

SKRIPT

# Umgang mit Evaluationsergebnissen

# Inhaltsverzeichnis

Einleitende Informationen.....	3
Modul 1: Auftakt & Einführung.....	5
Evaluation - eine Erfolgsgeschichte .....	5
Evaluation – zu Etymologie und Bedeutungsinhalt.....	6
Mögliche Handlungsfelder von Evaluation .....	8
Funktionen und Charakteristika von Evaluation .....	9
Evaluation an Hochschulen.....	11
Evaluation der Lehre.....	11
Modul 2: Lehrveranstaltungsevaluation (LVE) .....	16
Aktuelle Praxis der LVE .....	16
Zur Aussagekraft studentischer LVE .....	20
Evaluationsinstrumente (schriftlich und mündlich).....	25
Modul 3: Umgang mit Evaluationsergebnissen - Desk Research .....	28
Aspekte der Bewertung .....	30
Formen der Nutzung .....	32
Modul 4: Umgang mit Evaluationsergebnissen in der (LVE-)Praxis.....	37
Literatur.....	38

# Einleitende Informationen

Dieses Dokument dient der inhaltlichen Vorbereitung des Blended-Learning-Seminars „Umgang mit Evaluationsergebnissen“. Dabei werden die Themen in der gleichen Reihenfolge vorgestellt, wie sie auch im begleitenden Trainer:innen-Guide empfohlen werden<sup>1</sup>.

Modul	Thema	Unterthemen
Modul 1	Auftakt & Einführung	Kennenlernen: Wer ist alles im Blended-Learning-Seminar?  Erwartungen: Was erwartet ihr von diesem-Blended Learning-Seminar?  Organisatorisches: Wie funktioniert das Blended-Learning-Seminar?  Einführung: Evaluation an Hochschulen  Evaluation – eine Erfolgsgeschichte  Evaluation – zu Etymologie und Bedeutungsinhalt

<sup>1</sup> Dieses Werk „Umgang mit Evaluationsergebnissen“ ist eine Open Educational Ressource (OER) aus dem hochschuldidaktischen Verbundprojekt QUADIS – Qualität digital gestützter Lehre an bayerischen Hochschulen steigern (<https://quadis.profilehreplus.de/>). Es wurde 2024 von Dr. Elisa Knödler und Dr. Anna von Werthern an der Ludwig-Maximilians-Universität in München erstellt. Das gesamte Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Unbenommen davon und sofern nicht anders angegeben dürfen das Gesamtwerk und alle seine Teile im Rahmen der Creative-Commons-Lizenz „Namensnennung 4.0 International“ (CC BY 4.0) genutzt werden. Der Lizenztext ist abrufbar unter <https://creativecommons.org/licenses/by/4.0/legalcode.de>. Das Werk ist abrufbar unter [URL des Einzelmaterials bei der vhb].

		<p>Mögliche Handlungsfelder von Evaluation</p> <p>Funktionen und Charakteristika von Evaluation</p> <p>Evaluation an Hochschulen</p> <p>Wrap Up</p>
Modul 2	Lehrveranstaltungsevaluation (LVE)	<p>Lehrveranstaltungsevaluation (LVE) im Überblick</p> <p>Aktuelle Praxis der LVE</p> <p>Zur Aussagekraft studentischer LVE</p> <p>Evaluationsinstrumente (schriftlich und mündlich)</p> <p>Wrap Up &amp; Feedbackrunde</p>
Modul 3	Umgang mit Evaluationsergebnissen - Desk Research	<p>Umgang mit Evaluationsergebnissen – Desk Research</p> <p>Aspekte der Bewertung</p> <p>Formen der Nutzung</p> <p>Wrap Up</p>
Modul 4	Umgang mit Evaluationsergebnissen in der (LVE-)Praxis	<p>Von der Theorie in die (LVE-)Praxis - Evaluationsergebnisse und dann?</p> <p>„In einer idealen Welt...“</p> <p>Peer-to-Peer-Austausch</p> <p>Wrap Up und Abschluss</p>

# Modul 1: Auftakt & Einführung

## Evaluation - eine Erfolgsgeschichte

“[E]valuation has become a mantra of modernity.” (Pawson & Tilley 1997b, S. 2)

Der Evaluation wird weltweit ein „unvergleichliche[r] Siegeszug“ attestiert (Wittmann 2009, S. 59). Dieser lässt sich nicht nur anhand wissenschaftlicher Publikationen im Rahmen der Forschung über Evaluation, von Praxisberichten aus unterschiedlichen Anwendungsfeldern der Evaluation, Gründungen von wissenschaftlichen Evaluationsgesellschaften sowie dem Aufbau von Evaluationsabteilungen in internationalen und nationalen Organisationen nachvollziehen (vgl. ebd. 2009, S. 59; Hennefeld 2015, S. 7), sondern spiegelt sich dieser auch im Bereich des Alltäglichen wider: Sucht man beispielsweise online mithilfe von einer Suchmaschine den Begriff Evaluation, erhält man binnen kürzester Zeit knapp 500 Millionen Treffer (vgl. Google.de 2017).

Die Vielfältigkeit der angebotenen Definitionen lässt eine erste Erkenntnis zu: Eine häufige Nutzung des Begriffs steht unklaren respektive enorm heterogenen Bedeutungszuschreibungen gegenüber. Herausfordernd ist in diesem Zusammenhang, dass sich die Begriffsvielfalt nicht nur auf den Diskurs im Rahmen der einschlägigen Fachliteratur beschränkt, sondern sich der unterschiedliche Sprachgebrauch auch auf (teilweise) ähnliche Konzepte aus der Alltagssprachlichen beziehungsweise öffentlichen Diskussion bezieht (vgl. Brandt 2009, S. 69).

„In der deutschen Sprache breitet sich das Wort „Evaluation“ – besonders im Österreichischen auch in der Variante „Evaluierung“ – in den letzten zwanzig Jahren aus. Auch das Verb „evaluieren“ kann nicht mehr – wie vielleicht in den 1980er Jahren – nahezu exklusiv dem Wortschatz eines kleinen Kreises von Fachpersonen zugerechnet werden. Diese und davon abgeleitete Terme haben sich im Sprachgebrauch weit verbreitet und werden in Parlamentsdebatten, Tageszeitungen, Talkshows mit politischen oder gesellschaftlichen Inhalten, Radio- und Fernsehnachrichtensendungen genutzt“ (Beywl & Widmer 2009, S. 13).

Auch die DeGEval (Gesellschaft für Evaluation 2008) konstatiert ein Konglomerat von Definitionen:

„Aufgrund der Vielfalt von Evaluationen, ihrer Anwendungsbereiche, Aufgaben und zugrunde liegenden Konzepte, gibt es darüber hinaus zahlreiche Definitionen, die sich jeweils in einzelnen Aspekten unterscheiden. Der Begriff „Evaluation“ wird zunehmend auch in der Alltagssprache verwendet, jedoch häufig nicht in diesem professionellen Sinn, sondern es wird damit lediglich ausgedrückt, dass etwas in irgendeiner Weise geprüft, bewertet oder beurteilt wurde (Kromrey 2001)“ (ebd. 2008).

Grundsätzlich wurde und wird Evaluation in den USA bzw. Nordamerika als Begriff und Konzept geprägt und die Diskurse im Feld der Evaluation werden noch immer stark durch die nordamerikanische Evaluationscommunity bestimmt

Dennoch hat sich mittlerweile auch in Europa ein lebendiger Diskurs entwickelt (vgl. z. B. Giel 2013, Hense et al. 2013, Dahler-Larsen 2012; Widmer & DeRocchi 2012; Stockmann & Meyer 2010; Vedung 2010; Brandt 2009) und es ist festzuhalten, dass die Entwicklung der Evaluation stets von Land bzw. Kontinent und der dort vorzufindenden Situation geprägt ist (Mark et al. 2006) (von Werthern, 2020).

## Evaluation – zu Etymologie und Bedeutungsinhalt

Der Evaluationsbegriff wird in unterschiedlichen Anwendungskontexten divergent verwendet. Neben wissenschaftlichen Definitionen reihen sich Alltagsbedeutungen (Brandt, 2009), weshalb eine Begriffseingrenzung (Hochschulrektorenkonferenz, 1999) für den jeweiligen Anwendungskontext sinnvoll ist (Knödler, 2018).

Evaluation ist ein seit jeher ein bestehender Allerweltsvorgang, unter dem die Bewertung eines Prozesses oder Gegenstandes anhand bestimmter Kriterien verstanden wird – mit dem Ziel, daraus eine Entscheidung abzuleiten. Wird Evaluation jedoch als systematisches Verfahren verwendet, welches vorgegebenen wissenschaftlichen Regeln folgt, dann ist sie vielmehr ein modernes Phänomen, das mit politischer und sozialwissenschaftlicher Entwicklung einhergeht (Knödler, 2018).

Arbeitet oder lehrt man im Feld der Evaluation, ist es sinnvoll, sich eine Art Arbeitsdefinition zurechtzulegen, an der man sich in seiner Tätigkeit orientiert und

die man dann auch so weitergeben kann. Gut geeignet sind dafür z. B. die Definitionen der DeGEval (Gesellschaft für Evaluation), anbei zwei Annäherungen an den Evaluationsbegriff, die unterschiedliche Schwerpunkte setzen:

„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 quantitativen Daten beruhen“ (DeGEval 2008, S. 15).

„Evaluation kommt sowohl international als auch im deutschen Sprachraum eine zentrale Bedeutung zu. Sie ist in sehr unterschiedlichen Anwendungsfeldern etabliert, wie bspw. Entwicklungspolitik, Schulen und Hochschulen, berufliche und betriebliche Bildung, Strukturpolitik, Stadt- und Regionalentwicklung, Medien, Umwelt, Gesundheit, Forschung und Technologie, Verwaltung, Wirtschaft und Soziale Dienstleistungen. Damit geht eine Differenzierung von Evaluationsansätzen und Evaluationsmethoden einher. War Evaluation noch vor wenigen Jahrzehnten auf die Bewertung von Programmen und deren Wirkung begrenzt, so weist sie heute vielfältige Berührungspunkte mit Qualitätsmanagement und Organisationsentwicklung auf. Die Bedeutsamkeit des Themas sowie dessen ständige Ausdifferenzierung erfordert eine stetige Diskussion über Theorien, Modelle und Methoden von Personen, die über Evaluation forschen und lehren, Evaluationen durchführen und Evaluationen in Auftrag geben“ (DeGEval 2008, S. 2, zit. n. von Werthern 2011).

## Mögliche Handlungsfelder von Evaluation

Ein entscheidendes Charakteristikum von Evaluation ist, dass nahezu alles evaluiert werden kann, einschließlich einer Evaluation selbst. Gegenstände einer Evaluation können Organisationen, Personen, Produkte, Reformen, Gesetze, Maßnahmen, Projekte, Programme, Gruppen, Institutionen, Bildungssysteme, die Entwicklungshilfe oder ganze politische Systeme sein (vgl. Stockmann 2004, S. 2; von Werthern, 2020; Wittmann 2009, S. 59).



Abb. 1: Mögliche Handlungsfelder von Evaluation (eigene Darstellung)



## Funktionen und Charakteristika von Evaluation

Die fundamentale Frage nach Zielen und Funktionen von Evaluation lässt sich auf mehreren Ebenen beantworten. Zum einen besteht die Frage nach dem Ziel einer Evaluation im Einzelfall. Es wird diskutiert, was erreicht werden soll, welche Informationen zu welchem Zweck generiert werden sollen und wie man dieses Ziel erreichen kann. Auf einer anderen Ebene ist diese Frage nach den Funktionen von Evaluationen umfassender zu verstehen, im Sinne einer grundsätzlichen Sinnfrage nach den „reasons for doing evaluation“ (Mark et al. 2000, S. 12). Warum gibt es Evaluationen und was haben sie für einen Mehrwert für die Gesellschaft? Mark et al. (2000) formulieren dazu eine klare Antwort: „The ultimate goal of evaluation is social betterment [...]“ (ebd., S. 3). Diese Zielvorstellung des social betterment hebt auf eine Verbesserung der gesellschaftlichen Zustände ab:

„Social betterment refers to the improvement of social conditions. Put differently, in the context of democratic society, social betterment refers to bringing about a state that would be considered as better than the state that existed before, as judged through deliberation and by public opinion [...]“ (Henry & Mark 2003, S. 295).

Vier, für die Arbeitspraxis relevante und bekannte Funktionen nach Stockmann & Meyer (2014) sind: Erkenntnis, Kontrolle, Legitimation und Entwicklung.

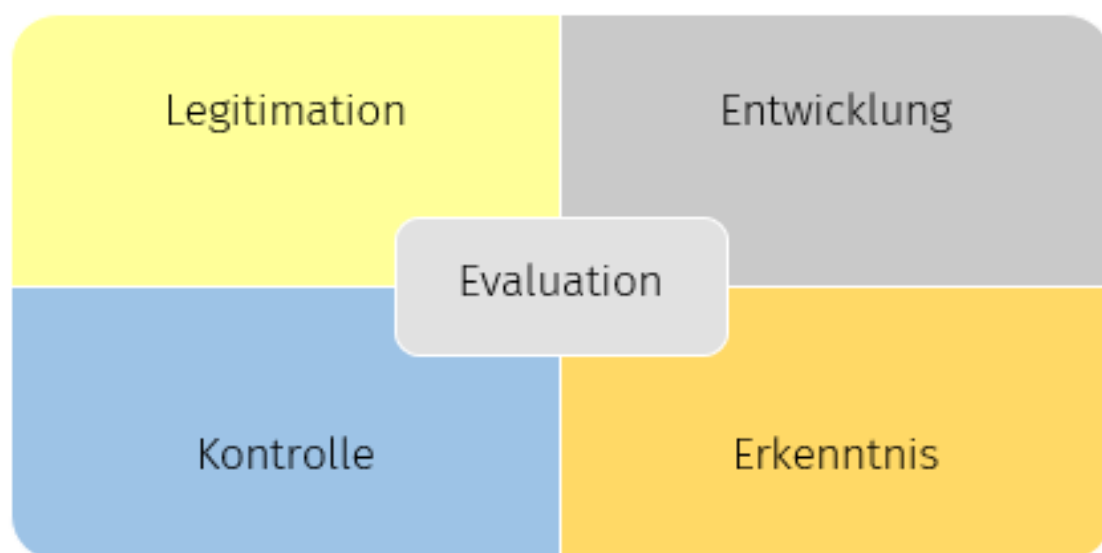


Abb. 2: Die vier Leitfunktionen nach Stockmann & Meyer 2014 (eigene Darstellung)

Eine weitere Unterscheidung wird häufig zwischen formativer und summativer Evaluation getroffen. Eine formative Evaluation ist am Prozess bzw. an der Gestaltung des Evaluationsgegenstandes orientiert und zielt vorrangig auf dessen Verbesserungen ‚on the way‘ ab. Die unterschiedlichen Akteursgruppen können mittels einer solch begleitenden Evaluation unterstützt werden, den Nutzen des Evaluationsgegenstandes immer wieder zu überprüfen und die Ressourcen auf unterschiedlichen Ebenen (beispielsweise personelle und monetäre Ressourcen) möglichst zielführend einzusetzen. Eine summative Evaluation zieht eine zusammenfassende Bilanz. Häufig sind an diese Bilanz grundlegende Entscheidungen über den Evaluationsgegenstand geknüpft. Neben diesen beiden Zielsetzungen ist auch ein Erkenntnisgewinn ohne unmittelbare Verwendungsabsicht ein weiterer möglicher Zweck von Evaluation. Dieses Ziel kann unter Rückgriff auf sowohl formative als auch summative Evaluation verfolgt werden. Wendet man den Blick dem praktischen Evaluationskontext zu, so hält die DeGEval (Gesellschaft für Evaluation 2008) im Rahmen ihres Standards N2 Hinweise zur Klärung der Evaluationszwecke bereit (von Werthern, 2020):

„Die Zwecke einer Evaluation sollen frühzeitig angesprochen und im Verlauf einer Evaluation ausgehandelt, schriftlich fixiert und so weit wie möglich den Beteiligten und Betroffenen bekannt gemacht werden, damit diese sich in die Evaluationsplanung einbringen können. Die Tätigkeit der beauftragenden Person bzw. Institution und des Evaluationsteams wird durch klare Zwecksetzungen erleichtert, da sie damit eine gute Grundlage haben, um die Verbreitung der Evaluationsergebnisse und deren Nutzung vorzubereiten“ (ebd., S. 24).

# Evaluation an Hochschulen

„Die Evaluation [an Hochschulen, Anm. d. Verf.] steht von verschiedenen Seiten unter Erwartungsdruck. Hochschulleitungen bzw. Fakultäts-, Departements- oder Abteilungsleitungen, strategische Programmleitungen, operative Studienleitungen, Lehrende oder Teilnehmende artikulieren teilweise unterschiedliche Ansprüche, stellen aber alle jeweils die Frage nach dem Nutzen einer bestimmten Evaluation. Dies ist durchaus berechtigt und nicht zuletzt durch Kostenüberlegungen begründet. Mit Evaluation ist meist ein gewisser Aufwand für alle Seiten verbunden“ (Friedrich, V.; Fischer, A. (Hg.; 2015, S. 7)).

Evaluationsverfahren stehen als Instrument der Qualitätssicherung im deutschen Hochschulsystem seit den 1980er Jahren auf der Tagesordnung und sind, durch die rechtliche Verankerung in den Hochschulgesetzen der meisten Bundesländer, an Hochschulen in einem breiten Ausmaß zu finden (Gollwitzer & Scholtz, 2003; Knödler, 2018). Dabei wird Evaluation nicht selten mit Qualitätssicherung gleichgesetzt. Während die Qualitätssicherung jedoch ein auf Dauer initiiertes System darstellt, ist Evaluation in der Regel ein zeitlich begrenztes und situativ verankertes (Bülow-Schramm, 2000) Verfahren. Qualitätssicherung kann folglich als System der fortgeführten Evaluation verstanden werden, Evaluation hingegen als Bestandteil des Qualitätssicherungssystems.

Das primäre Ziel der Evaluation (an Hochschulen) ist das Analysieren von Stärken und Schwächen und Herausarbeiten von deren Ursachen, um handlungsorientierte Empfehlungen zur Qualitätsverbesserung zu liefern. Evaluation kommt somit bei der Generierung von Handlungs- und Erklärungswissen zum Einsatz (Knödler, 2018).

## Evaluation der Lehre

Evaluationen an Hochschulen untersuchen definierte (Studien-)Programme und Maßnahmen im Sinne „praxisnaher Forschung“ (Voges, 2017, S. 184) und bewerten diese auf der Basis programm- und maßnahmenspezifischer Ansprüche (Knödler, 2018).

Bei einer systematischen Behandlung der Thematik „Evaluation der Lehre“ müssen aufgrund der spezifischen Organisationsstruktur deutscher Hochschulen drei Ebenen

differenziert werden: (1) die Organisation Hochschule in ihrer Grundgesamtheit, (2) die verschiedenen Fachbereiche, (3) die einzelnen Lehrveranstaltungen und Lehrende (el Hage, 1996a; Knödler, 2018; Reissert, 1992).

Bei der Evaluation der Organisation Hochschule im Gesamten (1) werden Leistungen in der Lehre und Forschung auf administrativer Ebene (z.B. Personaleinsatz, Mittelverwendung) erfasst. Die Evaluation fokussiert sich dabei vor allem auf die Darstellung der Lehr- und Studiensituation. Soll ein ganzer Fachbereich oder Studiengang (2) als Organisationseinheit (Makroebene) evaluiert werden, ist sowohl die administrative als auch die organisatorische Ebene zu untersuchen. Bezüglich der Lehre sind organisatorische Aspekte wie allgemeine Studienbedingungen sowie Daten zu erfassen, welche sich auf Auswirkungen und Gesamtsituation von Lehre beziehen. Die auf der Mikroebene durch studentische Lehrveranstaltungskritik stattfindende Lehrevaluation (3) beschäftigt sich mit der Qualität einzelner Lehrenden oder Lehrveranstaltungen und lässt sich hinsichtlich des Zeitpunktes und der Instrumente differenzieren (Döring, 2002; el Hage, 1996a).

### Zeitpunkt der Lehrevaluation

Die Unterscheidung bezüglich des Zeitpunktes zwischen formativer (während des Prozesses) und summativer (am Ende des Prozesses) Evaluation stammt ursprünglich von Scriven (1991a). Die *formative* Evaluation hat die Aufgabe, Stärken und Schwächen von Maßnahmen aufzudecken und dient der Unterstützung, Verbesserung und Weiterentwicklung laufender Projekte und Programme. Anhand prozessgeleiteter Evaluationen kann von Beginn an überprüft werden, ob Konzepte geeignet und realisierbar sind oder einer Überarbeitung bedürfen. Die Ergebnisse bleiben meist innerhalb der Institution und ihrer Programmentwicklung (Knödler, 2018). Die *summative* Evaluation liefert Daten über eine abgeschlossene Maßnahme und hat die Funktion der Bilanzierung, Bewertung sowie Überprüfung von Erfolg und Wirkung. So kann die Erreichung der Ziele nachgewiesen werden, um über mögliche Veränderungen zu entscheiden (Voges, 2017). Die Ergebnisse werden auch Interessenten außerhalb der Institution zur Verfügung gestellt und tragen zur

verbesserten Verwendung und zum besseren Verständnis des Programms bei (Knödler, 2018).

Eine Kombination beider Evaluationsformen ist möglich und häufige Praxis (Knödler, 2018; Scriven, 1972; Stamm, 2003).

### **Instrumente der Lehrevaluation**

In Bezug auf die Instrumente ist zwischen Lehrberichten, sowie interner und externer Evaluation zu differenzieren (Knödler, 2018).

*Lehrberichte:* Zur Erfassung der Lehrleistung werden an Hochschulen seit den 1990er Jahren von Seiten der Bundesländer verpflichtend sogenannte Lehrberichte eingesetzt, welche die aktuelle Lehrsituation auf Fachbereichs- und Hochschulebene darstellen (Bülow-Schramm, 2000). Dies erfolgt in standardisierter und klar gegliederter Form (Michalk, 2007) auf der Grundlage gesammelter und analysierter Daten (Bargel & el Hage, 2000; Webler et al., 1993). Die eigenverantwortliche Erstellung der Lehrberichte sowie das Ziel der Verbesserung von Kommunikations- und Entscheidungsprozessen und Akzeptanz hat zur Folge, dass die Hochschulen, beziehungsweise Fachbereiche selbst über die Konzeptentwicklung und Zusammenstellung der Kriterien entscheiden (Bargel & el Hage, 2000). Statistische Daten (Kennzahlen) und deskriptive Aussagen zu Studien-, Lehr-, Lern- und Rahmenbedingungen (Ausstattung, Arbeitsplätze, fachübergreifende Lehr- und Qualifikationsangebote, Öffnungszeiten), werden dabei durch Bewertungen von Lehrenden, Absolvent:innen und Studierenden ergänzt (Michalk, 2007; Webler et al., 1993). Die Hinzunahme von Leitbildern, Ziele der Ausbildung sowie Maßnahmen zur Qualitätssicherung in der Lehre sind möglich (Bargel & el Hage, 2000; el Hage, 1996a). Für die Erhebung der Informationen eignen sich Interviews, Dokumentenanalysen oder Fragebögen. Als Bezugspunkt bei der Dateninterpretation dienen Zeitreihenanalysen oder Vergleiche mit anderen Hochschulen oder Fachbereichen (Bargel & el Hage, 2000; Knödler, 2018).

Lehrberichte sind aufgrund der Selbstevaluation auf eine starke und aktive Beteiligung der Fachbereiche/Hochschule angewiesen (Bargel & el Hage, 2000). Die

Wirksamkeit der Lehrberichte und ihr Stellenwert sind für die Verbesserung der Lehrqualität bislang gering. Dies liegt mitunter daran, dass die Adressat:innen, Ziele sowie Konsequenzen, die aus den Berichten abzuleiten sind, in den Verordnungen und Gesetzen der Länder nicht klar definiert werden, was eine sinnvolle Verknüpfung der Lehrberichte mit anderen Qualitätssicherungsverfahren verhindert (Michalk, 2007). Hinzu kommt das Fehlen einer zentralen Koordinationsinstanz, welche sich der Auswertung und inhaltlichen Rückkopplung an die Fachbereiche verpflichtet (Bargel & el Hage, 2000).

*Interne Evaluation:* Interne Evaluationen werden auch Selbstevaluation genannt. Dabei untersuchen die Mitglieder das Handeln ihrer eigenen Organisation (Weiss, 1974). Interne Evaluator:innen verfügen über organisationsinternes Wissen (z.B. über Besonderheiten der Organisation) und kennen die Strukturen, werden jedoch nicht selten wegen mangelnder Objektivität kritisiert. Die interne Evaluation umfasst die Bestandsaufnahme und systematische Analyse von Studiums und Lehre (Reissert & Carstensen, 1998). Dabei dokumentiert ein Fachbereich in Eigenregie die aktuelle Situation, zukünftige strategische Ausrichtungen (Profil) sowie das bisher Geleistete. Dies geschieht in Form einer Selbstdokumentation (in der Literatur häufig auch Lehrbericht genannt), in der Stärken und Schwächen ebenso wie Maßnahmen für eine zukünftige Profilierung dargestellt werden. Dafür werden Instrumente wie die Dokumentenanalyse, moderierte Gruppengespräche, die Auswertung der amtlichen Hochschulstatistik und Studierenden- sowie Expert:innen-Befragungen eingesetzt. Absolvent:innenbefragungen liefern darüber hinaus Informationen zu beruflichen Aspekten (Verbleib, Übergang etc.; Knödler, 2018).

*Externe Evaluation:* Externe Evaluationen, auch Fremdevaluation genannt, werden von Fachleuten außerhalb der Organisation durchgeführt. Sie gelten als objektiver, da externe Evaluator:innen dem Untersuchungsobjekt distanzierter gegenüber stehen und anhand von (objektiven) Kriterien bewerten. Nicht selten werden sie deshalb zur Kontrolle und Legitimation herangezogen (Scriven, 1972). Externe Evaluationen sind daher eher für unabhängige Analysen im Dienste demokratischer Legitimierung oder gesellschaftlicher Aufklärung geeignet, für die

Programmsteuerung lassen sich hingegen interne wie auch externe Evaluationen einsetzen (Stockmann & Meyer, 2014). Die externe Evaluation ergänzt die innere Sicht in Form der Selbstdokumentation durch die Außensicht. Im Rahmen der externen Evaluation überprüft ein Gutachterteam (Peers) das Dokumentierte und führt eine (meist zweitägige) Vor-Ort-Begehung der Räumlichkeiten (Arbeitsräume, Labore, Bibliotheken etc.) der zu evaluierenden Einheit durch. Zudem werden Gruppen- und Einzelgespräche mit den unterschiedlichen Beteiligten (Studierende, Lehrende, Hochschulleitung etc.) geführt. In der Abschlussdiskussion erhält die evaluierte Einheit ein mündliches Feedback. Der schriftliche Gutachterbericht enthält Anregungen, Strategien und Empfehlungen zur Qualitätsverbesserung. Auf der Grundlage dieser Rückmeldungen nimmt der begutachtete Fachbereich in Form eines Maßnahmenkatalogs zur Umsetzung der Empfehlungen Stellung (Becker-Richter et al., 2002; Michalk, 2007; Reissert & Carstensen, 1998). Das Verfahren mündet in konkrete Zielvereinbarungen mit nachprüfbaren Maßnahmen und festgelegten Terminen. Follow-Ups, die in regelmäßigen Abständen vorgesehen sind, sollen einen kontinuierlichen Qualitätssicherungsprozess in Gang setzen (Kromrey, 2006; Knödler, 2018).

Die Kombination interner und externer Evaluationsverfahren (erst Selbst- dann Fremdevaluation) wird *Peer-Review-Verfahren* genannt (Stamm, 2003; Knödler, 2018). Das zweistufige Evaluationsverfahren (intern und extern; Reissert & Carstensen, 1998) ähnelt in seinem Vorgehen dem Akkreditierungsverfahren und stammt aus den Niederlanden. Es findet auf Empfehlung des Wissenschaftsrats (1996) und der Hochschulrektorenkonferenz (1995) Einzug in den deutschen Hochschulkontext und kommt seit Mitte der 1990er an Hochschulen zum Einsatz. Ziel des diskursiven und dynamischen Prozesses ist die Analyse von Stärken und Schwächen zur Schärfung des Qualitätsprofils, der Kommunikation über Qualität zwischen den Beteiligten (Studierende, Lehrende) sowie Anregungen zur Weiterentwicklung. Nachteil ist der hohe Aufwand an Personal, Zeit und Kosten (Becker-Richter et al., 2002; Michalk, 2007; Reissert & Carstensen, 1998; Knödler, 2018).

## Modul 2: Lehrveranstaltungsevaluation (LVE)

Evaluation hat sich an deutschen Hochschulen etabliert und ist mittlerweile ein fester Bestandteil der Qualitätssicherung (Bargel & el Hage, 2000). Werden in anderen Ländern verschiedene Evaluationsverfahren kombiniert, wird in Deutschland die Evaluation der Lehre recht einseitig betrieben. Die Probleme und Kritikpunkte anderer Evaluationsverfahren (Lehrberichte, Peer-Review-Verfahren etc.) sind mitunter der Grund, warum sich die Lehrevaluation an den meisten Universitäten auf Studierendenbefragungen zu Lehrenden und ihren Lehrveranstaltungen beschränken (Kromrey, 2006; Reissert & Carstensen, 1998; Knödler, 2018). Dies geschieht mittlerweile routinemäßig mithilfe studentischer Fragebogenerhebungen zur Verbesserung der Lehrveranstaltungen oder Lehrqualität (Aleamoni, 1987; Döring, 2002; Firth, 1979; Marsh & Roche, 1993a; Knödler, 2018).

### Aktuelle Praxis der LVE

Die Evaluationspraxis an Hochschulen besteht noch immer überwiegend aus studentischen Beurteilungen in Form von Fragebögen, dessen Ergebnisse (qualitativ und quantitativ) dem Lehrenden rückgemeldet werden.

Die deutsche Praxis der LVE lässt sich anhand

- ihrer Entwicklung und Verbreitung (1),
- der angewandten Messinstrumente (2),
- der Stichprobe und
- Datenerhebung (3)
- sowie der Auswertung und Interpretation der Daten (4)

darstellen (Knödler, 2018).



## 1 | Entwicklung und Verbreitung der LVE

Seit ihrem Einzug die Hochschulen in den 1960ern findet die LVE in Form der studentischen Veranstaltungskritik statt, durch die die Studierenden die Lehrveranstaltung aktiv beeinflussen und mitgestalten (Döring, 2002). Die studentische LVE beschränkt sich dabei auf konkrete Rückmeldungen an Lehrende (Kromrey, 1995). Erst durch die zunehmende Verwendung von Hochschulrankings in den 1980ern und gesetzliche Verankerung in den 1990ern wächst die Bedeutung von LVEs als wichtiges Mittel zum Nachweis der Qualität von Studium und Lehre (Döring, 2002). Obwohl die LVE häufig die einzige Evaluationsaktivität deutscher Hochschulen ist, kann derzeit nicht von einem flächendeckenden Einsatz von LVEs gesprochen werden. Im internationalen Vergleich hinkt Deutschland hinterher (Knödler, 2018).

## 2 | Evaluationsinstrumente

„Instrumente der LVE sollen Lehrverhalten so abbilden, dass eine verhaltensbasierte Rückmeldung ebenso möglich ist wie aussagekräftige Indikatoren für Mittelzuweisungen und Bewerbungsverfahren“ (Knödler, 2018, S. 114).

Unter Instrumenten zur LVE erscheinen diverse Angebote und Verfahren, die sich hinsichtlich ihrer Qualität, ihres Umfangs sowie ihres Einsatzfeldes unterscheiden. Neben elaborierten Instrumenten (z.B. HILVE; Rindermann & Amelang, 1994a), sind auch scheinbar beliebig zusammengestellte Fragebögen zu finden (siehe auch die Zusammenstellung von Reissert, 1992). Den meisten Instrumenten gemein sind eine Reihe formativer Fragen bezüglich einzelner Lehrveranstaltungsaspekte (Lehrmaterial, Tempo etc.) sowie Globalbewertungen (Zufriedenheit insgesamt). Meist werden quantitative Daten erhoben und durch wenige qualitative, offene Fragen (Lob, Verbesserungsvorschläge, Kritik) ergänzt (el Hage, 1996a). An den meisten Fachbereichen deutscher Hochschulen werden ad hoc und selbstentwickelte Instrumente verwendet. Derartige Fragebögen basieren weder auf einem theoretischen und elaborierten Modell der Qualität von Lehrveranstaltungen noch sind sie bezüglich ihrer testtheoretischen Gütemerkmale (v.a. hinsichtlich ihrer Validität und Reliabilität) geprüft (Knödler, 2018).

### 3 | Durchführung und Datenerhebung

LVE werden in der Regel von der Hochschule oder dem Fachbereich selbst durchgeführt. Die rein interne Evaluation von Lehrenden und Lehrveranstaltungen erfolgt typischerweise quasi-experimentell oder kasuistisch. Bei der Betrachtung einzelner Evaluationsobjekte wird von einer kasuistischen Evaluation gesprochen, beim Vergleich von (Objekt-)Gruppen von quasiexperimenteller Evaluation (Döring, 2002). Nur an wenigen Hochschulen wird die LVE von Evaluationszentren oder dafür abgestellten Expert:innen und Arbeitsgruppen organisiert. In der Regel werden die unterschiedlichen Teilprozesse häufig auf verschiedene Personengruppen aufgeteilt. Zum Beispiel können die Instrumente in einer Arbeitsgruppe oder einem Gremium entwickelt werden, die Ergebnisse wiederum werden meist von jedem Lehrenden selbst interpretiert. Die Studierenden sind die Informanten, welche die Daten liefern (Döring, 2002). Abhängig vom Zweck variiert der Erhebungszeitpunkt: Sollen die Ergebnisse für Personalentscheidungen oder Bewerbungsverfahren herangezogen werden, eignet sich eine summative Evaluation am Ende der Veranstaltung (wie gut/schlecht ist eine Veranstaltung insgesamt). Sollen jedoch Nachjustierungen im Laufe des Semesters möglich sein, empfiehlt sich eine formative Evaluation in der Mitte des Veranstaltungszeitraums (der Fokus hierbei: wo liegen Schwächen und wie lassen sie sich beheben; Marsh, 1984; Marsh & Roche, 1993b Covert & Mason, 2016). In der Regel werden Veranstaltung im Paper-Pencil-Verfahren durchgeführt (Covert & Mason, 2016; Marsh, 1984). Da so nur die in dieser Sitzung Anwesenden, sprich eine selektive Stichprobe, befragt werden können, was zu einer systematischen Verzerrung führen kann, wird die Datenerhebung in den letzten Jahren vermehrt online (z.B. mithilfe von EvaSys) erhoben (Biehl & Costanzo, 2014). Dadurch können alle in die Veranstaltung eingeschriebenen Studierenden erreicht werden, zudem ist eine automatische Analyse und Datenarchivierung möglich (Döring, 2002). Online-Umfragen haben den Vorteil, dass sie vorlesungszeitunabhängig durchgeführt werden können, hinzu kommt das vereinfachte Erhebungs- und Auswertungsverfahren. Nachteil ist die nachgewiesene geringere Rücklaufquote (Raser, 2016; Knödler, 2018).

## 4 | Auswertung und Interpretation

Die Skalenfragen werden in der Regel durch Mittelwertsbildung aller Studierenden ausgewertet. Die Gesamtbewertung am Schluss ist dabei nicht die Zusammenfassung der einzelnen formativ erhobenen Aspekte, sondern das Ergebnis der direkten summativen Frage nach der Gesamtzufriedenheit. Bewertet wird meist auf einer 5- bis 7-stufigen Skala oder anhand von Schulnoten. Die qualitativen Fragen werden in der Regel (pro Veranstaltung) zusammengefasst dargestellt.

Die Datenauswertung und Ergebnisinterpretation beschränken sich meist auf eine Rückmeldung von Mittelwerten an den Lehrenden, Konsequenzen sind selten (Knödler, 2018).

## Zur Aussagekraft studentischer LVE

Zahlreiche Autoren, wie Aleamoni (1987), Kromrey (1995) oder Rindermann und Amelang (1994b), beschäftigten sich mit der Aussagekraft studentischer LVE (Knödler, 2018).

### Studentisches Urteilsverhalten

Studierende sind durchaus in der Lage, unterschiedliche anforderungsspezifische und auf den Lehrenden bezogene Aspekte differenziert zu beurteilen, was sorgfältig konstruierte Messinstrumente voraussetzt, welche konkrete Veranstaltungsurteile mithilfe entsprechender Items ermöglichen (Rindermann, 2009). Die Bewertung unterschiedlicher Teilaspekte verlangt den Studierenden differenzierte Beurteilungen ab. Kromrey (1995) konnte neben den differenten Urteilen in einer Veranstaltung eine Intra-Befragten-Heterogenität (Unterschiede zwischen Global- und Detailurteilen eines Befragten) erkennen, was auf die Fähigkeit der Studierenden zu einer differenzierten Einschätzung schließen lässt. Beim Vergleich detaillierter Bewertungen (konkrete Items, die bestimmte Aspekte detailliert abfragen) mit Globalaussagen stellt Kromrey (2006) zudem fest, dass unter den Studierenden zwei gegensätzliche Evaluierungsweisen bestehen. Etwa die Hälfte der Studierenden einer Lehrveranstaltung beurteilt diese pauschal (das zusammenfassende Gesamturteil stimmt in der Tendenz der Urteile mit allen detaillierten Items einer Dimension überein), die andere Hälfte urteilt differenziert (mit einem Teilaspekt zufrieden, mit einem anderen unzufrieden und mit den nächsten mittelmäßig zufrieden) und bildet hinsichtlich der Gesamtbewertung einen subjektiven Mittelwert. Diese multidimensionale Bewertung wird in nationalen wie internationalen Studien mithilfe unterschiedlicher Instrumente empirisch belegt (Astleitner, 1991; Cadwell & Jenkins, 1985; Jenkins, 1987; Whitely & Doyle, 1976), ist allerdings in Praxis und Forschung weiterhin Diskussionsthema (Rindermann, 2009; Knödler, 2018).

Bledsoe (1978), Marsh und Overall (1981), Firth (1979), Marsh (1984) und Centra (1993) konnten in ihren Untersuchungen einen hohen Grad an Genauigkeit, Stabilität und

Übereinstimmung studentischer Urteile ermitteln. Auch weisen sie Stabilität über einen längeren Zeitraum nach (Knödler, 2018).

Selektionseffekte (Drop-Out-Effekte) sind in nahezu jeder Veranstaltung erkennbar. Da Gründe des Fernbleibens nicht erklärbar sind, kann nicht belegt werden, ob der Teilnehmerschwund an sich schon ein Indikator schlechter Lehre ist (Döring, 2002). Ebenso wenig lässt sich die Art der Verzerrung durch Drop-Outs erkennen, sprich ob anwesende Studierende tendenziell besser oder schlechter bewerten als abwesende. Die Auswirkung des Teilnehmerschwunds ist somit ebenso wenig kausal attribuierbar wie Unterschiede zwischen den An- und Abwesenden. Spekuliert wird diesbezüglich, dass die anwesende Gruppe eher aus zufriedenen Studierenden besteht, welche das Urteil tendenziell positiv verzerren (el Hage, 1996a). Um die Gründe des Fernbleibens und mögliche Verzerrungseffekte zu klären, empfiehlt sich eine zweite Erhebung zu Beginn der Vorlesungszeit, worüber auch die Drop-Out-Quote bestimmt werden kann (el Hage, 1996a). Dies ist allerdings mit einem hohen Kosten- und Zeitaufwand verbunden und in der Praxis kaum möglich. Für die Aussagekraft der LVE unter Berücksichtigung der Drop-Outs ist deshalb vielmehr die Anzahl der Bewertungen relevant. So konnte Rindermann (2009) bei einer Mindestzahl von zehn Studierenden eine relativ hohe Stabilität und Genauigkeit der Bewertungen ermitteln. McBean und Lennox (1985) fanden heraus, dass eine Rücklaufquote von 50% bei mehr als 30 Teilnehmern bereits für eine präzise Lehrbewertung ausreichen, bei weniger als 30 Studierenden sollten es ca. 80% sein, um den gleichen Grad an Genauigkeit zu erzielen (Knödler, 2018).

Studierende sind keine ausgebildeten Beobachter:innen oder Beurteiler:innen und bewerten in einer natürlichen Umgebung, geprägt von unterschiedlichen Einflüssen. Diese unterschiedlichen Einflüsse und Charakteristika der Studierenden erklären die Variationen bei den Beurteilungen (Feldmann, 1977). Dennoch sind Studierende als die direkt von der Lehre Betroffenen, durchaus in der Lage, die Lehrqualität (aus eigener Erfahrung) zu beurteilen, weshalb sich studentische LVE zur Rückmeldung an den Lehrenden eignen (Kromrey, 1995). Sie sind in der Lage, Urteile über die Relevanz und Attraktivität von Themen zu treffen, auch verhaltensbezogene Aussagen über

den Lehrenden und somit zur didaktischen Lehrqualität sind möglich. Was allerdings nicht durch die Studierenden beurteilt werden kann, sind die Inhalte der Veranstaltung und deren Umfang sowie die wissenschaftliche Fachkompetenz des Lehrenden (el Hage, 1996a; Marques, Lane & Dorfman, 1979; Rindermann, 2009). Das studentische Urteil muss somit als eine Zusammensetzung aus Charakteristika von Studierenden, Lehrenden und der Veranstaltung sowie dem Beurteilungskontext analysiert und interpretiert werden. Diese Einflüsse sind Bestandteil jeder Befragung und lassen sich nicht vermeiden (Feldmann, 1977; Knödler, 2018).

### **Einfluss lernangebotsunabhängiger Aspekte**

Bei der Suche nach Ursachen der unterschiedlichen oder gar gegensätzlichen Beurteilungen derselben Veranstaltung hinsichtlich der Lehrqualität und des Lernprozesses, kommen lehrangebotsunabhängige Aspekte, welche die Einschätzung der Studierenden beeinflussen, ins Spiel. Diese Aspekte sind vielzähliger Art. In diesem Zuge sind beispielsweise das Studienalter und Fachsemester, das Intelligenzniveau, geschlechtsspezifische Zusammenhänge, Vorkenntnisse, Fleiß, Anstrengungsbereitschaft, Interessen, Leistungsmotivation und Studienleistung der Studierenden als Einflussgrößen zu nennen (Süllwold, 1992). Hinzu kommen Motivation, Sympathie, demographische Daten des Lehrenden und Studierenden, das Thema, der Besuchsgrund oder das Studierverhalten.

Diese lernangebotsunabhängigen Aspekte, häufig auch Hintergrund- und Kontrollvariablen genannt, können verschiedene Dimensionen unterschiedlich beeinflussen (Marsh, 1984) und sind schwer isoliert voneinander zu messen.

Vor allem Lehrende gehen von hohen Effekten lehrunabhängiger Variablen aus (Marsh, 1984). In der Tat belegt die umfangreiche internationale Forschung, dass Bewertungen studentischer LVEs systematischen Verzerrungen unterliegen (Kromrey, 1994; Rindermann, 2009). Relevante lehrangebotsunabhängige Aspekte sind unter anderem (Knödler, 2018):

*Persönliche Merkmale:* Persönlichkeitsvariablen und sozialstatistische Merkmale auf Seiten der Studierenden führen zu keinen, allenfalls geringen Verzerrungen (Centra, 1980; el Hage, 1996b; Feldmann, 1977; Marsh, 1984; Rindermann, 1996; Knödler, 2018).

*Erster und genereller Eindruck:* Feldman (1977) konnte Auswirkungen des studentischen Eindrucks feststellen. Der Eindruck vor der Veranstaltung, zu Beginn der Veranstaltung und der generelle Eindruck der Studierenden in Bezug auf den Lehrenden und seine Veranstaltung korrelieren positiv mit dem studentischen Urteil. Der studentische Eindruck und der Ruf eines Lehrenden und seiner Veranstaltung wirken sich auch auf die Wahl des Kurses, auf das Interesse und die Motivation aus (Knödler, 2018).

*Besuchsgrund:* Ein freiwilliger Entschluss – hier die Entscheidung zum Besuch einer Veranstaltung – wirkt sich auf die Motivation (intrinsisch statt extrinsisch) und auf die Teilnahme und Mitarbeit aus (Kromrey, 2006). Da sich die Teilnahmemotive über die unterschiedlichen Veranstaltungen hinweg nicht gleichmäßig verteilen, bestimmt die Teilnahmestruktur in hohem Maße, ob die durchschnittliche Beurteilung tendenziell gut oder schlecht ausfällt. Empirisch wird dies bestätigt (siehe z.B. Kromrey, 1995; Rindermann, 2009). So bewerten Studierende über die Fakultäten hinweg einen verpflichtenden Veranstaltungsbesuch deutlich schlechter (Kromrey, 1995; Knödler, 2018).

*Interesse und Motivation:* Mit der Wahlmöglichkeit korreliert auch die Motivation. Neben der extrinsischen und intrinsischen Motivation und dessen „Ansteckungseffekt“ (Kromrey, 1995, S. 12), haben auch die individuelle Motivationslage und das individuelle Interesse einen starken Einfluss auf das Urteil (el Hage, 1996b; Kromrey, 1995). In welchem Ausmaß Interesse und Motivation allerdings durch den Lehrenden beeinflusst werden, ist nicht eindeutig belegt (Feldmann, 1977; Knödler, 2018).

*Studienerfahrung:* Der Einfluss der Studienerfahrung (erhoben durch das Fachsemester) liefert inkonsistente Ergebnisse (el Hagem 1996b; Feldmann, 1977), was unter anderem an unterschiedlichen Forschungsdesigns und Erhebungsmethoden

liegt. Somit ist bezüglich des Einflusses der Studienerfahrung auf das studentische Urteil keine Aussage zu treffen (Knödler, 2018).

*Studierendenverhalten – Vor- und Nachbereitung:* Der regelmäßige Besuch der Veranstaltung sowie eine kontinuierliche Vor- und Nachbereitung haben einen positiven Einfluss auf die studentische Bewertung (Kromrey, 1995). Ob die Studierenden in Teil- oder Vollzeit studieren, ist nicht entscheidend (Orpen, 1980; Knödler, 2018).

Häufig stellt sich die Frage, inwieweit die studentischen Urteile auf den Lehrenden (persönliche Eigenschaften) oder seine Lehre (didaktische Fähigkeiten) zurückzuführen sind – über einen existierenden Zusammenhang besteht kein Zweifel (Crull, 1982; Marsh & Overall, 1981.) Der Lehrende hat den größten Effekt auf das studentische Urteil (5- bis 10-mal größer). Auf welche Aspekte im Lehrendenverhalten dies genau zurückführbar ist, wird kontrovers diskutiert. So wurde der Dr. Fox Effekt mehrfach untersucht – mit unterschiedlichen Ergebnissen (siehe Marsh, 1984). Hinsichtlich des Geschlechts des Lehrenden und der beruflichen Position konnte ein geringer Einfluss auf das studentische Urteil festgestellt werden, wobei die Art des Einflusses variiert (Marsh, 1984). Das Alter und die Berufserfahrung haben keinen Einfluss auf das studentische Urteil (Cashin & Beneton, 2011). In Bezug auf die Validität und Fairness studentischer Beurteilungen ist an dieser Stelle der bisher weitgehend vernachlässigten Aspekte der Attraktivität oder Sympathie des Lehrenden zu nennen, welcher in einigen Untersuchungen nachgewiesen werden konnte (einen Überblick liefert Wolbring, 2013, S. 120; Knödler, 2018).



# Evaluationsinstrumente (schriftlich und mündlich)

## Schriftliche Methoden - Der Fragebogen

Fragebögen unterscheiden sich vor allem im Aufbau und Umfang sowie in der Wahl der Items.

Bezüglich des **Aufbaus** empfiehlt sich ein empirisch getestetes und zugleich flexibles Instrument (Webler, 1992), das sich aus einem festen (um die Veranstaltungen vergleichbar zu machen) und einem lehrform- oder veranstaltungsspezifischen Itempool (für verschiedene Veranstaltungsarten oder Fächer) sowie qualitativen Fragen zusammensetzt. In Hinblick auf die Länge eines Fragebogens sind 15 bis 35 Items zu empfehlen (Meredith, 2016; Knödler, 2018).

Die **Instruktion** fungiert als Einleitung und gibt Aufschluss über den Verwendungszweck der Evaluationsergebnisse. Da Instruktion und Verwendungszweck (z.B. administrative Zwecke oder Rückmelde- bzw. Verbesserungsfunktion) die Lehrbewertung beeinflussen, ist eine Standardisierung empfehlenswert (siehe diesbezüglich auch die Empfehlungen von Driscoll & Goodwin, 1979). In der Instruktion ist auf die anonyme Datenerhebung hinzuweisen (Orpen, 1980, Knödler, 2018).

**Ratingskalen** sollen äquidistant messen und zugleich hinreichend differenzieren (Rohrmann, 1978). In der Regel werden Ja-/Nein-Kategorien (dichotom) oder Likert-Skalen (mehrstufig) verwendet. Mehrstufige Antwortoptionen führen zu qualitativ hochwertigeren und informationshaltigeren Ergebnissen, weshalb sie zu empfehlen sind (Saris, Revilla, Krosnick & Shaeffer, 2010). Neumann und Neumann (1981) raten aufgrund ihrer Studien zu einer Skala zwischen 5 und 7 Ausprägungen (diese differenzieren am genauesten). Eine numerische und verbale Kategorisierung, die optisch verständlich dargestellt ist, eignet sich dafür am besten (Knödler, 2019).

**Itempool** – „Die Informationen, die durch eine LVE gezogen werden können, sind von den Items und ihren Formulierungen abhängig (Marsh, 1984). Es zeigt sich, dass eine Mischung aus Wertungen oder Beobachtungen sinnvoll ist. Wie bei allen Bewertungen werden auch bei der Beurteilung von Lehre Vergleichsmaßstäbe

respektive eine Bezugsnorm benötigt. Bei Items, für die eine Einschätzung des beobachtbaren Verhaltens nötig ist, kann entweder mithilfe eines sozialen Vergleichs oder anhand der Idealnorm verglichen werden, was sich durch entsprechende Formulierungen bewerkstelligen lässt (el Hage, 1996a). Pro Item sollte nur ein Sachverhalt der Dimension abgefragt werden (Spooren et al., 2013). Dabei ist die Formulierung entscheidend. Items sind deshalb leicht verständlich (angepasst an die Zielgruppe), kontextspezifisch und so präzise wie möglich zu formulieren. Von Verallgemeinerungen, langen Items, doppelten Verneinungen oder mehrdeutigen Begriffen ist abzusehen. Zeitspannen sollten genau definiert und Einheiten (z.B. Stunden pro Woche) benannt werden (Bühner, 2011). Spooren et al. (2013) empfiehlt eine Kombination aus spezifischen und globalen Items. Spezifische Items liefern wichtige Informationen über die Lehrqualität auf den unterschiedlichen Dimensionen, ein globaler Wert am Schluss kann einen einzelnen Wert der Lehrqualität erheben und ist vor allem für Personalentscheidungen nützlich (Aleamoni & Thomas, 1980; Covert & Mason, 2016; Marsh, 1982a; Marsh & Hocevar, 1984; Knöder, 2018; Spooren et al., 2013).

Das Erheben von Hintergrundvariablen, die sich auf soziodemografische Daten beziehen, und Kontrollvariablen (Interesse, Aufwand etc.), wodurch sich Urteile bei unterschiedlichen Auswertungszwecken kontrollieren lassen (el Hage, 1996b), ist empfehlenswert. Eine Gewichtung der Lehraspekte ist möglich, sollte jedoch empirisch und theoriebasiert erfolgen (Marsh, 1984). „Happy Sheets“ sind zu vermeiden, dennoch ist ein Mix aus negativ und positiv gepolten Items nicht zu empfehlen. Negative Items können zustimmendes Verhalten (Akquieszenz) zwar entgegenwirken, sind jedoch weniger reliabel und führen häufig zu Verwirrungen (Roszkowski & Soven, 2010). Hervorhebungen sind nur in geringem Maße hilfreich (Bühner, 2011; Knödler, 2018).

### **Mündliche Methoden**

Neben klassischen Fragebögen sind auch mündliche Methoden für eine LVE einsetzbar (von Werthern & Knödler, 2020).

Das *Ampelfeedback* als eine Form der Befragung erhebt schnelle quantitative Verteilung von Lernständen, Meinungen oder Präferenzen einer Gruppe. Mit gelben, roten und grünen Karten ausgestattet halten die Studierenden diese zu den gestellten Fragen gleichzeitig nach oben. Diese Methode (reine Erhebung ohne Einholen von Kommentaren) dauert ca. *zwei bis fünf Minuten* und eignet sich für kleine wie auch große Gruppen (bei Gruppen ab 30 Personen empfiehlt sich ein Assistent).

Das *Blitzlicht* ist das qualitative Pendant zum Ampelfeedback. Hier werden kurze Antworten der Studierenden zu auf einer Flipchart notierten Fragen erhoben. Jede\*r Studierende kommt pro Frage einmal zum Zuge – eine Bezugnahme zu anderen Antworten ist nicht erwünscht. Die Methode dauert ca. *10-20 Minuten* und eignet sich für kleine Gruppen (ca. 8-40 Personen).

*Aufstellungsbefragung:* Diese Befragung enthält Elemente der Beobachtung und eignet sich zur Reflexion oder Aktivierung der Studierenden und liefert vor allem quantitative Daten. Dabei stellen sich die Studierenden nach dem mündlichen Stellen einer Frage durch die Lehrperson zu der von ihnen ausgewählten Antwort in die jeweilige Ecke. Die Methode dauert ca. *15-20 Minuten* (für 5-8 Fragen) und eignet sich bei Gruppen von 9-60 Personen (ab 30 Personen empfiehlt sich ein Assistent).

*Plakate mit offenen Fragen:* Mit dieser Methode lassen sich gezielt Informationen über die Stärken und Schwächen einer Lehrveranstaltung einholen. Bei dieser Befragung entstehen überwiegend qualitative Daten. Jede Frage wird auf ein Flipchart geschrieben – die Studierenden bewegen sich im Raum und notieren auf allen Plakaten ihre Kommentare dazu. Die Dauer beläuft sich auf *10-20 Minuten*, die Gruppengröße kann zwischen acht und 80 Personen sein.

*Plakat mit Matrix-Fragen:* Der Plakat-Fragebogen ist wenig aufwändig und führt zu einer schnellen Meinungsabfrage, bei der quantitative Daten erzeugt werden. Mit Klebepunkten können die Studierenden die Fragen bewerten. Diese Methode dauert ca. *drei bis acht Minuten* und eignet sich für Gruppen zwischen sechs und 50 Personen.

# Modul 3: Umgang mit Evaluationsergebnissen

## - Desk Research

Während der Nutzen und die Nutzung von Evaluation auf theoretischer Ebene ein in der Evaluationscommunity vielbesprochenes Thema ist, verhält es sich in Bezug auf den ganz konkreten Umgang mit Evaluationsergebnissen in der Praxis ein wenig anders. Dennoch gibt es Möglichkeiten, sich dem Thema auf unterschiedlichen Ebenen anzunähern, anbei ein kurzer Überblick dazu:



Abb. 3: Umgang mit Evaluationsergebnissen. Oder: Was findet man, wenn man danach sucht? (eigene Darstellung)

Ohne zu tief in die Thematik Nutzen und Nutzung von Evaluation einzusteigen, anbei ein knapper Exkurs:

„Die Nutzung von Evaluationen wurde in den späten 1960er- beziehungsweise in den frühen 1970er-Jahren erstmals thematisiert. Michael Quinn Patton und Carol Weiss hatten festgestellt, dass ihre Evaluationen häufig nicht genutzt wurden. Dies bedeutet, dass weder die Ergebnisse noch die Empfehlungen von Evaluationen zu Veränderungen führten und umgesetzt wurden. Später hat Patton, mit seiner Schrift zur Utilization-Focused

Evaluation alle Arbeiten bei einer Evaluation dem Primat der Verwertung untergeordnet (Patton 2008). Unter Nutzung von Evaluationen versteht man die verschiedenen Arten der Verwendung und bezeichnet damit den beobachtbaren Gebrauch von Evaluationen. Bei der Nützlichkeit von Evaluationen soll geklärt werden, wie Evaluationen im Prozess gestaltet sein müssen, damit sie nützlich sind und von den Adressaten auch verwendet werden. Schliesslich beschreibt der Nutzen der Evaluation, welche Wirkung diese auf die Adressaten hat (Friedrich, V.; Fischer, A. (Hg.) (2015), S. 9).

Des Weiteren splitten die Autor:innen die Nutzung von Evaluationen in drei zentrale Elemente aufsplitten: Am Beginn der Nutzung von Evaluation steht der Evaluationsprozess, danach folgt die Verarbeitung der Evaluationen durch Organisationen und/oder Personen. Last but not least kommt es zur Verwendung oder Nutzung der Evaluationsergebnisse:

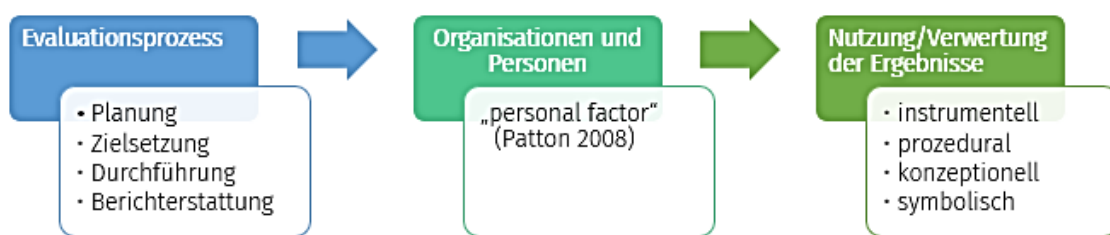


Abb. 4: Zentrale Elemente bei der Nutzung von Evaluation (Friedrich, V.; Fischer, A. (Hg.), 2015, S. 9)

## Aspekte der Bewertung

In der Regel werden an deutschen Hochschulen die LVE-Ergebnisse durch die Bildung des Mittelwerts der Skalenfragen aller Studierenden ausgewertet. Das arithmetische Mittel wird dabei oft als Qualitätsindikator interpretiert. Die Gesamtbewertung am Schluss erfasst über eine summative Frage die Gesamtzufriedenheit und ist deshalb nicht als die Zusammenfassung der einzelnen formativ erhobenen Aspekte zu verstehen (Döring, 2002; Kromrey, 1995). Qualitativen Fragen werden pro Veranstaltung zusammengefasst – eine systematische (z.B. inhaltsanalytische) Auswertung findet in der Regel nicht statt. Vorteil dieser Auswertungsstrategie ist die einfache Handhabung und leichte Vermittelbarkeit der Ergebnisse. Kritisch sind vor allem die Bildung des Mittelwerts sowie Qualitätskriterien und -standards als einzustufen (Knödler, 2018).

### 1 | Probleme bei der Bewertung durch Mittelwerte

Mittelwerte, bei der für jedes Item ebenso wie für die unterschiedlichen Dimensionen das arithmetische Mittel und die Standardabweichung angegeben werden, sind gängige Praxis bei Auswertung von LVE. Diese isolierten Antwortauszählungen und Mittelwertberechnungen sind in Bezug auf das heterogene Antwortverhalten der Studierenden durchaus kritisch zu sehen, da das differente Urteilsverhalten der Studierenden bei der Beurteilung der Fragebögen zwangsläufig zu unterschiedlichen Bewertungen führt.

Dennoch spricht die hohe Antwortvariation nicht per se gegen die Anwendung von Mittelwerten. Die Festlegung weniger restriktiver Bedingung wie die Abbildung tendenziell ‚guter‘ oder ‚schlechter‘ Lehre, trotz der Variation im Detail, ermöglicht zumindest eine grobe Klassifikation und Einordnung in wenige zweifelsohne ‚gute‘ beziehungsweise ‚schlechte‘ Veranstaltungen (Kromrey, 1995). Hinzu kommt, dass Lehrende und Vorgesetzte in der Regel interessierter an Informationen auf Veranstaltungsebene, als an LVE-Ergebnissen auf der individuellen Ebene (einzelne Studierende) sind. So Veranstaltungsmittelwerte zur Darstellung von LVE-Ergebnissen durchaus sinnvoll. Mittelwerte sind in aggregierter Form (z.B. mehrere

Items zu einem Faktor) reliabler und ermöglichen die Reduktion von Verzerrungen. Standardabweichungen und Konfidenzintervalle zeigen die Ausprägungen in der Bewertung und geben einen Anhaltspunkt über die Streuung des Urteils. Auch Vergleiche mit dem Mittelwert anderer Lehrender oder der Fakultät führen zu einer besseren Einordnung der einzelnen Ergebnisse (Abrami, 2001; Boysen, 2015). Beim Vergleich zwischen Lehrenden sind verhaltensbasierte Items zu berücksichtigen, welche zwischen ‚guter‘ und ‚schlechter‘ Lehre differenzieren können. Lehrunabhängige Items, die Wertungen (Selbstreferenz als Bezugspunkt) anstelle von Beobachtungen erheben, sind separat auszuwerten. Gesamtmittelwerte, die eine subjektive Gesamteinschätzung liefern, sollen ebenso rückgemeldet werden wie Profillinien, welche Ausschläge in der Bewertung veranschaulichen, um Stärken und Schwächen schneller zu erkennen (el Hage, 1996a; Knödler, 2018).

## 2 | Über die Notwendigkeit von Qualitätsstandards und -kriterien

Um anhand von LVE-Ergebnissen die Lehrqualität zu bewerten, sind Qualitätskriterien beziehungsweise -standards für die erhobenen Merkmale notwendig. Da nicht a priori festgelegt ist, ab welchem Durchschnittswert einer Lehrveranstaltung ‚gute‘ oder ‚schlechte‘ Qualität zugeschrieben wird, wird anstelle von kriteriumsorientierten Testen meist normiert vorgegangen. Das bedeutet, dass die Bewertung einer Veranstaltung mit dem Gesamtmittelwert aller evaluierten Veranstaltungen eines Semesters oder Studiengangs verglichen wird. Diese Interpretation rankt Lehrveranstaltungen oder Lehrende auf der Datengrundlage studentischer LVE (Döring, 2002). Eine normierte Interpretation von Evaluationsdaten, erscheint insofern problematisch, als dass die besten oder schlechtesten Lehrveranstaltungen nicht zwingend auf Lehrdefizite zurückgeführt werden können, sondern lediglich eine statistische Gesetzmäßigkeit darstellen. Aus statistischen Gründen würde so auch in einem hochschuldidaktisch optimalen Fachbereich mit hoher Lehrqualität ein Teil der Lehrveranstaltungen unterdurchschnittlich abschneiden, ohne diesen Lehrveranstaltungen schlechte Lehrqualität zu schreiben zu können (Döring, 2002). Bei der kriteriumsorientierten Testvariante ist zu kritisieren, dass die meisten Lehrveranstaltungen mit Durchschnittsnoten zwischen 2 und 3

bewertet werden, was eine statistisch signifikante Steigerung erschwert. Dass Verbesserungsinitiativen darüber hinaus „nur dann rational sind, wenn sie in vertretbarer Kosten-Nutzen-Relation stehen“ (Döring, 2002, S. 16), wird in der Diskussion über den Umgang mit den LVE-Ergebnissen nicht berücksichtigt (Döring, 2002; Knödler, 2018).

## Formen der Nutzung

Hinsichtlich der Nutzungsformen der Evaluationsergebnisse sind unterschiedliche Verfahren und Herangehensweisen möglich. Die aktuelle Praxis beschränkt sich meist auf die reine Ergebnismeldung an den Lehrenden, mit der Aufforderung, die LVE-Ergebnisse mit den Studierenden zu besprechen und in zukünftigen Planungen zu berücksichtigen. Offen bleibt dabei, wie dies konkret vonstattengehen soll (Döring, 2002). Negative Sanktionen (z.B. Gespräch mit dem Vorgesetzten) aufgrund unterdurchschnittlicher Ergebnisse, sind in Deutschland unüblich. Begründet wird dies damit, dass allein die Kenntnisnahme der LVE-Ergebnisse zur Verbesserung des Lehrverhaltens beiträgt, was Untersuchungen und Befragungen widerlegen. In Anlehnung an Gralki und Hecht (1992, S. 101) werden deshalb im Folgenden Nutzungsformen von LVE angeführt (Knödler, 2018).

### Transparenz- und Forschungsmodell

LVE-Ergebnisse liefern Informationen, welche für die Beteiligten von ganz unterschiedlichem Nutzen sind.

Der *Forschung* liefert die LVE neue Ergebnisse, die zur Weiterentwicklung und wissenschaftlichen Aufarbeitung (der Lehrqualität) beitragen können. So lassen sich Lehr-Hypothesen testen und auf der Basis empirischer Sozialforschung auf das Feld der LVE übertragen. In Deutschland sind derzeit vor allem die Vergleichbarkeit sowie Einflussfaktoren relevant, welche einiger empirischer Untersuchung bedürfen. Zudem können Methoden, Instrumente und Lehrbewertungsverfahren weiterentwickelt werden. Des Weiteren kann die empirische Forschung zur Versachlichung und Aufklärung von Kritik an der LVE beitragen (el Hage, 1996a; Knödler, 2018).



Einen Nutzen auf *hochschulpolitischer Ebene* haben Evaluationsergebnisse, wenn die unterschiedlichen Qualitätsmanagementaktivitäten zu einem hochschulischen Gesamtkonzept systematisch verzahnt werden (Jaeger & Kerst, 2010). Dabei sollte LVE auf der Makroebene durch die Lehrevaluation ergänzt werden, um die Sicherstellung der nachhaltigen und umfassenden Studienqualität zu gewährleisten. Damit gehen professionelle Meta-Evaluationen (Bewertung der Evaluation) einher. So können im Sinne der Qualitätsentwicklung schlussendlich hochschuldidaktische und individuelle Unterstützungsmöglichkeiten (z.B. Beratung, Betreuung) in den Hochschulalltag implementiert werden (Döring, 2002). Ein derartig ganzheitliches Verfahren kann als hochschulisches Steuerungsinstrument wirken (Barz & Miethig, 1992; Knödler, 2018).

Für die *Institution Hochschule* liefern LVE-Ergebnisse in aggregierter Form Hinweise über motivationsfördernden und beliebten Lernstoff, zum Bedarf an veranstaltungsübergreifenden zusätzlichen Arbeits- und Lernhilfen sowie zu notwendigen Rahmenbedingungen für gute Lehre (z.B. Ausstattung, zeitliche Taktung). Werden die LVE-Ergebnisse innerhalb der Veranstaltung besprochen, erhalten die Studierenden Einblicke in die durchschnittliche Wahrnehmung der Lehrveranstaltung ihrer Kommilitonen (Knödler, 2018).

Für den Nutzen von Evaluationsergebnissen auf *Fakultätsebene* ist die prozessbegleitende Kommunikation auf und zwischen allen Hierarchieebenen eine grundlegende Voraussetzung. Kommunikation ist die Basis eines erfolgreichen Evaluationsprozesses und für die anschließende Nutzung der Ergebnisse gleichermaßen Ziel und Voraussetzung (Spooren et al., 2013; Wehrlin, 2011b; Becker-Richter et al., 2002). Um Lehrende darüber hinaus zu Veränderungen im Lehrverhalten zu motivieren und eine ergebnisorientierte Verwendung zu ermöglichen, sind weitere Unterstützungsstrategien beim Umgang mit und der Nutzung von Evaluationsergebnissen zu implementieren (el Hage, 1996a). Dafür sind vor allem Rahmenbedingungen zu schaffen, die eine von allen am Evaluationsprozess Beteiligten akzeptierte und zielgerichtete Datenverwendung ermöglichen und studentische LVE als Teil hochschulischer Evaluationsmaßnahmen

zur Verbesserung der Lehre etablieren (el Hage, 1996a; Raser & Hense, 2017). Zudem können Evaluationsergebnisse auch für Bewerbungen und bei der Veranstaltungswahl durch Studierende herangezogen werden können. Eine Veröffentlichung von Evaluationsergebnissen bietet Studierenden die Möglichkeit, die Information bei der Wahl der Veranstaltung zu nutzen (Knödler, 2018).

Dem *Lehrenden* geben die Daten beispielsweise Aufschluss über die folgenden Punkte:

- die Zielgruppe (für wen er überhaupt lehrt), um die Lehre daran anzupassen;
- die Art der Wünsche, Interessen, Erwartungen aber auch Vorurteile, mit denen die Studierenden in die Lehrveranstaltung kommen, um die Lehre in geeigneter Form daran auszurichten;
- darüber, wie gut oder schlecht die Studierenden mit dem Lehrangebot zurechtkommen, um bei fehlenden (regulär vorausgesetzten) Vorkenntnissen gezielt Lern- und Arbeitsempfehlungen zu geben, zusätzliche Übungen anzubieten oder aber ein überzogenes Anforderungsniveau zurückzuschrauben;
- darüber, wie gut oder schlecht sie als Lehrperson auf die Studierenden wirken (z.B. (des-)interessiert, distanziert/ansprechbar, positiver/negativer Sprachstil, angemessene Beispiele);
- darüber, wie unterschiedlich die Wahrnehmungen und Beurteilungen der Studierenden ausfallen (können)
- einzelne Kritik oder einzelnes Lob ist nicht gleich die Meinung aller (Kromrey, 1995; Knödler, 2018).

## Qualifikationsmodell

Das Qualifikationsmodell fokussiert sich auf den Lehrenden und versteht die Qualifizierung des Lehrpersonals als Schlüssel zur Verbesserung von Lehre (Gralki & Hecht, 1992). Lehrende erhalten fachliche Unterstützung beim Umgang mit ihren Evaluationsergebnissen. Ziel dabei ist, identifizierte Schwächen in der Lehre in Weiterbildungsangebote zu übersetzen. Mit der Unterstützung durch die Hochschule wird Offenheit für Veränderungen im Lehrverhalten gefördert, um auch Abwehrmechanismen entgegenzuwirken. Hochschulen haben zwar den Auftrag, die Weiterbildung des Personals zu fördern und sicherzustellen, dass die Lehre methodischen und didaktischen Standards entspricht, eine verpflichtende didaktische Aus- und Weiterbildung ist an Hochschulen jedoch noch immer nicht vorhanden. Lehrenden fehlen hierbei weiterhin die nötigen Anreize für eine Weiterqualifikation (Knödler, 2018).

Unter anderem, deshalb fordert Rindermann (2009) in seinem Beratungsansatz eine individuelle und hochschuldidaktische Beratung für Lehrende, um adäquate Veränderungsstrategien zu entwickeln und in der Praxis umzusetzen. Auf der Basis möglichst detaillierter Auswertungen sollen Lehrende Unterstützung erfahren und auf die individuellen Stärken und Schwächen zugeschnittene Lernangebote erhalten (Döring, 2002; Knödler, 2018).

## Kommunikationsmodell

Kommunikation ist die Basis einer jeden Phase des Evaluationsprozesses (Bargel, 2001). In Form einer transparenten Prozessdarlegung, bei der alle Beteiligten von Beginn an partizipieren, motiviert sie mitunter zur Teilnahme sowie zur Umsetzung weiterer Maßnahmen und stellt zusätzlich eine Voraussetzung für die Akzeptanz und Verankerung des Verfahrens dar (Zieldefinition, Bewertung, Konsequenzen in Form von Verbesserungsmaßnahmen und Zielvereinbarungen; Preißer, 2002). In der Literatur wird häufig der Zwischenschritt einer Selbstreflexion des Lehrenden und eines anschließenden Dialogs zwischen Studierenden und Lehrenden als „Königsweg für die Sicherung der Lehrqualität propagiert“ (Döring, 2002, S. 18), der zur Verbesserung des Lehrenden-Lernenden-Verhältnisses beitragen soll (Döring, 2002). In diesem

kommunikativen Austausch können Vorstellungen gemeinsam präzisiert und Konsequenzen gezogen werden (Michalk, 2007; Reissert & Carstensen, 1998). Neben der Kommunikation zwischen Studierenden und Lehrenden ist auch die Kommunikation von Lehrenden untereinander und in entsprechenden Hochschulgremien von Bedeutung (el Hage, 1996a). Das Kommunikationsmodells zielt auf Akzeptanz anstelle von Sanktionen oder Veröffentlichungen von Ergebnissen, mit wenig Widerstand zu rechnen ist. Kritisiert wird, dass die Lehrverbesserung vom Lehrenden abhängig ist, da mündliches Feedback zur Veranstaltung und die Besprechung quantitativ erhobener Fragebogendaten von der Fähigkeit und Motivation des Lehrenden abhängig ist. Neben der Bereitschaft, Veränderungen vorzunehmen, sind dafür soziale und kommunikative Kompetenzen von Seiten des Lehrenden nötig (el Hage, 1996a; Preißer, 1992, 2002; Knödler, 2018).

### Steuerungsmodell

Das Steuerungsmodell hat das Ziel, durch Sanktion- und Anreizsysteme Negatives zu unterbinden und Erwünschtes (hier hinsichtlich Lehre und Lehrverhalten) zu fördern (el Hage, 1996a). Von staatlicher Seite erfolgt eine Steuerung durch rechtliche Vorschriften (direkt), hauptsächlich aber durch Mittelzuweisungen (indirekt). Wird die Mittelzuweisung an Evaluationsergebnisse gekoppelt, steigt die Bedeutung von Evaluationsaktivitäten und Lehre (el Hage, 1996a). Auch Hochschulen können den effizienten Einsatz von Mitteln und die Verbesserung der Lehre auf der Grundlage von Evaluationsergebnisse steuern. Unter anderem kann die Steuerung durch Anreizmechanismen (z.B. Sondermittel) und Sanktionen (z.B. Verleihung negativer Lehrpreise oder Mittelkürzungen), Veröffentlichungen von Evaluationsergebnisse sowie durch die Kopplung an Einkommen und Berufungsverfahren erfolgen (el Hage, 1996a; Knödler, 2018).

## Modul 4: Umgang mit Evaluationsergebnissen in der (LVE-)Praxis

Blickt man nun von der Theorie in die (LVE-)Praxis, stellt sich nach wie vor häufig die Frage „Evaluationsergebnisse und dann?“. Aus langjähriger Evaluationserfahrung im Hochschulbereich und im Speziellen in der LVE sowie zahlreichen Beratungen zum Thema (Lehrveranstaltungs-)Evaluation ergibt sich ein recht einheitliches Bild von den wiederkehrenden Herausforderungen hinsichtlich des Umgangs mit den Ergebnissen:



Abb. 5: Herausforderungen im Umgang mit LVE-Ergebnissen (eigene Darstellung)

Wie diesen Herausforderungen begegnet werden kann, stehen jedoch zahlreiche Möglichkeiten gegenüber - sinnvoll ist hier ein direkter Transfer aus der Praxis in die Praxis.

# Literatur

- Abrami, P. C. (2001). Improving Judgments About Teaching Effectiveness Using Teacher Rating Forms. *New Directions for Institutional Research*, 109, 59–87.
- Aleamoni, L. M. & Thomas, G. S. (1980). Differential Relationships of Student, Instructor, and Course Characteristics to General and Specific Items on a Course Evaluation Questionnaire. *Teaching of Psychology*, 7(4), 233–235.
- Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning*, 31, 25–31
- Astleitner, H. (1991). Studentische Einschätzungen von universitärem Lehrverhalten: Das Problem impliziter Theorien. *Psychologie in Erziehung und Unterricht*, 38, 116–122
- Bargel, T. (2001). Studentische Urteile zur Lehr- und Studienqualität - Erwartungen, Erfahrungen und Probleme. *Konstanzer-Online-Publikations System: Konstanz*.
- Bargel, T. & el Hage, N. (2000). Evaluation der Hochschullehre: Modelle, Probleme und Perspektiven. *Zeitschrift für Pädagogik*, 41, 207–224.
- Barz, A. & Miethig, T. (1992). Evaluation als Steuerungsinstrument. In D. Grün (Hrsg.), *Informationen aus Lehre und Forschung: Vol. 1992,2. Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes?* (S. 169–182). Berlin: Presse- und Informationsstelle der Freien Universität Berlin.
- Becker-Richter, M., Habel, E., Rinke, B. & Tegethoff, H. G. (2002). *Evaluation von Studium und Lehre: Verfahren - Methoden - Erfahrungen*. Wiesbaden, s.l.: VS Verlag für Sozialwissenschaften.
- Beywl, W. (1992). Perspektiven der Evaluation im Hochschulwesen - Plädoyer für responsive Evaluationen. In D. Grün (Hrsg.), *Informationen aus Lehre und Forschung: Vol. 1992,2. Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes?* (S. 227–238). Berlin: Presse- und Informationsstelle der Freien Universität Berlin.
- Beywl, W. & Widmer, T. (2009). Evaluation in Expansion: Ausgangslage für den intersektoralen Dreiländer-Vergleich. In: Widmer, T., Beywl, W. & Fabian, C. (Hrsg.). *Evaluation. Ein systematisches Handbuch*, S.13–23. Wiesbaden: VS.
- Biehl, P. & Costanzo, K. (2014). Online Course Evaluations: An Institutional Approach: Committee Executive Summary. Zuletzt verfügbar unter: [http://faculty-senate.buffalo.edu/documents/CourseEvaluationsCommitteeExecSumRpt-Append-3%2012%2014.pdf\\*21.03.18](http://faculty-senate.buffalo.edu/documents/CourseEvaluationsCommitteeExecSumRpt-Append-3%2012%2014.pdf*21.03.18).

- Bledsoe, J. C. (1978). Insight into one's own teaching: stability of students' evaluations across classes. *Psychological Reports*, 42, 1071
- Boyer, E. L., Altbach, P. G. & Whitelaw, M. J. (1994). *The Academic Profession: An International Perspective. A special report / The Carnegie Foundation for the Advancement of Teaching*. Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.
- Boysen, G. A. (2015). Uses and Misuses of Student Evaluations of Teaching. *Teaching of Psychology*, 42(2), 109–118.
- Brandt, T. (2009). *Evaluation in Deutschland: Professionalisierungsstand und -perspektiven. Sozialwissenschaftliche Evaluationsforschung: Band 7*. Münster, München [u.a.]: Waxmann.
- Bruton, B. T. & Crull, S. T. (1982). Causes and consequences of student evaluation of instruction. *Research in Higher Education*, 17(3), 195–206
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erw. Aufl.). PS Psychologie. München: Pearson Studium.
- Bülow-Schramm, M. (2000). Evaluation und Qualitätsmanagement ein strategisches Instrument der Hochschulentwicklung? In A. Hanft (Hrsg.), *Hochschulwesen, HSW Wissenschaft und Praxis. Hochschulen managen? Zur Reformierbarkeit der Hochschulen nach Managementprinzipien* (S. 170–190). Neuwied: Luchterhand.
- Cadwell, J. & Jenkins, J. (1985). Effects of the semantic similarity of items on student ratings of instructors. *Journal of Educational Psychology*, 77(4), 383–393.
- Cashin, W. E. & Beneton, S. L. (2011). *Student Ratings of Teaching: A Summary of Research and Literature*. IDEA Paper, 50, 1–13
- Centra, J. A. (1980). *Determining faculty effectiveness: Assessing Teaching, Research, and Service For Personnel Decisions and Improvement*. San Francisco: Jossey-Bass.
- Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness* (1st ed.). The Jossey-Bass higher and adult education series. San Francisco: Jossey-Bass.
- Covert, R. W. & Mason, E. J. (2016). Factorial Validity of a Student Evaluation of Teaching Instrument. *Educational and Psychological Measurement*, 34(4), 903–905.
- Gralki, H. O. & Hecht, H. (1992). Hochschuldidaktische Aspekte der Beurteilung von Lehrveranstaltungen durch Studenten. In D. Grün (Hrsg.), *Informationen aus Lehre und Forschung: Vol. 1992,2. Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes?* (S. 99–114). Berlin: Presse- und Informationsstelle der Freien Universität Berlin.

- Dahler-Larsen, P. (2012). The Evaluation Society. Stanford: Stanford University Press.
- Döring, N. (2002). Für Evaluation und gegen Evaluitis. Warum und wie Lehrevaluation an deutschen Hochschulen verbessert werden sollte. In B. Berendt, H.-P. Voss & J. Wildt (Hrsg.), Raabe - nachschlagen, finden. Neues Handbuch Hochschullehre. Lehren und Lernen effizient gestalten (2nd ed.). Stuttgart: Raabe.
- Driscoll, L. A. & Goodwin, W. L. (1979). The Effects of Varying Information About Use and Disposition of Results on University Students' Evaluations of Faculty and Courses. *American Educational Research Journal*, 16(1), 25–37.
- el Hage, N. (1996a). Lehrevaluation und studentische Veranstaltungskritik: Projekte, Instrumente und Grundlagen. Bonn: Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.
- el Hage, N. (1996b). Studentische Urteile zur Lehr- und Studienqualität. Analysen zu ihrer Validität anhand der Indikatoren und Skalen im Studierendensurvey (Hefte zur Bildungs- und Hochschulforschung). Universität Konstanz
- Feldmann, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: a review and analysis. *Research in Higher Education*, 6, 223–274.
- Firth, M. (1979). Impact of work experience on the validity of student evaluations of teaching effectiveness. *Journal of Educational Psychology*, 71(5), 726–730.
- Friedrich, V.; Fischer, A. (Hg.) (2015), Wie nützt Evaluation (nicht)?, zoom Nr. 5. Bern: Zentrum für universitäre Weiterbildung ZUW der Universität Bern.
- Gesellschaft für Evaluation (2008). Standards für Evaluation. Zuletzt verfügbar unter [http://www.degeval.de/fileadmin/user\\_upload/Sonstiges/STANDARDS\\_2008-12.pdf](http://www.degeval.de/fileadmin/user_upload/Sonstiges/STANDARDS_2008-12.pdf) [17.07.2017].
- Giel, S. (2013). Theoriebasierte Evaluation. Konzepte und methodische Umsetzung. Münster (u. a.): Waxmann.
- Gollwitzer, M. (2006b). Das "Trierer Inventar zu Lehrevaluation - modular" TRIL-MOD. Workshop "Qualitätssicherung der Lehre", Trier. Zuletzt verfügbar unter: [http://www.uni-trier.de/fileadmin/elearning/Pr\\_sentation\\_Trier.pdf](http://www.uni-trier.de/fileadmin/elearning/Pr_sentation_Trier.pdf)\*08.01.18.
- Gollwitzer, M. & Scholtz, W. (2003). Das "Trierer Inventar zur Lehrveranstaltungsevaluation" (TRIL): Entwicklung und erste testtheoretische Erprobungen. In G. Krampen (Hrsg.), Materialien der Sektion Aus-, Fort- und Weiterbildung / Berufsverband Deutscher Psychologen: Vol. 7. Neue Medien, Konzepte, Untersuchungsbefunde und Erfahrungen zur psychologischen Aus-, Fort- und Weiterbildung (S. 114–128). Bonn: Dt. Psychologen-Verl.



- Google (2017). Suchbegriff Evaluation. Zuletzt verfügbar unter [https://www.google.de/search?q=evaluation&ie=utf-8&oe=utf-8&gws\\_rd=cr&ei=\\_ZdsWeadGYrPwALY1JeIAQ](https://www.google.de/search?q=evaluation&ie=utf-8&oe=utf-8&gws_rd=cr&ei=_ZdsWeadGYrPwALY1JeIAQ) [17.07.17].
- Hennefeld, V., Meyer, W. & Silvestrini, S. (2015). Einleitung. In: Dies. (Hrsg.). Nachhaltige Evaluation? Auftragsforschung zwischen Praxis und Wissenschaft, S. 7-14. Münster (u. a.): Waxmann.
- Henry, G. T. & Mark, M. M. (2003b). Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions. In: American Journal of Evaluation, 24(3), S. 293-314.
- Hense, J., Rädiker, S., Böttcher, W. & Widmer, T. (2013) (Hrsg.). Forschung über Evaluation. Bedingungen, Prozesse und Wirkungen. Münster (u. a.): Waxmann.
- Hochschulkonferenz (HRK). (1995). Zu Evaluation im Hochschulbereich unter besonderer Berücksichtigung der Lehre: EntschlieÙung des 176. HRKPlenums vom 3.7.1995. Zuletzt verfügbar unter: [https://www.hrk.de/positionen/beschluss/detail/zur-evaluation-imhochschulbereich-unter-besonderer-beruecksichtigung-der-lehre/\\*21.03.18](https://www.hrk.de/positionen/beschluss/detail/zur-evaluation-imhochschulbereich-unter-besonderer-beruecksichtigung-der-lehre/*21.03.18).
- Hochschulrektorenkonferenz. (1999). Gemeinsame Ziele: Evaluation, Qualitätssicherung und Akkreditierung in Deutschland und der Mongolei. Beiträge zur Hochschulpolitik, 7, 1–79.
- Holtkamp, R. & Reissert, R. (1992). Evaluation der Lehre im Vergleich – Voraussetzungen und Verfahren an den Hochschulen der Bundesrepublik Deutschland, der Vereinigten Staaten, Großbritanniens und der Niederlande. In D. Grün (Hrsg.), Informationen aus Lehre und Forschung: Vol. 1992,2. Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes? (S. 61–73). Berlin: Presse- und Informationsstelle der Freien Universität Berlin
- Jaeger, M. & Kerst, C. (2010). Potentiale und Nutzen von Absolventenbefragungen für das Hochschulmanagement. Beiträge zur Hochschulforschung, 4, 8– 23.
- Jenkins, J. (1987). Implicit theories and semantic similarities: Reply to Marsh and Groves. Journal of Educational Psychology, 79(4), 490–493.
- Knödler, 2018. Evaluation an Hochschulen. Entwicklung und Validierung eines verhaltensbasierten Messinventars zur studentischen Lehrveranstaltungsevaluation. Wiesbaden: Springer VS.
- Kromrey, H. (1994). Wie erkennt man "gute Lehre"? Was studentische Vorlesungsbefragungen (nicht) aussagen. Empirische Pädagogik, 8(2), 153–168
- Kromrey, H. (1995). Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In P. P. Mohler & W. Bandilla (Hrsg.), Universität und Lehre. Ihre Evaluation als Herausforderung an die empirische Sozialforschung (2nd ed.). Münster: Waxmann.

- Kromrey, H. (2006). Qualität und Evaluation im System Hochschule. In R. Stockmann (Hrsg.), Sozialwissenschaftliche Evaluationsforschung: Vol. 1. Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder (3rd ed., S. 234–259). Münster: Waxmann.
- Kuzmanovic, M., Savic, G., Popovic, M. & Martic, M. (2013). A new approach to evaluation of university teaching considering heterogeneity of students' preferences. *Higher Education*, 66(2), 153–171.
- Mark, M. M., Henry, G. T. & Julnes, G. (2000). *Evaluation. An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs*. San Francisco: Jossey-Bass.
- Mark, M. M., Greene, J. C. & Shaw, I. F. (2006). The Evaluation of Policies, Programs and Practices. In: Shaw, I. F., Greene, J. C. & Mark, M. M. (Hrsg.). *Handbook of Evaluation*, S. 1–30. London (u. a.): SAGE.
- Marques, T. E., Lane, D. M. & Dorfman, P. W. (1979). Toward the development of a system for instructional evaluation: Is there consensus regarding what constitutes effective teaching? *Journal of Educational Psychology*, 71(6), 840–849.
- Marsh, H. W. (1982a). The Use of Path Analysis to Estimate Teacher and Course Effects in Student Ratings of Instructional Effectiveness. *Applied Psychological Measurement*, 6(1), 47–59.
- Marsh, H. W. & Hocevar, D. (1984). The Factorial Invariance of Student Evaluations of College Teaching. *American Educational Research Journal*, 21(2), 341–366.
- Marsh, H. W. & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching & Teacher Education*, 7(4), 303–314.
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76(5), 707–754.
- Marsh, H. W. & Overall, J. U. (1981). The Relative Influence of Course Level, Course Type, and Instructor on Student's Evaluations of College Teaching. *American Educational Research Journal*, 18(1), 103–112.
- Marsh, H. W. & Roche, L. (1993a). The Use of Students' Evaluations and an Individually Structured Intervention to Enhance University Teaching Effectiveness. *American Educational Research Journal*, 30(1), 217–251.
- Marsh, H. W. & Roche, L. (1993b). The Use of Students' Evaluations and an Individually Structured Intervention to Enhance University Teaching Effectiveness. *American Educational Research Journal*, 30(1), 217–251.
- Meredith, G. M. (2016). Preferred Length of Scales for Students' Evaluation of Instruction. *Perceptual and Motor Skills*, 53(2), 490.
- Michalk, B. (2007). *Verfahren der Qualitätssicherung und Qualitätsentwicklung: Projekt Qualitätsmanagement. Beiträge zur Hochschulpolitik: Vol. 2007,8*. Bonn: Hochschulrektorenkonferenz.

- Neumann, L. & Neumann, Y. (1981). Comparison of six lengths of rating scales: students' attitudes toward instruction. *Psychological Reports*, 48, 399–404.
- Orpen, C. (1980). The susceptibility of student evaluation of lecturers to situational variables. *Higher Education*, 9(3), 293–306
- Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed). Thousand Oaks, Calif.: Sage Publications.
- Pawson, R. & Tilley, N. (1997b). *Realistic Evaluation*. London (u .a.): SAGE.
- Preißer, R. (1992). Verwirklichungsbedingungen der Evaluation der Lehre und der Verbesserung der Lehre – Konsequenzen aus den bisherigen Erfahrungen mit Lehrveranstaltungskritiken. In D. Grün (Hrsg.), *Informationen aus Lehre und Forschung: Vol. 1992,2. Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes?* (S. 198–217). Berlin: Presse- und Informationsstelle der Freien Universität Berlin.
- Preißer, R. (2002). Evaluation der Hochschullehre und Selbststeuerung des Lernens. In B. Berendt, H.-P. Voss & J. Wildt (Hrsg.), *Raabe - nachschlagen, finden. Neues Handbuch Hochschullehre. Lehren und Lernen effizient gestalten* (2nd ed.). Stuttgart: Raabe
- Raser, M. (2016). "Zücken Sie ihr Smartphone, die Lehrevaluation steht an". DeGeval Jahrestagung, Salzburg.
- Raser, M. & Hense, J. (2017). Was geschieht nach der Ergebnissrückmeldung? In R. Kordts-Freudinger, D. Al-Kabbani & N. Schaper (Hrsg.), *Blickpunkt Hochschuldidaktik: v.131. Hochschuldidaktik im Dialog. Beiträge der Jahrestagung der Deutschen Gesellschaft für Hochschuldidaktik (dghd) 2015* (1st ed., S. 193–211). Bielefeld: Bertelsmann Verlag.
- Reissert, R. (1992). *HIS-Dokumentation: Evaluation der Lehre: Aktuelle Aktivitäten an deutschen Hochschulen*. Hannover: HIS.
- Reissert, R. & Carstensen, D. (1998). *Praxis der internen und externen Evaluation: Handbuch zum Verfahren*. Hannover: HIS.
- Rindermann, H. & Amelang, M. (1994a). *Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation: (HILVE) ; Handanweisung*. Heidelberg: Asang
- Rindermann, H. & Amelang, M. (1994b). Entwicklung und Erprobung eines Fragebogens zur studentischen Veranstaltungsevaluation. *Empirische Pädagogik*, 8(2), 131–151.
- Rindermann, H. (1996). *Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen*. Zugl.: Heidelberg, Univ., Diss., 1995. *Psychologie Pädagogische Psychologie: Vol. 6*. Landau: Verl. Empirische Pädagogik.

- Rindermann, H. (2009). *Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts* (2. Aufl.). Psychologie: Bd. 42. Landau: Empirische Pädagogik
- Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, 9, 222–245.
- Roszkowski, M. J. & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, 35(1), 113–130.
- Saris, W. E., Revilla, M., Krosnick, J. A. & Shaeffer, E. M. (2010). Comparing Questions with Agree/Disagree Response Options to Questions with ItemSpecific Response Options. *Survey Research Methods*, 4(1), 61–79.
- Schnell, R. & Kopp, J. (2000). Theoretische und methodische Diskussionen der Lehrevaluationsforschung und deren praktische Bedeutung. *KonstanzerOnline-Publikations-System*: Konstanz.
- Scriven, M. (1972). Die Methodologie der Evaluation. In Wulf & Christoph (Hrsg.), *Erziehung in Wissenschaft und Praxis*: Vol. 18. Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen (Vol. 18, S. 60-91). München: R. Piper & Co. Verlag; R. Piper.
- Scriven, M. (1991a). *Evaluation thesaurus* (4th ed., [Nachdr.]). Newbury Park, Calif: Sage Publications.
- Spooren, P., Brockx, B. & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching. *Review of Educational Research*, 83(4), 598–642.
- Stamm, M. (2003). Evaluation und ihre Folgen für die Bildung: Eine unterschätzte pädagogische Herausforderung. *Internationale Hochschulschriften*: Bd. 419. Münster, New York, München, Berlin: Waxmann.
- Stockmann, R. (2004). Was ist eine gute Evaluation? Einführung zu Funktionen und Methoden von Evaluationsverfahren, CEval-Arbeitspapiere, 9. Saarbrücken: Centrum für Evaluation. Zuletzt verfügbar unter [http://www.ceval.de/modx/fileadmin/user\\_upload/PDFs/workpaper9.pdf](http://www.ceval.de/modx/fileadmin/user_upload/PDFs/workpaper9.pdf) [18.07.2017].
- Stockmann, R. (2006). Evaluation und Qualitätsentwicklung: Eine Grundlage für wirkungsorientiertes Qualitätsmanagement. *Sozialwissenschaftliche Evaluationsforschung*: Bd. 5. Münster, New York, München, Berlin:
- Stockmann, R. & Meyer, W. (2014). *Evaluation: Eine Einführung* (2., überarb. und aktualisierte Aufl.). UTB Sozialwissenschaften: Vol. 8337. Opladen, Stuttgart: Budrich; UTB.
- Süllwold, F. (1992). Welche Realität wird bei der Beurteilung von Hochschullehrern durch Studierende erfaßt? *Mitteilungen des Hochschulverbandes*, 40, 34–35.

- von Werthern, A. (2020). Theoriebasierte Evaluation. Entwicklung und Anwendung eines Verfahrensmodells zur Programmtheoriekonstruktion. Wiesbaden (u. a.): Springer VS.
- von Werthern, A. & Knödler, E. (2020). Handreichung zur Online-Lehre. Evaluation von Hochschullehre. Verfügbar unter: [https://www.profil.uni-muenchen.de/profil-start/profil\\_hands-on-lehre\\_eval.pdf](https://www.profil.uni-muenchen.de/profil-start/profil_hands-on-lehre_eval.pdf) \*20.02.24.
- Vedung, E. (2010). Four Waves of Evaluation Diffusion. In: Evaluation, 16(3), S. 263-277.
- Voges, M. (2017). Die Bedeutung von Evaluationen für die Qualitätsentwicklung im Bildungswesen. Zeitschrift für Evaluation, 2, 183–186.
- Webler, W.-D. (1992). Evaluation der Lehre - Erfahrungen und Methodenhinweise. In D. Grünh (Hrsg.), Informationen aus Lehre und Forschung: Vol. 1992,2. Evaluation von Lehrveranstaltungen. Überfrachtung eines sinnvollen Instrumentes? (S. 143–161). Berlin: Presse- und Informationsstelle der Freien Universität Berlin.
- Webler, W.-D., Domeyer, V., Rommel, H. & Schiebel, B. (1993). Lehrberichte: Empirische Grundlagen, Indikatorenauswahl und Empfehlungen zur Darstellung der Situation der Lehre in Lehrberichten. Schriftenreihe Studien zu Bildung und Wissenschaft: Vol. 107. Bad Honnef: Bock.
- Wehrlin, U. (2011b). Hochschul-Change-Management: Veränderungsprozesse, Change Management, Organisationsentwicklung, Professionalisierung des Hochschulmanagements. Hochschul- und Wissenschaftsmanagement: Vol. 8. München: AVM.
- Weiss, C. H. (1974). Evaluierungsforschung: Methoden zur Einschätzung von sozialen Reformprogrammen. Studienbücher zur Sozialwissenschaft: Vol. 10. Opladen: Westdt. Verl.
- Whitely, S. E. & Doyle, K. O. (1976). Implicit Theories in Student Ratings. American Educational Research Journal, 13(4), 241–253.
- Widmer, T. & De Rocchi, T. (2012). Evaluation. Grundlagen, Ansätze und Anwendungen. Glarus: Ruegger.
- Wissenschaftsrat. (1996). Empfehlungen zur Stärkung der Lehre in den Hochschulen durch Evaluation. Zuletzt verfügbar unter: [https://www.wissenschaftsrat.de/download/archiv/2365-96.pdf\\*21.03.18](https://www.wissenschaftsrat.de/download/archiv/2365-96.pdf*21.03.18).
- Wittmann, W. W. (2009). Evaluationsmodelle. In: Holling, H. (Hrsg.). Grundlagen und statistische Methoden der Evaluationsforschung, S. 59-98. Göttingen (u. a.): Hogrefe.
- Wolbring, T. (2013). Fallstricke der Lehrevaluation: Möglichkeiten und Grenzen der Messbarkeit von Lehrqualität. Zugl.: München, Univ., Diss. 2013 u.d.T.: Wolbring, Tobias: Methodische Fallstricke bei der Evaluation universitärer Lehre (1. Aufl.). Hochschule und Gesellschaft: Vol. 1. Frankfurt am Main: Campus Verlag.

Wollmann, H. (2017). Entwicklungslinien von Evaluation und Evaluationsforschung in Deutschland. Zwischen Vorgeschichte, Aufbruch und Konsolidierung. In R. Stockmann, W. Böttcher, A. Caspari, G.-M. Hellstern, P. Mayring, H. Moosbrugger & C. Spiel (Hrsg.), Zeitschrift für Evaluation. Themenheft: Zwanzig Jahre DeGEval (S. 33–53). Waxmann Verl.