nachzuweisen, mit welchem Einsatz (Input) welche Wirkungen (Output) erzeugt werden oder welche Wirkungen eine Maßnahme über die Zeit entfaltet. Die Perspektive ist somit in der Regel eine rückwärtsgewandte. Legitimierende Evaluationen dienen beispielsweise als Beleg für oder gegen die Effizienz einer Maßnahme oder fungieren als Nachweis des erreichten Wirkungs-

[26] Dieser Aspekt wird auch durch die Hochschulrektorenkonferenz betont: „Wesentlich ist, dass Evaluationen zur Verbesserung der internen Kommunikation zwischen Lehrenden und Studierenden beitragen." (Hochschulrektorenkonferenz 1998, S. 10)

grades eines Programms.

Eine weitere Funktion, die Stockmann über die hier dargestellten hinaus anspricht, ist die *taktische Funktion*, in der Evaluationen dazu dienen sollen, politische Entscheidungen (ggf. auch nachträglich) zu legitimieren, gleichzeitig seitens der Auftraggeber aber kein weiteres echtes Interesse an der Nutzung der Ergebnisse besteht. Diese Funktion bezeichnet Stockmann als die „eher pathologische Seite" der Evaluation, da sie sich kaum mit dem eigentlichen Zweck von Evaluationen vereinbaren lässt (Stockmann 2006b, S. 21).

Das eigentliche, übergeordnete Ziel von Evaluationen an Hochschulen, das alle dargestellten (Teil-) Ziele in sich vereint, lässt sich – angelehnt an Münch und Pechmann (Münch/Pechmann 2009) – so umschreiben: Es geht darum, die geleistete Arbeit und ihre Qualität sichtbar zu machen sowie die Evaluationsgegenstände weiterzuentwickeln.

2.2.4. Probleme

Lehrevaluationen haben sich in unterschiedlicher Form in der Hochschule etabliert. Das bedeutet allerdings nicht, dass ihre Durchführung frei von Problemen oder sie nicht Gegenstand von Debatten wären. Im Gegenteil, die Liste der diagnostizierten Probleme, der zu diskutierenden Aspekte ist lang und betrifft sehr unterschiedliche Bereiche der Evaluation. Dennoch handelt es sich meist um keine unlösbaren Schwierigkeiten, sondern um solche, denen oft mit Maßnahmen wie erhöhter Transparenz der Prozesse, fairem Umgang mit den Stakeholdern und ihrer ehrlichen Berücksichtigung begegnet werden kann.

Schwerpunkte der Diskussion sind:

- Die Frage der Zuständigkeiten,
- negative Nebeneffekte von Evaluation,
- grundsätzliche Infragestellung von Evaluation,
- Nutzung der Evaluationsergebnisse,

- die Frage der Berücksichtigung der Stakeholder,
- unterschiedliche Ansprüche an Evaluationen sowie
- methodische Probleme.

Die jeweiligen Probleme und Schwierigkeiten werden hier nur in einem kursorischen Überblick dargestellt, denn zu jedem der Bereiche wird eine mehr oder weniger eigenständige und intensive Diskussion geführt. Die Zuordnung konkreter Probleme zu Problemfeldern ist außerdem nicht fix oder immer eindeutig, da Probleme oftmals eng miteinander verwoben sind. Am offensichtlichsten ist dies wohl bei der Frage nach der Berücksichtigung der Stakeholder, die in dieser Unterteilung als eigener Punkt erscheint, gleichzeitig aber auch alle übrigen Problemfelder direkt oder indirekt mitbetrifft. Dennoch soll diese Unterteilung helfen, einen Überblick über die wesentlichen Themen zu erlangen.

Die Frage der Zuständigkeiten

Das Problem der Zuständigkeiten ist in der einerseits zwar klar hierarchischen, andererseits jedoch hinsichtlich der Forschung und Lehre gleichzeitig autonomen Struktur der Hochschulen begründet: Die hierarchischen Strukturen reichen gewissermaßen nicht bis in den Bereich der konkreten inhaltlichen und didaktischen Gestaltung einer Lehrveranstaltung hinein. Somit ergibt sich das Problem der Zuständigkeiten in Form der Frage, wer wen mit welchen Zielen und Befugnissen evaluiert bzw. evaluieren sollte.

Innerhalb der Hochschulhierarchie können Lehrevaluationen prinzipiell von verschiedenen Ebenen ausgehen. Das Spektrum reicht von den Studierenden über zentrale Evaluationsstellen und übergeordnete staatliche Instanzen bis hin zu nicht direkt mit der konkreten Hochschule verbundenen Einrichtungen wie etwa dem Centrum für Hochschulentwicklung (CHE). Daraus ergeben sich einige der Probleme, auf die weiter unten eingegangen wird, nämlich einerseits, dass die Auftraggeber einer Evaluation nur in bestimmten Fällen Einfluss auf die konkrete Gestaltung der Lehre – i. S. v. Verbesserung, basierend auf den Ergebnissen der Bewertung – ausüben können. Den direktesten Einfluss auf die Lehre haben hier sicherlich die unmittelbar an einer Lehrveranstaltung

beteiligten Personen, also die Studierenden und Lehrenden. Daraus wiederum ergeben sich die Fragen, welche Methoden und Verfahren zweckmäßig und angemessen sind, wer Einblick in die Ergebnisse erhält und welche weiteren Ziele, Folgen und Wirkungen mit den durchgeführten Evaluationen für die Stakeholder verbunden sein könnten.[27] In der Folge ergeben sich schnell die bereits erwähnten Widerstände gegen eine Evaluation[28] und die eventuell aus ihr abgeleiteten Maßnahmen. Ein Problem entsteht in diesem Bereich vor allem dann, wenn die unterschiedlichen Gruppen wenig in den Evaluationsprozess einbezogen sind.

Negative Nebeneffekte von Evaluation

Jede Evaluation hat Nebeneffekte, die außerhalb des Rahmens ihrer definierten Zielsetzungen liegen. Einige dieser Nebeneffekte sind unproblematisch oder vielleicht sogar im Sinne der Evaluation, andere hingegen sind unerwünscht und mitunter gar kontraproduktiv. Folgende negative Nebeneffekte werden mit Blick auf Lehrevaluationen hauptsächlich diskutiert:

- **Anreizverzerrung** (Ernst 2008, S. 16, Frey 2006): Anreizverzerrung beschreibt den bereits in Kapitel 2.2 angesprochenen Effekt, dass der Anreiz zur Leistungserbringung nicht mehr innerhalb der eigentlichen, unmittelbaren Tätigkeit liegt, sondern von einem äußerlichen Anreiz, in diesem Falle also von einer Evaluation, ausgeht. Einerseits bedeutet dies die Verdrängung intrinsischer Anreize, denn das, was nicht gemessen wird, zählt auch nicht für die Bewertung. Eine Anreizverzerrung könnte also bedingen, dass der Arbeitseinsatz (und andere Ressourcen gleichermaßen) eher in die Dienstbarkeit einer möglichst positiven Bewertung gestellt als in die umfassende, gründliche Erfüllung der Aufgabe investiert wird. Andererseits besteht die Gefahr, dass Leistungskriterien oder ihre Erfüllung ma-

[27] Rindermann beispielsweise spricht sich gegen externe Maßnahmen im Sinne von Sanktionen und/oder Belohnung aus, um das Schönen von Berichten (siehe auch nächster Abschnitt) oder die Machterweiterung der Ministerialbürokratie zu verhindern (Rindermann 2009, S. 341).

[28] Bülow-Schramm weist etwa darauf hin, dass im Falle des Einsatzes von Gutachtern diese von den Begutachteten anerkannt werden müssen, weshalb sie ein Mitspracherecht bei der Gutachterwahl erhalten sollten (Bülow-Schramm 1995, S. 4/5). Ein solches Mitspracherecht ist sicherlich auch in anderen Frage der Planung einer Evaluation sinnvoll.

nipuliert werden. So wäre es beispielsweise denkbar, Studierende auf das Bestehen von Prüfungsleistungen hin zu trainieren.

- **Habitualisierung** (Münch/Pechmann 2009, S. 67, Frey 2006): Die andauernde Durchführung von Evaluationen birgt die Gefahr, dass sie zu Selbstläufern werden, deren Sinn wenig hinterfragt wird und deren Nutzen gering ausfällt. Power entwirft das Bild einer *Audit Society*, in der Evaluationen die Funktion einer vertrauensbildenden Maßnahme in einer Risikogesellschaft innehaben, womit sie sich von ihrer ursprünglichen Intention entfernen (Power 1999). Frey postuliert, dass Evaluationen sogar dann nicht seltener durchgeführt würden, wenn sie sich als weniger erfolgreich als zuvor angenommen herausstellen würden (*induzierte Verkrustung*).

- **Arbeitsbelastung** (Rindermann 2009, S. 341, DeGEval Gesellschaft für Evaluation 2008): Evaluationen bedürfen eines nicht zu unterschätzenden Einsatzes an Ressourcen. Dabei handelt es sich nicht nur um die von den Evaluatoren zu leistende Arbeit, sondern auch um den Aufwand für die Datengeber und für jene, die sich Gedanken über die Folgen von Evaluation bzw. über die Implementation von Änderungen machen. Somit bedeutet Evaluation eine mehr oder weniger erhöhte Arbeitsbelastung für alle Stakeholder. Diese Tatsache hat auch Eingang in die Evaluationsstandards der DeGEval (DeGEval Gesellschaft für Evaluation 2008) gefunden, in denen gefordert wird, Verfahren zu wählen, die gewährleisten, dass Belastungen und Nutzen in einem angemessenen Verhältnis stehen (D1) und dass auch der Gesamtaufwand in einem angemessenen Verhältnis zum Nutzen der Evaluation steht (D3). Und Kromrey formuliert:

  > „Außerdem darf die Evaluation den laufenden Betrieb nicht ‚stören' – schließlich ist das eigentliche Ziel der Hochschule die Sicherstellung eines geregelten Angebots für ein ordnungsgemäßes Studium und nicht dessen Evaluation."(Kromrey 2006b, S. 238)

Grundsätzliche Infragestellung von Evaluation

Evaluationen werden hin und wieder, meist durch bestimmte Stakeholder(-gruppen),

prinzipiell abgelehnt. Die grundsätzliche Infragestellung von Evaluationen hat mehrere Facetten. Zum einen spielen hier die bereits skizzierten Vorbehalte von Stakeholdern eine Rolle, etwa in Form der Sorge, dass der Evaluationsgegenstand prinzipiell nicht angemessen erfasst werden kann. Darüber hinaus gibt es auch elementare Vorwürfe gegen Evaluationen, die unterstellen, dass es sich um ein reines Ökonomisierungs- und Standardisierungsinstrument mit subtilen Machteffekten handelt oder dass es sich um eine Kontrolltechnologie handelt, die zu permanenter Selbstbeobachtung führe (Ernst 2008, S. 16). Münch und Pechmann konstatieren, dass die Kritik an Evaluationen und Ratings nirgendwo so groß ist wie innerhalb des wissenschaftlichen Feldes, und zwar nicht zuletzt deshalb, weil die Wissenschaft traditionell als ein Bereich angesehen werde, der außerhalb der Gesetze des wirtschaftlichen Marktes und des ökonomischen Denkens liege (Münch/Pechmann 2009, S. 67).

Nutzung der Evaluationsergebnisse

Die Nutzung der Ergebnisse von Evaluationen wird oftmals als unzureichend beschrieben (vgl. etwa Hense 2006, Rindermann 2009, S. 201, Schmidt 2009, S. 168), dies gilt auch für die Ergebnisse von Lehrevaluationen. Die mangelhafte Verwendung der Ergebnisse von Lehrevaluationen hat zwei zentrale Ursachen: Zum einen wirkt sich hier das weiter oben erwähnte Problem der nicht immer klaren Zuständigkeiten aus, zum anderen werden Evaluationen auch immer wieder in der ebenfalls bereits erwähnten *taktischen Funktion* eingesetzt, dienen also eher der Legitimation und Außendarstellung denn als Grundlage von Veränderung.

Ein weiterer Aspekt ist ein durch bestimmte Konstellationen ausgelöstes Misstrauen gegenüber Evaluationsergebnissen, was keine gute Basis für ihre Nutzung darstellt. Dieses Misstrauen wächst sicherlich in dem Maße, in dem die Integration der verschiedenen Stakeholdergruppen abnimmt. Lehrende, die mit Evaluationsergebnissen konfrontiert werden, an deren Ermittlung sie nicht mitgewirkt haben, werden ihnen ein erhebliches Maß an Skepsis entgegenbringen. Ein entscheidender Faktor für das Vertrauen in Evaluationsergebnisse ist das Vertrauen in die Kompetenz der Bewerter. Fehlt dieses

Vertrauen, werden auch die Ergebnisse vermutlich als wenig realitätsnah angesehen. Dieser Umstand wird beispielsweise dann offensichtlich, wenn Bewertete (z.B. Lehrende) das Urteil der Bewertenden (z.B. Studierende) in Zweifel ziehen, frei nach dem Motto: *Sie wissen es halt nicht besser, später werden sie mir dankbar sein.*

Die Frage der Berücksichtigung der Stakeholder

Evaluationen hängen von der Bereitschaft der Stakeholder zur Mitwirkung ab:

> „Nicht zuletzt ist sie [die Evaluation, Anm. d. A.] dabei mit zahlreichen Akteuren mit je unterschiedlichen Zielen und Vorstellungen konfrontiert, deren Handeln sämtlich über Erfolg und Misserfolg des zu evaluierenden Programms wie auch der Evaluation selbst mitentscheidet." (Kromrey 2006b, S. 238)

Nur wenn alle Beteiligten und Betroffenen bereit sind, ihren Teil zum Gelingen der Evaluation beizutragen, können belastbare Ergebnisse entstehen. Dazu kann es erforderlich sein, die Stakeholder in den Gesamtprozess einzubeziehen[29], etwa in die Phasen der Zweckbestimmung, der Erarbeitung bzw. Anpassung der Fragestellung (Lohnert/ Rolfes 1997, S. 12), der Methodenwahl und der Ergebniserarbeitung. So lehnen Studierende Evaluationen tendenziell eher ab, wenn sie nicht an der Planung beteiligt waren (Bülow-Schramm 1995, S. 5/6). Neben der Förderung der Bereitschaft zur Mitarbeit ist die Berücksichtigung der Stakeholder natürlich auch ein wichtiges Mittel zur Steigerung der Akzeptanz der Ergebnisse (Henninger/Balk 2001).

Unterschiedliche Ansprüche an Evaluationen

An Lehrevaluation an Hochschulen werden von den verschiedenen Stakeholdergruppen sich teilweisende widersprechende Ansprüche gestellt. Auch gibt es unterschiedliche Ansprüche innerhalb einer Stakeholdergruppe. Die Verschiedenartigkeit der Stakeholder, die möglicherweise mit Lehrevaluationen in Kontakt kommen oder durch sie betroffen sein können – von Eltern und Studieninteressierten über Studierende bis zu Mitarbeitern eines Ministeriums – bringt es mit sich, dass das Spektrum der Inter-

[29] Dies wird auch im DeGEval-Standard N1 empfohlen (DeGEval Gesellschaft für Evaluation 2008).

essen, die mit ihnen verbunden werden, ebenso vielfältig sein kann. Auch können die Erwartungen, die an eine Evaluation gerichtet werden, überhöht sein:

> „Natürlich möchte man neues empirisch abgesichertes Wissen darüber gewinnen, wovon erfolgreiches Lehren und Studieren abhängt und wie der Erfolg gefördert (wenn schon nicht garantiert) werden kann – insofern ist das Forschungsparadigma gefragt. Natürlich sollen zugleich Effektivität und Effizienz der Verwendung der in den Hochschulbereich fließenden öffentlichen Mittel kontrolliert werden, sollen die Hochschulen Rechenschaft über ihr Tun ablegen – also ist auch das Kontrollparadigma angesprochen. Und ebenso natürlich soll Evaluation dabei helfen, geeignete Maßnahmen zur Verbesserung der Qualität von Lehre und Studium zu konzipieren, zu implementieren und zu testen – womit schließlich das Entwicklungsparadigma zu seinem Recht kommt.“ (Kromrey 2006b, S. 238)

Zu Recht bezeichnet Kromrey derart komplexe Erwartungen als unrealistisch, zumal, wie er weiter ausführt, damit auch noch die Hoffnung einhergehe, dass eine solche Evaluation schnell, einfach und mit geringem Kosten- und Arbeitsaufwand zu realisieren sein solle und darüber hinaus den eigentlichen Studienbetrieb nicht stören dürfe (Kromrey 2006b, S. 238).

Methodische Probleme

Im Kern dieses Problem geht es darum, welche empirischen Methoden geeignet sind, den Evaluationsgegenstand angemessen zu erfassen und so eine faire und treffende Bewertung zu ermöglichen.

Eines der zentralen Themen in diesem Bereich ist sicherlich ein gewisses Theoriedefizit: Es gibt keine allgemeingültige Theorie *guter Lehre*, ebenso wenig gibt es eine ihrer Evaluation. Eine verbindliche Definition von Qualität ist kaum herzustellen, deshalb beruht das den Lehrevaluationen zugrunde liegende Verständnis von Qualität auf Plausibilitätsannahmen, was wiederum dazu führt, dass es in seinen Grundannahmen weder explizit noch transparent gemacht werden kann (Schmidt 2009, S. 168). Diese Tatsache wirft die Frage auf, ob der Evaluationsgegenstand überhaupt angemessen erfasst werden kann, welche in dieser Arbeit in Form der Frage nach der Gegenstsandsangemessenheit behandelt wird. Außerdem steht sie mit dem in Abschnitt *Ergebnisnutzung*

bereits angesprochenen Problem des Misstrauens gegenüber den Evaluationsergebnissen in Verbindung. Bülow-Schramm verweist allerdings darauf, dass methodische Argumente auch einen völlig anderen Ursprung als grundlegende Zweifel an der empirischen Vorgehensweise haben können. Sie merkt an, dass das Argument mangelnder Repräsentativität, das sowohl gegen freiwillige (und damit nicht die gesamte Stichprobe erreichende) Fragebogenuntersuchungen als auch gegen Interviews vorgebracht wird, eher auf Zweifel an der Wichtigkeit erhobener Inhalte verweise, einhergehend mit einer mangelnden Bereitschaft, Verantwortung für die festgestellten Schwächen zu tragen (Bülow-Schramm 1995, S. 6/7).

Die in Evaluationen häufig verwendeten Kennzahlen, d. h. Quantifizierungen von Leistungen, z. B. Absolventenzahlen oder die Anzahl von Publikationen, stellen ein weiteres wesentliches methodisches Problem dar. Der Kern dieses Problems ist die Schwierigkeit der Definition und Messung der Leistungen, die eine Hochschule erbringt (Riese 2007, S. 5). Forschung und Lehre sind kaum durch Leistungsmessungen abzubilden, ebenso wenig gibt es leicht quantifizierbare Zielsetzungen. Ein Verfahren, das sich stattdessen durchgesetzt hat, basiert auf dem Vergleich von Kennzahlen. Diese Kennzahlen werden dazu genutzt, das System Hochschule in verkürzter, modellhafter Form abzubilden (Riese 2007, S. 88/89). Häufig sind sie Ausdruck von Mengenleistungen in Forschung und Lehre. So werden im Bereich der Forschung etwa Publikations- und Vortragszahlen, Zitationshäufigkeiten und Drittmittelquoten verglichen. In der Lehre werden vorzugsweise Absolventenzahlen als Bewertungsgrundlage herangezogen. Gewichtige pragmatische Argumente für die Verwendung von Kennzahlen sind ihre relativ einfache Verfügbarkeit und ihre (scheinbare) Objektivität. Darüber hinaus sind Kennzahlen sehr einfach zu vergleichen, da ein komplexes System wie die Hochschule stark reduziert abgebildet wird und die Kennzahlen keiner weiteren Dekodierung bedürfen (Münch/Pechmann 2009, S. 69). Was eine Kennzahl jedoch nicht ausdrücken kann, ist die Qualität einer Leistung (Riese 2007, S. 90, Schmidt 2006a, S. 12): Ob eine hohe Anzahl Absolventen Ausdruck guter Leistung einer Hochschule oder eher geringer Qualitätsanforderungen ist, ist nicht unmittelbar ersichtlich. Eine umfassende Unter-

suchung bedeutet jedoch erheblichen methodischen Mehraufwand, da mit steigender Komplexität des Evaluationsgegenstandes auch die Anforderungen an das methodische Design wachsen.

3. Empirie, Gegenstandsangemessenheit und Hochschulevaluation

3.1. Empirie: die wissenschaftliche Erfahrung

3.1.1. Empirische Wissenschaft, nicht empirische Wissenschaft und alltagsweltliche Empirie

Evaluationsforschung bedient sich, wie im vorhergehenden Kapitel dargestellt, des Instrumentariums der empirischen Sozialforschung, um fundierte und tragfähige Ergebnisse zu erhalten. *Empirisch* bedeutet dabei, zurückgehend auf das griechische Wort *ém-peiros* (erfahren, kundig), *auf Erfahrung beruhend* (Dudenredaktion 2005). Empirische Sozialforschung ist somit eine Methode, die sich auf (sinnliche) Erfahrungen stützt, um wissenschaftlich abgesichertes Wissen über soziale Sachverhalte zu gewinnen. Empirischer Sozialforschung – oder allgemeiner empirischer Wissenschaft – liegt die Annahme zugrunde, dass Wissen über die tatsächliche Welt[30] nur durch Auseinandersetzung mit der Realität und durch ihre Beobachtung gewonnen werden kann (Kromrey 2006a, S. 29). Empirische Wissenschaft lässt sich dabei im Wesentlichen gegen zwei Bereiche abgrenzen: Zum einen gegen nicht empirische Wissenschaft, zum anderen gegen die alltagsweltliche Empirie.

Nicht empirische Wissenschaft gründet ihre Erkenntnisse nicht auf direkte Beobachtung und sinnliche Erfahrung, sondern bedient sich anderer Techniken. Von zentraler Bedeutung ist hier die *Logik* als Lehre des folgerichtigen Denkens (Dudenredaktion 2005), innerhalb derer Aussagen formuliert werden, deren Richtigkeit sich anhand formaler Kriterien untersuchen lässt. Klassische nicht rein empirische Wissenschaften

[30] Begriffe wie *tatsächliche Welt*, *Realität* etc. werden hier nicht positivistisch verwendet, sondern als Beschreibung des beobachtbaren Geschehens in der Welt verstanden, ohne eine Richtigkeit oder Fehlerhaftigkeit der jeweiligen Wahrnehmung zu implizieren.

sind beispielsweise die Mathematik und die Philosophie. Viele Wissenschaften bedienen sich allerdings in Abhängigkeit von den untersuchten Gegenständen sowohl empirischer als auch nicht empirischer Methoden.

Von *alltagsweltlicher Erfahrung* (alltagsweltlicher Empirie) unterscheidet sich wissenschaftliche Empirie in erster Linie durch eine systematische, gerichtete und selektive Vorgehensweise[31](Kromrey 2006a, S. 21 ff.), also durch die *Methode*, die von Lamnek folgendermaßen definiert wird:

> „Systematisches Verfahren bei der Entwicklung wissenschaftlicher Probleme, Aussagen etc., sowie deren Überprüfung an der Realität." (Lamnek 2005, S. 728)

Empirische Forschungsmethoden helfen dabei, Erfahrungen effizient zu sammeln und für das Erkenntnisinteresse relevante Aspekte gezielt und strukturiert zu erfassen. Somit leisten sie einen Beitrag dazu, den Untersuchungsbereich zu verstehen (Lueger 2000, S. 10). Wie im vorhergehenden Kapitel bereits dargestellt, führt der Einsatz von Forschungsmethoden nicht zwingend zu besseren (i. S. v. zutreffenderen) Ergebnissen, erfüllt aber einen wesentlichen Zweck, indem er die Erkenntnisse nachvollziehbar und auf ihre Stichhaltigkeit hin überprüfbar macht – ein wesentlicher Beitrag zu Sicherung der Qualität wissenschaftlicher Forschungsergebnisse. Schnell, Hill und Esser formulieren die Bedeutung von Methoden für die Wissenschaft wie folgt:

> „Das Kennzeichen der Wissenschaft ist die Methode. Wissenschaft lässt sich nicht über Inhalte definieren, sondern nur über die Vorgehensweise. Von der Vorgehensweise hängt die Gültigkeit der Schlussfolgerungen ab." (Schnell/Hill/Esser 2005, S. 6)

Dementsprechend ist das Kennzeichen der empirischen Sozialforschung die empirische Methode:

> „Sozialforschung ist die systematische Analyse von Fragestellungen unter Einsatz von empirischen Methoden (z. B. der Befragung, Beobachtung, Datenanalyse etc.) mit dem Ziel, verallgemeinerbare Aussagen empirisch begründet treffen oder überprüfen zu können." (Flick 2009b, S. 15)

[31] Die Betrachtungs- bzw. Untersuchungsgegenstände unterscheiden sich hingegen nicht unbedingt.

3.1.2. Was ist eine empirische Methode?

Bei empirischen Methoden handelt es sich um kodifizierte Handlungsanweisungen und Regeln, die dazu dienen, durch konkrete Erfahrungen gezielt Erkenntnisse zu gewinnen (Häder 2010, S. 20). Die in der empirischen Sozialforschung zur Anwendung kommenden Methoden stammen oftmals ursprünglich aus dem Alltag, was angesichts der Tatsache, dass sie der Erforschung des Alltags dienen, nicht verwundert (vgl. Häder 2010, S. 20, Friedrichs 1973). Methoden der wissenschaftlichen Empirie genügen dabei allerdings bestimmten Kriterien:

- Sie sind *regelgeleitet*, d. h. es existieren Regularien und Handlungsanweisungen, die das Vorgehen innerhalb der entsprechenden Methode beschreiben und festlegen.

- Die Vorgehensweise ist *explizierbar*, d. h. dass die Arbeitsschritte, die durchgeführt werden, und die getroffenen Entscheidungen einzeln und nachvollziehbar dargestellt werden können.

- Ihr Einsatz erfolgt *bewusst und zielgerichtet*, d. h. ihre Eignung für das jeweilige Forschungsvorhaben wird abgewogen und aufgrund dieser Abwägung werden Entscheidungen für oder wider eine Methode gefällt.

Methoden empirischer Sozialforschung lassen sich prinzipiell zwei unterschiedlichen Methodenfamilien zuordnen, nämlich *quantitativen* und *qualitativen* Verfahren (Lamnek 2005, S. 3 ff., Flick 2009b, S. 21 ff., Diekmann 2007, S. 18 ff.). *Quantitative* Verfahren transformieren die empirischen Erfahrungen in Zahlenwerte, d. h. sie *quantifizieren*. So würde beispielsweise die Zustimmung oder Ablehnung, die eine befragte Person bzgl. eines Sachverhaltes äußert, in einem Fragebogen erfasst und in Form einer Zahl wiedergegeben[32]. Die Fragen sind komplett vorformuliert und werden wörtlich als Stimuli

[32] Beispielsweise könnte die völlige Zustimmung durch eine 5 repräsentiert werden, die gänzliche Ablehnung durch eine 1. Die Zahlen von 2 bis 4 würden in diesem Beispiel Zwischenwerte repräsentieren, 2 könnte beispielsweise für „lehne eher ab“ stehen, 3 für eine neutrale Ausprägung und 4 für „stimme eher zu“.

präsentiert, die Antworten werden auf einer vorgegebenen Skala zugeordnet, d. h. *standardisiert* erfasst. Die Auswertung dieser numerischen Daten erfolgt üblicherweise mit Hilfe statistischer Rechenoperationen.

Qualitative Verfahren arbeiten nicht mit diesen Festlegungen, sondern zeichnen sich durch *Offenheit* aus, d. h. es werden weder feste Fragen noch Antworten vorgegeben, sondern die Datenerhebung erfolgt in freier(er) Form. Offenheit darf indes nicht mit Beliebigkeit verwechselt werden, denn auch hier muss selbstverständlich den Grundvoraussetzungen empirischer Sozialforschung Genüge geleistet werden, die Verfahren müssen regelgeleitet, bewusst und zielgerichtet eingesetzt werden. In Bereichen, in denen es sinnvoll ist, haben sich Verfahren zur Strukturierung der Datenerhebung (beispielsweise Beobachtungs- oder Interviewleitfäden) etabliert. Die Auswertung der Daten erfolgt anhand interpretativer Verfahren im weitesten Sinne.

Methoden alltagsweltlicher Empirie müssen gemäß dieser Unterscheidung anders als jene der wissenschaftlichen Empirie keinen festen Regeln folgen. Sie stellen damit eher eine Art gesellschaftlichen Common Sense dar, dessen Vorgehensweisen üblicherweise nicht expliziert werden und sie kommen ohne fundierte Eignungsabwägung zum Einsatz. Alltägliche Beobachtung, etwa in einer fremden Umgebung, findet üblicherweise ohne gesetzte Regeln statt. Sie ist nicht explizierbar und erfolgt auch nicht bewusst, sondern intuitiv und auf Basis des individuellen Weltwissens, das etwa Höflichkeitsregeln o. ä. beinhaltet. So gibt es beispielsweise keine Vorgaben zu den Fragen, warum beobachtet und nicht etwa ein Gespräch bevorzugt wird, in welcher Reihenfolge welche Handlungen oder Objekte zu beobachten sind oder wie sich die beobachtende Person verhalten soll. Innerhalb der wissenschaftlichen Methode *Beobachtung* hingegen werden genaue Anleitungen aufgestellt und bewusste Entscheidungen und Fokussierungen gefordert.

Über diese Punkte hinaus ist ein weiterer wesentlicher Aspekt, der wissenschaftliche von alltagsweltlichen empirischen Methoden unterscheidet, die Verfahrens*dokumentation*. Wissenschaftliche Methodik fordert, diesem für die Wissenschaftlichkeit entscheidenden Kriterium zu entsprechen. Deswegen fordern wissenschaftliche empi-

rische Methoden etwa dazu auf, die gemachten Erfahrungen strukturiert festzuhalten und Entscheidungen offenzulegen[33].

Die in der empirischen Sozialforschung zur Anwendung kommenden Methoden lassen sich drei funktionalen Gruppen zuordnen:

- Methoden zur *Stichprobenziehung* bzw. *Probandenauswahl*: Diese Methoden unterstützen bei der Auswahl der in die Untersuchung einzubeziehenden Personen. Prinzipiell lässt sich zwischen zufallsgesteuerten und nicht zufallsgesteuerten Verfahren unterscheiden.

- Methoden zur *Datenerhebung*: Diese Methoden dienen der Sammlung der eigentlichen Daten. Die grundlegenden Verfahren sind Befragungen (sowohl in Form von Fragebögen als auch in Gestalt von Interviews), Beobachtungen und Inhaltsanalysen[34] (Häder 2010, S. 21).

- Methoden zur *Datenauswertung*: Datenauswertungsmethoden leiten durch den Prozess der Ergebnisgenerierung aus den zuvor erhobenen Daten. Im Wesentlichen lässt sich zwischen statistischen (beschreibenden und schließenden) und interpretativen Methoden unterscheiden.

In der vorliegenden Arbeit wird die Gruppe der Datenerhebungsmethoden betrachtet, da sie jenen Teil der (Evaluations-)Forschung verkörpert, welcher die Stakeholder am unmittelbarsten betrifft und gewissermaßen die Schnittstelle zwischen Forschern und Probanden bildet. Dadurch verleiht sie den Befragten das nötige Gewicht für die Untersuchung und beeinflusst die Konstruktion von Gegenstandsangemessenheit innerhalb der Stakeholdergruppe maßgeblich, da es sich um den Teil der Untersuchung handelt,

[33] Reischmann nutzt dieses Kriterium der Dokumentation gar dazu, zwischen Evaluation und Evaluationsforschung, also zwischen nicht wissenschaftlicher und wissenschaftlicher Bewertung, zu unterscheiden (Reischmann 2003, S. 20).

[34] Die Methodengruppe der Inhaltsanalysen kann sowohl als Datenerhebungs- als auch als Datenauswertungsmethode betrachtet werden. In der Funktion, neue Daten zu generieren, die ihrerseits wieder ausgewertet werden – beispielsweise durch Wortschatzanalysen –, handelt es sich um eine Datenerhebungsmethode. Dieses Verfahren ist etwa im Rahmen quantitativer Sozialforschung üblich. In Rahmen qualitativer Sozialforschung dient sie hingegen dazu, latente Sinngehalte des Materials gezielt zu erschließen, ist also eine Daten*auswertungs*methode (vgl. Lamnek 2005, S. 480).

in dem sie direkt zu Wort kommen. Die Datenerhebungsmethoden sind auch die für die Probanden vorrangig sichtbaren Methoden, die in einem Forschungsvorhaben zum Einsatz kommen.

Kurzcharakterisierung der zentralen Datenerhebungsmethoden

Die Methoden Befragung, Beobachtung und Inhaltsanalyse[35] sind die grundlegenden Methoden wissenschaftlich-empirischer Datenerhebung.[36] Sie lassen sich wie folgt charakterisieren (Häder 2010, Kromrey 2006a, Lamnek 2005):

Befragung ist der Oberbegriff für Datenerhebungsmethoden, die auf der überwiegend sprachlichen Kommunikation zwischen mind. zwei Personen beruhen. Häder betont, dass es sich um *systematisch gesteuerte* Kommunikation handelt (Häder 2010, S. 187). Atteslander stellt diesem Kriterium außerdem das Kennzeichen der (theoriegeleiteten) *Kontrolle* der Situation an die Seite (Atteslander 2008, S. 103). In der Praxis gibt es eine Vielzahl unterschiedlicher Ausprägungen von Befragungen, die über jeweils eigene Mechanismen zur Steuerung der Kommunikation verfügen. Grundlegend kann zwischen Befragungen mittels Fragebögen, die durch die Probanden auszufüllen sind und somit auf indirekter Kommunikation beruhen, und gesprächsbasierten Befragungen unterschieden werden. Innerhalb dieser Ausprägungen können konkrete Methoden wiederum nach weiteren Charakteristika unterschieden werden, etwa nach der Art der Datenerfassung (standardisiert vs. offen, quantitativ vs. qualitativ), nach der Art der Präsentation bzw. der Kommunikationsvermittlung (Face-to-face, per Telefon oder via Computer), nach dem Grad der Strukturiertheit (maximal strukturiert, etwa unter

[35] Diekmann weist darauf hin, dass für das Verfahren der Inhaltsanalyse auch alternative Bezeichnungen wie Textanalyse, Dokumentenanalyse oder Bedeutungsanalyse verwendet werden, die sich aber weniger durchgesetzt haben (Diekmann 2007, S. 481).

[36] Michael Häder benennt diese Methoden als Grundmethoden der Datenerhebung in der empirischen Sozialforschung (Häder 2010, S. 21). Kromrey legt dar, dass wissenschaftliche Erfahrung letztlich immer auf Beobachtung beruht (Kromrey 2006a, S. 21 ff.), verwendet den Begriff Beobachtung dabei allerdings als Oberbegriff für jegliche sinnesbasierte Auseinandersetzung mit der Welt (Kromrey 2006a, S. 21 ff.; S. 37) und bezieht somit auch die Verfahren der Befragung und der Inhaltsanalyse mit ein. Auch die Analyse von Standardwerken zu Methoden der empirischen Sozialforschung, beispielsweise *Qualitative Sozialforschung* von Siegfried Lamnek (Lamnek 2005), bestätigt diese Einteilung.

Verwendung vorformulierter Fragen, bis völlig offen, d. h. ohne Steuerung durch die interviewende Person) oder nach der Zielsetzung der Erhebung (z.B. biographisches Interview, problemzentriertes Interviews). Außerdem sind vielfältige Kombinationen dieser Varianten denkbar.

Systematisch gesteuert bedeutet nicht, dass der Spielraum für die Antworten der befragten Personen unbedingt in besonderer Weise begrenzt wird, sondern vielmehr, dass die Kommunikationssituation planmäßig ist, d. h. bewusst gesucht und so gestaltet wird, dass sie im Sinne des Erkenntnisinteresses zielführend ist. Hierfür werden Techniken wie beispielsweise Erzählimpulse, Leitfäden oder eine Kombination aus offenen und standardisierten Fragen eingesetzt.

Beobachtung bezeichnet Verfahren zur gezielten Erfassung von für die Untersuchung wesentlichen Sachverhalten durch datenerhebende Personen, die Beobachter. Wissenschaftliche Beobachtungen sind systematisch und zielgerichtet, planmäßig und kontrolliert und heben sich dadurch von alltäglichen Beobachtungen ab.

Beobachtungen lassen sich v.a. nach der Rolle der Forschenden (teilnehmend vs. nicht teilnehmend), nach dem Gegenstand der Betrachtung (Selbst- oder Fremdbetrachtung), nach dem Beobachtungsort (Feld- oder Laborsituation), nach der Positionierung des Forschenden zum Gegenstand (offen oder verdeckt), nach dem Grad der Strukturierung (offen, teilstandardisiert, standardisiert) und der technischen Vermittlung (Unterstützung durch technische Hilfsmittel, wie beispielsweise Audio- oder Videoaufzeichnung) unterscheiden.

Von der Befragung unterscheidet sich die Beobachtung im Wesentlichen durch zwei Aspekte: zum einen beschränkt sie sich nicht schwerpunktmäßig auf die sprachliche Kommunikation, sondern kann auch Merkmale wie beispielsweise die Mimik oder soziales Verhalten mit berücksichtigen, zum anderen können Beobachtungen auch ohne direkte Interaktion zwischen Forschenden und Datengeber/-innen durchgeführt werden.

Inhaltsanalyse als Daten*erhebungs*methode ist quantitativer Natur[37] (Schnell/Hill/Esser 2005, S. 407). Mit ihrer Hilfe werden Inhalte von Aufzeichnungen und Dokumenten wie Texten, Bildern oder Filmen systematisch erfasst. Dabei werden die Eigenschaften und Merkmale der untersuchten Objekte quantifiziert. Ihr Ziel ist der Rückschluss vom vorliegenden Material auf individuelle und gesellschaftliche nichtsprachliche Phänomene (Mayntz/Holm/Hübner 1974, S. 151). Vier Formen der Inhaltsanalye können unterschieden werden (Diekmann 2007, S. 481 ff., Schnell/Hill/Esser 2005, S. 408 ff., Früh 2007, S. 25 ff.):

- Frequenzanalyse: Die Häufigkeit des Vorkommens bestimmter Merkmale, z. B. von Begriffen, wird erhoben.
- Valenzanalyse: Aufgrund der Häufigkeit bestimmter Wörter oder anderer Merkmale wird die Ausrichtung (positiv, neutral, negativ) des untersuchten Objekts bestimmt.
- Intensitätsanalyse: Zusätzlich zu den Bewertungen der Valenzanalyse wird auch deren Grad (Intensität) erfasst.
- Kontingenzanalyse: Überprüfung des Auftretens bestimmter Begriffe im Zusammenhang mit anderen.

Neben der Fokussierung auf inhaltliche Aspekte des Materials können auch formale Elemente Gegenstand der Analyse sein, etwa die Häufigkeit der Verwendung bestimmter grammatikalischer Formen[38].

Bei der Inhaltsanalyse handelt es sich, im Gegensatz zu den vorgenannten Techniken, um ein nicht reaktives Verfahren der Datenerhebung, d. h. dass die Urheber des zu untersuchenden Materials durch die Untersuchung nicht direkt betroffen sind und somit ihr Verhalten auch nicht beeinflusst wird, da das Material in

[37] Inhaltsanalysen schließen immer auch Aspekte der Datenauswertung mit ein, sind also keine reinen Datenerhebungsmethoden (vgl. Schnell/Hill/Esser 2005, S. 407, Diekmann 2007, S. 481). Als Methode der Daten*auswertung* existieren auch verschiedene qualitativ orientierte Varianten der Inhaltsanalyse (vgl. etwa Kuckartz 2012, Mayring 2010).

[38] Dieses Verfahren wurde beispielsweise im Rahmen der Studie *Die Arbeitslosen von Marienthal* (1933, Marie Jahoda, Paul Lazarsfeld und Hans Zeisel) eingesetzt, indem Schulaufsätze nach der Verwendung des Konjunktivs als Indikator für Resignation untersucht wurden (Diekmann 2007, S. 481).

der Regel nicht speziell zum Zwecke der Forschung und in Interaktion mit den Forschenden erstellt wurde.

Diese Unterteilung in drei zentrale Datenerhebungsmethoden stellt ein grundlegendes Ordnungsschema dar. Innerhalb jeder dieser Methoden existiert eine vielzahl spezifischer Ausprägungen des Verfahrens, die jeweils über einen besonderen Schwerpunkt bzw. über eigene Zielrichtungen und Charakteristika verfügen, sich teilweise ergänzen, aber auch widersprechen können. Für Forschende stellt sich nun das Problem, aus diesem reich gefüllten Werkzeugkasten das richtige Werkzeug zur Bearbeitung der gestellten Aufgabe, im Kern beschrieben und eingegrenzt durch Forschungsfrage und Forschungsgegenstand, auszuwählen. Das zentrale Konzept, nach dem sich die Methodenwahl bewerten lassen muss, ist das der *Gegenstandsangemessenheit*.

3.2. Die Auswahl von Datenerhebungsmethoden

3.2.1. Gegenstandsangemessenheit als Kriterium der Methodenwahl

In jeder empirischen Untersuchung stellt sich das Problem der Wahl einer geeigneten Datenerhebungsmethode, denn es gibt kein Patentrezept für die Durchführung empirischer Erhebungen. Die Auswahl der einzusetzenden Methoden muss vielmehr nach gründlicher Auseinandersetzung mit dem spezifischen Forschungsgegenstand und der Forschungsfrage erfolgen und möglichst gut zu ihnen passen (Kromrey 2006a, S. 12, Flick 2007, S. 478)[39]. Dieses Kriterium der *guten Passung*, das sowohl bei der Auswahl einer geeigneten Datenerhebungsmethode gilt als auch gleichzeitig ein Maß für die Güte dieser Wahl darstellt, wird durch das Konzept der *Gegenstandsangemessenheit* verkörpert (vgl. u. a. Flick 2007, S. 27, Helfferich 2009, S. 46, Brüsemeister 2008, S. 28). Gegenstandsangemessen bedeutet dabei, dass die gewählte(n) Methode(n) sämtliche für das jeweilige Forschungsvorhaben relevanten Daten und Informationen erfassen, gleichzeitig die Besonderheiten von Forschungsgegenstand (beispielsweise die Sprachkompetenz der Probanden) und Forschungsfrage (etwa die Einnahme eines speziellen Blickwinkels)

[39] Diese Festlegung ist die konkrete Umsetzung des Kennzeichens des bewussten und zielgerichteten Methodeneinsatzes empirischer Forschungsmethoden, siehe S. 55.

sowie weitere beeinflussende Faktoren (z. B. Zeit- und Ressourcenbeschränkungen) berücksichtigen[40]. Letztlich müssen die gewählten Methoden der Vielschichtigkeit des Gegenstandes gerecht werden (Brüsemeister 2008, S. 29). Oder anders formuliert: Sie müssen sowohl zur Forschungsfrage als auch zu den Eigenschaften des untersuchten Gegenstandes passen und die Verbindung zwischen beiden herstellen können.

Die Suche nach einer angemessenen Methode basiert auf der Annahme, dass keine universelle Methode existiert, die *jedem* Gegenstand und *jeder* Perspektive gleichermaßen gerecht werden kann[41]. Diese Sichtweise gründet auf der Grundüberlegung, dass unterschiedliche Datenerhebungsverfahren durch spezifische Vorteile, aber auch Grenzen gekennzeichnet sind und damit jeweils eine eigene Berechtigung besitzen. Anders formuliert: In einer monomethodischen Welt müsste die Frage der Angemessenheit einer Methode nicht gestellt werden. Die Annahme, dass unterschiedliche Verfahren über unmittelbar mit ihnen verbundene spezifische Eignungen verfügen, kann dabei von unterschiedlicher Reichweite sein: Einerseits kann sie sich auf Methoden beziehen, die innerhalb der gleichen Methodenfamilie angesiedelt sind, d. h. dass Forscher sich zwischen verschiedenen quantitativen *oder* verschiedenen qualitativen Verfahren entscheiden. Andererseits lässt sie sich auch umfassender und über die Grenzen der Methodenfamilien hinweg begreifen, so dass in die Betrachtung der individuellen Methodeneigenschaften das gesamte Spektrum der Methoden empirischer Sozialforschung mit einbezogen und der Weg zu ihrer Kombination geebnet wird[42].

Die Bedeutung der Gegenstandsangemessenheit für die Qualität einer empirischen Untersuchung ist allgemein anerkannt (vgl. etwa Brüsemeister 2008, Kelle 2008, Flick 2007, Kromrey 2006a, Flick/von Kardorff/Steinke 2000 oder Steinke 1999). Die Bestimmung

[40] Das Kriterium der Gegenstandsangemessenheit wird nicht allein auf die Wahl von Datenerhebungsmethoden angewendet, sondern für eine Vielzahl methodischer Entscheidungen, u. a. für die Wahl einer Datenauswertungsmethode oder das Sampling, also die Probandenauswahl.

[41] Uwe Flick empfiehlt in diesem Zusammenhang, ein Verfahren, das allgemein anwendbar sein soll, doppelt auf seine tatsächliche Eignung hin zu überprüfen (Flick 2007, S. 478).

[42] Diese familienübergreifende Sichtweise stellt in gewisser Weise eine Überwindung des lange geführten *Paradigmenstreits* (*paradigm wars*) dar, der zwischen den Vertretern der beiden Methodenfamilien geführt wurde. Außerdem steht sie der Diskussion über die Kombination von Verfahren beider Methodenfamilien, die unter verschiedenen Begriffen geführt wird – Mixed Methods, Triangulation, Methodenintegration – und mit jeweils eigenen Ansätzen verbunden ist, nahe. Vgl. hierzu Kelle 2008.

einer Fragestellung bzw. des Untersuchungsgegenstandes von der Methode aus wird in dieser Sichtweise abgelehnt, denn in dieser Logik würden nur Fragestellungen und Themen bearbeitet, die durch die präferierte Methode erfasst werden können, darüber hinausgehende Aspekte würden nicht berücksichtigt[43]. Dies würde schließlich dazu führen, dass Wissenschaft kaum in der Lage wäre, Entdeckungen außerhalb ihres eigenen durch die Instrumente determinierten Erwartungshorizontes zu machen.

Kriterien zur Bestimmung der Gegenstandsangemessenheit

Das Kriterium der Gegenstandsangemessenheit ist für die Wahl einer Datenerhebungsmethode zwar überzeugend und, wie eine Analyse der methodologischen Literatur gezeigt hat, akzeptiert, bleibt allerdings unscharf. Eine explizite Eingrenzung, wann eine Methode angemessen ist und wann nicht, wurde während dieser Literaturarbeit nicht sichtbar. Konkrete Hinweise darauf, wie und woran Gegenstandsangemessenheit tatsächlich bemessen werden kann, sind eher selten und – wenn sie denn existieren – in der Regel aufgrund der enormen Bandbreite der Möglichkeiten eher vage[44]. Gleichwohl ist eine Beschreibung der Faktoren, die Einfluss auf die Angemessenheit einer Methode ausüben, auf einer übergeordneten Ebene durchaus möglich. Um eine solche Beschreibung zu erhalten, wurden Werke, in denen die Frage der Methodenwahl thematisiert wird, systematisch auf Nennungen möglicher Beurteilungskriterien untersucht. Direkt oder indirekt benannte Bewertungsaspekte wurden gesammelt, nach Ähnlichkeit sortiert, abstrahiert und zusammengetragen. Auf diese Weise wurden folgende zentrale Kriterien zur Bestimmung oder Bewertung der Gegenstandsangemessenheit einer Methode erkennbar (siehe Häder 2010, Creswell 2009, Flick 2009b, Rindermann 2009, Atteslander 2008, Brüsemeister 2008, Kelle 2008, Diekmann 2007, Flick 2007, Kromrey 2006a, Bortz/Döring 2005, Lamnek 2005, Schnell/Hill/Esser 2005, Seale 2004, Flick/von Kar-

[43] Brüsemeister vergleicht die Frage nach der Angemessenheit von Forschungsmethoden mit den Kriterien der korrekten Operationalisierung und Messung in der rein quantitativen Forschung (Brüsemeister 2008, S. 48).

[44] Flick beispielsweise nennt drei Referenzpunkte, die zur Beurteilung der Angemessenheit konkreter Methoden herangezogen werden können, nämlich den Gegenstand, die Fragestellung und die untersuchten Subjekte (Flick 2007, S. 276). Über diese Nennung hinaus gibt Flick jedoch keine Hinweise zur konkreten Bewertung der Gegenstandsangemessenheit.

dorff/Steinke 2000, Tashakkori/Teddlie 2000, Seale 1999, Steinke 1999, Roth 1993, Garz/Kraimer 1991, Friedrichs 1973 sowie Blalock/Blalock 1968):

1. **Der Gegenstand**, seine Eigenschaften und Besonderheiten. Untersuchungsgegenstände verfügen über Eigenschaften, die sie einer Datenerhebungsmethode zugänglicher machen als einer anderen. So kann beispielsweise die Untersuchung der Veränderung von sozialen Netzwerken andere Datenerhebungsmethoden erfordern als etwa die Betrachtung der Interaktion zwischen Erzieherinnen und Kindergartenkindern. Um die Eigenschaften des Gegenstandes berücksichtigen zu können, ist es erforderlich, ihn möglichst exakt zu definieren und seine Besonderheiten herauszuarbeiten.

2. **Die Fragestellung** mit ihren spezifischen Schwerpunktsetzungen und Blickwinkeln. Jede Forschungsfrage hat eine eigene Zielrichtung und basiert auf einer besonderen Betrachtungsperspektive. Zielt sie etwa auf die Analyse von Einzelfällen ab, so sind andere Datenerhebungsmethoden angemessen, als wenn eine möglichst breit verallgemeinerbare Aussage Ziel der Untersuchung ist.

3. **Die untersuchten Subjekte**, ihre Merkmale und Kompetenzen. Dabei handelt es sich um die Grundgesamtheit der Personen, die als Datengeber fungieren. Jede Gruppe von Personen zeichnet sich ebenso wie jedes ihrer Individuen durch spezifische Eigenschaften aus, denen bei der Wahl der Datenerhebungsmethode Rechnung getragen werden muss. Beispiele für Einflussfaktoren sind etwa die Sprachkompetenzen der untersuchten Subjekte oder Vorbehalte gegenüber den Forschenden.

4. **Die Zielgruppe der Untersuchung** und ihr Erkenntnisinteresse bzw. ihre Informationsbedürfnisse. Richtet sich die Untersuchung etwa an ein Fachpublikum, das an speziellen Details interessiert ist, müssen andere Methoden angewendet werden, als wenn die Zielgruppe eher an überblicksartigem Wissen interessiert ist. Je nach Art der Untersuchung können Personen sowohl untersuchte Subjekte sein als auch gleichzeitig zur Zielgruppe der Untersuchung gehören.

5. **Die Historie der Methode** bzw. der Gegenstand, für den sie entwickelt wurde. Für beinahe jedes Datenerhebungsverfahren lässt sich zurückverfolgen, für welchen Forschungsgegenstand es entwickelt wurde (Flick/von Kardorff/Steinke 2000, S. 22). Natürlich kann eine Methode auch für einen anderen Gegenstand als für den ursprünglich intendierten eingesetzt werden, allerdings hilft die Kenntnis über die Entstehungszusammenhänge bei der Einschätzung der spezifischen Eignung und erklärt gleichzeitig Besonderheiten der Methode.

6. **Das Verhältnis von Aufwand, Ressourcen und Nutzen**. Dieses Kriterium fokussiert in erster Linie den Aufwand, der mit dem Einsatz von Datenerhebungsmethoden einhergeht, einerseits in Bezug auf die Gruppe der Forschenden, andererseits in Bezug auf den Forschungsgegenstand und die Untersuchten. Für Letztere bedeutet eine Datenerhebung in der Regel Zusatzaufwand. Der Aufwand muss mit den zur Verfügung stehenden Ressourcen – maßgeblich Geld und Zeit – und dem erwarteten Nutzen in einem sinnvollen Verhältnis stehen.

7. **Die Forschenden**, ihre Kompetenzen im Bereich der empirischen Methoden und ihre Erfahrungen mit dem Einsatz bestimmter Verfahren. In der Praxis verfügen vermutlich nur wenige Forschende über ideale Kompetenzen bzgl. jeder denkbaren Methode. Dementsprechend bestimmt sich die Angemessenheit auch über diese Einflussgröße.

8. **Die Offenheit der Methode(n)**. Dieses Kriterium bezieht sich auf die Frage, ob die zum Einsatz kommenden Methoden den untersuchten Gegenstand möglichst ganzheitlich erfassen, eine Infragestellung der Vorannahmen der Forschenden zulassen und ob sie den Äußerungen der Forschungsteilnehmer/-innen den zur Beantwortung der Forschungsfrage erforderlichen Freiraum einräumen.

Schnell wird ersichtlich, dass diese Liste keine Gewichtung enthält und konzeptbedingt auch nicht dazu geeignet ist, als Handlungsanweisung verstanden werden zu können. Vielmehr handelt es sich bei dieser Aufzählung um eine konkretisierte Sammlung aller Kriterien, die nach dem aktuellen Stand methodologischer Diskussion zur

Bewertung der Gegenstandsangemessenheit von Datenerhebungsmethoden herangezogen werden können. Dabei kann sie durchaus hilfreich für die Praxis der Bestimmung der Gegenstandsangemessenheit sein, wobei es nicht zwingend erforderlich ist, all diese Kriterien gleichermaßen zu berücksichtigen – manchmal ist es u. U. gar nicht möglich: eine maximal offene Methode ist ebensowenig automatisch die beste wie eine Methode, die hinsichtlich des Ressourcenverbrauchs besonders günstig ist. Vielmehr muss versucht werden, all diese Ebenen miteinander zu vereinen und so auszubalancieren, dass ein möglichst erfolgversprechendes Ergebnis erwartet werden kann. Das bedeutet, dass einzelne der aufgeführten Merkmale ggf. hinter anderen zurücktreten müssen. Diese Gewichtung muss allerdings wohlbegründet vorgenommen werden, stellt sie doch letztlich den Kern der Umsetzung der Gegenstandsangemessenheit dar.

3.2.2. Gegenstandsangemessenheit im Kontext von Evaluationen

Auch im Bereich der Evaluationsforschung sollte die Auswahl der zum Einsatz kommenden Datenerhebungsmethoden anhand des Kriteriums der Gegenstandsangemessenheit erfolgen. Diese Forderung lässt sich nicht nur aus der Verbindung der Evaluationsforschung zur empirischen Sozialforschung herleiten, sondern ist im Forschungsbereich Evaluation selbst fest verankert. Einerseits werden in vielen Evaluationsmodellen (siehe Abschnitt 2.1.3) die Einsatzmöglichkeiten bestimmter Methoden skizziert und im Kontext des jeweiligen Modells betrachtet, andererseits wird auch in den Standards der DeGEval (siehe Kap. 2.1.4) die Wahrung der Gegenstandsangemessenheit gefordert. Da die Standards – sowohl die der DeGEval als auch die des Joint Committee oder anderer Evaluationsgesellschaften – die Idee eines modellübergreifenden Qualitätsansatzes verkörpern und als solcher einzigartig und prinzipiell auch akzeptiert sind, stellen die in ihnen festgehaltenen Forderungen einen günstigen Ausgangspunkt für die Betrachtung der Gegenstandsangemessenheit im Kontext der Evaluationsforschung dar. Außerdem wird in den Standards der DeGEval, anders als in manchen Evaluationsmodellen, keine Präferenz für eine bestimmte Datenerhebungsmethode vertreten.

Insgesamt lassen sich aus 13 der 25 Standards der DeGEval Bezüge zur Gegenstands-

angemessenheit der Datenerhebungsmethode herstellen. Bei diesen Standards handelt es sich letztlich um konkretisierte Ausformulierungen von Teilaspekten des Konzepts der Gegenstandsangemessenheit, die im vorherigen Abschnitt (siehe S. 64 ff.) vorgestellt wurden:

N1 Identifizierung der Beteiligten und Betroffenen: „Die am Evaluationsgeschehen 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." (DeGEval Gesellschaft für Evaluation 2008)

Dieser Standard empfiehlt die Identifikation aller an der Evaluation beteiligten und durch sie betroffenen Personen, um sie so weit wie möglich in die Planung und Durchführung der Evaluation einbeziehen zu können. Gleichermaßen wird die Wichtigkeit der Ermittlung der Informationsbedürfnisse der Adressatinnen und Adressaten der Evaluation herausgestellt (DeGEval Gesellschaft für Evaluation 2008, S. 23). Somit bezieht sich N1 auf die Kriterien *Die untersuchten Objekte* (3) und *Die Zielgruppe der Untersuchung* (4).

N2 Klärung der Evaluationszwecke: „Es soll deutlich bestimmt sein, welche Zwecke mit der Evaluation verfolgt werden, so dass die Beteiligten und Betroffenen Position dazu beziehen können und das Evaluationsteam einen klaren Arbeitsauftrag verfolgen kann." (DeGEval Gesellschaft für Evaluation 2008)

Dieser Standard soll die Arbeit der Evaluierenden erleichtern und Grundlagen für die Verbreitung und Nutzung der Ergebnisse schaffen (DeGEval Gesellschaft für Evaluation 2008, S. 24). Er ist eng mit dem Kriterium *Die Fragestellung* (2) verbunden, da die Evaluationszwecke die Grundlage der Formulierung der Forschungsfrage(n) bilden.

N3 Glaubwürdigkeit und Kompetenz des Evaluators/der Evaluatorin: „Wer Evaluationen durchführt, soll persönlich glaubwürdig sowie methodisch und fachlich kompetent sein, damit bei den Evaluationsergebnissen ein Höchstmaß an Glaub-

würdigkeit und Akzeptanz erreicht wird.“ (DeGEval Gesellschaft für Evaluation 2008)

Dieser Standard verweist auf das Kriterium *Die Forschenden* (7). Die Bewertung der Glaubwürdigkeit und Kompetenz wird gemäß der Erläuterungen zu diesem Standard wiederum durch die Stakeholder der Evaluation durchgeführt. Somit existiert auch eine Verbindung zu Kriterium *Die untersuchten Subjekte* (3).

N4 Auswahl und Umfang der Informationen: „Auswahl und Umfang der erfassten Informationen sollen die Behandlung der zu untersuchenden Fragestellungen zum Evaluationsgegenstand ermöglichen und gleichzeitig den Informationsbedarf des Auftraggebers und anderer Adressaten und Adressatinnen berücksichtigen.“ (DeGEval Gesellschaft für Evaluation 2008)

Hier lassen sich Bezüge zu zwei Kriterien herstellen: *Die Fragestellung* (2) und *Die Zielgruppe der Untersuchung* (4). Dabei wird die Eignung der Methoden daran gemessen, wie gut sie in der Lage sind, Daten zu erheben, die der Beantwortung der *Fragestellung* dienen (2) und die gleichzeitig den Informationsbedürfnissen der Empfänger des Berichts entsprechen (DeGEval Gesellschaft für Evaluation 2008, S. 31), wodurch eine Verbindung zur *Zielgruppe der Untersuchung* (4) hergestellt wird.

D1 Angemessene Verfahren: „Evaluationsverfahren, einschließlich der Verfahren zur Beschaffung notwendiger Informationen, sollen so gewählt werden, dass Belastungen des Evaluationsgegenstandes bzw. der Beteiligten und Betroffenen in einem angemessenen Verhältnis zum erwarteten Nutzen der Evaluation stehen.“ (DeGEval Gesellschaft für Evaluation 2008)

In diesem Standard wird die Entscheidung für Methoden gefordert, die möglichst wenig Mehrbelastung erzeugen und ethisch vertretbar sind (DeGEval Gesellschaft für Evaluation 2008, S. 28). Als Bezugspunkte der Bewertung werden der Evaluationsgegenstand sowie die Stakeholder benannt. Damit stellt der Standard Bezüge zu folgenden Kriterien der Gegenstandsangemessenheit her: *Das Verhältnis von*

Aufwand, Ressourcen und Nutzen (6), *Der Gegenstand* (1) und *Die untersuchten Subjekte* (3).

D2 Diplomatisches Vorgehen: „Evaluationen sollen so geplant und durchgeführt werden, dass eine möglichst hohe Akzeptanz der verschiedenen Beteiligten und Betroffenen in Bezug auf Vorgehen und Ergebnisse der Evaluation erreicht werden kann." (DeGEval Gesellschaft für Evaluation 2008)

Im Fokus dieses Standards stehen die Stakeholder einer Evaluation. Er ist somit mit *Die untersuchten Subjekte* (3) verbunden. In der Erläuterung zu den Standards wird darauf verwiesen, dass die Verzerrung der Evaluationsergebnisse durch eine oder mehrere Stakeholderguppen zu vermeiden ist (DeGEval Gesellschaft für Evaluation 2008, S. 28). Dadurch besteht auch ein Bezug zum Kriterium 8, *Offenheit der Methoden.*

D3 Effizienz von Evaluation: „Der Aufwand für Evaluation soll in einem angemessenen Verhältnis zum Nutzen der Evaluation stehen." (DeGEval Gesellschaft für Evaluation 2008)

Durch diese Forderung steht der Standard D3 unmittelbar in Bezug zu Kriterium 6, *Verhältnis von Aufwand, Ressourcen und Nutzen.*

F3 Vollständige und faire Überprüfung: „Evaluationen sollen die Stärken und die Schwächen des Evaluationsgegenstandes möglichst vollständig und fair überprüfen und darstellen, so dass die Stärken weiter ausgebaut und die Schwachpunkte behandelt werden können." (DeGEval Gesellschaft für Evaluation 2008)

Standard F3 ist mit dem Kriterium *Offenheit der Methoden* (8) verbunden, denn nur, wenn die Untersuchung tatsächlich methodisch offen durchgeführt wird, können sowohl positive als auch negative Seiten des untersuchten Gegenstandes herausgearbeitet werden.

F4 Unparteiische Durchführung und Berichterstattung: „Die Evaluation soll unterschiedliche Sichtweisen von Beteiligten und Betroffenen auf Gegenstand und Er-

gebnisse der Evaluation in Rechnung stellen. Berichte sollen ebenso wie der gesamte Evaluationsprozess die unparteiische Position des Evaluationsteams erkennen lassen. Bewertungen sollen fair und möglichst frei von persönlichen Gefühlen getroffen werden.“ (DeGEval Gesellschaft für Evaluation 2008)

In diesem Standard wird die Forderung aufgestellt, dass Evaluationen nicht eine spezifische Sichtweise übernehmen sollen oder durch eine bestimmte Gruppe instrumentalisiert oder vereinnahmt werden (DeGEval Gesellschaft für Evaluation 2008, S. 31/32). Dies bedeutet für die Datenerhebung, dass die Stakeholder unparteiisch und möglichst offen zu betrachten sind. Dadurch entsteht eine Verbindung zu den Kriterien *Die untersuchten Subjekte* (3) und *Die Offenheit der Methode* (8).

G1 Beschreibung des Evaluationsgegenstandes: „Der Evaluationsgegenstand soll klar und genau beschrieben und dokumentiert werden, so dass er eindeutig identifiziert und so genau wie möglich zugänglich gemacht werden kann.“ (DeGEval Gesellschaft für Evaluation 2008)

Die in diesem Standard geforderte Beschreibung und Dokumentation des Evaluationsgegenstandes dient der Klärung von Zusammenhängen zwischen dem Gegenstand und seinen Wirkungen sowie dem Ziel, bisher unbeachtete Nebenwirkungen ausfindig zu machen (DeGEval Gesellschaft für Evaluation 2008, S. 33). Der Standard steht in Bezug zum Kriterium *Der Gegenstand* (1).

G2 Kontextanalyse: „Der Kontext des Evaluationsgegenstandes soll ausreichend detailliert untersucht und analysiert werden.“ (DeGEval Gesellschaft für Evaluation 2008)

Ziel dieses Standards ist es, die Rahmenbedingungen des Evaluationsgegenstandes zu berücksichtigen, um Erkenntnisse über die spezifischen Bedingungen zu erhalten und um die Übertragbarkeit der Ergebnisse auf andere Kontexte einschätzen zu können. Gleichzeitig wird davor gewarnt, die Kontextanalyse zu sehr auszudehnen, um nicht die erforderliche Bearbeitung der übrigen Evaluations-

schritte zu gefährden (DeGEval Gesellschaft für Evaluation 2008, S. 33). Somit ist der Standard G2 mit den Kriterien *Der Gegenstand* (1) und *Das Verhältnis von Aufwand, Ressourcen und Nutzen* (6) verbunden.

G3 Beschreibung von Zwecken und Vorgehen: „Gegenstand, Zwecke, Fragestellungen und Vorgehen der Evaluation, einschließlich der angewandten Methoden, sollen genau dokumentiert und beschreiben werden, so dass sie identifiziert und eingeschätzt werden können." (DeGEval Gesellschaft für Evaluation 2008)

Die Beschreibung der benannten Aspekte soll zur Steigerung der Transparenz der Evaluation beitragen (DeGEval Gesellschaft für Evaluation 2008, S. 34). Diese Explikation steht in Verbindung mit dem Kriterium *Die Fragestellung* (2). Zielgruppe dieser zu erzeugenden Transparenz sind neben etwaigen dritten Rezipienten der Evaluation v. a. die Beteiligten und Betroffenen, wodurch auch eine Verbindung zu den Kriterien *Die untersuchten Subjekte* (3) und *Die Zielgruppe der Untersuchung* (4) besteht.

G5 Valide und reliable Informationen: „Die Verfahren zur Gewinnung von Daten sollen so gewählt oder entwickelt und dann eingesetzt werden, dass die Zuverlässigkeit der gewonnenen Daten und ihre Gültigkeit bezogen auf die Beantwortung der Evaluationsfragestellungen nach fachlichen Maßstäben sichergestellt sind. Die fachlichen Maßstäbe sollen sich an den Gütekriterien quantitativer und qualitativer Sozialforschung orientieren." (DeGEval Gesellschaft für Evaluation 2008)

Standard G5 hat eine etwas herausgehobene Position inne: Er beschreibt keine konkrete Qualitätsforderung, sondern verweist vielmehr auf die Gütekriterien der qualitativen und quantitativen Sozialforschung. Dadurch erhalten prinzipiell alle Forderungen an die Wahl einer Datenerhebungsmethode Gültigkeit, die in der empirischen Sozialforschung allgemein erhoben werden. Somit bezieht sich dieser Standard letztlich auf alle acht Kriterien der Gegenstandsangemessenheit, also auch auf Kriterium 5, *Die Historie der Methode*, auf das sich die übrigen Standards nicht direkt beziehen.

Aus methodologischer Sicht referenzieren die *Standards für Evaluation* somit alle relevanten Aspekte zur Bestimmung der Eignung von Datenerhebungsmethoden. Dabei wurden die methodologischen Kriterien jedoch nicht einfach übernommen, sondern auf das Feld der Evaluationsforschung übertragen, an seine Besonderheiten angepasst und so weit als möglich konkretisiert. Werden nun die Anforderungen der Standards danach sortiert, welchen methodologischen Kriterien sie zuzuordnen sind, so ergibt sich folgende Anforderungsliste:

Der Gegenstand: Er soll durch die eingesetzten Datenerhebungsmethoden möglichst wenig belastet werden (D1) und inklusive seiner Zusammenhänge und Wirkungen (G1) und seines Kontextes (G2) exakt erfasst werden.

Die Fragestellung: Sie soll auf Grundlage der erhobenen Daten adäquat behandelt werden können (N4), sie soll zuverlässig auf der Basis von Daten, die nach den Maßstäben der Gütekriterien qualitativer und quantitativer Sozialforschung gültig sind, beantwortet werden können (G5), und die ausgewählten Methoden sollen anhand ihrer Passung zur Fragestellung bewertet werden können (G3). Zudem sollen die Zwecke der Evaluation, aus denen die Fragestellung und damit auch die methodische Vorgehensweise hervorgeht, deutlich bestimmt sein (N2).

Die untersuchten Subjekte: Ihre Interessen sollen so weit wie möglich in die Anlage der Evaluation einbezogen werden (N1). Die Subjekte sollen die fachliche und methodische Kompetenz (auch hinsichtlich der gewählten Erhebungsverfahren) der evaluierenden Personen einschätzen können (N3), ihre Belastung durch die gewählten Methoden soll so gering wie möglich gehalten werden (D1). Sowohl die Vorgehensweise (demnach auch die gewählten Methoden) als auch die Ergebnisse der Evaluation sollen bei den untersuchten Subjekten auf möglichst hohe Akzeptanz treffen können (D2), unterschiedliche Sichtweisen der untersuchten Subjekte sollen unparteiisch abgebildet werden können (F4). Schließlich sollen Hintergründe und gewählte Vorgehensweisen für sie nachvollziehbar und einschätzbar beschrieben werden (G3).

Die Zielgruppe der Untersuchung: Auch sie soll so weit wie möglich bei Planung Entwurf der Evaluation berücksichtigt werden (N1). Der Informationsbedarf der Zielgruppe wirkt sich auf Art und Umfang der zu erhebenden Informationen aus (N4) und auch ihr soll es möglich sein, Hintergründe und gewählte (methodische) Vorgehensweise einzuschätzen (G3).

Die Historie der Methode: Um die geforderte Validität und Reliabilität der erhobenen Daten (G5) sicherstellen zu können, kann die Entwicklungshistorie einer Methode – also eine Betrachtung der Problemstellungen, zu deren Lösung sie entwickelt wurde – als die Entscheidung unterstützendes Kriterium herangezogen werden.

Das Verhältnis von Aufwand, Ressourcen und Nutzen: Die Datenerhebungsmethoden sollen Beteiligte und Betroffene möglichst wenig belasten (D1), der für eine Evaluation betriebene Aufwand soll in einem angemessenen Verhältnis zu ihrem Nutzen stehen (D3) und auch die erforderliche Untersuchung des Kontextes des Evaluationsgegenstandes soll das Zeitkontingent nicht zu stark strapazieren (G2).

Die Forschenden: Sie sollen methodisch und fachlich kompetent sein, um die Glaubwürdigkeit der Evaluation zu befördern (N3), sie müssen also Methoden wählen, die sie beherrschen.

Die Offenheit der Methode(n): Datenerhebungsmethoden müssen so offen sein, dass sie unterschiedliche Sichtweisen gleichermaßen erfassen können und somit unverzerrte Daten und Ergebnisse sicherstellen (D2), sie sollen Stärken und Schwächen des Evaluationsgegenstandes gleichermaßen erfassen und abbilden (F3) sowie, ebenfalls durch ausreichende methodische Offenheit, sicherstellen, dass keine spezifische Sichtweise einer Partei übernommen, sondern unterschiedliche Positionen gleichermaßen erfasst werden (F4).

In Tabelle 3.1 (S. 74) sind diese Bezüge zwischen den Kriterien der Gegenstandangemessenheit und den Standards der DeGEval in kompakter Form dargestellt[45]. Eine

[45] Um die Tabelle übersichtlich zu halten, wurde der Standard G5 nur bei den Kriterien *Die Historie der Methode* und *Die Fragestellung* (zu der er in besonderer Weise gehört) vermerkt, wenngleich er, wie oben beschrieben, prinzipiell auch den anderen Kriterien zuzuordnen ist.

Kriterium der Gegenstandsangemessenheit	DeGEval-Standard
Der Gegenstand	D1, G1, G2
Die Fragestellung	N2, N4, G3, G5
Die untersuchten Subjekte	N1, N3, D1, D2, F4, G3
Die Zielgruppe der Untersuchung	N1, N4, G3
Die Historie der Methode	G5
Das Verhältnis von Aufwand, Ressourcen und Nutzen	D1, D3, G2
Die Forschenden	N3
Die Offenheit der Methode(n)	D2, F3, F4

Tabelle 3.1.: Zusammenhang zwischen den Kriterien der Gegenstandsangemessenheit und den Standards für Evaluation der DeGEval

Betrachtung der Häufigkeiten der Bezüge lässt dabei erkennen, welche Sichtweise auf die Konstruktion von Gegenstandsangemessenheit den Standards für Evaluation der DeGEval zugrunde liegt. Auffällig und gleichzeitig typisch für Evaluationsforschung ist die Orientierung an der Nutzbarkeit der erarbeiteten Ergebnisse. Um diese Nutzbarkeit zu gewährleisten, stellen die Stakeholder einen wichtigen Bezugspunkt hinsichtlich der Konstruktion von Gegenstandsangemessenheit dar:

Die mit sechs Bezügen zu den Kriterien der Gegenstandsangemessenheit besondere Betonung der Relevanz der untersuchten Subjekte ergibt sich aus der Tatsache, dass die Datengeber in zweifacher Hinsicht von zentraler Bedeutung für Evaluationen sind. Erstens gehen die Bewertungen, die im Zuge einer Evaluation vorgenommen werden, letztlich auf die Aussagen der Datengebenden zurück. Zweitens sind sie die Personen, die in den wohl allermeisten Fällen durch die Evaluation, bzw. die in ihrer Folge sich ergebenden Veränderungen, unmittelbar betroffen sind. Diese beiden Punkte zusammen bedingen, dass die untersuchten Subjekte ein natürliches Interesse daran haben, dass ihre Bewertungen richtig – also aus ihrer Sichtweise *angemessen* – erfasst und abgebildet werden. Dieses Interesse ist bei der Planung einer Evaluation grundlegend zu berücksichtigen, denn eine Evaluation, deren Datengrundlage bzw. -gewinnung für die Stakeholder nicht nachvollziehbar ist, wird bei ihnen nicht auf Akzeptanz stoßen[46],

[46] „Auch Wottawa und Thierau [. . .] erwähnen, daß bei Evaluationsvorhaben in umstrittenen Bereichen oder mit heiklen Ergebnissen häufig versucht wird, die erzielten Resultate hinsichtlich der verwen-

wodurch die Nutzung der Ergebnisse, die zentrales Kennzeichen von Evaluation ist (vgl. Kap. 2.1.2), erheblich erschwert oder gar unmöglich gemacht wird. Das Argument einer falschen Messung kann dementsprechend dazu führen, dass die gesamte Untersuchung hinfällig wird. Dem lässt sich nur durch eine ausreichende Berücksichtigung der untersuchten Subjekte entgegenwirken. Gleichzeitig schützt die grundlegende Integration der Sichtweise der Datengeber vor dem Vorwurf der Instrumentalisierung der Evaluation durch die Auftraggeber.

Neben der Beachtung der untersuchten Subjekte ist in diesem Zusammenhang der Nutzungsorientierung von Evaluationen auch die Beachtung der *Zielgruppe der Untersuchung* (Kriterium 4) bedeutsam, auf die drei Standards Bezug nehmen. Zwar gehört bei den meisten Evaluationen auch die Grundgesamtheit der Datengebenden zur Zielgruppe, darüber hinaus gehören ihr aber auch die Auftraggeber sowie ggf. weitere Interessenten an. Dabei existieren hier prinzipiell die gleichen Herausforderungen wie im Bereich der untersuchten Subjekte, denn auch die Auftraggeber werden eine Evaluation, mit deren Datengrundlage sie nicht einverstanden sind bzw. die sie als falsch ansehen, nicht akzeptieren. Deshalb sind auch die Interessen dieser Gruppe angemessen zu berücksichtigen. Allerdings ist der Einfluss der Auftraggeber auf den Verlauf einer Evaluation, insbesondere auf ihre prinzipielle Ausrichtung und Fragestellung, ungleich größer und wesentlich selbstverständlicher als jener der Datengeber. Aus diesem grundlegenden Ungleichgewicht heraus erklärt sich die auffällig starke Betonung der untersuchten Subjekte: Ihre ausführliche Berücksichtigung soll zu einem Ausgleich führen, der eine zentrale Basis für die Akzeptanz und somit für die Verwendung der Ergebnisse darstellt. Die in den Standards eingeräumte Bedeutung der Stakeholder für die Gegenstandsangemessenheit wird auch deutlich, wenn die Häufigkeit der Bezugnahmen betrachtet wird. Sieben der dreizehn Standards, die Bezüge zu Kriterien der Gegenstandsangemessenheit von Datenerhebungsmethoden beinhalten, berühren mindestens eine der beiden eben betrachteten Stakeholdergruppen.

deten Verfahren zu kritisieren, da diese ‚nicht idealen methodischen Anforderungen' entsprächen, was jedoch bei jeder Evaluationsstudie möglich ist [...]. Als Argumentationsfigur werden nach Wottawa und Thierau [...] gerne verwendet: ‚methodische Schwächen', ‚es wurde nicht alles berücksichtigt und ausgewertet', ‚fehlerhafte Operationalisierung', ‚interne Widersprüche', ‚singulär und nicht verallgemeinerbar' und ‚Widerspruch zu gesuchten Resultaten oder angesehenen Experten'." (Rindermann 2009, S. 201)

Ein weiteres Kriterium, das im Bereich der Evaluationsforschung eine gewichtige Rolle spielt, ist das *Verhältnis von Aufwand, Ressourcen und Nutzen* der Datenherhebungsmethode (Kriterium 6). Während dieser Aspekt in der idealtypischen Grundlagenforschung von eher nachgeordneter Bedeutung ist, hat er hier hohe Relevanz. Dies ist in den meist knappen Ressourcen, die für Evaluationen zur Verfügung stehen, und der damit einhergehenden Notwendigkeit eines effizienten Evaluationsprozesses begründet, wie bereits in Kapitel 2.1.2 dargestellt wurde.

Diese starke Betonung der Stakeholder und der Effizienz, die in der allgemeinen Literatur zur Gegenstandsangemessenheit so deutlich nicht hervorgehoben werden, sind, wie dargestellt, den Besonderheiten der Evaluationsforschung geschuldet. Die übrigen Bezüge der *Standards für Evaluation* zu Kriterien der Gegenstandsangemessenheit hingegen sind weniger auf Charakteristika dieses Bereichs zurückzuführen, sondern durch die Orientierung an empirischer Sozialforschung zu erklären:

Die Fragestellung ist mit vier Bezügen das insgesamt am zweithäufigsten berührte Kriterium der Gegenstandsangemessenheit. Dies wird vor dem Hintergrund der Übernahme von Regeln der empirischen Sozialforschung in das Feld der Evaluationsforschung verständlich. In der empirischen Sozialforschung stellt die Forschungsfrage das hauptsächliche Kriterium dar, das für das Design einer Untersuchung maßgeblich ist und anhand dessen die meisten der die Untersuchung betreffenden Entscheidungen zu begründen sind. In der Literatur dazu wird, wie bereits dargestellt, die Forschungsfrage als zentraler Bezugspunkt für die Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden benannt. Auch die Rückbindung der Gegenstandsangemessenheit an den *Untersuchungsgegenstand* (drei Bezüge) ist zentraler Bestandteil der Aussagen der Methodenliteratur bzgl. der Konstruktion von Gegenstandsangemessenheit. Gleiches gilt für die Frage der *Offenheit der Methoden* (Kriterium 8), *die Historie der Methode* (Kriterium 5) sowie das Kriterium der *Forschenden* (Kriterium 7). Diese Aspekte besitzen damit sowohl für empirische Sozialforschung im Allgemeinen als auch für Evaluationsforschung im Besonderen Gültigkeit.

Zusammenfassend lässt sich festhalten, dass die Konstruktion von Gegenstandsangemessenheit, wie sie im Rahmen der Standards für Evaluation vorgenommen wird, auf den in Abschnitt 3.2.1 dargestellten allgemeinen Kriterien der empirischen Sozialforschung aufbaut, aber einen besonderen Schwerpunkt auf die Stakeholder, insbesondere die untersuchten Subjekte, legt. Die Besonderheiten der Stakeholder und ihre Interessen stellen zusammen mit der Berücksichtigung des Verhältnisses von Aufwand, Ressourcen und Nutzen Kernkriterien zur Wahl angemessener Datenerhebungsmethoden im Kontext von Evaluationen dar.

3.3. Datenerhebungsmethoden in der Lehrevaluation an Hochschulen

Um die Gegenstandsangemessenheit von Datenerhebungsmethoden anhand der in Kapitel 3.2 dargestellten Kriterien einschätzen und abschließend bewerten zu können, ist es erforderlich, über die bereits erfolgte Betrachtung des Gegenstandsbereichs, in dem die Methoden eingesetzt werden sollen (siehe Kap. 2.2), hinaus auch die Methoden selbst näher zu betrachten. Demgemäß wird in den nachfolgenden Abschnitten zunächst dargestellt, welche Datenerhebungsmethoden im Kontext von Lehrevaluationen an Hochschulen gebräuchlich sind (Abschnitt 3.3.1), um anschließend zu analysieren, welche Eigenschaften diese Methoden auszeichnen und was sie hinsichtlich der Beantwortung der Frage nach der Gegenstandsangemessenheit implizieren. Danach werden schließlich, um das Bild des Evaluationsgegenstandes zu komplettieren, die Stakeholder in den Blick genommen. Denn während in Kapitel 2.2 die organisatorischen und strukturellen Rahmenbedingungen von Lehrevaluation an Hochschulen dargestellt wurden, blieb bislang eine genauere Betrachtung der Stakeholder außen vor, obwohl diese, wie gezeigt, bei Evaluationen von besonderer Bedeutung für die Einschätzung der Angemessenheit sind. Der Abschnitt 3.3.2 ist dementsprechend der näheren Betrachtung der Stakeholder von Lehrevaluationen an Hochschulen gewidmet.

3.3.1. Empirische Methoden in der Lehrevaluation und ihre spezifischen Eigenschaften

Das Spektrum der Datenerhebungsmethoden, die in der Lehrevaluation – also zur systematischen Untersuchung des Nutzens und Wertes von Lehrveranstaltungen an Universi-

täten – zum Einsatz kommen können, ist breit gefächert. Für beinahe jede existierende Datenerhebungsmethode ließe sich vermutlich ein Anwendungsbeispiel finden. Dennoch lässt sich ein gewisser Kanon von Verfahren, die in besonderem Maße zur Evaluation von Hochschullehre eingesetzt werden, erkennen. Dabei treten folgende Datenerhebungsmethoden in der Literatur, die sich mit der Beschreibung, Entwicklung und Diskussion von Verfahren der Hochschulevaluation befasst, aber auch in der Praxis[47] besonders hervor (Ernst 2008, S. 61 ff., Hochschulrektorenkonferenz 2006, Kromrey 2006b, Hochschulrektorenkonferenz 2004, Pasternack 2004, Becker-Richter et al. 2002, S. 25 ff., Barz 1998, Lohnert/Rolfes 1997, S. 31 ff., Schmidt 2009, S. 164 ff., Lewin et al. 2000, Richter 1994 oder Rindermann 2009) und stellen damit ein Standardrepertoire dar:

- *Fragebögen* als (teil-) standardisiertes Datenerhebungsinstrument,
- *Offene Interviews* als typisches Verfahren der offenen Datenerhebung,
- *Dokumenten- und Kennzahlenanalysen*,
- *Gruppendiskussionen*, etwa in Form offener Feedbackrunden, sowie
- Verfahren mit *Gutachtern*.

Kriterien zur Klassifikation von Datenerhebungsverfahren

Jedes dieser Verfahren zeichnet sich durch spezifische Eigenschaften aus, die zur Bewertung seiner Angemessenheit herangezogen werden können. Gegenstandsangemessenheit lässt sich über die Analyse der spezifischen Eigenschaften der Verfahren einerseits und der Anforderungen der Evaluation andererseits ermitteln. Zweckmäßige Verfahren sollten demnach Eigenschaften besitzen, die den Anforderungen der Evaluation entspre-

[47] Entsprechend des üblichen Charakters von Evaluation als *Tätigkeit im Auftrag*, bei der viele datenschutzrechtliche Aspekte relevant sind und deren Ergebnisse meist nur für bestimmte, definierbare Nutzerkreise von Interesse sind und nicht zuletzt deswegen selten einem breiten Publikum zugänglich gemacht werden, bleibt viel von der Arbeit, die geleistet wird – u. a. eben meist auch ihre konkrete Praxis – im Verborgenen. Somit gestaltet sich der Abgleich zwischen den in der Literatur betonten Aspekten und der tatsächlichen Evaluationspraxis oftmals schwierig. Dennoch konnte, insbesondere auf Tagungen sowie im Gespräch mit Evaluierenden, die nachfolgende Liste bestätigt werden.

chen. Die nachfolgend vorgestellten Kriterien wurden aus Standardwerken der sozialwissenschaftlichen Methodenliteratur (Atteslander 2008, Bortz/Döring 2005, Porst 2008, Diekmann 2007, Flick 2009b, Flick/von Kardorff/Steinke 2000, Friedrichs 1973, Garz/Kraimer 1991, Häder 2010, Aghamanoukjan/Buber/Meyer 2007, Kromrey 2006a, Lamnek 2005, Schnell/Hill/Esser 2005 und Silverman 2004) sowie aus den Standards für Evaluation (DeGEval Gesellschaft für Evaluation 2008) entwickelt bzw. abgeleitet. Sie nehmen die relevanten Aspekte von Datenerhebungsmethoden in den Fokus und dienen der Erfassung der inhärenten Eigenschaften von Methoden:

- Art der Datenerfassung,
- Grad der Strukturierung,
- Art der Kommunikationsvermittlung,
- Reaktivität,
- Größe der erfassbaren Stichprobe,
- Beteiligung von Personen,
- Zeitbedarf sowie
- Nähe zum Gegenstand.

Im Folgenden werden für jedes dieser Kriterien, basierend auf den Beschreibungen in der o. a. Literatur, Pole entwickelt, anhand derer die sich unterscheidenden Ausprägungen innerhalb des jeweiligen Charakteristikums dargestellt werden können. Jede der genannten Datenerhebungsmethode wird auf dem Kontinuum zwischen diesen Polen verortet, um sie auf diese Weise zu charakterisieren.

Art der Datenerfassung: Hier lässt sich zwischen zwei Polpaaren bzw. Unterdimensionen unterscheiden: 1. standardisiert vs. offen, 2. quantitativ vs. qualitativ (vgl. Lamnek 2005, S. 330 ff., Flick 2009b, S. 133 ff.).

Die Unterscheidung zwischen *standardisierter* und *offener* Befragung bezieht sich auf die (methodische) Gleichbehandlung aller Befragten. Bei einer gänzlich standardisierten Datenerhebung werden alle befragten Personen mit identischen Stimuli konfrontiert, erhalten also beispielsweise identische Fragebögen. Außerdem stehen allen Befragten exakt die gleichen Reaktionsräume zur Verfügung, etwa in der Form, dass alle Antwortmöglichkeiten vorweggenommen werden, d. h. dass auf die ausschließlich geschlossenen Fragen keine frei formulierten Antworten gegeben werden können. Auch die Situationen, in denen die Daten erhoben werden, sollten möglichst gleichförmig gestaltet werden. Diese Vereinheitlichung und methodische Gleichbehandlung – oder eben Standardisierung – soll dazu beitragen, die Antworten der befragten Personen maximal vergleichbar zu machen, indem individuelle Einflüsse möglichst kontrolliert werden sollen.

Dem gegenüber steht der Pol der offenen Datenerhebung. Bei einer vollkommen offenen Befragung werden keine vereinheitlichten Stimuli präsentiert, vielmehr entwickelt sich die Befragung situationsgebunden und wird damit auch durch die befragten Personen beeinflusst. Auch werden hier keine Reaktionsräume definiert, sondern die Befragten erhalten die Möglichkeit bzw. sind in der Pflicht, die Antworten zu äußern, die ihnen wesentlich erscheinen. Außerdem wird auf die Situation, in der die Datenerhebung stattfindet, möglichst kein Einfluss genommen. Dieses Verfahren soll es ermöglichen, die spezifische Sichtweise der befragten Personen ohne Einschränkungen durch die Fragebogendesigner zu erheben.

Das Gegensatzpaar *quantitativ* und *qualitativ* bezieht sich auf die Form, in der die beobachtete Realität festgehalten wird. Während sie in quantitativen Befragungen in Form von Zahlen festgehalten wird (etwa als Wert auf einer Skala von Eins bis Fünf zur Ermittlung von Zustimmung oder Ablehnung) wird sie in qualitativen Befragungen in einer nicht quantifizierenden Form aufgezeichnet, beispielsweise als detaillierte Beschreibung in Fließtextform.

Zwischen diesen Polen lassen sich zahlreiche Zwischenpositionen einnehmen, etwa teilstandardisierte Befragungen, in denen offene und standardisierte Einzelfragen

gemischt werden, oder Erhebungen, in denen sowohl quantifizierend als auch qualitativ gearbeitet wird.

Grad der Strukturierung: Der Grad der Strukturierung bezieht sich auf die Steuerung des Kommunikationsprozesses. Hier kann zwischen *strukturierten* und *unstrukturierten* (bzw. offenen) Verfahren unterschieden werden.

Bei strukturierten Verfahren folgt der Kommunikationsprozess einem definierten Schema. Die Befragung verläuft nach einem vorab festgelegten Kanon an Fragen bzw. Themen, der während der gesamten Erhebung nicht verlassen wird. Die Vorabfestlegung kann sich sowohl auf die Themen selbst als auch auf ihre Abfolge beziehen. Auch dieses Kriterium soll der Herstellung von Vergleichbarkeit zwischen den erhobenen Daten dienen.

Unstrukturierte Verfahren geben keine derartige Struktur vor. Der konkrete inhaltliche Ablauf ergibt sich erst unmittelbar im Prozess der Erhebung. Auch diese Variante soll dazu beitragen, den Befragten die Freiheit einzuräumen, ihre spezifische Sichtweise vorzutragen.

Ebenso wie bei der vorgenannten Art der Datenerfassung sind auch hier zahlreiche Zwischenpositionen denkbar. Verbreitet sind beispielsweise auch teilstrukturierte Verfahren, in denen zwar eine mehr oder weniger grobe Planung des Verlaufs existiert, die hinsichtlich der möglichen Inhalte der Antworten dann aber wiederum offen bzw. unstrukturiert sind.

Art der Kommunikationsvermittlung: Dieses Merkmal beschreibt den Weg, auf dem die Kommunikation zwischen den interagierenden Personen übermittelt wird. Hierbei lassen sich drei Unterscheidungsdimensionen festhalten: 1. Der Grad der Präsenz, 2. die Form der Kommunikation sowie 3. die Ausprägung der Gleichzeitigkeit.

Der Grad der Präsenz bezieht sich auf die räumliche Anwesenheit der Personen und lässt sich zwischen den Polen *persönlich* bzw. *Face-to-face* auf der einen Seite und *medial vermittelt* auf der anderen Seite verorten. In einer persönlichen Situa-

tion befinden sich alle beteiligten Personen zur selben Zeit am selben Ort. Die klassische Situation eines persönlichen Interviews etwa ist dieser Ausprägung zuzurechnen. Dem gegenüber stehen Kommunikationsformen, in denen (technische) Hilfsmittel zur Überbrückung der räumlichen Distanz zwischen den Beteiligten und zur Herstellung der Präsenz eingesetzt werden. Beispiele sind hier Telefon, Computer oder schriftlichen Kommunikationsformen, etwa ein Fragebogen.

Die Form der Kommunikation beschreibt die Art, in der der Informationsaustausch zwischen den beteiligten Personen erfolgt. Neben den Formen der *Mündlichkeit* und der *Schriftlichkeit* ist hier eine dritte Ausprägung denkbar: die *Visualität*. Im Falle einer Beobachtung erfolgt beispielsweise ein großer Teil der Datenerhebung durch die Anschauung.

Die Ausprägung der Gleichzeitigkeit schließlich bezieht sich auf die zeitliche Dimension der Kommunikation. Hier kann zwischen *synchronen* und *asynchronen* Kommunikationsformen unterschieden werden. Während bei synchronen Kommunikationsformen alle Beteiligten gleichzeitig miteinander kommunizieren, also unmittelbar auf einander reagieren können, findet die Kommunikation bei asynchronen Formen zeitversetzt statt, Reaktionen sind also mittelbar. Während mündliche Gespräche (unabhängig von ihrer jeweiligen Vermittlungsform) stets synchron stattfinden, ist beispielsweise bei Fragebögen keine Gleichzeitigkeit erforderlich.

Reaktivität: Dieses Kriterium bezieht sich auf den Grad, in dem der Einsatz einer Methode unmittelbar Reaktionen der Datenquelle(n) bedingt, wobei hier auch Probanden als Datenquellen betrachtet werden. Dabei sind zwei Dimensionen zu unterscheiden: Zum einen die Dimension der tatsächlichen Reaktivität der Methode, d. h. ihr Einfluss auf den Verlauf der Datenerhebung. Hier ist zwischen den Polen *reaktiv* und *nicht reaktiv* zu unterscheiden: Während die Daten, die mit reaktiven Methoden (beispielsweise teilnehmende Beobachtung, Befragungen) erhoben werden, bereits während des Messvorgangs etwa durch Antwortformulierungen verzerrt werden können, die nicht ehrlich sind, sondern sich an einer vermuteten sozialen Norm orientieren (soziale Erwünschtheit), zeichnen sich nicht

reaktive Methoden dadurch aus, dass diese Beeinflussung ausbleibt. Ein Beispiel für nicht reaktive Verfahren ist etwa die Analyse von Kennzahlen oder von bestehenden, unabhängig von dementsprechenden Forschungsvorhaben produzierten Dokumenten.

Die zweite Dimension dieses Kriteriums ist das Ausmaß der Interaktivität zwischen den beteiligten Personen. Während *interaktive* Methoden direktes Agieren und Reagieren zwischen den datenerhebenden und den datengebenden Personen gestatten, ist dies bei *nicht interaktiven* Methoden unmöglich.

Zwar sind nicht reaktive Verfahren prinzipbedingt auch nicht interaktiv. Unterschiedliche Formen reaktiver Verfahren können sich jedoch durch sehr verschiedene Interaktionsintensitäten auszeichnen.

Größe der erfassbaren Stichprobe: Methoden können sich hinsichtlich der Zahl der Probanden, deren Daten innerhalb einer bestimmten Zeit erhoben werden können, unterscheiden. Für die Erfassung der Daten einer *großen Stichprobe* sind oftmals Methoden besonders gut geeignet, die sich durch ein geringes Maß an erforderlicher persönlicher Präsenz sowie durch Asnychronität auszeichnen, da hier ohne die Anwesenheit der Forscher Daten erhoben werden können, was Arbeits- und Zeitersparnis bedeutet. Methoden, die Sychronität und ein hohes Maß persönlicher Präsenz erfordern, sind hingegen tendenziell eher für *kleine Stichproben* geeignet.

Methoden sind nicht unbedingt fest zwischen diesen Polen zu verorten. Durch Anpassungen einiger Parameter kann sich die Eignung der Methode zwischen den Polen durchaus verschieben.

Zu betonen ist, dass die Anzahl erfasster Probanden nicht zwingend in Relation zur erhobenen Datenmenge steht. Wenige sehr ausführliche Interviews können u. U. den gleichen Datenumfang produzieren wie sehr viele knapp gehaltene Fragebögen. Es können also auch auf Basis weniger Probanden viele Daten erzeugt werden – und umgekehrt.

Beteiligung von Personen: Auch hier sind zwei Dimensionen zu unterscheiden: 1. Die Anzahl der notwendigerweise in die Erhebung involvierten Personen und 2. die Erfordernis der Mitarbeit oder Kooperation Anderer. Während sich die erstgenannte Dimension mit den Polen *viele Personen* und *wenige Personen* auf die *Anzahl* der Personen bezieht, die bereit sein müssen, sich in irgendeiner Form an der Datenerhebung zu beteiligen (sei es als Datengeber/-in, als Forscher/-in oder in einer beliebigen anderen Rolle), fokussiert die zweite Dimension mit den Polen *abhängig* und *unabhängig* eben die Abhängigkeit des Gelingens der Erhebung von der Bereitschaft zur Mitarbeit oder Kooperation dieser oder auch bestimmter einzelner Personen.

Einerseits sind Szenarien denkbar, in denen die beiden Dimensionen positiv zusammenhängen, in denen also viele Personen involviert sind und es auf die Bereitschaft eben all dieser Personen ankommt. Andererseits gibt es aber auch Situationen, in denen die Kooperation weniger Personen – im Extremfall einer einzigen – über das Gelingen einer Erhebung entscheiden kann, beispielsweise wenn Gatekeeper[48] gewonnen werden müssen.

Der erstgenannte Aspekt, also die Anzahl der notwendigerweise in die Erhebung involvierten Personen, hängt maßgeblich mit der Größe der zu bearbeitenden Stichprobe zusammen: Je mehr Personen in die Untersuchung einbezogen werden, desto weniger hängt auch das tatsächliche Gelingen der Erhebung von der Bereitschaft zur Mitarbeit einer einzigen Person ab. Der letztgenannte Punkt hingegen ist eher in Verbindung mit den Eigenschaften der Grundgesamtheit der Datenerhebung zu sehen: Je enger und abgeschlossener eine Gruppe von Probanden ist – beispielsweise Angestellte einer bestimmten Abteilung eines Unternehmens –, desto abhängiger wird das Gelingen der Untersuchung von der Bereitschaft der Mitarbeit der Probanden. Dieser Effekt tritt auch bei sehr kleinen Grundgesamt-

[48] Als Gatekeeper oder Schlüsselpersonen werden Einzelpersonen oder ggf. auch Personengruppen bezeichnet, über die Zugang zum Feld erlangt werden kann. Gatekeeper können einerseits über die Öffnung des Feldes für die Forschenden entscheiden, andererseits können sie durch ihr Verhalten den Verlauf der Erhebung beeinflussen, etwa durch die ausdrückliche Unterstützung des Forschungsvorhabens die Rücklaufquote erhöhen (vgl. Roth 1993, S. 258 ff., Lamnek 2005, S. 603 ff.).

heiten auf, selbst wenn diese wenig abgeschlossen sind, etwa bei Experten für ein Nischenthema.

Zeitbedarf: Der Zeitbedarf einer Methode lässt sich zwischen den Ausprägungen *zeitintensiv* und *nicht zeitintensiv* verorten. Zeitintensive Methoden bedürfen eines vergleichsweise umfangreichen Zeiteinsatzes sowohl der datengebenden als auch der erhebenden Personen, um die Daten zu erheben – gemessen an der Gesamtzeit, die zur Durchführung der Evaluation/des Projekts zur Verfügung steht. So sind beispielsweise Methoden, die Synchronität erfordern, zumeist zeitintensiver als jene, die durch Asynchronität gekennzeichnet sind: Während in einem Face-to-face-Interview oftmals nur eine Person in einer bestimmten Zeit befragt werden kann, kann ein Fragebogen nach seiner Erstellung an beinahe beliebig viele Personen verteilt werden, die ihn auch gleichzeitig ausfüllen könnten – ohne zusätzlichen Einsatz von datenerhebenden Personen.

Nähe zum Gegenstand: Das Kriterium der Nähe zum Gegenstand zielt auf die Beziehung der Daten zum Untersuchungsgegenstand. Hier kann zwischen *ungefilterten* und *gefilterten* Daten unterschieden werden. Unter ungefilterten Daten werden solche verstanden, die unmittelbar am Gegenstand erhoben werden. Hierzu gehören beispielsweise Daten aus Beobachtungen des Gegenstandes oder (quantitativen oder qualitativen) Interviews mit Akteuren des Gegenstandes, in der Lehrevaluation etwa mit Studierenden, welche die zu bewertende Veranstaltung besucht haben. Gefilterte Daten hingegen werden nicht direkt durch die Forscher am Gegenstand erhoben, sondern durch Dritte produziert, durchlaufen also einen oder mehrere Filter, bevor die Erhebung abgeschlossen ist und sie den Evaluatoren/Forschern vorliegen. Hierzu zählen beispielsweise Datenerhebungen, die auf die Einschätzungen externer Gutachter zurückgreifen, oder auch Kennzahlen.

Anhand dieser Aspekte können die einzelnen Datenerhebungsmethoden auf ihre spezifische Eignung hin untersucht werden. Wesentlich ist dabei, dass nicht jedes dieser Kriterien für jede Datenerhebung gleichermaßen relevant ist. Bei empirischen Untersuchungen mit begrenzter Laufzeit ist etwa der Aspekt des Zeitbedarfs von größerer

Bedeutung als bei solchen, die keiner oder wenigstens keiner großen zeitlichen Restriktion unterliegen. Außerdem werden in der Praxis oftmals Kompromisse erforderlich sein, um den existierenden Ansprüchen gerecht werden zu können. So sind etwa Szenarien denkbar, in denen die Forschung zwar einer engen Zeitvorgabe unterliegt, aber dennoch relativ zeitintensive Datenerhebungsmethoden eingesetzt werden müssen, um die Forschungsfrage zufriedenstellend beantworten zu können. Dementsprechend ist die Relevanz jedes einzelnen Punktes dieser Kriterienliste jeweils einzelfallbezogen grundlegend abzuwägen, da jede Datenerhebungsmethode hier spezifische Eignungen aufweist, wie nachfolgend dargestellt wird.

Überblick: Die Methodengruppe der Befragungen

Drei der fünf oben aufgeführten, in Lehrevaluationen weit verbreiteten Methoden sind der Methodengruppe der Befragung[49] zuzuordnen: Fragebogen, Offenes Interview und Gruppendiskussion. Dementsprechend verfügen sie über einige Gemeinsamkeiten, die vor der Betrachtung der Spezifika der einzelnen Methode zu beleuchten sind.

Befragungen stellen das Standardinstrument der Datenerhebung in der empirischen Sozialforschung dar. Sie sind wohl die am meisten eingesetzten Verfahren und werden gelegentlich als Königsweg der Sozialforschung bezeichnet[50] (Atteslander 2008, S. 101 ff., Diekmann 2007, S. 371 ff., Kromrey 2006a, S. 358 ff.). Allgemein wird unter dem Begriff Befragung die verbale Kommunikation von mindestens zwei Personen gefasst (Diekmann 2007, S. 371, Atteslander 2008, S. 101, Abel/Möller/Treumann 1998, S. 52). In einer Befragung werden durch verbale Stimuli (oftmals Fragen Fragen) Reaktionen (Antworten) hervorgerufen (Atteslander 2008, S. 101). Somit sind Befragungen prinzipiell sehr nah an der alltäglichen Kommunikation, worin sicherlich ein Grund

[49] Die Begriffe *Befragung* und *Interview* werden häufig synonym verwendet. In dieser Arbeit wird zur Bezeichnung der Methoden*gruppe* der Begriff *Befragung* verwendet, um sie von der Methode des (quantitativen oder qualitativen) *Interviews*, bei der es sich um eine spezifische Form der Befragung handelt, abzusetzen.

[50] Einerseits trägt diese Bezeichnung natürlich der Relevanz des Verfahrens für die empirische Sozialforschung Rechnung. Andererseits dürfte es sich aber auch um einen augenzwinkernden Verweis auf René König handeln, einen der wichtigsten Begründer der modernen Sozialforschung im Nachkriegsdeutschland, der Bedeutendes zur Ausarbeitung der Methodologie der Befragung beigetragen hat (König 1974; Diekmann 2007, S. 371).

für Ihre Beliebtheit zu finden ist, da Forscher/-innen in ihrer Eigenschaft als sozial interagierende Menschen (vermeintlich) bereits über eine gute Kompetenz in den relevanten Bereichen verfügen. Wissenschaftliche Befragungen folgen jedoch den weiter oben in Kap. 3.1 dargestellten Regeln wissenschaftlicher Erfahrung und unterscheiden sich daher von alltäglicher Kommunikation in folgenden Punkten, wobei v. a. die beiden letztgenannten den Unterschied zwischen wissenschaftlicher und alltäglicher Form verdeutlichen[51]:

- Die interagierenden Personen sind in der Regel Fremde (Abel/Möller/Treumann 1998, S. 53).
- Die Beziehung zwischen Fragern und Befragten ist aufgrund der klaren Rollenverteilung asymmetrisch (Abel/Möller/Treumann 1998, S. 53). Üblicherweise agiert die fragende Person. Sie beeinflusst den Kommunikationsverlauf maßgeblich, während die befragte Person eher die Rolle eines Datenträgers anstatt eines klassischen Gesprächspartners innehat (Kromrey 2006a, S. 362).
- Die Ergebnisse der Kommunikation sind für die einzelnen Personen in der Regel folgenlos[52] (Abel/Möller/Treumann 1998, S. 53).
- Die wissenschaftliche Befragung unterscheidet sich von der alltäglichen Kommunikation durch die Kontrolliertheit jeder einzelnen Befragungsphase, also die systematische Steuerung (Atteslander/Kopp 1993, S. 148).
- Es handelt es sich um ein zielgerichtetes, formalisiertes Verfahren, um Informationen über interessierende Sachverhalte zu erheben (Kromrey 2006a, S. 360), d. h. die Fragen sind ausschließlich Mittel zum Zweck und folgen nicht unbedingt den üblichen sozialen Regeln der Kommunikation.

Zu den Vorteilen von Befragungen sind die aufgrund der Nähe zur Alltagskommunikation niedrige Schwelle sowohl für die Forscher/-innen als auch für die Probanden,

[51] Die nachfolgend genannten Punkte gelten prinzipiell für sämtliche Befragungsverfahren.
[52] Bei Evaluationen gilt dies nur unter der Voraussetzung der Vertraulichkeit bzw. einer den ethischen und datenschutztechnischen Standards und Erwartungen entsprechende Anonymisierung.

die vielfältige, sehr flexible Einsetzbarkeit des Verfahrens, die lange Anwendungserfahrung mit dieser Technik und das breite Spektrum unterschiedlichster erprobter und kodifizierter Befragungsverfahren zu zählen. Daneben sind Befragungen allerdings auch durch Nachteile gekennzeichnet. Die wesentlichsten Einwände sind, dass Befragungen erstens immer reaktiv sind, d. h. dass die Antworten u. a. von der Interviewsituation, dem Verhalten der interviewenden Person und den konkreten Messinstrumenten (Fragebogen oder Leitfaden) abhängen (Diekmann 2007, S. 371). Anders formuliert: Befragungen sind niemals neutral (Kromrey 2006a, S. 363). Zweitens lässt sich mit Befragungen zwar verbales, nicht aber soziales Verhalten insgesamt erfassen (Atteslander 2008, S. 101), denn die Äußerungen der Befragten lassen nicht unbedingt einen Rückschluss auf ihr tatsächliches Verhalten zu. Doch trotz dieser berechtigten Kritik stellt Diekmann fest, dass die Methode der Befragung nicht nur unverzichtbar ist, sondern dass wir ihr beispielsweise auch Kenntnisse über Sozialstruktur und soziale Schichtung, Bildungschancen sowie über andere sozialstrukturelle Merkmale und Zusammenhänge verdanken (Diekmann 2007, S. 371). Über diese Gemeinsamkeiten hinaus sind die einzelnen Befragungsverfahren – ebenso wie die übrigen Datenerhebungsverfahren – durch spezifische Eigenschaften gekennzeichnet, die durch die Konzeption des jeweiligen Verfahrens bedingt sind und nachfolgend betrachtet werden.

Quantitatives Interview/Fragebogen

Als quantitatives Interview werden Befragungsformen bezeichnet, die sich durch eine Erfassung der Kommunikation in Form von Quantifizierungen, also durch ihre Übertragung in Zahlenwerte, auszeichnen (Diekmann 2007, S. 373 ff., Kromrey 2006a, S. 358 ff., Roth 1993, S. 152 ff.). Üblich ist dabei die Umsetzung von Antworten in eine die jeweilige Antwort eindeutig repräsentierende Zahl. Ein prototypisches Verfahren des quantitativen Interviews ist der Fragebogen mit geschlossenen Antworten. Quantitative Interviews gehen in der Regel mit einer Standardisierung einher, u. a. deshalb, da andernfalls eine verlässliche, reproduzierbare Umsetzung der Antworten in Zahlenwerte nicht zu gewährleisten wäre.

Ein Beispiel für eine geschlossene Frage aus einem quantitativen (und standardisierten) Interviews ist die folgende Frage, die den Fragebögen zur universitätsweiten Lehrveranstaltungsevaluation der Philipps-Universität Marburg entnommen ist (FESEM, FEVOR, FEPRA, FEUEB: siehe Staufenbiel 2000, Klatt):

> „Der/Die Dozent/in fördert mein Interesse am Themenbereich".

Die Antwortmöglichkeiten zu dieser Frage lauten:

> „Stimmt nicht / Stimmt eher nicht / Stimmt eher / Stimmt"

Die eigentliche Quantifizierung wird nun bei der Datenübertragung vorgenommen. Jede der Antwortmöglichkeiten wird nach der Dateneingabe durch eine Zahl repräsentiert. „Stimmt nicht" würde also beispielsweise durch eine *1* repräsentiert, „Stimmt eher nicht" durch eine *2* und so weiter[53].

Beim o. a. Beispiel handelt es sich um eine *vollstandardisierte* Frage, bei der den Befragten nicht die Möglichkeit eingeräumt wird, eine Antwort zu geben, die außerhalb der Vorgaben liegt. Eine nicht den Vorgaben entsprechende Beantwortung der Frage – im obigen Beispiel etwa die Auswahl zweier Antwortmöglichkeiten – wird als Messfehler behandelt. Gelegentlich werden ergänzend auch offene Fragen mit in Fragebögen integriert, wodurch ein *teilstandardisierter* Fragebogen entsteht. Hier haben die Befragten die Möglichkeit, eine Antwort einzutragen, die bei der Erstellung des Fragebogens nicht antizipiert wurde. Prototypisch wäre hier etwa die Antwortmöglichkeit *Sonstiges*, gefolgt von Freiraum, in den die befragte Person einen frei formulierten Text eintragen kann (vgl. Atteslander 2008, S. 121 ff., Schnell/Hill/Esser 2005, S. 321 ff.; S. 423 ff.). Ein teilstandardisierter Fragebogen enthält demnach sowohl geschlossene als auch offene Fragen, es handelt sich dabei nicht mehr um ein rein quantitatives Interview.

Diese Form der Teilstandardisierung bietet den Probanden mehr Freiraum bei der Beantwortung der Frage, was v. a. dann von Interesse ist, wenn nicht alle möglichen Antworten vorweggenommen werden können. Allerdings bringt sie die Schwierigkeit mit

[53] Erfolgt die Dateneingabe manuell, so werden die Antworten durch die erfassenden Personen transformiert, bei automatisierten, computergestützten Erfassungsverfahren erfolgt die Transformation i. d. R. automatisch nach vorher festgelegten Regeln. Die Zuordnung von Zahlen zu Antworten erfolgte in diesem Beispiel willkürlich und orientiert sich nicht an den Vorgaben des Instruments.

sich, dass die Antworten ggf. nicht mehr sinnvoll in Zahlenwerte übertragen werden können, beispielsweise dann, wenn in das Freifeld viele verschiedene Antworten eingetragen werden, die sich nicht oder nur sehr selten wiederholen, es sich also vornehmlich um Einzelnennungen handelt. Solche Einzelnennung erschweren die Auswertung der Daten.

Zwar handelt es sich beim Fragebogen gewissermaßen um den Inbegriff des quantitativen Interviews, allerdings gibt es daneben auch andere Formen. So können beispielsweise auch verbal Interviews geführt werden, deren Kommunikation durch einen standardisierten Leitfaden strukturiert wird, der geschlossene Fragetypen enthält. In diesem Falle sitzen sich reale Personen gegenüber und die Antworten der Probanden werden durch die Interviewer direkt in Zahlenwerte übertragen und entsprechend festgehalten.

Werden die auf Seite 78 dargestellten Kriterien zur Klassifikation von Datenerhebungsmethoden auf das quantitative Interview angewandt, so ergibt sich das in Tabelle 3.2 dargestellte Profil[54].

Merkmal	**Verortung des *quantitativen Interviews***
Art der Datenerfassung	standardisiert oder teilstandardisiert, quantitativ
Grad der Strukturierung	strukturiert
Art der Kommunikationsvermittlung	medial vermittelt oder persönlich, mündlich oder schriftlich, u. U. visuell, synchron oder asynchron
Reaktivität	reaktiv, interaktiv oder nicht interaktiv
Größe der erfassbaren Stichprobe	sehr große Stichproben möglich
Beteiligung von Personen	Abhängig vom Umfang der Erhebung und ihrem organisatorischen Rahmen
Zeitbedarf	niedriger bis hoher Zeitbedarf
Nähe zum Gegenstand	ungefiltert

Tabelle 3.2.: Merkmale des quantitativen Interviews in Bezug auf die Gegenstandsangemessenheit

[54] Die Darstellung der Ausprägungen in dieser und den nachfolgenden thematisch gleichen Tabellen, bezieht sich auf *Tendenzen*. Durch Modifikationen der Datenerhebungsmethode können selbstverständlich Veränderungen der einzelnen Merkmalsausprägungen herbeigeführt werden.

Aus dieser Tabelle lässt sich ableiten, dass sich quantitative Interviews insbesondere für Datenerhebungen eignen, bei denen

- Sachverhalte betrachtet werden, die gut in quantifizierter Form abgebildet werden können,
- ein strukturierter Datenerhebungsprozess zur Anwendung kommen kann,
- die Reaktivität der Methode entweder wenig problematisch ist oder aus anderen Gründen keine Alternative existiert und
- ungefilterte Informationen, also primäre Daten erhoben werden sollen.

Hinsichtlich der übrigen oben dargestellten Aspekte ist das Verfahren des quantitativen Interviews sehr flexibel:

- Die mögliche Teilstandardisierung, also die Integration offener Elemente in Fragen, gestattet es, auch Sachverhalte zu bearbeiten, deren mögliche Ergebnisse nicht gänzlich vorweggenommen und in Antworten transformiert werden können.
- Quantitative Interviewverfahren sind mit beinahe jeder Art der Kommunikationsvermittlung kompatibel. Diese Vielseitigkeit erstreckt sich sowohl auf den Grad der Präsenz (räumliche Anwesenheit der beteiligten Personen), die Form der Kommunikation (mündlich/schriftlich/visuell) als auch die Ausprägung der Gleichzeitigkeit (synchron/asynchron).
- Die Größe der zu erfassenden Stichprobe ist beim standardisierten Interview von nachrangiger Bedeutung. Sehr kleine Stichproben können je nach eingesetztem Verfahren ebenso erfasst werden wie sehr große.
- Der reale Zeitbedarf hängt stark vom konkreten Einsatz ab, ist also in Grenzen dem jeweiligen Rahmen anpassbar.

Offenes Interview

Auch das offene oder nicht standardisierte Interview stellt eine Form der Befragung dar[55]. Im Gegensatz zum quantitativen Interview werden hier die erhobenen Daten jedoch nicht in verschlüsselter, quantifizierender, sondern in freier Textform[56] – *qualitativ* – erfasst. Dabei werden die Daten meist in persönlicher, mündlicher Form erhoben (Aghamanoukjan/Buber/Meyer 2007, S. 417), oftmals in einer Face-to-face-Situation oder per Telefon (Lamnek 2005, S. 330/331, Flick 2009b, S. 113 ff.).

Ein Kennzeichen offener Interviews ist der Verzicht auf eine Standardisierung der Datenerhebung (Lamnek 2005, S. 352), d. h. das eigentliche Interviewgespräch verläuft strikt nicht in vorher festgelegten Bahnen. Stattdessen erfolgt die Erhebung der Daten in freier Form, die eher einem alltäglichen Gespräch ähnlich ist und den befragten Personen Raum lässt, den Verlauf der Datenerhebung aktiv zu beeinflussen[57].

Hinsichtlich der konkreten Ausgestaltung offener Interviews kann zwischen einer Vielzahl von Verfahren unterschieden werden, die jeweils bestimmte Aspekte und Fragen betonen oder auch unterschiedliche Probandengruppen fokussieren (Aghamanoukjan/Buber/Meyer 2007, Helfferich 2009, S. 35 ff., Fontana/Frey 2005 oder Lamnek 2005, S. 330 ff.). Ein wichtiges Unterscheidungsmerkmal der verschiedenen Ansätze ist der Grad der Strukturierung des Gesprächs (Lamnek 2005, S. 356 ff.): Es existieren sowohl Verfahren, die auf einem gänzlich freien, nicht durch die datenerhebenden Personen aktiv beeinflussten Verlauf des Gespräches aufbauen, als auch Verfahren, bei denen der Verlauf des Gesprächs durch ein mehr oder weniger loses Themengerüst, oftmals in Form eines Gesprächsleitfadens, strukturiert wird. Dabei bezieht sich die Strukturierung lediglich auf den Verlauf der Datenerhebung, Antwortmöglichkeiten sind bei einem offenen Interview zu keiner Zeit vorgegeben. Weitere mögliche Unterscheidungs-

[55] In diesem Abschnitt werden Einzelinterviews betrachtet. Gruppendiskussionen, die je nach konkreter Umsetzung auch eine Art des offenen Interviews darstellen können, werden zu einem späteren Zeitpunkt separat untersucht.

[56] Der Begriff *Text* bezieht sich hier nicht ausschließlich auf schriftlich fixierte, sondern auf sprachliche Äußerungen jeder Art.

[57] Auch, wenn die Erhebungssituation einem alltäglichen Gespräch mitunter sehr nahe kommen kann, so behalten die in Abschnitt 3.3.1 aufgeführten Aspekte zu den Unterschieden zwischen alltäglicher und wissenschaftlicher Kommunikation uneingeschränkt ihre Gültigkeit.

dimensionen sind etwa die Art des produzierten Textes (beispielsweise *Erzählung* oder *Antwortsammlung*), die Art der Führung des Gesprächs (beispielsweise *interessiertes, passives Zuhören* oder *intendierte Beeinflussung des Gesprächsverlaufs*), die Fokussierung auf ein bestimmtes Thema (beispielsweise *Themengewichtung ausschließlich durch die befragte Person* oder *Themenfestlegung durch den/die Interviewer/-in*) oder die Art der zu gewinnenden Informationen (beispielsweise *subjektive Informationen aus Perspektive der befragten Person* oder *gezielte Beantwortung von Fragen der Forscher/-innen*).

Der Wert dieser Gruppe von Befragungsverfahren liegt vorrangig in der methodischen Offenheit der Antworten: Während geschlossene Fragen in vollstandardisierten Erhebungsverfahren ausschließlich die Aspekte erheben, die den Forscher/-innen bei der Erstellung des Erhebungsinstruments wichtig erschienen, ermöglichen offene Verfahren die Erfassung einer probandenzentrierten Sicht auf den Forschungsgegenstand bzw. das Forschungsproblem. Weitere Gründe für die Beliebtheit offener Interviews sind nach Lamnek 1. ein vergleichsweise leichter Zugang zum Feld, 2. der elaborierte Entwicklungsstand von Verfahren zur Analyse von Texten und 3. die vielfältigen und unkomplizierten Möglichkeiten zur unverfälschten, authentischen Aufzeichnung der erhobenen Informationen (Lamnek 2005, S. 329). Demgegenüber liegt das wohl wesentliche Problem offener Interviews in der in Abhängigkeit vom jeweiligen Interviewverfahren mitunter geringen Vergleichbarkeit der Fälle: Je freier ein Gespräch geführt wird, desto weniger lässt es sich mit einem anderen ebenso frei geführten Gespräch vergleichen[58]. Anhand der Kriterien zur Klassifikation von Datenerhebungsmethoden lässt sich das offene Interview verorten, wie in Tabelle 3.3 auf S. 94 dargestellt.

Demnach eignet sich die Methode des offenen Interviews besonders für Datenerhebungen, bei denen

- subjektive Sichtweisen, probandenzentrierte Problemsichten, in die Tiefe gehende Informationen im Zentrum stehen sollen,

[58] Dieses Problem der geringen Vergleichbarkeit ist kein an das offene Interview gebundenes, sondern generell reaktiven Verfahren mit nicht maximaler Standardisierung zu Eigen.

- ein strukturierter Datenerhebungsprozess entweder nicht erforderlich ist oder nicht zur Anwendung kommen könnte,
- Daten persönlich, mündlich und synchron erhoben werden können,
- die Reaktivität der Methode unproblematisch ist oder aus anderen Gründen keine Alternative existiert,
- keine großen Stichproben zu bearbeiten sind,
- der Zeitbedarf gedeckt werden kann und
- ungefilterte Daten erhoben werden sollen.

Der Vorteil des Verfahrens ist vor allem in der Möglichkeit zu sehen, mit Hilfe offener Interviews vielschichtige subjektive Sichtweisen zu gewinnen, die eher ins Detail gehen denn einen Überblick liefern. Ansonsten ist das offene Interview durch stärkere methodisch bedingte Restriktionen gekennzeichnet, als das quantitative Interview.

Merkmal	**Verortung des *offenen Interviews***
Art der Datenerfassung	offen oder teilstandardisiert, qualitativ
Grad der Strukturierung	offen oder teilstrukturiert
Art der Kommunikationsvermittlung	meist[59] persönlich, mündlich und synchron
Reaktivität	reaktiv, meist interaktiv
Größe der erfassbaren Stichprobe	eher für kleine bis mittlere Stichproben geeignet
Beteiligung von Personen	Abhängig vom Umfang der Erhebung und ihrem organisatorischen Rahmen
Zeitbedarf	mittlerer bis hoher Zeitbedarf
Nähe zum Gegenstand	ungefiltert

Tabelle 3.3.: Merkmale des offenen Interviews in Bezug auf die Gegenstandsangemessenheit

Dokumentenanalyse und Verwendung von Kennzahlen

Bei den Verfahren der Dokumentenanalyse und der Verwendung von Kennzahlen handelt es sich um unterschiedliche Techniken. Im Kontext der Lehrevaluation an Hoch-

[59] *Meist* bezieht sich auf die dominante (Aghamanoukjan/Buber/Meyer 2007, S. 417) Form einer Face-to-face- oder telefonischen Kommunikationssituation.

schulen werden sie jedoch häufig sehr ähnlich eingesetzt[60], so dass ihre gemeinsame Betrachtung zweckmäßig erscheint.

Bei der Dokumentenanalyse, die letztlich eine Spielart der Inhaltsanalyse darstellt, werden die auszuwertenden Informationen nicht kommunikativ erhoben, wie es bei den beiden vorgenannten Verfahren der Fall ist, sondern es wird auf bereits in Dokumenten fixierte Informationen zurückgegriffen[61] (Lamnek 2005, S.502). Ziel der Dokumentenanalyse ist die Extraktion der Informationen, die einen Beitrag zur Beantwortung der Forschungsfrage leisten können. Im Kontext von Lehrevaluationen könnten beispielsweise Dokumente wie Vorlesungsverzeichnisse, Seminarbeschreibungen, Seminarpläne oder bereits vorliegende Evaluationsergebnisse in die Analyse einbezogen werden. Die Erfassung der relevanten Merkmale der vorliegenden Dokumente kann dabei sowohl qualitativ als auch quantitativ erfolgen, je nach Erkenntnisinteresse und geplantem Fortgang der Datenauswertung.

Die Analyse von Kennzahlen bezeichnet Verfahren, bei denen quantitative Informationen über den zu untersuchenden Gegenstand gewonnen und ausgewertet werden. In Abgrenzung zum quantitativen Interview werden die Informationen jedoch – ebenso wie bei der Dokumentenanalyse – nicht durch Kommunikation produziert, sondern es wird auf Daten zurückgegriffen, die durch eine Betrachtung und Analyse des Gegenstandes selbst gewonnen werden können. Im Rahmen der Lehrevaluation könnten dies etwa Maßzahlen wie die Anzahl von Studierenden pro Veranstaltung, das Betreuungsverhältnis von Lehrenden zu Studierenden, Abbrecherquoten oder auch Absolventenzahlen sein. Viele dieser Zahlen werden oftmals ohnehin in Form von Hochschulstatistiken erhoben und liegen somit direkt in verwendbarer Form vor.

Im Gegensatz zu den vorgenannten Verfahren handelt es sich bei der Dokumentenanalyse und der Verwendung von Kennzahlen um nicht reaktive Verfahren der Datener-

[60] Im *Glossar der Evaluation – Eval-Wiki* etwa werden die Methoden unter dem Oberbegriff *Erfassung vorhandener Erzeugnisse und Verhaltensspuren* gebündelt (Beywl).

[61] Im Kontext dieser Arbeit wird der Begriff *Dokumentenanalyse* dann verwendet, wenn die Gewinnung von Informationen aus unabhängig von der konkreten Untersuchung produzierten Dokumenten betrieben wird. Daneben sind auch Szenarien denkbar, in denen die zu analysierenden Dokumente dediziert für die Untersuchung erstellt werden. In der hier vertretenen Lesart würde dieses Vorgehen jedoch als qualitatives, textbasiertes Interview klassifiziert. An dieser Stelle wird deutlich, dass die Benennung der einzelnen Verfahren an sich nicht trennscharf ist.

hebung. Gleichzeitig stellen beide Verfahren eine Mischform zwischen Datenerhebung und -auswertung dar, da die Daten durch die Auswertung bereits existierender Daten erhoben werden. Dabei ist die Auswertung eher mit einer systematischen Filterung als mit einer interpretativen Auswertung zu vergleichen und damit letztlich eine Analogie zum Leitfaden oder Fragebogen, der ebenfalls die prinzipiell verfügbaren Informationen gezielt filtert und erhebt.

Hinsichtlich der Kriterien zur Klassifikation von Datenerhebungsmethoden zeichnen sich die Dokumentenanalyse und die Verwendung von Kennzahlen durch folgende Merkmale aus:

Merkmal	**Verortung der *Dokumentenanalyse***	**Verortung der *Verwendung von Kennzahlen***
Art der Datenerfassung	1. Pol (standardisiert vs. offen) entfällt, da nicht reaktives Verfahren; qualitativ oder quantitativ	1. Pol (standardisiert vs. offen) entfällt, da nicht reaktives Verfahren; quantitativ
Grad der Strukturierung	nicht zutreffend, da kein aktiver Kommunikationsprozess	
Art der Kommunikationsvermittlung	medial vermittelt, schriftlich, asynchron	
Reaktivität	nicht reaktiv, nicht interaktiv	
Größe der erfassbaren Stichprobe	Abhängig vom konkreten Analyseverfahren sowohl für sehr kleine als auch für sehr große Stichproben geeignet	sehr kleine bis sehr große Stichproben
Beteiligung von Personen	wenige bis keine notwendigerweise in die Erhebung involvierte Personen, Abhängig von der Kooperation Anderer im Zusammenhang mit den Zugangswegen zu den Dokumenten	
Zeitbedarf	Abhängig von der Analysemethode	nicht zeitintensiv, falls die Daten bereits existieren
Nähe zum Gegenstand	gefiltert	

Tabelle 3.4.: Merkmale der Dokumentenanalyse und der Verwendung von Kennzahlen in Bezug auf die Gegenstandsangemessenheit

Die Verfahren der Dokumentenanalyse und der Analyse von Kennzahlen eignen sich demnach besonders gut für Szenarien, in denen

- bereits Daten in Form von Dokumenten oder Kennzahlen vorliegen bzw. in denen aus vorliegenden Informationen Kennzahlen generiert werden können,
- nicht reaktive Verfahren erforderlich sind,

- der Grad der Einbeziehung weiterer Personen gering gehalten werden soll und
- gefilterte Daten Verwendung finden können.

Der zentrale Vorteil dieses Datenerhebungsverfahren ist, dass eben auf bereits existierende Informationen zurückgegriffen werden kann. Gerade unabhängig von der konkreten Untersuchung produzierte Dokumente und Daten versprechen einen vergleichsweise unverfälschten Blick auf das Geschehen. Außerdem sind die Daten in vielen Fällen relativ leicht und schnell verfügbar. Demgegenüber steht die Voraussetzung, dass überhaupt Dokumente bzw. Daten existieren müssen, deren Analyse hinsichtlich des Untersuchungsziels ergiebig ist. Das Hauptproblem dieser Methode dürfte gerade darin liegen, dass die Passgenauigkeit zwischen existierenden Daten und intendierter Erkenntnis möglichst hoch sein muss, um relevante Ergebnisse erarbeiten zu können – auch hier zeichnet sich letztlich das Problem der Gegenstandsangemessenheit (hier: der existierenden Daten) ab: Da die Daten nicht zum ausdrücklichen Zweck der Beantwortung der bestimmten Forschungsfrage produziert wurden, muss gründlich überprüft werden, welchen Beitrag sie zur deren Beantwortung leisten können.

Gruppendiskussion

Bei der Analyse der Literatur zum Thema Gruppendiskussionen wird schnell ersichtlich, dass dieses Datenerhebungsverfahren in den letzten zehn Jahren mehr und mehr Beachtung in der Sozialforschung findet, und zwar auch außerhalb der Markt- und Meinungsforschung, wo es traditionell einen angestammten Platz im Kanon der verwendeten Datenerhebungsmethoden innehat (Lamnek 2005a, S. 11/12). Gleichzeitig wird deutlich, dass für dieses Verfahren eine Vielzahl unterschiedlicher Deutungen und praktischer Ausgestaltungen existiert. Im Rahmen dieser Arbeit wird der Datenerhebungsmethode Gruppendiskussion eine eher weit gefasste Definition zugrunde gelegt, die auf Morgan (1997) zurückgeht und beispielsweise auch von Lamnek präferiert wird: In Gruppendiskussionen werden Daten durch die Interaktionen von Gruppenmitgliedern gewonnen, wobei die Thematik durch das Interesse der Forscherin/des Forschers

gelenkt wird. In der Regel finden Gruppendiskussionen unter Laborbedingungen statt[62] (Lamnek 2005, S. 416, Lamnek 2005a, S. 27). Damit stellen Gruppendiskussionen eine besondere Ausprägung des Interviews dar. Allerdings erfassen sie nicht nur einfach mehrere Meinungen[63] auf einmal, sondern es lassen sich mit diesem Verfahren andere Erkenntnisse gewinnen als mit Methoden, bei denen immer die Meinung nur einer Person erfasst wird, sei es im Gespräch oder im Fragebogen. Während bei Einzelinterviews die besonderen Sichtweisen der einen befragten Person erhoben werden sollen, richtet sich das Erkenntnisinteresse der Gruppendiskussion auf andere Ebenen, die Lamnek in die Bereiche *Nicht-öffentliche Meinung*, *Informelle Gruppenmeinung*, *Situationskomponente der Gruppenmeinung* und *Ermittlung kollektiver Orientierungsmuster* einteilt (Lamnek 2005a, S. 51 ff). Anhand dieser Einteilung wird ersichtlich, dass sich Gruppendiskussionen sowohl zum Erfassen von in der Gruppe geäußerten Meinungen bezüglich des Gegenstandes der Diskussion als auch zur Beobachtung der Entstehungsprozesse von Meinungen oder auch Gruppenprozessen eignen. Auf dieser Basis lassen sich zahlreiche Einsatzmöglichkeiten für Gruppendiskussionen finden, so etwa Gruppendiskussionen als therapeutisches Instrument, als Pretest-Methode oder auch explizit als Evaluationsmethode (Lamnek 2005a, S. 59 ff). Auch lässt sich hinsichtlich der Zielsetzung der Gruppendiskussion zwischen ermittelnder (inhaltliche bzw. Sachinformationen evozierender) und vermittelnder (Veränderungen bei den Teilnehmenden hervorrufender) Ausrichtung unterscheiden (Lamnek 2005a, S. 59 ff). Entscheidend ist jedoch stets die Wirkung der Gruppensituation, in der beeinflussende Faktoren wie etwa argumentative Gruppenbildungen, strategisches Argumentieren oder auch soziale Erwünschtheit zum

[62] Dem gegenüber finden sich in der Literatur zahlreiche Abstufungen und Abgrenzungen zwischen verschiedenen Varianten der Gruppendiskussion. So wird von der Gruppendiskussion häufig die Focus Group abgesetzt, in deren Rahmen sich zwar auch mehrere Personen unter Laborbedingungen zu einem Thema unterhalten, die jedoch vor Beginn des Gesprächs beispielsweise durch einen filmischen Input o. ä. auf das Thema fokussiert werden. Auch wird der Verlauf einer Focus Group allgemein als weniger offen, sondern als eher ergebnisorientiert beschrieben (vgl. hierzu v. a. Blank 2007, S. 283 ff., Bohnsack/Przyborski 2007, S. 493 ff. sowie Lamnek 2005a, S. 26 ff.). Das Verfahren der *Online-Gruppendiskussion* (vgl. hierzu Kelle/Tabor/Metje 2009) wird hier ebenfalls nicht betrachtet, da es, so das Ergebnis der Analyse, im Kontext von Lehrevaluation bislang nahezu keine Rolle spielt.

[63] Die als ideal bezeichnete Größe einer Gruppe schwankt je nach Autor. Die breiteste Spanne reicht von drei Personen bis zu 20 Personen als möglicher Gruppenumfang, wobei diese Maximalgröße nicht als unbedingt ideal bezeichnet wird (Mangold 1973, S. 229).

Tragen kommen. Die Erfassung der Daten kann sowohl qualitativ als auch quantitativ erfolgen: Die quantifizierende Untersuchung von Redezeiten, Begriffsverwendungen oder der Reihenfolge der Wortergreifung kann ebenso interessant sein wie die offene Erfassung der Gesprächsinhalte.

Gruppendiskussionen lassen sich innerhalb der Kriterien zur Klassifikation von Datenerhebungsmethoden wie folgt verorten:

Merkmal	**Verortung der *Gruppendiskussion***
Art der Datenerfassung	standardisiert, teilstandardisiert oder offen; quantitativ oder qualitativ
Grad der Strukturierung	strukturiert oder unstrukturiert
Art der Kommunikationsvermittlung	persönlich; mündlich; synchron
Reaktivität	reaktiv; interaktiv
Größe der erfassbaren Stichprobe	je Diskussionsrunde kleinere Gruppen, insgesamt eher kleine Stichproben
Beteiligung von Personen	wenige notwendigerweise involvierte Personen; hohe Abhängigkeit von deren Bereitschaft zur Mitarbeit
Zeitbedarf	hoher Zeitbedarf
Nähe zum Gegenstand	ungefiltert

Tabelle 3.5.: Merkmale der Gruppendiskussion in Bezug auf die Gegenstandsangemessenheit

Anhand dieser Zuordnung wird ersichtlich, dass Gruppendiskussionen eine Spielart des Interviews sind, die je nach konkreter Ausgestaltung Merkmale sowohl der quantitativen als auch der offenen Variante in sich vereinen bzw. aufweisen kann. Die Besonderheiten dieser Methode ergeben sich demnach eben vor allem aus der Gruppensituation. Dieses Merkmal bestimmt auch für welche Art von Datenerhebungen Gruppendiskussionen besonders geeignet sind, nämlich für solche, in denen

- die Prozesse (etwa Meinungsbildungsprozesse, Argumentationsstrukturen, Untergruppenbildung, Effekte sozialer Erwünschtheit) innerhalb einer Gruppe von Interesse sind oder

- eine nivellierte Gruppenmeinung erhoben werden soll oder

- verschiedene Sichtweisen kontrastierend erhoben und ggf. gleichzeitig konfrontativ gegeneinander gestellt werden sollen und
- es möglich ist, die erforderlichen Personen auch tatsächlich zur Teilnahme zu bewegen[64].

Verfahren mit Gutachtern

Bei Verfahren mit Gutachtern werden Daten in Form der Urteile begutachtender Personen erhoben. Die Grundannahme, die hinter dieser Erhebungsform steht, ist, dass sachkompetente Personen in der Lage sind, ein treffendes und wertvolles Urteil über den begutachteten Gegenstand abzugeben. Um ein gewisses Maß an Intersubjektivität der Beurteilung herzustellen und ggf. extreme Einzelmeinungen zu relativieren, können im Rahmen solcher Verfahren mehrere Gutachter eingesetzt werden. Diese Verfahren sind im Kontext der Wissenschaftsevaluation sehr verbreitet und verfügen über eine lange Tradition, Bornmann und Daniel sprechen von mehr als 300 Jahren (Bornmann/ Daniel 2003, S. 207). Während es in der Schulevaluation wohl das Standardinstrument der Evaluation darstellt, ist es im Rahmen der Hochschulevaluation nicht allzu verbreitet, findet aber zunehmend Anwendung, beispielsweise in Verfahren der Akkreditierung neuer Studiengänge, in welchen sogenannte Akkreditierungskommissionen in der Regel eine wichtige Rolle spielen und die geplante Lehre auf Aspekte wie Relevanz, Studierbarkeit und Vollständigkeit der Themen hin inspizieren und beurteilen.

Aus methodischer Sicht kann diese Art von Gutachterverfahren, wie die Gruppendiskussion, auch als ein Sonderfall des Interviews betrachtet werden. Bei dieser Sichtweise stehen die Gutachter in der Rolle der befragten Personen, die ihr Urteil über den betrachteten Gegenstand äußern. Das Urteil kann dabei sowohl quantitativer Natur sein, etwa in Form von Benotungen, als auch in qualitativer Form, also offen bzw. nicht-quantifiziert, vorliegen. Viele der übrigen Aspekte dieser Variante der Datenerhebung sind variabel. So kann etwa der Prozess der Betrachtung des Gegenstandes sowohl

[64] Nachteilig könnte sich hier beispielsweise ein Machtgefälle zwischen den an der Diskussion beteiligten Personen auswirken und zwar vor allem dann, wenn das in der Diskussion Geäußerte negative Folgen für die Sprecherin bzw. den Sprecher nach sich ziehen könnte.

standardisiert als auch offen oder auch strukturiert oder unstrukturiert erfolgen. Die Begutachtung selbst kann sowohl als direkte Betrachtung, beispielsweise in Form von Hospitationen, erfolgen, es kann aber auch auf anderes Material wie etwa Dokumente zum Gegenstand, z. B. Lehrpläne oder Arbeitsmaterialen, zurückgegriffen werden. Somit erfolgt die Datenerhebung auf Seiten der Gutachter basierend auf den sonst üblichen Verfahren, etwa dem Interview, der Beobachtung oder der Dokumenten- und Kennzahlenanalyse.

Eine wesentliche Differenz zu anderen Formen des Interviews liegt in der Bedeutung, die den Ergebnissen der Erhebung, also hier den Urteilen der Gutachter zugesprochen wird: Während die Aussagen, die durch Interviews gesammelt werden, in aller Regel einer tieferen Analyse unterzogen und damit eher als Rohmaterial betrachtet werden, haben die Daten eines Gutachterverfahrens bereits den Status eines Ergebnisses. Zwar sind auch diese Ergebnisse und Daten üblicherweise weiter auszuwerten, allerdings oftmals eher rezipierend und Themen herausarbeitend denn klassisch interpretativ deutend und auslegend.

Der zentrale Vorteil dieses Verfahrens liegt darin, Urteile kompetenter Personen über den betrachteten Gegenstand zu erhalten. Der Wert dieser Urteile liegt in dem idealerweise fachlich qualifizierten Blick aus einer von außen kommenden Perspektive. Gerade in dieser Stärke liegt jedoch auch der zentrale Nachteil der Methode begründet, denn die Zumessung der erforderlichen Kompetenz ist nicht immer zweifelsfrei möglich. Befinden sich allerdings nicht kompetente Personen in der Rolle der Gutachter, so ist der Wert dieses Verfahrens stark geschmälert. Darüber hinaus besteht die Gefahr, dass durch gruppendynamische Prozesse innerhalb des Gutachtergremiums u. U. Urteilstendenzen entstehen, die den Ursprung nicht unbedingt zwingend in der Realität des betrachteten Gegenstandes haben müssen, sondern in Effekten wie beispielsweise der sozialen Erwünschtheit begründet sein können.

[65] Das Kriterium der Größe der erfassbaren Stichprobe ist auf dieses Verfahren nicht unmittelbar übertragbar. Betrachtet man die Gutachter als Stichprobe – da schließlich sie um ihre Einschätzung gebeten sind –, so lassen sich mit diesem Verfahren eher kleinere Stichprobengrößen bearbeiten, vor allem, da es sich um ein für die Gutachter durchaus zeitaufwändiges Verfahren handelt, ein gründliches Vorgehen vorausgesetzt.

Merkmal	**Verortung von *Gutachterverfahren***
Art der Datenerfassung	standardisiert, teilstandardisiert oder offen; quantitativ oder qualitativ
Grad der Strukturierung	strukturiert oder unstrukturiert
Art der Kommunikationsvermittlung	persönlich oder medial vermittelt; mündlich, schriftlich oder visuell; synchron oder asynchron
Reaktivität	reaktiv oder nicht reaktiv; interaktiv oder nicht interaktiv
Größe der erfassbaren Stichprobe	kleine Stichproben[65]
Beteiligung von Personen	wenige notwendigerweise involvierte Personen; hohe Abhängigkeit von deren Bereitschaft zur Mitarbeit
Zeitbedarf	variiert nach Art, Umfang und Tiefe des Gegenstandes und der Betrachtung
Nähe zum Gegenstand	gefiltert

Tabelle 3.6.: Merkmale von Gutachterverfahren in Bezug auf die Gegenstandsangemessenheit

Zusammenfassung: Empirische Methoden in der Lehrevaluation und ihre spezifischen Eigenschaften

Die genauere Betrachtung von unterschiedlichen Datenerhebungsmethoden ist ein wesentlicher Schritt bei der Prüfung der Eignung von Methoden für einen intendierten Erhebungszweck oder -kontext, da jede Methode über ihr eigenen Merkmale und Eigenschaften verfügt, die sich unmittelbar auf mögliche Einsatzzwecke auswirken. Schließlich steht hinter der Entwicklung jeder Datenerhebungsmethode auch der Anspruch, mit ihrem Einsatz einem bestimmten empirischen Problem begegnen zu können. Um diesen Vergleich zu ermöglichen, wurden auf Seite 78 ff. Kriterien entwickelt und vorgestellt, die für die Einschätzung der Gegenstandsangemessenheit einer Methode relevante Eigenschaften aufgreifen und für einen Vergleich nutzbar machen. Durch die Einordnung der Methoden hinsichtlich dieser Kriterien entstehen typische Profile, die helfen, die Eignung einer Datenherhebungsmethode für den beabsichtigten Zweck einzuschätzen.

Wesentlich ist, dass es sich bei der in diesem Kapitel vorgenommenen Darstellung der Methoden um eine prototypische handelt. In der Praxis ist es mitunter erforderlich, eine Datenerhebungsmethode zu verändern, um ihre Eignung für den konkret zu

erfassenden Gegenstand gewährleisten zu können. Dementsprechend existieren viele Spielarten der Methoden, die teilweise nur wenig, teilweise aber auch recht stark von den hier dargestellten Ausprägungen abweichen können. Aber gerade in dieser Möglichkeit der Modifikation einer Methode liegt die wichtige Chance, eine möglichst optimale Passung herzustellen. Als Grundlage und gleichzeitig Maßstab dieser Modifikationen sollten indes die in Kapitel 3.2, S. 61 ff. dargestellten Aspekte zur Bestimmung der Gegenstandsangemessenheit herangezogen werden, um eine möglichst zielführende Anpassung zu gewährleisten.

Neben der Anpassung einer Methode kann auch die Kombination verschiedener Methoden von besonderer Bedeutung sein. Die Grundidee ist dabei, die Nachteile bzw. weniger passende Aspekte einer infrage kommenden Datenerhebungsmethode durch den Einsatz einer weiteren aufzuwiegen. Dieser Ansatz der Verbindung verschiedener Methoden ist dabei keineswegs auf die Phase der Datenerhebung beschränkt, sondern erstreckt sich unter Bezeichnungen wie *Mixed Methods*, *Methodenintegration* oder *Triangulation* oftmals auf den gesamten Forschungsprozess. Beschreibungen verschiedener Ansätze und Verfahren zur planvollen und zweckmäßigen Kombination der unterschiedlichen Methoden sind in der Literatur ausgearbeitet (siehe etwa Tashakkori 2003, Kelle 2008, Flick 2008 sowie Creswell 2009). Vor allem diese Kombination unterschiedlicher Verfahren bietet die Möglichkeit, die spezifischen Vorteile der jeweiligen Methoden zu kombinieren, um so ein möglichst detailliertes und umfassendes Bild des Evaluationsgegenstandes mit all seinen Facetten zu zeichnen.

Wird in der Hochschule die Güte einer Lehrveranstaltung untersucht, so werden in der Regel die Studierenden als unmittelbar durch die Lehre betroffene Personen nach ihrem Urteil über eben diese Lehre gefragt[66]. Demnach werden viele der benannten Verfahren in Form von studentischer Veranstaltungskritik eingesetzt. Das hauptsächlich eingesetzte Instrument dürfte dabei der Fragebogen sein (Richter 1994, S. 50 ff.), aber auch

[66] Die Bezugnahme auf Studierende ist mehr oder weniger integraler Bestandteil eines Großteils der Publikationen zu diesem Thema. Siehe hierzu etwa Rindermann 2009, S. 26 ff., S. 201 ff., Pohlenz 2008, S. 76 ff., Pasternack 2004, S. 100 ff. und Rindermann 2003a.

andere Verfahren wie beispielsweise eine Art studentischer Gruppendiskussion über die Qualität der Lehre am Ende einer Veranstaltung sind mittlerweile weit verbreitet. Der organisatorische Rahmen dieser Evaluationsaktivitäten reicht dabei von wenig strukturierten, nicht weiter eingebundenen Gesprächen über die Lehrveranstaltung bis hin zu komplexen und hochschulweit einheitlich gestalteten und organisierten Prozessen. In den meisten Fällen der Bewertung der Lehre wird zur Gewinnung der zugrunde liegenden Daten auf die oben dargestellten Datenerhebungsmethoden zurückgegriffen, die dabei allerdings nicht zwingend in den engen Bahnen ihrer methodischen Definitionen gefasst werden: Gerade im Rahmen der weniger strukturierten Verfahren werden Varianten der Verfahren eingesetzt, die sich zwar an bestimmten Datenerhebungsmethoden orientieren oder ihnen zuordnen lassen, gleichzeitig aber freier und weniger strikt durchgeführt werden. So kann das bereits erwähnte wenig strukturierte Gespräch am Ende eines Semesters über den Verlauf, die Stärken und Schwächen einer Lehrveranstaltung zwar als verwandt mit der Gruppendiskussion betrachtet werden, gleichzeitig wird dabei jedoch nicht unbedingt im Sinne methodischer Strenge gehandelt, um mit der Methode verbundene Nachteile zu kontrollieren oder den Informationsgewinn aus dem Einsatz des Verfahrens zu maximieren. Allerdings dürfte auch diese Art Anpassung auf einer bestimmten Vorstellung der Angemessenheit der Methoden beruhen, die sich vermutlich jedoch wohl eher aus pragmatischen (*Was kenne ich? Was ist mit wenig Aufwand verbunden?* etc.) denn aus methodischen Überlegungen speisen dürfte. Durch eine solche Entscheidungsgrundlage sind Evaluationsergebnisse jedoch deutlich ihrer Aussagekraft und Glaubwürdigkeit eingeschränkt, da permanent der Vorwurf im Raum stehen kann, dass die Daten nicht geeignet sind, die gestellten Fragen zu beantworten.

3.3.2. Gegenstandsangemessen – aber für wen? Die Stakeholder von Lehrevaluation an Hochschulen

Evaluationen, deren Ergebnisse für Veränderungen genutzt werden sollen, haben in der Regel Einfluss auf oder sind von Interesse für einen mitunter weiten Personenkreis, die sog. *Stakeholder*. Als Stakeholder einer Evaluation werden, wie bereits in Abschnitt

2.1.2 erläutert, Personen bezeichnet, die Interesse an einer Evaluation bzw. an ihren Ergebnissen haben oder in anderer Weise durch sie berührt werden. Hierbei lässt sich zwischen *Beteiligten* und *Betroffenen* unterscheiden. Als Beteiligte werden jene Personen bezeichnet, die aktiv an der Evaluation mitwirken, beispielsweise in auftraggebender, planender, durchführender oder datengebender Funktion. Die Gruppe der Betroffenen umfasst jene Personen, auf die die Ergebnisse der Evaluation Auswirkungen entfalten, ohne dass sie selbst aktiv in den Prozess eingebunden sind.

Die Betrachtung der Stakeholder einer Evaluation ist im Kontext der Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden unerlässlich, denn Stakeholder werden sich in aller Regel ein Urteil über die Güte der Evaluation bilden. Die Daten, auf denen die Ergebnisse beruhen, sind hier von besonderem Interesse, da sie einen zentralen Angriffspunkt für mögliche Kritik durch die Stakeholder bieten: Argumente, dass beispielsweise die falschen Fragen gestellt oder relevante Personen (-gruppen) nicht angemessen berücksichtigt wurden, sind schwerwiegend und nach Abschluss der Evaluation wohl kaum noch einvernehmlich zu klären. Rindermann hat diese Problematik – etwas weiter gefasst – folgendermaßen dargestellt:

> „Auch Wottawa und Thierau [...] erwähnen, daß bei Evaluationsvorhaben in umstrittenen Bereichen oder mit heiklen Ergebnissen häufig versucht wird, die erzielten Resultate hinsichtlich der verwendeten Verfahren zu kritisieren, da diese ‚nicht idealen methodischen Anforderungen' entsprächen, was jedoch bei jeder Evaluationsstudie möglich ist [...]. Als Argumentationsfigur werden nach Wottawa und Thierau [...] gerne verwendet: ‚methodische Schwächen', ‚es wurde nicht alles berücksichtigt und ausgewertet', ‚fehlerhafte Operationalisierung', ‚interne Widersprüche', ‚singulär und nicht verallgemeinerbar'und ‚Widerspruch zu gesicherten Resultaten oder angesehenen Experten'." (Rindermann 2009, S. 201)

Gerade, wenn diese Form der Kritik, die sich gegen die Methodik der Evaluation richtet, durch Stakeholder hervorgebracht wird, die durch anstehende Veränderungen betroffen sind, entsteht ein deutliches Akzeptanzproblem. Stakeholder werden aufgrund ihres eigenen Interesses ein Urteil über die Angemessenheit der einer Evaluation zugrunde liegenden Daten fällen, das ggf. Auswirkungen auf die Nutzung der Ergebnisse hat. Wie wichtig die Berücksichtigung der Vorstellungen der Stakeholder eingeschätzt

wird, zeigt sich auch daran, dass, wie in Abschnitt 3.2.2 ausgeführt, die Stakeholder auch einen wesentlichen Bezugspunkt innerhalb der Standards für Evaluation darstellen: Sieben der insgesamt 25 Standards (vgl. Tabelle 3.1, S. 74) fokussieren die untersuchten Subjekte, die Forscher/-innen und/oder die Zielgruppe der Untersuchung. Diese Berücksichtigung der Stakeholder basiert auf methodischen Überlegungen und ist auf Fairness ausgerichtet. Vor diesem Hintergrund fordern Becker-Richter et al., die Betroffenen von Anfang an in die Evaluation einzubeziehen (Becker-Richter et al. 2002, S. 14).

Die Stakeholder von Lehrevaluationen an Hochschulen

Das Spektrum der Personen, die Interesse an der Evaluation universitärer Lehre haben oder haben könnten, ist breit gefächert. Aus dem in Abschnitt 2.2 dargestellten Hintergrund der Evaluation an Hochschulen lassen sich unterschiedliche Gruppen ableiten, die entweder von der Evaluation direkt betroffen sind oder aber zumindest Interesse an den Ergebnissen haben können:

- Die Studierenden in ihrer Eigenschaft als zentrale Klienten der Hochschullehre,
- die Lehrenden als die die Praxis der Lehre gestaltenden Personen,
- die Hochschulverwaltung als organisatorisch agierende Instanz,
- die Hochschulleitung als für die Hochschule und ihre Ausrichtung verantwortliche Einheit,
- die Politik als für die Bildungspolitik verantwortlicher Akteur,
- die Gesellschaft, einerseits als Finanzier der Bildung und andererseits als ihr Renditeempfänger,
- Arbeitgeber als auf die während des Studiums erworbenen Kompetenzen angewiesene Gruppe sowie
- Schulen mit ihren Schülern und Lehrern als zentrale, den Hochschulen vorgeschaltete Bildungsinstanz.

Aus dieser Liste lassen sich zwei Gruppen von Personen nach ihrer Distanz zum Evaluationsgegenstand unterscheiden: Einerseits jene, mit deren beruflicher oder hauptsächlicher Tätigkeit der Gegenstand der durchgeführten Evaluation unmittelbar in Zusammenhang steht, und andererseits die Gruppe derer, für die die Ergebnisse dieser Evaluation zwar von Interesse sind, aber eher informatorischen Zwecken dienen und ihren (Arbeits-) Alltag nicht betreffen. Bei der erstgenannten Gruppe handelt es sich somit um Personen, die in den Gegenstand der Evaluation *eingebunden* sind und sich durch eine geringe Distanz zu ihm auszeichnen, bei der zweiten Gruppe eher um *interessierte*, aber nicht unmittelbar in den Gegenstand involvierte Personen. Zur ersten Gruppe der Eingebundenen können die Studierenden und die Lehrenden als unmittelbar an der Hochschullehre beteiligte Personengruppen gezählt werden. Dies ist für die hier bearbeitete Forschungsfrage relevante Gruppe, da das Diskussionspotenzial um die Angemessenheit der Methoden hier naturgemäß am höchsten sein wird. Diese Gruppe ist in aller Regel sowohl durch die Erhebung der Daten als auch durch eventuelle Folgen einer Lehrevaluation direkt betroffen. Die übrigen der oben genannten Gruppen sind den Interessierten zuzuordnen, da sie am Lehr-/Lernprozess nicht unmittelbar beteiligt sind. Daraus folgt auf eine Formel gebracht: „Die via regia zur Bestimmung der Lehrqualität stellt somit die Heranziehung von Urteilen der Beteiligten dar.“ (Rindermann 2009, S. 26).

Diese beiden Gruppen der eingebundenen Stakeholder stellen die direkten Akteure universitärer Lehre dar. Die Studierenden stehen dabei in der Rolle der Empfänger und zentralen Adressaten der Lehre, deren Ziel es ist, sie auszubilden, ihr Wissen und ihre Fähigkeiten zu erweitern und ihr Denken zu schulen. Die Lehrenden hingegen sind die Personen, welche die Praxis dieser Lehre maßgeblich gestalten, sowohl hinsichtlich der Inhalte als auch der didaktischen Form. Auch wenn sicherlich jedes Individuum innerhalb dieser Gruppen eigene auf Erfahrungen und Erwartungen beruhende und damit sehr persönliche Sichtweisen auf die Lehre mitbringen wird, so ist doch zu erwarten, dass sich aufgrund der formalen Zugehörigkeit zu einer der Stakeholdergruppen durchaus auch gemeinsame, eben positionsabhängige Sichtweisen ergeben.

Die Unterscheidung nach der Nähe zum Evaluationsgegenstand liegt quer zur oben dargestellten Unterscheidung zwischen Beteiligten und Betroffenen, da in beiden Distanzgruppen beide Stakeholdertypen vertreten sein können. Es wäre nicht ungewöhnlich, wenn im Rahmen einer Lehrevaluation nur ein Sample der Studierenden und nicht alle befragt werden. Dabei würde die Gruppe der Eingebundenen sowohl aus beteiligten als auch betroffenen Personen bestehen. Ein anderes Beispiel könnte eine Evaluation sein, die durch eine Fachschaft angeregt und durchgeführt wird. In dieser Situation würden interessierte Personen zu Beteiligten, falls die Fachschaftsvertreter/-innen nicht ohnehin der Gruppe der Eingebundenen angehören. Somit ergibt sich eine Unterscheidung zwischen der formalen Position innerhalb der Hochschule, die in der Unterscheidung *Eingebunden – Interessiert* ihren Niederschlag findet, und der funktionalen Position innerhalb einer Evaluation, die sich in den Rollen der Beteiligten und Betroffenen ausdrückt.

3.4. Überblick: Gegenstandsangemessenheit und Lehrevaluation an Hochschulen

Die bisher erfolgte Darstellung verdeutlicht, dass die Bewertung der Gegenstandsangemessenheit von Datenerhebungsmethoden im Kontext von Lehrevaluationen fünf zentrale Bezugspunkte hat:

1. Den Rückgriff auf das methodische Instrumentarium der empirischen Sozialforschung und die damit einhergehenden Implikationen,
2. die Anforderungen der Standards für Evaluation als modellübergreifender und prinzipiell akzeptierter Qualitätsansatz,
3. die methodenimmanenten Eigenschaften,
4. den Rahmen, in dem die Evaluation stattfindet und die sich daraus ergebenden spezifischen Anforderungen sowie
5. die Stakeholder der Evaluation und ihre spezifische Sichtweise.

Der erstgenannte Aspekt – der in Kapitel 2.1.1 dargestellte Rückgriff der Evaluationsforschung auf das methodische Instrumentarium, aber auch die Logik der empirischen Sozialforschung – betrifft das grundlegende Vorgehen zur Gewinnung von Erkenntnissen. Hier ist vor allem das methodische und systematische Vorgehen hervorzuheben, das sich durch Regelgeleitetheit, Explizierbarkeit, Bewusstheit und zielgerichtetes Handeln auszeichnet.

Dieser Gesichtspunkt bestimmt noch nicht die Wahl einer konkreten Datenerhebungsmethode, steckt dafür aber sehr wohl einen Rahmen ab, da vorgegeben wird, *dass* eine empirische Methode einzusetzen ist. Gleichzeitig eröffnet sich ein Spektrum in Frage kommender Methoden zur Auswahl von Probanden, zur Erhebung und zur Auswertung von Daten. Für die Erhebung von Daten lassen sich drei Grundmethoden identifizieren, nämlich die Befragung, die Beobachtung und die Inhaltsanalyse. Über die Eingrenzung von Methoden hinaus entsteht durch diesen Rückgriff, wie in Kapitel 3.2 dargestellt, unmittelbar eine Grundlage zur Entwicklung von Kriterien für die Bewertung der Angemessenheit von Methoden. Dabei kann die Frage nach der Gegenstandsangemessenheit eine doppelte Funktion erfüllen: Sie kann sowohl als Entscheidungsgrundlage bei der Wahl von Methoden als auch als Gütekriterium zur nachträglichen Bewertung dieser Wahl herangezogen werden[67].

Bezogen auf die Frage nach der Gegenstandsangemessenheit von Datenerhebungsmethoden bedeutet *gegenstandsangemessen*, dass die gewählte(n) Methode(n) sämtliche für das jeweilige Forschungsvorhaben relevanten Daten und Informationen erfassen. Gleichzeitig sollen sie die Besonderheiten von Forschungsgegenstand (beispielsweise die Sprachkompetenz der Probanden) und Forschungsfrage (etwa die Einnahme eines speziellen Blickwinkels) sowie weitere beeinflussende Faktoren (z.B. Zeit- und Ressourcenbeschränkungen) berücksichtigen (siehe S. 61). Dieser methodenbezogenen Sichtweise liegt die Annahme zugrunde, dass jede Methode über ihr eigene spezifische Eigenschaften verfügt, aus denen sich Vor-, aber auch Nachteile für ein konkretes Forschungsvorhaben ergeben können. Gleichzeitig ist darauf hinzuweisen, dass die Wahrnehmung der

[67] Dieser Doppelcharakter gilt demnach auch für die konkreten Kriterien, die in den Kapiteln 3.2 und 3.3 vorgestellt wurden.

Besonderheiten von Forschungsvorhaben und -gegenstand positionsabhängig sind, d. h. dass verschiedene Stakeholder in Abhängigkeit von ihrer Position auch verschiedene Besonderheiten sehen können, was ggf. zu unterschiedlichen Bewertungen und Methodenpräferenzen führen kann.

Für die Bewertung der Angemessenheit von Datenerhebungsmethoden lassen sich verschiedene Kennzeichen heranziehen. Die Betrachtung folgender Bezugselemente ist besonders dazu geeignet, die Gegenstandsangemessenheit zu prüfen (siehe S. 63):

- Gegenstand
- Fragestellung
- Untersuchte Subjekte
- Zielgruppe der Untersuchung
- Historie der Methoden
- Verhältnis von Aufwand, Ressourcen und Nutzen
- Forschende
- Offenheit der Methode(n)

Hervorzuheben ist dabei, dass es sich nicht um zwingend gleichermaßen zu berücksichtigende Kriterien handelt, sondern um Ebenen, die bei der Beantwortung der Frage nach der Angemessenheit einer Datenerhebungsmethode sinnvoll in die Abwägung einbezogen werden können.

Der zweite Bezugspunkt zur Bewertung der Gegenstandsangemessenheit von Datenerhebungsmethoden, die Anforderungen der *Standards für Evaluation* als modellübergreifender und prinzipiell akzeptierter Qualitätsansatz, ist eng mit den Bestrebungen zur Professionalisierung von Evaluationsforschung verbunden. Mit den Standards für Evaluation hat die *DeGEval* einen Rahmen zur Steigerung, aber auch zur Beurteilung der Qualität einer Evaluation vorgelegt. 13 der 25 Standards beziehen sich direkt oder

indirekt auf die Frage der Gegenstandsangemessenheit (vgl. Tabelle 3.1), die damit bereits rein quantitativ einen besonderen Stellenwert hat. Bei genauerer Betrachtung dieser 13 Standards wird deutlich, dass die Konstruktion der Gegenstandsangemessenheit, wie sie in den Standards vorgenommen wird, auf die der allgemeinen empirischen Sozialforschung aufbaut, dabei aber einen evaluationsspezifischen Schwerpunkt auf die Stakeholder, insbesondere die untersuchten Subjekte, legt. Darüber hinaus haben auch die Zielgruppe der Untersuchung, das Verhältnis von Aufwand, Ressourcen und Nutzen sowie die Fragestellung zentrale Positionen inne. Diese Schwerpunktsetzungen ergeben sich aus den besonderen Anforderungen und Gegebenheiten im Bereich der Evaluationsforschung.

Für die Betrachtung des dritten Bezugspunktes, der immanenten Eigenschaften der zum Einsatz kommenden Datenerhebungsmethoden, ist es zunächst erforderlich, die in Frage kommenden Methoden zu bestimmen[68]. Folgende Methoden stellen gewissermaßen das Standardrepertoire[69] der Datenerhebungsverfahren in Lehrevaluation an Hochschulen dar:

1. Fragebogen
2. Offenes Interview
3. Dokumenten- und Kennzahlenanalyse
4. Gruppendiskussion
5. Verfahren mit Gutachtern

Jedes dieser Verfahren verfügt über charakteristische Eigenschaften. Jede dieser Eigenschaften wiederum hat Einfluss auf die Eignung einer Methode und somit auch auf

[68] Dies kann im Rahmen dieser Arbeit nur auf der vergleichsweise abstrakten Ebene von Datenerhebungsverfahrens*gruppen* erfolgen. Für eine detaillierte Betrachtung müssten konkrete Methoden in Erwägung gezogen werden können, wozu ein Forschungsprojekt unerlässlich wäre. Eine Betrachtung *aller* Datenerhebungsmethoden ist schlechterdings unmöglich.

[69] Ein tatsächlicher *Standard* (im Sinne von Vereinheitlichung) existiert in diesem Umfeld nicht. Jedoch hat sich über die Zeit ein Bestand von Methoden entwickelt, die besonders häufig zum Einsatz kommen. Diese zentralen Methoden werden hier näher betrachtet.

die Bewertung ihrer Angemessenheit. Zur Erfassung dieser inhärenten Eigenschaften werden folgende Kriterien vorgeschlagen (siehe S. 78 ff.):

- Art der Datenerfassung
- Grad der Strukturierung
- Art der Kommunikationsvermittlung
- Reaktivität
- Größe der erfassbaren Stichprobe
- Beteiligung von Personen
- Zeitbedarf
- Nähe zum Gegenstand

Entscheidend ist dabei, dass jede Methode zur Bestimmung ihrer Eignung individuell und genau begutachtet werden muss. So lassen sich beispielsweise aus der Betrachtung der Geschichte eines Verfahrens, v. a. aus dem Kontext, in dem es entwickelt wurde bzw. mit Blick auf das spezifische Problem, für dessen Lösung es entworfen wurde, Implikationen für die Frage nach der Passung zum zu untersuchenden Gegenstand ableiten. Auch wohl überlegte und begründbare Abwandlungen und Weiterentwicklungen einer bereits beschriebenen Methode können einen entscheidenden Beitrag zur Sicherstellung der Angemessenheit einer Methode darstellen. Gleiches gilt für die ergänzende Verwendung unterschiedlicher Methoden (Mixed Methods, Methodenintegration, Methodentriangulation) mit dem Ziel, die Nachteile einer durch die Vorteile einer anderen Methode auszugeichen bzw. spezifische Sichtweisen zu ergänzen.

Den vierten Bezugspunkt zur Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden im Kontext von Lehrevaluation bildet der Rahmen, in dem die Evaluation stattfindet, gemeinsam mit den sich daraus ergebenden spezifischen Anforderungen. Für die Lehrevaluation lassen sich, wie in Kapitel 2.2 dargestellt, auf struktureller Ebene drei zentrale Hintergründe feststellen:

1. Bestrebungen hochschulinterner Qualitätsentwicklung und -sicherung,
2. Controlling und Rechenschaftslegung gegenüber Ministerien und
3. Rechenschaftslegung bzw. Qualitätsnachweis gegenüber gegenwärtig Studierenden, Studieninteressierten und der Gesellschaft.

Neben diesem strukturellen gibt es noch den direkten Kontext. Von besonderer Bedeutung sind hier die individuellen Kennzeichen der zu evaluierenden Veranstaltung, so etwa die Organisationsform (Vorlesung, Übung etc.), die Anzahl der Veranstaltungsteilnehmer/-innen oder die didaktische Ausgestaltung (Präsenzveranstaltung, E-Learning-Veranstaltung o. ä.). Diese Merkmale üben direkten Einfluss auf die Wahl von Datenerhebungsmethoden aus. So bedarf etwa das Controlling, an dem Ministerien interessiert sind, anderer Daten als das Ziel der didaktischen Verbesserung eines Seminars. Es ist demnach erforderlich, Klarheit über die Zielsetzung und Verwendung der Evaluation zu erlangen. Anschließend können diese Überlegungen Eingang in den Entscheidungsprozess über angemessene Datenerhebungsmethoden finden.

Den fünften und letzten Bezugspunkt bilden, wie in Kapitel 3.3.2 dargestellt, die Stakeholder der Evaluation und ihre spezifische Sichtweise. Kritik der Stakeholder am Design und/oder den Methoden einer Untersuchung stellt ein ernsthaftes Hindernis für die Akzeptanz und Nutzung von Ergebnissen dar.

Innerhalb der Gesamtgruppe der Stakeholder kann einerseits zwischen Beteiligten und Betroffenen unterschieden werden (vgl. S. 15), andererseits zwischen Eingebundenen und Interessierten (vgl. S. 107 ff.). Für die in dieser Arbeit vorgenommene Untersuchung ist, wie gezeigt, die Gruppe der Eingebundenen von vorrangigem Interesse. Für den betrachteten Themenbereich setzt sie sich aus den Lehrenden und Studierenden zusammen.

In welcher Form die Stakeholder während der Planung und Durchführung einer Evaluation zu berücksichtigen sind, welche Themen und Fragen hier eine Rolle spielen, wird am besten durch die Analyse der Standards für Evaluation ersichtlich.

Die Wahl von Datenerhebungsmethoden findet im dargestellten Spannungsfeld zwischen den fünf Bezugspunkten zur Bewertung der Gegenstandsangemessenheit von Datenerhebungsmethoden in Evaluationen statt. Dass die verschiedenen, sich u. U. gar widersprechenden Anforderungen nicht ohne Weiteres zu vereinbaren sind, ist offensichtlich. Die Aufgabe der planenden Personen ist es, die unterschiedlichen Interessen, Fragen und Erfordernisse zu identifizieren, gegeneinander abzuwägen und auf dieser Basis eine Entscheidung zu treffen. Die Gewichtung der einzelnen Aspekte wird in jedem Fall individuell sein.

Die dargestellten Kriterien und Aspekte stellen eine gute Grundlage für die konkrete Auseinandersetzung mit der Frage nach der Angemessenheit dar. Dennoch bleibt an dieser Stelle eine für die Planenden einer Evaluation zentrale Frage offen, die sich aus der notwendigen und geforderten Berücksichtigung der Stakeholder ergibt und deutlich über die überblicksartige Darstellung in Kapitel 3.3.2 hinausgeht: *Welche Vorstellung von Gegenstandsangemessenheit herrscht innerhalb der Gruppe der Stakeholder vor?* Diese Frage besitzt besonderes Gewicht, da, wie ebenfalls in Kapitel 3.3.2 betrachtet, das Urteil der Stakeholder über die Angemessenheit einer Evaluation und ihrer Methoden sicher Einfluss auf die Akzeptanz haben wird.

Diese Forderung der Berücksichtigung der Stakeholder birgt selbstverständlich auch Probleme in sich. Zum einen werden die Stakeholder bzw. einzelne Stakeholdergruppen in den meisten Fällen eine spezifische Erwartung mit der Evaluation verknüpfen, beispielsweise die Verbesserung der eigenen Situation oder die Erleichterung des Arbeitsalltags. Neben diesen positiven Hoffnungen wird es auch sicher Befürchtungen geben, etwa die mögliche Beschränkung von Autonomien oder eine allgemeine Verschlechterung der persönlichen Situation. Es kann also nicht der Ausspruch *Stakeholder haben immer Recht* gelten. Vielmehr muss selbstverständlich auch hier darauf geachtet werden, dass keine Beeinflussung im Sinne einer Stakeholdergruppe erfolgt. Zum anderen dürften die Evaluationsteilnehmer/-innen wohl in den meisten Fällen nicht über fundiertes Wissen über Datenerhebungsmethoden verfügen. Dies bedeutet, dass sie ihr Ur-

teil eventuell auf Grundlage von Vermutungen, Halbwissen und/oder Vorurteilen fällen werden. Daraus folgt, dass die methodische Entscheidungskompetenz klar auf Seiten der Planenden bleiben muss. Es kann nicht darum gehen, eine Untersuchung zu entwerfen, deren Methoden sich vor allem oder ausschließlich am Interesse der untersuchten Subjekte messen. Gleichwohl ist es – wie dargestellt – erforderlich, die Betroffenen und Beteiligten zu hören, ernst zu nehmen und gründlich zu informieren. Die Partizipation darf jedoch nicht die Professionalität der Planenden untergraben, sondern soll vielmehr eine bessere Basis für die Durchführung der Evaluation und die Verwendung der Ergebnisse schaffen.

Diese Frage nach der Vorstellung der Gegenstandsangemessenheit kann nicht rein theoretisch beantwortet werden, sondern bedarf einer empirischen Untersuchung, die im nachfolgenden Kapitel vorgenommen wird.

4. Empirische Untersuchung zur Konstruktion von Gegenstandsangemessenheit

In den vorhergehenden Kapiteln standen der organisatorische Hintergrund von Lehrevaluation an Hochschulen sowie methodenbasierte Kriterien der Wahl angemessener Datenerhebungsverfahren im Fokus der Betrachtungen. Der folgende Abschnitt befasst sich mit den Stakeholdern einer Evaluation, da diese – wie gezeigt – eine wichtige Bezugsgröße bei der Bestimmung der Angemessenheit von Datenerhebungsmethoden darstellen und deswegen in besonderer Weise berücksichtigt werden sollten. Für die Planenden einer Evaluation bedeutet dies, dass die Frage nach der Sichtweise der Stakeholder auf die Methoden bei jeder Evaluation neu zu klären ist. Vorgaben der Standards für Evaluation und Grundsätze der empirischen Sozialforschung allgemein helfen zwar dabei, Aspekte zu identifizieren, die prinzipiell für die Stakeholder Relevanz besitzen können, konkretere Hinweise auf Kriterien und Ursachen für persönliche Bewertungen der Stakeholder bezüglich der Eignung einer Methode sind jedoch kaum auffindbar[70]. Um diese Lücke teilweise zu schließen und zu untersuchen, nach welchen Kriterien die Stakeholder einer Evaluation die Gegenstandsangemessenheit von Datenerhebungsverfahren für sich konstruieren, wurde eine empirische Untersuchung durchgeführt, deren Hintergrund, Verlauf und Ergebnisse nachfolgend dargestellt werden.

[70] Erfahrene Evaluatorinnen und Evaluatoren mögen über einen persönlichen Erfahrungsschatz bzgl. der Bedürfnisse, Vorstellungen, Erwartungen und Bewertungstendenzen von Stakeholdern hinsichtlich gewählter Methoden verfügen. Dieser wird aber in der Regel nicht offenbar.

4.1. Die Empirische Untersuchung – Beschreibung und Vorgehensweise

4.1.1. Ziele und Hintergrund

Thema dieser Arbeit ist die Frage, wie sich im Rahmen von Lehrevaluation an Hochschulen die Gegenstandsangemessenheit von Datenerhebungsmethoden bestimmen lässt. Diese Frage kann ohne die Berücksichtigung der Sichtweise der Stakeholder nur unzureichend bearbeitet werden. Deswegen wurde eine empirische Untersuchung durchgeführt, deren Ziel es war, Daten zu erheben, die helfen, die Vorstellungen von Stakeholdern bezüglich angemessener Datenerhebungsmethoden sichtbar zu machen. Anhand der Ergebnisse der empirischen Untersuchung sollen die in den vorhergehenden Kapiteln erarbeiteten theoretischen Erkenntnisse für die Evaluationspraxis nutzbar, d. h. möglichst verlässlich in die Planung einbeziehbar gemacht werden. Die empirische Untersuchung ist demzufolge der Frage gewidmet, nach welchen Kriterien Stakeholder einer Evaluation die Angemessenheit eingesetzter Datenerhebungsmethoden bewerten. Hier gilt es vorrangig zu untersuchen, mit welchen Argumenten Stakeholder eine Befürwortung oder Ablehnung einer Methode begründen, welche Methoden aus welchen Gründen bevorzugt werden und ob es Bezüge zwischen der Position einer Person gegenüber dem Gegenstand der Untersuchung und ihrer Vorstellung bezüglich der Angemessenheit eingesetzter Methoden gibt. Kenntnisse in diesen Bereichen können schließlich Personen, die eine Evaluation planen, dabei helfen, die Akzeptanz der gewählten Methoden einzuschätzen und ggf. auch zu steigern.

Der Versuch, die intrapersonalen Prozesse und Abläufe, die der Bewertung der Angemessenheit von Datenerhebungsmethoden zugrunde liegen, zu betrachten, birgt zwei wesentliche Problematiken: Zum einen handelt es sich vermutlich in der Mehrheit der Fälle um unbewusst ablaufende Prozesse, deren Regeln selbst durch die Betroffenen nur schwer ad hoc zu explizieren sein dürften. Eine Ausnahme wird bei Personen angenommen, die über entsprechende Vorerfahrungen verfügen und/oder bei denen die Entscheidungsprozesse eben nicht unbewusst, sondern bewusst ablaufen. Tatsächliche fachliche Vorbildung ist aber vermutlich in den meisten Fällen eher die Ausnahme denn

die Regel. Daran schließt sich die zweite Schwierigkeit an, nämlich der Umstand, dass praktisch *jede* Person zu einem Urteil über die Angemessenheit der eingesetzten Datenerhebungsmethoden kommen wird, und zwar auch dann, wenn ihr jegliche Basis für ein fundiertes Urteil fehlen sollte. In gewisser Weise befasst sich eine solche Art von Untersuchung also auch in großen Teilen mit *Vorurteilen* über Methoden. Gerade deshalb ist es wichtig, die Hintergründe der Konstruktion von Gegenstandsangemessenheit zu betrachten und mehr über sie zu erfahren. Schließlich bildet die persönliche Bewertung der eingesetzten Datenerhebungsmethoden eine wesentliche Grundlage des Urteils über die Evaluation und ihre Ergebnisse und ist damit entscheidend für deren Akzeptanz oder Ablehnung. Die methodische Herausforderung liegt nun darin, die Forschungsteilnehmer/-innen die unbewussten Prozesse explizieren, dabei die (Vor-) Urteile möglichst reflektieren und letztlich fundieren oder verwerfen zu lassen.

4.1.2. Rahmen der Untersuchung

Um eine Untersuchung mit der dargestellten Zielsetzung durchführen zu können, ist es angebracht, sie an eine real durchgeführte Evaluation anzubinden. So wird es möglich, einen relativ klaren Bezugspunkt für die erforderlichen Analysen zu schaffen und dem Problem rein hypothetischer Überlegungen entgegenzuwirken. Die Position einer Person gegenüber dem Gegenstand der Evaluation wird vor einem solchen realen Szenario bestimmbar. Gleichzeitig verfügen die Probanden als Stakeholder auch automatisch über Erfahrungen mit einer Evaluation, auf die für die Untersuchung zurückgegriffen werden kann. Überlegungen bzgl. der Angemessenheit bestimmter Methoden können direkt an einem Beispiel aus dem vertrauten Umfeld der Person diskutiert und erörtert werden. Darüber hinaus verfügen die Befragten über einen einheitlichen Bezugsrahmen, der es ermöglicht, ein gewisses Maß an Vergleichbarkeit herzustellen.

Die hier dargestellte empirische Untersuchung wurde im Kontext einer Evaluation umgesetzt, die im Zeitraum von Oktober 2009 bis Dezember 2010 durchgeführt wurde. Gegenstand der Evaluation war die fachdidaktische Ausbildung im Rahmen des Lehramtsstudiums an der Philipps-Universität Marburg. Die Evaluationsziele werden im Bericht zur Evaluation wie folgt dargestellt (Laging et al. 2010, S, 24)[71]:

[71] Dieser Bericht ist die Grundlage der gesamten nachfolgenden Darstellung der durchgeführten Eva-

„Das grundlegende Ziel ist die Generierung von Wissen über den Bestand und die Entwicklungsmöglichkeiten der Marburger Fachdidaktikausbildung, insbesondere zu

- den verschiedenen Umsetzungen der oben beschriebenen gesetzlichen Vorgaben,
- den Zusammenhängen zwischen der fächerinternen Fachdidaktikausbildung und den Fachwissenschaften sowie dem allgemeindidaktischen Bereich in den Erziehungswissenschaften und
- der personellen Ausstattung der Fachdidaktik.

Im Vordergrund steht die Verbesserung und Profilierung der Fachdidaktikausbildung in Marburg, womit verschiedene Teilaspekte verbunden werden können. [...] Ein weiteres Ziel zur nachhaltigen und angemessenen Etablierung der Fachdidaktiken in Marburg ist die Entstehung einer Konzeption zur zukünftigen Personalentwicklung.“

Es handelt sich demnach um breit angelegte und eher grundlegende Fragestellungen. Um die erforderlichen Antworten erhalten zu können, wurde eine umfassende Datenbasis angestrebt und es wurden verschiedene Datenquellen in die Erhebung einbezogen. Drei Datenerhebungsmethoden wurden aufeinander aufbauend und sich ergänzend eingesetzt:

1. Dokumentenanalysen,
2. Experteninterviews sowie
3. eine quantitative Online-Befragung.

Für die grundlegende Erfassung und Beschreibung der Inhalte der Fachdidaktik-Ausbildung wurden Dokumentenanalysen durchgeführt. Der formale Aufbau der Fachdidaktik-Ausbildung in den einzelnen Fächern, aber auch beispielsweise die Anzahl von Veranstaltungen, Semesterwochenstunden, die Verteilung von Leistungspunkten und die Anzahl von lehrenden Personen wurden den Modulhandbüchern und den während der Evaluation gültigen Vorlesungsverzeichnissen entnommen, die einer quantitativen Analyse unterzogen wurden. Ergebnis war eine tabellarische Übersicht der relevanten

luation.

Daten für alle Fächer. Daneben wurden der universitätsweiten Studierendenstatistik die Anzahl der Lehramtsstudierenden im Sommersemester 2009 und im Wintersemester 2009/2010 sowie häufige Studienfachkombinationen entnommen. Aus diesen Daten wurden fachspezifische Profilbeschreibungen erstellt.

Auf Grundlage dieser Daten wurden zwölf der insgesamt 20 Fächer, in denen an der Universität Marburg eine Lehramtsausbildung stattfindet, ausgewählt[72]. Im Rahmen qualitativer leitfadengestützter Interviews wurde je eine Expertin bzw. ein Experte für Fachdidaktik dieser Fächer befragt, i. d. R. die Person, die für die Koordination des fachdidaktischen Anteils des Studiums verantwortlich war. Die Interviews dienten dem Zweck, die Einschätzung der Experten bzgl. der Fachdidaktikausbildung ihres Faches zu erheben und die Hintergründe, Besonderheiten, Problemfelder sowie Verbesserungsmöglichkeiten der Fachdidaktikausbildung aus ihrer Sicht zu besprechen. In den verbleibenden acht Fächern wurden kürzere Gespräche geführt, unter anderem, um die Ergebnisse der Dokumentenanalyse zu validieren.

Die dritte Komponente der Datenerhebung bildete eine quantitative Online-Befragung von Studierenden. Alle im Sommersemester 2010 für ein Lehramtsstudium an der Philipps-Universität Marburg eingeschriebenen Studierenden, die mindestens im vierten und höchstens im zehnten Semester studierten, wurden per E-Mail eingeladen, den Fragebogen auszufüllen. Die Befragung wurde anonym durchgeführt und fokussierte verschiedene Aspekte der Fachdidaktik-Ausbildung in beiden Studienfächern. Sie beinhaltete u. a. Fragen zu den Themenkomplexen der subjektiven Wichtigkeit der Fachdidaktik-Ausbildung, nach der Wahrnehmung des Stellenwerts von fachdidaktischer und fachwissenschaftlicher Ausbildung und nach der Bewertung verschiedener Aspekte der konkreten Fachdidaktikausbildung, wie etwa dem theoretischen Anspruch oder der Relevanz für die spätere Berufspraxis. Ferner wurde nach der Einschätzung der eigenen fachdidaktischen Kompetenzen, aber auch nach der Zufriedenheit und Verbesserungsvorschlägen gefragt.

[72] Kriterien für die Auswahl von Fächern waren einerseits deren Größe (in Form der Anzahl an Lehramtsstudierenden), andererseits aber auch etwa ihre inhaltliche Ausrichtung (Berücksichtigung verschiedener wissenschaftlicher Richtungen) oder organisatorische Besonderheiten (Organisation und Personalbestückung des fachdidaktischen Ausbildungszweiges).

Diese Evaluation der fachdidaktischen Ausbildung im Rahmen der Lehramtsausbildung an der Philipps-Universität Marburg bildete den Rahmen, innerhalb dessen die empirische Untersuchung zu den Vorstellungen der Stakeholder bezüglich angemessener Datenerhebungsmethoden durchgeführt wurde. So bildete die Grundgesamtheit der Evaluation gleichzeitig die Grundgesamtheit für die hier vorgestellte Untersuchung. Die Themen der Fachdidaktik-Evaluation stellten den Kontext dar, der in der Untersuchung zur Gegenstandsangemessenheit aufgegriffen und als Bezugspunkt für die hier relevanten Themen genutzt wurde. Die Rollen, die die Stakeholder innerhalb der Evaluation innehatten, dienten hier als zentrales Kriterium zur Selektion der Probanden (siehe hierzu Abschnitt 4.1.4). Auf diese Weise wurde es möglich, die oben beschriebenen Vorteile der Anbindung der geplanten Untersuchung an eine real durchgeführte Evaluation auch tatsächlich nutzbar zu machen.

4.1.3. Forschungsmethoden und -design

Wie bereits dargestellt, lag die methodische Herausforderung dieser Untersuchung darin, den befragten Personen die meist unbewusst ablaufenden Prozesse ins Bewusstsein zu rufen. Dabei sollten sie ihre (Vor-) Urteile reflektieren und letztlich fundieren , verändern oder verwerfen. Datenerhebungsmethoden, die hierzu geeignet sein können, sollten drei Erfordernissen genügen:

- Sie bieten die Möglichkeit komplexer verbaler Interaktion.
- Die befragten Personen können frei und auf Basis ihrer persönlichen Sichtweise antworten.
- Die erhobenen Daten müssen Vergleiche ermöglichen.

Die *Möglichkeit komplexer verbaler Interaktion* ist wichtig, um die relevanten, vermutetermaßen unbewusst ablaufenden Prozesse sichtbar machen und Daten zur Beantwortung der Forschungsfragen evozieren zu können. Ein solches Vorhaben erfordert, dass die befragten Personen sich eingehender mit dem für sie in dieser Form oftmals

neuen Themengebiet auseinandersetzen können. Es ist beispielsweise damit zu rechnen, dass Fragen oder Unklarheiten auftreten, dass eine Person weitergehende Informationen benötigt oder mit den Anliegen des Interviewers zunächst nicht viel anzufangen weiß und deshalb weiterer Erläuterungen bedarf. Da die Findung des persönlichen Standpunkts durch die Probanden aller Voraussicht nach kein einfacher, schnell zu klärender Vorgang sein wird, muss zudem die Möglichkeit bestehen, das Urteil nach und nach in der Auseinandersetzung mit dem Thema zu konkretisieren und ggf. auch wieder zugunsten einer anderen Entscheidung verwerfen zu können. Diese Vorgänge erfordern eine datenerhebende Person, die ggf. im Gespräch unterstützt, nachfragt, erläutert und subsummiert, d. h. mit der befragten Person interagiert. Aufgrund des eventuell sehr unterschiedlichen Vorwissens der Probanden muss die Möglichkeit zu individuell den Erfordernissen angepasster verbaler Interaktion bestehen.

Die Erfordernis, dass die Befragten *frei und auf Basis der persönlichen Sichtweise antworten*, ist für das Forschungsvorhaben essenziell, denn es stehen gerade die individuellen Abläufe im Zentrum der Untersuchung.

Die prinzipielle *Möglichkeit des Vergleichs der erhobenen Daten* ist bedeutsam, damit aus den erhobenen Daten Aussagen generiert werden können, deren Bedeutung über den einzelnen Fall hinausreicht. Wesentlich ist dabei, dass die Forschungsteilnehmer/-innen während der Datenerhebung möglichst dieselben Themen in den Blick nehmen. Da bei wenig bewussten Themen nicht davon auszugehen ist, dass die Personen von sich aus auf relevante Aspekte zu sprechen kommen, sollte eine geeignete Datenerhebungsmethode die thematische Führung des Gesprächs erlauben, ohne den Verlauf oder gar bestimmte Antworten zu determinieren.

Geeignete Verfahren zu Datengewinnung müssten sich demnach durch folgende der auf S. 78 ff. vorgestellten Kriterien auszeichnen[73]:

- Art der Datenerfassung: offen
- Grad der Strukturierung: teilstrukturiert

[73] Nicht alle entwickelten Kriterien werden hier referenziert, sondern lediglich jene, die als Bedingung für die zu wählende Methode zu betrachten sind.

- Art der Kommunikationsvermittlung: persönlich, mündlich, synchron
- Reaktivität: interaktiv
- Nähe zum Gegenstand: ungefiltert

Angesichts dieser Erfordernisse wurde als hauptsächliche Datenerhebungsmethode das Verfahren offener, leitfadengestützter Einzelinterviews ausgewählt. Um auch die unter Umständen relevanten *Gruppen*prozesse der Meinungsbildung zu erfassen wurde, wie im Abschnitt 4.1.3 dargestellt wird, zusätzlich das Verfahren der Gruppendiskussion eingesetzt.

Forschungsdesign[74]

Die Untersuchung wurde im Querschnittsdesign realisiert, d. h. die Datenerhebung wurde einmalig durchgeführt. Die Datenerhebung fand nach Abschluss der Fachdidaktik-Evaluation im Zeitraum von Februar bis November 2011 statt. In die Datenerhebung einbezogen wurden Personen aus folgenden Gruppen:

- Beteiligte Studierende
- Betroffene Studierende
- Beteiligte Lehrende
- Betroffene Lehrende
- Planende

Bei den Personen dieser Gruppen handelt es sich einerseits um die eingebundenen Stakeholder einer Evaluation (vgl. S. 107 ff.) sowie andererseits um jene, die für die methodische Umsetzung der Untersuchung verantwortlich waren.

[74] Die nachfolgende Darstellung des Forschungsdesigns gibt einen Überblick über den groben Ablauf der empirischen Untersuchung. Eine detailliertere Darstellung der Arbeitsschritte und der eingesetzten Methoden findet sich in den Abschnitten *Die Probanden: Sampling und Stichprobe* und *Beschreibung und Dokumentation der Durchführung.*

Die Berücksichtigung der beiden Gruppen der Studierenden und der Lehrenden ist von Bedeutung, um untersuchen zu können, ob die Position einer Person gegenüber dem Untersuchungsgegenstand Einfluss auf ihre Vorstellung von der Angemessenheit der Datenerhebungsmethoden ausübt. Die Positionen sowohl der Studierenden als auch der Lehrenden zeichnet sich durch eine große Nähe zum Gegenstand der Fachdidaktik-Ausbildung aus, gleichzeitig unterscheiden sich die Blickwinkel auf die Materie jedoch deutlich: Während die Studierenden in erster Linie Adressaten der Ausbildung sind, stehen die Lehrenden in der Rolle der die Ausbildung verantwortenden Personen.

Die Unterscheidung zwischen beteiligten und betroffenen Personen ist von Belang, da die Möglichkeiten zur Einflussnahme auf die Evaluation und die Auseinandersetzung mit dem Gegenstand, wie dargestellt, Einfluss auf das Wissen über und das Verständnis für die Evaluation ausüben wird. Auch wenn die Betroffenen nicht an der Datenerhebung selbst teilgenommen haben, werden sich die eventuellen Ergebnisse und Folgen einer Evaluation in der gleichen Weise auf sie auswirken wie auf die Beteiligten. Die Unterschiede zwischen diesen Gruppen liegen also einerseits im Maß der persönlichen Einflussmöglichkeiten auf die Evaluation und andererseits darin, wie sie in die Erzeugung der Ergebnisse der Evaluation eingebunden wurden.

Die Sicht der planenden Personen schließlich komplettiert die Untersuchung. Hier sind vor allem die Fragen interessant, aus welchen Gründen Entscheidungen zugunsten der eingesetzten Datenerhebungsmethoden fielen und wie sich die Bewertung der Methoden bzw. die Intention ihres Einsatzes im Vergleich zur Sichtweise der übrigen Gruppen der Befragten darstellen.

Bei den befragten Gruppen handelt es sich summarisch um die zentralen, eingebundenen Stakeholder der Fachdidaktik-Evaluation zuzüglich der Planenden.

Aus jeder der o. g. Gruppen wurden sechs Personen in Einzelinterviews befragt. Eine Ausnahme bildeten die Planenden, da es sich hier nur um zwei Personen handelte, die beide in die Untersuchung einbezogen werden konnten. Insgesamt wurden 26 Einzelin-

terviews geführt. Die Interviews wurden durch Leitfäden[75] strukturiert. Die Leitfäden beinhalteten eine Sammlung relevanter Themen, die im Verlauf der Interviews, die ca. 30 – 45 Minuten dauern sollten, anzusprechen waren sowie Anmerkungen für den Interviewer. Drei Leitfadenvarianten wurden erstellt: Eine für Interviews mit Beteiligten, eine für Interviews mit Betroffenen sowie eine für Interviews mit Planenden. Die Beteiligten und die Betroffenen wurden im Verlauf des Interviews darum gebeten, im Rahmen eines Rollenspiels die Position der die Fachdidaktik-Evaluation planenden Person zu übernehmen und aus diesem Blickwinkel heraus die Eigenschaften von Datenerhebungsmethoden zu bewerten. Während die Leitfäden für die Beteiligten und die Betroffenen bis auf wenige Unterschiede identisch waren[76], musste sich der Leitfaden für die Interviews mit den planenden Personen unterscheiden, da hier das Rollenspiel nicht anwendbar war. Hier wurde ein Leitfaden eingesetzt, der sich stärker auf die Grundlagen der Entscheidung für die tatsächlich gewählten Methoden bezog.

In Ergänzung zu den Einzelinterviews wurde eine ermittelnde Gruppendiskussion durchgeführt, zu der Angehörige der vier erstgenannten Gruppen eingeladen wurden. Während in den Einzelinterviews die persönlichen Argumente, Sichtweisen und Bewertungen von Datenerhebungsmethoden im Mittelpunkt standen, diente die Gruppendiskussion dem Ziel, einen Einblick in die in einer Gruppe während der Diskussion über die Angemessenheit von Datenerhebungsmethoden ablaufenden Prozesse zu erhalten. Damit verbunden war die Erwartung, dass in einer solchen Gesprächssituation die Argumente verbalisiert und gegen andere abgegrenzt werden und somit eine andere Art von Einblick in die Genese der Urteile bzgl. der Angemessenheit von Methoden möglich wird. Vorrangig diente sie einer Perspektivvariation sowie dem Zwecke der Komplementierung und/oder Korrektur bzw. der Verdeutlichung der durch die Einzelinterviews erhobenen Informationen (vgl. Lamnek 2005a, S. 59 ff.).

[75] Die Interviewleitfäden sind im Anhang auf S. 238 ff. angefügt.
[76] Für eine detaillierte Darstellung der Leitfäden siehe S. 129 ff.

Die Probanden der Einzelinterviews und der Gruppendiskussion rekrutierten sich aus der Grundgesamtheit der Fachdidaktik-Evaluation sowie zusätzlich aus der Gruppe der Planenden. Zur Auswahl der Studierenden und Lehrenden konnte auf die Datensätze und Probandendefinitionen der Fachdidaktik-Evaluation zurückgegriffen werden. Diese Anbindung ermöglichte es, die Personen auch danach zu unterscheiden, ob sie an der Datenerhebung teilgenommen hatten oder nicht, ob es sich also um Beteiligte oder Betroffene handelte. Aus diesen Gruppen wurde per Zufallsauswahl eine Stichprobe gezogen. Außerdem wurden die beiden planenden Personen gebeten, an der Untersuchung teilzunehmen. Im Anschluss an die Datenerhebung wurde das Material transkribiert.

Die Auswertung der transkribierten Interviews erfolgte mit Hilfe von strukturierenden inhaltsanalytischen Techniken. Zunächst wurden Kategorien gebildet, anhand derer die Inhalte des Materials für die weitere Analyse zielgerichtet erfasst werden konnten. Die Kategorien wurden sowohl den Transkripten selbst entnommen (Induktion) als auch aus den Vorüberlegungen und -annahmen abgeleitet (Deduktion). Zum Einsatz kamen Kategorien, in denen Materialteile mit gleichen inhaltlichen Bezügen gesammelt wurden, die also der Strukturierung, Sammlung und Ordnung der Inhalte dienten. Außerdem wurden skalierende Kategorien verwendet, die es gestatteten, Aussagen der befragten Personen auf einer (nominalen) Skala einzuordnen und so eine überblicksartige, aber gleichzeitig differenzierte Sichtweise auf die Inhalte des codierten Materials ermöglichten. Durch die Kombination dieser Codierverfahren konnten sowohl die Besonderheiten der einzelnen Fälle erfasst als auch eine fallübergreifende Vergleichbarkeit hergestellt werden.

Die gewählten Methoden ermöglichten es, den Ansprüchen der Forschungsfrage und des Gegenstandes gerecht zu werden: Es gelang, die Balance zwischen der unabdingbaren Freiheit und Flexibilität und der analytisch erforderlichen Strukturierung herzustellen, und zwar sowohl in der Datenerhebung als auch in der Auswertung.

4.1.4. Die Probanden: Grundgesamtheit, Sampling und Stichprobe

Die Grundgesamtheit der vorliegenden Untersuchung definierte sich über jene der Fachdidaktik-Evaluation. Sie bestand demnach aus allen Personen, die im Sommersemester 2010 für ein Lehramtsstudium an der Philipps-Universität Marburg eingeschrieben waren und mindestens im vierten, maximal jedoch im zehnten Semester studierten, sowie aus den Personen, die in den Fächern, in denen eine Lehramtsausbildung erfolgt, Veranstaltungen anboten, die der Fachdidaktik-Ausbildung zuzurechnen waren. Zusätzlich zu diesen Personengruppen wurden auch die beiden die Fachdidaktik-Evaluation planenden Personen in die Untersuchung einbezogen.

Die Grundgesamtheit der Studierenden bestand aus 1705 Personen, von denen 670 den Online-Fragebogen ausgefüllt und somit aktiv an der Fachdidaktik-Evaluation teilgenommen hatten. Für die hier durchgeführte Untersuchung besteht die Grundgesamtheit der *beteiligten* Studierenden somit aus 670 Personen, die der *betroffenen* Studierenden aus 1035. Die Ziehung der Stichprobe erfolgte zufallsgesteuert, über die Einteilung anhand der formalen Stellung gegenüber der Fachdidaktik-Evaluation hinaus wurden keine weiteren Selektionskriterien angewendet. Die Probanden wurden auf einer Liste gesammelt, fortlaufend nummeriert und per Zufallszahlengenerator wurden 20 Personen aus jeder der beiden Gruppen ausgelost. Diese Personen wurden per personalisierter E-Mail kontaktiert. Die E-Mail enthielt grundlegende Informationen zu Form und Inhalt der Untersuchung sowie die Bitte, in einer kurzen Antwort mitzuteilen, ob die Person bereit sei, an der Untersuchung teilzunehmen oder nicht. Fünf bis sieben Tage nach der ersten Kontaktaufnahme wurden Personen, die bis dahin nicht geantwortet hatten, erneut angeschrieben und darum gebeten, kurz zu antworten und über Teilnahme oder Nichtteilnahme zu informieren. Eine weitere Kontaktaufnahme erfolgte nicht. In dem Fall, dass nach der ersten Anschreibewelle noch keine sechs Personen pro Gruppe für Interviews zur Verfügung standen, wurde das Verfahren so lange wiederholt, bis die angestrebte Quote erfüllt war. Sowohl in der Gruppe der beteiligten als auch der betroffenen Studierenden wurde das Verfahren je drei mal durchgeführt, es wurden also je 60 Personen kontaktiert[77].

[77] Von einer Erhöhung der Zahl der auf einmal kontaktierten Personen je Gruppe wurde abgesehen,

Für die Gruppe der Lehrenden lag keine komplette Personenliste vor. Zwölf Personen aus dieser Gruppe wurden im Rahmen der Fachdidaktik-Evaluation in ausführlichen Einzelinterviews befragt, weitere acht in Form kurzer Gespräche – diese Personen waren somit bekannt. Von dieser Gruppe waren lediglich die zwölf ausführlich interviewten Personen von Interesse, da bei Ihnen die Beteiligung an der Evaluation deutlich intensiver war, als bei den übrigen acht. Diese zwölf intensiv Befragten bildeten die Grundgesamtheit der *beteiligten* Lehrenden. Zur Identifikation der *Betroffenen* wurden aus dem Vorlesungsverzeichnis des betreffenden Semesters (Sommersemester 2010) Personen ermittelt, die Lehrveranstaltungen im Bereich der Fachdidaktik anboten. Das weitere Verfahren war identisch: Die Personen der Gruppen wurden auf einer Liste vermerkt, ihnen wurde eine Kennzahl zugeordnet und per Zufallszahlengenerator wurden die anzuschreibenden Personen ausgelost, per E-Mail kontaktiert und nach fünf bis sieben Tagen erinnert. Allerdings wurden hier in der ersten Kontaktierungswelle lediglich sechs Personen je Gruppe ausgelost. Da auch hier die Quoten nicht direkt erfüllt werden konnten, wurden im Anschluss weitere Personen angeschrieben, allerdings immer maximal so viele, wie zur angestrebten Stichprobengröße je Gruppe noch fehlten.

Für die Gruppendiskussion wurde prinzipiell identisch vorgegangen. Lediglich für die Gruppe der beteiligten Lehrenden wurde ein anderer Weg beschritten: Da sich die Zahl der Personen, die noch für eine Teilnahme an der Gruppendiskussion in Frage kamen, stark reduziert hatte – Voraussetzung war, dass die hier teilnehmenden Personen nicht auch schon in einem Einzelinterview befragt wurden –, wurde zunächst persönlich Kontakt zu den verbleibenden Personen aus dieser kleinsten der Gruppen aufgenommen. Nach der Zustimmung einer Person wurde ein Termin für die Durchführung der Gruppendiskussion vereinbart, der vom Tag der Zusage aus gerechnet rund fünf Wochen in der Zukunft lag. In der Zeit zwischen der Terminvereinbarung mit der Person aus der Gruppe der beteiligten Lehrenden und der Durchführung der Gruppendiskussion wurden Angehörige der drei übrigen Gruppen mit dem oben beschriebenen Verfahren kontaktiert. Hier konnte bereits bei der ersten Kontaktierung eine Person aus jeder der drei verbleibenden Gruppen für die Teilnahme gewonnen werden.

um im Falle der Bereitschaft zur Teilnahme möglichst keine Absagen erteilen zu müssen.

Für die Gruppe der Planenden wurde keine besondere Auswahltechnik eingesetzt. Beide Personen wurden persönlich kontaktiert und um Mithilfe gebeten, beide Personen willigten ein.

Nachdem eine Person in die Teilnahme an einem Interview oder der Gruppendiskussion eingewilligt hatte, wurde per E-Mail ein Termin für das Gespräch vereinbar. Außerdem wurde ihr ein Informationsanschreiben zugesendet, dass weiterführende Informationen zum Hintergrund und zum Ablauf der Untersuchung enthielt[78].

Die für die Durchführung der Untersuchung erforderlichen Personen konnten gewonnen werden. Für die Teilnahme an einem Einzelinterviews erklärten sich jeweils sechs beteiligte Studierende, betroffene Studierende, beteiligte Lehrende und betroffene Lehrende sowie die beiden für die Planung der Fachdidaktik-Evaluation verantwortlichen Personen bereit. Für die Gruppendiskussion konnte jeweils ein/e Vertreter/-in der vier erstgenannten Gruppen gewonnen werden.

4.1.5. Beschreibung und Dokumentation der Durchführung

Datenerhebung: Einzelinterviews

Alle Interviews wurden durch den Verfasser dieser Arbeit geführt. Die Einzelinterviews mit den Studierenden und ein Interview mit einer planenden Person wurden im Büro des Interviewers durchgeführt, die Gespräche mit den Lehrenden und das Gespräch mit der zweiten planenden Person fanden in den Büroräumen der Gesprächspartner statt. Die Aufzeichnung der Gespräche erfolgte mit Hilfe eines digitalen Audioaufzeichnungsgerätes. Der Ablauf der 26 geführten Einzelinterviews war stets identisch und folgte den im Leitfaden festgelegten Themen. Die Leitfäden für die Beteiligten und Betroffenen waren bis auf wenige Formulierungen identisch[79] und beinhalteten folgende Elemente:

[78] Die Informationsanschreiben, die den teilnehmenden Personen zugesendet wurden, sind im Anhang auf S. 251 (Einzelinterviews) und S. 252 (Gruppendiskussion) abgedruckt.

[79] Im Leitfaden für die Gespräche mit Betroffenen wurde lediglich auf drei Aspekte der Einbeziehung in die durchgeführte Fachdidaktik-Evaluation verzichtet: Erstens auf die Frage, wie die Person über die Ziele der Fachdidaktik-Evaluation informiert wurde (zugunsten der Formulierung, *ob* die Person informiert wurde), zweitens auf die Frage, wie sie in die Evaluation einbezogen wurde und drittens auf die Frage, ob die Person weiß, welche welche Rolle die damals erhobenen Daten innerhalb der gesamten Evaluation spielten.

1. Begrüßung und Einführung
2. Rekapitulation der Fachdidaktik-Evaluation
3. Erläuterung des Konstruktes *Gegenstandsangemessenheit*
4. Rollenspiel zum Thema Gegenstandsangemessenheit
5. Stellungnahme zu den tatsächlich eingesetzten Methoden
6. Stellungnahme zu den allgemein in Lehrevaluationen an Hochschulen eingesetzten Verfahren
7. Resümee
8. Offene Abschlussfrage

Die in den Leitfäden gesammelten anzusprechenden Themen wurden aus der theoretischen Bearbeitung des Themas der Gegenstandsangemessenheit abgeleitet. Die Leitfäden wurden in vier Pretest-Interviews, die nicht mit in die Analyse einbezogen wurden, getestet.

Den Auftakt der Einzelinterviews bildete zur Begrüßung und Einführung ein lockeres Gespräch, das vor allem der erneuten Darstellung des Forschungsvorhabens, seines Kontextes, der dahinter liegenden Intention sowie der Klärung eventuell noch offener Fragen diente. Dieser Teil des Gesprächs wurde nicht aufgezeichnet, war aber dennoch Bestandteil der Interviewleitfäden, da die Befragten hier auf das Thema der Gesprächs eingestimmt und erneut über die Datenerhebung, -erfassung und -verarbeitung informiert wurden. Somit diente es einerseits der Fokussierung, andererseits aber auch noch einmal der Erneuerung der informierten Einwilligung durch die Befragten.

Während der nächste Themenblock, die Rekapitulation, vor allem dazu diente, Hintergrundinformationen über den Wissensstand der befragten Person bezüglich der Fachdidaktik-Evaluation im Speziellen, aber auch hinsichtlich der Evaluationsforschung allgemein zu erheben, handelt es sich beim anschließenden Rollenspiel um den Kern der

Interviews. Die befragten Personen wurden nach einer präzisierenden Erläuterung gebeten, sich in die Rolle der die Fachdidaktik-Evaluation planenden Person zu versetzen. Aus dieser Rolle heraus sollten sie frei von Zwängen eine aus ihrer persönlichen Sicht möglichst gute Evaluation der Fachdidaktik-Ausbildung an der Philipps-Universität Marburg planen. Im Zentrum standen hier die Fragen, wen (welche Personengruppen) die Person befragen würde, welche Informationen diese Gruppen zur Bewertung beisteuern könnten und wie die Daten erhoben werden könnten. Dieser Teil des Leitfadens widmete sich dem persönlichen Verständnis der Angemessenheit von Methoden.

Der nächste Teil, für den die Befragten weiterhin aus der fiktiv eingenommenen Rolle einer planenden Person heraus urteilen sollten, befasste sich mit den tatsächlich im Kontext der Fachdidaktik-Evaluation eingesetzten Instrumenten (vgl. Abschnitt 4.1.2). Die Interviewpartner wurden gebeten, jede der Methoden hinsichtlich ihrer Eignung einzuschätzen. Dieser Teil des Gesprächs richtete das Vorverständnis der befragten Person auf die real durchgeführte Evaluation.

Daran anschließend wurden die Personen – immer noch in ihrer Eigenschaft als fiktiv die Evaluation Planende – gebeten, die Eignung der im Rahmen der Lehrevaluation an Hochschulen üblichen Methoden (vgl. Abschnitt 3.3.1) zu bewerten. Damit weitete sich der Blickwinkel auf den Kanon gängiger Methoden.

Für das Resümee wurden die Probanden gebeten, die Position einer planenden Person zu verlassen, um abschließend ein Urteil aus Sicht einer an einer Evaluation teilnehmenden Person abzugeben: Vor dem Hintergrund des Gesprächs sollten sie versuchen, Kriterien zu formulieren, nach denen sie Datenerhebungsmethoden als angemessen bewerten würden. Ziel war es, die in den vorherigen Teilen des Interviews erarbeiteten, jedoch i. d. R. lose nebeneinander stehenden Kriterien zusammenzuführen und dabei ggf. eine Gewichtung vorzunehmen.

Die offene Abschlussfrage bot den Befragten die Möglichkeit, Themen, die im Interview nicht zur Sprache kamen, ihnen in diesem Kontext jedoch wichtig waren, zur Sprache zu bringen.

Die Aufzeichnung des Gesprächs wurde beendet, nachdem die jeweils befragte Person keine weiteren Ergänzungen zu den Inhalten des Gesprächs mehr vornahm.

Etwas anders wurde in den Interviews mit den planenden Personen verfahren, da hier der fiktive Rollentausch in der Form, wie er bei den übrigen Befragten angewandt wurde, unpassend gewesen wäre. Demnach fokussierten sich diese beiden Gespräche vorrangig auf die Gründe für die Wahl der eingesetzten Methoden. Der Interviewleitfaden war wie folgt aufgebaut:

1. Begrüßung und Einführung
2. Rekapitulation der Fachdidaktik-Evaluation
   a) Verfolgte Ziele
   b) Kriterien der Probandenwahl
   c) Erhobene Daten sowie Gründe für die Wahl der zu erhebenden Daten
   d) Datenerhebungsmethoden und Gründe für ihre Wahl
3. Frage nach der vermuteten Zufriedenheit der Probanden mit den gewählten Methoden
4. Stellungnahme zu den allgemein in Lehrevaluationen an Hochschulen eingesetzten Verfahren
5. Resümee
6. Offene Abschlussfrage

Die Teile der Begrüßung und Einführung (1), der Stellungnahme zu den allgemein in Lehrevaluationen an Hochschulen eingesetzten Verfahren (4), des Resümees (5) sowie der offenen Abschlussfrage (6) waren identisch mit jenen der Interviews mit Beteiligten und Betroffenen. Die zentralen Teile der Rekapitulation der Fachdidaktik-Evaluation (2) und der Frage nach der vermuteten Zufriedenheit der Probanden mit den gewählten Methoden (3) hingegen dienten dem Zweck, die Beweggründe der Planenden für die Wahl und ihre Sichtweise auf die eingesetzten Datenerhebungsmethoden zu erfassen.

Die Leitfäden wurden so frei gehandhabt, dass die Befragten stets die Möglichkeit hatten, eigene Schwerpunkte zu setzen. Dennoch wurde in keinem der Interviews aus der Struktur des Leitfadens ausgebrochen[80], es gab lediglich in einigen Gesprächen Rückgriffe in Form von Relativierungen oder Präzisierungen von bereits Gesagtem. Ergänzend zu den im Interview erhobenen Informationen wurde zu jedem Gespräch ein Postskriptum angefertigt, in dem Besonderheiten des Gesprächs sowie ggf. Gedanken bezüglich des Inhalts festgehalten wurden.

Ein technischer Defekt verhinderte die Aufzeichnung des Interviews mit einer der planenden Personen. In diesem Fall wurde unmittelbar nach Abschluss des Interviews ein auf den während des Gesprächs angefertigten Notizen und dem Gedächtnis beruhendes Gesprächsprotokoll erstellt.

Die Dauer der Einzelinterviews bewegte sich zwischen 19 Minuten und 49 Sekunden und 52 Minuten und 06 Sekunden, sie lag im Durchschnitt bei 36 Minuten und 50 Sekunden. Alle Interviewpartner/-innen zeigten eine große Bereitschaft, sich mit den Fragestellungen des Gesprächs auseinander zu setzen und ernsthaft auf sie einzugehen. Somit konnten in allen geführten Interviews die erforderlichen Daten erhoben werden.

Datenerhebung: Gruppendiskussion

Für die Durchführung der Gruppendiskussion ergab sich ein Problem: Am Tag vor dem Gespräch musste leider die Person, die als beteiligte/r Lehrende/r eingeladen war, aufgrund einer unaufschiebbaren Verpflichtung die Teilnahme am Gespräch absagen. Dennoch wurde die Gruppendiskussion durchgeführt, wenngleich eine für das Gespräch als zentral vermutete Stakeholderrolle unbesetzt blieb.

Der Ablauf der Gruppendiskussion, die in einem Seminarraum der Universität Marburg stattfand und durch den Verfasser dieser Arbeit moderiert wurde, orientierte sich an jenem der Einzelinterviews. Der Leitfaden zur Gruppendiskussion gliederte sich in folgende thematische Blöcke:

[80] Hierfür werden zwei Gründe als ursächlich betrachtet: Erstens lag das Thema, wie bereits herausgestellt, den meisten Personen recht fern, so dass sie die thematische Führung durchaus als hilfreich empfanden, zweitens bauten die Teile des Leitfadens inhaltlich aufeinander auf.

1. Begrüßung und Einführung
2. Erläuterung des Konstruktes *Gegenstandsangemessenheit*
3. Rollenspiel zum Thema Gegenstandsangemessenheit
4. Stellungnahme zu den allgemein in Lehrevaluationen an Hochschulen eingesetzten Verfahren
5. Offene Abschlussfrage

Begrüßung und Einführung sowie die offene Abschlussfrage wurden hier in gleicher Weise wie in den Einzelinterviews durchgeführt. Anders organisiert war in der Gruppendiskussion das Rollenspiel: Die Anwesenden sollten sich in die Rolle eines Gremiums versetzen, dessen Aufgabe die Planung einer Evaluation der Fachdidaktik ist. Die Themen, zu denen dieses Gremium Stellung beziehen sollte, waren die gleichen wie in den Einzelinterviews. Es sollte diskutiert werden, welche Personengruppen zu welchen Themen und mit welchen Methoden befragt werden sollten. Um die Frage nach Argumentationsmustern zu untersuchen, stand hierbei jedoch vor allem das Gespräch der Teilnehmenden untereinander im Zentrum der Beobachtung.

Als wesentlich stellten sich während der Diskussion die unterschiedlichen Sichtweisen heraus, die sich in immer neuen Ideen und/oder spezifischen Sichtweisen auf die Eignung von Methoden und auf zu betrachtende Elemente der Fachdidaktik-Ausbildung äußerten. Diesem Sachverhalt ist die Tatsache geschuldet, dass der Moderator vom inhaltlichen Kanon des Leitfadens abwich und die Frage der Stellungnahme zu den allgemein in Lehrevaluationen an Hochschulen eingesetzten Verfahren in der im Leitfaden vorgesehenen Deutlichkeit ausließ. Diese Informationen wurden in ausreichender Deutlichkeit während des Gesprächs sichtbar.

Die teilnehmenden Personen äußerten im Anschluss an die Gruppendiskussion, dass sie es als erhellend und bereichernd empfunden hätten, die Perspektive der anderen Vertreter/-innen sehen und verstehen zu können. Somit hatte die Diskussion für sie – anders, als für die Betrachtung der Diskussion aus analytischer Perspektive – eher vermittelnden Charakter.

Die Gruppendiskussion dauerte knapp eineinhalb Stunden. Die Teilnehmer/-innen verhielten sich ebenso wie jene der Einzelinterviews sehr kooperativ und haben sich während des Gesprächs eingehend mit seiner Thematik auseinandergesetzt.

Transkription

Die aufgezeichneten Interviews und die Gruppendiskussion wurden vollständig transkribiert. Die Definition der Transkriptionsregeln folgte der Maßgabe, einen gut lesbaren Text zu erzeugen, der alle relevanten Informationen beinhaltet. Für den Zweck der Untersuchung standen dabei die konkreten inhaltlichen Informationen an erster Stelle. Nonverbale Kommunikation, Dialekte, Lautäußerungen etc. wären nur dann von Belang gewesen, wenn sie eine Bedeutung über die konkrete Aussage hinaus beinhaltet hätten. Dementsprechend wurden sie nur dann transkribiert, wenn sie diesem Kriterium entsprachen. Außerdem wurde die Sprache geglättet, d. h. an die Schriftsprache angenähert, um eine möglichst gute Lesbarkeit und Verständlichkeit zu erreichen[81].

Acht der insgesamt 24 Einzelinterviews wurden ebenso wie die Gruppendiskussion durch den Autor dieser Arbeit transkribiert, die übrigen 16 wurden durch andere Personen transkribiert. Jedes Transkript wurde im Anschluss an die Abschrift komplett Korrektur gelesen, bei den fremdtranskribierten Interviews wurden zudem Kontrollen anhand der Audioaufzeichnung durchgeführt. Die Audioaufzeichnung wurde bei allen unklaren oder fragwürdigen Stellen des Transkriptes als Grundlage der Korrektur herangezogen. Daneben erfolgten stichprobenartige Kontrollen jedes Transkriptes, indem zwecks Vergleich der Inhalte die Wiedergabe der Aufnahme zur korrespondierenden Stelle des Transkriptes erfolgte.

Den Probanden wurde zugesichert, dass die im Gespräch erhobenen Daten Dritten ausschließlich in anonymisierter Form zugänglich gemacht würden. In den eigentranskribierten Interviews wurde die Anonymisierung direkt während der Abschrift vorgenommen, in den fremdtranskribierten erfolgte sie während des Korrekturvorganges. Im Rahmen der Anonymisierung wurden sämtliche Informationen, die Rückschlüsse auf

[81] Ein Abdruck der detaillierten Transkriptionsregeln findet sich im Anhang auf S. 253.

einzelne Personen erlauben würden, so umschrieben, dass Einzelpersonen nicht mehr identifizierbar sind. Hierzu gehören beispielsweise Aussagen, die das Geschlecht der interviewten Person erkennen ließen, Aussagen über Funktionen innerhalb der Hochschule oder auch über bestimmte Studienfächer. In diesem Zuge wurden auch die Verbindungen zwischen Transkript und Audiodatei, die sog. Timestamps, entfernt.

Im Anschluss an die Korrektur der Transkripte wurden sie den befragten Personen zur Kenntnisnahme zugeschickt. Eine Korrektur der Transkripte durch die befragten Personen war nicht vorgesehen und wurde auch nicht nachgefragt.

Datenauswertung

Die hier beschriebene empirische Untersuchung ging der Frage nach, nach welchen Kriterien Stakeholder einer Evaluation die Angemessenheit eingesetzter Datenerhebungsmethoden bewerten. Wie bereits in Abschnitt 4.1.1 beschrieben sollten Argumente der Stakeholder für oder gegen bestimmte Datenerhebungsmethoden ebenso untersucht werden wie die Fragen, welche Methoden präferiert werden und ob sich Bezüge zwischen der Position einer Person gegenüber dem Evaluationsgegenstand und ihrer Vorstellung von angemessenen Methoden erkennen lassen. Auf der Grundlage dieses Erkenntnisinteresses lassen sich wesentliche Anforderungen an Datenauswertungsmethoden ableiten:

- Sie sollten ermöglichen, die zielgerichtete Struktur der Gesprächsleitfäden in ihrer Intention zu erfassen und abzubilden, gleichzeitig sollten sie für Themen und Aspekte, die nicht antizipiert wurden, offen sein (gleichzeitige methodische Fokussierung und Offenheit). Diese Anforderung ist darin begründet, dass der Interviewleitfaden zwar Antworten auf konkrete Fragen hervorbringen, aber auch Freiräume für im Sinne der Fragestellung wichtige individuelle Schwerpunktsetzungen und das Einbringen neuer Themen lassen sollte.

- Die Datenauswertungsmethoden sollten die Struktur des Gesprächsleitfadens in einer Art und Weise erfassen, die es ermöglicht, die einzelnen Interviews (= Fälle) miteinander zu vergleichen. Zu diesem Zweck ist eine analytische Struktur dienlich, die sich auf alle Interviews gleichermaßen anwenden lässt und inhalt-

liche Aussagen thematisch bündelt. Gleichzeitig sollte das Material aber nicht vollkommen zergliedert werden, sondern auch in seiner ursprünglichen Form erhalten bleiben, um jenseits der Vergleiche auch die Besonderheiten des jeweiligen Falls mit in den Blick nehmen zu können. Ziel ist hier die Ermöglichung der Betrachtung von Einzelfällen und fallübergreifenden Themen gleichermaßen.

- Die Erfassung der durch den Leitfaden erhobenen Daten sollte nicht nur im Sinne der in den betreffenden Textstellen angesprochenen thematischen Aspekte neutral sammelnd, sondern – wo immer sinnvoll möglich – gleich in gewichtender Form erfolgen können. Auf diese Weise können nicht nur inhaltliche Aussagen, sondern darüber hinaus auch gleich ihre Tendenz, etwa Zustimmung oder Ablehnung, erfasst werden, was beispielsweise für Vergleiche zwischen Fällen dienlich sein kann.

In Anbetracht dieser Erfordernisse fiel die Wahl auf eine inhaltsanalytisch orientierte Vorgehensweise (siehe v. a. Kuckartz 2012, Schreier 2012, Schmidt 2010, Mayring 2010). Kern des analytischen Instrumentariums war, wie in inhaltsanalytischen Auswertungsansätzen üblich, ein Kategoriensystem (siehe Kuckartz 2012, S. 40 ff., Mayring/Brunner 2010, S, 325). Die für die Analyse relevanten Kategorien wurden teils aus theoretischen Überlegungen abgeleitet (Deduktion, Schluss von Allgemeinen auf das Spezielle), teils unmittelbar aus dem Material der Interviews heraus erzeugt (Induktion, Schluss vom Speziellen auf das Allgemeine). Die deduktiv gebildeten Kategorien wurden auf der Grundlage des Interviewleitfadens gebildet, der seinerseits den Rahmen relevanter Themen für die Datenerhebung vorgab. Diese deduktive Kategorienbildung fand zu zwei Zeitpunkten statt: Eine erste Kategorienbildung wurde direkt nach der Erstellung des Leitfadens vorgenommen, um die relevanten Themen möglichst unmittelbar und gemäß ihrer ursprünglichen Intention festzuhalten. Ein zweiter Durchgang, während dessen die bereits gebildeten Kategorien überprüft und wenn erforderlich ergänzt wurden, schloss sich unmittelbar an die Phase der Datenerhebung an. Durch diesen zeitlichen Abstand konnten gleichzeitig die Gültigkeit der Kategorien und ein gleichbleibendes Verständnis sichergestellt werden (Intra-Coder-Realiabilität, Mayring/

Brunner 2010, S. 326). Die zweite Phase deduktiver Kategorienbildung brachte keine Änderungen des in der ersten Phase produzierten Kategoriensystems hervor.

Ergänzend zur deduktiven Kategorienbildung wurden während des Analyseprozesses der Daten für Themen, die für die Beantwortung der Forschungsfrage relevant sind, jedoch von noch keiner Kategorie erfasst wurden, induktiv neue Kategorien erzeugt. Auch diese Phase gliederte sich in mehrere Unterphasen: Zunächst wurden acht Interviews ausgewählt, per Zufallszahlengenerator wurden je zwei Interviews aus den vier zentralen Stakeholdergruppen (Beteiligte Studierende, Betroffene Studierende, Beteiligte Lehrende, Betroffene Lehrende) ausgelost. Diese Interviews wurden unter Anwendung der deduktiv gebildeten Kategorien bearbeitet. Für Textstellen, die im Sinne des Erkenntnisinteresses zu unspezifisch oder gar nicht durch das bestehende Kategoriensystem erfasst werden konnten, wurden neue Kategorien gebildet. Zu diesem Zeitpunkt erfolgte jenseits der Erzeugung neuer Kategorien noch keine dauerhafte Zuordnung von Textstellen zu Kategorien (Codierung), sondern die Interviews wurden exemplarisch bearbeitet. An diese Phase anschließend wurde das nun aus deduktiv und induktiv gebildeten Kategorien bestehende Kategoriensystem erstmals zur Codierung verwendet. Hierzu wurden aus den verbleibenden Interviews wiederum per Zufall zwölf bisher nicht bearbeitete ausgewählt, diesmal drei aus jeder der vier zentralen Stakeholdergruppen. Nachdem diese Interviews codiert waren, wurde erneut überprüft, ob alle relevanten Textstellen zufriedenstellend erfasst wurden oder das Kategoriensystem weiterer Änderungen bedurfte, was nicht erforderlich war. Außerdem wurden die Kategorien entsprechend ihrer inhaltlichen Bedeutungen in Haupt- und Subkategorien geordnet. Anschließend wurden die verbleibenden Interviews codiert. Nach Abschluss der Codierphase wurde das Kategoriensystem einer finalen Reorganisation unterzogen, während derer Haupt- und Subkategorien entsprechend ihrer Bedeutungen endgültig angeordnet wurden. Das finale Kategoriensystem enthielt folgende Hauptkategorien [82]:

1. *Vorverständnis Evaluation/Grundlagen/Erfahrungen:* Die Informationen, die dieser Kategorie zugeordnet wurden, dienten dazu, fokussiert auf das Thema der

[82] Eine Übersicht über das gesamte finale Codesystem findet sich im Anhang ab S. 254.

Erhebung etwas über den Hintergrund der Interviewpartnerin/des Interviewpartners zu erfahren. Hier wurde etwa erfasst, ob die Person bereits über Evaluationserfahrungen über die Fachdidaktik-Evaluation hinaus verfügte, wie gut sie sich über die Evaluation informiert fühlte oder welche Ziele sie wahrgenommen hat. Die hier enthaltenen Daten liefern somit wichtige Informationen, vor denen die Aussagen der Probanden besser eingeschätzt werden können, da der persönliche fragestellungsrelevante Hintergrund der Personen sichtbar gemacht wird.

2. *Vorgeschlagene Probandengruppen:* Diese Kategorie bezieht sich auf das Gedankenexperiment, das mit den Probanden/-innen im Laufe des Interviews in Form der fiktiven Einnahme der Rolle einer eine Fachdidaktik-Evaluation planenden Person durchgeführt wurde. Hier wurde ermittelt, welche Personengruppen aus Sicht der befragten Person sinnvollerweise in eine Evaluation der Fachdidaktik-Ausbildung an der Universität Marburg einbezogen werden sollten.

3. *Beiträge der Probandengruppen:* In dieser Kategorie wurde erfasst, welche Beiträge die vorgeschlagenen (und in der vorigen Kategorie gesammelten) Gruppen zur Bewertung der Güte der Fachdidaktik-Ausbildung der Meinung der Probanden nach liefern könnten. Die hier enthaltenen Aussagen präzisieren folglich die Ideen, welche die befragte Person mit dem Vorschlag der Berücksichtigung einer Stakeholdergruppe verband.

4. *Datenerhebungsmethoden:* Bei dieser Kategorie handelt es sich um jene mit den meisten Unterkategorien. Sie enthält die relevanten Aussagen der befragten Personen zu Datenerhebungsmethoden. Die Aussagen wurden thematisch gruppiert in folgenden Subkategorien erfasst:

   Vorgeschlagene Methoden nach Stakeholdergruppen: Mit welchen Methoden sollten die Informationen, welche die Stakholdergruppen nach Ansicht der Probandin/des Probanden beisteuern können, erfasst werden?

   Bewertung der selbst vorgeschlagenen Methoden: Hier wurden Aussagen zugeordnet, die Auskunft über Einschätzungen (positive wie negative) der vor-

geschlagenen Methode enthalten.

Bewertung der tatsächlich eingesetzten Methoden: Bestandteil des Leitfadens war die Frage nach der Bewertung der Methoden, die im Rahmen der Fachdidaktik-Evaluation tatsächlich eingesetzt wurden. Die Einschätzungen der Probanden wurden in dieser Kategorie erfasst.

Bewertung der üblichen Methoden: Es wurden Aussagen gebündelt, in denen die befragten Personen Bewertungen der allgemein in Lehrevaluationen an Hochschulen übliche Methoden vornehmen.

Diskussion über Methoden in der Gruppendiskussion: Dieser Kategorie wurden alle Passagen der Gruppendiskussion zugeordnet, in denen sich die teilnehmenden Personen über die Eignung von Methoden unterhalten.

5. *Persönliche Begründungen zur Angemessenheit:* Aussagen, aus denen erkennbar wird, nach welchen Kriterien die Interviewpartner/-innen die Güte von Datenerhebungsmethoden im Kontext von Lehrevaluationen an Hochschulen persönlich bewerten, wurden dieser Kategorie zugeordnet. Sie bezieht sich damit auf den letzten Abschnitt des Interviewleitfadens, in dem die Probandinnen und Probanden gebeten wurden zu explizieren, wann sie mit den Datenerhebungsmethoden einer Evaluation zufrieden wären.

6. *Teilnehmer/-innen Gruppendiskussion:* Diese letzte Hauptkategorie diente dazu, die Sprecher/-innen der Gruppendiskussion zu erfassen und für die weitere Analyse nutzbar zu machen. Sie hat folglich keine inhaltliche, sondern eine rein formale Funktion.

Im Sinne der für die Inhaltsanalyse typischen Reduktion der Datenmenge (Schreier 2012, S. 3 ff.) war die wesentliche Maßgabe für die Bildung der Kategorien die Anforderung, dass jede einen direkten Beitrag zur Beantwortung der Forschungsfrage leisten kann. Auf diese Weise wurden alle inhaltlich relevanten Textstellen in Kategorien gesammelt.

Im endgültigen Kategoriensystem können zwei Arten von Kategorien unterschieden werden, nämlich *sammelnde* und *bewertende*. Die sammelnden Kategorien dienten im Sinne der von Kuckartz beschriebenen *inhaltlich strukturierenden qualitativen Inhaltsanalyse* (Kuckartz 2012, S. 77 ff.) dazu, Textstellen, deren Inhalt etwas zum durch die jeweilige Kategorien erfassten Themenbereich beiträgt, zu sammeln und so die verschiedenen Facetten der entsprechenden Thematik zu bündeln. Die bewertenden Kategorien erlaubten im Sinne von Kuckartz' *evaluativer qualitativer Inhaltsanalyse* (Kuckartz 2012, S. 98 ff.), die Textstellen gleichzeitig zu gewichten bzw. eine Deutung oder Bewertung der Inhalte in Kategorienform abzubilden.

Ein Beispiel für sammelnde Kategorien ist die Kategorie *Wahrgenommene Ziele*, eine Subkategorie von *Vorverständnis Evaluation/Grundlagen/Erfahrungen*. Ihr wurden alle Aussagen zugeordnet, in der sich die befragten Personen über ihre Wahrnehmung der Ziele der Fachdidaktik-Evaluation äußern. Die Kategorie *Offene Interviews* (eine Subkategorie der Kategorie *Bewertung der tatsächlich eingesetzten Methoden*) hingegen ist ein Beispiel für bewertende Kategorien: Hier wurden Textstellen, die inhaltlich in diese Kategorie gehören, ihr nicht einfach zugeordnet, sondern es wurde direkt festgehalten, ob die befragte Person dieses Datenerhebungsverfahren einsetzen oder nicht einsetzen würde. Ebenso wurde erfasst, wenn die befragte Person keine klare Aussage bezüglich des Einsatzes der Methode getroffen hat. Um diese Tendenz der Aussagen der befragten Person festzuhalten[83] wurden der Kategorie *Offene Interviews* drei Subkategorien zugeordnet: *Einsetzen*, *Nicht einsetzen* sowie *Keine klare Bewertung*. Ausschlaggebend für die Zuordnung zu einer dieser bewertenden Subkategorien war nun die Gesamtposition der befragten Person zur jeweiligen Methode: Oftmals nannten die Probanden/-innen sowohl Vor- als auch Nachteile einer Methode, und zwar auch dann, wenn sie sich insgesamt für oder gegen ihren Einsatz entschieden hatten. In einem Fall, in dem sich eine befragte Person gegen den Einsatz einer bestimmten Methode im Rahmen der fiktiven Fachdidaktik-Evaluation entschieden hatte, aber dennoch Vorteile dieser Me-

[83] Letztlich handelt es sich bei dieser Zuordnung um eine Interpretation der Aussage durch die codierende Person. Zwar war der Leitfaden so angelegt, dass die befragten Personen um eine Aussage bzgl. des Einsatzes einer Datenerhebungsmethode gebeten wurden, diese Aussage erfolgte jedoch nicht immer in direkter Form.

thode benannte, wurden all ihre Aussagen über diese Methode – positive wie negative – demnach gemäß der Codierregeln der Kategorie *Nicht einsetzen* zugeordnet. Auf diese Weise wird anhand der Codierungen die Kernaussage der jeweiligen Person sichtbar. Daneben ist es möglich, die Summe der Argumente einer Person für oder gegen eine Methode immer unmittelbar im Kontext ihrer Gesamtbewertung zu betrachten.

Um die Auswertung möglichst effizient und transparent zu gestalten, wurde für die Analyse die Software MAXQDA eingesetzt. Der Einsatz von solch spezieller Software zur Analyse qualitativer Daten ermöglicht einen überaus flexiblen Umgang mit den Daten. So wird es beispielsweise mit wenig Aufwand möglich, Personengruppen zu vergleichen oder zielgerichtet Annahmen zu überprüfen. Auch kann der oben formulierten Anforderung nach Etablierung einer neuen analytischen Struktur (in Form analytischer Kategorien, denen Textstellen zugeordnet werden) und gleichzeitiger Wahrung der ursprünglichen Form (Originalsequenz der Interviews) der Daten entsprochen werden, da Textstellen sowohl im Kontext ihrer Kategorien als auch im Originalkontext des Interviews betrachtet werden können. Schließlich ermöglicht die Software die unmittelbare Integration unterschiedlicher Methodenfamilien, etwa in Form der Bereitstellung einer Fülle quantitativer Informationen über die qualitativ-interpretative Tätigkeit des Forschers. Neben diesen forschungspraktischen Komponenten stellt sie auch einen Beitrag zur Sicherung der Qualität der Analyse dar, da sie eine engmaschige Dokumentation ermöglicht und die Daten in einem für andere Personen zugänglichen Format abspeichert, was im Sinne der Nachvollziehbarkeit von großer Bedeutung ist.

4.2. Ergebnisse der Untersuchung

Nachfolgend werden die Ergebnisse der empirischen Untersuchung dargestellt und eingeordnet. Zunächst werden die Antworten der Befragten[84] – der Logik und dem The-

[84] Um Unklarheiten bezüglich der vom Autor angesprochenen Personengruppe zu vermeiden, werden im Folgenden die Bezeichnungen *Probanden*, *befragte Personen* und *Befragte* für jene Personen verwendet, welche in die hier vorgestellte Untersuchung einbezogen wurden. Die Beteiligten der Fachdidaktik-Evaluation, auf die sich diese Untersuchung bezieht, werden entsprechend anders bezeichnet.

menverlauf des Interviewleitfadens folgend – betrachtet, d. h. der Fokus wird zuerst auf die von den Befragten vorgeschlagenen Stakeholder (-gruppen), danach auf die von ihnen genannten Datenerhebungsmethoden und schließlich auf die Kriterien zur Bewertung der Angemessenheit gerichtet. Das Hauptaugenmerk liegt dabei auf den Argumenten und Kriterien zur Begründung der Sichtweise der Probanden/-innen auf die Angemessenheit von Datenerhebungsmethoden. Abschließend wird dargestellt, welche Implikationen sich aus diesen Aussagen für die Frage nach den Bewertungskriterien für die Angemessenheit von Datenerhebungsmethoden in Evaluationen ergeben.

Die Probanden entstammen, wie in Abschnitt 4.1.4 dargestellt, unterschiedlichen Personengruppen: Bei zwölf Befragten handelt es sich um Stakeholder, die im Rahmen der an der Universität Marburg durchgeführten Fachdidaktik-Evaluation befragt und somit in die Evaluation einbezogen worden waren, also um Beteiligte. Sechs dieser Personen waren Studierende, die übrigen sechs Lehrende. Zwölf weitere befragte Personen aus dem gleichen Personenkreis waren nicht in die Fachdidaktik-Evaluation einbezogen worden, somit handelt es sich um Betroffene. Auch hier wurden sechs Studierende und sechs Lehrende befragt. Diese vier Personengruppen werden hier, unter Bezugnahme auf ihre Rolle in der real durchgeführten Evaluation der Fachdidaktikausbildung, unter der Bezeichnung *(potenzielle) Evaluationsteilnehmer/-innen* zusammengefasst. Neben dieser ersten und größten Gruppe wurden die beiden Personen befragt, welche die Evaluation geplant hatten. Im Rahmen der Gruppendiskussion trafen die unterschiedlichen Stakeholdergruppen mit Ausnahme der Planenden aufeinander[85].

Die drei Erhebungseinheiten *(potenzielle) Evaluationsteilnehmer/-innen*, *Planende* und *Gruppendiskussion* werden nachfolgend zunächst nicht gleichzeitig betrachtet, da jede dieser Gruppen mit einer anderen Intention in die Untersuchung einbezogen wurde und sich deshalb sowohl die Interviewleitfäden als auch die Verläufe der Interviews zwischen diesen drei Gruppen unterschieden, die Analyse der Daten wird somit zuerst nach eben diesen drei Gruppen getrennt durchgeführt. Eine Integration und Weiterführung der Ergebnisse erfolgt im Anschluss an die Analyse der einzelnen Gruppen.

[85] Im Anhang auf Seite 258 ist eine Übersicht über die Verortung der Interviewpartner/-innen innerhalb der verschiedenen Erhebungseinheiten abgedruckt.

4.2.1. Die Sicht der (potenziellen) Evaluationsteilnehmer/-innen

Die Gruppe der (potenziellen) Evaluationsteilnehmer/-innen ist die mit der größten Nähe zum Evaluationsgegenstand, da es sich bei den Studierenden und Lehrenden um die unmittelbaren Akteure universitärer Lehre handelt. Die ihr zugehörigen Personen waren im Rahmen der Fachdidaktikevaluation befragt worden (dies ist die Gruppe der Evaluationsteilnehmer/-innen) oder gehörten doch wenigstens zu einer der prinzipiell befragten Gruppen, wenngleich sie – aus welchen Gründen auch immer[86] – persönlich keine Daten beigesteuert hatten (Gruppe der potenziellen Evaluationsteilnehmer/-innen). Nachfolgend werden die Aussagen der dieser Gruppe zuzuordnenden Probanden zusammenfassend dargestellt.

Vorgeschlagene Stakeholdergruppen

Betrachtet man die Vorschläge der Probanden für die in die fiktive Evaluation einzubeziehenden Stakeholdergruppen, so lässt sich eine klare Tendenz feststellen: Alle Befragten sind sich einig, dass die Studierenden für die Datenerhebung zu berücksichtigen wären (24 Nennungen)[87]. Ähnlich deutlich fällt das Votum für die Einbeziehung der Lehrenden aus, hier haben sich 23 der 24 befragten (potenziellen) Stakeholder dafür ausgesprochen, diese als Datengeber/-innen einzuplanen. Auf Rang drei dieser Vorschlagsliste rangieren die Referendare/-innen bzw. Absolventen/-innen, die nach der Meinung von 16 Probanden ebenfalls einen wertvollen Beitrag zur Bewertung der Fachdidaktik-Evaluation leisten könnten. Diese drei Gruppen wurden nicht nur am häufigsten, sondern auch sehr gleichverteilt vorgeschlagen, d. h., dass sowohl die befragten Studierenden als auch die befragten Lehrenden diese Gruppen gleichermaßen in die Evaluation einbeziehen würden. Auch die Gruppe der Referendare/-innen bzw. Absolventen/-innen wurde von jeder Probandengruppe gleich häufig, nämlich je achtmal, vorgeschlagen.

[86] Bei den Studierenden liegen die Gründe für die Nichtteilnahme in der Regel vermutlich im persönlichen Bereich, da alle Studierenden per E-Mail zur Teilnahme an der Evaluation eingeladen wurden und ihnen die Teilnahme somit prinzipiell möglich war. Auf Seiten der Lehrenden hingegen wurde eine Stichprobe bestimmter Personen gezogen, die Übrigen wurden nicht in die Datenerhebung einbezogen und hatten somit keine Möglichkeit zur Teilnahme.

[87] Tabellarische Übersichten über die Häufigkeiten der Nennungen finden sich im Anhang auf S. 258 ff.

Deutlich weniger häufig genannt wurden Stakeholdergruppen, die außerhalb des unmittelbaren Lehr-/Lernkontextes der Hochschule verortet sind. So wurde die Gruppe der Seminarleiter/-innen bzw. der Ausbilder/-innen im Referendariat, die am vierthäufigsten vorgeschlagen wurde, lediglich fünfmal als relevant benannt. Eine ähnlich häufig genannte Gruppe stellen die Lehrerkollegien an den Schulen dar, in denen die Referendare weiter ausgebildet oder die jungen Lehrer arbeiten werden. Diese Gruppe wurde viermal vorgeschlagen. Es ist auffällig, dass diese beiden Stakeholdergruppen lediglich von Lehrenden, nicht aber von Studierenden als relevant genannt wurden. Somit wird ihnen zumindest innerhalb einer der befragten Probandengruppen Relevanz beigemessen.

Die am seltensten benannten Gruppen sind die Schüler der Absolventen (drei Nennungen), der administrative Bereich mit Verwaltung, Schulamt und Politik (zwei Nennungen) sowie sonstige Außenstehende (Nennungen). Hier gibt es keine Häufung der Nennungen innerhalb einer der Gruppen der befragten Personen.

Setzt man diese Häufigkeiten der Nennungen in einen erweiterten Kontext, so fällt auf, dass den unmittelbar am Lehr-/Lernprozess beteiligten Personen – den Lernenden und den Lehrenden, also den *Eingebundenen* (siehe S. 107) – unzweifelhaft und übereinstimmend die größte Relevanz für die Bewertung der Fachdidaktik-Ausbildung zugesprochen wird. Diese Tatsache ist wenig verwunderlich und lässt sich auf verschiedene Weise erklären. Einerseits scheint es naheliegend zu sein, dass die Personen vorgeschlagen wurden, die im Alltag der Befragten als Akteure innerhalb des Evaluationsgegenstandes am stärksten in Erscheinung treten und damit auch die stärkste Präsenz in der Wahrnehmung besitzen, und zwar auch gerade deshalb, weil die Interviewsituation auf die spontane und ggf. naive Vorstellung von geeigneten Stakeholdergruppen (und damit die unmittelbar präsenten Elemente) abzielte. Schließlich üben diese beiden Gruppen auch den sichtbarsten Einfluss auf den Evaluationsgegenstand aus[88]. Neben

[88] Interessant ist in diesem Kontext die Argumentation der Person, die die Lehrenden nicht als datengebende Gruppe in die Evaluation einbeziehen würde. Sie begründet die bewusste Nichtberücksichtigung der Lehrenden damit, dass diese ohnehin entsprechend ihrer Intentionen handelten und somit das Urteil der Studierenden einen wesentlicheren Beitrag zur Bewertung leiste als die

dieser eher organisationspragmatischen Erklärung ist aber auch eine inhaltliche möglich, denn schließlich verfügt keine andere Stakeholdergruppe über ähnlich fundierte Kenntnisse bezüglich der tatsächlichen Inhalte und Praxis des Evaluationsgegenstandes. Es gibt also gute Gründe für die Einbeziehung dieser beiden Gruppen. Nimmt man nun noch in den Blick, welche inhaltlichen Beiträge sie aus Sicht der Probanden leisten könnten, so wird ersichtlich, dass die jeweils subjektive Sichtweise auf das Studium als besonders relevant eingestuft wird. So sollten nach Meinung der Probanden/-innen die Studierenden vor allem nach ihrer Sicht auf die Studienpraxis befragt werden. Dabei sollte zwischen einem subjektiven Zufriedenheitsurteil auf der einen und der Bewertung der Studienorganisation und der Rahmenbedingungen auf der anderen Seite unterschieden werden. Im Bereich der Zufriedenheitsurteile stünden individuelle Bewertungen der Studieninhalte im Zentrum. Allgemeiner gesprochen geht es hier um die Zufriedenheit der Studierenden mit den Inhalten, der Qualität des Studiums, aber auch um eine abschließende Beurteilung durch Studierende in höheren Semestern bzw. mit Praxiserfahrung, ob das, was an der Universität gelehrt wurde, auch tatsächlich Praxisrelevanz besitzt (vgl. bspw. B2, 26; B7, 42; B8, 26; B9, 50; B18, 18 – 20; B20, 24 – 25; B22, 18). Die Lehrenden würden ebenfalls in erster Linie nach ihrer Sicht auf das Studium und die Lehrveranstaltungen befragt und auch hier stünde die individuelle, durch die Position gegenüber dem Evaluationsgegenstand bedingte Sichtweise im Zentrum des Interesses. Hierbei wurden auch Fragen nach den Lehrveranstaltungen, ihrer Umsetzung, Organisation und Praxis als relevant vorgeschlagen (vgl. etwa B5, 44; B6, 54; B8, 50, B17, 40). Vor diesem Hintergrund ist die Nennung der Studierenden und der Lehrenden als besonders geeignete Stakeholder aus doppelter Sicht naheliegend und nachvollziehbar.

Jenseits dieser Begründungen ist interessant, dass, wie anhand der wichtigsten vorgeschlagenen Stakeholdergruppen erkennbar ist, eine auf der Innensicht der direkten Akteure basierende Evaluation breite Zustimmung finden würde. Eine Erweiterung der zu befragenden Stakeholder hin zu Personen, die nicht (mehr) aktiv am zu evaluierenden Geschehen beteiligt sind, nämlich den Referendaren/-innen bzw. Absolventen/-innen,

Selbstbewertung der Lehrenden, die demnach verzichtbar sei (B11, 46).

wurde zwar noch von zwei Dritteln der Probanden vorgeschlagen. Auch hier handelt es sich noch immer um Personen mit einer großen Nähe zum Evaluationsgegenstand. So wird ihr möglicher positiver Beitrag zur Bewertung vor allem darin gesehen, dass sie den Evaluationsgegenstand aufgrund des abgeschlossenen Studiums und des somit gänzlich durchlaufenen fachdidaktischen Studienanteils komplett im Blick haben, also über Überblickswissen verfügen. Gleichzeitig wird als Vorteil gesehen, dass sie sich an der Schnittstelle zur Praxis befinden und somit die im Studium bearbeiteten Inhalte aktiv zur Anwendung bringen müssen, was wiederum als gute Grundlage für die Bewertung eben dieser Inhalte angesehen wird (vgl. bspw. B5, 32; B12, 47 – 48; B20, 16).

Zwar stellt die vorgeschlagene Berücksichtigung von ehemaligen Studierenden bereits eine Öffnung hin zu nicht mehr aktiv am Studiengeschehen beteiligten Personen dar, jedoch handelt es sich auch hier um Personen, die dem direkten Ausbildungskontext zuzuordnen oder wenigstens sehr intensiv mit ihm vertraut sind. Die Vorschläge, auch die Seminarleiter/-innen bzw. die Ausbilder/-innen im Referendariat sowie die Kollegien der Schulen als Datenquellen mit in die Evaluation einzubeziehen, stellen hingegen eine neue Qualität dar, da hier die Sicht von Personen als Bewertungsgrundlage herangezogen wird, die aktiv nichts mit der fachdidaktischen Ausbildung der angehenden Lehrer/-innen an der Universtiät zu tun haben, sondern etwas von ihr distanziert sind, jedoch die weiteren Anforderungen aus professioneller Sicht kennen. Somit stellt der Vorschlag, eine oder mehrere dieser Gruppen zu befragen, ein Plädoyer für die Berücksichtigung von Außenperspektiven dar. Auffällig ist an dieser Stelle, wie bereits erwähnt, die Tatsache, dass diese Stakeholdergruppen lediglich von befragten Lehrenden eingebracht wurden. Hier ist anzunehmen, dass Vertreter dieser externen Gruppen eine andere Sicht auf das Studium und die nachuniversitäre Ausbildung der Studierenden haben als die Lehrenden selbst: Aus ihrer Sicht handelt es sich bei den Angehörigen der beiden hier besprochenen Stakeholdergruppen um Personen, die auf die Arbeit der Lehrenden an den Hochschulen aufbauen und auf sie angewiesen sind. So würde auf gewisse Weise auch die Frage nach der praktischen Verwendbarkeit und des

konkreten Nutzens der erlernten Inhalte zum Gegenstand der Evaluation, und zwar jenseits der Selbsteinschätzung der Lehrenden und der Studierenden. Ob Studierende diese Sichtweise ebenfalls teilen würden, mag unter anderem vom jeweiligen Stadium ihrer Ausbildung abhängen. Die hier befragten Studierenden waren alle noch an der Hochschule, so dass ihnen konkrete Erfahrungen in diesem weiterführenden Ausbildungsbereich fehlten. Das Interesse, das die Lehrenden mit den Daten dieser beiden Stakeholdergruppen verbinden, wird deutlich, wenn die konkreten Vorschläge von Themen, nach denen diese befragt werden sollten, untersucht werden. Im Zentrum stehen hier Fragen nach einer möglichst guten Abstimmung von Inhalten zwischen Hochschulen und Studienseminaren (etwa B12, 26 oder B23, 52) und nach der Nützlichkeit des im Rahmen der universitären Ausbildung erworbenen Wissens (z. B. B21, 49 – 50; B22, 18).

Die übrigen drei möglichen Stakeholdergruppen, die zur Bewertung der Fachdidaktik-Ausbildung vorgeschlagen wurden, sind die Schüler/-innen, die durch die im Rahmen des Evaluationsgegenstandes ausgebildeten Lehrer unterrichtet werden (3 Nennungen), der administrative Sektor, hier verkörpert durch Verwaltung, Schulamt und Politik (2 Nennungen) sowie ganz allgemein außenstehende Dritte (2 Nennungen). Dabei sind die möglichen Beiträge dieser Gruppen in den Interviews jedoch nur sehr vage umrissen worden (beispielsweise B10, 20; B14, 18). Es handelte sich eher um vorsichtige Überlegungen, ob die betreffenden Gruppen gewinnbringende Daten liefern könnten oder nicht, sowie um eine auf die entsprechende Frage des Interviewers hin geäußerte Suche nach möglichen anderen Stakeholdergruppen.

Zwar berühren die hier dargestellten Vorschläge der Befragten für zu berücksichtigende Stakeholdergruppen noch nicht direkt das Gebiet der Datenerhebungsmethoden. Allerdings stellen die Personen, die im Rahmen einer Datenerhebung berücksichtigt werden sollen, mit ihren spezifischen Eigenschaften – wie in Kapitel 3.2 dargestellt – einen wesentlichen Bezugspunkt für die Wahl von Datenerhebungsmethoden dar. So lassen sich für die Frage nach der Gegenstandsangemessenheit von Datenerhebungsme-

thoden zwei wesentliche Aspekte aus den Aussagen über die vorgeschlagenen Stakeholdergruppen ableiten:

1. Der Erfassung der Sichtweise (bzw. hier konkreter der persönlichen Zufriedenheit mit der fachdidaktischen Ausbildung) der Personen in unmittelbarer Nähe zum Evaluationsgegenstand wird ein hoher Stellenwert eingeräumt. Ausschlaggebend dürfte hier die angenommene Expertise dieser Personen für den Evaluationsgegenstand sein. Dabei sprechen sich die hier befragten Personen prinzipiell für die Berücksichtigung mehrerer unterschiedlicher Perspektiven, hier mindestens jener der Studierenden und der Lehrenden, aus. Somit müssten Datenerhebungsmethoden zum Einsatz kommen, die es gestatten, die individuellen Sichtweisen unterschiedlicher Stakeholdergruppen zu erfassen.

2. Die Sicht auf die Relevanz bestimmter Stakeholdergruppen wird durch die Tiefe der Kenntnisse des Evaluationsgegenstandes beeinflusst. Je komplexer die Sicht auf den Gegenstand, je tiefer und weitreichender die Kenntnisse, desto mehr Stakeholdergruppen werden als relevant identifiziert werden können. Relevanz kann dabei in Form möglicher Einflussnahme auf den sowie in Form direkter Beteiligung am Evaluationsgegenstand bemessen werden. Dies bedeutet, dass ein breites (aber gleichzeitig nicht ausuferndes) Spektrum an Stakeholdern als Datengeber in die Evaluation einbezogen werden sollte, um größtmögliche Akzeptanz herstellen zu können, denn nicht jede Stakeholdergruppe gesteht jeder anderen die gleiche Wichtigkeit zu[89]. Für die einzusetzenden Datenerhebungsmethoden heißt dies, dass sie so flexibel sein müssen, den Besonderheiten der unterschiedlichen Gruppen entsprechen zu können. Hier kann z. B. die möglicherweise von einer Stakeholdergruppe zu erhebende Tiefe der Daten ebenso einen Einflussfaktor darstellen wie die Anzahl der zu befragenden Personen.

[89] An dieser Stelle soll kurz der Auswertung der Gruppendiskussion (S. 185) vorgegriffen werden, denn dort hat sich gezeigt, dass Personen, die eine bestimmte Stakeholdergruppe nicht selbst vorgeschlagen haben, prinzipiell empfänglich für begründete Argumente sind und damit durchaus offen für eine Berücksichtigung der fraglichen Gruppen sind.

Für die Auswahl von Datenerhebungsmethoden ergibt sich aus der Betrachtung der vorgeschlagenen Stakeholdergruppen vor allem die Implikation, dass die zum Einsatz kommenden Methoden unterschiedlichsten Personengruppen gerecht werden können sollten. Das Spektrum der hier potenziell relevanten Faktoren ist dabei breit gefächert und erstreckt sich von Aspekten wie der Personenstärke einer Gruppe über die Art möglicherweise zu erfassender Informationen bis hin zu ethischen Fragen, die sich beispielsweise bei der Befragung von Kindern und Jugendlichen ergeben könnten.

Datenerhebungsmethoden

Von den Probanden vorgeschlagene Datenerhebungsmethoden

Um einen Einblick in die tatsächliche Vorstellung der befragten Personen bezüglich der Angemessenheit von Datenerhebungsmethoden zu erhalten, wurde in den Interviews auf drei unterschiedlichen Ebenen nach der Eignung von Datenerhebungsmethoden gefragt. Zunächst wurden die Befragten aufgefordert, im Rahmen des Rollenspiels (siehe S. 129 ff.) völlig frei Methoden vorzuschlagen, die ihrer Ansicht nach dazu geeignet wären, die für die Bewertung des Evaluationsgegenstandes erforderlichen Informationen angemessen zu erfassen. Im zweiten Schritt sollten sie die Methoden, die im Kontext der real durchgeführten Fachdidaktik-Evaluation eingesetzt worden waren, hinsichtlich ihrer Angemessenheit bewerten. Abschließend wurden sie gebeten, die häufig für Evaluationen universitärer Lehre eingesetzten Datenerhebungsmethoden (siehe S. 77 ff.) dahingehend zu beurteilen, ob sie sie für die Evaluation der Fachdidaktik-Ausbildung für angemessen halten oder nicht. Auf diese Weise wurde es möglich, sowohl die naiven (in Abgrenzung zu professionell methodologisch fundierten) Vorstellungen von Methoden und ihrer Eignung als auch die Sichtweisen auf Verfahren, die den Probanden vorerst nicht bekannt oder gegenwärtig waren, zu betrachten.

Bezüglich der Datenerhebungsmethoden, die von den Probanden/-innen vorgeschlagen werden, lässt sich als übergeordneter Trend erkennen, dass der Einsatz sowohl offener als auch standardisierter Datenerhebungsmethoden (Methodenmix/Methodenintegration[90]) bevorzugt wird. Die Vorschläge wurden dabei nicht frei, sondern stets an

[90] In der Literatur zur Verbindung quantitativer und qualitativer Forschungsmethoden werden mit

die vorher im Gespräch vorgeschlagenen Stakeholdergruppen gebunden erhoben. Somit lässt sich diese Tendenz auch mit Fokus auf die meistgenannten Gruppen (Studierende, Lehrende und Referendare/innen bzw. Absolventen/-innen) bestätigen: Für die Erfassung der Daten der Studierenden wurde zehnmal ein Methodenmix vorgeschlagen (daneben zehn mal standardisierte und drei mal offene Verfahren), für die Lehrenden 13-mal ein Methodenmix (versus fünfmal offene und viermal standardisierte Methoden) und für die Gruppe der Referendare/innen bzw. Absolventen/-innen siebenmal ein Methodenmix (im Vergleich zu je vier Vorschlägen für standardisierte und offene Verfahren.) Somit entfallen 30 der insgesamt 60 Vorschläge auf Verfahren, die offene und standardisierte Verfahren kombinieren, standardisierte Methoden wurden 18-mal und offene 12-mal vorgeschlagen.

Auf der Ebene der Frage, aus welcher Probandengruppe die Vorschläge stammen und welche Methoden eine Person für die verschiedenen Stakeholdergruppen vorschlägt, wird das Bild zwar heterogener, die Affinität zum Mix von Methoden wird jedoch noch etwas deutlicher: Alle relevanten Stakeholdergruppen nur mit standardisierten Datenerhebungsmethoden zu befragen, wurde insgesamt von fünf Personen in Betracht gezogen, ausschließlich offene Datenerhebungsmethoden einzusetzen, wurde von zwei Befragten vorgeschlagen. Die verbleibenden 17 Probanden/-innen (über 70%) schlugen entweder für die verschiedenen genannten Stakeholdergruppen unterschiedliche Methoden vor oder bevorzugten auch innerhalb einer Gruppe die Integration unterschiedlicher Methoden.

Bei der Analyse der Begründungen für den Einsatz bestimmter Methoden lassen sich stets ähnliche Argumente entdecken. Als Hauptgründe für den Einsatz standardisierter Methoden werden die Möglichkeit, auch die Daten großer Stichproben mit vertretbarem

bestimmten Begriffen (im Wesentlichen *Methodentriangulation*, *Methodenintegration*, *Mixed Methods*) spezielle, ausgearbeitete Konzepte und Modelle der Kombination der jeweiligen Methoden verbunden. In dieser Arbeit wird aufgrund der Art der erhobenen Informationen nicht weiter zwischen diesen Konzepten und ihren Spezifika unterschieden. Stattdessen werden die verschiedenen Begriffe synonym verwendet und bezeichnen jegliche Art des gemeinsamen (aufeinander folgenden, sich ergänzenden, gleichzeitigen...) Einsatzes von Datenerhebungsmethoden, die den beiden unterschiedlichen Gruppen zuzurechnen sind.

Aufwand erfassen zu können, ein auch bei großen Datenmengen überschaubarer Auswertungsaufwand (Zeiteffizienz, und zwar für alle beteiligten Parteien), die Möglichkeit, trotz der Masse an Daten bzw. Probanden zielgruppenspezifische Erhebungsinstrumente einzusetzen, eine niedrige Beteiligungshürde für die Befragten, Zeitflexibilität für die Befragten hinsichtlich der Beteiligung an der Untersuchung sowie die Möglichkeit der Verwirklichung anonymer Datenerhebungen als zentrale Begründungen angeführt. Eine weitere, aber weniger dominante Begründung ist die Vergleichbarkeit der erhobenen Daten, die durch den Einsatz standardisierter Instrumente erreicht werden kann. Demgegenüber werden als wesentliche Einwände die Schwierigkeit der Antizipation tatsächlich relevanter Inhalte (Ungenauigkeit), die sich daraus unter Umständen ergebende mangelnde Passgenauigkeit zur spezifischen Situation der Datengeber, die Schwierigkeit der Erstellung eines methodisch sauberen Fragebogens, die Erfordernis einer klaren Anleitung sowie das Risiko, dass die Daten eher oberflächlich gegeben, etwa die Fragebögen schnell und ungenau ausgefüllt werden, genannt.

Für den Einsatz offener Verfahren sprechen aus Sicht der Befragten die Möglichkeit, individuelle Sichtweisen zu erfassen (größere Freiheit als bei standardisierten Verfahren, differenziertere Antworten, vielschichtigeres Bild des Evaluationsgegenstandes), eine höhere Verbindlichkeit (aufgrund der von den Probanden präferierten Interviewsituation), die Chance, während der Beschäftigung mit dem Thema neue Gedanken zu entwickeln, Interaktionsmöglichkeiten (z. B. für Rückfragen) sowie die mögliche Tiefe der Informationen, von der sich die Probanden neue und unerwartete Erkenntnisse erhoffen. Probleme offener Methoden werden hauptsächlich im erforderlichen Aufwand für die Erhebung der Daten als auch für deren Auswertung, in der dadurch begrenzten möglichen Stichprobengröße und in dem Risiko, relevante Themen durch wenig pointierte Aussagen der Befragten zu verpassen, gesehen.

Vor dem Hintergrund dieser Argumente wird auch schnell deutlich, welche Vorteile sich die Probanden/-innen vom Rückgriff auf Mixed Methods versprechen: Im Zentrum steht hier das Bestreben, die Vorteile der einen Methode mit jener der anderen zu verbinden und so auch spezifische Nachteile auszugleichen. Stark vereinfacht lässt sich der

Wunsch formulieren, eine Balance zwischen Zeiteffizienz und Antworttiefe herzustellen. Wie bereits dargestellt, wird dabei durchaus differenziert geurteilt: Die Befragten schlagen nicht einfach pauschal die Kombination verschiedener Methoden vor, oftmals lassen sich durch die Analyse der Interviews Begründungen für diese Urteile finden. Die zentralen Fragen, anhand derer sich die befragten Personen in diesem Teil des Interviews für oder gegen bestimmte Methoden entschieden haben, sind demnach die nach der Anzahl der für die Datenerhebung zu berücksichtigenden Personen, (Wenige zu befragende Personen zeitigen eine Tendenz zu offenen Verfahren, viele hingegen zu standardisierten.), die nach Art und Qualität der Informationen, die erhoben werden sollen (Handelt es sich um persönliche Erfahrungen? Oder liegen die Informationen bereits vor, beispielsweise in Form von Lehrberichten? Werden individuelle, wichtige Begründungen für Aussagen erwartet? Handelt es sich um einen Bereich, dessen relevante Inhalte nicht zu antizipieren sind?) und schließlich die Frage danach, ob eine Stakeholdergruppe Überblicks- oder Detailinformationen beisteuern soll. In Einzelfällen spielen auch wissenschaftliche Verortungen der Probanden eine Rolle, so beispielsweise bei B23, die/der prinzipiell qualitative Methoden bevorzugt[91].

Diese Erkenntnisse über die differenzierte Sichtweise der Befragten zeigen, dass die Stakeholder einer Evaluation sowohl bereit als auch in der Lage sind, Argumentationen für oder gegen Datenerhebungsmethoden zu folgen oder sogar selbst zu entwerfen. Eine auf Bequemlichkeit gründende Entscheidung für die Integration von Methoden aus unterschiedlichen Familien im Sinne der fiktiven Aussage „Alle Methoden auf einmal sind immer richtig!“ spielte – zumindest im hier betrachteten Setting – keine Rolle. Direkte Verbindungen zu den evaluationsspezifischen Vorerfahrungen einer Person sind aus dem vorliegenden Material ebenfalls nicht ersichtlich[92]. Vielmehr lässt sich aus den erhobe-

[91] „Also da bin ich jetzt sehr parteiisch, weil ich glaube, dass man vieles besser mit qualitativen als mit quantitativen Befragungen herausbekommt.“ (B23, 62)

[92] Die absolute Mehrheit der hier befragten Personen verfügt über zumindest grundlegende Erfahrungen mit Evaluationen. So haben 14 der insgesamt 24 hier betrachteten Personen selbst schon einmal Evaluationen durchgeführt oder in Auftrag gegeben, 21 wurden schon einmal als Datengeber/-innen in Evaluationen einbezogen. Lediglich eine Person verfügte über keinerlei Erfahrungen mit Evaluationen. Natürlich ist hierzu anzumerken, dass das Niveau der Erfahrungen sehr unterschiedlich sein kann. Die Durchführung von Evaluationen kann dabei der Organisation einfachen Feedbacks (B3,

nen Daten, wie bereits beim Thema der potenziell zu befragenden Stakeholdergruppen, wieder eine Verbindung zwischen der Komplexität des Wissens über den Evaluationsgegenstand bzw. seiner Wahrnehmung und der Wahl von Methoden ableiten: Wenn komplexes Wissen über den Gegenstand vorliegt oder wenn der Gegenstand als komplex wahrgenommen wird, werden tendenziell eher offene Methoden bevorzugt. Aus den Interviews wird ersichtlich, dass diese Folgerung in der Regel mit den sich eröffnenden größeren Freiheitsgraden dieser Verfahren begründet wird. Diese Sichtweise betont somit die Möglichkeit zur sich aus der konkreten Erhebungssituation heraus ergebenden Schwerpunktsetzung, zur Berücksichtigung des Unerwarteten als Stärke offener Verfahren. Dabei würden diese Verfahren nicht zwingend alleine eingesetzt, sondern häufig als Ergänzung zu standardisierten Verfahren.

Neben dieser Argumentationslinie lassen sich aus den hier besprochenen frei formulierten Vorschlägen der Befragten noch zwei weitere für die Methodenwahl relevante Aspekte ableiten. Einerseits lässt sich eine Art Begründungszwang für Planende erkennen. Die angeführten Gründe sind, abgesehen von auch teilweise erkennbaren prinzipiellen Verortungen der vorschlagenden Probanden/-innen, durchaus rationaler und nicht emotionaler Art. Das bedeutet gleichzeitig, dass die Stakeholder vermutlich auch für Argumentationswege von Planenden empfänglich sein werden, sofern diese gut begründet sind und die Stakeholder überhaupt über ein Mindestmaß von Interesse bezüglich des Evaluationsgegenstandes verfügen. Auf der anderen Seite bilden die ins Feld geführten Begründungen die in der Literatur diskutierten Vor- und Nachteile spezifischer Methoden durchaus gut ab, wenn – in Abhängigkeit von der fachlichen Vorbildung der betreffenden Personen – auch auf einem anderen Niveau. Hier wird wieder ersichtlich, dass Stakeholder (zumindest im hier betrachteten Bereich einer Hochschul-Lehrevaluation) durchaus in der Lage sein können, sich im Kontext einer Evaluation und unabhängig von ihren individuellen Evaluationserfahrungen mit den Eigenschaften von Methoden auseinanderzusetzen.

Zusammengefasst lassen sich aus diesem Bereich folgende Erkenntnisse über die Vorstellung der Probanden/-innen über die Gegenstandsangemessenheit von Datenerhe-

2) von bis zur Erstellung eigener oder zumindest Abwandlung bestehender Instrumente (B19, 2).

bungsmethoden sowie für deren Auswahl ableiten:

1. Stakeholder können sich relativ fundiert mit der methodischen Seite einer Evaluation auseinandersetzen. Persönliche Merkmale der Stakeholder spielen hier selbstverständlich eine Rolle, denn über diese Fähigkeit wird sicher nicht jede beliebige Stakeholdergruppe verfügen. Ihre Sicht sollte nicht unter dem alleinigen Verweis auf mangelnde Vorerfahrungen unberücksichtigt bleiben[93].

2. Die durch die Stakeholder angeführten Begründungen sind prinzipiell auf einer Linie mit jenen der Fachdiskussion. Dies ist auch ein Grund, ihre Urteile und Sichtweisen nicht pauschal zu verwerfen, sondern bewusst wahrzunehmen. Es sollte aber, wie in Abschnitt 3.4, S. 115 dargestellt, vermieden werden, dass die Professionalität der Planenden untergraben wird.

3. Zunehmende Komplexität des Evaluationsgegenstandes bzw. seiner Wahrnehmung führen in der Tendenz dazu, dass Methoden mit offenen Anteilen oder Methodenkombinationen, in denen offene Methoden eine tragende Rolle spielen, bevorzugt werden.

4. Für die Planenden einer Evaluation ergibt sich die Erfordernis, den Einsatz der gewählten Methoden sorgfältig darzustellen und zu begünden, da fundierte Begründungen vermutlich helfen werden, die Akzeptanz der eingesetzten Datenerhebungsmethoden und damit der gesamten Evaluation zu steigern.

Bewertung der tatsächlich eingesetzten Datenerhebungsmethoden

Anschließend an die Frage nach den Datenerhebungsmethoden, die die Probanden/-innen selbst für die Erfassung der relevanten Informationen der von ihnen vorgeschlagenen Stakeholdergruppen einsetzen würden, wurden sie gebeten, die Methoden, die im Rahmen der tatsächlich durchgeführten Evaluation der Fachdidaktik-Ausbildung

[93] Selbstverständlich ist davon auszugehen, dass vorhandene Vorerfahrungen die Güte der Betrachtung, Argumentation und Beurteilung von Methoden positiv beeinflussen werden. Allerdings ist das Fehlen eben dieser kein Hinweis darauf, dass die Sichtweise der Personen nicht doch treffend sein könnte.

eingesetzt worden waren, zu bewerten. Dazu wurden die drei verwendeten Methoden *Dokumenten- und Kennzahlenanalyse*, *Interviews mit Fachdidaktik-Verantwortlichen ausgewählter Fachbereiche* sowie *Teilstandardisierte Online-Datenerhebung unter allen als relevant identifizierten Studierenden* (siehe auch S. 118 ff.) einzeln abgefragt. Die Probandinnen/ Probanden wurden jeweils gebeten zu beurteilen, ob sie die entsprechende Methode für geeignet halten, die Ziele der Fachdidaktik-Evaluation zu erreichen.

Die Sichtweise der befragten Personen auf diese Thematik ist eindeutig: Insgesamt treffen die eingesetzten Methoden auf hohe Akzeptanz[94]. Jeweils 18 Personen hielten die Methoden der Dokumenten- und Kennzahlenanalyse und des Online-Fragebogens für geeignet, sogar 22 Personen die offenen Interviews. Die übrigen Probanden/-innen würden die jeweilige Methode nicht einsetzen oder gaben keine klare Bewertung ab[95]. Zu dieser deutlichen Zustimmung (75% bzw. 92%) ist jedoch anzumerken, dass in den Interviews häufig Rückfragen zum konkreten Einsatz der jeweiligen Methode gestellt wurden. Erst nach einer Darlegung ihrer Verwendung und ggf. nach kurzen Gesprächen darüber fällten die Befragten ihre Urteile. Es ist gleichfalls interessant, dass keiner Person alle eingesetzten Datenerhebungsmethoden als ungeeignet erschienen und lediglich eine Person (B15) zwei Verfahren, nämlich die Dokumenten- und Kennzahlenanalyse sowie die standardisierte Online-Datenerhebung, ablehnte.

Werden die hier über die tatsächlich eingesetzten Methoden gefällten Urteile den vorher gemachten Vorschlägen einzusetzender Datenerhebungsmethoden gegenübergestellt, so wird ersichtlich, dass die Urteile sich nicht unbedingt mit den eigenen Vorschlägen decken. Dieser Effekt wird darauf zurückzuführen sein, dass Angaben über die tatsächlich eingesetzten Methoden und die Frage nach ihrer Sinnhaftigkeit für den Evaluationszweck in weiterführende Informationen eingebettet wurden, wodurch eine Einschätzungsgrundlage entstand. Zusätzlich konnten die Methoden aufgrund des Aufbaus des Interviews vor dem Hintergrund der vorher angestellten eigenen Überlegungen

[94] Eine Übersichtstabelle findet sich im Anhang auf S. 259.

[95] Im Interview B1 wird diese Frage für die Methode der Online-Datenerhebung nicht beantwortet, so dass hier nur 23 Antworten vorliegen.

betrachtet werden. Diese forcierte Auseinandersetzung mit den Datenerhebungsmethoden führte zu einem positivien Effekt hinsichtlich ihrer Akzeptanz. Auch hier lässt sich wieder eine Bereitschaft zur Reflexion und Auseinandersetzung mit den methodenbezogenen Begründungen erkennen.

Mit Blick auf die Frage nach der Vorstellung der Gegenstandsangemessenheit von Datenerhebungsmethoden kann festgehalten werden, dass die Vorstellung der Stakeholder einer Evaluation nicht unveränderbar festgelegt ist, sondern durch fundierte Informationen beeinflusst werden kann, allerdings nicht im Sinne der „Produktion einer erwünschten Sichtweise“, sondern der Veränderung durch überzeugende Argumente. Ob die Stakeholder argumentativ von einer Methode, die ihnen grundsätzlich nicht geeignet erscheint, überzeugt werden können, wurde hier nicht untersucht. Hier wird ersichtlich, dass Methoden nicht aus sich selbst heraus verständlich sind, sondern dass es vielmehr erforderlich ist, durch Informationen über die gewählten Methoden Akzeptanz zu schaffen.

Bewertung der in Lehrevaluationen an Hochschulen üblichen Methoden

Zum Abschluss des Interviewteils über Datenerhebungsmethoden wurden die Probanden/-innen gebeten, die Datenerhebungsmethoden, die häufig zur Evaluation der Lehre an Hochschulen eingesetzt werden (vgl. S. 77 ff.), zu bewerten. Auch hier diente die fiktive Evaluation der fachdidaktischen Ausbildung als Grundlage für die Bewertungen. Neben der Aussage, ob die befragten Personen die Methoden einsetzen würden oder nicht, wurde auch erfasst, ob sie die jeweilige Methode eher zur Erfassung von Daten in standardisierter oder in offener Form verwenden und welche praktische Umsetzungsform (z. B. Papierfragebogen, Face-to-face-Interview, Online-Fragebogen oder Interview in einer Face-to-face-Situation, per Telefon, Mail o. ä.) sie bevorzugen würden.

Standardisierte Datenerhebungsverfahren würden von vielen der Befragten zumindest als ein Element der Evaluation eingesetzt, 21 der 24 Personen sprachen sich dafür aus, keine Person lehnte dieses Verfahren gänzlich ab. Allerdings beziehen sich diese Aussagen ausschließlich auf teilstandardisierte Varianten dieses Verfahrens, denn keine

der befragten Personen sprach sich für die Umsetzung in vollstandardisierter Form aus. Als Hauptgrund dafür wurde immer wieder der Wunsch geäußert, auch individuelle Sichtweise zu erfassen, die mit einer Vollstandardisierung aus Sicht der Befragten nicht abgebildet werden könnten. So formulierte etwa B20: „Ja, weil das halt nun einfach immer eine sehr individuelle Geschichte ist und diese standardisierten Dinge, damit kann man schon relativ viel herausfinden, aber das Andere [teilstandardisierte Verfahren, Anm. d. A.] ist sicherlich differenzierter." (B20, 70). Gleichwohl wird auch gesehen, dass die Auswertung teilstandardisiert erhobener Daten aufwändiger werden kann als die vollstandardisiert erhobener[96].

Wird die bevorzugte Präsentationsform der Erhebungsinstrumente betrachtet, so sprach sich die Mehrheit der Probanden/-innen für eine Online-Umsetzung aus (16 Befürwortungen, drei Ablehnungen), gefolgt vom klassischen Papierfragebogen (zwölf Befürwortungen, zwei Ablehnungen). Am wenigsten gewünscht wurde eine Durchführung der Datenerhebung als quantitatives Interview, also in einer Face-to-face-Situation (zwei Befürwortungen, zehn Ablehnungen). Die wichtigsten Argumente für den Rückgriff auf Online-Verfahren stellten die gewonnene Flexibilität sowohl für die Befragenden als auch für die Beantwortenden sowie der geringe organisatorische Aufwand für die Verteilung und den Rückerhalt der Fragebögen dar. Außerdem wurde mehrfach erwähnt, dass dieses Medium der Lebenswelt und der Arbeitsweise der Personen im Hochschulkontext entspräche. Die beiden Hauptargumente für den Einsatz eines Papierfragebogens hingegen waren eine im Vergleich zur Online-Datenerhebung vermuteterweise etwas höhere Verbindlichkeit der Ausfüllsituation, die im hier betrachteten Bereich in der Regel Teil einer Lehrveranstaltung wäre, sowie individuelle mediale Vorlieben, also etwa die persönliche Bevorzugung von Drucksachen gegenüber am Computerbildschirm dargestellten Texten. Vor allem die Personen, die mediale Vorlieben als Grund für die Bevorzugung des Papierfragebogens anführten, könnten sich aber auch mit einer anderen medialen Präsentationsform arrangieren[97]. Das quantitative Interview, also eine

[96] B16 formuliert hier etwa: „Und mit diesem offenen, da ist auch das Problem, dass eben dann die Bewertung, also die Auswertung von den Evaluationsbögen, die man dann am Ende hat, die ist eindeutig umfangreicher und aufwendiger, als eben die andere." (B16, 78)

[97] Ein plakatives Beispiel ist hier eine Aussage von B10: „Ich persönlich bin vielleicht etwas traditionell

Face-to-face-Situation, wurde meist mit dem Hinweis abgelehnt, dass die Datengeber/-innen den Fragebogen auch selbst ausfüllen könnten und dass die Befragten sich die Situation zwischen interviewender und befragter Person eher unangenehm vorstellten. Als positives Gegenargument wurde angeführt, dass die entstehende soziale Situation und die Möglichkeit, ggf. Nachfragen stellen zu können, sich positiv auf die Daten auswirken könnten.

Auch *offene Datenerhebungsverfahren* würden von der Mehrheit der befragten Personen eingesetzt. 21 Personen würden teilstrukturierte Verfahren zur Datenerhebung wählen, zwei von ihnen zusätzlich auch gänzlich offene. Die verbleibenden drei Personen sprechen sich indes auch nicht gegen den Einsatz teilstrukturierter Verfahren aus, sondern kamen lediglich zu keinem klaren Urteil. Acht der Befragten allerdings sprachen sich ausdrücklich gegen gänzlich offene Verfahren aus. Allgemein wird zwar durchaus ein möglicherweise wertvoller Beitrag der gänzlich offenen Verfahren gesehen, nämlich dass die Probanden die aus ihrer Sicht für die Evaluation insgesamt tatsächlich wichtigen Dinge zur Sprache bringen können. Genau in diesem Punkt wurde aber auch die größte Gefahr der gänzlich offenen Verfahren gesehen, und zwar die prinzipiell gegebene Möglichkeit, dass die Datengeber/-innen zwar interessante, im Sinne der Datenerhebung jedoch wenig oder gar nicht relevante Inhalte beitragen, was als nicht effizienter Zeit- und Ressourcenverbrauch angesehen wird. Die Teilstrukturierung der Daten, im Gespräch oftmals durch Rückgriff auf Leitfäden charakterisiert[98], wurde jedoch als Möglichkeit gesehen, den Themenverlauf des Interviews zielgerichtet zu gestalten, ohne die Aussagen der Befragten zu beeinflussen, und so dem Problem der für die Evaluation nicht inhaltstragenden Informationen zu begegnen.

Die von den Befragten bevorzugte Präsentationsform offener Datenerhebungsmetho-

und halte immer noch eine ganze Menge vom Fragebogen, den ich in Papierform vor mir habe, auch wenn ich weiß, dass natürlich eine Befragung am Computer viele Vorzüge hat. Aber es ist vielleicht eine Gewohnheitssache, ob Sie dann lieber vor dem Bildschirm sitzen und darüber nachdenken oder ob Sie das Ding auf den Schreibtisch legen, sich mit einer Tasse Kaffee zurücklehnen oder vielleicht in der Couch sitzen und dann darüber nachdenken.“ (B10, 68)

[98] Diese Empfehlung wurde eventuell auch durch die Interviewsituation selbst, in der ein Leitfaden zur Strukturierung genutzt wurde, befördert.

den stellt das Face-to-face-Gespräch dar, das von 20 der 24 Personen als positiv eingeschätzt wurde. Eine telefonische Datenerhebung wurde dagegen lediglich von zwei Personen als positiv bewertet, 20 hingegen lehnten diese Variante ab. Eine Umsetzung der Datenerhebung in schriftlicher Form, beispielsweise per E-Mail, wurde von drei Personen positiv, von zehn hingegen negativ bewertet. Längerfristige Verfahren schließlich, etwa ein individuell geführtes Fachdidaktik-Tagebuch, würden zwei Personen befürworten, sieben hingegen ablehnen.

Als wesentliche Vorteile eines Face-to-face-Interviews wurden der hohe Grad an sozialer Interaktion, etwa mit der Möglichkeit nachzufragen, der für die befragte Person vergleichsweise geringe Aufwand (abgesehen von den erforderlichen Terminvereinbarungen), die Möglichkeit der Steuerung des Interviews und des Feedbacks durch die interviewende Person, aber auch die Reichhaltigkeit der Kommunikation (Mimik, Gestik etc.) benannt. Negativ wurde, wie bereits angemerkt, vor allem der mit der Vereinbarung und der Anreise zum Treffpunkt verbundene relativ hohe Organisationsaufwand gesehen. Aber auch die sich für die Forschenden aus der Auswertung der in einem Interview erhobenen offenen Daten ergebende Arbeit, die als intensiver eingeschätzt wurde als jene, die zur Analyse standardisierter Daten erforderlich ist, wurde kritisch betrachtet. Bei der Abwägung der genannten Vor- und Nachteile einer Face-to-face-Datenerhebung überwiegen vor allem die Vorteile.

Als Gründe für die Ablehnung telefonischer Datenerhebungen wurden teilweise persönliche mediale Vorlieben, im Vergleich zu einer Face-to-face-Gesprächssituation fehlende Kommunikationskanäle, aber auch eine generell negative Konnotation aufgrund von persönlichen Erfahrungen mit Telefonwerbung und kommerziellen Telefonumfragen angeführt. Fehlende Kommunikations-, darüber hinausgehend aber auch fehlende direkte Interaktionsmöglichkeiten wurden auch von jenen Personen angeführt, die schriftliche Verfahren zur Erhebung offener Daten wie beispielsweise E-Mail oder Chat ablehnten. Darüber hinaus wurden der für die schreibende Person relativ hohe Arbeitsaufwand und damit erforderliche Einsatz sowie altersbedingte mediale Vorlieben bzw. altersspezifische Umgangsformen und Vertrautheiten mit den genannten Medien genannt. Sowohl

bei den telefonischen als auch bei den schriftlichen Verfahren wurde hingegen die mögliche Flexibilität als zentraler Vorteil gesehen, der in den Augen der Befragten jedoch die Nachteile nicht überwog. Werden schließlich noch die Gründe für die Ablehnung längerfristiger Verfahren, wie beispielsweise das Verfassen eines Fachdidaktik-Tagebuchs, betrachtet, so wurde auch hier als zentraler Ablehnungsgrund der durch die Befragten zu erbringende Zeiteinsatz ins Feld geführt. Als wieder nicht die Nachteile aufwiegender Vorteil wurde vor allen Dingen die mögliche Tiefe und Plastizität der so gewonnen Daten gesehen.

Im Unterschied zu den Bewertungen des Einsatzes der standardisierten und offenen Datenerhebungsmethoden ergibt sich bezüglich der drei verbleibenden Verfahren, die im Interview besprochen wurden, ein weniger einheitliches Bild. Sowohl *Gruppendiskussion* als auch *Dokumenten- und Kennzahlenanalyse* und *Gutachterverfahren* wurden seitens der Probanden/-innen positive wie auch negative Seiten abgewonnen.

Die meisten Befürworter/-innen fand das Verfahren der Dokumenten- und Kennzahlenanalyse: Elf Personen schätzten es als geeignet ein, sechs Personen sprachen sich gegen seinen Einsatz aus, sieben trafen keine abschließende Bewertung. Generell wurde das Verfahren relativ differenziert gesehen. Zum einen wurde von den Befürworter/-innen stets einschränkend geäußert, dass nicht jede Zahl im Sinne des fiktiven Evaluationsziels aussagekräftig sein müsse. Zwar sei es beispielsweise durchaus wichtig zu wissen, wie viele Personen eines Studienjahrgangs ein Studium nicht mit dem angestrebten Abschluss verlassen (Abbrecherquote), unbeantwortet bliebe aus Sicht der Befragten dabei jedoch die Frage, ob die Fachdidaktik-Ausbildung mit dem Studienabbruch überhaupt in Verbindung stehe (vgl. etwa B5, 120 oder B7, 100). Dies führte wiederum zu der Forderung, ausgewählte durch Dokumenten- und Kennzahlenanalyse gewonnene Zahlen zwar durchaus zu nutzen, jedoch nicht als alleinige Datenquelle, sondern eben als Ergänzung zu anderen Methoden und Daten. In dieser Kombination wurde das Verfahren durchaus als gewinnbringend eingeschätzt. Der Hauptgrund für die generelle Ablehnung dieses Verfahrens bezog sich indes auf den gleichen Umstand, der auch zur

Begründung der eingeschränkten Tauglichkeit dieses Verfahrens herangezogen wurde. Die das Verfahren ablehnenden befragten Personen wandten ein, dass die so zu gewinnenden Daten nichts über die Hintergründe der Ausbildung aussagten und somit letztlich verzichtbar seien. Den Vorschlag einer Ergänzung brachten sie nicht ein. Die Frage, ob die ablehnenden, aber auch die unschlüssigen Personen das Verfahren unter der Bedingung, dass es nur zur Bereitstellung ergänzender Informationen genutzt würde, eher akzeptieren würden, konnte aus dem Material nicht eindeutig ermittelt werden. Etwas überraschend ist, dass sich hier sieben Personen weniger positiv bezüglich dieses Datenerhebungsverfahrens äußerten, als der in der tatsächlich durchgeführten Evaluation eingesetzten Variante der Analyse der Modulbeschreibungen zugestimmt hatten (11 vs. 18). Mögliche Gründe für diese Differenz könnten die im Interview angeführten Beispiele sein, die den Befragten im Kontext der fiktiven Evaluation unter Umständen ungeeignet erschienen sein könnten[99]. Diese Tatsache könnte aber auch darauf zurückzuführen sein, dass die Dokumenten- und Kennzahlenanalyse der real durchgeführten Fachdidaktik-Evaluation in einen Kontext eingebunden war und aus diesem heraus logisch erschien.

Auch für den Einsatz des Verfahrens der Gruppendiskussion, das ebenfalls zur Diskussion gestellt wurde, lassen sich unter den befragten Personen keine klaren Mehrheiten erkennen. Jeweils neun Personen würden das Verfahren einsetzen oder gaben keine klare Bewertung dazu ab, sechs Personen sprachen sich gegen seine Nutzung im Kontext der fiktiven Fachdidaktik-Evaluation aus. Der wesentliche Zugewinn, den die Befürworter/-innen dieses Verfahren durch seinen Einsatz erwarten würden, ist, dass verschiedene Personengruppen, die durch den Evaluationsgegenstand betroffen sind, miteinander ins Gespräch kommen könnten. In diesem Gespräch könnten demnach Standpunkte diskutiert, Positionen überdacht und ein größeres Verständnis erreicht werden (vgl. etwa B7, 95 – 98 oder B18, 80). Somit handelt es sich in der Vorstellung der Befragten weniger um ein rein der Informationsgewinnung dienendes Verfahren als

[99] Als Beispiele für mögliche Kennzahlen bzw. Erkenntnisse aus Dokumentenanalysen waren im Leitfaden die Analyse von Modulbeschreibungen, die Anzahl von Studierenden in einem Seminar („Seminarteilnahmen“) sowie absolvierte Prüfungsleistungen verankert.

eher um eine Abstimmung, um die Erweiterung und ggf. Korrektur der Sichtweisen der befragten Personen, also um ein im Sinne Lamneks eher *vermittelndes* als *ermittelndes* Verfahren (vgl. Lamnek 2005a, S. 29 ff.). Während ein Teil der befragten Personen in der unmittelbaren sozialen Situation also einen Vorteil sieht, stellt diese aus Sicht der Personen, die keine klare Bewertung abgaben oder die Gruppendiskussion ablehnten, auch die zentrale Quelle für Risiken dar. Sie führten an, dass möglicherweise entstehende Streitgespräche jenseits der inhaltlich relevanten Themen und Meinungen die gesamte Diskussionsrunde dominieren und die aufgrund eines möglichen Machtgefälles der Beteiligten (z. B. Lehrende und ihre Studierenden in einer Gruppendiskussion) eventuell taktisch ausgerichtete Kommunikation den Wert der so gewonnen Daten erheblich schmälern könnten. Darüber hinaus sahen sie die datenerhebenden Personen hohen Anforderungen gegenüber, vor allem im Bereich der Auswertung der erhobenen Daten, aber auch in Bezug auf die terminliche Organisation.

Bei der Frage nach der Eignung von Gutachterverfahren schließlich lässt ebenfalls sich keine klare Tendenz erkennen: Jeweils acht Personen schätzten diese Methode als sinnvoll ein, lehnten sie ab oder kamen zu keiner klaren Bewertung. Das hauptsächliche Argument für den Einsatz eines solcher Verfahren stellte der Wert dar, welcher der Außensicht zugestanden wird, die die Gutachter/-innen einbringen. So wurden Verfahren dieser Gruppe als geeignet angesehen, Betriebsblindheit zu kompensieren oder ihr zumindest ein andere Sichtweise zur Seite stellen zu können. Die zentrale Problematik wurde in den begutachtenden Personen und ihrer Auswahl gesehen. Die Gutachten, so eine geäußerte Sorge, könnten unter Umständen interessengeleitet sein. Auch könnte es schwierig sein, Personen zu finden, die über die erforderlichen Kompetenzen verfügen. Wie bereits bei Verfahren der Dokumenten- und Kennzahlenanalyse wurde der Wert der durch Gutachterverfahren gewonnenen Daten vor allem in der möglichen Ergänzung und Bereicherung der mit anderen Methoden erhobenen Informationen gesehen – diese Einschränkung wurde auch von sechs der acht das Verfahren befürwortenden Personen geäußert.

Auffällig ist, dass die drei letztgenannten Methoden deutlich weniger einheitlich bewertet wurden als Verfahren zur Erhebung standardisierter und offener Daten. Ein Grund könnte in dem Grad der Vertrautheit mit den Verfahren liegen: Während die beiden erstgenannten allen befragten Personen mehr oder weniger vertraut sein dürften, handelt es sich bei den letztgenannten wohl eher um Datenerhebungsmethoden, mit denen nur wenige Personen konkrete Erfahrungen sammeln konnten. Dies würde zu weniger konkreten Vorstellungen von den Möglichkeiten und Grenzen der Verfahren führen, somit letztlich auch zu einer stärkeren Abhängigkeit der Bewertung von den individuellen Vorurteilen. Bei den Gutachterverfahren hingegen könnte eventuell ein genau gegenteiliger Effekt eingetreten sein: Bei den Probanden/-innen handelte es sich ausnahmslos um Lehramtsstudierende, in deren Ausbildung Unterrichtshospitationen ein wichtiges Element darstellen, aber insgesamt eher kritisch gesehen zu werden scheinen[100]. In beiden Fällen würde die Erfahrung mit den jeweiligen Methoden Einfluss auf ihre Bewertung ausüben.

Auf einer Metaebene lassen sich aus diesen Aussagen folgende Erkenntnisse über die Sichtweise der befragten Stakeholder auf die Eignung von Methoden ableiten:

1. Generell werden Methoden bevorzugt, die eine Mischung aus Struktur und Freiheit bieten. Die beiden Extreme der gänzlich standardisierten Datenerhebung und der komplett offenen Datenerhebung werden gleichermaßen kritisch gesehen. Es besteht demnach der Wunsch, dass die Datenerhebung im Sinne des Erkenntnisinteresses zwar durchaus gerichtet stattfindet, gleichwohl aber Freiräume zur individuellen Schwerpunktsetzung erhalten bleiben bzw. eröffnet werden. Teilstandardisierte Fragebögen werden hier ebenso als adäquates Mittel angesehen wie leitfadengestützte Interviews.

2. Der erforderliche Aufwand für die Datenerhebung fließt in die Urteile der Befragten ein. Als wünschenswert werden Methoden skizziert, bei denen der für alle

[100] Die empfundenen Parallelen zwischen Gutachterverfahren und Unterrichtshospitationen wurden im Anschlussgespräch an das Interview von vielen Personen dargestellt, ebenso die Skepsis und die Vorbehalte gegenüber diesem Verfahren.

Beteiligte entstehende Aufwand in einem günstigen Verhältnis zu den erhobenen Daten steht. Als negatives Beispiel kann hier das quantitative Interview angeführt werden, das aus Sicht der Befragten keine oder doch nur sehr geringe Vorteile gegenüber einem klassischen Fragebogen mit sich bringt, mit dem aber ein deutlich höherer Organisations- und u. U. auch Zeitaufwand einhergeht.

3. Soziale Präsenz während der Erhebung wird nicht generell gewünscht, jedoch durchaus als hilfreich angesehen und akzeptiert, wenn sie zu einem Mehrwert führen kann. Gleichzeitig scheint es so, als sei die Stärke des Wunsches nach sozialer Präsenz auch durch individuelle Medien-, aber auch Kommunikationsvorlieben beeinflusst.

4. Das tatsächliche Wissen über ein Datenerhebungsverfahren scheint Einfluss auf seine Wahrnehmung auszuüben. Verfahren, die den Stakeholdern vertraut sind, können differenzierter bzw. kritischer betrachtet werden als Verfahren, die ihnen nicht vertraut sind, über die sie nur Mutmaßungen anstellen und bei deren Bewertung sie vorrangig auf Vorurteile bzw. eine naive Sichtweise zurückgreifen können. Dieser eher augenscheinliche Sachverhalt kann dabei in zwei Richtungen wirken, einerseits in Richtung einer positiven Wahrnehmung und eines konstruktiven, ggf. auch kreativen Umgangs mit den bekannten Methoden, andererseits jedoch auch in negativer Richtung, verbunden mit einer eher prinzipiellen Ablehnung, wie beispielsweise bei den Verfahren der Dokumenten- und Kennzahlenanalyse oder bei Gutachterverfahren.

5. Datenerhebungsverfahren, denen gegenüber Stakeholder prinzipiell kritisch eingestellt sind, werden nicht kategorisch abgelehnt. Die Eignung zur Ergänzung von Daten, die mit anderen Verfahren erhoben wurden, wird ihnen zumindest teilweise zugestanden. Der Einsatz solcher Verfahren wird dann eher akzeptiert, wenn die mit ihrer Hilfe gewonnen Daten nicht zur alleinigen Grundlage einer Bewertung werden, sondern zur Ergänzung anderer Daten dienen, also in einem ggf. korrigierenden oder erläuternden Kontext stehen.

Benannte Kriterien zur Bewertung der Angemessenheit

Zum Abschluss des Interviews wurden die Probanden/-innen gebeten, auf einer abstrakten Ebene zu beschreiben, welche Kriterien ihnen nach dem Gespräch als wesentlich erscheinen, die Angemessenheit von Datenerhebungsmethoden in Evaluationen zu beurteilen. Für die Beantwortung dieser Frage sollten die Befragten die fiktive Rolle einer evaluationsplanenden Person verlassen und stattdessen als „normale/r“ Stakeholder/-in antworten. Ziel dieser Frage war die Abstraktion der bisherigen Überlegungen zur Gegenstandsangemessenheit und, soweit möglich, das Vornehmen einer individuellen Gewichtung zwischen den unterschiedlichen genannten Aspekten.

Die individuellen Vorschläge der Probanden wurden analysiert und gemäß ihrer Zielrichtungen zu 19 Beurteilungsfaktoren zusammengefasst. Anschließend wurden ähnliche Faktoren zu thematischen Gruppen gebündelt. Es ergaben sich die folgenden fünf Gruppen (in den Klammern hinter den einzelnen Beurteilungsfaktoren ist jeweils die Anzahl der unterschiedlichen Personen, auf deren Vorschläge das entsprechende Kriterium zurückzuführen ist, vermerkt)[101]:

Ergebnisnutzung und Konsequenzen:

- Konsequenzen aus der Evaluation, Umsetzung und/oder Weiterverwendung der Ergebnisse (13)
- Rückmeldung der Ergebnisse und Erkenntnisse an die Stakeholder (11)
- Zielgerichtete (im Sinne von Forschungsfrage und Datenverwertung) Datenerhebung, Passgenauigkeit der eingesetzten Verfahren zum Erkenntnisinteresse (7)
- Persönlicher Nutzen aus der Beteiligung (2)

[101] Einige der in der Liste aufgeführten Beurteilungsfaktoren weisen Bezüge zu mehr als einer Gruppe auf. So könnte beispielsweise der Aspekt der *Rückmeldung der Ergebnisse und Erkenntnisse an die Stakeholder* neben der Gruppe *Ergebnisnutzung und Konsequenzen* auch der Gruppe *Transparenz und fairer Umgang mit den Datengeber/-innen* zugerechnet werden. Um jedoch die Übersicht zu wahren wurden alle Kriterien lediglich einer der Gruppen zugeordnet, nämlich stets jener, welche sich am deutlichsten in den Aussagen der Probanden/-innen widerspiegelt.

Individuelle Berücksichtigung und Freiheitsgrade:

- Ausreichend Antwortfreiheitsgrade während der Datenerhebung (11)
- Intensive Berücksichtigung jeder einzelnen befragten Person, verbunden mit der Möglichkeit, deren individuelle Sichtweisen und eventuell durch sie wahrgenommene Besonderheiten des Evaluationsgegenstandes darstellen zu können (5)
- Wertschätzung der befragten Personen und ihrer Beiträge (1)

Genauigkeit und Richtigkeit:

- Genauigkeit der Abbildung des Gegenstandes (4)
- Erzeugung eines möglichst umfassenden Bildes des Evaluationsgegenstandes (2)
- Saubere Methodik (1)
- Wertfreiheit der Evaluation (1)
- Die Herstellung eines größeren Bezugsrahmens, nicht nur Fokussierung auf einzelne Personen (1)
- Ausreichende Beteiligung von Datengeber/-innen, ausreichende breite Datenbasis (1)
- Transparenz der Bewertung (1)

Effizienz:

- Angemessener Zeitaufwand (5)
- Erkennbare und das Vorhaben unterstützende Struktur (3)

Transparenz gegenüber den Datengeber/-innen und fairer Umgang mit ihnen:

- Die gute Verfügbarkeit von Informationen über Ziele und Methoden der Evaluation (4)
- Ausreichende Information der potenziellen Probanden/-innen (2)

- Anonymität (2)

Diese fünf Gruppen umfassen die Hauptaspekte, die von den Befragten als relevant für die Beurteilung der Angemessenheit von Datenerhebungsmethoden im Kontext von Lehrevaluationen an Hochschulen benannt wurden. Eine zentrale Gruppe stellt die der *Ergebnisnutzung und Konsequenzen* dar. In 20 der hier analysierten 24 Interviews wurden Kriterien geäußert, die dieser Gruppe zuzurechnen sind. Der gemeinsame Wunsch, der hinter den dieser Gruppe zugeordneten Kriterien steht, ist, dass eine Evaluation tatsächlich zu Veränderungen führen soll[102]. Der Gegenentwurf zu diesem Anspruch wäre demnach eine Evaluation, deren Ziel nach erfolgter Datenerhebung nicht weiter verfolgt wird, deren Ergebnisse nicht kommuniziert werden oder aus der, beispielsweise trotz identifizierter Missstände, keine Veränderungen (hin zum Positiven) hervorgehen. B13 bringt diese Sichtweise sehr pointiert zum Ausdruck:

> „Also ich möchte einmal die Auswertung der Evaluation danach sehen oder irgendwie mitbekommen. Und ich möchte, dass das auch Konsequenzen hat. Und diese Konsequenzen möchte ich auch spüren.“ (B13, 142)

Etwas weniger verwendungsorientiert ist der Wunsch, dass den Datengeber/-innen und Stakeholdern die Ergebnisse nicht vorenthalten, sondern zugänglich gemacht und zurückgemeldet werden sollen. Hier lassen sich zwei unterschiedliche Teilaspekte erkennen: Einerseits möchten die Stakeholder gerne informiert werden, sowohl um mehr über den Evaluationsgegenstand und eventuell sich ergebende (oder auch ausbleibende) Veränderungen zu erfahren als auch um sich selbst innerhalb der Gesamtheit der Stakeholder verorten zu können[103]. Andererseits dient eine solche Ergebnisrückmeldung aber auch der Schaffung einer Diskussionsgrundlage, da so die Basis der getroffenen

[102] Hier ist anzumerken, dass die Probanden prinzipiell von Veränderungen auszugehen scheinen, die entweder in ihrem Sinne positiv sind und von ihnen wahrgenommene Schwierigkeiten beheben oder die ihnen doch zumindest plausibel und vor dem Hintergrund der insgesamt erhobenen Daten sinnvoll erscheinen.

[103] Ein Beispiel für die erwünschte Selbstverortung wird von B2 eingebracht: „Wichtig ist ja aber auch, jetzt neben diesem persönlichen Interview, dass man auch merkt: Okay, da wurden noch mehr Leute befragt, also diese Massenerhebung, weil man vielleicht... vielleicht hat man ja selber so ein ganz negatives Bild und man weiß vielleicht, dass man selber so ein bisschen eher destruktiv oder negativ eingestellt ist, dass man eben weiß: Da werden noch mehr Leute befragt, das wird ein bisschen abgeglichen. Vielleicht bin ich ja nur so unzufrieden und andere Studierende zum Beispiel sehen das anders. Ja, dass man dann hinterher auch bei der Ergebnispräsentation sieht: Okay, da

Entscheidungen überprüfbar wird, gleichzeitig aber auch Veränderungen evident werden. Dieser Punkt verdeutlicht auch den Wunsch nach Begründungen. Hier besteht auch eine Verbindung zum Aspekt des persönlichen Nutzens aus der Beteiligung, der die eigene Person noch etwas mehr in den Mittelpunkt stellt. Dabei geht es beispielsweise darum, sich seiner selbst und seiner Position im Vergleich zu anderen Stakeholdern zu vergewissern (B12, 120) oder konkrete Anstöße zu Veränderungen im persönlichen Einflussbereich zu bekommen (B22, 112). Das Kriterium der zielgerichteten Datenerhebung und der Passgenauigkeit der eingesetzten Verfahren zum Erkenntnisinteresse steht in direktem Bezug zur Verwertung der Daten. Diesem Kriterium folgend müssten die erhobenen Daten möglichst eindeutig und direkt zur Erfüllung des Evaluationszwecks beitragen. Sehr deutlich ist hier eine Aussage von B14:

> „[. . .] in einer Evaluation sollten eigentlich nicht Fragen sein, wo man sich dann denkt: ‚Hä? Wozu wollen die das jetzt wissen? Das ist doch vollkommen unnötig, das zu wissen¡.“ (B14, 114)

Zusammenfassend kann festgehalten werden, dass angemessene Methoden nach Auffassung der befragten Personen in erster Linie dazu beitragen sollten, begründete Entscheidungen bezüglich des Evaluationsgegenstandes zu unterstützen oder zu ermöglichen, also letztlich möglichst konkrete Ergebnisse zu produzieren. Dabei kann zwischen einer evaluationsbezogenen und einer persönlichen Ebene, auf der die erwarteten Änderungen und/oder Wirkungen angesiedelt sein können, unterschieden werden.

Im Block *individuelle Berücksichtigung und Freiheitsgrade* geht es um die Möglichkeiten der Datengeber/-innen einer Evaluation, ihre individuellen Standpunkte und ihre spezifischen Situationen während der Datenerhebung darstellen zu können. B9 formuliert diesen Aspekt folgendermaßen:

> „Dass eben genügend Freiheit auch in dieser Evaluation sozusagen vorhanden ist. Also dass ich auch die Möglichkeit habe, meine eigene Meinung vielleicht noch mal hinzuschreiben oder darzulegen.“ (B9, 172)

wurden mehrere Leute befragt und ja, andere sehen das vielleicht ganz anders. Das ging mir jetzt bei einer, ich glaube bei einer [Thema]-Umfrage so, da habe ich dann gemerkt: ‚Huch, viele Leute sehen das gar nicht so kritisch wie ich¡oder so.“ (B2, 112)

An dieser Stelle bestätigt sich, was sich bereits im Abschnitt *Vorgeschlagene Stakeholdergruppen* auf S. 144 ff. andeutete, nämlich, dass der Erfassung der jeweils individuellen Sichtweisen der Befragten ein hoher Stellenwert zugebilligt wird. Aus diesem Anspruch lässt sich allerdings nicht der Wunsch nach dem vorrangigen Einsatz offener Datenerhebungsverfahren ableiten. Vielmehr muss er, wie ein Blick auf die oben dargestellten Ergebnisse bezüglich der präferierten Datenerhebungsmethoden klarstellt, als Argument gegen eine vollständige Standardisierung gewertet werden, für die sich keine der befragten Personen ausgesprochen hat. Doch auch diese Sichtweise auf die verschiedenen Eigenschaften von Datenerhebungsverfahren unterschiedlicher Familien (qualitativ/quantitativ) ist noch nicht umfassend genug. So hat etwa B5 darüber hinaus den Aspekt eingebracht, dass nicht nur die Verfahren, sondern auch die konkreten Erhebungsinstrumente angemessen sein müssen, und zwar den Besonderheiten der jeweils zu erfassenden Probanden/-innengruppe:

> „Ja, was ich immer gut finde, dadurch, dass ich ja jetzt hauptsächlich Fragebögen gemacht habe und so, und dann Sachen..., wenn die ziemlich individuell sind. Jetzt nicht ganz allgemein: ‚Wir fragen die Fachdidaktik ab von zwanzig Fächern.'. Sondern ich finde es eigentlich immer sehr gut, wenn die möglichst auf das einzelne Fach zugeschnitten ist, weil ich finde, die meisten kann man nicht vergleichen." (B5, 126)

In diesem Kriterienblock wird nochmals das Bewusstsein der Datengeber/-innen für die eigenen Besonderheiten deutlich und es wird gleichzeitig die empfundene Relevanz eben dieses speziellen individuellen Wissens für eine Evaluation betont. Als Anforderung an Datenerhebungsmethoden lässt sich daraus erneut ableiten, was ebenfalls bereits im Abschnitt *Vorgeschlagene Stakeholdergruppen* auf S. 144 ff. – dort jedoch auf anderer Ebene – besprochen wurde, nämlich dass die einzusetzenden Methoden flexibel und in der Lage sein müssen, den möglichen Beiträgen der datengebenden Gruppen angepasst und aufgrund der gebotenen Freiheitsgrade den individuellen Sichtweisen gerecht werden zu können.

Die Kriteriengruppe zum Thema *Genauigkeit und Richtigkeit* bezieht sich auf Faktoren, anhand derer sichergestellt werden könnte, dass die erhobenen Daten den Evalua-

tionsgegenstand genau, neutral, umfassend und nachvollziehbar beschreiben. Hier geht es u. a. darum, die unterschiedlichen Elemente und Bereiche des Evaluationsgegenstandes zusammenzufassen, beispielsweise die organisatorischen Rahmenbedingungen eines Studiengangs einerseits und die erlebte Studienrealität auf der anderen Seite (B3, 127). Ein anderer Aspekt ist etwa die Schaffung einer ausreichend breiten Datenbasis, die ebenfalls dazu dienen kann, ein umfassendes Bild zu erzeugen (B8, 126). Auf diese Weise wird gleichzeitig sichergestellt, dass die einzelnen erhobenen Datensätze in einen größeren Kontext eingebunden werden.

Die übrigen Kriterien führen bereits in die Richtung allgemeiner Qualitätskriterien, so etwa, wenn die Probanden/-innen eine saubere Methodik in Form einer sorgfältigen Arbeits- und einer logischen Vorgehensweise fordern (B10, 100), die prinzipielle Wertfreiheit der Evaluation anführen (B6, 159 – 161) oder Transparenz und Nachvollziehbarkeit der Ergebnisse und ihrer Produktion erwarten (B22, 113 – 118).

Für die Frage der Angemessenheit von Datenerhebungsmethoden lässt sich hieraus ableiten, dass geeignete Methoden einen detaillierten und präzisen Blick auf den Evaluationsgegenstand und einer genügend großen Anzahl von Personen die Teilnahme ermöglichen sollten. Zusätzlich müssten sie explizierbar sein, um Transparenz zu gewährleisten und um eine Überprüfung sowohl der Vorgehensweise als auch der erhobenen Daten zu ermöglichen.

Bei den Kriterien zum Thema *Effizienz* werden zwei wesentliche Teilaspekte ersichtlich. Einerseits sprechen die Probanden die tatsächlich durch sie in die Datenerhebung investierte Zeit an, wobei hier kein absolutes, sondern ein relatives Maß genannt wird: Der Zeitaufwand (sowohl der vorher erforderlichen Planungen als auch der eigentlichen Erhebung) sollte in einem guten Verhältnis zum erwarteten Ergebnis stehen. Gleichzeitig geht es nicht darum, möglichst wenig Zeit zu investieren, wie beispielsweise von B17 (Absatz 108) angeführt wird. Andererseits bezieht sich diese Gruppe auch auf die Struktur von Datenerhebungsmethoden, da diese dazu geeignet sein muss, die eigentliche Erhebung effizient zu gestalten. So formuliert etwa B4:

> „Mit einer Evaluation, die eine gewisse Freiheit einem lässt zur Beantwortung von Fragen, die aber auch eine gewisse Struktur hat, so dass sich das nicht im Nirvana verliert.“ (B4, 103)

Als angemessen beurteilte Datenerhebungsmethoden zeichnen sich demnach dadurch aus, dass sie den eigentlichen Erhebungsprozess durch eine Strukturierung zielgerichtet unterstützen und den erforderlichen Zeitaufwand so gering wie möglich halten.

Der letzte Block *Transparenz gegenüber und fairer Umgang mit den Datengeber/-innen* schließlich beinhaltet Kriterien, die sich auf die Themen der Information der Befragten und ihre Anonymität beziehen. Einerseits wird genannt, dass die Datengeber/-innen einer Evaluation in die Lage versetzt werden sollten, sich über die Ziele, Methoden und Hintergründe einer Evaluation fundiert informieren zu können. Dazu gehört auch ein Offenlegen der Bewertungsgrundlagen, nach denen der Evaluationsgegenstand evaluiert wird. Der andere Aspekt bezieht sich auf die Anonymität der Teilnehmer/-innen, die gewährleistet sein sollte, um negativen Auswirkungen durch die Teilnahme vorzubeugen. Datenerhebungsmethoden müssten es demnach einerseits ermöglichen, mit eventuell auftretenden Rückfragen umgehen zu können und andererseits müssten sie gewährleisten, dass die Identitäten der datengebenden Personen im weiteren Evaluations- und Auswertungsprozess unbekannt bleiben.

Im nächsten Schritt werden nun alle Äußerungen der Probanden über mögliche Kriterien und Indikatoren zur Bewertung der Angemessenheit von Datenerhebungsmethoden im Kontext von Lehrevaluationen an Hochschulen weiter integriert. Daraus ergeben sich folgende Anforderungen an *angemessene* Methoden:

- Sie sollen die Nutzung der Evaluationsergebnisse fördern und dazu beitragen, begründete Konsequenzen aus der Evaluation zu erarbeiten. Die erhobenen Daten sollten dementsprechend als Grundlage solch fundierter Entscheidungen dienen können. Auf konkreter Ebene bedeutet dies, dass die erhobenen Daten eine möglichst direkte inhaltliche Verbindung zur Fragestellung bzw. zum Evaluations-

zweck haben sollten und dass keine unnötigen Daten erhoben werden. Darüber hinaus sollte die Struktur der gewonnen Daten so geartet sein, dass die enthaltenen relevanten Informationen den Auswertungsmethoden gut zugänglich sind.

- Angemessene Datenerhebungsmethoden sind flexibel. Sie eröffnen den Befragten Freiraum, individuelle Schwerpunktsetzungen vorzunehmen sowie aus ihrer Sicht relevante Themen in die Evaluation einzubringen. Dabei ist die tatsächliche Anzahl der erforderlichen Freiheitsgrade nicht absolut, sondern immer im Kontext des jeweiligen Vorhabens zu bewerten. Nicht jeder Probandengruppe müsste also die gleiche Möglichkeit der individuellen Schwerpunktsetzung eingeräumt werden – jedoch ganz auf sie zu verzichten wird nicht gewünscht. Daneben sollten die Datenerhebungsinstrumente den jeweiligen Gruppen, innerhalb derer sie zur Anwendung kommen, angepasst werden können, etwa in Form gruppenspezifischer Fragebögen.

- Methoden, die als angemessen wahrgenommen werden, erzeugen ein detailliertes, nachprüfbares und präzises Bild des Evaluationsgegenstandes. Dazu ist es erforderlich, dass sie nachvollziehbaren und möglichst explizierbaren Regeln folgen und darüber hinaus sauber angewendet werden. Die mit ihnen erhobenen Daten sind überprüfbar, was zu Steigerung der Glaubwürdigkeit beiträgt. Methodenkombinationen werden hier als adäquates Mittel wahrgenommen.

- Angemessene Datenerhebungsmethoden helfen, den erforderlichen Zeitaufwand gering zu halten. Zu diesem Zweck ist eine zielführende Strukturierung des Erhebungsprozesses wünschenswert. Dabei muss die Balance zwischen leitender Strukturierung und gebotenen Freiheitsgraden gefunden und es muss vermieden werden, dass die Struktur die Ergebnisse bestimmt.

- Schließlich stellen angemessene Datenerhebungsmethoden die Anonymität der Datengeber/-innen sicher, um sie vor eventuellen negativen Auswirkungen ihrer Beteiligung zu schützen. Weiterhin ermöglichen sie Rückfragen, um Informationsdefizite reduzieren zu können.

Darüber hinaus wünschen die befragten Personen, über die eingesetzten Methoden und die mit ihnen verbundenen Intentionen informiert zu werden. Dabei handelt es sich zwar um eine Anforderung, die nicht die Eigenschaft einer Datenerhebungsmethode anspricht, die aber dennoch entscheidend für das Erreichen einer hohen Akzeptanz der gewählten Vorgehensweise durch die betroffenen Personen ist.

Zusammenfassung: Die Sicht der (potenziellen) Evaluationsteilnehmer/-innen

Werden die Erkenntnisse aus den Interviews mit (potenziellen) Evaluationsteilnehmer/-innen zusammengefasst, so ergeben sich drei Themenkomplexe, die aus ihrer Perspektive für die Beurteilung der Angemessenheit von Datenerhebungsmethoden und damit für Evaluationsplanende relevant sind:

1. **Spezifizität, Offenheit, Flexibilität und Genauigkeit.** Die gewählten Methoden zur Datenerhebung sollen sensibel für die Besonderheiten der Evaluation sein. Dazu gehören im Wesentlichen die Berücksichtigung der von den Befragten wahrgenommenen Charakteristika des Evaluationsgegenstandes sowie der besonderen Merkmale und Eigenschaften der zu befragenden Personen (-gruppen). Die Methoden sollen den Befragten Freiräume eröffnen, die eigene Sichtweise in die Datenbasis der Evaluation einzubringen. Gleichzeitig sollen sie so flexibel sein, dass sich die konkreten Erhebungsinstrumente den einzelnen Teilgruppen anpassen lassen. Zielvorstellung ist, dass das Bild, das aus den erhobenen Daten gezeichnet wird, detailgetreu, nachvollziehbar, überprüfbar und idealerweise auch multiperspektivisch sein sollte. Offene Anteile in der Datenerhebung werden prinzipiell begrüßt, gänzlich standardisierte Verfahren ebenso wie gänzlich offene eher kritisch gesehen. Die Verbindung von Verfahren der quantitativen und der qualitativen Familie wird als Möglichkeit zur Erfassung von Komplexität und zur Erreichung der Evaluationszwecke wahrgenommen. Die Datenerhebung sollte stets anonym erfolgen.

2. **Effizienz der Erhebung und Verwendung der Ergebnisse.** Die datengebenden Personen erwarten einen Nutzen von der Beteiligung an der Datenerhebung, und

zwar auf einer evaluationsbezogenen und/oder einer persönlichen Ebene. Die Datenerhebungsmethoden und die durch sie erhobenen Daten sollten somit klar der Erreichung des Evaluationszwecks und der Beantwortung der grundlegenden Fragen dienen. Hierzu kann gehören, dass die Datenerhebung in einer den geplanten Nutzen unterstützenden Art und Weise – beispielsweise durch einen zielführenden Interviewleitfaden – strukturiert wird. Verfahren, aber auch Fragen, deren Sinn nicht ersichtlich ist, sollten nicht eingesetzt werden. Daten sollten nach ihrer Erhebung in einer Form vorliegen, die ihre Anschlussverwendung ermöglicht und befördert.

3. **Transparenz.** Die Gründe für den Einsatz der gewählten Datenerhebungsmethoden sollten transparent gemacht und offengelegt werden. Den Stakeholdern soll die Möglichkeit geboten werden, sich über die mit dem Einsatz der Methoden verknüpften Intentionen und Begründungen zu informieren. So wird es den Stakeholdern ermöglicht, sich mit der Evaluation auseinanderzusetzen.

Diese Sichtweise der für die Untersuchung befragten Personen ist gleichermaßen differenziert wie pragmatisch. Hervorzuheben ist dabei, dass die Stakeholder prinzipiell dazu bereit zu sein scheinen, ihr Urteil zu überprüfen. In den Interviews hat sich gezeigt, dass die Probanden fundierte Begründungen zum Einsatz von Methoden akzeptieren, und zwar auch dann, wenn ihnen selbst die Verwendung einer der fraglichen Methoden zunächst nicht plausibel erschien oder sie die Methode anfänglich ablehnten. Es zeigt sich hier die Bedeutung der Berücksichtigung der Stakeholder und ihrer Sichtweise. Damit geht gleichzeitig ein weiterer Effekt einher: Durch die ausführliche Information der Stakeholder wird ihre Wahrnehmung der gesamten Evaluation gestärkt, ihr Engagement kann dadurch positiv beeinflusst werden. Eine Berücksichtigung dieser Aspekte wird der Akzeptanz der Evaluation zugutekommen.

4.2.2. Die Planenden

Über die Gruppe der (potenziellen) Evaluationsteilnehmer/-innen hinaus wurden auch die Personen, welche die real durchgeführte Fachdidaktikevaluation verantworteten, in

die empirische Untersuchung für die vorliegende Arbeit einbezogen. Dabei handelt es sich um zwei Personen, von denen eine schwerpunktmäßig für die inhaltliche (B25), die andere hingegen für die methodische Seite der Untersuchung verantwortlich zeichnete (B26).

Die Berücksichtigung der Sichtweisen der Planenden erfolgte mit dem Ziel, ihre Beweggründe für die Wahl der eingesetzten Datenerhebungsmethoden und ihre Sichtweise darauf zu erfassen. Da es sich lediglich um zwei Interviews handelte und die Herstellung einer Vergleichbarkeit oder die Bündelung von Inhalten hier weniger von Interesse war, sondern vielmehr die individuelle Expertensicht auf die real durchgeführte Evaluation im Fokus stand, wurde das zuvor angewendete Auswertungsverfahren abgeändert. Um einen Überblick über die Inhalte zu erlangen, wurden die Dokumente zunächst anhand des für die übrigen Einzelinterviews entworfenen Kategoriensystems auf einer allgemeinen, nicht detaillierten Ebene codiert. Die eigentliche Auswertung erfolgte einzelfallbezogen, indem die Inhalte jedes der beiden Interviews vornehmlich individuell-interpretativ und nur in Einzelfällen anhand der vergebenen Codierungen bearbeitet wurden.

Nachfolgend werden zunächst die Einschätzungen der Planenden der in der Fachdidaktik-Evaluation eingesetzten, aber auch der nicht eingesetzten Datenerhebungsmethoden dargestellt. Anschließend wird untersucht, welche Kriterien der Angemessenheit aus dieser Darstellung ersichtlich werden.

Einschätzungen der Datenerhebungsmethoden

Beide befragte Personen führten in den Gesprächen als zentrale Entscheidungsgrundlage für die Wahl der eingesetzten Methoden immer wieder die Fragestellung und das Ziel der Evaluation an. Der von ihnen benannte Hauptzweck der durchgeführten Evaluation war eine überblicksartige Erfassung, eine Bilanzierung der Fachdidaktik-Ausbildung gewesen, woraus sich für die Planenden klare Implikationen für die Wahl von Methoden ergaben. Dabei wurden folgende Argumente in den Interviews als ausschlaggebend für die methodischen Entscheidungen benannt:

- Die Dokumenten- und Kennzahlenanalyse wurde eingesetzt, um die formale Struktur des Studiums abbilden und bilanzieren zu können (B25, 3-4; B26, 39 – 40). Erfasste Informationen, wie etwa die Beschreibungen von Lehrinhalten im Modulhandbuch und die Anzahl der Lehrenden im fachdidaktischen Studienteil, wurden als gute Basis für die weitere Betrachtung der fachdidaktischen Ausbildung und als wertvolle Argumentationsgrundlage (B25, 7) gewertet. Zudem waren die Daten vergleichsweise einfach zu erheben (B25, 7).

- Das Verfahren der Interviews mit Fachdidaktikverantwortlichen wurde eingesetzt, um Details und Besonderheiten der in den Modulbeschreibungen abgebildeten Rahmenbedingungen und der dahinter stehenden Planungen ergründen zu können. Dabei handelte es sich um Informationen, die nicht vorwegzunehmen waren und deshalb den Einsatz einer offeneren Datenerhebungsmethode erforderten (B25, 3). Somit standen die befragten Fachdidaktik-Verantwortlichen einerseits in der Rolle von Experten für die Fachdidaktik-Ausbildung im jeweiligen Fach, gleichzeitig fungierten sie jedoch als Korrektiv der Dokumenten- und Kennzahlenanalyse (B26, 24). Zusätzlich wurden die Interviews auch als Zugeständnis an die Belastung und die Mitarbeitsbereitschaft der befragten Lehrenden gesehen: Eine Erfassung der in den Interviews erhobenen Informationen beispielsweise mit Hilfe eines Fragebogens wurde von B26 als unrealistisch eingeschätzt, da hierfür eine intensive, der Beantwortung der Fragen vorausgehende Auseinandersetzung der Probanden mit den Ergebnissen der Dokumenten- und Kennzahlenanalyse erforderlich gewesen wäre, die als kaum erfüllbar angesehen wurde (B26, 30-32). Während der Interviewsituation hingegen konnten fragliche Aspekte direkt angesprochen und erläutert werden. Zudem bot dieses Verfahren den Befragten die Möglichkeit, die aus ihrer Sicht besonderen Elemente hervorzuheben und zu erläutern. Somit stellte dieses Verfahren aus Sicht von B26 einen Beitrag zur Sicherstellung der Validität der erhobenen Daten dar (B26, 30 – 32).

- Die Entscheidung für die Umsetzung der Befragung der Studierenden in Form einer Online-Datenerhebung wurde vor allem mit organisatorischen Aspekten be-

gründet. Die Umsetzung in Form eines Face-to-face-Verfahrens hätte einen zu hohen zeitlichen Aufwand bedeutet, Papierfragebögen hätten in dem speziellen Setting vermutlich einen niedrigeren Rücklauf zur Folge gehabt (B26, 34). Beide Befragte führen zudem den nicht zu bewältigenden Aufwand der Verteilung von Papierfragebögen an Studierende in 20 Fächern an, die in unterschiedlichen Fachsemestern studieren und keine verpflichtenden gemeinsamen Veranstaltungen besuchen (B25, 4; B26, 34). Die Tatsache, dass die Gebäude der Universität Marburg zudem über das gesamte Stadtgebiet verstreut liegen, hätte die Organisation von Verteilung und Rückerhalt von Papierfragebögen zusätzlich verkompliziert. Auch die bei Papierfragebögen erforderliche Dateneingabe und die mit ihr verbundenen Kosten wurden als Argument für die Online-Umsetzung des Instruments angeführt. Über diese organisatorischen und finanziellen Aspekte hinaus führte B25 an, dass die mediale Form der Online-Fragebögen den Studierenden durch ihre Lebens- und Arbeitswelt vertraut sei und die zeitlich und räumlich flexiblen Ausfüllmöglichkeiten die Rücklaufquote vermutlich positiv beeinflussen würden (B25,4). Eine andere Seite der sich durch den Einsatz des Verfahrens bietenden Flexibilität wird von B26 hervorgehoben, nämlich die Möglichkeit, komplexe und zielgenaue Filterführungen zu integrieren, die beispielsweise auch dazu geeignet seien, in komplexeren Szenarien[104] individuell passgenaue Fragebögen zu erzeugen (B26, 36 – 38).

Die Zielsetzung der Evaluation, einen Überblick über die Realität und Praxis der Fachdidaktik-Ausbildung an der Universität Marburg zu ermöglichen, führte über die oben dargestellten, direkt methodenbezogenen Entscheidungen hinaus zu weiteren Entscheidungen, die ihrerseits wiederum Einfluss auf die Wahl der Methoden ausübten. Von besonderem Interesse sind im hier betrachteten Zusammenhang die Fragen der Auswahl der Grundgesamtheit und der daraus zu befragenden Personen. Wesentlich ist dabei, dass inhaltliche Fragen wie beispielsweise die Überprüfung der im Studium erworbenen

[104] Als konkretes Beispiel wird im Interview die Anzahl möglicher Kombinationen angeführt, die sich ergibt, wenn Personen befragt werden, die jeweils zwei aus insgesamt 20 zur Verfügung stehenden Fächern beinahe beliebig kombinieren können.

Kompetenzen nicht zur Zielsetzung der Evaluation zählten (B25, 2; B26, 2 – 8). Daraus wurde eine Eingrenzung des Stakeholderkreises auf die von der Thematik betroffenen Studierenden und Lehrenden der Universität Marburg abgeleitet. Weitere, externe Gruppen wie etwa Seminarleiter/-innen, Referendare/-innen oder Schüler/-innen wurden aus diesem Grunde nicht in die Datenerhebung einbezogen. Die zu befragenden studierenden Probanden/-innen wurden, wie in Abschnitt 4.1.2 dargestellt, anhand der zum Zeitpunkt der Datenerhebung absolvierten Fachsemester ausgewählt, um sicherzustellen, dass sie die in der Evaluation zu bewertenden Elemente bereits im Studium kennengelernt hatten.

Neben den bisher dargestellten positiven Argumenten äußerten sich die Planenden hinsichtlich der von ihnen eingesetzten Methoden teilweise auch kritisch. Wenn die wahrgenommenen Negativseiten die Positiven auch nicht überwogen, so ist es für die Frage nach der Gegenstandsangemessenheit dennoch wichtig, sie zu berücksichtigen, da sich durch sie das Bild komplettieren lässt. Folgende mögliche Einschränkungen der Methoden wurden benannt:

- Das Hauptrisiko der Dokumenten- und Kennzahlenanalyse wurde in der möglichen Übergewichtung und/oder alleinigen Betrachtung der durch sie gewonnen Daten gesehen. Als problematisch wurde dabei vor allem die für detaillierte Betrachtungen zu geringe Differenziertheit dieser Art von Daten angesehen (B25, 7). Als Lösungsvorschlag wurde eine Einbettung dieser Daten in einen weiteren Kontext also ihre Kombination mit zusätzlichen Daten vorgeschlagen. Auch B26 hebt hervor, dass Hintergründe von mit Dokumenten- und Kennzahlenanalysen erhobenen Daten nicht aus sich heraus sichtbar würden, sondern weiterer Analysen bedürften (B26, 26), was in der Evaluation in Form von Interviews mit Fachdidaktik-Verantwortlichen realisiert wurde. Für den Evaluationsgegenstand geht B26 von einer gewissen Differenz zwischen der Realität und den in den Dokumenten festgehaltenen Vorgaben aus (B26, 40).

- Bezüglich der Interviews mit Fachdidaktik-Verantwortlichen wurden zwei wesent-

liche Kritikpunkte geäußert (B25, 3). Erstens wurde das Verfahren als zeitintensiv charakterisiert, was dazu führte, dass nur vergleichsweise wenige Personen tatsächlich in Form von Interviews befragt werden konnten. Zweitens wurde auf die Gefahr hingewiesen, dass die in solchen Interviews erhobenen Antworten aufgrund verschiedener Motivationen nicht unbedingt wahrheitsgetreu seien. Als mögliche Beispiele wurden einerseits beschönigende, andererseits jedoch auch übertriebene Darstellungen von Schwächen des Evaluationsgegenstandes angeführt, die aufgrund taktischer Überlegungen geäußert werden könnten.

- Hinsichtlich der Online-Befragung der Studierenden vermutete B25, dass der Rücklauf des Online-Fragebogens u. U. geringer ausgefallen sein könnte, als wenn die Befragung in Form klassischer Papierfragebögen realisiert worden wäre. Als Grund wird angeführt, dass aufgrund der individualisierten Ausfüllsituation ein gewisser Handlungszwang durch ebenfalls den Fragebogen ausfüllende Kommilitonen/-innen entfalle[105] (B25, 4). Allerdings gewann B25 diesem Gesichtspunkt auch positive Seiten ab, nämlich, dass sich aufgrund des fehlenden Zwangscharakters der Ausfüllsituation von selbst eine Selektion der Ausfüllenden ergebe (nur jene mit realem Interesse beteiligen sich) und dass für die Planenden und Durchführenden der Evaluation eine gesteigerte Notwendigkeit guter und verstärkter Information und Argumentation entstünde, um Probanden für die Teilnahme zu gewinnen.

Über diese konkret auf einzelne Methoden bezogenen Negativaspekte hinaus sah B25 vor allem in der nicht optimalen Tiefe der Daten mögliche Ansatzpunkte für Kritik durch die Stakeholder (B25, 5). Dabei wurden allerdings die Rahmenbedingungen der Evaluation als Faktoren angeführt, die den Spielraum der planenden Personen begrenzten. B26 hingegen vermutete mögliche Äußerungen von Kritik der Stakeholder eher in Bezug auf die Ergebnisse der Evaluation, nämlich den Mangel an aus ihnen hervorgehenden (positiven) Veränderungen des Evaluationsgegenstandes für die Stakeholder (B26, 54 – 62).

[105] Dieser vermutete Effekt stellt für die befragte Person gewissermaßen die Kehrseite der positiv hervorgehobenen Ausprägungen der Individualisierung der Ausfüllsituation dar, namentlich der Möglichkeit, die Beantwortung flexibel nach den eigenen Bedürfnissen vornehmen zu können.

Zusätzlich zur Frage nach Argumenten für oder gegen den Einsatz der gewählten Methoden wurden die Planenden auch bezüglich der allgemein in Lehrevaluationen an Hochschulen üblichen, in der besprochenen Fachdidaktik-Evaluation jedoch nicht eingesetzten Datenerhebungsverfahren befragt. Dazu nahmen sie folgende Einschätzungen vor:

- Das Verfahren der Gruppendiskussion wurde als organisatorisch aufwändig, als zeit- und ressourcenintensiv beschrieben (B25, 5; B26, 68). B26 wertete es zudem als eher für einen Blick auf Details denn für einen Überblick gewinnbringend, der jedoch nicht zur Zielsetzung der hier betrachteten Fachdidaktik-Evaluation gehört habe (B26, 68). B25 benennt als möglichen positiven Nutzen das Entstehen eines Dialogs zwischen den Beteiligten, wodurch die Praxis des Fachdidaktik-Studiums realistisch erfahrbar und eine Reflexion der Beteiligten angeregt hätte werden können (B25, 6). B25 begreift die Gruppendiskussion somit sowohl als er- als auch als vermittelndes Instrument.

- Zu Gutachterverfahren wurden drei wesentliche Aspekte vorgebracht. Zum einen wurden sie als mit hohen Kosten für die Durchführung des Begutachtungsverfahrens verbunden eingeschätzt (B26, 72), zum zweiten wurde die mögliche Schwierigkeit, für das spezifische Thema der Evaluation kompetente Gutachter/-innen zu finden, thematisiert (B26, 74 – 78). Schließlich wurde von B25 die Frage des Umgangs mit den Ergebnissen der Begutachtung angesprochen. Hier wurden Szenarien als positiv skizziert, in denen die Hoheit über die Verwendung der Ergebnisse der Begutachtung auf Seiten der Begutachteten verbliebe, sich also keine Zwangsfolgen ergäben, und in denen die mit diesem Verfahren erhobenen Daten zudem nicht als alleinige Grundlage für Entscheidungen genutzt würden (B25, 8).

Zufriedenheit der Datengeber/-innen

Zum Ende der Interviews wurden die Planenden gebeten zu beschreiben, welchen Kriterien die Datenerhebung in einer Evaluation aus ihrer Sicht entsprechen müsste, um

beteiligte und/oder betroffene Personen zufriedenzustellen. Diese Frage zielte auf vermutete Bedürfnisse der Stakeholder. Folgende Aspekte wurden von den Befragten geäußert:

- Nach Ansicht von B26 stellt stets der Gesamtprozess der Evaluation die Grundlage der Zufriedenheit der Stakeholder dar. Es wird hervorgehoben, dass für eine Zufriedenheit der Stakeholder die gegebenen Daten auch zu Veränderungen führen müssten. Der insgesamt empfundene Nutzen für die eigene Person wird als ausschlaggebend für eine positive Sichtweise gewertet (B26, 80 – 88).

- Die Evaluationsfragestellung wird von beiden Befragten als wesentliches Kriterium angeführt, an dem die Stakeholder die Eignung der eingesetzten Methoden überprüfen können (B25, 9; B26, 92 – 96).

- Eine weitere Facette stellt aus Sicht von B25 die Frage dar, ob die Ergebnisse der Evaluation im Horizont der Erwartungen der befragten Personen liegen (B25, 9). Unerwartete Ergebnisse könnten dazu führen, dass die Erhebung (und mit ihr die gesamte Evaluation) nicht ernst genommen und/oder ihre Methodik in Frage gestellt würde.

Keiner dieser Punkte bringt grundlegend neue Aspekte in die bisher erarbeitete Kriterienliste zur Beurteilung der Gegenstandsangemessenheit ein, der letztgenannte führt aber zu einer weiteren Begründung für die Wahl von Methoden, die den Befragten Freiraum zum Äußern der eigenen Meinung eröffnen, um sicherzustellen, dass in die Evaluation einfließen kann, was den Befragten wichtig ist. Außerdem handelt es sich um ein weiteres Argument für eine ausführliche Begründung der Evaluationsergebnisse, da auf diese Weise auch Unerwartetes für die Befragten verständlich und akzeptabel werden könnte.

Aspekte der Gegenstandsangemessenheit

Aus dieser Darstellung der Sichtweisen der Planenden lassen sich für die Frage nach der Gegenstandsangemessenheit von Datenerhebungsmethoden im Kontext von Lehre-

valuationen an Hochschulen folgende Kriterien ableiten:

- Zentrales Kriterium für die Bestimmung der Angemessenheit von Datenerhebungsmethoden stellt aus Sicht der Planenden die der Untersuchung zugrunde liegende Forschungs- bzw. Evaluationsfragestellung und damit der avisierte Zweck der Evaluation dar. Aus der Fragestellung gehen unmittelbare Anforderungen an Datenerhebungsmethoden hervor, beispielsweise in Form der Anzahl zu berücksichtigender Stakeholder, des erforderlichen Detailgrades der zu erhebenden Informationen oder auch organisatorischer Rahmenbedingungen, wie etwa der räumlichen Verteilung der zu befragenden Personen.

- Die Methoden müssen es ermöglichen, die existierenden Zeit- und Finanzvorgaben zu erfüllen. Hier handelt es sich, wie in Kapitel 2.1.2 bereits dargestellt, um einen für Evaluationen typischen Aspekt mit besonderem Gewicht. Die Auswahl geeigneter Methoden bewegt sich folglich im Spannungsfeld zwischen Zeit- und Aufwandsanforderungen und dem für die Beantwortung der Forschungsfrage erforderlichen Detailgrad der Datenerfassung.

- Die Orientierung an den Ressourcen und Bedürfnissen der Stakeholder ist eine wichtige Einflussgröße zur Bestimmung der Angemessenheit. Verfahren, die mit der prinzipiellen Arbeitsweise der zu befragenden Personen vereinbar sind und auf ihre Kompetenzen aufbauen, helfen, die Akzeptanz der Datenerhebung zu steigern und erhöhen dadurch den Rücklauf inhaltlich wertvoller Antworten[106].

- Die Bereitstellung von Freiheitsgraden für die Antworten der Befragten ist aus Sicht der Planenden nicht von prinzipieller Wichtigkeit, sondern jeweils im Kontext der Fragestellung und der Kompetenzen der Datengeber/-innen zu bewerten.

- Es ist erforderlich, eine ausreichende Tiefe der erhobenen Daten sicherzustellen, die es erlaubt, fundierte Schlüsse zu ziehen. Deshalb werden beispielsweise Daten

[106] Als Beispiele können hier die Begründungen für den Einsatz des Verfahrens der Online-Datenerhebung und der Interviews mit Lehrenden herangezogen werden.

aus Dokumenten- und Kennzahlenanalysen als durch zusätzlich zu erhebende Daten erläuterungsbedürftig eingeschätzt.

- Die Datenerhebungsmethoden müssen dazu geeignet sein, eine ausreichend breite Datenbasis im Sinne eines hohen Rücklaufs sicherstellen zu können. Gleichzeitig sollen sie dazu beitragen, die Datenqualität (im Sinne ernsthaft gegebener und nicht unüberlegter/vorschneller Antworten) zu gewährleisten.

- Um die Zufriedenheit der Stakeholder zu befördern, sollten die eingesetzten Methoden den Evaluationsgegenstand in ausreichender Breite detailliert erfassen können. Zusätzlich sollten die erhobenen Daten zu sicht- und/oder spürbaren Veränderungen des Gegenstandes führen.

Subsumierend kann festgehalten werden, dass die Sicht der Planenden auf die Gegenstandsangemessenheit von Datenerhebungsmethoden deutlich durch den Rahmen der Evaluation bestimmt wird. Die beiden Planenden sind theoretisch wie praktisch gleichermaßen sehr evaluations- und methodenerfahren. Vor dem Hintergrund dieser Erfahrungen konnten sie entsprechend fundierte methodische Entscheidungen treffen und deren Begründungen anführen. Für die Frage nach der Angemessenheit von Datenerhebungsmethoden ergeben sich aus ihren Äußerungen Kriterien, die den vorher aus den Interviews mit den Beteiligten erarbeiteten sehr ähnlich oder sogar mit ihnen identisch sind. Der wesentliche Unterschied liegt in der Einschätzung der Wichtigkeit der zur Beantwortung zur Verfügung gestellten Freiheitsgrade: Während für die (potenziellen) Evaluationsteilnehmer/-innen gerade dieser Aspekt von besonderer Wichtigkeit ist, hat er für die Planenden keinen besonderen Stellenwert und wird von ihnen in eine Reihe mit den übrigen Aspekten gestellt.

Die Überlegungen, aus denen heraus sich diese Sichtweise der Planenden auf die Datenerhebungsmethoden und ihre Angemessenheit ergibt, sind grundlegend andere als bei den (potenziellen) Evaluationsteilnehmer/-innen und sie sind mitunter auch anders konnotiert. Somit entsteht durch die Sichtweise der Planenden ein zusätzliches Begründungsmuster für die Berücksichtigung der bisher erarbeiteten zentralen Kriterien.

4.2.3. Gruppendiskussion

An der Gruppendiskussion nahmen eine beteiligte studierende, eine betroffene studierende sowie eine betroffene lehrende Person teil. Um die Erkenntnisse aus den Einzelinterviews zu komplettieren, zu komplementieren oder zu untermauern, wurden für die Auswertung der Gruppendiskussion Codierungen nur zu dem Zweck eingesetzt, die Themen des Gesprächs zu identifizieren, gezielt aufrufen und somit zu den Inhalten der Einzelinterviews in Bezug setzen zu können. Die inhaltliche Auswertung erfolgte klassisch intrepretativ.

Werden die Interaktionsprozesse innerhalb der Gruppe in den Blick genommen, so ist festzustellen, dass der Gesamtverlauf der Diskussion durch ein generelles Wohlwollen der befragten Personen füreinander geprägt war. Besondere Vorkommnisse oder relevante Verhaltensmuster jenseits der vermutlich für die Person charakteristischen[107] waren nicht zu bemerken. So gab es keine übermäßige Ehrfurcht, keine sichtbar gewordenen Ressentiments und kein prinzipielles Infragestellen der Sichtweisen der Angehörigen der übrigen Gruppen. Die Diskutanten/-innen hörten einander vielmehr aufmerksam zu und befassten sich ernsthaft mit den wechselseitig vorgebrachten Argumenten, prüften sie und versuchten, gemeinsam zu Urteilen zu kommen. Mit Blick auf das Erkenntnisinteresse, das mit der Gruppendiskussion verbunden war, lässt sich dabei festhalten, dass die Sichtweisen der Angehörigen anderer Stakeholdergruppen grundsätzlich ernst genommen wurden. Hat sich der Nutzen eines Vorschlags den übrigen Personen nicht direkt erschlossen, so wurde ggf. mehrfach nachgefragt, um die mit dem Vorschlag verbundenen Intentionen zu verstehen (bspw. Gruppendiskussion[108], 29 – 34). Es ist festzuhalten, dass die eingebrachten Vorschläge ernsthaft auf den mit ihnen verbundenen Mehrwert hin untersucht und, wenn dieser nicht entdeckt werden konnte, auch kritisch betrachtet wurden (bspw. GD, 36 – 51). Die Stakeholder waren demnach nicht primär

[107] Eine der Personen, B27, war etwas zurückhaltender, während die beiden anderen Teilnehmer/-innen einen höheren Redeanteil hatten. Die Beiträge von B27 wurden allerdings von den Beteiligten nicht anders behandelt als die übrigen.

[108] Nachfolgend werden Referenzen auf das Transkript der Gruppendiskussion mit dem Kürzel *GD* bezeichnet.

daran interessiert, alle im Gespräch vorgeschlagenen Methoden zu integrieren, etwa um Aushandlungskonflikten aus dem Weg zu gehen, sondern wollten tatsächlich den mit einer Methode verbundenen Nutzen erkennen können. Einige der Ideen, welche die anwesenden Personen zunächst nicht überzeugen konnten, wurden beispielsweise nicht einfach abgelehnt, sondern gemeinsam weiterentwickelt, um sie aus einer gemeinsamen Perspektive heraus gewinnbringend einsetzen zu können (etwa GD, 59 – 76). Aspekte, die sich als nicht konsensfähig erwiesen[109], wurden im Verlauf des Gesprächs immer wieder aufgegriffen, um zu versuchen, eine gemeinsame Beurteilung zu erreichen[110]. Die große Bereitschaft zur Konsensfindung, zur kritischen und gleichzeitig wohlwollend-gründlichen Auseinandersetzung mit den vorgebrachten Äußerungen war hinsichtlich der Interaktionsprozesse in der Gruppe auffällig.

Eine weitere Frage, zu deren Beantwortung die Gruppendiskussion dienen sollte, war jene nach den durch die Beteiligten vorgebrachten Argumenten und Begründungen für oder gegen den Einsatz bestimmter Methoden. Die Aussagen der teilnehmenden Personen konnten den folgenden fünf Gruppen zugeordnet werden:

- Evaluationszweck/Forschungsfrage: Dieser Gruppe wurden Argumente zugeordnet, die als Kriterium zur Einschätzung des Nutzens einer Datenerhebungsmethode auf die grundlegende Fragestellung und damit die Zielsetzung der Evaluation insgesamt zurückgreifen und diese als Entscheidungskriterium heranziehen (siehe etwa GD, 47 – 51).

- Zeiteffizienz: Argumente, die das Verhältnis zwischen der Zeit, die für den Einsatz einer Methode erforderlich ist, und den zu erwartenden Ergebnissen bzw. deren Wert und Beitrag zur Erfüllung des Evaluationszwecks aufgreifen (etwa GD, 279 ff.; GD, 289).

[109] Ein Beispiel für einen nicht direkt akzeptierten und wiederholt diskutierten Vorschlag ist die Frage, ob Professoren/-innen in die Datenerhebung einzubeziehen wären. Es konnte keine Einigkeit über die möglichen hilfreichen Beiträge dieser Personengruppe zum Ziel der Evaluation, also der Bewertung der Fachdidaktik-Ausbildung, erreicht werden.

[110] Eine tatsächliche Einigung der Befragten wurde durch den Moderator nicht forciert, wäre aber vermutlich möglich gewesen.

- Methodenimmanente Eigenschaften: In dieser Gruppe wurden Argumente zusammengefasst, die sich auf die einer Methode eigenen Charakteristika und die sich daraus ergebenden (vermuteten) Auswirkungen auf die Daten/die Gesamtevaluation beziehen (siehe beispielsweise GD, 320 – 328; GD, 292 – 297). Hierzu gehört etwa der Bezug auf Aspekte wie Offenheit oder Grad der Standardisierung einer Methode mit ihren jeweils spezifischen Vor- und Nachteilen.

- Zu befragende Personengruppen: Argumente dieser Gruppe nehmen die datengebenden Personen und ihre Eigenschaften in den Blick. Hierzu gehören Argumente, die sich auf die Anzahl der zu befragenden Personen (etwa GD, 292 – –297) oder die Homo- bzw. Heterogenität einer Probandengruppe (GD, 291) stützen, ebenso wie Aussagen, die den Bedarf einer gezielten Unterstützung der Probanden, so etwa die Erfordernis zur Aktivierung von Wissensbeständen durch die Datenerhebungsmethode (GD, 299 – 312), erkennen lassen.

- Verfügbarkeit von Bezugsgrößen, die die Daten erklären: Diese Gruppe schließlich umfasst Argumente, in denen hervorgehoben wird, dass die erhobenen Daten eines verfügbaren Kontextes bedürfen, vor dem sie eingeordnet und auf die Güte der Abbildung des Evaluationsgegenstandes hin überprüft werden können (beispielsweise GD, 356 – 402; GD, 482 – 498).

Wird der Blick auf die inhaltliche Ebene der Äußerungen gerichtet, lässt sich erkennen, dass sich die Befragten hinsichtlich der einzusetzenden Datenerhebungsmethoden nach kurzem Gespräch darauf einigen konnten, dass als Haupterhebungsinstrument teilstandardisierte Fragebögen einzusetzen wären, die gegebenenfalls, etwa bei Auffälligkeiten in den erhobenen Daten, durch Interviews ergänzt werden könnten. Einige, v. a. kleinere Stakeholdergruppen könnten alternativ auch gänzlich durch Interviews erfasst werden. Schüler sollten in Form von Gruppendiskussionen einbezogen werden, um in den Gesprächen ihr Potenzial nutzbar zu machen und die Reflexion über das vergleichsweise abstrakte Thema *Fachdidaktik* zu fördern. Dokumenten- und Kennzahlenanalysen schließlich sollten helfen, die tatsächliche Studienrealität mit den zu erfüllenden Vorgaben abzugleichen. Insgesamt sind die Argumente und Begründungen,

die in der Gruppendiskussion für oder gegen den Einsatz von bestimmten Datenerhebungsmethoden angeführt wurden, in der Summe mit jenen, die in den Einzelinterviews geäußert wurden, identisch.

Hinsichtlich der Zielsetzung der Gruppendiskussion hat sich vor allem die Frage nach den zu berücksichtigenden Stakeholdergruppen als relevant herausgestellt. Hierüber wurde intensiver diskutiert als über die Datenerhebungsmethoden. Bei den Gesprächen über die zu berücksichtigenden Stakeholdergruppen wurde ersichtlich, dass sich der Wunsch nach einer umfassenden Betrachtung des Evaluationsgegenstandes auch in der Gruppendiskussion wiederfinden lässt. Die vorgeschlagenen Stakeholdergruppen (Studierende in verschiedenen Ausbildungsstadien, Schüler/-innen, Absolventen/-innen, Berufseinsteiger/-innen, Studienseminarleiter/-innen und Lehrende an der Universtität, in dieser Eigenschaft auch Professoren/-innen) erstrecken sich über die Kerngruppen hinaus auch auf Personenkreise, die Auskunft über das tatsächlich erworbene Wissen und seine Anwendung geben könnten. Demgegenüber wurden Personen, die nach Meinung der hier Befragten einen starken Einfluss auf den Aufbau und die Organisation des fachdidaktischen Curriculums ausüben, nicht eindeutig als relevant für die Evaluation identifiziert. Bezüglich der Berücksichtigung einiger Personengruppen wurde keine Einigung erreicht. So blieb die Beteiligung der Gruppen der Lehrbeauftragten, der Studierenden, die noch nach einer alten, nicht modularisierten Studienordnung studieren, sowie der Professoren/-innen in ihrer Eigenschaft als die Fachdidaktik-Ausbildung planende Personen (in Abgrenzung zu *lehrenden* Professoren/-innen) ungeklärt. Während bei den Gruppen der Lehrbeauftragten und der Professoren/-innen sowohl unterstützende als auch hemmende Argumente gefunden wurden und sich die Befragten letztlich vor allem nicht ob der Wichtigkeit dieser Argumente einig wurden, stellte sich die Situation bezüglich der Gruppe der nach alter Studienordnung Studierenden etwas anders dar: Diese Gruppe wurde vor allem von den beiden Studierenden vorgeschlagen. Der dahinter liegende Wunsch war gemäß ihrer Äußerungen, den durch eine Änderung des Studiengangs erwirkten Mehrwert, den beide empfanden, sicht- und messbar zu machen. Allerdings wurde hier auch das Gegenargument, dass dieses Ansinnen jenseits der zu besprechenden Zielsetzung der fiktiven Evaluation läge, anerkannt.

Durch diesen Verlauf des Gesprächs wird erneut ersichtlich, dass rollenspezifische Sichtweisen der Beteiligten eine wichtige Entscheidungsgröße sind, jedoch nicht prinzipiell egoistisch konnotiert sind, sondern durchaus in einen Kontext gestellt werden und durch Argumente beeinflussbar sind. Die inhaltlichen Unterschiede, die sich erkennen lassen – so etwa die bereits erwähnte Unentschiedenheit der Teilnehmer/-innen, ob denn nun Professorinnen und Professoren in die Datenerhebung einzubeziehen seien – sind einerseits im Rahmen der Schwankungen, die über den gesamten hier besprochenen Datensatz hinweg auftreten und andererseits sicherlich auch durch die gemäß der Zielsetzung der Gruppendiskussion zurückhaltende und wenig eingreifende und keine Ergebnisse forcierende Moderation bedingt. Entscheidend für die Akzeptanz eines Vorschlages scheint vor allem die Plausibilität, die der vorgeschlagenen Methode und der mit ihr verbundenen Argumentation von den übrigen Beteiligten beigemessen wird.

Ein Aspekt, der über die bisher vorgenommene inhaltlichen Auswertung der Gruppendiskussion hinaus von Interesse war, ist, dass sich die beteiligten Personen im Anschluss an das Gespräch ungefragt als positiv überrascht von der Gruppendiskussion und ihrem Verlauf zeigten. Im Anschluss an die Gruppendiskussion äußerten sie einvernehmlich, dass sie sich inhaltlich bereichert fühlten und ihnen die spezifische Sichtweise der Angehörigen anderer Gruppen sehr viel klarer geworden sei. Sie hätten mehr über Praxis und Hintergründe des Studiums erfahren, was eine Bereicherung darstelle. Die Beteiligten empfanden also die eigentlich als ermittelndes Instrument angelegte Gruppendiskussion als ein Verfahren mit zumindest teilweise vermittelndem Charakter – ganz so, wie das Verfahren in den oben dargestellten Einzelinterviews konnotiert wurde.

Werden diese Erkenntnisse aus der Analyse der Gruppendiskussion auf die Frage der Konstruktion von Gegenstandsangemessenheit bezogen, so ergeben sich zwei zentrale Erkenntnisse:

1. Sollen Methoden, aber auch Stakeholdergruppen von Stakeholdern als relevant anerkannt werden, so bedarf es nachvollziehbarer und fundierter Begründungen. Die Stakeholder sollten in die Lage versetzt werden, die mit den Methoden bzw.

mit den zu berücksichtigenden Stakeholdern verbundenen Intentionen zu erkennen und bei Bedarf auch zu hinterfragen.

2. Die Tatsache, dass in der Gruppendiskussion prinzipiell die gleichen Argumente und Aussagen in Erscheinung treten wie bereits in den Einzelinterviews, bestätigt die Relevanz der bisher identifizierten Aspekte (Bestätigung durch Perspektivvariation). Gleichzeitig wird die Annahme unterstützt, dass die angeführten Argumente tatsächlich für diesen Kontext relevant sind.

Selbstverständlich sind diese Gruppendiskussion und ihre Ergebnisse nicht ohne Weiteres auf andere Bereiche zu übertragen. Zu speziell und mit Blick auf vorhandene Kompetenzen selektiv mögen hier die Rahmenbedingungen sein, zu ausgeprägt die Bereitschaft der Teilnehmer/-innen zur konstruktiven Mitwirkung. Unter anderen Bedingungen, in denen beispielsweise Machtgefälle oder wechselseitige Abhängigkeiten der Teilnehmer/-innen untereinander eine Rolle spielen, in denen emotionale Beteiligung ausgeprägter ist, in denen von einem für die einzelne Person möglicherweise ungünstigen Evaluationsergebnis handfeste negative Folgen ausgehen könnten oder in denen eine prinzipiell negative Stimmung zwischen den Beteiligten herrscht, kann die Bereitschaft, die Argumente anderer Personen aufzunehmen und ernsthaft zu überdenken, deutlich geringer sein. Dennoch wird erneut ersichtlich, dass begründete Argumentationen fruchten und helfen können, das Verständnis der Stakeholder für die Methoden einer Evaluation zu befördern und damit letztlich die Akzeptanz ihrer Ergebnisse zu steigern.

4.3. Zusammenfassung: Nach welchen Kriterien bewerten Stakeholder Gegenstandsangemessenheit?

Ziel der beschriebenen empirischen Untersuchung war es, zu analysieren, nach welchen Kriterien die Stakeholder einer Evaluation die Angemessenheit von Datenerhebungsmethoden bewerten. Zudem sollte betrachtet werden, welche Methoden aus welchen Gründen bevorzugt werden, ob Bezüge zwischen der Position einer Person gegenüber

dem Gegenstand der Untersuchung und ihrer Vorstellung bezüglich der Angemessenheit eingesetzter Methoden existieren und welche Folgerungen sich aus all diesen Erkenntnissen für die Arbeit von Personen, die eine Evaluation planen, ergeben können. Nachfolgend werden die ersten drei Aspekte aufgegriffen, die vierte Frage nach den Folgerungen für die Arbeit planender Personen wird schwerpunktmäßig im nächsten Kapitel im Rahmen eines Kriterienkatalogs beantwortet.

Werden die Ergebnisse der einzelnen vorangehenden Abschnitte zusammenfassend betrachtet, um die erste Frage zu beantworten, so lassen sich folgende Kriterien identifizieren, nach denen Stakeholder die Angemessenheit von Datenerhebungsmethoden bewerten:

Passung zu den Evaluationszwecken/der Fragestellung: Ein zentrales Kriterium zur Bewertung der Gegenstandsangemessenheit stellt die Zielsetzung bzw. die Fragestellung der Evaluation dar. Einzusetzende Datenerhebungsmethoden sollten in jedem Falle Daten erheben, die zur Erreichung der intendierten Ziele und Zwecke beitragen. Wichtig ist, welchen konkreten Beitrag eine gewählte Methode dazu leisten kann. Diese Informationen sowie die Zielsetzung der Evaluation sollten gleichzeitig transparent gehandhabt werden, um es den Stakeholdern zu ermöglichen, sich mit den gewählten Methoden auseinanderzusetzen.

Erfassung der Sichtweisen datengebender Personen: Die individuellen Sichtweisen der Datengeber/-innen einer Evaluation werden von den hier Befragten als eine wichtige Grundlage für die Bewertung des Evaluationsgegenstandes betrachtet. Der Fokus liegt dabei auf der Innensicht der zu befragenden Personen als Experten für den jeweiligen Evaluationsgegenstand. Um diese Sichtweisen möglichst unverfälscht zu erfassen, sollten für die Datenerhebung Verfahren eingesetzt werden, die eine individuelle Berücksichtigung der jeweils befragten Person ermöglichen, also ausreichend Freiheitsgrade für die Beantwortung zur Verfügung stellen. Hervorzuheben ist dabei, dass dieser Bewertungsaspekt nicht per se den Einsatz möglichst offener Datenerhebungsmethoden impliziert. Vielmehr zielt er darauf ab, dass den Datengebern/-innen ermöglicht werden sollte, ihre individu-

elle Sichtweise in den erhobenen Daten tatsächlich abbilden zu können, was aus Sicht der hier Befragten prinzipiell auch mit (teil-) standardisierten Methoden erreicht werden kann.

Komplexität des Evaluationsgegenstandes: Die durch die Stakeholder wahrgenommene Komplexität eines Evaluationsgegenstandes wirkt sich auf die Präferenzen für Methoden aus: Stakeholder, denen der Evaluationsgegenstand komplexer erschien, neigten eher dazu, offene Methoden als angemessen zu betrachten als jene, die ihn als weniger komplex wahrgenommen haben[111]. Mit Zunahme der wahrgenommenen Komplexität geht die Annahme einher, dass nicht alle für seine Bewertung relevanten Aspekte eines Gegenstandes durch die planenden Personen antizipiert und passend operationalisiert werden können.

Effizienz: Stakeholder bewerten Datenerhebungsmethoden nach ihrer Effizienz. Dabei steht die Frage im Zentrum, wie viel Arbeitsaufwand der Einsatz einer Methode erfordert. In der Tendenz wird Verfahren der Vorzug gegeben, die den Erhebungsprozess strukturieren und damit zielführend leiten. Dabei geht es bei diesem Aspekt nicht ausschließlich darum, möglichst wenig Arbeit und/oder Zeit in die Untersuchung investieren zu müssen. Vielmehr werden gleichzeitig der Nutzen einer Methode bzw. der mit ihr erhobenen Daten und die für die Stakeholder verbleibenden Freiheitsgrade im Blick behalten. Dieser Aspekt lässt sich als Wunsch nach einer ausbalancierten Mischung zwischen zielführender und zeitsparender Strukturiertheit auf der einen und ausreichend Freiheitsgraden für die Datengeber/-innen auf der anderen Seite paraphrasieren. Aus Sicht der Personen, die die Evaluation planen und für ihre Durchführung verantwortlich sind, kommt ein weiterer Aspekt hinzu: Die einzusetzenden Methoden dürfen den für die Durchführung der Untersuchung zur Verfügung stehenden Zeit-, Ressourcen- und Finanzrahmen nicht sprengen.

[111] Hierin liegt ein Unterschied zur Sichtweise der Planenden, aus deren Sicht die Komplexität des Evaluationsgegenstandes ein Kriterium ist, das gegenüber anderen – etwa der Zeiteffizienz – eher zurücktritt.

Verwendungsorientierung: Den befragten Personen ist es wichtig, dass die erhobenen Daten auch tatsächlich Verwendung finden. Der zentrale Wunsch ist dabei, dass sich aus der Evaluation tatsächlich Konsequenzen ergeben, die aus den erhobenen Daten ableitbar sind. Für die Wahl von Datenerhebungsmethoden bedeutet dies, dass die gewählten Methoden möglichst Daten hervorbringen sollen, die unmittelbar zur Erfüllung der Ziele einer Evaluation beitragen[112].

Genauigkeit und Richtigkeit der gewonnenen Daten: Dieses Kriterium nimmt die Qualität der erhobenen Daten in den Blick. Die Befragten erwarten, dass die erhobenen Daten die Sicht der Datengeber/-innen und damit letztlich den Evaluationsgegenstand exakt und korrekt abbilden. Das Bild des Gegenstandes, das auf Grundlage dieser Daten erzeugt wird, sollte damit sowohl detailliert als auch nachprüfbar sein. Dabei zielt dieses Kriterium nicht auf ein prinzipielles *so genau wie möglich*, sondern die Befragten behalten durchaus die eigentliche Fragestellung im Blick. So lautet die Forderung eher *so richtig wie möglich und so genau, wie für die Erfüllung der Evaluationszwecke erforderlich.*

Fundiertheit der Daten: Über die Genauigkeit und Richtigkeit hinaus sollten die Daten, die durch die ausgewählten Erhebungsmethoden gewonnen werden, ausreichend fundiert sein, um gültige Aussagen über den Evaluationsgegenstand zu ermöglichen. Dies bezieht sich auf die Tiefe der Daten, d. h. auf den Detailgrad der verfügbaren Informationen, dieses Kriterium verweist jedoch auch auf die Breite der Datenbasis, die so beschaffen sein sollte, dass die erhobenen Daten ein facettenreiches Bild des Evaluationsgegenstandes zeichnen können und nicht nur wenige, schlechtestenfalls willkürlich gewählte Perspektiven berücksichtigt werden.

Ressourcen und Bedürfnisse der Stakeholder: Die einzusetzenden Datenerhebungsmethoden sollten sich an den Ressourcen und Bedürfnissen der zu befragenden Stakeholder orientieren. In einer Evaluation können Stakeholdergruppen mit sehr

[112] Dieser Aspekt ist unmittelbar mit der *Passung zu den Evaluationszwecken/der Fragestellung* verbunden.

unterschiedlichen Eigenschaften – beispielsweise hinsichtlich ihrer Kenntnisse, des Abstraktionsniveaus dieser Kenntnisse, ihrer Sprachkompetenz, des Alters etc. – befragt werden. Datenerhebungsmethoden sollten diesen Besonderheiten Rechnung tragen, sie berücksichtigen und derart gestaltet sein, dass den Anforderungen, die sich aus diesen Eigenschaften ergeben, entsprochen werden und so eine hochwertige Datenerhebung durchgeführt werden kann.

Transparenz und fairer Umgang: Dieses Kriterium bezieht sich auf den Umgang mit den datengebenden Personen. Es wird erwartet, dass ausreichende und leicht zugängliche Informationen über Ziele, Methoden und Ergebnisse der Evaluation verfügbar sind und dass die Probanden möglichst aktiv über diese Aspekte informiert werden. Darüber hinaus wird erwartet, dass die Mitwirkung an einer Evaluation anonym erfolgen kann.

Über diese sich unmittelbar auf die Beurteilung der Angemessenheit von Datenerhebungsmethoden beziehenden Kriterien hinaus hat die Analyse drei weitere Aspekte hervorgebracht, auf die sich Stakeholder bei ihrer Urteilsfindung stützen und die teilweise auch auf die verbleibenden beiden Fragen wirken:

Naheliegende Stakeholder sind wichtige Stakeholder: Aus Perspektive der Befragten besitzt die Sicht von eingebundenen Personen, die sich in direkter Nähe zum Evaluationsgegenstand befinden, einen hohen Stellenwert für seine Beurteilung. Im Bereich der Lehrevaluation an Hochschulen handelt es sich dabei um die Personen, die unmittelbar am Lehr-/Lernprozess beteiligt sind. Das unbegründete Ignorieren einer solchen dem Evaluationsgegenstand nahen Gruppe könnte Zweifel an der Güte der Evaluation hervorrufen.

Stakeholder wünschen Argumente und sind ihnen auch prinzipiell zugänglich: Die (naiven) Vorstellungen von Stakeholdern einer Evaluation bezüglich der Angemessenheit von Datenerhebungsmethoden sind durch stichhaltige Argumentationen veränderbar. So ist beispielsweise – zumindest innerhalb der hier betrachteten Probandengruppe – eine anfängliche Skepsis gegenüber einer bestimmten

Datenerhebungsmethode nicht mit einer kategorischen Ablehnung gleichzusetzen, sondern kann durch den Austausch fundierter Argumente oder die Bereitstellung von Hintergrundinformationen in Akzeptanz verwandelt werden. Dies gilt in gleicher Weise auch für die Diskussion der Auswahl von Personen(gruppen), die im Rahmen der Evaluation als Datengeber/-innen fungieren. Hieraus lässt sich ableiten, dass fundierte Begründungen bezüglich der Methodik und des Samplings einer Evaluation ihre Akzeptanz steigern können. Dabei wird die Bereitschaft zur Auseinandersetzung mit diesen Themen von der Stärke des Eigeninteresses der Stakeholder abhängen: Je stärker dieses Eigeninteresse von Stakeholdern an einer Evaluation bzw. ihrem Gegenstand ist, desto eher werden sie bereit sein, sich mit methodischen Fragen der Evaluation auseinanderzusetzen, desto höher wird jedoch auch gleichzeitig die Erwartung an argumentative Begründungen durch die Planenden sein.

Vorwissen und Vorerfahrung der Stakeholder beeinflussen ihr Urteil: Das methodische Vorwissen einer Person übt Einfluss auf die Wahrnehmung und Einschätzung von Erhebungsverfahren aus. Verfahren, über die eine Person fundiertes Wissen besitzt, werden differenzierter und präziser betrachtet und bewertet als solche, über die sie nur Vermutungen anstellen oder zu deren Beurteilung sie nur auf Alltagswissen und -vorstellungen zurückgreifen kann – ein weiteres Argument für die fundierte Information der Stakeholder über die eingesetzten Datenerhebungsmethoden.

Diese Kriterien sind in die Planung einer Evaluation mit einzubeziehen, um die Sichtweise der Stakeholder bei der Planung einer Evaluation zu berücksichtigen. Dabei ist darauf hinzuweisen, dass die hier erarbeiteten Anforderungen Maximalanforderungen darstellen: Es wird, wie es auch für die *Standards für Evaluation* (siehe Kapitel 2.1.4) gilt, nicht immer möglich sein, alle Kriterien bei der Auswahl einzusetzender Datenerhebungsmethoden gleichermaßen zu berücksichtigen. So kann beispielsweise die Forderung nach Effizienz dem Anspruch der Erfassung der individuellen Sichtweisen datengebender Personen oder auch jenem nach Fundiertheit der Daten entgegenstehen. Wie auch

bei den *Standards für Evaluation* wird man die einzelnen Kriterien gegeneinander abwägen müssen.

Bezüglich der zweiten Frage, welche Methoden Stakeholder aus welchen Gründen bevorzugen, ist festzuhalten, dass prinzipielle, nicht im Persönlichen begründete Präferenzen für bestimmte Methoden nur sehr begrenzt festgestellt werden konnten. Festzustellen ist jedoch, dass die Affinität zu offenen Methoden mit der wahrgenommenen Komplexität des Evaluationsgegenstandes bzw., anders formuliert, mit zunehmender Bewusstheit der Begrenzung der eigenen Sichtweise auf den Gegenstand zunimmt. Zudem lässt sich eine klare generelle Tendenz zur Verbindung und Integration unterschiedlicher Erhebungsverfahren, zu Mixed Methods, erkennen. Dass einige Methoden von den Befragten häufiger vorgeschlagen wurden als andere, hängt vermutlich eher mit der Bekanntheit der Methoden und dem mit der Untersuchung erfassten Erfahrungsschatz hinsichtlich der Durchführung von Datenerhebungen und weniger mit in die Tiefe gehenden methodologischen Überlegungen zusammen: Methoden, die im Kontext von Datenerhebungen insgesamt seltener oder weniger prominent eingesetzt werden, so beispielsweise die Dokumenten- und Kennzahlenanalyse oder längerfristige Verfahren, sind in den frei formulierten Vorschlägen ebenfalls weniger präsent.

Die dritte Frage, ob sich die Position einer Person gegenüber dem Evaluationsgegenstand auf ihre Vorstellung von der Angemessenheit der Datenerhebungsmethoden auswirkt, kann dahingehend beantwortet werden, dass dies auf indirektem Wege geschieht. Die Position einer Person lässt hier keinen Einfluss auf die eigentlichen, persönlichen Maßstäbe zur Bestimmung der Angemessenheit erkennen – zumindest lässt sich Entsprechendes für die Varianzen, die in den Daten auftreten, nicht gesichert nachweisen. Vielmehr dürften diese Varianzen auf persönliche Erfahrungen mit Evaluationen und/oder Datenerhebungen, die Genauigkeit der Kenntnis der Materie oder Ähnliches zurückzuführen sein. Aber genau auf dieser Ebene beeinflusst die Position einer Person gegenüber dem Evaluationsgegenstand indirekt ihre Vorstellung der Angemessenheit:

Oftmals ist die spezifische Position einer Person gegenüber dem Evaluationsgegenstand mit bestimmten Kompetenzen, Erfahrungen oder Eignungen verbunden, die wiederum Einfluss auf das hier relevante Wissen sowohl über den Gegenstand selbst als auch über Datenerhebungsmethoden ausüben. Im vorliegenden Fall etwa zeichnen sich die Studierenden in der Regel durch einen anderen Kenntnisstand, etwa in Bezug auf die organisatorischen Aspekte des Studiums, aus als die Lehrenden. Diese für den Gruppendurchschnitt durchaus charakteristische Wissensbasis führt wiederum, wie dargestellt, im Vergleich zu den Lehrenden zu unterschiedlichen Bewertungen und Einschätzungen einzelner Methoden und ihrer Eignung. Diese Bewertungen und Einschätzungen sind nicht unumstößlich, sondern lassen sich durch gezielte Information ausgleichen, da sich die eigentlichen Bewertungs*maßstäbe* zwischen den Gruppen nicht prinzipiell unterscheiden.

Die hier herausgearbeiteten Kriterien gestatten es, einen Einblick in die Sichtweise der befragten Personen auf die Angemessenheit von Datenerhebungsmethoden im Kontext von Evaluationen zu erhalten. Um sie jedoch zu einer verlässlichen Grundlage für die Arbeit von Personen, die Evaluationen planen, werden zu lassen, müssen diese Sichtweisen in einen erweiterten theoretischen Kontext gestellt werden, der auch die methodologische Sicht auf die Frage der Gegenstandsangemessenheit, die in Kapitel 3 herausgearbeitet wurde, integriert. Im nachfolgenden Kapitel werden die in diesem Kapitel erarbeiteten Sichtweisen der Stakeholder den zuvor identifizierten methodologischen Anforderungen gegenübergestellt und zu einer praktisch anwendbaren Kriterienliste verdichtet.

5. Die Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden

5.1. Ergebnisse im Kurzüberblick

Zur Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden im Kontext von Lehrevaluationen an Hochschulen lassen sich zwei wesentliche Kriterienquellen heranziehen, fachliche Standards auf der einen Seite und die Sichtweise der Stakeholder einer Evaluation auf der anderen. Innerhalb dieser zwei Quellen lässt sich jeweils eine weitere Unterteilung in bedeutsame Untergruppen vornehmen. Bei den fachlichen Standards ist zwischen den allgemein für empirische Sozialforschung geltenden Standards, die einen grundsätzlichen Rahmen für die Wahl von Datenerhebungsmethoden definieren, und fachspezifischen Standards wie etwa den *Standards für Evaluation* der DeGEval, die in Deutschland eine wichtige Rolle spielen, zu unterscheiden. Innerhalb der Gruppe der Stakeholder ist zwischen den Ansprüchen der Gruppe der Beteiligten und Betroffenen auf der einen Seite und denen der Evaluationsplanenden auf der anderen Seite zu unterscheiden.

Die Anforderungen an die Angemessenheit von Datenerhebungsmethoden, die aus jeder dieser vier einzelnen Kriterienquellen abgeleitet wurden, stehen sich nicht entgegen, sondern lassen sich aufeinander beziehen und ergänzen sich. Die Grundlage stellen dabei die allgemeinen Ansprüche empirischer Sozialforschung an die Gegenstandsangemessenheit von Datenerhebungsmethoden dar, die im Rahmen dieser Arbeit aus der allgemeinen methodologischen Literatur zur empirischen Sozialforschung erarbeitet wurden (siehe Kap. 3.2.1). Die Analyse dieser Literatur hat gezeigt, dass das Konstrukt der Gegenstandsangemessenheit als Indikator zur Wahl konkreter Datenerhebungsmethoden allgemein akzeptiert ist und beinahe immer benannt wird, auf konkreter Ebene

aber mit einigen wenigen Ausnahmen unbestimmt und sehr vage bleibt. Um dieses Konstrukt greifbarer zu machen, wurden die verstreut und in unterschiedlichsten Kontexten – z. B. in Beschreibungen einzelner Methoden, als Hinweise in Werken zur Qualität von Forschung oder in Übersichtswerken – genannten Einzelaspekte zusammengetragen, sortiert, abstrahiert und integriert. Auf diese Weise entstand aus den Fragmenten die in Kapitel 3.2.1 vorgestellte Zusammenstellung allgemeiner Ansprüche empirischer Sozialforschung an die Gegenstandsangemessenheit von Datenerhebungsmethoden. Die in dieser Form vorliegenden Kriterien können nicht als ein abgeschlossenes Konzept oder gar als Handlungsanweisung verstanden werden, sondern stehen zunächst vielmehr ungewichtet nebeneinander. Es handelt es sich um die grundlegende Sammlung von Kriterien, die nach aktuellem Stand der Methodologie zur Betrachtung der Frage nach der Gegenstandsangemessenheit von Datenerhebungsmethoden herangezogen werden können. Demgegenüber sind die Anforderungen, die sich aus den drei übrigen Kriterienquellen ergeben, deutlich spezifischer auf den hier bearbeiteten Bereich der Evaluationsforschung ausgerichtet. Sie sind sehr viel konkreter und setzen individuelle Schwerpunkte, haben teilweise normativen Charakter, aber lassen sich dennoch sämtlich auf die auf S. 63 ff. vorgestellten allgemeinen Kriterien empirischer Sozialforschung zur Bestimmung der Gegenstandsangemessenheit zurückführen und sind somit theoretisch fundiert.

Die *Standards für Evaluation* der DeGEval, die zweite als bedeutsam identifizierte Quelle, baut in jenen Teilen, in denen sie Hinweise zur Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden enthält, auf die identifizierten Ansprüche der allgemeinen empirischen Sozialforschung auf. So lassen sich Bezüge zu allen acht ihrer Einflussgrößen nachweisen. Allerdings stehen die einzelnen Aspekte hier nicht mehr nur mehr oder weniger gleichgeordnet nebeneinander, sondern es werden für das Feld der Evaluationsforschung charakteristische Schwerpunktsetzungen ersichtlich. Im Zentrum steht dabei die Nutzbarkeit der durch die Evaluation erarbeiteten Ergebnisse. Dieser übergeordneten Anforderung müssen demnach auch die zum Einsatz kommenden Datenerhebungsmethoden Rechnung tragen. Ein zweiter wesentlicher Aspekt innerhalb

der in den *Standards* vorgenommenen Gewichtung ist die Betonung der Bedeutung der Stakeholder einer Evaluation, und zwar sowohl der Gruppe der untersuchten Subjekte als auch der Zielgruppe der Untersuchung. Es werden also Interessensgruppen als relevant definiert, deren Position in einem klassischen Prozess empirischer Forschung oftmals eher schwach ist oder denen ggf. jenseits ihrer Eigenschaft als Datengeber/-innen oder Ergebnisempfänger/-innen gar keine besondere Rolle zugedacht wird. Ein dritter Schwerpunkt wird in den *Standards* auf die Einhaltung der gesetzten Rahmenbedingungen gelegt. Dabei wird einem allgemeinen Charakteristikum angewandter Sozialforschung Rechnung getragen, die – anders als klassische Grundlagenforschung – als gescheitert zu betrachten ist, wenn sie innerhalb der gesetzten (zeitlichen, finanziellen, personellen etc.) Rahmenbedingungen nicht zu einem verwertbaren Ergebnis kommt. Insgesamt wird in den Standards eine Konstruktion von Gegenstandsangemessenheit vorgenommen, die maßgeblich auf die Verwendbarkeit der Ergebnisse, auf die Berücksichtigung der Ansprüche und Vorstellungen relevanter Stakeholdergruppen sowie auf Methodeneffizienz ausgerichtet ist.

Auch die Kriterien, die von der dritten Quelle, den Beteiligten und Betroffenen einer Evaluation, ausgehen, zeichnen sich durch eine spezifische Schwerpunktsetzung aus. So ist dieser Personengruppe, wie in Kapitel 4.2.1 herausgearbeitet wurde, wichtig, dass die Datenerhebung gewährleistet, dass der Evaluationsgegenstand mit seinen Besonderheiten idealerweise multiperspektivisch erfasst wird, dass sich die Methoden den Evaluationsteilnehmer/-innen und ihren Anforderungen flexibel anpassen lassen und dass ihnen die Möglichkeit geboten wird, ihre individuelle Sicht auf den Evaluationsgegenstand in die Datenbasis der Evaluation einzubringen. Gleichzeitig soll die Datenerhebung möglichst effizient sein und Ergebnisse hervorbringen, die dazu angetan sind, tatsächlich im Sinne der Evaluationszwecke genutzt zu werden sowie idealerweise einen Nutzen für die Personen dieser Gruppe hervorbringen. Im Zentrum steht hier also die Forderung, dass die Datenerhebungsmethoden präzise, umfassend, offen, zielführend und auf die Verwendung der Daten ausgerichtet sein sollen. Die Balance zwischen Detail- und Überblick, Effizienz und Nutzbarkeit sind die entscheidenden Momente.

Eine Lösung, wie diesen keineswegs trivialen Anforderungen entsprochen werden könnte, wird von den Angehörigen dieser Stakeholdergruppe mehrheitlich im Einsatz von Mixed Methods gesehen: Die Kombination unterschiedlicher Erhebungsverfahren, auch und gerade über die Grenzen der Paradigmen *quantitativ* und *qualitativ* hinweg, verspricht aus der Perspektive der Befragten, diesen breiten Anforderungen gerecht werden zu können und wird somit aus ihrem Blickwinkel zum Baustein der Sicherstellung der Gegenstandsangemessenheit. Jenseits dieser gruppenspezifischen Schwerpunkte lassen sich die Forderungen, die von dieser Stakeholdergruppe ausgehen, mit den Kriterien, die von den beiden anderen oben aufgeführten Quellen ausgehen, zur Deckung bringen.

Die Gewichtung innerhalb der Kriterien zur Bestimmung der Gegenstandsangemessenheit, die schießlich von den Evaluationsplanenden, der vierten und letzten relevanten Kriterienquelle, ausgeht, führt erneut in eine andere Richtung, denn für sie stehen die Effizienz und die Gewährleistung der Durchführbarkeit einer Evaluation im Zentrum. Angemessene Methoden dienen für sie der Erfüllung der Evaluationszwecke, ohne die zur Verfügung stehenden Ressourcen überzubeanspruchen. Zwar ist auch die Akzeptenz der Vorgehensweise durch die anderen Stakeholder für diese Gruppe wichtig, allerdings wird der diesbezügliche Handlungsspielraum als durch den jeweils zur Verfügung stehenden Rahmen begrenzt eingeschätzt. In der Folge wird auch der Einsatz von Mixed Methods pragmatisch und ressourcenorientiert betrachtet. Insgesamt liegen die Kriterien zur Bewertung der Angemessenheit, die von der Gruppe der Evaluationsplanenden genannt werden, sehr nah an den Ansprüchen der allgemeinen empirischen Sozialforschung und der *Standards für Evaluation*, wenngleich auch nicht alle Aspekte von ihnen aufgegriffen wurden.

Werden die in dieser Arbeit betrachteten Quellen zur Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden im Vergleich betrachtet, so zeigen sich unterschiedliche Schwerpunktsetzungen, die sich in Form jeweils spezifischer Gewichtungen der grundlegenden Kriterien äußern. Diese Kriterien und ihre Gewichtungen werden in Bezug auf die dieser Arbeit zugrunde liegenden Forschungsfrage, wie sich

im Rahmen von Lehrevaluation an Hochschulen die Gegenstandsangemessenheit von Datenerhebungsmethoden bestimmen lässt, nachfolgend zu einem kompakten und für die Evaluationspraxis nutzbaren Kriterienkatalog zusammengeführt.

5.2. Kriterienkatalog zur Bestimmung der Gegenstandsangemessenheit von Datenerhebungsmethoden

5.2.1. Zielsetzung und Hintergründe

Der vorliegende Kriterienkatalog stellt eine praxis- und anwendungsorientierte Zusammenstellung der Erkenntnisse dieser Arbeit dar. Er fokussiert ausschließlich die Frage der Gegenstandsangemessenheit von Datenerhebungsmethoden und ist damit als eine sich auf einen bestimmten Teilbereich der insgesamt bei einer Evaluation zu bearbeitenden Schritte konzentrierende Ergänzung bzw. als Erweiterung zu Kriteriensammlungen wie den *Standards für Evaluation* der DeGEval (DeGEval Gesellschaft für Evaluation 2008) konzipiert.

Ziel des Kriterienkatalogs ist es, Evaluationsplanende bei der Identifikation von Datenerhebungsmethoden zu unterstützen, die möglichst gegenstandsangemessen sind. *Gegenstandsangemessenheit* wird dabei verstanden als die Eigenschaft von Datenerhebungsmethoden, sämtliche für das Evaluationsvorhaben relevanten Daten und Informationen zu erfassen und dabei die Besonderheiten von Evaluationsgegenstand, Fragestellung und Kontext (im Sinne weiterer beeinflussender Faktoren) zu berücksichtigen. Auf diese Weise soll eine möglichst tragfähige Datengrundlage für die Evaluation geschaffen werden, wobei Evaluation im Sinne der DeGEval (DeGEval Gesellschaft für Evaluation 2008, S. 15) als „systematische Untersuchung des Nutzens oder Wertes eines Gegenstandes“ begriffen wird.

Die nachfolgend aufgeführten Kriterien wurden im Bereich der Lehrevaluation an Hochschulen entwickelt. Wie an verschiedenen Stellen dieser Arbeit dargestellt wurde, handelt es sich bei diesem Feld um eines, das sich durch bestimmte Besonderheiten etwa hinsichtlich seiner formalen Struktur auszeichnet, die in dieser Form sicher nicht

in jedem anderen möglichen Evaluationsfeld anzutreffen sind. Um den Katalog auch für Bereiche jenseits seines direkten Entstehungskontextes nutzbar zu machen, wurden die einzelnen Aspekte abstrahiert, so dass sie möglichst auch auf andere Evaluationsfelder übertragen werden können. Dennoch ist es möglich, dass Faktoren, die in einem bestimmten Evaluationsrahmen von besonderer Bedeutung sind, im vorliegenden Kriterienkatalog aufgrund seiner konzeptionellen Offenheit nicht oder nur unzureichend abgebildet sind. Für solche spezifische Kontexte kann demnach eine Erweiterung des Katalogs erforderlich sein. Zudem gilt, dass die einzelnen Fragen nicht per se gleich gewichtet sind, sondern dass je nach Art des Evaluationsvorhabens einzelne Kriterien stärkeren Einfluss ausüben können als andere, so dass eine kontextbezogene, aber stets dokumentierte Gewichtung durch die Evaluationsplanenden unerlässlich ist. In diesem Sinne ist der Katalog nicht als fixes normatives Konstrukt zu sehen.

5.2.2. Verwendung des Kriterienkatalogs

Bevor mit der Bearbeitung des Kriterienkatalogs begonnen werden kann, ist eine Grundauswahl von Datenerhebungsmethoden zu treffen, die für die Evaluation in Betracht kommen können. Für jede dieser Datenerhebungsmethoden ist zu bestimmen, durch welche Charakteristika sie sich auszeichnet. Dazu sind die Methoden anhand der ab S. 78 detailliert vorgestellten *Kriterien zur Klassifikation von Datenerhebungsverfahren* (Art der Datenerfassung, Grad der Strukturierung, Art der Kommunikationsvermittlung, Reaktivität, Größe der erfassbaren Stichprobe, Beteiligung von Personen, Zeitbedarf, Nähe zum Gegenstand) zu analysieren; auf S. 88 ff. ist eine Analyse der für Datenerhebungen in Evaluationen zentralen Methoden *Quantitatives Interview/Fragebogen*, *Offenes Interview*, *Dokumenten- und Kennzahlenanalyse*, *Gruppendiskussion* und *Gutachterverfahren* anhand dieser Kriterien zu finden. Grundsätzlich ist zu berücksichtigen, dass unterschiedliche Methoden oft auch sinnvoll in Kombination eingesetzt werden können.

Der Kriterienkatalog selbst ist in drei zentrale Bereiche unterteilt, nämlich in 1. Kriterien zu Evaluationsgegenstand und Evaluationszweck/Fragestellung, 2. Kriterien zu

den Stakeholdern sowie 3. Kriterien zum Rahmen der Evaluation. Nach der Analyse der Charakteristika der fraglichen Datenerhebungsmethoden ist im nächsten Schritt zu überprüfen, ob sie den Kriterien der drei Gruppen des Katalogs entsprechen. Dabei sollten alle Aspekte des Kriterienkatalogs abgearbeitet werden. Die verantwortlichen Personen sollten zu jedem einzelnen Punkt Stellung beziehen, offene Fragen notieren, Probleme und Lösungsvorschläge aufzeichnen sowie Gründe für das besondere Gewicht eines Aspektes im spezifischen Kontext oder auch für den Umstand, dass ein Punkt des Kriterienkatalogs in der speziellen Situation nicht angewendet werden kann, festhalten. Auf diese Weise entsteht eine Sammlung, die einen wesentlichen Baustein zur Erreichung von Transparenz darstellt und gleichzeitig eine wichtige Grundlage fundierter Dokumentation ist – nicht zuletzt gegenüber den Stakeholdern. Hervorzuheben ist, dass dieser Prozess der Kriterienanalyse und -gewichtung für jede Evaluation individuell vorzunehmen ist.

Der vorliegende Kriterienkatalog deckt, da er sich ausschließlich mit der Frage der Gegenstandsangemessenheit befasst, nur einen Teilbereich des gesamten Evaluationsprozesses ab und sollte deshalb nicht allein zum Einsatz kommen. Vielmehr ist zusätzlich sicherzustellen, dass die Anlage der gesamten Evaluation den akzeptierten fachlichen Standards, etwa den *Standards für Evaluation* der DeGEval (DeGEval Gesellschaft für Evaluation 2008) sowie den Ansprüchen empirischer Sozialforschung und ihren spezifischen Qualitäts- und Gütekriterien entspricht. Hilfreich kann es zudem sein, zu überprüfen, ob bereits Evaluationsmodelle beschrieben wurden, die einen geeigneten Rahmen für die vorgesehene Untersuchung bieten (siehe auch Kap. 2.1.3). Daran anschließend kann anhand der ab S. 78 aufgeführten Kriterien untersucht werden, welche Datenerhebungsmethoden für das geplante Vorhaben geeignet sind.

Der Kriterienkatalog selbst kann zu verschiedenen Zeitpunkten und ggf. auch wiederholt zur Anwendung gebracht werden. So bietet es sich an, die Kriterien zu Beginn, während der Festlegung der empirischen Vorgehensweise, zu bearbeiten. Ebenso können sie während des Designs eingesetzt werden, um wesentliche Entscheidungen zu treffen oder zu fundieren, oder sie können im Anschluss an die Designphase zur Überprüfung

ihrer Ergebnisse verwendet werden. Anhand des Kriterienkatalogs kann aber auch die Methodik bereits abgeschlossener Evaluationen hinsichtlich ihrer Angemessenheit bewertet werden. Insgesamt dient die Bearbeitung des Katalogs der Sicherstellung der Gegenstandsangemessenheit der zum Einsatz kommenden Datenerhebungsmethoden, also ihrer Passung zu Forschungsfrage/Evaluationszweck, Untersuchungsgegenstand und, falls existent, weiteren beeinflussenden Faktoren.

5.2.3. Kriterienkatalog

Kriterien zu Evaluationsgegenstand und Evaluationszweck/Fragestellung

- **Evaluationsgegenstand**

  EF-1 Die gewählten Datenerhebungsmethoden passen zu den Anforderungen des Evaluationsgegenstandes, eventuelle Probleme und mögliche Lösungen für diese werden beschrieben.

  EF-2 Die Wahl der Datenerhebungsmethoden trägt der Komplexität der Evaluation Rechnung: Das Design der Datenerhebung stellt sicher, dass das durch die erhobenen Daten gezeichnete Bild des Evaluationsgegenstandes detailgetreu, nachvollziehbar, überprüfbar, multiperspektivisch und facettenreich ist.

  EF-3 Die Anlage der Datenerhebung ist unparteiisch und dazu geeignet, sowohl Stärken als auch Schwächen des Evaluationsgegenstandes abzubilden.

  EF-4 Bildet eine Kombination von Methoden (-familien) den Evaluationsgegenstand besser ab, als der Einsatz einer einzigen Methode, so wird diese Kombination eingesetzt.

  EF-5 Die Datenerhebungsmethoden belasten den Evaluationsgegenstand möglichst wenig.

- **Evaluationszweck/Fragestellung**

  EF-6 Die einzusetzenden Verfahren sind hinsichtlich der Zielsetzung der Untersuchung zielführend.

EF-7 Die Datenerhebungsmethoden sind effizient.

EF-8 Die erhobenen Daten sind ausreichend genau, um auf ihrer Grundlage die Evaluationszwecke zu erreichen. Sie bringen Daten von ausreichender Breite und Tiefe hervor, um Bewertungen fundiert vornehmen bzw. die Ausgangsfrage fundiert beantworten zu können.

EF-9 Die erhobenen Daten sind verwendungsorientiert, d. h. sie bzw. die Form, in der sie vorliegen, befördern ihre Verwendung.

Kriterien zu den Stakeholdern

- **Auswahl der Stakeholder**

S-1 Die (aus der Eigenperspektive der Stakeholder) *naheliegenden* und die (formal) *eingebundenen* Stakeholder werden in die Datenerhebung einbezogen.

- **Passung zu Ressourcen und Anforderungen**

S-2 Die individuellen Sichtweisen der Datengeber/-innen sind in ausreichender Weise erfasst. Die Evaluationsteilnehmer/-innen können ihre individuelle Perspektive und unterschiedlichste Positionen angemessen darstellen. Gegebenenfalls werden hierzu Methodenkombinationen eingesetzt.

S-3 Die gewählten Datenerhebungsmethoden passen zu den Arbeitsweisen und Gewohnheiten der Evaluationsteilnehmer/-innen, sie tragen ihren Besonderheiten, Bedürfnissen, Ansprüchen, zur Verfügung stehenden Ressourcen und Eigenschaften Rechnung und tragen so dazu bei, einen möglichst hohen Rücklauf zu erzeugen.

S-4 Der Aufwand, der allen Beteiligten durch die Datenerhebung entsteht, steht in einem angemessenen Verhältnis zum Nutzen der Evaluation, die Datenerhebungsmethoden belasten Beteiligte und Betroffene möglichst wenig.

S-5 Die Vorteile, welche die gewählten Datenerhebungsmethoden für die einzelnen Stakeholdergruppen mit sich bringen, überwiegen die Nachteile.

S-6 Für jede Stakeholdergruppe wurde eine adäquate Datenerhebungsmethode oder eine Kombination unterschiedlicher Methoden (-familien) gefunden, die richtige (i. S. v. *die Sichtweise der Datengeber/-innen exakt und korrekt abbildende*) Daten hervorbringt.

S-7 Die Datenerhebung erfolgt anonym.

- **Information**

S-8 Es wird für die Stakeholder nachvollziehbar begründet, mit welcher Intention und mit welchen Zielen die gewählten Methoden eingesetzt werden.

S-9 Den Stakeholdern sind alle Informationen, die sie benötigen, um sich ein fundiertes Bild von den eingesetzten Methoden, aber auch der Gesamtevaluation zu machen, bekannt und leicht zugänglich.

- **Nutzen für die Stakeholder**

S-10 Aus den erhobenen Daten können sicht- und/oder spürbare Veränderungen für die Stakeholder hervorgehen.

S-11 Die erhobenen Daten ermöglichen den Teilnehmer/-innen der Evaluation, aus ihrer Beteiligung einen evaluationsbezogenen und/oder persönlichen Nutzen zu ziehen.

Kriterien zum Rahmen der Evaluation

R-1 Es wird sichergestellt, dass die gewählten Methoden den gesetzten Ressourcen-, Zeit- und Finanzrahmen der Evaluation einhalten.

R-2 Alle die Datenerhebung durchführenden und an ihr beteiligten Personen sind methodisch kompetent und beherrschen die einzusetzenden Verfahren sicher.

5.3. Ausblick

Das Konzept der Gegenstandsangemessenheit von Datenerhebungsmethoden ist in der methodologischen Diskussion tief verankert. Eine anwendungsorientierte Ausarbeitung

dieses Konzeptes fehlte allerdings bislang. In der vorliegenden Arbeit wurde der Versuch unternommen, diese Lücke zu bearbeiten. Durch die Konkretisierung der methodologischen Anforderungen ermöglicht der erarbeitete Kriterienkatalog die Erhöhung der Transparenz einer Untersuchung, die ein Gütekritierum empirischer Forschung darstellt. Zudem soll er zur Anregung der Diskussion über und Auseinandersetzung mit der Frage der Gegenstandsangemessenheit in der Praxis dienen. In diesem Sinne kann er als ein Beitrag zur Verbesserung von Evaluationen bzw. ihrer Datenerhebungen, aber auch allgemein von empirischen Datenerhebungen gesehen werden.

Die Suche nach konkretisierten Kriterien kann nicht maximale Vereinfachung zum Ziel haben. Da es sich um ein nicht triviales, sondern sehr vielschichtiges Themengebiet handelt, ist Komplexitätsreduktion nur in sehr begrenztem Rahmen möglich und sinnvoll. Konkretisierte Formulierungen können jedoch, wie in der vorliegenden Arbeit gezeigt, durchaus auch erreicht werden, ohne die aus methodologischer Sicht so wichtigen Freiheiten der Methodenwahl von vornherein einzuengen. Das Ziel eines solchen Kataloges sollte demnach eine Strukturierung des Entscheidungsprozesses sein – ein Ansinnen, das mit dem Anspruch der Wissenschaft, regelgeleitet und systematisch vorzugehen, konform ist, ja eigentlich essenziell dazugehört. Strukturierte Entscheidungsprozesse erleichtern die korrekte Umsetzung und können durch Nachvollziehbarkeit, Transparenz und Fundiertheit einen Beitrag zur Verbesserung von Entscheidungen und somit zur Steigerung der Qualität empirischer Untersuchungen leisten.

Die Auswahl von Datenerhebungsmethoden ist dabei lediglich eine jener Entscheidungen, die Planende und Durchführende von empirischen Untersuchungen im Laufe des Forschungsprozesses treffen müssen und die sich auf den Verlauf der Untersuchung, ihre Ergebnisse sowie die allgemeine Güte auswirken. Eine zweite sehr zentrale Frage ist jene nach der Auswahl geeigneter Auswertungsverfahren, für deren Auswahl aus methodologischer Perspektive ebenfalls das (auch hier vage) Primat der Angemessenheit anzusetzen ist. Insgesamt ist die Situation in diesem Bereich sehr ähnlich: Es existiert ein äußerst breites Spektrum unterschiedlichster Auswertungsverfahren, die jeweils über eigene Entstehungszusammenhänge und Zielsetzungen verfügen, sich teil-

weise entgegenstehen, manchmal aber auch ergänzen. Und auch hier lassen sich durchaus Quellen finden, aus denen Kriterien zur Bestimmung der Angemessenheit abgeleitet werden können. So wäre es auch für den Bereich der Auswertungsverfahren möglich, wünschenswert und hilfreich, einen Katalog von Kriterien zu erarbeiten, der bei der Auswahl angemessener Auswertungsverfahren unterstützt und so zur Erhöhung von Transparenz und somit auch von Qualität beiträgt.

Die begründete Auswahl von Datenerhebungsmethoden ist auch im meist sehr eng bemessenen Zeitrahmen von Evaluationen wichtig, handelt es sich doch um einen Kernaspekt empirischer Forschung. In dieser Arbeit konnte gezeigt werden, dass Stakeholder eigene Vorstellungen von gelungenen Datenerhebungen haben, die aus methodologischer Sicht durchaus Berechtigung besitzen. So ist festzuhalten, dass Personen, die Evaluationen planen und durchführen, Forschungsdesigns nicht ausschließlich auf der Basis des Evaluationsgegenstandes, methodischer Überlegungen und formaler Restriktionen entwerfen sollten. Vielmehr sind über diese Aspekte hinaus die Stakeholder und ihre Vorstellungen zu berücksichtigen. Erst so kann sichergestellt werden, dass eine Evaluation und ihre Ergebnisse nicht nur den relevanten formalen Kriterien entsprechen, sondern auch die Genauigkeit besitzen und die Akzeptanz erfahren können, die sie benötigen, um die intendierten Ziele zu erreichen und Wirkungen auch tatsächlich entfalten zu können.

Literaturverzeichnis

Abel, Jürgen/Möller, Renate/Treumann, Klaus Peter (1998): Einführung in die Empirische Pädagogik. Stuttgart: Kohlhammer.

Aghamanoukjan, Anahid/Buber, Renate/Meyer, Michael (2007): Qualitative Interviews. In **Buber, Renate/Holzmüller, Hartmut M. (Hrsg.) (2007):** Qualitative Marktforschung. Konzepte – Methoden – Analysen. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 415 – 435.

Alkin, Marvin (2004): Evaluation Roots. Thousand Oaks: Sage.

Altrichter, Herbert/Brüsemeister, Thomas/Wissinger, Jochen (Hrsg.) (2007): Educational Governance. Handlungskoordination und Steuerung im Bildungssystem. Wiesbaden: VS Verlag für Sozialwissenschaften.

American Evaluation Association (2007): Guiding Principles for Evaluators. In American Journal of Evaluation, 28, Nr. 4, S. 397 – 398.

Atria, Moria/Strohmeier, Dagmar/Spiel, Christiane (2006): Der Einsatz von Vignetten in der Programmevaluation – Beispiele aus dem Anwendungsfeld „Gewalt in der Schule“. In **Flick, Uwe (2006c):** Qualitative Evaluationsforschung. Konzepte - Methoden - Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 233 – 249.

Atteslander, Peter (2008): Methoden der empirischen Sozialforschung. 12. Auflage. Berlin: Erich Schmidt Verlag.

Atteslander, Peter/Kopp, Manfred (1993): Befragung. In **Roth, Erwin (1993):** Sozialwissenschaftliche Methoden. München u. a.: Oldenbourg, S. 144 – 172.

Balk, Michael (2000): Evaluation von Lehrveranstaltungen. Die Wirkung von Evaluationsrückmeldung. Frankfurt am Main: Peter Lang.

Balzer, Lars (2005): Wie werden Evaluationsprojekte erfolgreich? Landau: Verlag Empirische Pädagogik.

Balzer, Lars/Frey, Andreas/Nenniger, Peter (1999): Was ist und wie funktioniert Evaluation. In Empirische Pädagogik, 4, Nr. 13, S. 393 – 413 ⟨URL: `http://www.lars-balzer.info/publications/pub-balzer_1999-02_PUB_EP1999-13(4)_balzer-frey-nenniger.pdf`⟩.

Barlösius, Eva (2006): Wissenschaft evaluiert – praktische Beobachtungen und theoretische Betrachtungen. In **Flick, Uwe (Hrsg.):** Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 385–404.

Barz, Andreas (1998): Evaluation im deutschen Hochschulsystem: Ziele, Instrumente, Erfahrungen, Trends. Köln ⟨URL: `http://www.degeval.de/koeln1998/barz.htm`⟩ – Zugriff am 23.10.2009.

Becker-Richter, Marion et al. (2002): Evaluation von Studium und Lehre. Verfahren – Methoden – Erfahrungen. Opladen: Leske + Budrich.

Beywl, Wolfgang: Eval-Wiki: Glossar der Evaluation. Köln ⟨URL: `http://www.eval-wiki.org/glossar/`⟩ – Zugriff am 17.08.2011.

Beywl, Wolfgang (1988): Zur Weiterentwicklung der Evaluationsmethodologie. Grundlegung, Konzeption und Anwendung eines Modells der responsiven Evaluation. Frankfurt am Main: Peter Lang.

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

Beywl, Wolfgang et al. (2007): Evaluation Schritt für Schritt: Planung von Evaluationen. Darmstadt: Hiba-Verlag.

Beywl, Wolfgang/Schepp-Winter, Ellen (2000): Zielgeführte Evaluation von Programmen. Ein Leitfaden. Berlin: Bundesministerium für Familie, Senioren, Frauen und Jugend.

Blalock, Hubert M./Blalock, Ann B. (1968): Methodology in Social Research. New York u. a.: McGraw-Hill Book Company.

Blank, Renate (2007): Gruppendiskussionsverfahren. In **Naderer, Gabriele/ Balzer, Eva (Hrsg.) (2007):** Qualitative Marktforschung in Theorie und Praxis. Grundlagen, Methoden und Anwendungen. Wiesbaden: Gabler Verlag, S. 280 – 301.

Bock, Karin/Miethe, Ingrid (Hrsg.) (2010): Handbuch Qualitative Methoden in der Sozialen Arbeit. Opladen & Farmington Hills: Verlag Barbara Budrich.

Bogumil, Jörg (1998): Kommunale Verwaltungsreform. In **Andersen, Uwe (Hrsg.):** Kommunalpolitik in Nordrhein-Westfalen im Umbruch. Köln, Stuttgart, Berlin: Kohlhammer, S. 82 – 108.

Bogumil, Jörg/Grohs, Stephan (2009): Von Äpfeln, Birnen und Neuer Steuerung. In **Bogumil, Jörg/Heinze, Rolf G. (Hrsg.) (2009):** Neue Steuerung von Hochschulen. Eine Zwischenbilanz. Berlin: Edition Sigma, S. 139 – 149.

Bogumil, Jörg et al. (2007a): Zehn Jahre Neues Steuerungsmodell. Berlin: Edition Sigma.

Bogumil, Jörg/Heinze, Rolf G. (2009): Einleitung. In **Bogumil, Jörg/Heinze, Rolf G. (Hrsg.) (2009):** Neue Steuerung von Hochschulen. Eine Zwischenbilanz. Berlin: Edition Sigma, S. 7 – 12.

Bogumil, Jörg/Heinze, Rolf G. (Hrsg.) (2009): Neue Steuerung von Hochschulen. Eine Zwischenbilanz. Berlin: Edition Sigma.

Bogumil, Jörg et al. (2007b): Perspektiven kommunaler Verwaltungsmodernisierung. Praxiskonsequenzen aus dem Neuen Steuerungsmodell. Berlin: Edition Sigma.

Bohnsack, Ralf/Marotzki, Winfried/Meuser, Michael (2011): Hauptbegriffe Qualitativer Sozialforschung. 3. Auflage. Opladen & Farmington Hills: Verlag Barbara Budrich.

Bohnsack, Ralf/Przyborski, Aglaja (2007): Gruppendiskussionsverfahren und Focus Groups. In **Buber, Renate/Holzmüller, Hartmut M. (Hrsg.) (2007):** Qualitative Marktforschung. Konzepte – Methoden – Analysen. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 491 – 506.

Bornmann, Lutz Bornmann/Daniel, Hans-Dieter (2003): Begutachtung durch Fachkollegen in der Wissenschaft. Stand der Forschung zur Reliabilitat, Fairness und Validitat des Peer-Review-Verfahrens. In **Schwarz, Stefanie/Teichler, Ulrich (Hrsg.) (2003):** Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung. Frankfurt am Main: Campus, S. 207 – 225.

Bortz, Jürgen/Döring, Nicola (2005): Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler. 3. Auflage. Heidelberg: Springer.

Boruch, Robert (2007): Encouraging the flight of error: Ethical standards, evidence standards, and randomized trials. In New Directions for Evaluation, 2007, Nr. 113, S. 55 – 73.

Böttcher, Wolfgang/Holtappels, Heinz Günter/Brohm, Michaela (Hrsg.) (2006): Evaluation im Bildungswesen. Eine Einführung in Grundlagen und Praxisbeispiele. Weinheim und München: Juventa.

Brüsemeister, Thomas (2008): Qualitative Forschung. Ein Überblick. 2. Auflage. Wiesbaden: Westdeutscher Verlag.

Buber, Renate/Holzmüller, Hartmut M. (Hrsg.) (2007): Qualitative Marktforschung. Konzepte – Methoden – Analysen. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bülow-Schramm, Margret (1995): Wer hat Angst vor den Evaluatoren? Der Umgang mit Akzeptanzproblemen von Evaluationsverfahren. In **Verbeek, David/ Balogh, Heike (Hrsg.) (1995):** Evaluation der Lehre. Ziele – Akzeptanz – Methoden. Band 1, Handbuch Hochschullehre Highlights. Stuttgart u.a.: Raabe, S. 1 – 20.

Chase, Susan E. (2005): Narrative Inquiry: Multiple Lenses, Approaches, Voices. In **Denzin, Norman K./Lincoln, Yvonna S. (Hrsg.) (2005):** The Sage handbook of qualitative Research. 3. Auflage. Thousand Oaks: Sage, S. 651 – 679.

Chelimsky, Eleanor (1999): Evaluation for the 21st century. A handbook. Thousand Oaks: Sage.

Cook, Thomas D/Leviton, Laura C/Shadish, William R (2000): Foundations of Program Evaluation. Theory of Practice. Newbury Park: Sage.

Creswell, John (2009): Research design. Qualitative, quantitative, and mixed methods approaches. Thousand Oaks: Sage.

Cropley, Arthur J. (2005): Qualitative Forschungsmethoden. Eine praxisnahe Einführung. 2. Auflage. Eschborn: Verlag Dietman Klotz.

DeGEval Gesellschaft für Evaluation (2008): Standards für Evaluation. 4. Auflage. Alfter: DeGEval Gesellschaft für Evaluation ⟨URL: `http://www.degeval.de/calimero/tools/proxy.php?id=19076`⟩.

DeGEval Gesellschaft für Evaluation (2011): Einladungs-E-Mail zur 14. DeGEval-Jahrestagung 2011 in Linz, Thema: „Partizipation – dabei sein ist alles!?“. No address in ⟨URL: `http://us2.campaign-archive1.com/?u=3b8f5b3690f13ad672ebc8fc6&id=a11911a2bb&e=3bd964bebd`⟩ – Zugriff am 12.05.2011.

DeGEval Gesellschaft für Evaluation (2012): Partizipation in der Evaluation. Positionspapier der DeGEval – Gesellschaft für Evaluation. No address in ⟨URL: `http://www.degeval.de/images/stories/Publikationen/Positionspapier_Partizipation_in_der_Evaluation.pdf`⟩ – Zugriff am 30.01.2012.

Denzin, Norman K./Lincoln, Yvonna S. (Hrsg.) (2005): The Sage handbook of qualitative Research. 3. Auflage. Thousand Oaks: Sage.

Denzin, Norman K./Lincoln, Yvonna S. (Hrsg.) (2008): Collecting and Interpreting Qualitative Materials. 3. Auflage. Thousand Oaks: Sage.

Diekmann, Andreas (2007): Empirische Sozialforschung. 18. Auflage. Reinbek bei Hamburg: Rowohlt.

Dudenredaktion (1989): Das Herkunftswörterbuch. 2. Auflage. Mannheim: Dudenverlag.

Dudenredaktion (2005): Das Fremdwörterbuch. 8. Auflage. Mannheim: Dudenverlag.

Ebert, Thomas (2007): Qualitative Evaluation universitärer Online-Lehre. Diplomarbeit, Philipps-Universität Marburg.

Eichler, Dirk/Merkens, Hans (2006): Organisationsforschung mit qualitativen Methoden – Erfahrungen aus der Evaluation eines freien Jugendhilfeträgers. In **Flick, Uwe (Hrsg.):** Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 301–318.

Erdogan, Gülten (2001): Die Gruppendiskussion als qualitative Datenerhebung im Internet. Ein Online-Offline-Vergleich. In kommunikation@gesellschaft, Jg. 2, Beitrag 5.

Ernst, Stefanie (2006): Die Evaluation von Qualität – Möglichkeiten und Grenzen von Gruppendiskussionsverfahren. In **Flick, Uwe (Hrsg.):** Qualitative Evalua-

tionsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 183–213.

Ernst, Stefanie (2008): Manual Lehrevaluation. Wiesbaden: VS Verlag für Sozialwissenschaften.

Eubel, Klaus-Dieter/Brüsemeister, Thomas (2008): Evaluationsbasierte Steuerung, Wissen und Nichtwissen – Einführung in die Thematik. In **Eubel, Klaus-Dieter/Brüsemeister, Thomas (Hrsg.):** Evaluation, Wissen und Nichtwissen. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 7 – 15.

European Association for Quality Assurance in Higher Education (2005): Standards and Guidelines for Quality Assurance in the European Higher Education Area. 2005.

European Association for Quality Assurance in Higher Education (2006): Standards und Leitlinien für die Qualitätssicherung im Europäischen Hochschulraum. Band 9, Beiträge zur Hochschulpolitik. Bonn: Hochschulrektorenkonferenz.

Fetterman, David M. (2001): Foundations of empowerment evaluation. Thousand Oaks: Sage.

Flick, Uwe (1995): Qualitative Forschung. Theorie, Methoden, Anwendung in Psychologie und Sozialwissenschaften. Reinbek bei Hamburg: Rowohlt.

Flick, Uwe (2006a): Interviews in der qualitativen Evaluationsforschung. In **Flick, Uwe (2006c):** Qualitative Evaluationsforschung. Konzepte - Methoden - Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 214–232.

Flick, Uwe (2006b): Qualität in der qualitativen Evaluationsforschung. In **Flick, Uwe (2006c):** Qualitative Evaluationsforschung. Konzepte - Methoden - Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 424–443.

Flick, Uwe (2006c): Qualitative Evaluationsforschung. Konzepte - Methoden - Umsetzungen. Reinbek bei Hamburg: Rowohlt.

Flick, Uwe (2006d): Qualitative Evaluationsforschung zwischen Methodik und Pragmatik – Einleitung und Überblick. In **Flick, Uwe (2006c):** Qualitative Evaluationsforschung. Konzepte - Methoden - Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 9–29.

Flick, Uwe (2007): Qualitative Sozialforschung. Eine Einführung. Reinbek bei Hamburg: Rowohlt.

Flick, Uwe (2008): Triangulation. Eine Einführung. Wiesbaden: VS Verlag für Sozialwissenschaften.

Flick, Uwe (2009a): An Introduction to Qualitative Research. 4. Auflage. Thousand Oaks: Sage.

Flick, Uwe (2009b): Sozialforschung. Methoden und Anwendungen. Ein Überblick für die BA-Studiengänge. Reinbek bei Hamburg: Rowohlt.

Flick, Uwe/Kardorff, Ernst von/Steinke, Ines (2000): Qualitative Forschung. Ein Handbuch. Reinbek bei Hamburg: Rowohlt.

Fontana, Andrea/Frey, James H. (2005): The Interview: From Neutral Stance to Political Involvment. In **Denzin, Norman K./Lincoln, Yvonna S. (Hrsg.) (2005):** The Sage handbook of qualitative Research. 3. Auflage. Thousand Oaks: Sage, S. 695 – 727.

Frey, Bruno (2006): Evaluitis – Eine neue Krankheit. Univerität Zürich. ⟨URL: `http://homepage.univie.ac.at/Eveline.Christof/evaluation09/Evaluitis.pdf`⟩.

Friedrichs, Jürgen (1973): Methoden empirischer Sozialforschung. Reinbek bei Hamburg: Rowohlt.

Froschauer, Ulrike/Lueger, Manfred (2003): Das qualitative Interview. Zur Praxis interpretativer Analyse sozialer Systeme. Wien: WUV.

Froschauer, Ulrike/Lueger, Manfred (2006): Qualitative Prozessevaluierung in Unternehmen. In **Flick, Uwe (Hrsg.):** Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 319–338.

Früh, Werner (2007): Inhaltsanalyse. Theorie und Praxis. Konstanz: UVK.

Garz, Detlef/Kraimer, Klaus (Hrsg.) (1991): Qualitativ-Empirische Sozialforschung. Konzepte, Methoden, Analysen. Opladen: Westdeutscher Verlaghabe.

Gollwitzer, Mario/Jäger, Reinhold S. (2007): Evaluation. Weinheim: Beltz Psychologie Verlags Union.

Gruber, Hans/Mandl, Heinz/Renkl, Alexander (1999): Was lernen wir in Schule und Hochschule: träges Wissen? Institut für Pädagogische Psychologie und Empirische Pädagogik, Lehrstuhl Prof. Dr. Heinz Mandl, München.

Grüner, Herbert (1993): Evaluation und Evaluationsforschung im Bildungswesen. In Pädagogische Rundschau, 47. Jahrgang, Nr. 1/1993, S. 29 – 52.

Haase, Volker (2009): Ökonomische Transformationen. Soziale Arbeit im Kontext der Neuen Steuerung. Diplomarbeit, Philipps-Universität Marburg.

Habel, Edna (1995): Hochschulen zum Rapport??? Erfahrungen mit internen Lehrberichten und Lehrberichten nach Universitätsgesetz an der Universität Dortmund. In **Verbeek, David/Balogh, Heike (Hrsg.) (1995):** Evaluation der Lehre. Ziele – Akzeptanz – Methoden. Band 1, Handbuch Hochschullehre Highlights. Stuttgart u.a.: Raabe, S. 1 – 28.

Habermehl, Werner (1992): Angewandte Sozialforschung. München u. a.: Oldenbourg.

Häder, Michael (2010): Empirische Sozialforschung. Eine Einführung. Band 2., überarbeitete Auflage, Wiesbaden: VS Verlag für Sozialwissenschaften.

Hamburger, Franz (2010): Grundlagenforschung und Praxisforschung: Gegensatz oder unverzichtbares Wechselverhältnis? In Praxisforschung in der Kinder-und Jugendhilfe,, S. 71 – 78.

Helfferich, Cornelia (2009): Die Qualität qualitativer Daten. Manual für die Durchführung qualitativer Interviews. 3. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften.

Henninger, Michael/Balk, Michael (2001): Integrative Evaluation: Ein Ansatz zur Erhöhung der Akzeptanz von Lehrevaluation an Hochschulen. Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie der Ludwig-Maximilians-Universität München.

Hense, Jan (2006): Selbstevaluation. Erfolgsfaktoren und Wirkungen eines Ansatzes zur selbstbestimmten Qualitätsentwicklung im schulischen Bereich. Frankfurt am Main: Peter Lang.

Hense, Jan (2008): Online-Wörterbuch Evaluation. No address in ⟨URL: `http://www.evoluation.de/glossary`⟩ – Zugriff am 12.12.2008.

Hesse-Biber, Sharlene Nagy/Leavy, Patricia (2006): Emergent Methods in Social Research. Thousand Oaks: Sage.

Hirschauer, Stefan (2006): Wie geht Bewerten? – Zu einer anderen Evaluationsforschung. In **Flick, Uwe (Hrsg.):** Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 405–423.

Hochschulrektorenkonferenz: Qualitätsmanagement HRK. No address in ⟨URL: `http://www.hrk.de/de/projekte_und_initiativen/121_4074.php`⟩ – Zugriff am 10.07.2009.

Hochschulrektorenkonferenz (Hrsg.) (1998): Evaluation – Sachstandsbericht zur Qualitätsbewertung und Qualitätsentwicklung in deutschen Hochschulen. Bonn: Hochschulrektorenkonferenz.

Hochschulrektorenkonferenz (Hrsg.) (2001): Wettbewerb – Profilbildung – Evaluation. Qualitätssicherung von Lehre und Studium in Gegenwart und Zukunft. Band 6/2001, Bonn: Hochschulrektorenkonferenz.

Hochschulrektorenkonferenz (Hrsg.) (2003): Evaluation und ihre Konsequenzen. 3. Berliner Evaluationstagung an der Technischen Fachhochschule Berlin, 21./22. Februar 2002. Band 2/2003, Bonn: Hochschulrektorenkonferenz.

Hochschulrektorenkonferenz (Hrsg.) (2004): Evaluation – ein Bestandteil des Qualitätsmanagements an Hochschulen. Band 9/2004, Bonn: Hochschulrektorenkonferenz.

Hochschulrektorenkonferenz (Hrsg.) (2005): Hochschule entwickeln, Qualität managen: Studierende als (Mittel)punkt. Die Rolle der Studierenden im Prozess der Qualitätssicherung und -entwicklung. Band 10/2005, Bonn: Hochschulrektorenkonferenz ⟨URL: `http://www.hrk.de/de/download/dateien/2005-10_Hochschule_entwickeln_Qualitaet_managen.pdf`⟩.

Hochschulrektorenkonferenz (Hrsg.) (2006): Qualitätsentwicklung an Hochschulen. Erfahrungen und Lehren aus 10 Jahren Evaluation. Band 8/2006, Bonn: Hochschulrektorenkonferenz ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩.

Holla, Bernhard (2002): Qualitätsentwicklung in der Weiterbildung durch praxisorientierte Evaluation. Frankfurt am Main: Peter Lang.

Holstein, James A./Gubrium, Jaber F. (2004): The active interview. In **Silverman, David (Hrsg.) (2004):** Qualitative Research. Theory, Method and Practice. 2. Auflage. Thousand Oaks: Sage, S. 140 – 161.

Hopf, Christel/Schmidt, Christiane (1993): Zum Verhältnis von innerfamilialen sozialen Erfahrungen, Persönlichkeitsentwicklung und politischen Orientierungen. Dokumentation und Erörterung des methodischen Vorgehens in einer

Studie zu diesem Thema. Hildesheim ⟨URL: `http://w2.wa.uni-hannover.de/mes/berichte/rex93.htm`⟩ – Zugriff am 10.06.2009.

Hyman, Herbert H. et al. (2004): Interviewing in Social Research. In **Seale, Clive (Hrsg.) (2004):** Social Research Methods. A Reader. London: Routledge Students Readers, S. 88 – 95.

Julnes, George/Rog, Debra J/American Evaluation Association (2007): Informing Federal Policies on Evaluation Methodology: Building the Evidence Base for Method Choice in Government Sponsored Evaluations. San Francisco, Fairhaven: Jossey-Bass.

Karbach, Manfred (1998): Anmerkungen zum Wort Evaluation. Lünen ⟨URL: `http://evaluelab.de/extra/evalety.html`⟩ – Zugriff am 31.07.2008.

Kardorff, Ernst von (2006): Zur gesellschaftlichen Bedeutung und Entwicklung (qualitativer) Evaluationsforschung. In **Flick, Uwe (Hrsg.):** Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 63–91.

Kelle, Udo (2008): Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte. 2. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften.

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

Kelle, Udo/Tabor, Alexandra/Metje, Brigitte (2009): Qualitative Evaluationsforschung im Internet — Online-Foren als Werkzeuge interpretativer Sozialforschung. In **Jackob, Nikolaus/Schoen, Harald/Zerback, Thomas (Hrsg.):** Sozialforschung im Internet. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 181 – 195.

Kempfert, Guy/Rolff, Hans-Günter (2005): Qualität und Evaluation. Ein Leitfaden für Pädagogisches Qualitätsmangement. 4. Auflage. Weinheim und Basel: Beltz.

Kirkpatrick, Donald L. (Hrsg.) (1987): More evaluating training programs – A collection of articles from Training and Development Journal. Alexandria: American Society for Training and Development.

Klatt, Anja: Informationen zur Lehrveranstaltungsevaluation an der Philipps-Universität Marburg. No address in ⟨URL: `http://www.uni-marburg.de/studium/qs/lehrevaluation`⟩ – Zugriff am 29.04.2011.

Klusmeyer, Jens (2005): Beratung – Evaluation – Transfer. Oldenburg: BIS-Verlag.

Kommunale Gemeinschaftsstelle für Verwaltungsmanagement (KGSt) (Hrsg.) (1993): Das Neue Steuerungsmodell. Begründungen, Konturen, Umsetzung. Band 5/1993, Köln: Kommunale Gemeinschaftsstelle für Verwaltungsmanagement (KGSt) ⟨URL: `http://www.tvoed.info/kgst_nsm_1993.pdf`⟩.

Kommunale Gemeinschaftsstelle für Verwaltungsmanagement (KGSt) (Hrsg.) (1996): KGSt-Politikerhandbuch zur Verwaltungsreform. Köln: Kommunale Gemeinschaftsstelle für Verwaltungsmanagement (KGSt).

König, René (Hrsg.) (1974): Das Interview. Köln: Kiepenheuer und Witsch.

Kracht, Stefan (2006): Das Neue Steuerungsmodell im Hochschulbereich. Zielvereinbarungen im Spannungsverhältnis zwischen Konsens und hierarchischem Verwaltungsaufbau. Baden-Baden: Nomos.

Kromrey, Helmut (2003): Evaluation in Wissenschaft und Gesellschaft. In ZfEv, Zeitschrift für Evaluation, 1/2003, S. 114 – 145 ⟨URL: `www.hkromrey.de/Kromrey_EvalWissG.pdf`⟩.

Kromrey, Helmut (2005): Evaluation – Ein Überblick. In **Schöch, Heidrun (Hrsg.):** Was ist Qualität. Die Entzauberung eines Mythos. Berlin: Wissen-

schaftlicher Verlag, Schriftenreihe Wandel und Kontinuität in Organisationen 6, S. 31 – 85.

Kromrey, Helmut (2006a): Empirische Sozialforschung. 11. Auflage. Stuttgart: Lucius & Lucius.

Kromrey, Helmut (2006b): Qualität und Evaluation im System Hochschule. In **Stockmann, Reinhard (2006c):** Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder. 3. Auflage. Münster: Waxmann, S. 234 – 259.

Kromrey, Helmut (2006c): Studierendenbefragungen in Lehrveranstaltungen: Instrument der Evaluation oder „nur" der Qualitätsentwicklung? In **Hochschulrektorenkonferenz (Hrsg.) (2006):** Qualitätsentwicklung an Hochschulen. Erfahrungen und Lehren aus 10 Jahren Evaluation. Band 8/2006, Bonn: Hochschulrektorenkonferenz ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩, S. 72 – 71 ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩.

Kromrey, Helmut (2007): „Die empirische Erfassung von Qualität: Evaluationsforschung, Qualitätssicherung und Qualitätsstandards" – Veranstaltung der DGS-Sektion Methoden der empirischen Sozialforschung am 11.10.2006. In Zeitschrift für Evaluation, 1/2007, S. 111/112.

Krueger, Richard A. (1994): Focus groups. A practical guide for applied research. Thousand Oaks: Sage Publications.

Kuckartz, Udo (2006): Quick and dirty? – Qualitative Methoden der drittmittelfinanzierten Evaluation in der Umweltforschung. In **Flick, Uwe (Hrsg.):** Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 267–283.

Kuckartz, Udo (2007): Einführung in die computergestützte Analyse qualitativer Daten. 2. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuckartz, Udo (2009): Inhaltsanalyse. In **Westle, Bettina (Hrsg.):** Methoden der Politikwissenschaft. Baden-Baden: Nomos, S. 334 – 344.

Kuckartz, Udo (2012): Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. Weinheim und Basel: Beltz Juventa.

Kuckartz, Udo et al. (2008): Qualitative Evaluation. Der Einstieg in die Praxis. Band 2., aktualisierte Auflage, Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuper, Harm (2005): Evaluation im Bildungssystem. Eine Einführung. Stuttgart: Kohlhammer.

Labaf-Wiltzsch, Behjat (2006): Evaluation der Lehre. Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen.

Laging, Ralf et al. (2010): Evaluation der fachdidaktischen Ausbildung im Rahmen der Lehramtsausbildung an der Philipps-Universität Marburg – Bericht zur Evaluation. Marburg.

Lamnek, Siegfried (2005): Qualitative Sozialforschung. Lehrbuch. Weinheim und München: Beltz Psychologie Verlags Union.

Lamnek, Siegfried (2005a): Gruppendiskussion. Theorie und Praxis. Weinheim und Basel: Beltz.

Laux, Eberhard/Teppe, Karl (Hrsg.) (1998): Der neuzeitliche Staat und seine Verwaltung. Beiträge zur Entwicklungsgeschichte seit 1700. Stuttgart: Franz Steiner.

Lee, Barbara (2006): Theories of Evaluation. In **Stockmann, Reinhard (2006c):** Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder. 3. Auflage. Münster: Waxmann, S. 137 – 176.

Lewin, Karl et al. (2000): Evaluation der Praxissemester an den Fachhochschulen des Landes Nordrhein-Westfalen. Hannover: HIS.

Löffler, Elke (1998): Die Diskussion über Führung und Management in der öffentlichen Verwaltung als zentrales Anliegen der Verwaltungswissenschaften. In **Laux, Eberhard/Teppe, Karl (Hrsg.) (1998):** Der neuzeitliche Staat und seine Verwaltung. Beiträge zur Entwicklungsgeschichte seit 1700. Stuttgart: Franz Steiner, S. 333 – 348.

Lohnert, Beate/Rolfes, Manfred (1997): Handbuch zur Evaluation von Lehre und Studium an Hochschulen. Ein praxisorientierter Leitfaden. Hannover: Zentrale Evaluationsagentur der niedersächsischen Hochschulen.

Lüders, Christian (2006): Qualitative Evaluationsforschung – was heißt hier Forschung? In **Flick, Uwe (Hrsg.):** Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 33–62.

Lueger, Manfred (2000): Grundlagen qualitativer Feldforschung. Wien: UTB.

Mangold, Werner (1973): Gruppendiskussionen. In **König, René (Hrsg.):** Handbuch der empirischen Sozialforschung. Band 2: Grundlegende Methoden und Techniken der empirischen Sozialforschung, 3. Auflage. Stuttgart: Deutscher Taschenbuch-Verlag, S. 228 – 259.

Mayerhofer, Wolfgang (2007): Das Fokusgruppeninterview. In **Buber, Renate/Holzmüller, Hartmut M. (Hrsg.) (2007):** Qualitative Marktforschung. Konzepte – Methoden – Analysen. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 477 – 490.

Mayntz, Renate/Holm, Kurt/Hübner, Peter (1974): Einführung in die Methoden der empirischen Soziologie. Opladen: Westdeutscher Verlag.

Mayring, Philipp (2010): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 11. Auflage. Weinheim: Beltz.

Mayring, Philipp/Brunner, Eva (2010): Qualitative Inhaltsanalyse. In **Friebertshäuser, Barbara/Langer, Antja/Prengel, Annedore (Hrsg.):** Handbuch

Qualitative Methoden in der Erziehungswissenschaft. 3. Auflage. Weinheim und München: Juventa, S. 323 – 333.

Mensching, Anja (2006): Zwischen Überforderung und Banalisierung – zu den Schwierigkeiten der Vermittlungsarbeit im Rahmen qualitativer Evaluationsforschung. In **Flick, Uwe (Hrsg.):** Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt, S. 339–362.

Merten, Klaus (1995): Inhaltsanalyse. Einführung in Theorie, Methode und Praxis. 2. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften.

Meuser, Michael (2011): Inhaltsanalyse. In **Bohnsack, Ralf/Marotzki, Winfried/Meuser, Michael (Hrsg.):** Hauptbegriffe Qualitativer Sozialforschung. Opladen und Farmington Hills: Verlag Barbara Budrich, S. 89 – 91.

Meyer, Wolfgang/Höhns, Gabriela (2002): Was ist Evaluation? Saarbrücken ⟨URL: `http://www.bibb.de/dokumente/pdf/a11_transform_was-ist-evaluation.pdf`⟩ – Zugriff am 09.06.2009.

Miles, Matthew B./Huberman, A. Michael (1999): Qualitative data analysis. An expanded Sourcebook. 2. Auflage. Thousand Oaks: Sage.

Miller, Jody/Glassner, Barry (2004): The „inside“ and the „outside“: finding realities in interviews. In **Silverman, David (Hrsg.) (2004):** Qualitative Research. Theory, Method and Practice. 2. Auflage. Thousand Oaks: Sage, S. 125 – 139.

Mittag, Sandra/Bornmann, Lutz/Daniel, Hans-Dieter (2003a): Evaluation von Studium und Lehre an Hochschulen. Handbuch zur Durchführung mehrstufiger Evaluationsverfahren. Münster: Waxmann.

Mittag, Sandra/Bornmann, Lutz/Daniel, Hans-Dieter (2003b): Mehrstufige Verfahren für die Evaluation von Studium und Lehre – Eine Zwischenbilanz europäischer Erfahrungen. In **Schwarz, Stefanie/Teichler, Ulrich (Hrsg.) (2003):** Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung. Frankfurt am Main: Campus, S. 183 – 205.

Morgan, David L. (1997): Focus groups as qualitative research. Band 16, 2. Auflage. Thousand Oaks: Sage Publications.

Moser, Sir Claus/Kalton, Graham (2004): Questionnaires. In **Seale, Clive (Hrsg.) (2004):** Social Research Methods. A Reader. London: Routledge Students Readers, S. 73 – 87.

Müller-Böling, Detlef (2000): Die Entfesselte Hochschule. Gütersloh: Bertelsmann-Stiftung.

Münch, Richard/Pechmann, Max (2009): Der Kampf um Sichtbarkeit. Zur Kolonisierung des wissenschaftsinternen Wettbewerbs durch wissenschaftsexterne Evaluationsverfahren. In **Bogumil, Jörg/Heinze, Rolf G. (Hrsg.) (2009):** Neue Steuerung von Hochschulen. Eine Zwischenbilanz. Berlin: Edition Sigma, S. 67 – 92.

Naderer, Gabriele/Balzer, Eva (Hrsg.) (2007): Qualitative Marktforschung in Theorie und Praxis. Grundlagen, Methoden und Anwendungen. Wiesbaden: Gabler Verlag.

Naschold, Frieder/Bogumil, Jörg (2000): Modernisierung des Staates. New Public Management in deutscher und internationaler Perspektive. Opladen: Leske + Budrich.

Nentwig-Gesemann, Iris (2010): Das Gruppendiskussionsverfahren. In **Bock, Karin/Miethe, Ingrid (Hrsg.):** Handbuch Qualitative Methoden in der Sozialen Arbeit. Opladen und Farmington Hillstasse: Verlag Barbara Budrich, S. 259 – 268.

Nickel, Sigrun (2004): Dezentralisierte Zentralisierung. Die Suche nach neuen Organisations- und Leitungsstrukturen für Fakultäten und Fachbereiche. In die hochschule. journal für wissenschaft und bildung, 1, S. 87 – 99.

OECD (2002): FRASCATI-Manual. Proposed Standard Practice for Surveys on Research and Experimental Development. OECD ⟨URL: `http://browse.oecdbookshop.org/oecd/pdfs/free/9202081e.pdf`⟩ – Zugriff am 21.10.2011.

Pasternack, Peer (2004): Qualitätsorientierung an Hochschulen. Verfahren und Instrumente. In Arbeitberichte 5/04.

Patton, Michael Q. (1992): How to use qualitative methods in evaluation. 2. Auflage. Newbury Park: Sage.

Patton, Michael Q. (2002): Qualitative research & evaluation methods. Thousand Oaks: Sage.

Patton, Michael Quinn (1997): Utilization-Focused Evaluation. The New Century Text. Thousand Oaks: Sage Publications.

Peter, Lothar/Wawrzinek, Andreas (1995): Dialogische Evaluation. Ein studentisches Evaluationsverfahren. In **Verbeek, David/Balogh, Heike (Hrsg.) (1995):** Evaluation der Lehre. Ziele – Akzeptanz – Methoden. Band 1, Handbuch Hochschullehre Highlights. Stuttgart u.a.: Raabe, S. 1 – 18.

Pohlenz, Philipp (2008): Datenqualität als Schlüsselfrage der Qualitätssicherung von Lehre und Studium an Hochschulen. Dissertation, Potsdam.

Porst, Rolf (2008): Fragebogen. Ein Arbeitsbuch. Wiesbaden: VS Verlag für Sozialwissenschaften.

Power, Michael (1999): The Audit Society. Rituals of Verification. Band 2. Auflage, Oxford: Oxford University Press.

Prior, Lindsay (2003): Using documents in social research. Thousand Oaks: Sage.

Puchta, Claudia/Wolff, Stephan (2004): Gruppendiskussion in zwei Welten: Gute Gründe für schlechte Praktiken. Vortrag auf der Jahrestagung der Sektion „Methoden der qualitativen Sozialforschung“ der Deutschen Gesellschaft für Soziologie in Tübingen.

Raithel, Jürgen (2006): Quantitative Forschung. Ein Praxiskurs. Wiesbaden: VS Verlag für Sozialwissenschaften.

Reischmann, Jost (2003): Weiterbildungsevaluation. Lernerfolge messbar machen. Kriftel: Luchterhand.

Richter, Roland (Hrsg.) (1994): Qualitätssorge in der Lehre. Leitfaden für die studentische Lehrevaluation. Neuwied: Luchterhand.

Riese, Karina (2007): Kriterien zur Ressourcensteuerung an Hochschulen. Wiesbaden: Deutscher Universitäts-Verlag.

Rindermann, Heiner (2003a): Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. In Zeitschrift für Evaluation,, Nr. 2/2003, S. 233 – 256 ⟨URL: `http://www.zfev.de/fruehereAusgabe/ausgabe2003-2/artikel/ZfEv2-2003_5-Rindermann.pdf`⟩.

Rindermann, Heiner (2003b): Methodik und Anwendung der Lehrveranstaltungsevaluation für die Qualitätsentwicklung an Hochschulen. In Sozialwissenschaften und Berufspraxis, 26, Nr. 4, S. 401 – 413.

Rindermann, Heiner (2009): Lehrevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen. Landau: Verlag Empirische Pädagogik.

Rossi, Peter H./Lipsey, Mark W./Freeman, Howard E. (2004): Evaluation. A systematic approach. 7. Auflage. Thousand Oaks: Sage.

Rössler, Patrick (2005): Inhaltsanalyse. Konstanz: UVK.

Roth, Erwin (1993): Sozialwissenschaftliche Methoden. München u. a.: Oldenbourg.

Sanders, James R./American Association of School Administrators/Joint Committee on Standards for Educational Evaluation (1994): The Program Evaluation Standards. How to Assess Evaluations of Educational Programs. Thousand Oaks: Sage Publications.

Sanders, James R./Beywl, Wolfgang (2006): Handbuch der Evaluationsstandards. Die Standards des Joint Committee on Standards for Educational Evaluation. 3. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften.

Schedler, Kund/Proeller, Isabella (2006): New Public Management. Bern, Stuttgart, Wien: Haupt.

Schimank, Uwe (2009): Governance-Reformen nationaler Hochschulsysteme. Deutschland in internationaler Perspektive. In **Bogumil, Jörg/Heinze, Rolf G. (Hrsg.) (2009):** Neue Steuerung von Hochschulen. Eine Zwischenbilanz. Berlin: Edition Sigma, S. 123 – 137.

Schmidt, Christiane (2010): Auswertungstechniken für Leitfadeninterviews. In **Friebertshäuser, Barbara/Langer, Antja/Prengel, Annedore (Hrsg.):** Handbuch Qualitative Methoden in der Erziehungswissenschaft. 3. Auflage. Weinheim und München: Juventa, S. 473 – 486.

Schmidt, Hans Heinrich (2006a): Universitätsreform und New Public Management. In Quo vadis universitas? Kritische Beiträge zur Idee und Zukunft der Universität,, Nr. 6, S. 1 – 16 ⟨URL: `http://www.unizh.ch/forschung/quovadis.html`⟩.

Schmidt, Uwe (2005): Zwischen Messen und Verstehen. Anmerkungen zum Theoriedefizit in der deutschen Hochschulevaluation. In evaNet-Positionen 06/2005.

Schmidt, Uwe (2006b): Hochschulentwicklung und Evaluation: Perspektiven, Beteiligung und Verantwortung Studierender. In **Hochschulrektorenkonferenz (Hrsg.) (2006):** Qualitätsentwicklung an Hochschulen. Erfahrungen und Lehren aus 10 Jahren Evaluation. Band 8/2006, Bonn: Hochschulrektorenkonferenz ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩, S. 39 – 46 ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩.

Schmidt, Uwe (2009): Evaluation an deutschen Hochschulen – Entwicklung, Stand und Perspektiven. In **Widmer, Thomas/Beywl, Wolfgang/Fabian, Carlo (Hrsg.) (2009):** Evaluation. Ein Systematisches Handbuch. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 163 – 169.

Schnell, Rainer/Hill, Paul B./Esser, Elke (2005): Methoden der empirischen Sozialforschung. 7. Auflage. München u. a.: Oldenbourg.

Schreier, Gerhard/Litz, Raimund (Hrsg.) (1998): Evaluation und Qualitätssicherung an den Hochschulen in Deutschland – Stand und Perspektiven. Bonn.

Schreier, Margit (2012): Qualitative Content Analysis in Practice. Los Angeles, London, New Delhi, Singapore, Washington DC: Sage.

Schwarz, Stefanie/Teichler, Ulrich (Hrsg.) (2003): Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung. Frankfurt am Main: Campus.

Scriven, Michael (1973): Goal-free evaluation. In **House, Ernest R. (Hrsg.):** School evaluation: The politics and process. Berkeley: McCutchan, S. 219 – 328.

Scriven, Michael (2012): Key Evaluation Checklist. Michigan ⟨URL: `http://michaelscriven.info/images/KEC.11.17.12.pdf`⟩ – Zugriff am 30.12.2012.

Seale, Clive (1999): The Quality of Qualitative Research. Thousand Oaks: Sage Publications.

Seale, Clive (Hrsg.) (2004): Social Research Methods. A Reader. London: Routledge Students Readers.

Seale, Clive et al. (Hrsg.) (2004): Qualitative Research Practice. Thousand Oaks: Sage.

Serrano-Velarde, Kathia (2008): Evaluation, Akkreditierung und Politik. Zur Organisation von Qualitätssicherung im Zuge des Bolognaprozesses. Wiesbaden: VS Verlag für Sozialwissenschaften.

Sheehan, Kim/Hoy, Mariea (2004): On-Line Surveys. In **Seale, Clive (Hrsg.) (2004):** Social Research Methods. A Reader. London: Routledge Students Readers, S. 105 – 110.

Silverman, David (Hrsg.) (2004): Qualitative Research. Theory, Method and Practice. 2. Auflage. Thousand Oaks: Sage.

Soellner, Renate/braun, Edith/Gusy, Burkhard (2006): Lehrevaluation aus pädagogisch-psychologischer Sicht. Das Berliner Evaluationsinstrument für Kompetenzen. In **Hochschulrektorenkonferenz (Hrsg.) (2006):** Qualitätsentwicklung an Hochschulen. Erfahrungen und Lehren aus 10 Jahren Evaluation. Band 8/2006, Bonn: Hochschulrektorenkonferenz ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩, S. 72 – 80 ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩.

Spencer, Liz et al. (2003): Quality in Qualitative Evaluation: A framework for assessing research evidence. London: National Centre for Social Research, Cabinet Office.

Stake, Robert E. (1975): Program evaluation particularly responsive evaluation. No address in ⟨URL: `http://www.wmich.edu/evalctr/pubs/ops/ops05.pdf`⟩ – Zugriff am 30.04.2010.

Staufenbiel, Thomas (2000): Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. In Diagnostica, 46, Nr. 4, 169–181.

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

Stock, Manfred (2004): Steuerung als Fiktion. Anmerkungen zur Implementierung der neuen Steuerungskonzepte an Hochschulen aus organisationssoziologischer Sicht. In die hochschule. journal für wissenschaft und bildung, 1, S. 30 – 48.

Stockmann, Reinhard (2006a): Evaluation in Deutschland. In **Stockmann, Reinhard (2006c):** Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder. 3. Auflage. Münster: Waxmann, S. 15 – 46.

Stockmann, Reinhard (2006b): Evaluation und Qualitätsentwicklung. Eine Grundlage für wirkungsorientiertes Qualitätsmanagement. Münster: Waxmann.

Stockmann, Reinhard (2006c): Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder. 3. Auflage. Münster: Waxmann.

Stockmann, Reinhard (2007a): Einführung in die Evaluation. In **Stockmann, Reinhard (2007b):** Handbuch zur Evaluation. Eine praktische Handlungsanleitung. Münster: Waxmann, S. 24 – 70.

Stockmann, Reinhard (2007b): Handbuch zur Evaluation. Eine praktische Handlungsanleitung. Münster: Waxmann.

Stockmann, Reinhard/Meyer, Wolfgang (2010): Evaluation. Eine Einführung. Opladen & Farmington Hills: Verlag Barbara Budrich.

Strauss, Anselm/Corbin, Juliet (1996): Grounded Theory: Grundlagen Qualitativer Sozialforschung. Weinheim: Beltz.

Stufflebeam, Daniel L. et al. (1971): Educational Evaluation & Decision Making. 4. Auflage. Itasca: F. E. Peacock.

Suchman, Edward Allen (1967): Evaluative Research. Principles and Practice in Public Service & Social Action Programs. New York: Russel Sage Foundation.

Tashakkori, Abbas (2003): Handbook of mixed methods in the social & behavioral research. Thousand Oaks: Sage.

Tashakkori, Abbas/Teddlie, Charles (2000): Mixed methodology. Combining qualitative and quantitative approaches. 4. Auflage. Thousand Oaks: Sage.

Vedung, Evert (2006): Evaluation Research and Fundamental Research. In **Stockmann, Reinhard (2006c):** Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder. 3. Auflage. Münster: Waxmann, S. 113 – 136.

Verbeek, David/Balogh, Heike (Hrsg.) (1995): Evaluation der Lehre. Ziele – Akzeptanz – Methoden. Band 1, Handbuch Hochschullehre Highlights. Stuttgart u.a.: Raabe.

Verbund Norddeutscher Universitäten/Hochschulrektorenkonferenz (Hrsg.) (2004): Evaluation – ein Bestandteil des Qualitätsmanagements an Hochschulen. Bonn: Verbund Norddeutscher Universitäten.

Weber, Regina (2006): Studentische Lehrevaluation: Anforderungen aus studentischer Sicht. In **Hochschulrektorenkonferenz (Hrsg.) (2006):** Qualitätsentwicklung an Hochschulen. Erfahrungen und Lehren aus 10 Jahren Evaluation. Band 8/2006, Bonn: Hochschulrektorenkonferenz ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩, S. 52 – 57 ⟨URL: `http://www.hrk.de/de/download/dateien/2006-08_Qualitaetsentwicklung_an_Hochschulen.pdf`⟩.

Weber, Robert Philip (2004): Content Analysis. In **Seale, Clive (Hrsg.) (2004):** Social Research Methods. A Reader. London: Routledge Students Readers, S. 117 – 124.

Widmer, Thomas (2006): Qualität der Evaluation – Wenn Wissenschaft zur praktischen Kunst wird. In **Stockmann, Reinhard (2006b):** Evaluation und Qualitätsentwicklung. Eine Grundlage für wirkungsorientiertes Qualitätsmanagement. Münster: Waxmann, S. 85 – 112.

Widmer, Thomas/Beywl, Wolfgang/Fabian, Carlo (Hrsg.) (2009): Evaluation. Ein Systematisches Handbuch. Wiesbaden: VS Verlag für Sozialwissenschaften.

Widmer, Thomas/Landert, Charles/Bachmann, Nicole (2000): Evaluations-

Standards der Schweizerischen Evaluationsgesellschaft (SEVAL-Standards). Freiburg: Schweizerische Evaluationsgesellschaft (SEVAL).

Wilkinson, Sue (2004): Focus group research. In **Silverman, David (Hrsg.) (2004):** Qualitative Research. Theory, Method and Practice. 2. Auflage. Thousand Oaks: Sage, S. 177 – 199.

Wottawa, Heinrich/Thierau, Heike (2003): Lehrbuch Evaluation. 3. Auflage. Bern: Huber.

A. Anhang

Inhaltsverzeichnis

A.1. Interviewleitfaden Betroffene

Leitfaden Einzelinterviews mit Betroffenen

✓	Thema	Anmerkungen
☐	**Begrüßung** • Bedanken • Kurzvorstellung meiner Person und meines Vorhabens: Untersuchung über Forschungsmethoden, die im Rahmen von Evaluationen an Hochschulen zum Einsatz kommen • Erläuterung der Rolle der Interviews für die Dissertation: Sicht der Teilnehmenden erfasssen • Hinweis auf die Weiterverarbeitung der erhobenen Daten und die Strategien zur Wahrung der Anonymität	• Wahrung der Anonymität durch komplette Anonymisierung: ◦ Vergabe von Decknamen ◦ Umschreibung von Informationen, die eine Identifikation ermöglichten • Kein Rückgriff auf die Daten der FD-Eval., keine Verbindung zwischen den Daten • Weitergabe der Daten ausschließlich in transkribierter und anonymisierter Form
☐	**Aufnahmegerät einschalten**	
☐	**Rekapitulation der Fachdidaktik-Evaluation (Stichworte/Themen):** • Haben Sie bereits Erfahrung mit Evaluationen, sowohl als Teilnehmer/in und/oder als planende Person? • Was verstehen Sie unter Evaluation? • Wurden Sie über die Ziele und Verfahrensweisen der Fachdidaktik-Evaluation informiert? Wie? • Welche Ziele verfolgte die Fachdidaktik-Evaluation aus Ihrer Wahrnehmung? (*Für mich:* was kam bei Ihnen an?) • Entsprach die FD-Evaluation Ihrem Verständnis von Evaluation? • Wurden Sie darüber informiert, welche Daten erhoben wurden und wie mit den gesammelten Daten weiter verfahren wurde? • Fühlen Sie sich ausreichend informiert?	• **Hauptpunkt 1:** Messung der Güte der Fachdidaktik-Ausbildung • **Hauptpunkt 2:** Verbesserung der Ausbildung • gleichzeitig Generierung von Wissen über die Zusammenhänge • Grob: Bestand und Entwicklungsmöglichkeiten der Marburger Fachdidaktik-Ausbildung
☐	**Das Konstrukt der Gegenstandsangemessenheit.** **Input:** Gegenstandsangemessenheit ist eines der zentralen Kriterien für die Wahl einer Datenerhebungsmethode in der Sozialforschung, also auch bei Evaluationen. Die Daten werden ja erhoben, um eine konkrete Frage oder Forschungsfrage zu beantworten. Und die Methode, mit der die Daten erhoben werden, soll möglichst gut zum Untersuchungsgegenstand, also dem untersuchten Objekt – hier der Fachdidaktikausbildung –, passen. (**Für mich**: Gibt es bis hier evtl. Nachfragen?) Nun möchte ich Sie bitten, sich in die Situation zu versetzen, die Fachdidaktik-Evaluation planen zu müssen. Dabei geht es nicht um die Kritik der durchgeführten Evalua-	**Ziel dieses Teils:** Annäherung an das Konzept der Gegenstandsangemessenheit Bei den Begründungen geht es um Metabeschreibungen: „Muss zeiteffizient sein.“ „Muss die richtigen Antworten liefern.“ etc. **Wichtig:** Reale Zwänge wie Geld- oder Sachmittel spielen hier **keine Rolle**: Es geht um den (fiktiven) Idalfall

1

✓	Thema	Anmerkungen
	tion, sondern um methodische Überlegungen! (**Ggf. Hinweis:** Die Daten werden – wenn überhaupt – nur anonymisiert weitergegeben!) 1. Wer müsste denn Ihrer Meinung nach beteiligt werden, um die Fachdidaktikausbildung angemessen bewerten zu können? Warum?*(Mitnotieren!)* (**Für mich:** In welcher Rolle?) 2. Was, also welche Daten, müssten unbedingt erhoben werden? Warum? (Präzisierung: Wonach sollten die genannten Gruppen gefragt werden?) 3. Wie, also mit welchen Methoden, ließen sich die Daten denn erheben und die benannten Personen einbeziehen? Warum! („Wie könnten denn die Daten erhoben werden?“)	**Immer nachfragen:** • **Was ist an der Person/engruppe, den Daten … wichtig?** • **Was würde sonst fehlen?** • ***Etc. - an jedem Punkt!*** Was meint die Person genau mit den genannten Methoden? Beschreiben lassen.
☐	**Gegenstandsangemessenheit der Methoden:** Wie schätzten Sie die Angemessenheit der in der Fachdidaktikevaluation tatsächlich eingesetzten Methoden – so weit Sie darüber informiert sind – ein? • Ist die Methode *Offenes Interview mit Fachdidaktik-Verantwortlichen* geeignet, die Ziele der Evaluation zu erreichen? • Ist die Methode *Dokumentenanalyse* (Analyse der Inhalte der Modulbeschreibungen) geeignet, die Ziele der Evaluation zu erreichen? • Ist die Methode *Online-Datenerhebung mit Studierenden* geeignet, die Ziele der Evaluation zu erreichen? • Wo liegen Probleme der eingesetzten Methoden? • Wo liegen Stärken der Methoden? • Warum könnte sie gewählt worden sein?	Auch abfragen, wenn zwar bereits etwas dazu gesagt wurde, aber nur Grundlegendes. **Nach gründlicher vorheriger Besprechung nur noch zusammenfassen.** Wichtig: **Warum?**

☐	**Einschätzung von Methoden:** *Nur noch die Methoden abfragen, die noch nicht genannt wurden.* Welche Datenerhebungsmethoden wären geeignet, um den Zielen der Fachdidaktikevaluation gerecht zu werden? • *Standardisierte Datenerhebung* ◦ Vollstandardisiert, d.h. alle Fragen und Antwortmöglichkeiten sind vorgegeben ◦ Teilstandardisiert, d.h. Mischung standardisiert und nicht standardisierter, offener Fragen ◦ Papierfragebogen ◦ Face to face-Erhebung ◦ Onlinefragebogen • *Offene Datenerhebung/Interview* ◦ gänzliches offenes Interview, bei dem keinerlei Fragen und Antworten vorgegeben sind ◦ vorstrukturiertes Interview, das – so, wie dieses Gespräch – durch einen Leitfaden gegliedert wird, der relevante Themen beinhaltet; gleichzeitig bleibt das Interview aber offen für spezielle Schwerpunktsetzungen ◦ Face to face/mündlich ◦ per Telefon/mündlich ◦ per Chat/Mail/schriftlich ◦ per Fragebogen/Tagebuch/schriftlich • *Gruppendiskussion:* Mehrere Personen diskutieren mit einander über das gleiche, durch einen Moderator vorgegebene und strukturierte Themengebiet, Moderator kann das Gespräch unterschiedlich stark lenken • *Dokumenten- und Kennzahlenanalyse:* (inerpretative) Auswertung bereits vorliegenden Materials (z.B. Modulbeschreibungen) oder Rückgriff auf Kennzahlen wie z.B. Seminarteilnahmen oder positiv bewältigte Prüfungsleistungen • *Verfahren mit Gutachtern:* Bewertung der Fachdidaktikausbildung durch Fachkundige und besonders dafür geschulte Personen, beispielsweise entsprechend qualifizierte Personen einer anderen Universität	Wichtig ist es, bei allen (Alternativ-) Vorschlägen, nach dem **Warum** der Eignung zu fragen. Genannte Methoden möglichst gut beschreiben lassen. Es geht um die *Gedanken zur Angemessenheit.*

3

☐	**Resümee: Was macht eine Methode zu einer geeigneten Methode?** • Wann wären Sie als befragte Person mit einer Evaluation zufrieden? Aspekte: • *formale Kriterien:* Welchen technischen Kriterien (Arbeitsbelastung, zahlenmäßige Reichweite, Schwierigkeit der Durchführung etc.) muss sie genügen? Welche Wertigkeit gibt es innerhalb der genannten Aspekte • *subjektive Kriterien:* Wann kann man als Befragter mit einer Methode zufrieden sein?	Ggf. nach Kriterien fragen, anhand derer Planende entscheiden können, ob eine Methode angemessen ist. **Welche Merkmale sind Ihnen als Proband wichtig?** **Priorisierung** innerhalb der genannten Punkte?
☐	**Aufnahmegerät ausschalten,** Danken.	
	Anmerkungen der befragten Person zum Interview (lautes Denken, „Was geht Ihnen gerade hinsichtlich dieses Interviews durch den Kopf?“): **Postskriptum:**	

A.2. Interviewleitfaden Beteiligte

Leitfaden Einzelinterviews mit Beteiligten

✓	Thema	Anmerkungen
☐	**Begrüßung** • Bedanken • Kurzvorstellung meiner Person und meines Vorhabens: Untersuchung über Forschungsmethoden, die im Rahmen von Evaluationen an Hochschulen zum Einsatz kommen • Erläuterung der Rolle der Interviews für die Dissertation: Sicht der Teilnehmenden erfasssen • Hinweis auf die Weiterverarbeitung der erhobenen Daten und die Strategien zur Wahrung der Anonymität	• Wahrung der Anonymität durch komplette Anonymisierung: ◦ Vergabe von Decknamen ◦ Umschreibung von Informationen, die eine Identifikation ermöglichten • Kein Rückgriff auf die Daten der FD-Eval., keine Verbindung zwischen den Daten • Weitergabe der Daten ausschließlich in transkribierter und anonymisierter Form
☐	**Aufnahmegerät einschalten**	
☐	**Rekapitulation der Fachdidaktik-Evaluation (Stichworte/Themen):** • Haben Sie bereits Erfahrung mit Evaluationen, sowohl als Teilnehmer/in und/oder als planende Person? • Was verstehen Sie unter Evaluation? • Wie wurden Sie über die Ziele und Verfahrensweisen der Evaluation informiert? • Welche Ziele verfolgte die Fachdidaktik-Evaluation aus Ihrer Wahrnehmung? (*Für mich:* was kam bei Ihnen an?) • Entsprach die FD-Evaluation Ihrem Verständnis von Evaluation? • Wie wurden Sie in die Evaluation einbezogen? • Wurden Sie darüber informiert, welche Daten sonst noch erhoben werden und wie mit den gesammelten Daten weiter verfahren wird und/oder welche Rolle das mit Ihnen geführte Interview insgesamt spielt? • Fühlen Sie sich ausreichend informiert?	• **Hauptpunkt 1:** Messung der Güte der Fachdidaktik-Ausbildung • **Hauptpunkt 2:** Verbesserung der Ausbildung • gleichzeitig Generierung von Wissen über die Zusammenhänge • Grob: Bestand und Entwicklungsmöglichkeiten der Marburger Fachdidaktik-Ausbildung

1

✓	Thema	Anmerkungen
☐	**Das Konstrukt der Gegenstandsangemessenheit.** **<u>Input:</u>** Gegenstandsangemessenheit ist eines der zentralen Kriterien für die Wahl einer Datenerhebungsmethode in der Sozialforschung, also auch bei Evaluationen. Die Daten werden ja erhoben, um eine konkrete Frage oder Forschungsfrage zu beantworten. Und die Methode, mit der die Daten erhoben werden, soll möglichst gut zum Untersuchungsgegenstand, also dem untersuchten Objekt – hier der Fachdidaktikausbildung –, passen. (**Für mich**: Gibt es bis hier evtl. Nachfragen?) Nun möchte ich Sie bitten, sich in die Situation zu versetzen, die Fachdidaktik-Evaluation planen zu müssen. Dabei geht es nicht um die Kritik der durchgeführten Evaluation, sondern um methodische Überlegungen! (**Ggf. Hinweis:** Die Daten werden – wenn überhaupt – nur anonymisiert weitergegeben!) 1. Wer müsste denn Ihrer Meinung nach beteiligt werden, um die Fachdidaktikausbildung angemessen bewerten zu können? Warum?*(Mitnotieren!)* (**Für mich:** In welcher Rolle?) 2. Was, also welche Daten, müssten unbedingt erhoben werden? Warum? (Präzisierung: Wonach sollten die genannten Gruppen gefragt werden?) 3. Wie, also mit welchen Methoden, ließen sich die Daten denn erheben und die benannten Personen einbeziehen? Warum! („Wie könnten denn die Daten erhoben werden?“)	**Ziel dieses Teils:** Annäherung an das Konzept der Gegenstandsangemessenheit Bei den Begründungen geht es um Metabeschreibungen: „Muss zeiteffizient sein.“ „Muss die richtigen Antworten liefern.“ etc. **Wichtig:** Reale Zwänge wie Geld- oder Sachmittel spielen hier **keine Rolle**: Es geht um den (fiktiven) Idalfall **Immer nachfragen:** • **Was ist an der Person/engruppe, den Daten … wichtig?** • **Was würde sonst fehlen?** • ***Etc. - an jedem Punkt!*** Was meint die Person genau mit den genannten Methoden? Beschreiben lassen.
☐	**Gegenstandsangemessenheit der Methoden:** Wie schätzten Sie die Angemessenheit der in der Fachdidaktikevaluation tatsächlich eingesetzten Methoden – so weit Sie darüber informiert sind – ein? • Ist die Methode *Offenes Interview mit Fachdidaktik-Verantwortlichen* geeignet, die Ziele der Evaluation zu erreichen? • Ist die Methode *Dokumentenanalyse* (Analyse der Inhalte der Modulbeschreibungen) geeignet, die Ziele der Evaluation zu erreichen? • Ist die Methode *Online-Datenerhebung mit Studierenden* geeignet, die Ziele der Evaluation zu erreichen? • Wo liegen Probleme der eingesetzten Methoden? • Wo liegen Stärken der Methoden? • Warum könnte sie gewählt worden sein?	Auch abfragen, wenn zwar bereits etwas dazu gesagt wurde, aber nur Grundlegendes. **Nach gründlicher vorheriger Besprechung nur noch zusammenfassen.** Wichtig: **Warum?**

☐	**Einschätzung von Methoden:** *Nur noch die Methoden abfragen, die noch nicht genannt wurden.* Welche Datenerhebungsmethoden wären geeignet, um den Zielen der Fachdidaktikevaluation gerecht zu werden? • *Standardisierte Datenerhebung* ◦ Vollstandardisiert, d.h. alle Fragen und Antwortmöglichkeiten sind vorgegeben ◦ Teilstandardisiert, d.h. Mischung standardisiert und nicht standardisierter, offener Fragen ◦ Papierfragebogen ◦ Face to face-Erhebung ◦ Onlinefragebogen • *Offene Datenerhebung/Interview* ◦ gänzliches offenes Interview, bei dem keinerlei Fragen und Antworten vorgegeben sind ◦ vorstrukturiertes Interview, das – so, wie dieses Gespräch – durch einen Leitfaden gegliedert wird, der relevante Themen beinhaltet; gleichzeitig bleibt das Interview aber offen für spezielle Schwerpunktsetzungen ◦ Face to face/mündlich ◦ per Telefon/mündlich ◦ per Chat/Mail/schriftlich ◦ per Fragebogen/Tagebuch/schriftlich • *Gruppendiskussion:* Mehrere Personen diskutieren mit einander über das gleiche, durch einen Moderator vorgegebene und strukturierte Themengebiet, Moderator kann das Gespräch unterschiedlich stark lenken • *Dokumenten- und Kennzahlenanalyse:* (inerpretative) Auswertung bereits vorliegenden Materials (z.B. Modulbeschreibungen) oder Rückgriff auf Kennzahlen wie z.B. Seminarteilnahmen oder positiv bewältigte Prüfungsleistungen • *Verfahren mit Gutachtern:* Bewertung der Fachdidaktikausbildung durch Fachkundige und besonders dafür geschulte Personen, beispielsweise entsprechend qualifizierte Personen einer anderen Universität	Wichtig ist es, bei allen (Alternativ-) Vorschlägen, nach dem **Warum** der Eignung zu fragen. Genannte Methoden möglichst gut beschreiben lassen. Es geht um die *Gedanken zur Angemessenheit.*

3

☐	**Resümee: Was macht eine Methode zu einer geeigneten Methode?** • Wann wären Sie als befragte Person mit einer Evaluation zufrieden? Aspekte: • *formale Kriterien:* Welchen technischen Kriterien (Arbeitsbelastung, zahlenmäßige Reichweite, Schwierigkeit der Durchführung etc.) muss sie genügen? Welche Wertigkeit gibt es innerhalb der genannten Aspekte • *subjektive Kriterien:* Wann kann man als Befragter mit einer Methode zufrieden sein?	Ggf. nach Kriterien fragen, anhand derer Planende entscheiden können, ob eine Methode angemessen ist. **Welche Merkmale sind Ihnen als Proband wichtig?** **Priorisierung** innerhalb der genannten Punkte?
☐	**Aufnahmegerät ausschalten,** Danken.	
	Anmerkungen der befragten Person zum Interview (lautes Denken, „Was geht Ihnen gerade hinsichtlich dieses Interviews durch den Kopf?“): **Postskriptum:**	

A.3. Interviewleitfaden Planende

Leitfaden Einzelinterviews mit Planenden

Frage	Notizen
Begrüßung • Bedanken • Kurzvorstellung meiner Person und meines Vorhabens: Untersuchung über Forschungsmethoden, die im Rahmen von Evaluationen an Hochschulen zum Einsatz kommen • Input: Gegenstandsangemessenheit ist eines der zentralen Kriterien für die Wahl einer Datenerhebungsmethode in der Sozialforschung, also auch bei Evaluationen. Die Daten werden ja erhoben, um eine konkrete Frage oder Forschungsfrage zu beantworten. Und die Methode, mit der die Daten erhoben werden, soll möglichst gut zum Untersuchungsgegenstand, also dem untersuchten Objekt – hier der Fachdidaktikausbildung –, passen. (Für mich: Gibt es bis hier evtl. Nachfragen?) • Erläuterung der Rolle der Interviews für die Dissertation: Sicht der Planenden erfassen • Hinweis auf die Weiterverarbeitung der erhobenen Daten und die Strategien zur Wahrung der Anonymität	• Anonymisierung: ◦ Vergabe von Decknamen ◦ Umschreibung von Informationen, die eine Identifikation ermöglichten • Weitergabe der Daten ausschließlich transkribiert und anonymisiert
Aufnahme starten	
• Welche Ziele verfolgte die Fachdidaktik-Evaluation?	
Ich möchte Sie bitten, die für die Evaluation der Fachdidaktik-Ausbildung gewählten Verfahren zur Datenerhebung hinsichtlich ihrer Angemessenheit zu erläutern. Dabei geht es nicht um die Kritik der durchgeführten Evaluation und ihrer Ergebnisse, sondern darum, aufgrund welcher methodischen Vorüberlegungen die Datenerhebungsmethoden gewählt wurden und ob Sie die Eignung der Methoden noch mal genauer erläutern könnten.	
• Aus welchen Überlegungen heraus wurden die Stakeholdergruppen der Lehrenden und Studierenden als Datengeber ausgewählt? o Warum nicht auch bspw. Absolventen?	
• Welche Daten bzw. Informationen wurden von den Gruppen erfragt? Mit welchem Ziel? (Präzisierung: *Wonach* wurden die genannten Gruppen gefragt?)	
• Wie, also mit welchen Methoden, wurden die Daten erhoben? Warum?	
Wenn eine der folgenden Methoden nicht angesprochen wurde gezielt nach ihrer Angemessenheit fragen: • *Dokumentenanalyse* (Analyse der Inhalte der Modulbeschreibungen) • *Offenes Interview mit Fachdidaktik-Verantwortlichen* • *Online-Datenerhebung mit Studierenden* • Noch einmal zusammengefasst: o Wo liegen in Ihren Augen die Stärken der eingesetzten Methoden? o Wo liegen eventuelle methodische Probleme der eingesetzten Verfahren?	
• Gehen Sie davon aus, dass die Befragten mit der	

Evaluation zufrieden sind? • Wo bestehen Ansatzpunkte für Kritik an den Datenerhebungsmethoden durch die Befragten?	
Nach der Gegenstandsangemessenheit folgender Methoden im Kontext der FD-Evaluation fragen – **so weit sie noch nicht genannt wurden**: • *Standardisierte Datenerhebung* o Vollstandardisiert, d.h. alle Fragen und Antwortmöglichkeiten sind vorgegeben o Teilstandardisiert, d.h. Mischung standardisiert und nicht standardisierter, offener Fragen o Papierfragebogen o Face to face-Erhebung o Onlinefragebogen • *Offene Datenerhebung/Interview* o gänzliches offenes Interview, bei dem keinerlei Fragen und Antworten vorgegeben sind o vorstrukturiertes Interview, das – so, wie dieses Gespräch – durch einen Leitfaden gegliedert wird, der relevante Themen beinhaltet; gleichzeitig bleibt das Interview aber offen für spezielle Schwerpunktsetzungen o Face to face/mündlich o per Telefon/mündlich o per Chat/Mail/schriftlich o per Fragebogen/Tagebuch/schriftlich • *Gruppendiskussion:* Mehrere Personen diskutieren mit einander über das gleiche, durch einen Moderator vorgegebene und strukturierte Themengebiet, Moderator kann das Gespräch unterschiedlich stark lenken • *Dokumenten- und Kennzahlenanalyse:* (interpretative) Auswertung bereits vorliegenden Materials (z.B. Modulbeschreibungen) oder Rückgriff auf Kennzahlen wie z.B. Seminarteilnahmen oder positiv bewältigte Prüfungsleistungen • *Verfahren mit Gutachtern:* Bewertung der Fachdidaktikausbildung durch Fachkundige und besonders dafür geschulte Personen, beispielsweise entsprechend qualifizierte Personen einer anderen Universität	
Losgelöst von der bisher besprochenen Evaluation und auf einer allgemeinen Ebene: • Woran lässt sich die Eignung einer Methode erkennen? • Welchen Kriterien muss die Datenerhebung einer Evaluation entsprechen, damit man als **beteiligte oder betroffene** Person damit zufrieden sein kann?	
Ende	

A.4. Interviewleitfaden Gruppendiskussion

Leitfaden Gruppendiskussion

✓	Thema	Anmerkungen
☐	**Begrüßung** • Bedanken • Kurzvorstellung meiner Person und meines Vorhabens: Untersuchung über Forschungsmethoden, die im Rahmen von Evaluationen an Hochschulen zum Einsatz kommen • Erläuterung der Rolle der Interviews für die Dissertation: Sicht der Teilnehmenden erfasssen • Hinweis auf die Weiterverarbeitung der erhobenen Daten und die Strategien zur Wahrung der Anonymität	• Wahrung der Anonymität durch komplette Anonymisierung: ◦ Vergabe von Decknamen ◦ Umschreibung von Informationen, die eine Identifikation ermöglichten • Kein Rückgriff auf die Daten der FD-Eval., keine Verbindung zwischen den Daten • Weitergabe der Daten ausschließlich in transkribierter und anonymisierter Form
☐	**Aufnahmegerät einschalten**	
	Einleitung • **Erläuterung durch Moderator**: Evaluation als Bewertung. • Haben Sie bereits Erfahrung mit Evaluationen, sowohl als Teilnehmer/in und/oder als planende Personen?	GENERELL: Probleme/Streitpunkte extrahieren und zur Diskussion stellen!
☐	**Das Konstrukt der Gegenstandsangemessenheit.** **Input:** Gegenstandsangemessenheit ist eines der zentralen Kriterien für die Wahl einer Datenerhebungsmethode in der Sozialforschung, also auch bei Evaluationen. Die Daten werden ja erhoben, um eine konkrete Frage oder Forschungsfrage zu beantworten. Und die Methode, mit der die Daten erhoben werden, soll möglichst gut zum Untersuchungsgegenstand, also dem untersuchten Objekt – hier der Fachdidaktikausbildung –, passen. (**Für mich**: Gibt es bis hier evtl. Nachfragen?) Nun möchte ich Sie bitten, sich in die Situation zu versetzen, die Fachdidaktik-Evaluation planen zu müssen. Dabei geht es nicht um die Kritik der durchgeführten Evaluation, sondern um methodische Überlegungen! (**Ggf. Hinweis:** Die Daten werden – wenn überhaupt – nur anonymisiert weitergegeben!) 1. Wer müsste denn Ihrer Meinung nach beteiligt werden, um die Fachdidaktikausbildung angemessen bewerten zu können? Warum?*(Mitnotieren!)* (**Für mich:** In welcher Rolle?)	**Ziel dieses Teils:** Annäherung an das Konzept der Gegenstandsangemessenheit Bei den Begründungen geht es um Metabeschreibungen: „Muss zeiteffizient sein.“ „Muss die richtigen Antworten liefern.“ etc. **Wichtig:** Reale Zwänge wie Geld- oder Sachmittel spielen hier **keine Rolle**: Es geht um den (fiktiven) Idealfall **Immer nachfragen:** • **Was ist an der Personengruppe, den Daten … wichtig?** • **Was würde sonst fehlen?** • ***Etc. - an jedem Punkt!***

✓	Thema	Anmerkungen
	2. Was, also welche Daten, müssten unbedingt erhoben werden? Warum? (Präzisierung: Wonach sollten die genannten Gruppen gefragt werden?) 3. Wie, also mit welchen Methoden, ließen sich die Daten denn erheben und die benannten Personen einbeziehen? Warum! („Wie könnten denn die Daten erhoben werden?“)	Was meint die Person genau mit den genannten Methoden? Beschreiben lassen.
	Diskussionsthemen: • Welche Anforderungen stellt eigentlich der Gegenstand „Fachdidaktik-Ausbildung“ an eine Methode? Was ist das Besondere, das eine Methode berücksichtigen sollte? • Was „kann“ eine ideale Methode, wann wäre eine Methodenwahl ideal? • Wie kann man eine gute (i.S.v. geeignete) von einer schlechten (also ungeeigneten) Methode unterscheiden? • Was genau macht die Passung von Methode und Gegenstand aus?	**Ziel:** nicht Abfrage von Einzelmeinungen, sondern möglichst Austausch und Aushandlungsprozesse.
☐	**Liste gängiger Methoden und ihrer Ausprägungen:** • Standardisierte Datenerhebung ◦ Vollstandardisiert, d.h. alle Fragen und Antwortmöglichkeiten sind vorgegeben ◦ Teilstandardisiert, d.h. Mischung standardisiert und nicht standardisierter, offener Fragen ◦ Papierfragebogen ◦ Face to face-Erhebung ◦ Onlinefragebogen • Offene Datenerhebung/Interview ◦ gänzliches offenes Interview, bei dem keinerlei Fragen und Antworten vorgegeben sind ◦ vorstrukturiertes Interview, das – so, wie dieses Gespräch – durch einen Leitfaden gegliedert wird, der relevante Themen beinhaltet; gleichzeitig bleibt das Interview aber offen für spezielle Schwerpunktsetzungen ◦ Face to face/mündlich	Wichtig ist es, bei allen (Alternativ-) Vorschlägen, nach dem **Warum** der Eignung zu fragen. Genannte Methoden möglichst gut beschreiben lassen. Es geht um die Gedanken zur Angemessenheit.

✓	Thema	Anmerkungen
	◦ per Telefon/mündlich ◦ per Chat/Mail/schriftlich ◦ per Fragebogen/Tagebuch/schriftlich • Gruppendiskussion: Mehrere Personen diskutieren mit einander über das gleiche, durch einen Moderator vorgegebene und strukturierte Themengebiet, Moderator kann das Gespräch unterschiedlich stark lenken • Dokumenten- und Kennzahlenanalyse: (inerpretative) Auswertung bereits vorliegenden Materials (z.B. Modulbeschreibungen) oder Rückgriff auf Kennzahlen wie z.B. Seminarteilnahmen oder positiv bewältigte Prüfungsleistungen • Verfahren mit Gutachtern: Bewertung der Fachdidaktikausbildung durch Fachkundige und besonders dafür geschulte Personen, beispielsweise entsprechend qualifizierte Personen einer anderen Universität	
☐	**Aufnahmegerät ausschalten, Danken.**	
	Anmerkungen der befragten Person zum Interview (lautes Denken, „Was geht Ihnen gerade hinsichtlich dieses Interviews durch den Kopf?“): **Postskriptum:**	

A.5. Informationsanschreiben Einzelinterviews

Informationen zum Interview

Vielen Dank, dass Sie sich bereit erklärt haben, meine Promotion durch ein Interview zu unterstützen. Mit diesem Schreiben möchte ich Sie über Hintergrund und Ablauf des Interviews sowie über die Verwendung der erhobenen Daten informieren.

Thema der Arbeit, Zielsetzung der Interviews

Meine Dissertation befasst sich mit der Frage nach der *Gegenstandsangemessenheit* von Datenerhebungsmethoden, die im Kontext von Evaluationen an Hochschulen zum Einsatz kommen. Den Interviews, die mit Beteiligten und Betroffenen der Fachdidaktik-Evaluation geführt werden, kommt dabei die Rolle zu, die Frage zu klären, wie diese Stakeholder die Angemessenheit von Datenerhebungsmethoden einschätzen.

Dauer des Interviews

Das Interview wird ca. 30 bis max. 45 Minuten dauern.

Aufzeichnung des Interviews

Das Interview wird mit Hilfe eines digitalen Audioaufnahmegerätes aufgezeichnet.

Stellung des Interviews in der Dissertation

Das Interview, das ich mit Ihnen führen werde, ist eines von insg. 30, die sich auf verschiedene Stakeholdergruppen (Lehrende, Studierende, Planende) verteilen. Darüber hinaus ist eine Gruppendiskussion mit sechs Teilnehmer/-innen geplant.

Was geschieht mit den erhobenen Daten?

- Die durch das Interview erhobenen Daten werden wörtlich transkribiert.
- Während der Transkription erfolgt eine *Anonymisierung*, d.h.:
  - Die Namen der befragten Personen sowie Namen, die während des Interviews genannt werden, werden durch Decknamen ersetzt. Ausnahmen betreffen hierbei u.U. Personen, die eine zentrale Bedeutung für die Fachdidaktik-Evaluation haben und auch bei der Ersetzung des Namens unmittelbar erkennbar blieben, bspw. Prof. Laging.
  - Allgemeine Informationen, die eine Identifikation von Personen ermöglichen – etwa Aussagen über einen konkreten Fachbereich o.ä. –, werden ebenfalls verschleiert, i.d.R. durch Umschreibungen. Auch hierbei kann allerdings unter den oben genannten Umständen von einer Anonmyisierung abgesehen werden.
- *Vertraulichkeit*: Die Daten müssen im Rahmen der Promotion mindestens den Gutachtern zugänglich gemacht werden. Sämtliche Informationen werden jedoch selbstverständlich vertraulich behandelt, d.h. dass sie prinizpiell nicht als Einzeldaten veröffentlicht oder weitergegeben werden. Jede Weitergabe erfolgt außerdem *ausschließlich* in transkribierter und anonymisierter Form.
- Die im Interview erhobenen Daten werden *nicht* mit denen der eigentlichen Fachdidaktik-Evaluation verbunden, etwa um zu untersuchen, ob die Aussagen dieses Interviews mit jenen der eigentlichen Evaluation korrespondieren.
- Schließlich werden die Daten *ausgewertet*, d.h. die Interviews werden mit Hilfe von Methoden der qualitativen Sozialforschung (interpretative Verfahren) bearbeitet, um die Forschungsfrage zu beantworten. Dabei wird in erster Linie nach Faktoren gesucht, die Hinweise auf die Konstruktion von Gegenstandsangemessenheit liefern können.

Zum Abschluss möchte ich mich nochmals herzlich für Ihre Bereitschaft zu einem Interview bedanken. Falls Sie weitere Fragen haben sollten, so bitte ich Sie, sich mit mir in Verbindung zu setzen. Sie erreichen mich entweder per E-Mail unter der Adresse stefer@staff.uni-marburg.de oder telefonisch – am besten vormittags – unter der Nummer 06421/28-24924.

Mit freundlichen Grüßen,
Claus Stefer

A.6. Informationsanschreiben Gruppendiskussion

Informationen zur Gruppendiskussion am 18.04.2011

Vielen Dank, dass Sie sich bereit erklärt haben, meine Promotion durch die Teilnahme an einer Gruppendikussion zu unterstützen. Mit diesem Schreiben möchte ich Sie (knapp) über Hintergrund und Ablauf des Gesprächs sowie über die Verwendung der erhobenen Daten informieren.

Thema der Arbeit, Zielsetzung der Gruppendiskussion

Meine Dissertation befasst sich mit der Frage nach der *Gegenstandsangemessenheit* von Datenerhebungsmethoden, die im Kontext von Evaluationen an Hochschulen zum Einsatz kommen. Der Gruppendiskussion, an der Sie teilnehmen werden, kommt dabei die Rolle zu, die Frage zu klären, wie die beteiligten Personen die Angemessenheit von Datenerhebungsmethoden einschätzen.

Ablauf der Gruppendiskussion

An dem Gespräch werden insgesamt 5 Personen teilnehmen: ein/e Dozent/-in und ein/e Studierende/r, die im Rahmen der Fachdidaktik-Evaluation befragt wurden sowie ein/e Dozent/-in und ein/e Studierende/r, die nicht an der Befragung zur Fachdidaktik-Evaluation teilgenommen haben. Moderiert wir das Gespräch durch mich, Claus Stefer.

Anhand eines Leitfadens werden in dieser Expertenrunde Fragen der Auswahl von Datenerhebungsmethoden erörtert. *Eine inhaltliche Vorbereitung ist ihrerseits nicht erforderlich.*

Dauer der Gruppendiskussion

Das Interview wird ca. 60 Minuten dauern.

Aufzeichnung des Gesprächs

Das Interview wird mit Hilfe eines digitalen Audioaufnahmegerätes aufgezeichnet.

Was geschieht mit den erhobenen Daten?

- Die aufgezeichneten Daten werden wörtlich transkribiert (verschriftlicht).
- Während der Transkription erfolgt eine *Anonymisierung*, d.h.:
  - Die Namen der beteiligten Personen sowie Namen, die während des Interviews genannt werden, werden durch Decknamen ersetzt. Ausnahmen betreffen hierbei u.U. Personen, die eine zentrale Bedeutung für die Fachdidaktik-Evaluation haben und auch bei der Ersetzung des Namens unmittelbar erkennbar blieben, bspw. Prof. Laging.
  - Allgemeine Informationen, die eine Identifikation von Personen ermöglichen – etwa Aussagen über einen konkreten Fachbereich o.ä. –, werden ebenfalls verschleiert, i.d.R. durch Umschreibungen. Auch hierbei kann allerdings unter den oben genannten Umständen von einer Anonmyisierung abgesehen werden.
- *Vertraulichkeit*: Die Daten müssen im Rahmen der Promotion mindestens den Gutachtern zugänglich gemacht werden. Sämtliche Informationen werden jedoch selbstverständlich vertraulich behandelt, d.h. dass sie prinizpiell nicht als Einzeldaten veröffentlicht oder weitergegeben werden. Jede Weitergabe erfolgt außerdem *ausschließlich* in transkribierter und anonymisierter Form.
- Schließlich werden die Daten *ausgewertet*, d.h. die Interviews werden mit Hilfe von Methoden der qualitativen Sozialforschung (interpretative Verfahren) bearbeitet, um die Forschungsfrage zu beantworten. Dabei wird in erster Linie nach Faktoren gesucht, die Hinweise auf die Konstruktion von Gegenstandsangemessenheit liefern können.

Zum Abschluss möchte ich mich nochmals herzlich für Ihre Bereitschaft, an der Gruppendiskussion teilzunehmen, bedanken. Falls Sie weitere Fragen haben sollten, so bitte ich Sie, sich mit mir in Verbindung zu setzen. Sie erreichen mich entweder per E-Mail unter der Adresse stefer@uni-marburg.de oder telefonisch – am besten vormittags – unter der Nummer 06421/28-24924.

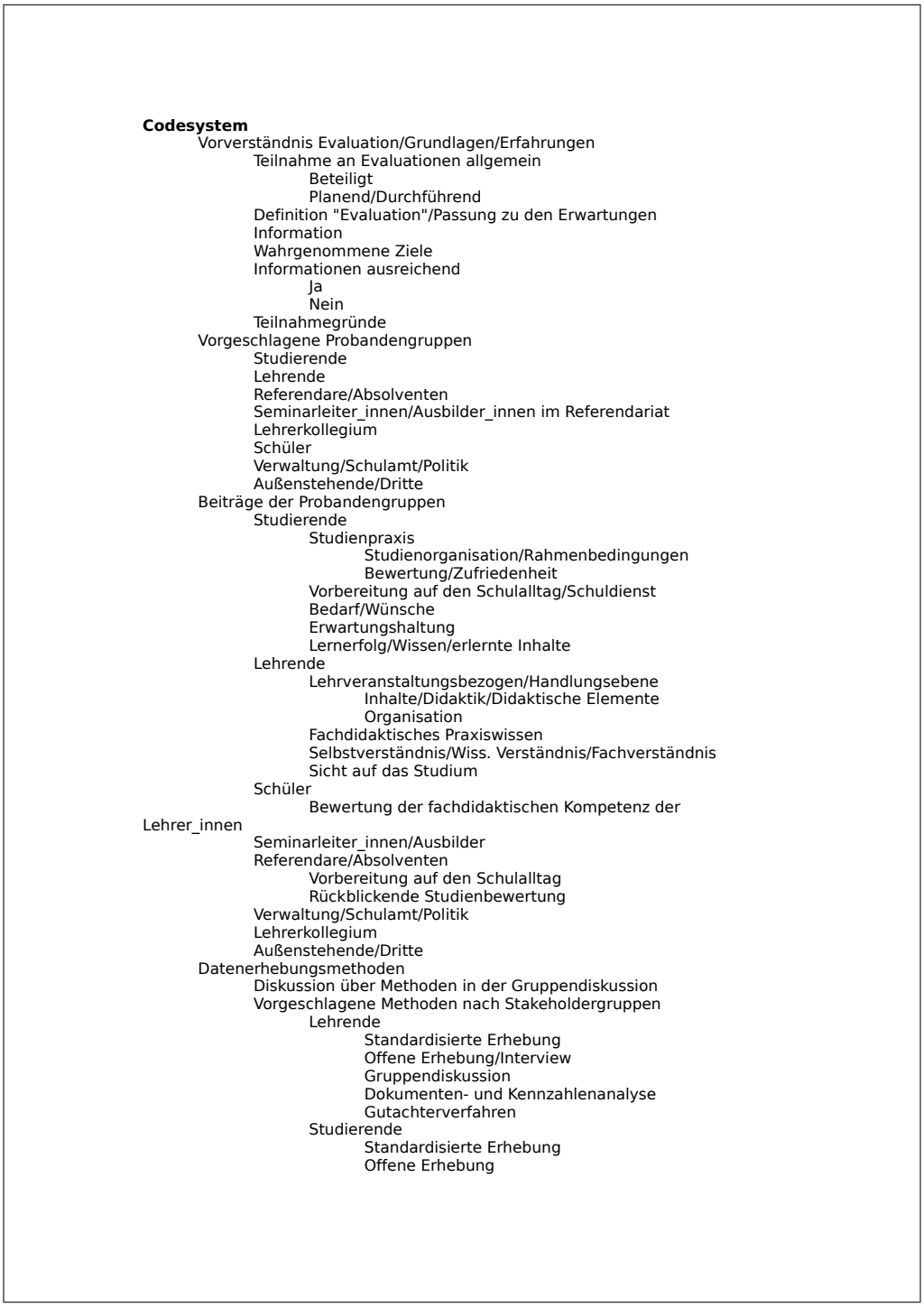
Mit freundlichen Grüßen,
Claus Stefer

A.7. Transkriptionsregeln

Transkriptionsregeln

- Die Transkription erfolgt wörtlich (nicht lautsprachlich oder zusammenfassend). Dialekte werden nicht mit transkribiert.
- Die interviewende Person wird durch ein „I“, die befragte Person durch ein „B“, gefolgt von ihrer Kennnummer, gekennzeichnet (etwa „B4“).
- Sprache und Interpunktion werden geglättet und an die Schriftsprache angenähert.
- Füllwörter wie "Äh" und "Ähm" werden nicht mit transkribiert.
- Worte, die mit dem Begriff "Fachdidaktik" beginnen, werden, um die Lesbarkeit zu verbessern, mit Bindestrich geschrieben (Fachdidaktik-Ausbildung, Fachdidaktik-Evaluation, Fachdidaktik-Verantwortliche etc.).
- Angaben, die Rückschlüsse auf eine befragte Person ermöglichen würden, werden anonymisiert, d.h. durch Decknamen, Umschreibungen oder Neutralisierungen des grammatischen Geschlechts ersetzt.
- Besonders betonte Begriffe werden durch Unterstreichungen markiert.
- Wird ein Wort nicht exakt verstanden, so wird die wahrscheinlichste Deutung in geschweiften Klammern notiert.
- Einwürfe der jeweils anderen Person werden in Klammern gesetzt.
- Zustimmende bzw. bestätigende Lautäußerungen (Mhm, Aha etc., auch "Ja") sowie Worte, die keine inhaltliche Aussage transportieren, sondern Teil der Technik des aktiven Zuhörens sind, werden nicht mit transkribiert, sofern sie den Redefluss der befragten Person nicht unterbrechen. Ausnahme: Wird mehr als eine Person befragt und drückt einer der gerade nicht sprechenden Probanden seine Zustimmungen zu einer Äußerungen durch eine solche Lautäußerung aus, wird sie gemäß den Regeln zur Notation von Einwürfen festgehalten.
- Lautäußerungen der befragten Person, welche die Aussage unterstützen oder verdeutlichen (etwa Lachen oder Seufzen), werden in Klammern notiert.
- Verständnishinweise werden in eckige Klammern gesetzt (etwa [hat Dokumente vorliegen, durchsucht sie]).
- Worte und Anfänge von Nebensätzen, die durch die Sprecherin/ den Sprecher sofort korrigiert werden, werden nur in der korrigierten Form niedergeschrieben.
- Werden Worte während des Sprechens abgebrochen und nicht weiter korrigiert oder vervollständigt, so wird der gesprochene Wortteil notiert. An der Stelle, an welcher der Redefluss abbricht, wird ein Bindestrich notiert (z.B. „Auswert–“).
- Werden Sätze inhaltlich nicht zu Ende geführt, so wird dies durch Auslassungspunkte gekennzeichnet (Das wäre mir ein zu...in dem Modell war das so, aber mir wäre das Modell dann zu ungenau.).
- Unterbrechungen werden mit kurzer Angabe des Grundes in eckigen Klammern notiert ([Kurze Unterbrechung durch einen Telefonanruf, der unmittelbar wieder beendet wurde.]).
- Pausen im Redefluss werden nicht notiert.

A.8. Finales Kategoriensystem



Codesystem
- Vorverständnis Evaluation/Grundlagen/Erfahrungen
  - Teilnahme an Evaluationen allgemein
    - Beteiligt
    - Planend/Durchführend
  - Definition "Evaluation"/Passung zu den Erwartungen
  - Information
  - Wahrgenommene Ziele
  - Informationen ausreichend
    - Ja
    - Nein
  - Teilnahmegründe
- Vorgeschlagene Probandengruppen
  - Studierende
  - Lehrende
  - Referendare/Absolventen
  - Seminarleiter_innen/Ausbilder_innen im Referendariat
  - Lehrerkollegium
  - Schüler
  - Verwaltung/Schulamt/Politik
  - Außenstehende/Dritte
- Beiträge der Probandengruppen
  - Studierende
    - Studienpraxis
      - Studienorganisation/Rahmenbedingungen
      - Bewertung/Zufriedenheit
    - Vorbereitung auf den Schulalltag/Schuldienst
    - Bedarf/Wünsche
    - Erwartungshaltung
    - Lernerfolg/Wissen/erlernte Inhalte
  - Lehrende
    - Lehrveranstaltungsbezogen/Handlungsebene
      - Inhalte/Didaktik/Didaktische Elemente
      - Organisation
    - Fachdidaktisches Praxiswissen
    - Selbstverständnis/Wiss. Verständnis/Fachverständnis
    - Sicht auf das Studium
  - Schüler
    - Bewertung der fachdidaktischen Kompetenz der Lehrer_innen
  - Seminarleiter_innen/Ausbilder
  - Referendare/Absolventen
    - Vorbereitung auf den Schulalltag
    - Rückblickende Studienbewertung
  - Verwaltung/Schulamt/Politik
  - Lehrerkollegium
  - Außenstehende/Dritte
- Datenerhebungsmethoden
  - Diskussion über Methoden in der Gruppendiskussion
  - Vorgeschlagene Methoden nach Stakeholdergruppen
    - Lehrende
      - Standardisierte Erhebung
      - Offene Erhebung/Interview
      - Gruppendiskussion
      - Dokumenten- und Kennzahlenanalyse
      - Gutachterverfahren
    - Studierende
      - Standardisierte Erhebung
      - Offene Erhebung

- Gruppendiskussion
- Dokumenten- und Kennzahlenanalyse
- Gutachterverfahren
- Längerfristige Verfahren

Referendare
- Standardisierte Erhebung
- Offene Erhebung
- Gruppendiskussion
- Dokumenten- und Kennzahlenanalyse
- Gutachterverfahren

Seminarleiter_innen
- Standardisierte Erhebung
- Offene Erhebung
- Gruppendiskussion
- Dokumenten- und Kennzahlenanalyse
- Gutachterverfahren

Schüler
- Standardisierte Erhebung
- Offene Erhebung
- Gruppendiskussion
- Dokumenten- und Kennzahlenanalyse
- Gutachterverfahren

Verwaltung/Schulamt/Politik
- Standardisierte Erhebung

Lehrerkollegium

Außenstehende/Dritte

Bewertung der selbst vorgeschlagenen Methoden

Standardisierte Erhebung
- positiv
- negativ
- keine klare Bewertung

Offene Erhebung
- positiv
- negativ
- keine klare Bewertung

Gruppendiskussion
- positiv
- negativ
- keine klare Bewertung

Dokumenten- und Kennzahlenanalyse
- positiv
- negativ
- keine klare Bewertung

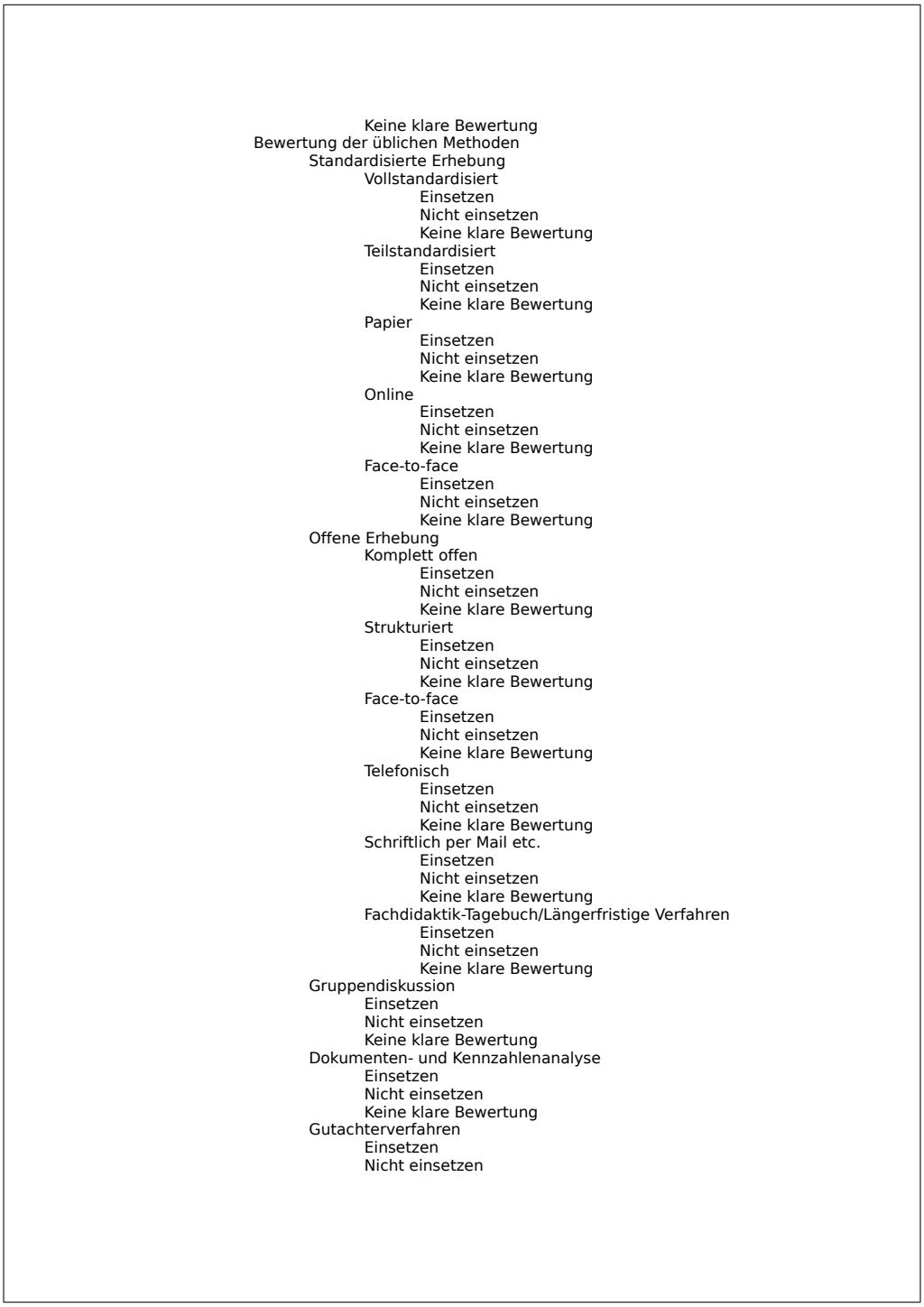
Gutachterverfahren
- positiv
- negativ
- keine klare Bewertung

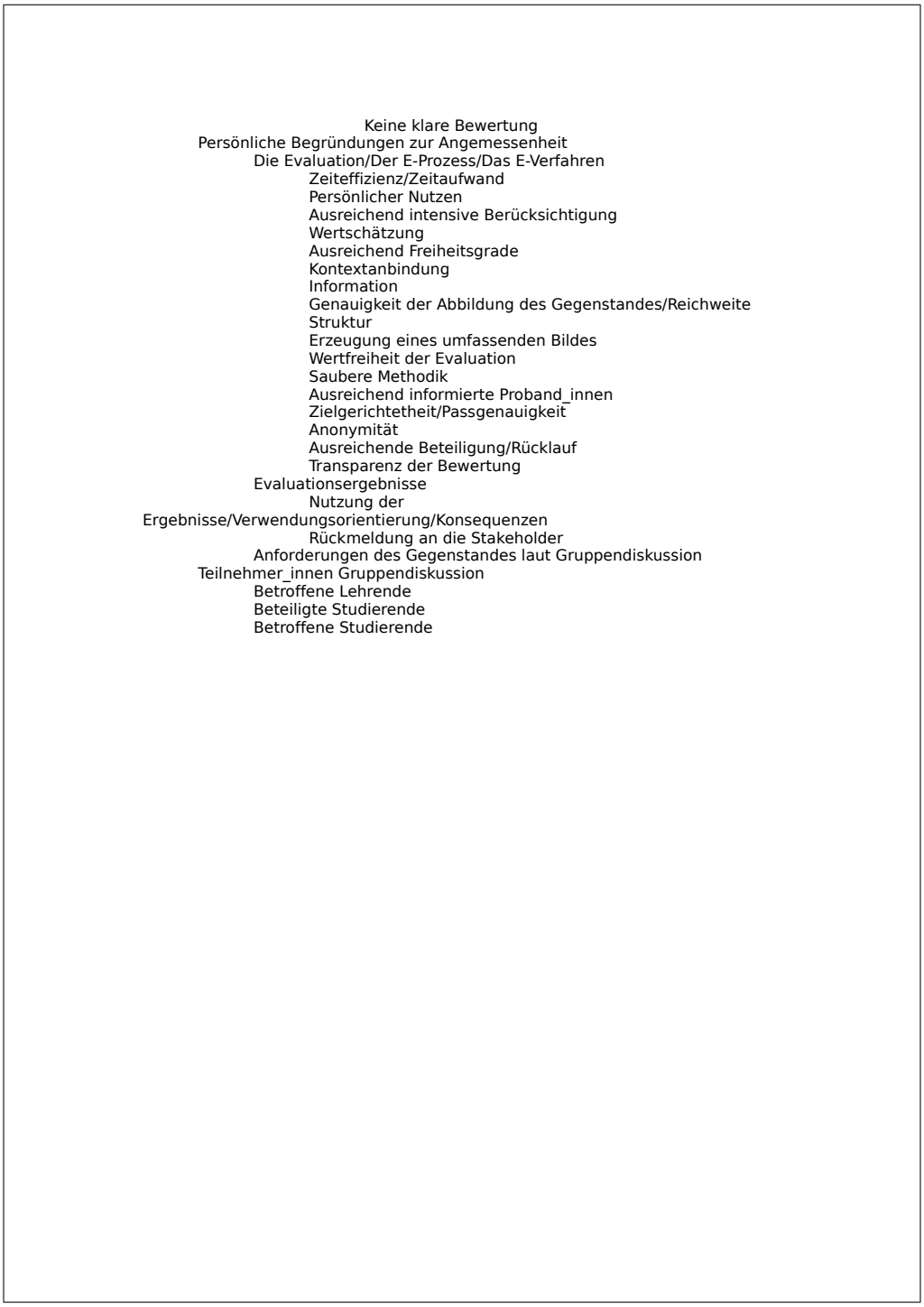
Bewertung der tatsächlich eingesetzten Methoden

Dokumenten- und Kennzahleneanalyse
- Einsetzen
- Nicht einsetzen
- Keine klare Bewertung

Offene Interviews
- Einsetzen
- Nicht einsetzen
- Keine klare Bewertung

Online-Fragebogen
- Einsetzen
- Nicht einsetzen


Keine klare Bewertung
Bewertung der üblichen Methoden
Standardisierte Erhebung
Vollstandardisiert
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Teilstandardisiert
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Papier
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Online
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Face-to-face
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Offene Erhebung
Komplett offen
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Strukturiert
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Face-to-face
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Telefonisch
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Schriftlich per Mail etc.
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Fachdidaktik-Tagebuch/Längerfristige Verfahren
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Gruppendiskussion
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Dokumenten- und Kennzahlenanalyse
Einsetzen
Nicht einsetzen
Keine klare Bewertung
Gutachterverfahren
Einsetzen
Nicht einsetzen


Keine klare Bewertung
Persönliche Begründungen zur Angemessenheit
Die Evaluation/Der E-Prozess/Das E-Verfahren
Zeiteffizienz/Zeitaufwand
Persönlicher Nutzen
Ausreichend intensive Berücksichtigung
Wertschätzung
Ausreichend Freiheitsgrade
Kontextanbindung
Information
Genauigkeit der Abbildung des Gegenstandes/Reichweite
Struktur
Erzeugung eines umfassenden Bildes
Wertfreiheit der Evaluation
Saubere Methodik
Ausreichend informierte Proband_innen
Zielgerichtetheit/Passgenauigkeit
Anonymität
Ausreichende Beteiligung/Rücklauf
Transparenz der Bewertung
Evaluationsergebnisse
Nutzung der
Ergebnisse/Verwendungsorientierung/Konsequenzen
Rückmeldung an die Stakeholder
Anforderungen des Gegenstandes laut Gruppendiskussion
Teilnehmer_innen Gruppendiskussion
Betroffene Lehrende
Beteiligte Studierende
Betroffene Studierende

A.9. Übersichtstabellen

Erhebungseinheit	Interviewbezeichnung
Beteiligte Studierende	B2, B5, B7, B8, B9, B11
Betroffene Studierende	B13, B14, B15, B16, B17, B18
Beteiligte Lehrende	B1, B3, B4, B6, B10, B12
Betroffene Lehrende	B19, B20, B21, B22, B23, B24
Planende	B25, B26
Gruppendiskussion	B27, B28, B29

Tabelle A.1.: Übersicht über die Verortung der Interviews in den einzelnen Erhebungseinheiten

Probandengruppe	Befragte Studierende	Befragte Lehrende	Gesamt
Studierende	12	12	24
Lehrende	11	12	23
Referendare/Absolventen	8	8	16
Seminarleiter/-innen, Ausbilder/-innen im Referendariat	0	5	5
Lehrerkollegium	0	4	4
Schüler	2	1	3
Verwaltung, Schulamt, Politik	1	1	2
Außenstehende, Dritte	1	1	2

Tabelle A.2.: Anzahl der Vorschläge zu befragender Probandengruppen (Zeilen) nach Statusgruppe

Probandengruppe	Methodenmix	Standardisierte Verfahren	Offene Verfahren
Studierende	10	10	3
Lehrende	13	4	5
Referendare/Absolventen	7	4	4

Tabelle A.3.: Vorgeschlagene Methoden (Spalten) nach Stakeholdergruppen, für die sie vorgeschlagen wurden (Zeilen)

	Anzahl Vorschläge Studierende	Anzahl Vorschläge Lehrende
M für Studierende und Lehrende	4	6
S für Studierende und Lehrende	1	3
O für Studierende und Lehrende	1	1
S für Studierende, M für Lehrende	3	-
S für Studierende, O für Lehrenden	1	1
O für Studierende, M für Lehrende	-	1
M für Studierende, O für Lehrende	1	-
Nur Studierende befragen, S	1	-

Tabelle A.4.: Verteilung der vorgeschlagenen Methodenkombinationen (Zeilen) nach vorschlagenden Probanden/-innengruppen. S = Standardisierte Datenerhebungsmethoden, O = Offene Datenerhebungsmethoden, M = Methodenmix

	Geeignet	Ungeeignet	Keine klare Bewertung
Dokumenten- und Kennzahlenanalyse	18	3	3
Online-Fragebögen	18	3	2
Offene Interviews	22	-	2

Tabelle A.5.: Bewertung der Eignung der in der tatsächlich durchgeführten Evaluation eingesetzten Datenerhebungsmethoden

	Geeignet	Ungeeignet	Keine klare Bewertung
Standardisierte Verfahren	21	-	3
Offene Verfahren	21	-	3
Dokumenten- und Kennzahlenanalyse	11	6	7
Gruppendiskussion	9	6	9
Gutachterverfahren	8	8	8

Tabelle A.6.: Bewertung der Eignung der üblicherweise in Lehrevaluationen an Hochschulen eingesetzten Datenerhebungsmethoden

Erklärung

Ich erkläre, dass ich die Dissertation unter Verwendung keiner anderen als der angegebenen Hilfsmittel angefertigt habe.

(Claus Stefer)