

Zimmermann, Tobias

Leistungsbeurteilungen an Hochschulen lernförderlich gestalten. Prüfen, Beurteilen und Rückmelden von Lernleistungen

Opladen • Berlin • Toronto : Verlag Barbara Budrich 2024, 404 S.



Quellenangabe/ Reference:

Zimmermann, Tobias: Leistungsbeurteilungen an Hochschulen lernförderlich gestalten. Prüfen, Beurteilen und Rückmelden von Lernleistungen. Opladen • Berlin • Toronto : Verlag Barbara Budrich 2024, 404 S. - URN: urn:nbn:de:0111-pedocs-308719 - DOI: 10.25656/01:30871; 10.3224/84743045

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-308719>

<https://doi.org/10.25656/01:30871>

in Kooperation mit / in cooperation with:



<https://www.budrich.de>

Nutzungsbedingungen

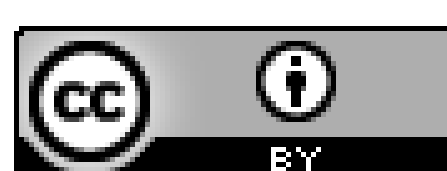
Dieses Dokument steht unter folgender Creative Commons-Lizenz: <http://creativecommons.org/licenses/by/4.0/deed.de> - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License: <http://creativecommons.org/licenses/by/4.0/deed.en> - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to the public as long as you attribute the work in the manner specified by the author or licensor.

By using this particular document, you accept the above-stated conditions of use.



Kontakt / Contact:

peDOCS

DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation

Informationszentrum (IZ) Bildung

E-Mail: pedocs@dipf.de

Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

Tobias Zimmermann

Leistungsbeurteilungen an Hochschulen lernförderlich gestalten

Prüfen, Beurteilen und Rückmelden von Lernleistungen



Tobias Zimmermann
Leistungsbeurteilungen an Hochschulen
lernförderlich gestalten

Tobias Zimmermann

Leistungsbeurteilungen an Hochschulen lernförderlich gestalten

Prüfen, Beurteilen und Rückmelden
von Lernleistungen

Verlag Barbara Budrich
Opladen • Berlin • Toronto 2024

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

© 2024 Dieses Werk ist bei der Verlag Barbara Budrich GmbH erschienen und steht unter der Creative Commons Lizenz Attribution 4.0 International (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/>
Diese Lizenz erlaubt die Verbreitung, Speicherung, Vervielfältigung und Bearbeitung unter Angabe der Urheber*innen, Rechte, Änderungen und verwendeten Lizenz.
www.budrich.de



Die Verwendung von Materialien Dritter in diesem Buch bedeutet nicht, dass diese ebenfalls der genannten Creative-Commons-Lizenz unterliegen. Steht das verwendete Material nicht unter der genannten Creative-Commons-Lizenz und ist die betreffende Handlung gesetzlich nicht gestattet, ist die Einwilligung des jeweiligen Rechteinhabers für die Weiterverwendung einzuholen. In dem vorliegenden Werk verwendete Marken, Unternehmensnamen, allgemein beschreibende Bezeichnungen etc. dürfen nicht frei genutzt werden. Die Rechte des jeweiligen Rechteinhabers müssen beachtet werden, und die Nutzung unterliegt den Regeln des Markenrechts, auch ohne gesonderten Hinweis.

Dieses Buch steht im Open-Access-Bereich der Verlagsseite zum kostenlosen Download bereit (<https://doi.org/10.3224/84743045>).
Eine kostenpflichtige Druckversion (Print on Demand) kann über den Verlag bezogen werden. Die Seitenzahlen in der Druck- und Onlineversion sind identisch.

ISBN 978-3-8474-3045-2 (Paperback)
eISBN 978-3-8474-1979-2 (PDF)
DOI 10.3224/84743045

Umschlaggestaltung: Bettina Lehfeldt, Kleinmachnow – www.lehfeldtgraphic.de
Titelbildnachweis: KI-generiert von DALL-E /OpenAI
Satz: Angelika Schulz, Zülpich
Lektorat: Dr. Daniela Gastell, Berlin

„Lernen ergibt sich aus dem, was die Studierenden tun und denken, und nur aus dem, was die Studierenden tun und denken. Lehrende können das Lernen nur fördern, indem sie Einfluss darauf nehmen, was die Studierenden tun, um zu lernen.“

Herbert A. Simon (zitiert nach Ambrose et al. 2010, 1, Übers. tz)

„Wettbewerbe und Prüfungen sind heutzutage in unterschiedlichsten Formaten in fast allen Lebensbereichen institutionalisiert. Sie zielen darauf ab, individuelle Leistungsunterschiede vermeintlich objektiv zu ermitteln und zu belohnen. Und genau damit stellen sie solche Unterschiede maßgeblich her.“

Nina Verheyen (2014, 67)

Dank

Das vorliegende Buch habe ich im Sinne des Urheberrechts allein verfasst. Dennoch hätte es ohne die Unterstützung durch viele Menschen nicht entstehen können. All diesen Personen gebührt mein herzlicher Dank.

Ein durch meine Arbeitgeberin, die Pädagogische Hochschule Zürich, gewährtes Weiterbildungs- und Forschungssemester verschaffte mir die nötige Zeit, um das Buch zu schreiben. Ebenso danke ich dem Verlag Barbara Budrich für die Möglichkeit, dieses Buch zu veröffentlichen, und dem institutionellen Open-Access-Fonds der PH Zürich (iOAF) für die Übernahme der Kosten für die Open-Access-Publikation.

Besonders danken möchte ich:

Geri Thomann, der mir über Jahre hinweg die Auseinandersetzung mit dem Themenfeld der Leistungsbeurteilung ermöglicht hat und dessen Gedanken zum Thema für mich wegweisend waren. Ich danke ihm zudem für die Durchsicht des gesamten Buchmanuskripts und die daraus resultierenden wertvollen Anregungen.

Markus Weil, der mir als Vorgesetzter während meiner schreibbedingten Abwesenheit den Rücken freigehalten und das Erscheinen dieses Buches auf verschiedenen Ebenen unterstützt hat.

Meinen Kolleginnen und Kollegen aus dem Zentrum für Hochschuldidaktik und -entwicklung der Pädagogischen Hochschule Zürich: Dagmar Engfer, Mònica Feixas, Gabriel Flepp, Simone Heller-Andrist, Petra Weiss und Franziska Zellweger Moser. Dem regen fachlichen Austausch mit ihnen verdanke ich viele Einsichten, die dieses Buch geprägt haben. Zudem haben sie mir wertvolle Anregungen zum Kapitel über Grundlagen der Beurteilung gegeben, Gabriel Flepp außerdem auch zum KI-Kapitel.

Wolfgang Bühner, der große Teile des Buchmanuskripts aus der Sicht eines Hochschuldozenten kommentiert und somit geholfen hat, das Buch an den Bedürfnissen von Lehrenden aus verschiedensten wissenschaftlichen Disziplinen auszurichten.

Viviane Zimmermann für die präzise Lektüre und kritische Diskussion mehrerer Buchkapitel sowie die Unterstützung bei sozialwissenschaftlichen und statistischen Fragen.

Urs Siegfried für die anregenden Kommentare und Diskussionen zum Kapitel über Leistung und Gesellschaft sowie für unsere Gespräche über Gott und die Welt.

Susan D. Blum, Gavin T.L. Brown, Sean Kearney und Philippe Wampfler, mit denen ich parallel zum Verfassen dieses Buches im Rahmen einer Delphi-Studie inspirierende Interviews führen durfte zur Frage, wie sich

künstliche Intelligenz auf die Leistungsbeurteilung an Hochschulen auswirken wird.

Daniela Gastell für das kompetente Lektorat und Korrektorat des gesamten Manuskripts und Paula Schmieding vom Verlag Barbara Budrich für die zuverlässige Betreuung des Buchprojekts von der ersten Anfrage bis zur Veröffentlichung.

Zudem danke ich meiner Frau Viviane, unseren Kindern Nina und Jonas und meinen Freunden herzlich für ihre Unterstützung, ihre Geduld angesichts meines großen Zeitaufwands und ihre Offenheit gegenüber meiner Begeisterung für das Buchthema und das Schreiben.

Tobias Zimmermann

Inhaltsverzeichnis

Dank	7
1 Einleitung: Leistungsbeurteilung und Leistungsrückmeldung	13
1.1 Herausforderungen der Leistungsbeurteilung	13
1.2 Die Beurteilung und ihre Reflexion	18
Teil I: Grundlagen	23
2 Leistung in Gesellschaft und Bildungswesen	25
2.1 Leistung als Begriff und Wahrnehmungskonzept	25
2.2 Leistung und Meritokratie	30
2.3 Noten machen Leistungen	37
2.4 Metrische Fixierung und die Zertifikatsgesellschaft	45
2.5 Soziale Konstruktion von Lernleistung und ihrer Beurteilung	53
3 Funktionen von Leistungsbeurteilung	61
3.1 Gesellschaftliche Funktionen von Leistungsbeurteilung – Berechtigung	62
3.2 Didaktisch-lernbezogene Funktionen von Leistungsbeurteilung – Fördern	64
3.3 Dilemmata bei der Leistungsbeurteilung	67
4 Was ist Lernen und wie funktioniert es?	79
4.1 Zwei Lernexperimente	79
4.2 Lernen ist Konstruktion in einem Netzwerk	83
4.3 Neuronale Netzwerke	87
4.4 Drei zentrale Lernmechanismen in neuronalen Netzwerken	90
4.5 Gefühle, Motivation und Bedeutsamkeit	95
4.6 Aufbau komplexer Wissensstrukturen: Zeit und Expertise	101
4.7 Argumentatives Schlussfolgern als sozialer Prozess	109
4.8 Überleitung zum zweiten Teil des Buches	113

Teil II: Anwendung 115

5 Was sollen Studierende eigentlich können? Lernziele im Constructive Alignment 117

- 5.1 Lehre vom Ziel her denken: Constructive Alignment 118
- 5.2 Planungsvorlage für Lernziele 125
- 5.3 Übergeordnete Lernziele identifizieren und formulieren 127
- 5.4 Spezifische Lernziele identifizieren und formulieren 138
- 5.5 Epilog: Von Lernzielen und Kompetenzen 151

6 Voraussetzungen einer hochwertigen Leistungsbeurteilung 155

- 6.1 Wissen und Können als latente Konstrukte 155
- 6.2 Bezugsnormen der Beurteilung 158
- 6.3 Gütekriterien der Leistungsbeurteilung 162
- 6.4 Kognitive Verzerrungen und Beurteilungsfehler 170
- 6.5 Qualität der Leistungsbeurteilung sicherstellen 176

7 Formate von Leistungsnachweisen 181

- 7.1 Grundeigenschaften von Leistungsnachweisen 181
- 7.2 Aufgabenstellungen 187
- 7.3 Aufgabenkombinationen und Sammelnachweise 196
- 7.4 Strukturierungsansätze 199
- 7.5 Beschreibungsschema für Leistungsnachweisformate 202
- 7.6 Beispiele für „klassische“ Formate 204
- 7.7 Beispiele für innovativere Formate 218
- 7.8 Auflistung weiterer innovativer Formate 226

8 Formative Leistungsrückmeldung 229

- 8.1 Was ist eine formative Leistungsrückmeldung? 229
- 8.2 Leistungsrückmeldungen und Motivation 232
- 8.3 Inhaltliche Eigenschaften von Leistungsrückmeldungen 239
- 8.4 Feedbackinstanzen 246
- 8.5 Medialität von Feedback 258
- 8.6 Feedbackbasiertes Handeln: Schließen des Kreislaufs 265

9	Summativ beurteilen: Bewerten und Benoten	271
9.1	Die Wirkung von Bewertungen auf Motivation und Selbstregulation	272
9.2	Verschiedene Konzeptionen von Leistungsbeurteilung	274
9.3	Beurteilungsraster und andere Beurteilungsinstrumente	281
9.4	„Berechnung“ von Noten.....	298
9.5	Weniger ist mehr: Basisbewertungen und Prädikate	316
9.6	Grundlegendere Alternativen zur klassischen summativen Beurteilung	317
	Teil III: Spotlights.....	327
10	Leistungsnachweise im Curriculum	329
10.1	Bewertungen weisen Wert zu	329
10.2	Deklaratives und prozedurales Wissen im Studienverlauf.....	331
10.3	Vertikale und horizontale Integration	331
11	Leistungsbeurteilung und Prüfungsangst	335
11.1	Definition und Häufigkeit von Prüfungsangst	335
11.2	Die wichtigsten Ursachen von Prüfungsangst	337
11.3	Behandlungsmöglichkeiten von Prüfungsangst	338
11.4	Angstauslösende und angsthemmende Aspekte von Leistungsbeurteilung.....	339
12	Wissenschaftliche Arbeiten begleiten und beurteilen	343
12.1	Besonderheiten beim Begleiten wissenschaftlicher Arbeiten	343
12.2	Betreuungsverhältnis kontraktieren	344
13	Multiple-Choice-Prüfungen.....	347
13.1	Zum Begriff „Multiple Choice“	347
13.2	Was leisten MC-Prüfungen – und was nicht?	348
13.3	Struktur und Typen von MC-Aufgaben	349
13.4	Ungewollte Lösungshinweise vermeiden	352
13.5	Auswertung von MC-Prüfungen.....	353
13.6	MC-Tests als formative Selbsttests.....	354

14 Bildungsgeschichtliche Vorläufer von Leistung als schulischem Leitkonzept	357
14.1 Begriffsgeschichte	357
14.2 Ehre und Ehrtrieb als Leitkonzept der Jesuitenschulen	357
14.3 Meritentafeln und der Übergang zu Leistung als Leitkonzept.....	358
15 Künstliche Intelligenz und menschliche Leistung	361
15.1 Kulturelle Evolution oder der Mensch als Cyborg	363
15.2 Neue Arbeitsteilung zwischen Mensch und Maschine	364
15.3 Veränderungen im institutionell organisierten Lehren und Lernen.....	366
15.4 Veränderungen in der Beurteilungskultur.....	369
15.5 Offene ethische Fragen	370
16 Auflösung und Erläuterungen zu den Aufgaben in Kapitel 4.1.1.....	373
16.1 Auflösung Aufgabe 1b.....	373
16.2 Auflösung Aufgabe 2b.....	373
Literatur	375



QR-Code: Zugang zur Seite der zusätzlichen Materialien:
<https://doi.org/10.3224/84743045A>

1 Einleitung: Leistungsbeurteilung und Leistungsrückmeldung

Das Beurteilen von Leistungen, das Erteilen von Noten und das Formulieren von Leistungsrückmeldungen an Studierende sind anspruchsvolle Tätigkeiten. Viele Dozierende führen nicht gerne Leistungsnachweise durch und mögen die Notengebung noch weniger. Studierenden Feedback zu ihrem Lernfortschritt zu geben, stellt zwar für viele Lehrende kein Reizthema dar, aber im Alltag fehlt ihnen doch oft die Zeit für differenzierte Rückmeldungen. Die nachfolgenden Fallvignetten bieten anhand konkreter Beispiele Einblick in die wichtigsten Themen des vorliegenden Buches (Kap. 1.1). Im zweiten Teil der Einleitung werfe ich einen ersten kritischen Blick auf die Themen Leistung und Beurteilung, um darauf aufbauend die wichtigsten Ziele des Buches zu erläutern (Kap. 1.2).

1.1 Herausforderungen der Leistungsbeurteilung

Der Einstieg über eine Reihe von Fallvignetten gibt der sozialen Realität Raum, in der Beurteilungssituationen für alle Beteiligten auf vielschichtige Weise eingebettet sind. Die Vignetten dienen dabei als praxisnahe Beispiele, die typische Herausforderungen der Leistungsbeurteilung und Leistungsrückmeldung an Hochschulen illustrieren. Die Fälle sind bewusst gewählt, um die Dynamik und Mehrdeutigkeit der hochschulischen Lebens- und Erfahrungswelt an den Beginn dieses Buchs zu stellen und zur Reflexion anzuregen. Dies soll eine direkte Verbindung zu Ihrer eigenen Praxis herstellen und die Relevanz der Thematik greifbar machen.

Wenn Wissen nicht ausreicht

Im vergangenen Semester unterrichtete ich erstmals ein Aufbaumodul in Sozialstatistik (3. Semester, Bachelor in Politikwissenschaft). Mein Unterrichtsziel war es, den Studierenden nicht nur das Verständnis statistischer Konzepte zu vermitteln, sondern auch deren Anwendung auf reale Forschungsfragen zu fördern. Das Modul verlief angenehm, die Studierenden arbeiteten engagiert. Entsprechend hatte ich hohe Erwartungen an die Abschlussprüfung.

Die Prüfungsauswertung war deshalb für mich ein Schock: Die Leistungen der Studierenden lagen weit unter meinen Erwartungen. Die meisten Studierenden zeigten zwar ein ansprechendes Wissen über Begriffe, Konzepte und rechnerische Verfahren, hatten aber große Schwierigkeiten bei der Anwendung der Verfahrensweisen auf reale Fragestellungen. Rückblickend wurde mir bewusst, dass die Studierenden bereits im Laufe des Semesters einige Mühe gezeigt hatten, die Lerninhalte auf reale Forschungsfragen anzuwenden. Da sie jedoch die Übungen mehrheitlich rechtzeitig und vollständig bearbeiteten und gut vorbereitet in die Modulsitzungen kamen, hatte ich diesem Umstand keine große Beachtung geschenkt.

Das Prüfungsergebnis wirft für mich Fragen zur Wirksamkeit meines Lehransatzes auf: Macht mein Fokus auf Anwendung und kritisches Denken überhaupt Sinn? Was habe ich übersehen? Was kann ich beim nächsten Mal anpassen, damit die Studierenden besser für Anwendungsaufgaben gerüstet sind?

Felicia Graf, Assistentin für Politikwissenschaft

Dass Frau Graf aufgrund der schlechten Prüfungsergebnisse ihr ganzes Vorgehen hinterfragt, ist nachvollziehbar. Wenn wir einmal die Möglichkeit ausklammern, dass ihre Erwartungen an die Anwendungsfähigkeiten schlicht zu hoch waren, bieten die in diesem Buch erläuterten Konzepte folgende Erklärungsmöglichkeiten:

- Die beschriebenen Studierenden scheinen sich auf das Lernen von Begriffs- und Konzeptwissen zu fokussieren. Offenbar sind sie nicht gewohnt, erworbenes Wissen praktisch anzuwenden. Vermutlich haben sie nicht verstanden, dass dies in der Prüfung entscheidend sein würde. Hier zeigt sich, dass Lernziele definiert, kommuniziert und mit den Studierenden besprochen werden müssen. Nur so können diese wissen, welches Wissen und Können sie sich aneignen müssen (siehe Kap. 5). Damit verbunden ist das Definieren und Besprechen klarer Beurteilungskriterien, die auf diesen Lernzielen beruhen (siehe Kap. 9). Die Kenntnis und das Verstehen von Lernzielen und Beurteilungskriterien sind notwendige Voraussetzungen für Studierende, um sich gezielt auf eine Prüfung vorzubereiten. Dies trifft umso mehr zu, wenn die Anforderungen vom vertrauten Muster abweichen – so wie im geschilderten Fall, bei dem die Studierenden offenbar vor allem Prüfungen kennen, bei denen sie deklaratives Wissen auswendig lernen und wiedergeben müssen. Auswendiglernen ist aber bei Leistungsnachweisen, die auf tiefgreifendes Verstehen zielen, keine sinnvolle Lernstrategie (siehe Kap. 4.2).
- Die Fallbeschreibung lässt offen, ob und in welcher Weise die Studierenden Feedback zu den erwähnten Übungen erhalten haben. Da sie auch am

Semesterende noch große Mühe mit Anwendungsaufgaben hatten, war ihnen vermutlich die Bedeutung dieser Anforderung nicht bewusst. Das bedeutet, dass sie entweder zu wenig oder zu unklares Feedback zu den gelösten Übungen erhalten haben. Für den Lernprozess sind formative Leistungsrückmeldungen aber zentral (siehe Kap. 3 und 8).

- Eine weitere Erklärungsmöglichkeit besteht darin, dass die Anwendungsaufgaben der Abschlussprüfung deutlich von den Übungsaufgaben abweichen. In diesem Fall ergäbe sich eine Diskrepanz zwischen Lehr-Lern-Aktivitäten und Leistungsnachweis (Bruch im Constructive Alignment, vgl. Kap. 5.1).

Aus diesen Erklärungen ergeben sich einige konkrete Anpassungsideen: Felicia Graf könnte bei der nächsten Durchführung des Moduls zu Beginn des Semesters mehr Zeit aufwenden, um mit den Studierenden die Lernziele (siehe Kap. 5.2 bis 5.4) und die Beurteilungskriterien (siehe Kap. 9.3) zu besprechen. Zudem dürfte es sinnvoll sein, den Studierenden mehr Möglichkeiten zu geben, formative Rückmeldungen zu ihrem Lernen zu erhalten, sowie Anreize zu schaffen, dieses Feedback auch umzusetzen. Dabei sollte der Schwerpunkt auf Anwendungsaufgaben liegen, und insbesondere Peer-Feedback könnte eine zusätzliche Lernwirkung erzeugen (siehe Kap. 8).

Neues Format, neue Muster?

Thea Schleiden, Professorin für Zellbiologie, ist bekannt für ihre begeisterten Lehrveranstaltungen. Sie möchte den Studierenden ein tiefes Verständnis für die Zellbiologie vermitteln, führt inspirierende Diskussionen und bewirkt ein hohes Engagement der Studierenden. Diese schätzen, dass die erfolgreiche Forscherin ihnen auf Augenhöhe begegnet. An Klausuren hat Frau Schleiden allerdings wenig Interesse, da sie am Zusammenhang zwischen Klausuren und Lernfortschritt zweifelt. Sie lässt sie deshalb größtenteils von ihren Mitarbeitenden durchführen.

Als die Fakultät eine neue Studien- und Prüfungsordnung einführt, wird es möglich, Vorlesungen mit einer Posterkonferenz als Leistungsnachweis abzuschließen. Thea Schleiden sieht darin eine Möglichkeit, Klausuren zu vermeiden, und informiert die Studierenden in der ersten Vorlesungswoche über das neue Vorgehen. Während des Semesters freut sie sich, dass die Studierenden keine Fragen mehr zum Prüfungsstoff stellen.

Am Semesterende ist sie jedoch enttäuscht: Viele Poster weisen Ungenauigkeiten bei zentralen Konzepten der Zellbiologie auf und machen die zu präsentierenden Laborexperimente kaum verständlich. Dies steht im Widerspruch zur engagierten Mitarbeit der Studierenden während des Semesters. Spontan fragt die Professorin die Studierenden am Ende der Posterkonfe-

renz, ob sie ihre Poster gelungen fänden. Die teilweise emotionalen Reaktionen der Studierenden deuten darauf hin, dass viele nicht verstanden haben, wie die Poster zu gestalten sind und wie die Beurteilung erfolgt. Es wird auch deutlich, dass viele zum ersten Mal ein Poster erstellt haben und sich bei dessen Gestaltung unsicher fühlten.

Die Besprechung mit den Studierenden zeigt, dass diese nicht aus Faulheit oder mangelnder Motivation Poster gestaltet haben, die Thea Schleidens Erwartungen nicht entsprachen. Auch hier bieten sich verschiedene Erklärungsansätze an:

- Im Denken und Handeln der Professorin deutet sich ein Zielkonflikt an zwischen dem ökonomischen Einsatz ihrer Zeitressourcen und dem Gestalten eines Leistungsnachweis-Settings, das die Gütekriterien für Leistungsbeurteilungen in ansprechendem Ausmaß erfüllt (siehe Kapitel 6.3). Es liegt überdies nahe, dass Thea Schleiden die Posterkonferenz als neues Leistungsnachweisformat nicht sehr sorgfältig eingeführt hat, wodurch die Studierenden nicht genau wussten, was von ihnen erwartet wird (die Vielfalt sowie die Vor- und Nachteile von Leistungsnachweisformaten werden in Kap. 7 erläutert). Analog zur vorangehenden Fallvignette wäre es auch hier wichtig gewesen, mit den Studierenden Lernziele und Beurteilungskriterien zu besprechen (Kap. 5 und 9).
- Eine Einführung in das Gestalten und die Qualitätsmerkmale von Postern wäre notwendig gewesen, wenn viele Studierende dieses Format nicht kennen. Wie schon im vorangehenden Punkt zeigt sich hier die Wichtigkeit des Constructive Alignments: Lernziele, Leistungsnachweis und die auf den Nachweis vorbereitenden Lehr-Lern-Aktivitäten müssen aufeinander ausgerichtet sein und zueinander passen (siehe Kap. 5.1). Andernfalls ist es für Studierende schwierig, die an sie gerichteten Leistungserwartungen zu erfüllen.
- Da viele Studierende nicht mit dem wissenschaftlichen Poster als „Textsorte“ und somit auch der Posterkonferenz als Leistungsnachweisformat bekannt sind, scheint es besonders wichtig, vor einer summativen Bewertung (Note, Prädikat „bestanden/nicht bestanden“, siehe Kap. 3.1) Übungsgelegenheiten zu schaffen, bei denen die Studierenden formatives Feedback erhalten (siehe Kap. 3.2).

Die Analyse legt verschiedene Anpassungsideen nahe: Thea Schleiden könnte bei der nächsten Durchführung des Moduls zu Beginn des Semesters mehr Zeit aufwenden, um mit den Studierenden die Posterkonferenz als Format des Leistungsnachweises, die Lernziele (siehe Kap. 5.2 bis 5.4) und die Beurteilungskriterien (siehe Kap. 9.3) zu besprechen. Mit diesem bescheidenen Zusatzaufwand kann sie den Studierenden ermöglichen, wesentlich zielgerichte-

ter zu lernen – und gibt dem neuen Leistungsnachweisformat eine echte Chance, lernförderliche Wirkung zu entfalten.

Auch sollten die Studierenden genügend Möglichkeiten erhalten, das Erstellen von Postern zu üben bzw. formative Rückmeldungen zu Entwürfen des Endprodukts zu erhalten. Ähnliches gilt für das Präsentieren von Postern. Um den knappen Zeitressourcen der Professorin Rechnung zu tragen, empfiehlt sich zu diesem Zweck insbesondere der Einsatz von Peer-Feedback oder anderen Feedbackinstanzen (vgl. Kap. 8.4). Das Gespräch mit den Studierenden zeigte außerdem, wie wichtig es ist, dass Dozierende ihre Leistungsbeurteilungspraxis reflektieren und anpassen (Kap. 2 und 3).

Wessen Leistung beurteile ich?

Vor einiger Zeit begann ich, Bachelorarbeiten im Studiengang Physiotherapie zu betreuen. Viele Studierende möchten ihre Abschlussarbeit einfach möglichst effizient und mit einer passablen Note hinter sich bringen. Doch kürzlich traf ich auf eine außergewöhnlich motivierte Studentin. Nuriya wollte in ihrer Bachelorarbeit die Auswirkungen spezifischer motorischer Übungen auf die neuromuskuläre Koordination untersuchen. Ich warnte sie, dass ihr sehr forschungsorientiertes Vorhaben für eine Bachelorarbeit aufwändig sei. Aber sie war begeistert vom Thema und der Aufwand schreckte sie nicht ab.

Nuryias Engagement war bemerkenswert. Sie recherchierte unermüdlich, besuchte Schulungen in Forschungsmethodik und führte Tests mit selbst rekrutierten freiwilligen Probanden durch. Regelmäßig kam sie in meine Sprechstunde. Ich beantwortete ihre offenen Fragen, wies sie auf neue Publikationen hin und unterstützte sie bei der methodischen Planung der Tests.

Als sie ihre Arbeit einreichte, war ich zuerst begeistert von deren Qualität. Doch als ich das Erstgutachten zur Arbeit verfassen sollte, wurde ich unsicher. Aufgrund unseres intensiven Austausches fiel es mir schwer, die Grenze zwischen ihrem eigenen Beitrag und meinen Ratschlägen zu erkennen. Hatte ich als Dozent zu stark in den Arbeits- und Lernprozess der Studentin eingegriffen? Inwiefern konnte ich ihre Arbeit noch unabhängig beurteilen und bewerten? Und sollte ich mich künftig beim Begleiten von Bachelorarbeiten stärker abgrenzen?

Mateo Precioso, Dozent für Neuromotorik

Die Zweifel von Herrn Precioso verweisen auf grundlegende Herausforderungen im Zusammenhang mit dem Beurteilen von Studierendenleistungen:

- Zwischen dem Beraten und Begleiten von Studierenden und dem abschließenden Bewerten ergibt sich ein Rollenkonflikt: Als Berater (wie auch im Lehralltag) soll ein Dozent die Studierenden unterstützen, als Bewerter muss er Stärken und Schwächen möglichst unabhängig feststellen und mitunter harte „Urteile“ fällen (siehe Kap. 3.3 zu Rollenkonflikten).
- Die Studentin war offenbar sehr motiviert und engagiert, hat aber gleichzeitig extensiv auf die Unterstützung des Dozenten zurückgegriffen. Dies wirft Fragen danach auf, wie Studierende dazu ermutigt und befähigt werden können, ihre eigenen Lernprozesse zu reflektieren und dadurch ihre Metakognition und Selbststeuerung weiterzuentwickeln (siehe Kap. 8).

Viele Lehrende machen sich gelegentlich ähnliche Gedanken wie Mateo Precioso. Darin äußern sich Widersprüche, die Lehrsituationen strukturell inne- wohnen und sich nicht auflösen lassen, etwa der Widerspruch, dass Studie- rende zum selbständigen fachlichen Denken *geführt* werden sollen. So fragte der Philosoph Immanuel Kant (vgl. 1803, 32) schon vor mehr als 200 Jahren, ob Menschen durch Zwang zur Freiheit erzogen werden könnten. Das vorlie- gende Buch möchte Ihnen in diesem Zusammenhang helfen, Ihren eigenen, bewussten Umgang mit diesen strukturell bedingten Herausforderungen zu finden (siehe Kap. 3).

Die Erfahrung von Mateo Precioso zeigt des Weiteren die Komplexität von Betreuungsverhältnissen auf, etwa bei der Begleitung wissenschaftlicher Arbeiten oder Projekte. Es lohnt sich in solchen Fällen, die gegenseitigen Erwartungen und Verantwortlichkeiten durch eine schriftliche Vereinbarung vorab zu klären (siehe Kap. 12).

1.2 Die Beurteilung und ihre Reflexion

Welche Erlebnisse kommen Ihnen zuerst in den Sinn, wenn Sie das Wort „Prüfungen“ hören? Schulische Situationen? Lebensprüfungen, die Ihnen vom Schicksal zugemutet wurden? Die Prüfung von Bilanzen und Geschäfts- unterlagen? Was verstehen Sie unter einer Leistung? Wer erbringt Ihres Er- achtens eine Leistung, und wie kann man sie erfassen, bewerten, vergleichen, beurteilen?

Begriffsklärung

Im Zusammenhang mit Leistungsbewertung an Hochschulen ist oft von „Prüfungen“ die Rede. Dieser Begriff beinhaltet aber diverse Einschränkungen – bereits eine wissenschaftliche Arbeit dürfte kaum als Prüfung bezeichnet werden. Deshalb wird in diesem Buch der Begriff *Leistungsnachweis* als Oberbegriff für alle Formen von Leistungsüberprüfung verwendet.

Die Kognitionen und Emotionen, die Sie mit diesen ganz persönlichen Schlüsselerlebnissen verbinden, sowie Ihre bewussten und unbewussten Schlussfolgerungen daraus dürften Ihren Umgang mit Beurteilungssituationen wesentlich prägen. Betrachten Sie Leistungsnachweise (siehe obenstehende Begriffsklärung) als lästiges Übel, das sowohl Sie als auch Ihre Studierenden irgendwie hinter sich bringen müssen? Oder glauben Sie, dass Leistungsnachweise eine wichtige Rolle spielen, indem sie die schlechten Studierenden aussortieren helfen, so dass nur die geeigneten übrigbleiben? Oder sehen Sie sich eher in der Rolle der Beraterin, welche die Studierenden beim Lernen und beim Erfüllen der durch die Institution Hochschule gestellten Anforderungen unterstützt? Gibt es bestimmte Formen von Leistungsnachweisen oder Aufgabentypen, die Sie in Ihrem Fach für besonders (un)geeignet halten, oder empfehlen Sie Ihren Studierenden bestimmte Vorgehensweisen zur Prüfungsvorbereitung?

Auch wenn Sie diese Haltungen und Überzeugungen auf Ihre Erfahrungen als lehrende und/oder lernende Person sowie auf Wissen über Lehren, Lernen und Prüfen stützen können: Subjektiv geprägte, teilweise unbewusste Einschätzungen und die mit ihnen verbundenen Emotionen prägen unser Denken und Handeln stark – oft mehr als kognitive Wissensbestände. Das gilt selbst dann, wenn dieses Wissen über lange Zeit in einem fachlichen und/oder professionellen Kontext aufgebaut wurde und relativ weit reicht (vgl. Wahl 2001). Dies erklärt, weshalb selbst eine Professorin für Erziehungswissenschaft eine monologische Vorlesung über den Wert dialogischer Lernformen halten kann – die langjährige akademische Prägung führt zu einem unhinterfragten Bild, wie eine Vorlesung abzulaufen hat, und erzeugt eine kognitive Dissonanz zwischen vermitteltem Inhalt und Vermittlungsform.¹

- 1 Das in Kap. 5.1.2 vorgestellte Modell des Constructive Alignments ermöglicht, das geschilderte Problem didaktisch genauer zu fassen: Sollen Studierende den Wert dialogischer Lernformen erkennen, korrespondiert dieses Lernziel nicht mit der Unterrichtsmethode des frontalen Vortrags – es liegt ein Bruch im Alignment vor.

Das vorliegende Buch möchte deshalb mehr sein als nur eine „Best-Practice“-Anleitung zum effizienten und lernförderlichen Durchführen von Leistungsbeurteilungen. Es möchte Sie darüber hinaus dazu auffordern, Ihre Einstellungen und Praktiken im Umgang mit Leistungsbeurteilungen sowie beim Erteilen von Leistungsrückmeldungen an Studierende kritisch zu hinterfragen. Das beginnt schon bei der Frage danach, was im Zusammenhang mit Lernen überhaupt unter einer „Leistung“ zu verstehen ist (siehe Kap. 2). Hier ist auch die Beziehung zu den Menschen im Auge zu behalten, um deren Lernen es geht: „Lernen ergibt sich aus dem, was die Studierenden tun und denken, und nur aus dem, was die Studierenden tun und denken. Lehrende können das Lernen nur fördern, indem sie Einfluss darauf nehmen, was die Studierenden tun, um zu lernen“ (Herbert A. Simon, zitiert nach Ambrose et al. 2010, 1, Übers. tz).

Deshalb ist zu beachten, dass Leistungsnachweise nicht nur zur Bilanzierung von Lernleistungen dienen, sondern ganz wesentlichen Einfluss darauf ausüben, wie Studierende lernen (siehe Kap. 5). Damit sind wiederum Fragen nach der gesellschaftlichen Aufgabe von Hochschulen als Institutionen verbunden. Denn das Lernen an Hochschulen endet mit der Verleihung von Studienabschlüssen, welche zur Übernahme als besonders relevant eingeschätzter gesellschaftlicher Rollen berechtigen, wie Ärztinnen, Juristen, Ingenieurinnen oder Lehrer. Unsere Beurteilungspraxis bewegt sich deshalb zwischen dem Fördern von Lernen, der Sicherstellung des Erwerbs von Wissen und Können sowie dem Erteilen von gesellschaftlichen Berechtigungen (siehe Kap. 3).

Das kritische Hinterfragen Ihrer Beurteilungspraxis macht Sie im Erfolgsfall zum „Reflective Practitioner“ im Sinne von Schön (1991). Indem Sie Ihren Umgang mit Leistungsnachweisen und Leistungsrückmeldungen auf verschiedenen Ebenen reflektieren können, erweitern Sie Ihre Handlungsalternativen: Sie können beim Entwickeln von Leistungsnachweisen situationsadäquat Formate und Aufgaben auswählen und gestalten, auf der institutionellen Ebene an der Erarbeitung von lernförderlichen Regelungen in Bezug auf die Bewertung und Beurteilung von Leistungen mitwirken und Studierende wirksam bei der Vorbereitung auf Leistungskontrollen begleiten.

Bei aller kritischer Distanz zum Leistungsbegriff sowie zur Bewertung und Beurteilung von Leistungen sei somit gesagt: Wenn in diesem Buch der Leistungsbegriff hinterfragt und dysfunktionale Prüfungs- und Bewertungspraktiken (z. B. das Erteilen von Noten) moniert werden, dann immer mit dem Ziel, durch eine ganzheitlichere Praxis der Leistungsbeurteilung und -rückmeldung *an Hochschulen mehr und besseres Lernen zu ermöglichen*. Ich bin überzeugt, dass die Hochschulen dadurch auch ihre gesellschaftlichen Funktionen als Bildungs- und Forschungsstätten sowie als Qualifikationsinstanzen noch wirkungsvoller wahrnehmen.

So ist mein unmittelbares Ziel, Ihnen mit diesem Buch hilfreiche praktische Ideen für die Gestaltung von Beurteilungssituationen und den Umgang mit ihnen an die Hand zu geben. Darüber hinaus hege ich die vielleicht verwegene Hoffnung, dass es ein wenig dazu beiträgt, Ihre persönliche Beurteilungspraxis – und mittelbar vielleicht diejenige ganzer Bildungsinstitutionen – lernförderlicher, aussagekräftiger und gerechter zu machen.

Teil I: Grundlagen

2 Leistung in Gesellschaft und Bildungswesen

Sie möchten wissen, wie Sie die Leistungen Ihrer Studierenden effizienter oder zuverlässiger erfassen und beurteilen können und/oder wie Sie den Studierenden möglichst lernwirksames Feedback zu den gezeigten Lernleistungen geben können? Das Wissen, das Sie sich in diesem Kapitel aneignen können, wird einerseits vermutlich Ihre Vorstellungen von „Zuverlässigkeit“ im Bereich der Leistungsmessung und -beurteilung verändern. Andererseits ermöglicht es Ihnen ein neues Bild davon, wie studentische Leistungen zustande kommen und interpretiert werden. Dies ist eine notwendige Vorbedingung dafür, diese Leistungen angemessen erfassen, beurteilen, bewerten und Feedback dazu geben zu können.

Lassen Sie sich deshalb auf eine Reise durch die Begriffsgeschichte der Leistung sowie gesellschaftlicher Leitkonzepte im Umgang mit Leistung mitnehmen. Wir gehen zuerst der Rolle von Leistung als Wahrnehmungskonzept in unserer Gesellschaft und speziell im Bildungswesen auf den Grund (Kap. 2.1). Auf dieser Basis können wir das Leitkonzept der Meritokratie kritisch analysieren (Kap. 2.2) und die Zuverlässigkeit und Auswirkung von Noten hinterfragen (Kap. 2.3). Dadurch stoßen wir auf Phänomene, die sich durch eine ausgeprägte gesellschaftliche Orientierung an Messbarkeit und Vergleichen sowie dem Erstellen von Rangordnungen erklären lassen (Kap. 2.4). Dabei werden auch Zusammenhänge zwischen der Beurteilung von studentischen Leistungen und der Beurteilung von Hochschulen als Institutionen oder von Forschungsleistungen sichtbar. Der kritische Blick auf Leitkonzepte wie Meritokratie, Metrisierung und Kredentialismus ermöglicht es schließlich, den Leistungsbegriff einer für Lehr- und Lernzwecke konstruktiven Sichtweise zuzuführen (Kap. 2.5).

2.1 Leistung als Begriff und Wahrnehmungskonzept

Überraschenderweise zeigt die Begriffsgeschichte, dass der Leistungsbegriff im Bildungskontext noch nicht so lange verwendet wird, wie man aufgrund seiner heutigen Allgegenwärtigkeit vermuten könnte. Er erweist sich nicht nur als eng verbunden mit technologischem Fortschritt, sondern auch mit Vorstellungen von Mess- und Vergleichbarkeit. „Leistung“ ist heute zwar ein vieldeutiger Begriff. Diese Vieldeutigkeit ist aber keineswegs beliebig, sondern führt ins Zentrum unseres gesellschaftlichen Selbstverständnisses. „Leistung“ kann als ein Wahrnehmungskonzept beschrieben werden, das

wesentliche Vorgänge in Gesellschaft und Bildungswesen steuert (Reh, Berdelmann u. Scholz 2015, 37).

Heute umfasst der Begriff „Leistung“ im allgemeinen Sprachgebrauch gewöhnlich vier Bedeutungselemente, die je nach Verwendungszusammenhang mehr oder weniger im Vordergrund stehen (vgl. Schlie 1988, 63f.):

- Leisten ist zielgerichtetes Handeln, wobei das Ziel durch die Handelnden oder von außen gesetzt sein kann.
- Damit verbunden ist eine gewisse Anstrengung, eine Leistung stellt sich nicht einfach von selbst ein.
- Als Ergebnis des Leistens wird ein bestimmtes Resultat erwartet, zu dem die zielgerichtete Anstrengung führen soll.
- Das Ergebnis wird in der Regel beurteilt, meistens anhand der gesetzten Ziele, auf die sich das Leisten bezieht, und manchmal auch aufgrund des Grades der investierten Anstrengung. Leistung kann folglich anhand von Gütekriterien beurteilt werden.

Diese Elemente treten auch – bis in die begriffliche Ebene – in der heutigen Konzeption von Schulleistung hervor, in der von Lernzielen, learning outcomes und Gütekriterien der Beurteilung die Rede ist (siehe Kap. 5 und 6). Das ist kein Zufall und zeigt: Der heutige Leistungsbegriff ist auch für den Bildungsbereich prägend. Das gesellschaftliche und das bildungsbezogene Leistungsverständnis sind sogar untrennbar miteinander verwoben, wie die folgenden Seiten deutlich machen.

Dass Leistung nicht immer so stark mit Fragen der Bildung gekoppelt war, zeigt ein Blick in die Geschichte des Schulwesens, den Kapitel 14 im Rahmen eines bildungshistorischen Exkurses vertieft. Erst im Laufe des 19. Jahrhunderts rückte die auf das fachliche Lernen bezogene „Leistung“ in den Fokus. So heißt es in der „Enzyklopädie des gesamten Erziehungs- und Unterrichtswesens“ 1887, die Zensur (=Note) diene dazu, die Schüler „zu einem richtigen Urteil über sich selbst, [ihre] Fähigkeiten und *Leistungen* anzuleiten“ (zit. nach Ziegenspeck 1999, 75).² Im Rahmen dieser Entwicklung wurden schulische Leistungen im Deutschland des 19. Jahrhunderts zunehmend in Zeugnissen abgebildet, wobei die Zahl der Abstufungen noch variierte, um sich dann im 20. Jahrhundert zu stabilisieren (vergleichbar mit den Entwicklungen in den USA, vgl. Brookhart et al. 2016, 805).³

- 2 Als eigenständiges Stichwort tritt „Leistung“ allerdings erst seit 1930 in pädagogischen Handbüchern und Enzyklopädien auf (vgl. Furck 1972, 140, FN 34).
- 3 Die heute in Deutschland noch gültige sechsstufige Skala von 1 (=sehr gut) über 4 (=ausreichend) bis 6 (=ungenügend) wurde schließlich 1938 in der Zeit des Nationalsozialismus eingeführt und erstaunlicherweise seither in der Bundesrepublik Deutschland beibehalten. In Österreich wird dagegen heute eine Skala

Die Entstehung von Schulnoten im Zeitraum von 1800 bis ca. 1950 ist Symptom eines grundsätzlichen Wandels im Bildungswesen. Erst durch die Einführung und nationale Standardisierung von Schulnoten wurde eine Art Währung eingeführt, mit der Schulleistungen scheinbar unabhängig vom Leistungsniveau einer Klasse oder Schule vergleichbar wurden (vgl. Verheyen 2018, 63).

Interessanterweise verlief die Entwicklung des modernen Leistungssports parallel zur Standardisierung der Schulnoten. Auch im Sport wurden Leistungen zwar schon früher verglichen (vgl. Mallinckrodt u. Schattner 2016). Aber erst ab der zweiten Hälfte des 19. Jahrhunderts wurde zunehmend situationsübergreifend gemessen und verglichen. „Damit schien es möglich, einen Wettbewerb überall auf der Welt und beliebig häufig zu wiederholen und *alle* erreichten Ergebnisse von *allen* Sportlern – Frauen waren mal wieder größtenteils ausgeschlossen – miteinander zu vergleichen, ganz egal, wo und wann sie erbracht worden waren, ob die Sonne schien oder den Athleten Regen ins Gesicht peitschte. (...) Das ging bis hinauf zum ‚Weltrekord‘ (...)“ (Verheyen 2018, 87, Hervorhebungen von Verheyen).

Hier zeigt sich eine klare Gemeinsamkeit von sportlicher und schulischer Leistungsmessung: Beide zielen auf eine globale Vergleichbarkeit der gemessenen Leistungen (vgl. Heintz 2016). Diesen Megatrend erkannte schon der Philosoph Friedrich Nietzsche, als er 1878 die Moderne zum „Zeitalter der Vergleichung“ ausrief (Nietzsche 1999 [1878], 45). Und der Trend beschränkte sich mitnichten auf Schule und Sport, sondern war eng verbunden mit technischen, wirtschaftlichen und politischen Entwicklungen. Dank gesellschaftlicher und technologischer Entwicklungen waren nicht nur immer genauere Messungen von physikalischen und anderen Werten möglich, sondern diese konnten durch zunehmend leistungsfähigere Kommunikationstechnologien immer schneller und globaler übermittelt und verglichen werden.⁴ Solche Prozesse ermöglichten auch die Entstehung von Börsenindizes

von 1 (=sehr gut) über 4 (=genügend) bis 5 (=nicht genügend) verwendet (vgl. Ziegenspeck 1999, 76–79 u. 95). Die Deutschschweiz benotet hingegen in umgekehrter Richtung und mit anderer Gewichtung von 6 (=sehr gut) über 4 (=genügend) bis zu 1 (=schlecht). Man beachte, dass der Anteil der ungenügenden Noten am gesamten Notenband von 25 % im österreichischen bis zu 60 % im Deutschschweizer System variiert. Dies führt z. B. in der Schweiz dazu, dass der ungenügende Teil der Bewertungsskala kaum genutzt wird – Noten unter 3 sind sehr selten und zeigen außerordentlich schwache Leistungen an.

- 4 Die Zusammenhänge von Ursache und Wirkung dieser bis heute anhaltenden Dynamik werden schnell komplex und unübersichtlich: Haben gesellschaftliche Entwicklungen erst den rasanten technologischen Fortschritt ab dem 19. Jh. ermöglicht, etwa die Einführung des Meters als Maßeinheit in der Folge der Französischen Revolution (vgl. Konersmann 2021, 205 und 298–300 (FN 117))? Oder waren die technologischen Fortschritte, im 19. Jh. z. B. industriell produ-

wie dem Dow Jones oder die Abwicklung von Außenhandelsgeschäften ohne persönliche Vertrauensbeziehungen (vgl. Gill 2017, 393; Ambrosius 2018, 285).

Gerade auch aufgrund der Wirtschaftsnähe des Leistungsdenkens wird oft postuliert, dass es aus der Verbürgerlichung der Gesellschaft hervorging, als „bürgerliches Leistungsdenken“. Die Zusammenhänge sind freilich wesentlich komplexer, zumal die Entwicklungen in verschiedenen Gesellschaften auf unterschiedlichen Kontinenten fast zeitgleich stattfanden (vgl. für nähere Erläuterungen Verheyen 2018, 2014). Sapolsky (2018, 296f.) weist darauf hin, dass die im 19. Jahrhundert einsetzende massive Urbanisierung, die von 1870–1914 einen ersten Höhepunkt erreichte, anthropologisch weitreichende Folgen hatte.

Laut dem Anthropologen Sapolsky ist es unter Primaten erstmalig, dass sich regelmäßig Fremde begegnen, die sich nie wieder sehen – eine Großstadtbewohnerin begegnet täglich wesentlich mehr Fremden als Bekannten. Dies habe zur Erfindung der anonymen Handlung geführt. Henrich et al. (2010) weisen in diesem Zusammenhang darauf hin, wie erstaunlich es ist, dass in anonymen Massengesellschaften Fremde mit hoher Zuverlässigkeit miteinander Transaktionen zum gegenseitigen Vorteil durchführen.⁵ Die Entwicklung von übergreifenden Normen und Institutionen dürfte nötig geworden sein, um Fairness in solchen flüchtigen Austauschbeziehungen zu gewährleisten. Ein Aspekt davon betrifft die zunehmende Organisation gesellschaftlichen Austausches als Markt: In der Tat geht laut Henrich et al. (2010) eine größere Vermarktlichung von Gesellschaften mit mehr Fairness im Verhalten bei an Marktmechanismen orientierten Spielen einher. Märkte wiederum leben vom Vergleich und legen damit Quantifizierungen nahe: Die zum Vergleich „synonyme Formulierung ‚aneinander messen‘ stellt klar, worum es geht. Dinge vergleichen heißt, sie unter dem Aspekt ihrer Gleichartigkeit zu betrachten, und das bedeutet: sie schon mit der Art des Zugriffs vergleichbar zu machen“ (Konersmann 2021, 52). Die Leistung als quantifiziertes Maß menschlicher Tätigkeit wurde somit zur Basis vieler gesellschaftlicher Bezie-

zierte Uhren und die Telegraphie, verantwortlich für gesellschaftliche Entwicklungen wie die Urbanisierung? Diese beiden Aspekte scheinen untrennbar miteinander verwoben.

- 5 Auch die Erfindung der Kriminalliteratur, die in die gleiche Zeit fällt, sieht Sapolsky (2018, 296) als Auswirkung der neuartigen Urbanisierung: Erst das Setting von anonymen, städtisch geprägten Gesellschaften macht die Frage spannend, wer ein bestimmtes Verbrechen begangen hat – in auf persönlichen Kontakten basierenden Gesellschaftsordnungen in weniger dicht besiedelten Gegenden wissen in der Regel alle gegenseitig über ihre Handlungen Bescheid.

hungen – mit weitreichenden Folgen, von denen wir in diesem Kapitel nur die Spitze des Eisbergs erkunden können.⁶

Zugleich ist es wichtig zu beachten, dass viele der entstandenen Maße für die Messung und den Vergleich von Leistungen keine wirklich objektiven und absoluten Vergleiche ermöglichen – die Genauigkeit der Maße täuscht über die Unterschiedlichkeit der Bedingungen hinweg, unter denen die Leistungen entstanden sind. Um nochmals zum Sport zurückzukehren: Es spielt für die Gültigkeit eines Marathonweltrekords keine Rolle, welche Temperatur- oder Windverhältnisse bei seiner Erbringung geherrscht haben. Und innerhalb eines Radzeitfahrens oder Skirennens können ganz unterschiedliche Wind-, Niederschlags- und Lichtverhältnisse herrschen – die Zeiten werden dennoch auf Hundertstelsekunden genau gemessen und verglichen. Falls Sie nun einwenden möchten, dass im Sport auch die Show im Vordergrund steht und die Leistungsmessung und -beurteilung im Bildungsbereich gewiss seriöser erfolgt: Leider muss ich Sie enttäuschen. Die aktuelle Praxis der Beurteilung und des Vergleichs von Lernleistungen – nicht nur, aber ganz besonders durch Noten – ist noch viel unzuverlässiger. Die widersprüchlichen Zusammenhänge von Messung, Beleg und Vergleich von Leistungen, die sich in Gesellschaft und Bildungswesen etabliert haben, zeige ich nachfolgend auf. Am Ende des Kapitels wird auf dieser Basis dargelegt, inwiefern Lernleistungen durch Bildungseinrichtungen wie Hochschulen (auch) sozial konstruiert werden. Dadurch lässt sich eine differenzierte Definition des Begriffs „Lernleistung“ vornehmen (siehe Kap. 2.5.2).

- 6 Sapolsky weist angesichts der Studie von Henrich, Joseph et al. (2010) darauf hin, dass wir dazu neigen, Marktmechanismen als den Gipfel der Regulation von komplexen menschlichen Bedürfnissen zu betrachten – alles kann zu einem Geldwert abstrahiert werden. Doch anthropologisch betrachtet sei menschliche Gegenseitigkeit ursprünglich wesentlich intuitiver und ganzheitlicher: „Der Typ hier ist ein großartiger Jäger; der andere dort kann das weniger gut, aber hält dir den Rücken frei, wenn ein Löwe in der Nähe ist; jene Frau dort ist enorm gut darin, die besten Mongongo-Nüsse zu finden, während diese ältere Frau alles über Heilkräuter weiß, und die seltsame Type dort drüben kann die besten Geschichten erzählen. Wir wissen, wo die anderen leben, mit der Zeit gleichen sich die Schuldigkeiten aus, und wenn jemand wirklich das System missbraucht, werden wir uns gemeinsam darum kümmern“ (Sapolsky 2018, 500, Übers. tz). Deshalb sieht er Marktinteraktionen als Verarmung der Gegenseitigkeit menschlichen Handelns.

2.2 Leistung und Meritokratie

Bevor wir uns dem Thema Meritokratie widmen, versuchen Sie bitte, die folgenden vier Fragen spontan und möglichst schnell zu beantworten. Selbstverständlich brauchen Sie Ihre Antworten niemandem mitzuteilen, aber merken Sie sich diese bitte.

In welchem Ausmaß sind arme Menschen dafür verantwortlich, dass sie arm sind?

(1 = überhaupt nicht verantwortlich, 7 = gänzlich verantwortlich)

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐

In welchem Ausmaß sind blinde Menschen dafür verantwortlich, dass sie blind sind?

(1 = überhaupt nicht verantwortlich, 7 = gänzlich verantwortlich)

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐

In welchem Ausmaß sind ungebildete Menschen dafür verantwortlich, dass sie ungebildet sind?

(1 = überhaupt nicht verantwortlich, 7 = gänzlich verantwortlich)

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐

In welchem Ausmaß sind dicke Menschen dafür verantwortlich, dass sie dick sind?

(1 = überhaupt nicht verantwortlich, 7 = gänzlich verantwortlich)

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐

Kuppens et al. (2018) haben in einer Reihe von Studien Menschen in Europa und den USA zu ihren Einstellungen gegenüber verschiedenen bekanntermaßen sozial benachteiligten Gruppen wie körperlich Behinderten oder ethnischen Minderheiten befragt. Im gleichen Zug erfragten sie auch die Einstellungen gegenüber bildungsnahen und bildungsfernen Personen. Die Resultate waren frappant: Bildungsnahe („gut gebildete“) Befragte aus Europa und den USA zeigten gegenüber keiner Gruppe weniger Sympathien als gegenüber den bildungsfernen („wenig gebildeten“) Personen. Die Studienautoren kommen zum wenig schmeichelhaften Schluss, dass das Betonen des eigenen Bildungserfolgs die letzte Bastion sei, wo bildungsnahe Kreise sich das Hegen von Vorurteilen zugestehen (Kuppens et al. 2018, 444).

Die Studienautoren wollten zudem wissen, wie die Verantwortung für den Bildungserfolg eingeschätzt wird, wozu sie unter anderem Fragen wie die

vorhin gestellten verwendeten.⁷ Das Resultat ist eindeutig: Im Vergleich etwa zur Arbeiterklasse wurden Bildungsferne als stärker selbst schuldig an ihrer nachteiligen Situation wahrgenommen (wie lautete Ihre Einschätzung?). Diese negative Einschätzung bildungsferner Menschen ist nicht nur bei den bildungsnahen Studienteilnehmenden zu finden, sondern wird auch von den weniger gebildeten Befragten selbst geteilt. Sie scheinen also die negativen Zuschreibungen ihrer Gruppe gegenüber internalisiert zu haben (Kuppens et al. 2018, 443f.).

Dass gebildete Menschen gegenüber Ungebildeten gewisse Dünkel empfinden, mag naheliegend sein, auch wenn das Ausmaß erstaunt. Aber wie ist es möglich, dass sogar bildungsferne Personen überwiegend der Meinung sind, sie seien für ihren niedrigen gesellschaftlichen Status selbst verantwortlich? Die Antwort liegt in einer gesellschaftlich tief verankerten Überzeugung: „Nicht soziale Herkunft, die Hautfarbe oder das Geschlecht, sondern allein individuelle Leistungen sollen darüber entscheiden, welches Ansehen, welchen Einfluss und welchen materiellen Wohlstand eine Person erringen kann“ (Dröge, Marrs u. Menz 2008, 7). Oder in der amerikanischen, vom Bild der Tellerwäscherkarriere geprägten Vorstellung: „Those who work hard and play by the rules should be able to rise, as far as their talents will take them“ (Sandel 2021, 23).

Das Leistungsprinzip, das sich in den zitierten Maximen äußert, hat sich parallel zu der in Kapitel 2.1 skizzierten Fokussierung des Messens und Vergleichens von Leistungen etabliert. Dass Menschen ihre gesellschaftliche Stellung unabhängig von ihrer sozialen Schicht, Ethnie oder ihrem Geschlecht einnehmen sollen, gilt als ein Grundprinzip liberaler Demokratien. Wenn Anstellungen, Lohn und Ansehen ausschließlich aufgrund von Leistung vergeben werden, so scheint dies gerechter als eine auf vererbten Privilegien basierte Feudalgesellschaft. Ungleiche Leistungen legitimieren damit soziale Ungleichheiten etwa in der Entlohnung oder im Prestige einer Ausbildung oder eines Berufes. Man bezeichnet eine solche Vorstellung von der gesellschaftlichen Ordnung auch als Meritokratie (vgl. Schäfer 2015, 17; Becker u. Hadjar 2009, 36).

Definition

Der *sozioökonomische Status (SES)* ist ein zusammenfassendes Maß für den wirtschaftlichen Zugang einer Person oder Familie zu Ressourcen und für ihre soziale Stellung im Verhältnis zu anderen. Bildungsabschlüsse spielen eine große Rolle für den SES: Je höher der höchste Bildungsabschluss einer Person, desto höher ihr SES.

7 Bei den oben formulierten Fragen handelt es sich um ungefähre Übersetzungen von Fragen aus Kuppens et al. (2018, 439).

„Das Bildungssystem ist in diesem Sinne die zentrale Rechtfertigungsfabrik sozialer Ungleichheit in der modernen Gesellschaft. Es verwandelt nach den Maßstäben ‚individueller Leistung‘, individuell zurechenbar und nachvollziehbar Gleiche in Ungleiche (in Rang, Bezahlung usw.), und zwar so, dass die Benachteiligten in der Geltung des Gleichheitsprinzips ihre Benachteiligung akzeptieren“ (Beck 1988, 265). Damit lassen sich auch die oben berichteten negativen Einschätzungen von bildungsfernen Menschen erklären: Körperliche Behinderungen oder Zugehörigkeit zu einer benachteiligten Klasse werden mindestens teilweise als unverschuldet betrachtet, während niedrige Bildungsabschlüsse für eine Folge mangelnder Anstrengung und damit für ein individuelles Versagen gehalten werden. Diese Ansicht ist gesellschaftlich so stark etabliert, dass sie sogar bildungsferne Menschen verinnerlicht haben (vgl. Sandel 2021, 96).

Halten diese Überzeugungen einer näheren Überprüfung stand? Sind in unserer Gesellschaft – mit Ausnahme von Menschen mit offensichtlichen körperlichen oder kognitiven Benachteiligungen – alle ihres eigenen Glückes Schmied? Wir können diese weit gefasste Frage für den Bildungsbereich eingrenzen und in die wissenschaftliche Form einer Hypothese überführen. Sie lautet: Der sozioökonomische Status der Eltern hat keinen Einfluss auf den Bildungserfolg ihrer Kinder (siehe Definition). Diese Hypothese wurde in der Bildungsforschung vielfach überprüft. Wir werfen nachfolgend einen Blick auf drei aufschlussreiche Studien.

Studie 1 (Schütz, Ursprung u. Wössmann 2008)

Ausgewertet wurden Daten der internationalen Vergleichsstudie TIMSS (Trends in International Mathematics and Science Study) aus den Jahren 1995 und 1999. Für rund 300.000 Schülerinnen und Schüler des 8. Schuljahres (Alter ca. 13 Jahre) aus 54 Ländern wurde ein Gesamtwert ihrer Kenntnisse in Mathematik und Naturwissenschaften gebildet. Dieser Wert kann als Bildungserfolg im mathematisch-naturwissenschaftlichen Bereich betrachtet werden. In Bezug gesetzt wurde er mit dem familiären Bildungshintergrund der untersuchten Kinder. Die Auswertung weist für alle Länder der OECD einen deutlichen Zusammenhang zwischen der familiären Herkunft und der Leistung aus (vgl. Abb. 1). In Deutschland ist dieser Zusammenhang besonders ausgeprägt, in Österreich und der Schweiz etwas weniger, aber immer noch stark. Sogar in der Schweiz als dem egalitärsten dieser drei Länder erreicht ein Schüler aus der bildungsfernensten Stufe (FB 1) nur knapp mehr als die Hälfte der Testleistung einer Schülerin aus der bildungsnächsten Stufe (FB 5); in Deutschland sind es 40 %.

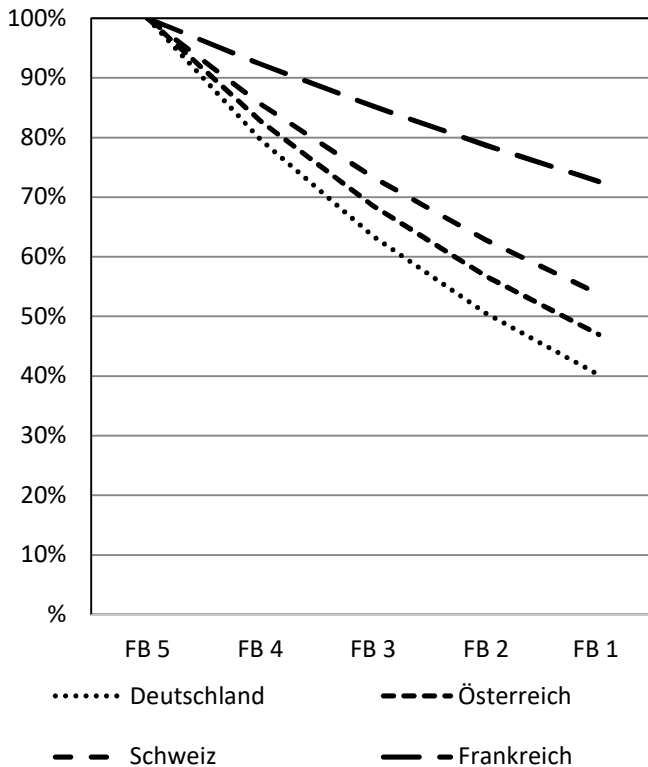


Abb. 1: Mathematisch-naturwissenschaftlicher Bildungserfolg im 8. Schuljahr, in Abhängigkeit vom familiären Bildungshintergrund; FB 5 = bildungsnah, FB 1 =bildungsfern. Quelle: Eigene Abbildung, basierend auf eigenen Berechnungen anhand von Daten aus Schütz et al. 2008, 291f.

Dass die Unterschiede auch geringer ausfallen können, zeigen Länder wie Frankreich und Kanada. Aber auch in Frankreich, das den geringsten Zusammenhang zwischen familiärem Bildungshintergrund und Bildungserfolg aufwies, erreichen Kinder aus bildungsfernen Familien lediglich 72 % des mathematisch-naturwissenschaftlichen Bildungserfolgs von Kindern aus besonders bildungsnahen Familien.

Nun könnte man einwenden, dass die TIMSS-Daten nur eine Momentaufnahme darstellen und noch nicht die definitiven Bildungsabschlüsse auf der Sekundarstufe 2 oder der für dieses Buch besonders relevanten Tertiärstufe erfassen. Dies leistet die folgende Studie.

Studie 2 (Bukodi, Erikson u. Goldthorpe 2014)

Ausgewertet wurden Bildungsverläufe von sieben Jahrgangskohorten aus Großbritannien und Schweden (Geburtsjahre 1946, 1958 und 1970 in Großbritannien, 1948, 1953, 1967 und 1972 in Schweden, $n=48.889$). Untersucht wurde, wie die familiäre Herkunft und allgemeine kognitive Fähigkeiten⁸ die Bildungsabschlüsse der in diesen Jahren geborenen Menschen beeinflusst haben. Die Zahlen der verschiedenen Jahrgänge unterscheiden sich da und dort, aber das grundsätzliche Bild ist in beiden Ländern und allen Jahrgängen das Gleiche. Der Einfachheit halber greifen wir an dieser Stelle die beiden jüngsten Jahrgänge heraus: Personen des Jahrgangs 1970 aus Großbritannien und solche des Jahrgang 1972 aus Schweden, die aus Familien mit höherem sozioökonomischem Status stammen, hatten eine fast doppelt so hohe Chance, einen höheren Abschluss auf Sekundarstufe II zu erreichen als Gleichaltrige aus Familien mit tieferem SES. Dies gilt unabhängig von ihren kognitiven Fähigkeiten.⁹ Abbildung 2 vereinfacht die tatsächlichen Zahlen in einer schematisierten Darstellung.

Diese Befunde sind umso bemerkenswerter, als bereits die kognitiven Fähigkeiten im Alter von 10–13 Jahren deutlich durch den sozioökonomischen Status geprägt sind: Kinder aus Familien mit hohem SES haben schon im Alter von 2 Jahren einen durchschnittlich um 6 Punkte höheren IQ-Wert, und bis zum Alter von 16 Jahren verdreifacht sich dieser Unterschied fast. Die scheinbar unabhängige Variable der kognitiven Fähigkeiten ist also de facto erheblich vom SES abhängig (vgl. Stumm u. Plomin 2015).

- 8 Die kognitiven Fähigkeiten wurden im Alter von 10–13 Jahren gemessen. Die verwendeten Verfahren erlauben eine gute Annäherung an IQ-Werte, die sich auf den allgemeinen Faktor Intelligenz beziehen lassen. Sie wurden aber für die Zwecke der berichteten Studie in eine relative Ordnung gebracht, indem sie für jeden Jahrgang in Quintile ordinalisiert wurden (vgl. Bukodi, Erikson u. Goldthorpe 2014, 296f.).
- 9 In den beiden Ländern verlaufen die Unterschiede leicht anders. Auf die gesamte Kohorte betrachtet liegen aber auch die schwedischen Personen bei einem Faktor von ungefähr 2. Schweden weist insgesamt einen etwas höheren Anteil an abiturähnlichen Abschlüssen auf als Großbritannien (vgl. Bukodi, Erikson u. Goldthorpe 2014, 302–5).

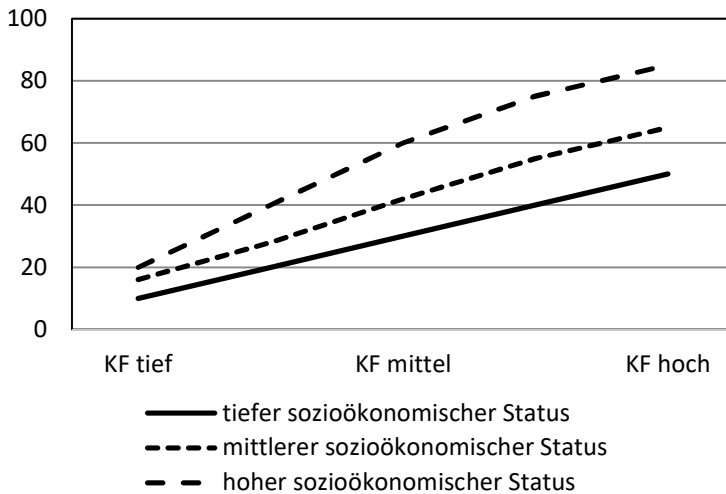


Abb. 2: Anteil höherer Abschlüsse auf Sekundarstufe II für 1970/1972 in Großbritannien/Schweden geborene Personen, unterschieden nach sozioökonomischem Status und kognitiven Fähigkeiten (KF). Quelle: Eigene Abbildung (schematisierte Darstellung von Angaben und Grafiken aus Bukodi et al. 2014, 302-305)

Studie 3 (Carnevale u. Rose 2004)

Die ersten beiden Studien zeigen, dass sowohl Schulleistungen als auch Bildungsabschlüsse zwischen Personen mit tiefem und solchen mit hohem sozioökonomischem Status ohne Weiteres um den Faktor 2 variieren können. Geradezu extrem fallen demgegenüber die Zahlen für die Hochschulabschlüsse in den USA aus. Dazu ist zu betonen, dass das US-amerikanische Hochschulsystem stark stratifiziert ist, also geschichtet in Hochschulen mit unterschiedlichem Prestige. So unterscheiden Carnevale u. Rose (2004) zwischen vier Stufen von Colleges, von der Stufe „wenig bis gar nicht wettbewerbsfähig“ bis zur Stufe „höchst wettbewerbsfähig“. Je höher die Prestige-Stufe, desto selektiver ist die Aufnahme. So nahmen 2019 verschiedene Top-Universitäten weniger als 10 % der Studierenden auf, die sich bei ihnen bewarben (vgl. Sandel 2021, 176).

Die Studierendenschaft dieser vier Stufen ist sozioökonomisch sehr stark segregiert. Im Jahr 1988 wurden zum Beispiel in die 146 Hochschulen der obersten Stufe 24-mal mehr Studierende aus dem obersten sozioökonomischen Quartil aufgenommen als aus dem untersten Quartil – die Studierendenschaft dieser Elite-Universitäten stammte zu 74 % aus dem obersten sozioökonomischen Viertel und zu 3 % aus dem untersten Viertel. Aber selbst an

den 429 Universitäten der prestigeärmsten Stufe stammen noch zwei Drittel der Studierenden aus der oberen sozioökonomischen Hälfte. Weiter verschärft wird die Situation dadurch, dass die Quote der Studierenden, die das Studium erfolgreich abschließen, mit abnehmendem sozioökonomischem Status sinkt (vgl. Carnevale u. Rose 2004, 106–9). Diese Zahlen sind so drastisch, dass sich eine grafische Darstellung erübrigt. Stattdessen bietet sich ein Bild von Sandel (2021, 169) an: Die amerikanische Hochschulbildung ist wie ein Aufzug in einem Gebäude, das die meisten Menschen im obersten Stockwerk betreten. Zumindest in der Tendenz trifft dieses Bild, wie vor allem Studie 2 belegt, auch für das Hochschulwesen Mitteleuropas zu, das weniger stratifiziert ist. In der Soziologie spricht man auch von einem Matthäus-Effekt: Wer hat, dem wird gegeben.

Erinnern Sie sich noch an die Ausgangshypothese zur Meritokratie in unseren heutigen westlich orientierten Gesellschaften? Sie lautete: Der sozioökonomische Status der Eltern hat keinen Einfluss auf den Bildungserfolg ihrer Kinder. Diese Hypothese kann aufgrund der dargestellten drei Studien – und zahlloser weiterer Untersuchungen – in aller Deutlichkeit verworfen werden. Das Gegenteil ist richtig: Der sozioökonomische Hintergrund, also das Einkommen und das Bildungsniveau der Eltern haben einen massiven Einfluss auf den Bildungserfolg der Menschen – sowohl auf die effektiven Fähigkeiten wie auch auf den Erwerb von Bildungsabschlüssen.¹⁰ Mit anderen Worten: Die Bildungschancen sind ungleich verteilt und sinken mit tieferem SES der Herkunftsfamilie. Umfassende Vergleichsstudien wie TIMSS oder PISA gibt es auf Hochschulstufe leider nicht. Die vorhandenen Daten deuten allerdings darauf hin, dass der SES auch an den Hochschulen weiterhin mit dem Studienerfolg korreliert. Wie oben anhand der Studie von Carnevale u. Rose (vgl. 2004, 108f.) gezeigt, sinkt in den USA die Quote der Studierenden, die das Studium erfolgreich abschließen, mit abnehmendem sozioökonomischem Status. Daten aus Europa zeigen den gleichen Zusammenhang (Larsen, Sommersel u. Larsen 2013, 35–37).

Die Leistungen bis hin zum Erwerb von offiziellen Abschlüssen, welche die Studierenden an Hochschulen zeigen, sind folglich durch ihre familiäre Herkunft deutlich beeinflusst. Die Meritokratie ist somit eher eine kollektive Illusion, die der Rechtfertigung sozialer Ungleichheiten dient. Ihr eigenes Glück schmieden können vor allem jene, die aus einer bildungsnahen Familie stammen (vgl. Becker u. Hadjar 2009, 55).

10 Diese Unterschiede werden nach den formalen Abschlüssen noch weiter verstärkt. Denn je höher ihr Bildungsniveau, desto eher bilden sich Menschen weiter. So besuchten 2016 rund 80 % der Schweizer Bevölkerung mit einem Tertiärabschluss in den vorangegangenen 12 Monaten eine Weiterbildung, während es bei den Personen ohne Abschluss auf Sekundarstufe II lediglich rund 30 % waren (vgl. Wolter et al. 2018, 291–94).

2.3 Noten machen Leistungen

Nun bliebe noch die Hoffnung, dass wenigstens die Noten, die wir im Bildungswesen den Lernenden erteilen, den fachlichen Wissensstand der Lernenden verlässlich wiedergeben. Wir können dies wiederum als Hypothese formulieren: Noten geben den fachlichen Wissensstand von Lernenden so verlässlich wieder, dass gleiche Noten auf einen vergleichbaren Stand des Wissens und Könnens schließen lassen.

Wie bereits angesprochen: Mehr als fünfzig Jahre Forschung haben diese Hypothese falsifiziert. Da Noten auch für die Leistungsbewertung an Hochschulen eine große Rolle spielen, lohnt es sich, genau hinzusehen, wo denn die Schwierigkeiten liegen. Wir ziehen zu diesem Zweck wiederum drei Studien heran.

Studie 1 (Ingenkamp 1971)

Die wohl immer noch bekannteste Untersuchung zur Vergleichbarkeit von Schulnoten im deutschsprachigen Raum stammt von Ingenkamp (1971). Ihre Resultate wurden inzwischen mehrfach bestätigt. Ingenkamp untersuchte 37 Schulklassen eines Westberliner Bezirks. Die Schülerinnen und Schüler legten einen unabhängigen, lehrplanbasierten Rechtschreib- und Rechentest ab, dessen Resultate mit ihren Noten („Zensuren“) verglichen wurden. Dabei zeigten sich sowohl auf Ebene der einzelnen Schülerinnen als auch auf Ebene der Klassen markante Diskrepanzen.

Vergleicht man z. B. Schüler, die in ihrer Klasse die Rechennote 2 (=gut) erhalten haben, so zeigt der unabhängige Test enorme Leistungsunterschiede. Die 2-Schüler aus Klasse 7 erzielten im Durchschnitt ca. 53 Testpunkte, während die 2-Schüler aus Klasse 16 durchschnittlich ca. 31 Testpunkte erzielten. Auch die Rangplätze der Klassen unterschieden sich deshalb erheblich, je nachdem, ob man den unabhängigen Rechentest oder die Rechennote heranzog. So wies etwa Klasse 8 den drittbesten Notenschnitt aller 37 Klassen auf, erreichte aber anhand der Testresultate lediglich Rang 27 (vgl. Abb. 3). Die hier berichteten Beispiele sind zwar die ausgeprägtesten Fälle, aber die Unterschiede in den Testleistungen waren über alle 37 Klassen hinweg extrem. So „zeigte sich, dass in jeder der drei schlechtesten Klassen drei Viertel aller Schüler nur Testleistungen erreichten, die unter dem niedrigsten in den beiden besten Klassen überhaupt vorkommenden Wert lagen“ (Ingenkamp 1971, 157). Sämtliche Schüler der beiden besten Klassen hätten somit in den beiden schlechtesten Klassen die Bestnote erhalten müssen. Diese Resultate führen zu zwei Schlüssen:

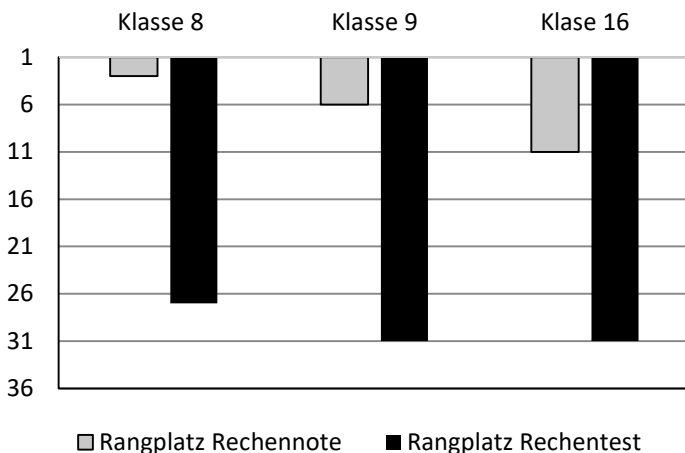


Abb. 3: Rangplatzunterschiede nach Noten und unabhängigen Testresultaten am Beispiel von 3 Klassen. Quelle: Eigene Abbildung, basierend auf Daten aus Ingenkamp 1971, 159

Die Lehrpersonen bilden mit hoher Zuverlässigkeit Rangreihen innerhalb ihrer jeweiligen Klasse (siehe Kap. 6.2.3: kollektive Bezugsnorm). Hingegen hängt es vor allem von der zufälligen Einteilung in eine Klasse ab, ob z. B. eine Schülerin mit mittlerem Leistungsniveau eine 1 oder eine 6 erhält.

Mit anderen Worten: „Die Zensuren haben über den Rahmen einer Klasse hinaus überhaupt keinen Vergleichswert. (...) Diese Ergebnisse besagen nicht mehr und nicht weniger, als dass für unser gesamtes schulisches Berechtigungswesen keine sachliche Rechtfertigung besteht. Die Annahme, der Schüler, der seiner guten Zensuren wegen von der Aufnahmeprüfung [für das Gymnasium, tz] dispensiert wird, weise bessere Schulleistungen auf als der Sitzbleiber in einer anderen Klasse, ist bloße Spekulation“ (Ingenkamp 1971, 161f.). Nun könnte sich seit den 1960er Jahren viel geändert haben. Wenden wir uns deshalb einer aktuelleren Studie zu.

Studie 2a (Kronig 2007)

Diese Studie gleicht der zuvor berichteten: Es werden Schulnoten in den Fächern Deutsch und Mathematik verglichen mit den Resultaten aus unabhängigen Leistungstests in den beiden Fächern. Die Datengrundlage ist aber breiter: Erhoben wurden 112 sechste Schulklassen (Durchschnittsalter ca. 12 Jahre) mit 2104 Schülerinnen und Schülern in 21 deutschsprachigen Schweizer Kantonen im Schuljahr 2000/2001.

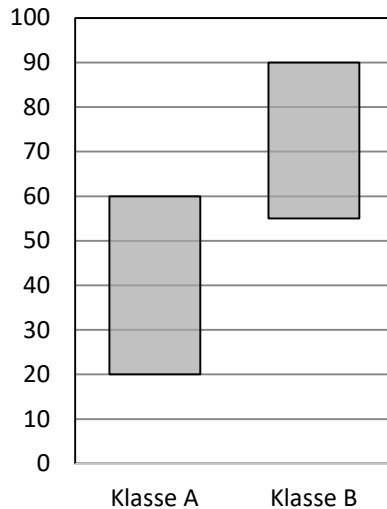


Abb. 4: Bandbreite der unabhängig gemessenen Leistungen zweier Deutschschweizer Schulklassen im Fach Deutsch. Quelle: Eigene Abbildung, basierend auf Kronig 2007, 199f.

Die Studie zeigt im Prinzip die gleichen Resultate wie jene von Ingenkamp (1971). Sowohl in den Fächern Mathematik wie Deutsch ist es möglich, dass Schülerinnen mit der gleichen Schulleistung alle Zeugnissnoten aus dem Band von 4 (=genügend) bis 6 (=sehr gut) erhalten.¹¹ Auch der Vergleich ganzer Klassen fällt problematisch aus, wie sich anhand eines Extrembeispiels aus dem Fach Deutsch zeigt (vgl. Abb. 4). Der schwächste Schüler aus Klasse A löste 20 % der unabhängigen Testaufgaben richtig, die beste Schülerin 60 %. In Klasse B kann bereits die schwächste Schülerin 55 % der Testaufgaben richtig lösen, der stärkste Schüler 90 %.

Die Leistungsspektren der beiden Klassen zeigen also objektiv betrachtet nur eine minimale Überschneidung – obwohl die Klassen sich nach äußeren Merkmalen sehr ähnlich sind: Beide befinden sich in einer Kleinstadt und stehen kurz vor dem Übertritt in die Sekundarstufe. In Klasse A sind 20 Schülerinnen und Schüler, davon 5 aus zugewanderten Familien, in der anderen Klasse sind es 22, davon 4 zugewanderte. Der beste Schüler aus Klasse A

11 Das sind 5 Noten: 4, 4.5, 5, 5.5, und 6. Die Bandbreite der Leistungen, die sowohl mit einer 4 als auch einer 6 benotet werden können, ist dabei in der Mathematik sehr groß und variiert von knapp unter 60 % bis knapp über 80 % im unabhängigen Test richtig gelösten Aufgaben. In Deutsch umfasst diese Bandbreite ca. 10 %.

würde also in Klasse B zu den schlechtesten gehören. Die lebenspraktischen Folgen eines solchen Unterschieds sind massiv: „Der schwächste Schüler aus der Klasse B muss – obschon seine Leistungen in der anderen Klasse mit Bestnoten bewertet würden – um seine Versetzung bangen. Seine Note weicht mit 3.25 deutlich vom geforderten Mindestniveau für ein bestandenes Schuljahr ab, und er wird auf jeden Fall später einen anspruchsräreren Schultyp besuchen müssen. In der anderen Schulklasse hingegen würde seine Bildungsbiografie mit größter Wahrscheinlichkeit einen völlig anderen Verlauf nehmen“ (Kronig 2007, 199f.).

Studie 2b (Kronig 2007)

Bei dieser Studie handelt es sich um eine Auswertung zusätzlicher Daten aus dem gleichen Datensatz wie in Studie 2a. Die untersuchten sechsten Klassen wurden nämlich zweimal unabhängig getestet, einmal am Anfang und einmal am Ende des Schuljahres. Das ermöglichte eine äußerst aufschlussreiche Analyse zur Wirkung, die erhaltene Noten auf die weitere Leistungsentwicklung von Lernenden haben.

Verglichen wurden Schülerpaare, die im unabhängigen Test zu Beginn des Schuljahres den gleichen Leistungsstand gezeigt hatten. Nur hatte jeweils eines der beiden Kinder im vorangehenden Zeugnis eine Vier und das andere eine Fünf erhalten. Dank der umfassenden Datenerhebung konnte sichergestellt werden, dass die gebildeten Paare jeweils bezüglich ihres Geschlechts und Herkunftslandes (und ggf. des Einwanderungszeitpunkts) gleich waren. Zudem waren sie bezüglich nur ungefähr erfassbarer Variablen wie kognitiven Grundfähigkeiten oder dem SES vergleichbar. Die Resultate waren frappant: Sowohl in Deutsch als auch in Mathematik entwickelte sich die Leistung jener Schülerinnen, die mit einer Fünf benotet worden waren, signifikant besser als die Leistung jener Schüler, die nur eine Vier erhalten hatten (vgl. Abb. 5). Grob gesagt erreichten Schülerinnen, die im vorangehenden Zeugnis eine Fünf erhalten hatten, nach einem knappen Schuljahr in Deutsch 11 % und in Mathematik 12 % mehr richtig gelöste Testaufgaben.¹² Der Effekt der Notengebung auf die Leistung ist somit in beiden Fällen groß (vgl. Kronig 2007, 206–9).

12 Aus Abb. 5 lässt sich der Unterschied in Testpunkten nicht direkt ablesen. Für statistisch Interessierte: Die Leistungswerte wurden z-transformiert, um sie vergleichbar zu machen. Eine Standardabweichung im Deutschttest (n=74 Paare) betrug ca. 14 % richtig gelöste Aufgaben, im Mathematiktest (n=91) ca. 19 %. Aufgrund dieser Daten lassen sich die obigen Schätzungen berechnen.

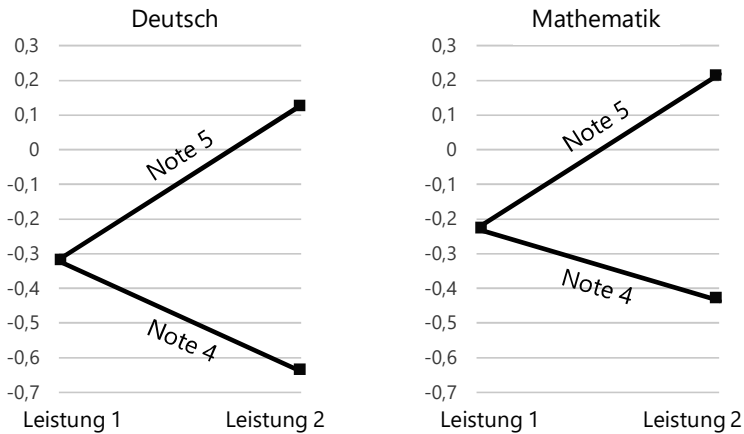


Abb. 5: Auswirkung unterschiedlicher Noten auf die Leistungsentwicklung innerhalb eines Schuljahrs bei Schülerinnen mit anfänglich gleichem Leistungsstand. Quelle: Eigene Abbildung, basierend auf Daten und Darstellungen aus Kronig 2007, 206–208

Dies bedeutet: Noten stellen de facto nicht nur fest, wie gut eine schulische Leistung ausfällt. Sie üben gleichzeitig einen erheblichen Einfluss darauf aus, wie die nachfolgende Lernleistung der benoteten Personen ausfällt. In Anlehnung an den Titel der Novelle von Gottfried Keller (1975 [1874]) „Kleider machen Leute“ ließe sich deshalb auch pointiert sagen: Noten machen Leistungen.

Betrachten wir nochmals die Hypothese, die zu prüfen war: Noten geben den fachlichen Wissensstand von Lernenden so verlässlich wieder, dass gleiche Noten auf einen vergleichbaren Stand des Wissens und Könnens schließen lassen.

Die Studien 1 und 2a widerlegen diese Hypothese klar – sie zeigen, dass Noten viel stärker von der zufälligen Zusammensetzung der Lernendengruppe abhängen als von der tatsächlichen fachlichen Leistung der Lernenden. Hierbei handelt es sich um einen so genannten Referenzfehler (siehe Kap. 6.4.1): Lehrpersonen wenden unabhängig von der effektiven Leistung ihrer Schulklasse stets ungefähr das gleiche Notenband an (vgl. Kronig 2007, 199–201). Der Hauptgrund dafür liegt darin, dass bei der Leistungsbewertung innerhalb einer Gruppe – gleich, ob Schulklasse oder Teilnehmende der gleichen Vorlesung – in erster Linie auf eine kollektive Bezugsnorm (siehe Kap. 6.2) abgestellt wird. Dies ist selbst dann der Fall, wenn Lehrende sich bewusst an fachlichen Kriterien statt an einem Vergleich innerhalb der Lernendengruppe zu orientieren versuchen. An vielen Orten werden aber bestimmte Notendurchschnitte auch durch Schulleitungen, Behörden oder ande-

re Instanzen verlangt, zumindest inoffiziell.¹³ Das hat Folgen, die man eigentlich nur als absurd bezeichnen kann: „Entgegen allen Beteuerungen ist deshalb vermutlich der Einfluss von pädagogischen Maßnahmen auf die Bildungsbiografie begrenzt. Denn potenziell kann ein Wechsel des Wohnorts eine größere Veränderung der Leistungsbewertung zur Folge haben, als sie mit einer erhöhten Leistungsbereitschaft, intensivem Arbeiten und einem erfolgreichen Unterricht jemals zu erreichen wäre“ (Kronig 2007, 214).

Studie 2b zeigt, dass das Problem nochmals um eine Dimension größer ist. Denn wir erwarten von Noten nicht nur eine diagnostische, sondern auch eine prognostische Validität. Sie sollen also nicht nur eine gültige Diagnose der aktuellen Leistung darstellen, sondern auch eine Prognose der künftigen Leistungen erlauben (Ingenkamp u. Lissmann 2008, 150). Wenn aber die Note selbst massive Auswirkungen darauf hat, wie sich Lernende im weiteren Lernverlauf verhalten, dann wird die prognostische Validität tautologisch: Die Note wird – mindestens in Teilen – zur selbsterfüllenden Prophezeiung.

In den Schulsystemen Deutschlands, Österreichs und der Schweiz werden ab der Sekundarstufe verschiedene Schulniveaus geführt, so dass diese problematischen Wirkungen von Noten früher und stärker greifen als in anderen Ländern. So zeigen die Daten von Kronig für den Übertritt von der Primar- in die Sekundarstufe: „In acht von zehn Fällen gibt es im Anschluss an den Selektionsentscheid einen Schüler, der vergleichbare Leistungen hat, jedoch an den gegenüberliegenden Schultyp überwiesen worden ist“ (Kronig 2007, 217). Die Noten, welche die Schülerinnen erhalten, führen also zu mehrheitlich zufälligen Übertritten in die verschiedenen Anforderungstypen der folgenden Schulstufe. Das wiederum hat gewiss einen noch stärkeren Einfluss auf die künftige Leistungsentwicklung als die Wirkung von Noten im Verlauf eines Schuljahres an der gleichen Schule (leider gibt es m. W. keine Studien dazu).

Zu den berichteten Befunden passen auch Metaanalysen zum Zusammenhang zwischen Abiturnoten und Studienabschlussnote. In Ländern wie Deutschland, Österreich und der USA erklären die Abiturnoten durchschnittlich ungefähr 10 % der Unterschiede in den Studienabschlussnoten – die anderen 90 % der Unterschiede können die Abiturnoten nicht erklären (vgl. Trapmann et al. 2007 für Deutschland und Österreich, Burton u. Ramist 2001

13 Das Schulrecht in Deutschland gibt freilich das Primat klar der Kriterialnorm (siehe Kap. 6.2), mit landesspezifischen Unterschieden in der Deutlichkeit (vgl. Kostorz 2016, 288).

für die USA¹⁴).¹⁵ Und bei den erklärten 10 % ist zu beachten, dass ein Teil davon durch die „leistungsmachende“ Wirkung der Noten hervorgebracht wird, wie sie anhand von Studie 2b gezeigt wurde.

Ähnlich wie bei der Frage nach dem Einfluss des SES auf den Bildungserfolg gibt es auch zum Zusammenhang zwischen fachlichem Wissen und Können von Studierenden und ihren Noten keine umfangreichen Vergleichsstudien wie jene von Ingenkamp (1971) oder Kronig (2007).¹⁶ Allerdings deuten ausgewählte Daten auf gleich gelagerte Schwierigkeiten hin:

- Eine exploratorische Analyse von Aufsatzbewertungen im Rahmen der Aufnahmeprüfung an eine Schweizer Hochschule von internationalem Rang zeigte, dass verschiedene Expertinnen die Aufsätze sehr unterschiedlich bewerteten. So wurden für drei Aufsätze Notenbänder von 3 bis 6, 4 bis 5.5 und 3 bis 5.5 gefunden. Die Autoren machen darauf aufmerksam, dass diese großen Unterschiede für alle drei Bewerbenden den Ausgang des Aufnahmeverfahrens entscheidend beeinflussen können (vgl. Eberle u. Metzger 1988, 69–85, bes. 75). Dieses nicht repräsentative Resultat deutet zumindest an, dass sich die für die tieferen Schulstufen gezeigten Probleme auch beim Übertritt in die Hochschule fortsetzen.
- Eine ältere Studie aus einer deutschen Universität offenbart anhand der „ersten Lehrerprüfung“, dass unterschiedliche Fächer ungleich streng benotet werden. In den musischen Fächern waren die Noten am besten, und „wie in der Schule wurden also die klassischen Auslesefächer Deutsch und Mathematik auch an der Hochschule durchschnittlich am schlechtesten bewertet“ (Ziegenspeck 1999, 247, für die ganze Studie vgl. 238–265,

14 Hinweis zu Burton u. Ramist (2001): Für die Bachelorabschlüsse (undergraduate grades) wurde $r=.42$ gefunden, für Masterabschlüsse (graduate grades) war $r=.29$. Das ergibt eine Varianzaufklärung von 17 % bei den Bachelor- und von 8 % bei den Masterabschlüssen.

15 Auf einer statistischen Ebene mag man mit solchen Werten zufrieden sein, da bisher keine Variablen gefunden wurden, die den Studienerfolg wesentlich besser erklären, so etwa die Einschätzung von Trapmann et al. (2007, 24f.). Allerdings muss es bei der Bewertung von Lernleistungen letztlich um für die Entwicklung der einzelnen Studentin gültige Prognosen gehen, und dafür sind solche Werte ungenügend (vgl. Ingenkamp u. Lissmann 2008, 150f.). Deshalb kann ich aus pädagogischer Perspektive keine Begründung dafür erkennen, die Reliabilitätswerte statistisch zu korrigieren, wie dies Trapmann et al. (2007) tun (wodurch sie auf eine Varianzaufklärung in der Größenordnung von 25 % kommen).

16 Dies liegt auch – aber nicht nur – daran, dass Studiengänge verschiedener Hochschulen aufgrund unterschiedlicher Curricula und anderer lokaler Besonderheiten schwieriger zu vergleichen sind als etwa Schulklassen, in denen nach dem gleichen Lehrplan unterrichtet wird.

Erstpublikation der Resultate 1974). Weiter stieg der Notendurchschnitt aller Fächer innerhalb von 7 Jahren von 2.6 auf 2.3 an, was kaum darauf zurückzuführen sein wird, dass die Studierenden so schnell so viel besser wurden. Insgesamt kommt Ziegenspeck (1999, 242, Hervorhebungen von Ziegenspeck) zum deutlichen Schluss: „Hochschulprüfungen sollen eine *Diagnose* der im Studium erworbenen Fähigkeiten und Kenntnisse und eine *Prognose* im Hinblick auf die berufliche Tätigkeit erlauben. Diesen beiden Aufgaben kann die herkömmliche Prüfung kaum gerecht werden, weil sie weder objektiv ist noch den Kriterien der Reliabilität und Validität standzuhalten vermag.“

- Die prognostische Validität von Studienabschlussnoten für den Berufserfolg ist bescheiden. Daten aus den USA zeigen, dass die Abschlussnoten maximal 10 % der Unterschiede in den beruflichen Leistungen erklären. Dieser Wert wird nach einem Jahr Berufstätigkeit erreicht und sinkt danach stark ab (vgl. Roth et al. 1996). Auch wenn man berücksichtigt, dass Hochschulabschlüsse je nach Fachrichtung unterschiedlich ausgeprägt auf eine konkrete Berufsausübung vorbereiten und einige Hochschulabsolvierende Stellen antreten, die nur teilweise mit ihrem Studienfach zusammenhängen, stimmt diese geringe prognostische Validität doch sehr bedenklich. So überrascht es auch nicht, dass unstrukturierte Bewerbungsgespräche den beruflichen Erfolg gleich gut vorhersagen wie Studienabschlussnoten – und dass strukturierte Bewerbungsgespräche und Intelligenztests dies sogar deutlich besser können (vgl. Roth et al. 1996, 553).

Vielleicht fragen Sie sich jetzt, welchen Sinn das Durchführen von Leistungsnachweisen an Hochschulen angesichts dieser Datenlage überhaupt ergibt. Das ist freilich eine Frage der Perspektive: Als objektive Messung von Leistungen sind Leistungsbeurteilungen im Bildungswesen angesichts der vorgestellten Befunde in der Tat äußerst fragwürdig. Allerdings kann ihre Qualität und Aussagekraft durch ein angemessenes Vorgehen durchaus erhöht werden; und die Kenntnis ihrer Schwächen ermöglicht einen angemesseneren Umgang mit Leistungsbeurteilungen – diesen Zwecken ist das vorliegende Buch verpflichtet. Zudem sind die Leistungsnachweise und -beurteilungen als subjektive Standortbestimmungen und zur intersubjektiven Verständigung zwischen Lehrenden und Lernenden generell sinnvoll. Auch dies ist ein wesentlicher, in diesem Buch erläuteter Aspekt.

Möglicherweise stellen Sie sich auch noch eine andere Frage: Weshalb wird denn an Noten von der Primarschule bis zur Hochschule derart festgehalten, wenn die Datenlage so eindeutig ist, dass Experten wie Ingenkamp und Kronig die Notengebungen außerhalb einzelner Klassenzimmer schlichtweg als untauglich bezeichnen – und dies in einer Deutlichkeit, die im ge-

wöhnlich zurückhaltenden wissenschaftlichen Sprachgebrauch ihresgleichen sucht? Dieser Frage widmet sich das folgende Unterkapitel.

2.4 Metrische Fixierung und die Zertifikatsgesellschaft

In Kapitel 2.1 habe ich aufgezeigt, wie ab der Mitte des 19. Jahrhunderts das Messen und Vergleichen verschiedenartigster Gegenstände, Tätigkeiten und Artefakte aufkam, das uns heute so selbstverständlich ist – von der Abmessung von Möbeln über die Sportresultate bis zu Börsenindizes und Hochschulratings. Bei dieser zunehmenden Vermessung der Welt spielt auch das Bildungswesen eine entscheidende Rolle. Für das Verständnis, weshalb Schulen nicht von Noten wegkommen oder weshalb unsere Gesellschaft trotz aller gegenteiligen Belege sich als Meritokratie inszeniert, analysieren wir diese Rolle nun genauer.

Dabei ist zunächst zu unterscheiden zwischen der Gesellschaft als strukturierter Form menschlichen Zusammenlebens und nicht regulierten Aspekten des sozialen Zusammenlebens. Gemeinsam bilden sie „ein soziales Plasma, in dem praktisches Wissen und soziale Ressourcen mobilisiert werden, um Subjekte zu schmieden und zu einer Ordnung zu verfestigen“ (Angermüller u. Maeße 2015, 64). Seit dem 17. Jahrhundert sind die Prozesse, die das Soziale in gesellschaftliche Ordnungen überführen, zunehmend abstrakter und unpersönlicher geworden. Dies war gewiss bedingt durch die zivilisatorischen Prozesse im Rahmen der zunehmenden Urbanisierung, die in Kapitel 2.1 erläutert wurden. Der Philosoph Michel Foucault argumentierte, dass der Übergang von der Vormoderne zu marktwirtschaftlichen Gesellschaften durch zwei neue Strukturen geprägt ist: den Markt und den Staat. Dabei greift der moderne Staat nicht mehr primär durch direkte Disziplinierung wie körperliche Strafen ein, sondern indem er indirekt auf ökonomische Bereiche und individuelle Lebenskonzeptionen einwirkt (vgl. Foucault 2004, 367–98). Gemäß Foucault beeinflussen Markt und Staat nicht nur die wirtschaftlichen Aspekte, sondern strukturieren auch soziale Ungleichheit. Markt und Staat entwickeln dazu Verfahren, um die verschiedenen und vielfältigen Elemente der Gesellschaft (scheinbar) vergleichbar und gleichförmig zu machen. Zum Beispiel ermöglichen es Normen wie Meter und Kilogramm oder komplexe Messungen wie Kardiogramme oder Raumbewegungen durch Geoinformationssysteme, Dinge zu vergleichen und zu bewerten. Messungen und Vergleiche orientieren sich dabei stark an ökonomischem und mathematischem Wissen und sind vor allem für Marktmechanismen wichtig. Der Staat hingegen nutzt ordnende Mechanismen wie Schulnoten oder Zertifikate, um das Soziale zu strukturieren. Diese Praktiken helfen, Personen, Gegenstände und Ereignisse in eine geordnete Form zu bringen und Entscheidungen zu treffen.

Dieses Regime technologisierter Praktiken des Regierens (...) bezeichnen wir als numerokratischen Markt-Staat. Ganz gleich in welcher Ausprägung numerokratisches Regieren daher kommt – als Staatswirtschaft im Realsozialismus oder als Wettbewerbsstaat im Neoliberalismus – im Ergebnis werden mit großen Zahlen, Maßstäben, Benchmarkings, Kenngrößen, Standards, Rankings, Skalen, Matrizen große Bereiche des Sozialen geordnet, unter Kontrolle gestellt und in die Struktur der Gesellschaft gebracht. (Angermüller u. Maeße 2015, 66)

Die geschilderte Numerokratisierung hat in den letzten Jahrzehnten zwei weitere Schübe erhalten. Einerseits wurde sie begünstigt durch die verstärkte Orientierung an freien Märkten seit den 1980er Jahren. Dies führte in den meisten westlich orientierten Ländern zur Privatisierung von staatlichen Dienstleistungen. Deren Qualität wird in der Regel anschließend durch gewinnorientierte Wirtschaftsprüfungsfirmen geprüft (vgl. Shore u. Wright 2015, 427). Andererseits erhielt die Numerokratisierung durch den immer weiterreichenden Einsatz von Software-Algorithmen einen zusätzlichen Schub. Dies gilt sowohl für die marktwirtschaftliche als auch die staatliche Sphäre: In beiden Sphären metrisieren Algorithmen immer mehr Bereiche. Im Bereich der Marktwirtschaft ist etwa an die Algorithmen zu denken, mit denen Social Media wie Facebook oder TikTok bestimmen, wem wann welche Mitteilungen angezeigt werden. Im staatlichen Bereich bestimmen zum Beispiel Algorithmen zur Vorhersage von Kriminalität, in welchen Gebieten wann wieviel Polizeipräsenz erfolgen soll. Die Bereiche vermischen sich aber auch, vor allem findet eine starke Individualisierung der Prozesse statt. Unsere täglich zurückgelegten Schritte, die Herzfrequenz, unser berufliches Netzwerk oder Ausgabeverhalten werden nicht nur uns persönlich gespiegelt – ganz unabhängig vom Grad ihrer Genauigkeit. Auch Institutionen oder ihren Vertreterinnen wie Versicherungen oder Ärzten werden diese Daten gemeldet und sollen Zeugnis davon ablegen, wer wir sind und wie wir uns verhalten. Was sich dabei manchmal wie eine spielerische Selbstdiagnose anfühlt, ist eigentlich ein Zustand permanenter Selbstüberprüfung, der durch immer neue Indikatoren und ständige Feedbackschleifen zwischen Mensch und Maschine angetrieben wird (vgl. Burrell u. Fourcade 2021, bes. 228f.).

In der gesellschaftlichen Handhabung von Bildung ist insbesondere die ordinale Logik von Bildungszertifikaten maßgeblich. Diese wird in Anlehnung an die Theorie von Collins (1979) als Kredentialismus bezeichnet: „Erworbene Bildungspatente (‘credentials’) regeln demnach die Zugangsmöglichkeiten zu den höheren und besser bezahlten Berufspositionen und prägen Karriere- und Mobilitätschancen in bedeutsamer Weise“ (Becker u. Hadjar 2009, 54). Bildung wird somit gemäß Collins zunehmend als Instrument zum Erwerb von Bildungszertifikaten denn im Hinblick auf inhaltliche Lerngelegenheiten betrachtet. Auch wenn bei Weitem nicht alle Menschen Bildungsgänge primär aus solchen Motiven absolvieren: Kredentialistische Mechanismen sind in wesentlicher Weise dafür verantwortlich, dass sozioökonomi-

sche Herkunftsbedingungen immer noch starken Einfluss auf die gesellschaftlichen Erfolgschancen haben (siehe Kap. 2.2).

Denn Bildungszertifikate dienen dazu, soziale Unterschiede bezüglich Status oder Einkommen nicht nur aufrechtzuerhalten, sondern auch zu legitimieren. Die notwendigen Vergleichs- und Prüfungsverfahren „sind heutzutage in unterschiedlichsten Formaten in fast allen Lebensbereichen institutionalisiert. Sie zielen darauf ab, individuelle Leistungsunterschiede vermeintlich objektiv zu ermitteln und zu belohnen. Und genau damit stellen sie solche Unterschiede maßgeblich her“ (Verheyen 2014, 67). Sie sind deshalb auch als Techniken der sozialen Exklusion zu verstehen (ebd., 84). Damit ergibt sich außerdem eine enge Verbindung zwischen Bildungsabschlüssen und wirtschaftlichen Erfolgschancen (vgl. Hadjar u. Becker 2009, 201; Konietzka 2009, 262).

Der Kredentialismus beschränkt sich längst nicht mehr auf individuelle Praktiken, sondern hat auf private und staatliche Institutionen übergegriffen. Power (1997) hat dafür anhand einer Analyse von Buchprüfungspraktiken den Begriff der „Audit Society“ geprägt, der im Deutschen als Prüfungsgesellschaft übersetzt werden könnte: Immer mehr Firmen und Institutionen müssen über immer mehr Prozesse und Produkte Rechenschaft ablegen. Das beginnt dabei, dass sich Kleinbetriebe mit ISO-Zertifikaten schmücken und fast jedes mögliche Produkt durch die Stiftung Warentest geprüft werden kann, und endet bei Prozessen wie der Akkreditierung von Gesundheitseinrichtungen oder Hochschulen (siehe auch Abb. 6).¹⁷ Die entsprechenden Zertifikate sind entweder Voraussetzung dafür, ein Angebot oder Produkt bereitstellen zu dürfen oder dienen als Verkaufsargument. Insbesondere im Fall von Hochschulen gehen dabei Kredentialismus und Numerokratie miteinander einher.

Das Phänomen, dass mehr wert ist, was geprüft wird, zeigt sich übrigens auch innerhalb des Bildungssystems: Werden bestimmte Fächer oder Veranstaltungen anhand anspruchsvoller Leistungsnachweise bewertet, so verlieren

17 Dabei findet oft keine direkte Prüfung konkreter Produkte oder Prozesse mehr statt, sondern eine Prüfung der institutionalisierten Prüfverfahren wie der Qualitätssicherung. Dieses Prinzip der Meta-Prüfung bezeichnet Power (1997, 12) als „control of control“. Das leistet einer weiteren Depersonalisierung der Prozesse Vorschub – theoretisch könnten künftig auch Algorithmen geprüft werden, und diese wiederum könnten die Prüfung vornehmen etc.

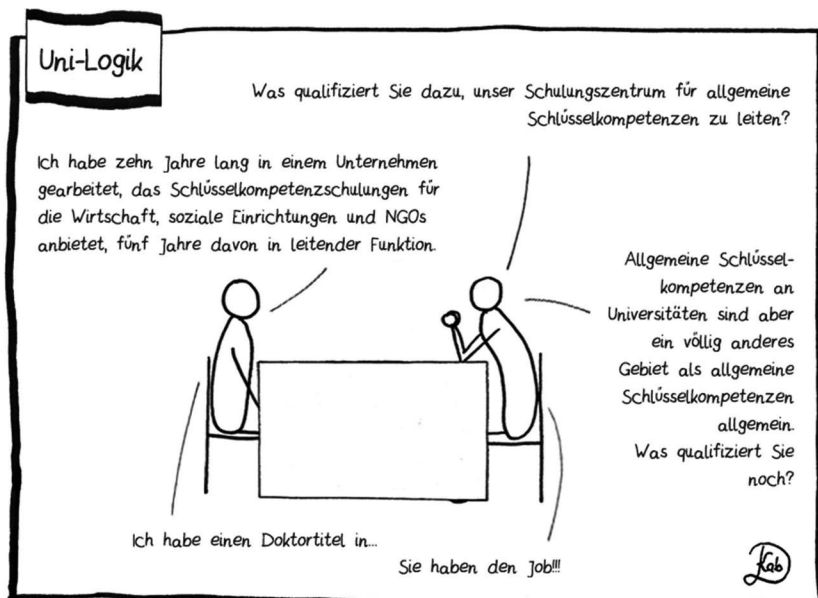


Abb. 6: Kredentialismus existiert auch in Hochschulen. Quelle: Constructive Amusement/Daniel Al-Kabbani, <http://www.constructive-amusement.de/comics/uni-logik>

nicht oder weniger anspruchsvoll bewertete Bereiche an Wert.¹⁸ Dieser kredentialistische Effekt muss insbesondere bei der Gestaltung von Curricula bedacht werden, weil Leistungsnachweise ganz wesentlich steuern, was, wie und wie viel Studierende lernen (siehe Kap. 5.1 zum Constructive Alignment und Kap. 10 zur Curriculumentwicklung).

Geprüft werden aber nicht nur einzelne Institutionen oder Firmen, sondern auch ganze Systeme. Im Bildungsbereich zum Beispiel führt die OECD die PISA-Rangliste der öffentlichen Bildungssysteme, und verschiedene –

¹⁸ Das kann bisweilen zu bizarren Auswüchsen führen, wie Thomann (2019, 250) anhand eines Beispiels von Heid zeigt: Da stellt ein Religionslehrer am ersten Schultag neu eingeschulter Gymnasiastinnen sein Fach Religion mit der Bemerkung vor, dies sei ein Fach, in dem man auch durchfallen könne. „Um die pädagogische ‚Ungeheuerlichkeit‘ einer solchen ‚Eröffnung‘ richtig einschätzen zu können, möge man sich zum Vergleich vorstellen, ein Chirurg würde, um die Qualität und Effektivität seines Berufes zu verdeutlichen, seinen (versammelten) Patienten mit Stolz eröffnen: ‚In meiner Praxis kann man auch sterben!‘“ (Heid 1989, 47, zit. nach Thomann).

meist gewinnorientierte – Anbieter erstellen jährliche Hochschulrankings, etwa das World University Ranking von Times Higher Education. Gerade Hochschulen sind heute mit einer verwirrenden Vielzahl von Messungen und Vergleichen konfrontiert, die unterschiedlichste Aspekte bewerten. Das reicht vom Erfolg bei der Einwerbung von Drittmitteln über ihre ökologische Nachhaltigkeit bis zur Bewertung der studentischen Erfahrung oder des Rankings ihres Forschungsausgangs inklusive Zitationsmetriken (vgl. Shore u. Wright 2015, 428). Nicht selten geht es bei diesen Rankings darum, die „Exzellenz“ der jeweiligen Universität unter Beweis zu stellen – ein Diskurs, der das meritokratische Prinzip auf die Hochschulen als Institutionen zu übertragen versucht (vgl. Maasen u. Weingart 2008, 145–48). Der Philosoph Konrad Paul Liessmann (2006, 126) merkt dazu kritisch an: „Die Forderung nach Elite und Exzellenz dient allzu schnell dazu, unliebsam gewordene Forschungsbereiche und Studienrichtungen zuerst finanziell auszuhungern und dann, wegen mangelnder Leistungsfähigkeit, zu schließen.“

Es ist überdies bestens belegt, dass viele dieser Messungen unzuverlässig sind und unerwünschte Effekte haben (vgl. z. B. Muller 2018, Shore u. Wright 2015, Sauder u. Espeland 2009, Maasen u. Weingart 2008 und Strathern 2000). Darin besteht eine große Gemeinsamkeit mit den Schulnoten und auf ihrer Basis beruhenden Bildungszugängen. Die Gründe dafür können hier nicht ausgelotet werden, hingegen werden einige Folgen der metrischen Anreizsteuerung an Hochschulen und vergleichbaren Institutionen aufgezeigt. Viele davon haben Gemeinsamkeiten mit problematischen Verhaltensweisen von Studierenden und Dozierenden im Umgang mit Leistungsnachweisen.

Bevor Sie die folgende Liste durchlesen, lade ich Sie ein, einige Fragen durchzudenken: Wie reagieren Sie auf Zahlenvorgaben Ihrer Hochschule? Wie würden Sie die Leistungen von Hochschulen evaluieren? Welche Taktiken im Umgang mit Leistungsnachweisen und Bewertungen sind Ihnen von Ihren Studierenden bekannt? Und welche davon haben Sie auch schon selbst in Ihrem Studium angewendet?

Werden Firmen oder staatliche Institutionen an quantitativen Indikatoren gemessen, führt dies zu folgenden Effekten:

- „Gaming the system“: Je mehr ein quantitativer sozialer Indikator für das Fällen von Entscheidungen verwendet wird, umso mehr kommt er unter Korruptionsdruck und umso mehr verzerrt er die sozialen Prozesse, die er eigentlich überwachen soll (Campbells Gesetz, Campbell 1976, 49). Vor allem führen solche Indikatoren sehr schnell dazu, dass das System ausgenutzt wird. Dazu gehören verschiedene Taktiken, etwa die Fokussierung auf möglichst einfache Fälle, zum Beispiel wenn Krankenhäuser komplizierte Fälle vermeiden. Aber auch das Senken von Standards gehört dazu, etwa zur Verbesserung der Abschlussquoten von Hochschulen. Das Beschönigen von Zahlen ist eine weitere Möglichkeit, beispielsweise wenn Schulen ihre Notendurchschnitte verbessern, indem sie schwächere

Schüler als behindert deklarieren (vgl. Muller 2018, 22–24 u. 92; Shore u. Wright, Susan 2015, 424).¹⁹

Ähnliche Strategien sind auch im Umgang mit Leistungsnachweisen bekannt: Das Suchen nach möglichst einfachen Aufgabenstellungen oder anspruchlosen Dozierenden auf der Seite Studierender oder das Anpassen von Standards seitens der Hochschule, um irgendwelchen Metriken gerecht zu werden. Alle Fälle von „Gaming the system“ sind letztlich auf den korumpierenden Effekt extrinsischer Motivationsanreize respektive Kontrollsysteme zurückzuführen (siehe Kap. 4.5 zur Rolle der Motivation beim Lernen).

- **Verhinderung von Innovation:** Wird im Voraus eine begrenzte Palette an (angeblich) messbaren Zielen vorgegeben, so verhindert dies unternehmerisches Handeln innerhalb eines Betriebs oder einer Institution. Denn das Angehen vielversprechender neuer Ziele und von unternehmerischen/wissenschaftlichen Innovationen werden unattraktiv, wenn sie durch die herrschenden Zielvorgaben nicht belohnt oder gar bestraft werden. Parallel dazu reduzieren von außen gesetzte Ziele die intrinsische Motivation, die für die Produktivität im Allgemeinen und ganz besonders für kreative und riskante Unterfangen wichtig ist (vgl. Muller 2018, 61; Amabile 1998; Deci, Koestner u. Ryan 1999, 658).

Im Bereich von Leistungsnachweisen verhindert eine zu starke Orientierung an quantitativen, von außen gesetzten Zielen vor allem tiefenorientiertes Lernen. Dieses zeichnet sich durch echtes Einlassen auf einen sachlichen Gegenstand aus, welches durch Ergebnisoffenheit geprägt ist. Die durch extern gesetzte Ziele verstärkte extrinsische Motivation begünstigt dagegen das Verfolgen von oberflächlichen Lernstrategien (Schiefele u. Schreyer 1994; siehe Kap. 4.5 zur Lernmotivation).

- **Verstärken der Ungleichheit:** Die Anreizsteuerung via metrische Indikatoren benachteiligt oft jene Institutionen, die am weitesten davon entfernt sind, die verlangten Werte zu erreichen. Dies trifft beispielsweise bei

19 Einen Schritt weiter geht der Betrug. Auch dieser kann durchaus eine Folge sozialer Indikatoren sein. So führte die „No Child Left Behind“-Politik in den USA zu einer verstärkten Bewertung der Schulen durch standardisierte Leistungstests. Das führte zu diversen Betrugsfällen, bei denen die Zahlen gefälscht wurden (Muller 2018, 88–93). Für den akademischen Bereich führt Hucho (2010, 73) ein eindrückliches Beispiel an: Ein junger Physiker fälschte Publikationen, die in hochdotierten wissenschaftlichen Journals veröffentlicht wurden. „Die Fälschungen wurden nicht von Verwaltern aufgedeckt (...). Es waren nicht die Kennzahlen oder ‚Leistungsindikatoren‘, die Hinweise auf das Fehlverhalten gaben. Die Ergebnisse hielten der Intuition erfahrener Wissenschaftler und der Neugierde von Kolleginnen und Kollegen nicht stand. Die Selbstkontrolle der Wissenschaft hat funktioniert.“

Krankenhäusern zu, die nicht die gewünschten Heilungsquoten erreichen oder bei Schulen, die bestimmte Testwerte nicht erreichen. Wenn diese Institutionen dann als Folge des Anreizsystems weniger Mittel zugesprochen erhalten als die gemäß Zahlenwerten erfolgreicher Institutionen, verstärkt das die Ungleichheiten, die angeblich bekämpft werden sollen. Dazu gehört auch Verhalten, das nur dem Erreichen der Indikatorenwerte, aber nicht des eigentlichen Ziels dient. Im schulischen Kontext ist dies etwa der Fall, wenn Lernende nur auf das Bestehen bestimmter Prüfungen vorbereitet werden und dadurch der fachliche Lerngewinn reduziert wird (vgl. Muller 2018, 120–23; Jennings u. Lauen 2016).

Im Bereich des Umgangs mit Leistungsnachweisen wirken sich solche Indikatoren auf zwei Ebenen aus. Aufnahmequoten wie ein Numerus Clausus oder bestimmte Testwerte, die erreicht werden müssen, bevorzugen Lernende aus Familien mit höherem SES. Ihnen können Eltern direkt oder durch das Finanzieren von Nachhilfeunterricht helfen, die für das Erreichen der Vorgaben erforderlichen Leistungen zu erbringen. Und wenn die Qualität von Schulen, die schlechtere Indikatorenwerte erreichen, aufgrund der oben beschriebenen Effekte abnimmt, dann wird diese Scherenwirkung des SES noch verstärkt: An solchen Schulen wird die individuelle Erfolgchance der einzelnen Lernenden noch mehr vom familiären Hintergrund abhängen.

- **Zahlenglaube:** Der Glaube an Zahlen hat zumindest irrationale Züge und grenzt teilweise an magisches Denken. Zahlen werden in diesem Denkstil als authentischer betrachtet als die durch sie erfassten Phänomene. Damit einher geht die Vorstellung, dass metrische Indikatoren menschliches Urteilsvermögen ersetzen. Irrationale Züge hat dieses Denken, wenn an ihm trotz zahlreicher Belege für unerwünschte Konsequenzen festgehalten wird – z. B. Gaming the System, Verhinderung von Innovation oder Verstärkung von Ungleichheit (vgl. Muller 2018, 16–18; Maasen u. Weingart 2008, bes. 158).²⁰ Liessmann (2006, 83) schreibt dazu in polemischer Zuspitzung: „Die Fetischisierung der Rangliste ist Ausdruck und Symptom einer spezifischen Erscheinungsform von Unbildung: mangelnde Urteilskraft.“

Der Zahlenglaube äußert sich auch im Umgang mit Leistungsnachweisen: auf Seiten von Lernenden zum Beispiel, wenn die Einschätzung ihrer eigenen Lernfähigkeiten auf Noten oder anderen Testwerten basiert (siehe Kap. 8.2 und 9.1). Aber auch Dozierende können irrationalen Zah-

20 Konersmann (2021, 216) formuliert gar die These, dass die in digitalen Algorithmen gipfelnde Weltvermessung die Stellung einer Religion eingenommen hat: „Die digitale Infrastruktur erneuert das Angebot, mit dem vorzeiten der Monotheismus die Herzen der Menschen gewann: das Angebot, von einer väterlichen Übermacht, die alles sieht und alles weiß, bedingungslos angenommen zu sein.“

lenglauben aufsitzen, etwa wenn sie davon ausgehen, dass die Leistungsverteilung in einer Gruppe der gaußschen Normalverteilung entsprechen muss (siehe dazu Kap. 5.1.3 und 9.2.1) oder wenn sie Leistungen ebenso akribisch wie scheinogenau in Zehntelsnoten unterteilen (Kronig 2007, 64 u. 224).²¹

Die beschriebenen Folgen der Steuerung über metrische Anreize entbehren nicht einer gewissen Ironie. Obwohl sie oft von an freier Marktwirtschaft orientierten Politikern und anderen Entscheidungsträgern etabliert und gerühmt werden, führen sie zu ähnlichen Problemen, wie sie im Realsozialismus osteuropäischer Prägung bestanden: So wie die Bürokratie in der Planwirtschaft für jeden Sektor und jede Fabrik Produktionsziele festlegte, so legen heute Technokraten messbare Leistungsziele für Krankenhäuser, Polizei oder Unternehmen fest. Und so wie im Realsozialismus minderwertige Waren produziert wurden, um die von der Obrigkeit gesetzten numerischen Ziele zu erfüllen, so finden auch heutige Institutionen Wege, die Quoten durch minderwertige Produkte zu erfüllen: indem sie komplizierte medizinische Fälle nicht mehr übernehmen, indem sie schwerwiegende zu geringfügigen Delikten umklassieren oder Scheinkonten für Bankkunden eröffnen (vgl. Muller 2018, 61; Angermüller u. Maeße 2015, 66). Im Fall von Bildungseinrichtungen ist besonders der Umgang mit Aufnahme- und Abschlussquoten sowie mit Schulleistungstudien wie PISA von diesen Mechanismen betroffen. Leistungsnachweise spielen hierbei naturgemäß eine zentrale Rolle.

Man könnte nun einwenden, dass der geschilderte Umgang mit Leistungen Hochschulabsolvierende somit gut auf das Leben in der Numerokratie vorbereitet.²² Aber sich damit zufrieden zu geben, wäre nicht nur zynisch, sondern als Angehörige von Hochschulen würden wir uns aus unserer fachlichen wie gesellschaftlichen Verantwortung stehlen (siehe zu den verschiedenen Rollen von Hochschuldozierenden Kapitel 3.3.1). Deshalb geht dieses Buch von folgenden Prämissen aus:

- Zum einen kann Numerokratiebewältigung nicht das Ziel sein, wenn wir unseren Studierenden tatsächlich fachliches Wissen und Können beibringen wollen (vgl. Kap. 3.2 zur pädagogischen Funktion von Leistungsnachweisen).

21 Es entbehrt auch nicht einer gewissen Ironie, dass die Argumentation in diesem Kapitel durch viel Zahlenmaterial gestützt wird. Andererseits ist der Glaube an Zahlen vermutlich am ehesten zu überwinden, indem die absurden Auswirkungen von Zahlenfixierung anhand von (evidenzbasierten) Zahlen vor Augen geführt werden.

22 So auch die – nicht positiv gemeinte – Einschätzung von Sacher (2014, 24): „In gewisser Weise ist die Notenbürokratie ein heimlicher Lehrplan zur Einübung in die bürokratische Gesellschaft.“

- Und zum anderen können wir an den Hochschulen durch eine sinnvolle Praxis der Leistungsbeurteilung einen wertvollen gesellschaftlichen Beitrag über die fachliche Befähigung hinaus leisten: Indem wir die Absolvierenden nicht mit blindem Zahlenglauben – oder als zynische Systemprofiteure – in die Gesellschaft entlassen, sondern sie auch im Umgang mit ihren Lernleistungen die Wichtigkeit begründeter Urteile erfahren lassen. Denn wie Muller (2018, 176, Übers. tz) festhält, sind Zahlen keine Alternative zum Urteilsvermögen, sondern sie erfordern es. Dieses verlangt „Urteile darüber, ob gemessen werden soll, was gemessen werden soll, wie die Bedeutung dessen, was gemessen wurde, bewertet werden soll, ob die Ergebnisse mit Belohnungen und Strafen verbunden werden sollen und wem die Messungen zur Verfügung gestellt werden sollen.“

Die in diesem Buch dargebotenen Grundlagen und Gestaltungsanregungen für den Umgang mit Leistungsnachweisen einschließlich der Rückmeldung und Bewertung von Leistungen tragen diesen Prämissen Rechnung. Dabei wird eine Verschiebung von der bilanzierenden Bewertung zur förderorientierten Leistungsrückmeldung angestrebt. Das ist nicht nur didaktisch, sondern auch gesellschaftlich relevant: Einmal wäre es im Interesse der Gesellschaft als Ganzes, die Leistungsnachweiskultur so zu optimieren, dass möglichst viele Menschen möglichst viel lernen können. Dies gilt zumindest unter der Annahme, dass Wissen und Erkenntnisgewinn die Lebensumstände sowohl der einzelnen Menschen wie einer Gesellschaft verbessern – wirtschaftlich und auch soziokulturell. Andererseits dürfte sich ein lerngerechterer Umgang mit Leistungsnachweisen an Hochschulen – und im Bildungsweisen insgesamt – auch positiv auf die Kommunikation über und die Beurteilung von Leistung in Wirtschaft und Gesellschaft auswirken.

2.5 Soziale Konstruktion von Lernleistung und ihrer Beurteilung

Bisher haben wir uns noch nicht direkt mit der Frage nach dem Lernen befasst. Wir sind der Entstehung des Leistungsbegriffs, der gesellschaftlichen Vorstellung von Meritokratie, der Wirkung von Noten und der Funktion von Zahlen und Rangordnungen in der Gesellschaft auf den Grund gegangen. Dabei haben wir immer wieder Aspekte des Lernens gestreift: Wie hängen Wissenserwerb und gesellschaftliche Position zusammen? Wie hängen Noten und Lernleistung zusammen? Wie wird die Leistung von Individuen und Institutionen mit Zahlen fassbar gemacht? Nicht gefragt haben wir hingegen, was denn eigentlich eine Lernleistung ausmacht und was ihr zugrunde liegt. Das holen wir nun nach.

Dabei ist zunächst zu unterscheiden zwischen der Lernleistung und ihrer Beurteilung, die wir als Leistungsbeurteilung bezeichnen. Beide sind in hohem Maße abhängig von den sozialen Situationen, in denen sie hervorgerufen werden. Wir schauen uns in Kapitel 2.5.1 die Leistungsbeurteilung an und in Kapitel 2.5.2 die Lernleistung.

2.5.1 Soziale Konstruktion von Leistungsbeurteilung

Die bisherigen Ausführungen haben bereits deutlich gemacht, dass die Beurteilung von Leistungen sozial konstruiert ist. So sind zum Beispiel Noten nicht Eigenschaften von Menschen, sondern haben den Status von Etikettierungen, die den gezeigten Leistungen von Lernenden durch Lehrpersonen oder Expertinnen zugeschrieben werden – gleichgültig, wie differenziert die Bewertungskriterien formuliert sind.

Jede einzelne Beurteilungshandlung stellt eine solche soziale Konstruktion dar, bei der einem gezeigten Verhalten oder Verhaltensbündel ein Leistungsniveau zugeschrieben wird. Überdies nehmen Lehrende, die über längere Zeit mit Lernenden arbeiten, oft auch zeitlich überdauernde Zuschreibungen vor wie „du bist eine Zweierschülerin“ (vgl. Breidenstein u. Thompson 2014). Damit wird einer Person ein stabiles Leistungsvermögen zugeschrieben – das es aber nicht gibt, wie Lernpsychologie und Hirnforschung zeigen (siehe Kapitel 4).

Schon einzelne Zeugnisnoten können die künftig gezeigte Leistung deutlich beeinflussen, wie in Kap. 2.3 dargelegt wurde – stabile Zuschreibungen gewiss noch mehr. Zudem gilt die von der tatsächlich jeweils gezeigten Leistung unabhängige Zuschreibung eines bestimmten Leistungsniveaus an eine bestimmte Person als eine der verbreitetsten kognitiven Verzerrungen (Bestätigungstendenz, vgl. Kap. 6.4).

Sogar die fachlichen (und ggf. überfachlichen) Kriterien, an denen sich Leistungsbeurteilungen orientieren, sind durch eine wissenschaftliche Gemeinschaft sozial konstruiert. Auch die Lehrpersonen oder Experten, welche die Beurteilungen vornehmen, sind deshalb in der Festlegung von Beurteilungskriterien nicht „frei“, sondern Teil einer fachlichen Gemeinschaft: „Indem aber Leistungsmaßstäbe soziale Maßstäbe sind, sind sie dem Individuum in der Regel nicht ‚rein individuell‘ verfügbar“ (Schlie 1988, 63). Beurteilungskriterien spiegeln deshalb in der Regel den Stand des fachlichen Diskurses. Dies gilt auch dann, wenn es innerhalb eines Diskurses widerstreitende Positionen oder Lücken gibt: Auch solche werden in Beurteilungskriterien sichtbar.

All diese Aspekte zeigen auf, dass die Beurteilung einer Leistung nicht nur ein sozialer Vorgang ist, sondern auch auf sozialen Konventionen beruht – im Fall von Hochschulen auf den Konventionen der jeweiligen fachlichen

Diskurse. Diese sozialen Konventionen kommen bei der Beurteilung umso mehr zum Tragen, als sowohl die Lernleistung (siehe Kap. 2.5.2) wie auch das Wissen und Können einer Person latente Konstrukte sind: Sie sind keiner direkten Beobachtung zugänglich. Beobachten können wir nur die Performanz, also wie eine Person bestimmte Aufgabenstellungen bearbeitet, wie sie fachliche Probleme löst (vgl. Zlatkin-Troitschanskaia u. Seidel 2011, 220; eine nähere Erläuterung des Unterschieds zwischen Kompetenz und Performanz erfolgt in Kap. 6.1). Dieses beobachtete performative Verhalten schließlich messen wir an den fachlichen Kriterien und kommen aufgrund des Abgleichs von Performanz und Kriterien zu einer bestimmten Beurteilung des gezeigten Verhaltens. Im Falle von summativen Beurteilungen machen wir in der Regel Aussagen darüber, wie umfassend jemand ein bestimmtes Lernziel erreicht hat (kriteriale Norm, es sind aber auch andere Normorientierungen möglich, siehe die verschiedenen Bezugsnormen in Kap. 6.2).

2.5.2 Lernleistung als soziales Konstrukt

Die Beurteilung ist also anerkanntermaßen sozial, aber noch immer wird Leistung zumindest im Bildungsbereich einzelnen Personen zugeschrieben. Leistung wird somit als individuelle Kategorie behandelt. Das ist nicht unbedingt falsch, aber es greift zu kurz:

Zweifelsohne kann man messen, was jemand als Einzelperson weiß oder kann. Man gestaltet dabei das Setting so, dass die zu prüfende Person allein (ggf. auch ohne Hilfsmittel) handelt. Anschließend wird das Resultat beurteilt und ausschließlich als Folge der Handlungen dieser einzelnen Person betrachtet.²³ Genau dies ist auch im Hochschulbereich noch immer der unmarkierte Normalfall. Bereits Gruppenleistungsnachweise bedürfen in der Regel – auch gegenüber Studierenden – einer besonderen Rechtfertigung. In Anlehnung an Ricken (2018, 52) ließe sich der soziale Vorgang, der zu einer solchen Kon-

23 Eine strikte Zuschreibung von Leistungen an einzelne ist nur mit bestimmten Formaten wie z. B. mündlichen Prüfungen oder schriftlichen Klausuren möglich; Portfolios, Laborjournale etc. eignen sich offensichtlich nicht zu dieser Abschließung (verschiedene Formate von Leistungsnachweisen und ihre Vor- und Nachteile werden in Kap. 7.6 und 7.7 erläutert). Einen interessanten Grenzfall stellen längere schriftliche Arbeiten dar – hier ist allen klar, dass kaum ein Student diese gänzlich allein verfasst, aber das stört im Gegensatz zu Klausuren in aller Regel nicht. Das Zuschreibungsproblem taucht stattdessen in Form der Angst vor Plagiaten auf, wobei in der Regel seltsam unterbelichtet bleibt, dass wissenschaftliche Texte stets auch Intertexte sind und es einen Übergangsbereich zwischen Intertextualität und Plagiat gibt (vgl. Blum 2009; Blühdorn 2006; Jakobs 1993).

struktion einer individuellen hochschulischen Lernleistung führt, in folgenden Ablauf gliedern:

1. Lernleistung setzt eine Gelegenheit zum Lernen voraus.²⁴ Diese kann informell aus Eigeninitiative geschaffen werden, etwa durch Lesen eines Buches, Durcharbeiten von Lernvideos oder Partizipation in Aktivitäten einer (fachlichen) Gemeinschaft. Im Hochschulkontext ist diese Gelegenheit aber meist ziemlich strukturiert, vom Präsenzunterricht bis zu Phasen selbstorganisierten Lernens. Hochschulische Lerngelegenheiten sind sozial gestaltet – von einer oder mehreren Lehrenden für die Studierenden, basierend auf einem ganzen Fachdiskurs, der selbst ein soziales Konstrukt ist.
2. Um zeigen zu können, was sie gelernt haben, brauchen Studierende passende Gelegenheiten. Diese Gelegenheiten müssen sozial arrangiert werden.
3. Die Studierenden müssen im Rahmen dieser Gelegenheiten eine Handlung oder Handlungsfolge vollziehen: z. B. Verfassen eines schriftlichen Produkts, mündliche Aussagen in einem Gespräch, künstlerische Artefakte oder auch körperliche Vorführungen etwa in Sport und Kunst. Je nachdem wird die Handlung an sich oder ihr Produkt (z. B. schriftliches Produkt, künstlerisches Artefakt) beurteilt.
4. Diese Handlungen oder Produkte werden einzelnen Personen – allenfalls auch einer Gruppe – zugeschrieben, die als Urheber betrachtet werden.
5. Produkte oder Handlungen können dann anhand von Kriterien beurteilt werden, wobei oft auch ein sozialer Vergleich mit anderen Angehörigen einer Studierendenkohorte erfolgt.
6. Erst die abschließende Bewertung, meist in Form einer Note oder eines Prädikates wie „bestanden/nicht bestanden“, macht die von den Studierenden vollzogenen Handlungen zu einer individuellen *Lernleistung*.²⁵

Denken wir an einen typischen Arbeitskontext: In einem heutigen Dienstleistungsunternehmen kann etwa jemand Produktmanagerin sein und hat ein gro-

24 Streng genommen beginnt dieser soziale Konstruktionsprozess noch viel früher: Letztlich fließt stets die gesamte Lernbiografie einer Person ein und bestimmt in großem Maße mit, auf welche Weise und in welchem Ausmaß jemand eine Lerngelegenheit kognitiv, emotional und motivational ergreifen kann. Sozioökonomischer Status, Beziehungen zu Eltern, Lehrpersonen und Peers etc. – all diese sozialen Faktoren spielen dabei eine große Rolle.

25 Wie in Kap. 2.3 gezeigt, hat die Bewertung insbesondere in Form einer Note wiederum einen prägenden Charakter für das weitere Lerngeschehen, so dass sich hier der Kreis schließt: Die Bewertung bestimmt mit, wie und wie nachhaltig ein Student in der nächsten Iteration des geschilderten Ablaufs lernen wird.

Bes fachliches und überfachliches Wissen und Können dazu. Dennoch kann sie das Produkt nicht allein managen: Sie verfügt weder über alles Wissen und Können, um das Produkt von Grund auf herzustellen, noch um es zu vermarkten, noch um die Finanz- und Steuerabläufe rund um den Verkauf zu bewältigen. Ähnliches gilt beispielsweise auch in Anwaltskanzleien oder naturwissenschaftlichen Forschungsgruppen, deren Arbeit so komplex ist, dass sie nicht mehr von einer einzelnen Person hervorgebracht werden könnte (vgl. Mohammed u. Dumville 2001, 95). In wissenschaftlichen Publikationen wird die immer stärkere Verteilung von Wissen und Können auf verschiedene Personen anhand der steigenden Zahl der Autorinnen sichtbar.²⁶

In all den genannten Arbeitskonstellationen ist es kaum möglich, die Leistung einer einzelnen Person isoliert zu betrachten. Dies ist keineswegs eine neue Erkenntnis: Schon 1970 stellte Offe fest, „dass die fortgeschrittenen Formen industrieller Arbeit die Kategorie kompetitiv offenbaren, individuellen Leistungsvermögens irrelevant machen“ (Offe 1970, 166). Im Kontext schulischer und hochschulischer Leistungsbeurteilung bleibt dieser gesellschaftliche Zusammenhang bisher freilich weitgehend ausgeblendet. Insbesondere durch das Erteilen von Noten und deren selektive Wirkung stellen die Schulen das Lernen und damit den Lernvorgang eigentlich auf den Kopf: Ziel ist für die benoteten Schülerinnen und Schüler oft nicht mehr das Lernen bzw. die Beschäftigung mit relevanten fachlichen Fragen, sondern die Note – und das Lernen ist nur noch Mittel zum Zweck. Ein Bildungswesen, das Noten ins Zentrum stellt, verdrängt das Lernen aus diesem. Die Leistungsfixierung führt so ins Paradoxe und reduziert die Leistung sowohl des Bildungswesens als Ganzem als auch der meisten Lernenden (Schüler:innen, Studierende, Auszubildende, mitunter sogar Weiterbildungsteilnehmende). Dabei wird in der Regel übersehen, „dass individuelle Leistung nicht per se existiert, sondern immer sozial konstruiert werden muss, wozu in einem ersten Schritt Arbeitsergebnisse künstlich isoliert werden, um sie in einem zweiten Schritt einem einzelnen Menschen als ‚seine‘ Leistung zuzuschreiben“ (Verheyen 2014, 66).

Auch in schulischen Kontexten sind Leistungen nicht einfach etwas, was Schülerinnen allein „tun“. Vielmehr gibt es einen „gesamten sozialen Prozess der Hervorbringung von Leistungen, an dem auch Lehrkräfte beteiligt sind, indem sie Aufgaben stellen, Beurteilungen vornehmen, Zensuren vergeben usw., kurz gesagt: indem sie Leistungen individuell zurechnen“ (Nicht u. Müller 2017, 72).

26 In naturwissenschaftlichen Studien werden teilweise hunderte von Autorinnen und Autoren genannt, etwa in den GWAS-Studien (vgl. Bulik-Sullivan u. Sullivan 2012). Unklar ist, ob tatsächlich die Interdependenzen verteilten Wissens immer größer werden oder ob dieser Umstand lediglich verstärkt in den Blick rückt.

Definition

Bei der Leistungsbeurteilung erschließen Lehrende aus wahrnehmbaren Handlungen oder Produkten einer Person deren Lernstand und die zu dessen Erreichung erfolgte Lernleistung. Die beurteilten Handlungen und Produkte bringen die Studierenden in einem Leistungsnachweis-Setting hervor, das als soziale Situation durch Lehrende und Institutionen vorgegeben ist. *Lernleistung* ist deshalb ein mittelbar anhand einer unter spezifischen institutionellen und sozialen Bedingungen hervorgebrachten Handlung erschlossenes Konstrukt. Sie existiert nicht objektiv, sondern lediglich als Produkt von Kommunikations- und Interpretationsvorgängen.

Deshalb macht es Sinn, diese Umstände auch beim Lernen an Hochschulen zu berücksichtigen und Leistungsnachweise und Leistungsbeurteilung entsprechend zu gestalten. Es soll hier nicht in Abrede gestellt werden, dass Lernvorgänge in einzelnen Lernenden stattfinden. Selbstverständlich ist Lernen in der Gemeinschaft nur denkbar, indem die Individuen etwas lernen, aus denen sie sich zusammensetzt. Aber angesichts der großen Interdependenzen zwischen verschiedenen Menschen sowohl beim Erwerb neuen Wissens und Könnens als auch bei dessen Abrufung und Umsetzung erscheint es sinnvoll, Lernleistungen als soziales Konstrukt zu betrachten. Dies bedeutet, sich den oben aufgeführten Ablauf, der zur sozialen Konstruktion von Leistungen führt, als ganzen bewusst zu machen und bewusst zu halten. Das führt dazu, dass wir nicht mehr so tun können, als seien Leistungen einzig auf das Handeln einer einzelnen Studentin zurückzuführen. Vielmehr müssen der vorangegangene Unterricht, die Lernmaterialien, der Austausch unter den Studierenden, die Gestaltung der Leistungsnachweise inklusive Randbedingungen (wie erlaubte Hilfsmittel oder Kooperation), die Wirkung der Beurteilungskriterien und die eigene Rolle als Beurteilungsinstanz mitbedacht werden.

Sofern es der institutionelle Kontext vorgibt, wird man weiterhin Noten und Prädikate erteilen müssen. Aber die Erteilung dieser Bewertungen und das Formulieren von Rückmeldungen an Studierende dürfte im Bewusstsein der Konstruiertheit der Leistungsbeurteilung doch anders ausfallen. Unter anderem müsste dadurch auch klar werden, dass sowohl die Hochschule als Institution als auch die einzelnen Lehrenden einen erheblichen Einfluss auf die zu beurteilenden Leistungen haben. Streng genommen stellt die Beurteilung von selbst unterrichteten Studierenden einen ähnlichen Fall dar wie ein Unternehmen, das seine eigene Leistung evaluiert. Das ist nicht verboten und zur Optimierung der eigenen Tätigkeiten sicherlich sinnvoll. Aber die meisten von uns dürften es angebracht finden, dass z. B. die Einhaltung arbeitsrechtlicher Vorgaben in Betrieben oder hygienischer Standards in Gesundheits- oder Gastronomiebetrieben (auch) von unabhängigen Instanzen überprüft wird. Der Normalfall der Beurteilung von studentischen Lernleistungen

durch Lehrende erscheint so betrachtet nicht weniger rechtfertigungsbedürftig als die Einschätzung des Lernstandes durch Mitstudierende (Peer-Assessment oder Peer-Feedback) oder studentische Selbsteinschätzungen (siehe Kap. 8.4 zu den verschiedenen Akteuren, die Beurteilungen vornehmen können).

Meine Hoffnung ist deshalb, dass sich durch das Verständnis von Leistung als sozialem Konstrukt eine erkenntnistheoretische Bescheidenheit einstellt, die es erlaubt, den Blick gemeinsam mit den Studierenden stärker auf das eigentliche Lerngeschehen und die Lerninhalte zu richten. Die Aspekte der Bewertung, der Selektion und Qualifikation dürfen demgegenüber ruhig etwas stärker in den Hintergrund des Bewusstseins rücken. Im nächsten Kapitel schauen wir uns diese Zusammenhänge unter anderer Perspektive noch einmal an, bevor wir sie einen Moment zur Seite legen, um dann ab Kapitel 6 wieder auf sie zurückzukommen.

3 Funktionen von Leistungsbeurteilung

Eine allgemein gültige Definition, was einen Leistungsnachweis oder eine Prüfung ausmacht, gibt es nicht. Dies gilt jenseits aller konkreten Realisierungsformen, denn was einen Leistungsnachweis konstituiert, ist stets abhängig vom Kontext. Welche der beteiligten Personen verfolgen im Rahmen der Überprüfung einer Leistung welche Ziele, und welche unterschiedlichen Perspektiven haben diese Personen? Wie ist die Beurteilungssituation in die Institution Hochschule eingebettet und welche Ziele verfolgt sie damit? Wenn Sie nur schon versuchen, für einen konkreten Leistungsnachweis diese Fragen zu beantworten, erkennen Sie sogleich die Komplexität dieser Kontextfaktoren. Aus dem gleichen Grund ist auch Lernleistung keine objektiv messbare Größe, sondern kann nur durch Kommunikations- und Interpretationsvorgänge in einem spezifischen Kontext mittelbar erschlossen werden (siehe die Definition in Kap. 2.5.2).

Deshalb ist es einerseits wichtig, Leistungsnachweise jeweils so gut als möglich an die aktuellen Bedingungen anzupassen. Diese reichen von curricularen Lernzielen bis hin zu Anforderungen des zukünftigen Berufsfeldes der Studierenden. Es sind also stets Interessen und Bedürfnisse verschiedener Instanzen und Personengruppen zu bedenken, ähnlich wie bei einer Stakeholderanalyse im betrieblichen Kontext. Andererseits sollte unabhängig von diesen spezifischen Kontextbedingungen das Hauptziel sein, die Durchführung eines Leistungsnachweises so zu gestalten, dass er Ihnen den Lernstand der Studierenden zu ermitteln erlaubt. Dazu gehört in aller Regel auch, den Studierenden Hinweise für die Gestaltung ihres weiteren Lernens zu ermöglichen, wobei dieses Feedback nicht nur durch Lehrende, sondern auch durch Peers oder als Selbsteinschätzung erfolgen kann (siehe Kap. 8).

Diese verschiedenen impliziten und expliziten Zweckbestimmungen von Leistungsnachweisen und ihren Beurteilungen lassen sich als unterschiedliche Funktionen beschreiben, die zu zwei Bündeln zusammengefasst werden: den gesellschaftlichen Funktionen von Leistungsbeurteilung einerseits und den didaktisch-lernbezogenen Funktionen andererseits (siehe Kap. 3.1 und 3.2).

Aus den unterschiedlichen Funktionen von Leistungsbeurteilung ergeben sich einige Dilemmata im Umgang mit Beurteilungssituationen. Diese hängen zusammen mit unterschiedlichen Rollen, die Lehrpersonen im Kontext von Leistungsbeurteilung einnehmen können. Zudem liegen hinter diesen Dilemmata und Rollenkonflikten strukturelle Spannungsfelder institutionalisierter Bildung, die sich auf Beurteilungshandlungen auswirken. Dieser Themenkomplex ist in Kapitel 3.3 dargelegt.

3.1 Gesellschaftliche Funktionen von Leistungsbeurteilung - Berechtigung

Indem sie Studiengänge anbieten, übernehmen Hochschulen in verschiedener Hinsicht gesellschaftliche Verantwortung. Sie sind ein erheblicher Teil der Lebenswelt ihrer Studierenden, tragen zur Heranbildung des wissenschaftlichen Nachwuchses bei und gewährleisten für einen substanziellen Teil der Bevölkerung die weiterführende Bildung als Grundlage für die Berufsausübung²⁷ (vgl. Huber 1995, 127). Sowohl das Wissenschaftssystem als auch die Gesellschaft insgesamt möchten sich darauf verlassen können, dass die Studienabgängerinnen und -abgänger in den Fachgebieten, in denen sie ein Studium erfolgreich abgeschlossen haben, wirklich kompetent sind. Diese eher fachorientierten gesellschaftlichen Anliegen vermischen sich dabei teilweise mit grundsätzlicheren Strukturierungstendenzen, die sich mit Begriffen wie Numerokratie oder Kredentialismus fassen lassen (siehe Kap. 2.4). So können die folgenden gesellschaftlichen Funktionen von Leistungsnachweisen ausgemacht werden, die sich allerdings teilweise überlagern und keine trennscharfen Kategorien darstellen (vgl. Sacher 2014, 22–31; Prahl 1995, 442–45; Ingenkamp u. Lissmann 2008, 33–38):

- **Selektion:** Die Hochschule kontrolliert anhand von Leistungsnachweisen, dass ihre Abgängerinnen und Abgänger ein bestimmtes Niveau an Wissen und Können erreichen. Die Abschlüsse und akademischen Grade sollen verlässliche Erwartungen an die Fähigkeiten der Abgehenden ermöglichen. Diese Qualitätskontrolle ist gewiss sinnvoll – aber mit den bereits erläuterten Problemen verbunden: Die zur Selektion verwendeten Noten erlauben kaum sichere Rückschlüsse auf das tatsächliche Wissen und Können und sind außerdem systematisch schichtspezifisch verzerrt (siehe Kap. 2.2 und 2.3).
- **Legitimation und Kontrolle der Hochschule:** Die Selektion hat über die reine Auslese hinaus auch die Funktion, die Gesellschaft davon zu überzeugen, dass die Hochschule qualitativ ansprechende Arbeit leistet und hohe Ansprüche an ihre Absolvierenden stellt. Auf tieferer Ebene legitimieren sich oft auch einzelne Studiengänge, Institute oder Lehrende über die Selektionsfunktion. Tun sie es anhand von Noten, ist dies problematisch.

27 Der Berufsbezug verschiedener Studienfächer ist traditionell sehr unterschiedlich, kann sich im Lauf der Zeit aber auch wandeln. Medizin ist wohl der Inbegriff eines Studiums, das direkt auf eine Berufsbefähigung hinführt, Betriebswirtschaft, Recht oder Ingenieurwissenschaften haben ebenfalls klare Professionsbezüge, während die geisteswissenschaftlichen Fächer am anderen Ende dieser Skala rangieren und ohne klare Berufsbilder allgemein wissenschaftlich ausbilden (vgl. Hessler u. Oechsle 2012, 115).

tisch, da es die in Kapitel 2.3 herausgearbeitete Tendenz zum Referenzgruppenfehler bei der Notenerteilung verstärkt (siehe Kap. 5.1.3 zur Fehlannahme einer Normalverteilung): Indem sie dafür sorgen, dass der Durchschnitt der erteilten Noten nicht zu hoch ausfällt, zeigen die Studiengänge oder Lehrenden, dass sie von den Studierenden Leistung einfordern. Und indem die Noten nicht zu schlecht ausfallen, wird gezeigt, dass die Studierenden etwas gelernt haben und ein ansprechendes Niveau erreichen.²⁸ Diese beiden Tendenzen führen dazu, dass Notenbewertungen unabhängig von den effektiven Lernleistungen der Studierenden in jeder Gruppe relativ ähnlich ausfallen und eigentlich primär die Rangreihe innerhalb der Gruppe abbilden.

- Leistungssozialisation und Disziplinierung der Studierenden: Das regelmäßige Bewerten ihrer Leistungen führt die Studierenden auch an die in unserer Gesellschaft herrschende Leistungskultur heran, wie sie in Kapitel 2 ausführlich beschrieben wurde. Wie dort besprochen ist dies insoweit problematisch, als dadurch ein übermäßiger Glaube an die Messbarkeit menschlichen Verhaltens befördert wird und im Gegenzug die nicht durch Leistung begründbaren Aspekte von Bewertungen verdeckt werden. Zudem gehören Noten und Zeugnisse zu extrinsischen Anreizen und unterminieren damit die lernwirksamere intrinsische Motivation (siehe Kap. 4.5). Soweit die Leistungsbeurteilung die Studierenden bei der Internalisierung von Qualitätsansprüchen unterstützt, ist die Leistungssozialisation hingegen durchaus sinnvoll. Dies dürfte allerdings eher zum Bündel der didaktisch-lernbezogenen Funktionen gehören.
- Information über die aktuellen und Prognose zu den künftigen Leistungen der Studierenden: Noten und andere Bewertungen – insbesondere Studienabschlüsse – informieren nicht nur über den aktuellen Leistungsstand der Studierenden. Ihnen wird oft auch unterstellt, dass sie Annahmen über die künftigen Leistungen zu treffen erlauben. Insbesondere diese prognostische Gültigkeit ist allerdings kaum gegeben, wie in Kapitel 2.3 herausgearbeitet wurde.

Insgesamt kann man aus diesen Schilderungen schließen, dass die gesellschaftlichen Funktionen von Leistungsbeurteilung in erster Linie der Berechtigung dienen (siehe Definition). Sie berechtigen etwa zum Besuch weiterer Module oder zum Erhalt eines akademischen Grades. Akademische Grade und konkrete Studienabschlüsse wiederum berechtigen zur Ausübung bestimmter Berufe oder anderer gesellschaftlicher Funktionen sowie zum Besuch weiterführender Ausbildungen, wie zum Beispiel für medizinische und

28 Dementsprechend gibt es in vielen Institutionen offizielle oder inoffizielle Regelungen, in welcher Bandbreite der Notendurchschnitt bei bestimmten Prüfungen oder Zeugnissen liegen darf.

juristische Spezialisierungen in den deutschsprachigen Ländern weitgehend üblich. Insbesondere unter Verweis auf den in Kapitel 2.4 erläuterten Kredentialismus gilt es für diese berechtigenden Aspekte festzuhalten: „Beurteilungen dienen in dieser Relevanzform dazu, Lebens- und Berufschancen zuzuweisen“ (Thomann 2019, 250).

Definition

Leistungsbeurteilungen, die primär der Qualifikation und Selektion der Lernenden dienen, bezeichnet man als *summative Beurteilungen* (Zimmermann 2018, 52). Das Adjektiv „summativ“ rührt daher, dass solche Beurteilungen in der Regel am Schluss einer Lerneinheit erfolgen und als Abschluss eines Lernprozesses betrachtet werden. Diese Art der Beurteilung kann auch als Bewertung bezeichnet werden, da Lernleistungen durch sie ein bestimmter Wert zugeschrieben wird.

3.2 Didaktisch-lernbezogene Funktionen von Leistungsbeurteilung - Fördern

Wie notwendig die Berechtigungsaspekte aus gesellschaftlicher Sicht auch sein mögen – für die Didaktik stehen primär andere Funktionen von Leistungsbeurteilung im Vordergrund, nämlich die Ermöglichung eines Lehr-Lern-Prozesses und seine Steuerung. Lernen soll ermöglicht und Lernprozesse sollen optimiert werden. Wesentliche Bezugspunkte für die Beurteilung von Lernleistungen sind in diesem Kontext das Wissen über das menschliche Lernen (siehe Kap. 4) und die Wirkung von Feedback zu erbrachten Lernleistungen (siehe Kap. 8). Zu beachten sind dabei auch die Möglichkeiten, verschiedene Akteursgruppen in die Beurteilungs- und Feedbackprozesse einzubinden: Das sind neben Lehrenden auch externe Expertinnen, Peers und die Leistungserbringenden selbst (siehe Kap. 8.4).

Unter einer didaktisch-lernbezogenen Perspektive lassen sich die folgenden Funktionen ausmachen, die sich wie die gesellschaftlichen Funktionen teilweise überlagern und keine trennscharfen Kategorien darstellen (vgl. Sacher 2014, 22–31; Prahl 1995, 442–45; Ingenkamp u. Lissmann 2008, 20f. u. 33–38; Zimmermann, Tobias 2018, 52–54):

- **Lehr- und Lerndiagnose:** Die Diagnose des Lernstandes mit dem Ziel, das Lernen zu verbessern, ist aus Sicht von Pädagogik und Didaktik grundlegender Bestandteil jeder geplanten Lehrtätigkeit. Leistungsbeurteilung zielt in diesem Sinne nicht auf eine objektivierende Vermessung oder

Rangierung von Lernleistungen. Vielmehr geht es darum, anhand von studentischen Leistungsnachweisen den Lernstand zu eruieren und diesen den Studierenden zurückzumelden. Dabei sind Fragen zu stellen wie: Was weiß und kann der Student schon? Wo kann er sich noch verbessern? Welche Lernhandlungen könnten hilfreich sein, um eine entsprechende Verbesserung zu erzielen? Antworten auf solche Fragen – in Form einer Leistungsrückmeldung bzw. eines Feedbacks – sind einerseits für die Studierenden als Lernende zentral, weil sie mit ihrer Hilfe ihr Lernverhalten besser steuern können (allerdings bewirkt längst nicht jede erhaltene Leistungsrückmeldung ein solches „Acting upon“, siehe Kap. 8.6). Das Erheben und Beurteilen von Lernleistungen ist aber auch direkt für das Handeln der Lehrenden relevant. Denn die Einschätzung des Wissens- und Könnensstandes einer Studierendengruppe ermöglicht es, den weiteren Unterrichtsverlauf zu planen und das Lehrhandeln an den Lernstand der Studierenden anzupassen. Leistungsbeurteilungen mit dem Ziel der Lernoptimierung sollten laufend durchgeführt werden und nicht nur am Ende einer Lehrveranstaltung, da die daraus gewonnenen Erkenntnisse sonst kaum zur Anpassung des Lehr- oder Lernverhaltens genutzt werden können.²⁹

- Differenzierte Rückmeldung: Damit die Lernenden die Leistungsbeurteilung tatsächlich für die Steuerung ihres weiteren Lernens nutzen können, benötigen sie eine hinreichend differenzierte Rückmeldung. Inhaltsleere Vermessungen wie eine Punktzahl, Note oder das Prädikat „bestanden/nicht bestanden“ reichen dazu nicht aus. Vielmehr müssen die Studierenden verstehen können, in welchen Bereichen sie welchen Lernstand erreicht haben, wo sie noch welches Entwicklungspotenzial haben und mit welchen Lern- und Übungsstrategien sie dieses am besten ausschöpfen können.
- Motivationale Funktion: Gerade die inhaltlichen Rückmeldungen dienen auch zur Bestätigung erfolgreicher Lernschritte und sollen die Studierenden motivieren, sich weiter mit einem Thema zu beschäftigen und/oder neue Herausforderungen in Angriff zu nehmen. Damit die Studierenden motiviert bleiben, sich mit den angestrebten Lernzielen auseinanderzusetzen, ist auch von entscheidender Bedeutung, dass Leistungsnachweise wirklich prüfen, was die Lernenden lernen sollen (das ist weniger trivial,

29 Dies ist ein gutes Beispiel für die Überlagerung verschiedener Funktionen von Beurteilungen, denn wie Sacher feststellt, ist bei solchen Verhaltensanpassungen immer auch ein prognostischer Anteil enthalten: „Wenn im Anschluss an eine Diagnose des Lernstandes unterrichtliche Maßnahmen ergriffen werden, dann offensichtlich in der prognostischen Annahme, dass diese zu Lernerfolgen führen“ (Sacher 2014, 29). Diese Art von Prognose steht allerdings stärker im Dienst der Lernförderung als der gesellschaftlichen Berechtigung.

als es scheint, siehe die Erläuterungen zum Gütekriterium der Validität in Kap. 6.3.1 und 6.3.2). Andernfalls unterlaufen die Leistungsnachweise die eigentlichen Lernziele und damit auch die Lernmotivation, was in der Realität leider oft geschieht.

- **Anregung zur Selbststeuerung:** Indem sie regelmäßig Leistungen erbringen und Rückmeldungen zu ihrem Leistungsstand erhalten, sollen die Studierenden nicht nur die Anforderungen an das Wissen und Können in ihrem Fachgebiet kennen lernen. Darüber hinaus sollen sie auch lernen, diese Kriterien auf ihren eigenen Lernfortschritt anzuwenden und damit zunehmend Verantwortung für ihr eigenes Lernen zu übernehmen. Leistungsnachweise haben insofern – zumindest teilweise – Zwangscharakter, sollen aber zur Eigenständigkeit führen. Dieses Paradox ist eine zentrale Herausforderung, wenn Leistungsbeurteilung das Lernen fördern soll, und wird in Kapitel 3.3.2 noch näher beleuchtet.

Definition

Leistungsbeurteilungen dienen aus didaktischer Perspektive der laufenden Entwicklung und Verbesserung des Lehrens und Lernens. Deshalb bezeichnet man Beurteilungen, die primär der Lernoptimierung dienen, als *formative Beurteilungen* (Zimmermann 2018, 54). Sie erfolgen bereits während einer Lerneinheit und geben den Lernenden in der Regel Feedback zu ihrem Lernstand.

Aus den geschilderten Funktionen geht hervor, dass die didaktisch-lernbezogenen Funktionen von Leistungsbeurteilung in erster Linie der laufenden Steuerung und Optimierung des Lehrens und Lernens dienen (siehe Definition). Zentraler Bestandteil solcher formativen Beurteilung ist, die Studierenden mit Angaben zu ihrem Lernstand darin zu unterstützen, ihr weiteres Lernen zu gestalten. Dies wird Leistungsrückmeldung genannt. Dazu gehört der Abgleich mit Lernzielen und eine Einschätzung, über welche Aspekte des angestrebten Wissens und Könnens die Studierenden bereits gut verfügen und wo noch Lücken bestehen, die sie schließen müssen (vgl. Ambrose et al. 2010, 139 u. 206). Wie wir in Kapitel 9.1 noch genauer betrachten werden, gibt es zwischen dieser formativen Leistungsrückmeldung und der Erteilung von summativen³⁰ Beurteilungen wie Noten oder Prädikaten starke Interferenzen: Werden Noten oder Prädikate und differenziertere formative Leis-

30 Das Begriffspaar *summativ – formativ* wurde von Bloom, Hastings u. Madaus (1971, 117) erstmals in dieser Bedeutung verwendet. Sie unterschieden zwischen „*summative evaluation*“, die der Benotung und Zertifizierung von Lernenden dient, und „*formative evaluation*“, die der Optimierung von Lehr- und Lernprozessen dient.

tungsrückmeldungen gleichzeitig erteilt, schenken Studierende in aller Regel der formativen Rückmeldung keine Beachtung mehr. Dies ist – neben der logischen Folgerung, dass Lernen vor allem während des Lernprozesses und weniger nach dessen Abschluss optimiert werden kann – ein weiterer wichtiger Grund, weshalb formative Beurteilungen und Leistungsrückmeldungen während eines Lernprozesses erfolgen sollen. Leistungsbeurteilung steht in dieser Perspektive nicht am Schluss einer Lernhandlung, sondern ist ein laufender Prozess, in dem sich Lehrende und Lernende immer wieder über die Lerninhalte, die Lernfortschritte und die Lernhandlungen verständigen.

3.3 Dilemmata bei der Leistungsbeurteilung

Wie bereits angesprochen, bestehen verschiedene Gegensätze zwischen den didaktisch-lernbezogenen und den gesellschaftlichen Funktionen von Leistungsbeurteilung. Einige dieser Gegensätze sind, teilweise abhängig vom Kontext, vereinbar, einige grundsätzlich unvereinbar. Thomann (2019, 261f.) führt solche Dilemmata auf, die auch im vorliegenden Buch wiederholt aufscheinen und in Tabelle 1 in einer eigenen Fassung dargelegt werden.

In diesen Dilemmata zeigen sich unter anderem auch Rollenkonflikte, etwa zwischen jener der Beraterin und der Beurteilerin, und Antinomien, etwa zwischen den beiden gleichermaßen berechtigt scheinenden Werten des Forderns und Förderns. Diese beiden Aspekte werden im Folgenden näher beleuchtet, da sie für die Reflexion sowohl Ihrer eigenen Handlungsweisen als auch jenen von Studierenden oder anderen Hochschulangehörigen hilfreich und in einem gewissen Maß auch entlastend sein können.

Tab. 1: Dilemmata im Umgang mit Leistungsnachweisen (in Anlehnung an Thomann 2019, 262, und Zimmermann, Tobias 2018, 54)

Dilemmata im Umgang mit Leistungsnachweisen		
Selektion	⇔	Unterstützung
Summative Beurteilung	⇔	Formative Beurteilung
Fordern	⇔	Fördern
Lenkung	⇔	Begleitung
Objektivität & Messgenauigkeit	⇔	Validität
Qualifikation	⇔	Beratung

3.3.1 Rollen von Hochschuldozierenden und Rollenkonflikte

Der Rollenbegriff ist relativ anwendungsorientiert und somit geeignet, „um Lehrhandeln im sozialen Kontext der Hochschule als Bildungs- und Forschungseinrichtung zu thematisieren“ (Weil 2020, 84). Dabei entstammt der Begriff der Rolle ursprünglich dem Theater, wo er den früher in einer Schriftrolle vorgegebenen Text bezeichnete, den ein Schauspieler zu sprechen hatte. Aus soziologischer Sicht wird unter Rolle ein Bündel von Erwartungen verstanden, die an eine Person gerichtet werden. Diese Erwartungen können explizit sein, sind aber oft auch implizit. Entscheidend ist, dass sie zwar an eine einzelne Person gerichtet werden, sich aber stets auf deren soziale Position beziehen. Die Verhaltenserwartungen sind stark gesellschaftlich geprägt, allerdings nicht bis ins letzte Detail definiert. Das führt dazu, dass Rollen eine Art Bindeglied zwischen einzelnen Individuen und Organisationen darstellen, etwa zwischen Studierenden, Dozierenden und ihrer Hochschule (vgl. Thomann 2019, 29).

Rollen werden also zugewiesen, wobei man zuweisende Instanzen als Rollensender bezeichnet. Die Zuweisung kann von außen erfolgen. Wenn

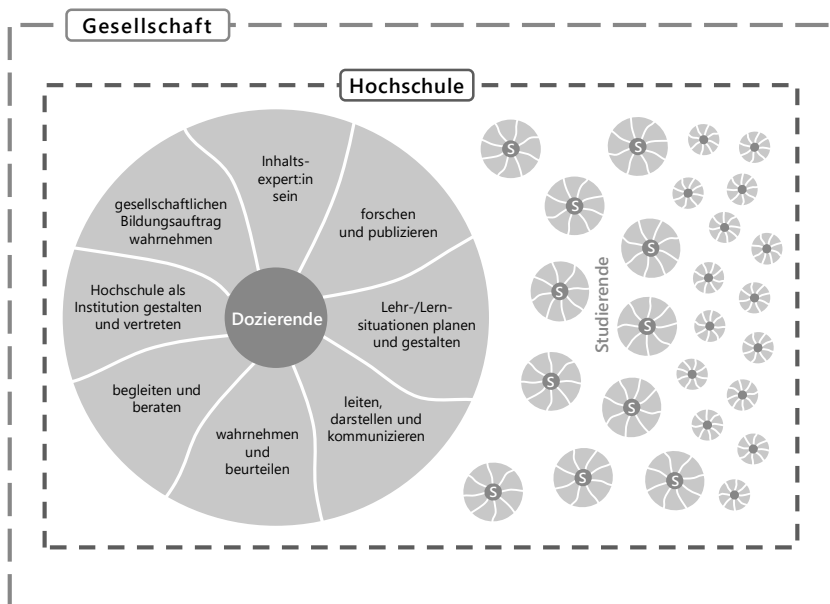


Abb. 7: Der Rollenstrauch von Dozierenden. Quelle: Eigene Abbildung, in Anlehnung an Thomann 2019, 31

etwa eine Dozentin durch hochschulische Reglemente zur Abnahme bestimmter Leistungsnachweise verpflichtet ist, wird sie zur Beurteilerin. Oder jemand kann sich eine Rolle selbst zuschreiben, beispielsweise wenn sich ein Dozent als Berater seiner Studierenden versteht. Ferner können solche Rollen situativ bedingt oder zeitlich überdauernd sein und die Erwartungen können mehr oder weniger verbindlich sein (vgl. Thomann 2019, 29; Schimank 2010, 60f.).

Beruflich bedingte Rollen enthalten in der Regel wesentliche Bestandteile, die mittels Erwartungen der Arbeitgeberin und/oder der Gesellschaft den Funktionsträgerinnen zugewiesen werden. Dies gilt auch für Personen, die Lehrfunktionen an Hochschulen übernehmen, wobei Hochschuldozierende in der Regel einen verhältnismäßig großen Spielraum in der Selbstkonzeption ihrer Rolle haben. Sie können die Ausgestaltung ihrer Dozierendenrolle also zu einem gewissen Grad mitbestimmen. Dennoch gibt es wesentliche Teilerollen, deren Übernahme von Hochschule, Studierenden und Gesellschaft erwartet wird. Teilweise sind diese Rollenerwartungen eher informell, andere Aspekte sind sogar öffentlich-rechtlich oder arbeitsrechtlich verbindlich festgehalten, etwa in Personalverordnungen (vgl. zu diesen Ausführungen Thomann 2019, 29–33; Weil 2020).

Die Rollenvielfalt von Hochschuldozierenden in der Dyade Dozierende – Studierende lässt sich wie in Abbildung 7 dargestellt fassen, wobei Thomann (2019, 29) von einem „Rollenstrauß“ spricht. Dabei ist festzuhalten, dass diese Darstellung dozierendenzentriert ist: Einerseits wäre für eine umfassende Darstellung der involvierten Rollen auch die Rollenvielfalt der Studierenden zu erfassen, andererseits stellen Rollen wie Dozierende oder Studierende immer nur einen Ausschnitt aus der Rollenvielfalt einer Person dar. Wir alle nehmen zugleich eine Vielzahl von Rollen in verschiedenen Lebensbereichen ein, sind zum Beispiel auch noch Mittelstürmerin im Fußballverein, Vater, Vereinsmitglied etc. Zwischen diesen Rollen kann es nun verschiedene Arten von Konflikten geben, von denen sich einige im Umgang mit Leistungsbeurteilung kristallisieren.

Interrollenkonflikt: Beraten versus Beurteilen

In den meisten Fällen, in denen Dozierende die Rolle der Beurteilenden übernehmen, haben sie parallel dazu auch andere Rollen. Dies betrifft vor allem die Rollen des Gestaltens von Lehr-Lern-Situationen und des Begleitens/Beratens. Insbesondere bei summativen Beurteilungen, also wenn abschließende Bewertungen inklusive Noten, Berechtigungen oder Qualifikationen erteilt werden, zeichnet sich dabei ein Interrollenkonflikt ab, also ein Konflikt zwischen verschiedenen (Teil-)Rollen.

Während wir als Dozierende beim Gestalten und Durchführen von Unterricht und fast noch mehr beim Begleiten und Beraten die Studierenden in

ihrem Lernen zu unterstützen und zu fördern versuchen, müssen wir beim Beurteilen Leistung von ihnen fordern und diese möglichst wertfrei einschätzen. Dieser Rollenkonflikt wird für Lehrende an Hochschulen besonders im Zusammenhang mit Projektarbeiten, schriftlichen Arbeiten und Abschlussarbeiten spürbar: „Der pädagogisch orientierten Begleitung folgt in der Regel die summative Beurteilung von Leistungen (...). Die Nähe der begleitenden formativen Beurteilung wird durch die Distanz der summativen Beurteilung in Frage gestellt, Enttäuschungen können auf beiden Seiten entstehen“ (Thomann 2017a, 28).³¹ Dozierende hegen dabei gemäß Thomann verschiedene Sorgen (ebd., 29):

- Die Beurteilungsfunktion gefährde die Beratungsbeziehung.
- Die erwartete Beurteilung verunmögliche das Vertrauen der Studierenden in die Lehrenden.
- Die Machtzuweisung der Beurteilungsfunktion erschwere einen fachlichen Dialog auf Augenhöhe mit den Studierenden.
- Studierende könnten nach einem quasi gleichberechtigten Beratungsprozess die Beurteilung nicht mehr akzeptieren.
- Der Beratungsprozess erschwere eine anschließende unvoreingenommene Beurteilung.

Diese Sorgen sind alle berechtigt und lassen sich aufgrund des geschilderten Rollenkonflikts Beraten – Beurteilen nicht eliminieren. Wichtig ist ein transparenter Umgang damit. Indem man sich die konflikthafter Aspekte selbst bewusst macht und sie mit den Studierenden bespricht, kann man zur Rollenklarheit beitragen. Dies kann auch explizit erfolgen, indem man etwa in einem Gespräch mit Studierenden sagt: „Ich äußere dies jetzt aus der Perspektive der Beurteilerin...“.

Werden Lehrende gegenüber Studierenden beratend tätig, haben sich Beratungsvereinbarungen als wertvolle Instrumente erwiesen. Sie helfen, gegenseitige Erwartungen oder Aspekte der Vertraulichkeit zu klären (vgl. Engfer 2018, 148). Am häufigsten kommt ein solches Contracting bei der Betreuung wissenschaftlicher Arbeiten und Projekte zum Einsatz: In einer Vereinbarung halten die Betreuerin und der Student die wichtigsten Aspekte ihres Arbeits-

31 Thomann (ebd.) weist darauf hin, dass gelegentlich innerhalb der Hochschule die Beurteilungsfunktion personell von jener der Beratung getrennt wird. Damit verlagere sich in der Regel die intrapersonelle Spannung in eine interpersonelle, indem z. B. Dozierende mit Beurteilungsfunktion von Studierenden einen höheren Status zugewiesen bekommen, was jene mit beratender Funktion frustrieren kann.

bündnisses schriftlich fest. In Kapitel 12.2 wird konkret erläutert, wie eine Kontraktierung erfolgt.

Intrarollenkonflikte und Intrasenderkonflikt

Konflikte können aber auch innerhalb einer (Teil-)Rolle entstehen. So kann eine Studien- oder Prüfungsordnung Vorgaben zum Beurteilen machen, die ich als Dozent problematisch oder sogar fachlich falsch finde. Auch solche Intrarollenkonflikte treten häufig auf und sind nicht leicht zu lösen. Der Weg, problematische Vorgaben auf dem Instanzenweg zu ändern, kann sehr anstrengend sein und ein Erfolg ist alles andere als garantiert. Kreative Auslegungen von Vorgaben können didaktisch sinnvoll sein, sind aber nicht frei von Risiken. Sich im Stillen darüber zu ärgern ist aus psychologischer Sicht wenig empfehlenswert. Insofern empfiehlt sich auch hier ein möglichst transparenter Umgang mit festgestellten Konflikten, sei es gegenüber Studierenden, anderen Lehrenden oder Vorgesetzten (vgl. Thomann 2017a, 30).

Konflikte innerhalb einer Person können auch bezüglich ihrer verschiedenen Rollenanteile auftreten. Ein solcher Intrasenderkonflikt kann etwa dann vorkommen, wenn man von sich in der Beurteilung von Leistungsnachweisen große Sorgfalt erwartet und zugleich unbedingt genügend Zeit für die eigene Forschungstätigkeit aufwenden möchte. Hier zeigt sich der an Hochschulen oft auftretende Konflikt zwischen Lehr- und Forschungstätigkeiten, der über unsere Thematik hinausführt (vgl. Weil 2020, 93).

Als Grundstrategie bei allen Formen von Rollenkonflikten empfiehlt es sich, einen Abgleich zwischen dem eigenen Rollenbild mit Fremdbildern von Studierenden, Vorgesetzten sowie institutionellen Vorgaben vorzunehmen. Dies ermöglicht eine aktive Gestaltung der eigenen Rolle bei gleichzeitig hoher Kongruenz von Fremdbildern und Selbstbild (vgl. Weil 2020, 100).

3.3.2 *Antinomien des Lehrhandelns*

Der strukturtheoretische Ansatz gemäß Helsper (2016, erstmals 1996) hebt die Konflikte, die sich anhand des Rollenstraußes von Dozierenden zeigen, auf eine noch grundsätzlichere Ebene. Er geht davon aus, dass Lehrhandeln prinzipiell nicht vollständig planbar und steuerbar ist.³² Gemeint ist, dass sich Unterrichtsplanungen nie vollständig umsetzen lassen: Was in sozialen Interaktionen genau geschieht, ist nicht antizipierbar. Daraus ergeben sich fundamentale Widersprüche, die Lehrhandeln riskant machen (vgl. Böing 2009,

32 Helsper bezieht sich auf schulischen Unterricht in der Primar-/Grund- und Sekundarstufe, die beschriebenen strukturellen Herausforderungen stellen sich in teilweise anderer Nuancierung aber auch auf der Tertiärstufe.

44f.). Denn „faktisch sind alle Mitglieder der Schulgemeinschaft mitverantwortlich für die Lehr-Lern-Prozesse, haben aber weder eine Zielkontrolle noch können sie die Wege dahin allein bestimmen und für das Ergebnis keine Garantie übernehmen“ (Stangl 2023).

Aus dieser Situation, die Nähe und Distanz, rollenbezogenes und rollenüberwindendes Verhalten verlangt, ergeben sich immer wieder strukturell bedingte Spannungen. Diese bezeichnet Helsper als Antinomien: „Bezogen auf Bildungs- und Erziehungsprozesse ist eine Antinomie dadurch bestimmt, dass für das professionelle pädagogische Handeln widerstreitende Orientierungen vorliegen, die entweder beide Gültigkeit beanspruchen können oder die nicht prinzipiell aufzuheben sind“ (Helsper 2016, 111). Antinomien sind demzufolge gleichsam die strukturellen Mechanismen, die Dilemmata wie die in Tabelle 1 dargelegten oder die in Kapitel 3.3.1 erläuterten Rollenkonflikte hervorbringen.

Jene Antinomien, die zentral sind für die Dilemmata und Rollenkonflikte im Umgang mit Leistungsbeurteilungen, werden nachfolgend kurz erläutert. Ergänzt wird zudem eine Antinomie, die für den Hochschulkontext spezifisch ist. Uns diese Antinomien bewusst zu machen, soll die Erwartungen nicht ins Unermessliche steigern. Im Gegenteil: Indem wir uns bewusst machen, wie „unmöglich“ Lehrtätigkeit oder spezifisch der Umgang mit der Beurteilung von Lernleistungen ist, können wir uns entlasten. Herausforderungen, die sich auf diese grundsätzlichen Antinomien zurückführen lassen, sind weder durch uns als Dozierende noch durch unsere jeweilige Hochschule verursacht, sondern liegen gleichsam in der gesellschaftlichen Organisationsform institutionalisierter Bildung begründet. „Ein wichtiges Element in der Entwicklung der Lehrkompetenz von Hochschullehrenden ist es daher, mit diesen Widersprüchen umgehen zu lernen“ (Wegner u. Nückles 2012, 65).

Subsumtionsantinomie

Um Lehr-Lern-Situationen organisieren zu können, benötigen Lehrende Handlungsrouninen, sollen aber dennoch in der Lage sein, stets das Besondere der einzelnen Lernenden erkennen zu können. Umgekehrt lässt sich sozialwissenschaftlich gesehen die Spezifik des Einzelfalls keinen abstrakten Regeln oder allgemeinen Methodiken unterwerfen, obwohl allgemeine Erklärungsmuster nötig sind, um Einzelfälle verstehen zu können. Sie müssen immer auf ihre jeweilige Fallangemessenheit überprüft und angepasst werden (vgl. Helsper 2016, 112; Böing 2009, 50).

Paradoxerweise wird die Subsumtionsantinomie umso ausgeprägter, je mehr ein Dozent über ein lernendes Gegenüber weiß – denn umso mehr wird er „mit Reflexionsaufforderungen über die Angemessenheit seines Handelns befrachtet. Umfassende Verstehensversuche (...) reduzieren somit nicht, sondern erhöhen erst einmal die Ungewissheit“ (Helsper 1996, 533). Dieses

Paradox wird für Hochschuldozierende besonders in längerfristigen Betreuungssituationen wie Projektarbeiten, schriftlichen Arbeiten und Abschlussarbeiten inklusive Dissertationen sichtbar.

Symmetrieantinomie

Lehrhandeln ist durch eine Asymmetrie gekennzeichnet: Lehrende besitzen größeres Wissen, mehr Erfahrungen und Kenntnisse im Fachbereich, den sie unterrichten, als Studierende. Ohne dieses Gefälle könnten sie keine Lehrfunktion übernehmen. Verschärft wird diese Asymmetrie dadurch, dass Lehrende auch Anordnungs-, Zuweisungs- oder Sanktionsmacht besitzen – das beginnt bei der Erteilung des Rederechts im Unterricht und endet beim Verhängen von Strafen oder der summativen Beurteilung von Leistungen. Andererseits lassen sich – insbesondere tiefenorientierte – Lernprozesse eben gerade nicht durch Macht erzwingen. Im Gegenteil: Wenn die Dominanz der Lehrperson „zu stark, der Zwang zu deutlich und der Wissensvorsprung demonstrativ zu Geltung gebracht werden, erzeugt dies Probleme in der Auseinandersetzung mit der Sache“ (Helsper 2016, 112). Die strukturelle Asymmetrie ist also notwendig für die Entstehung eines Lehrverhältnisses, während zugleich laufend Symmetrisierungen im Austausch zwischen Lehrenden und Lernenden nötig sind, um Lernprozesse zu ermöglichen, zu unterstützen und zu fördern (vgl. ebd.).

Die Symmetrieantinomie zeigt sich im hochschulischen Beurteilungskontext insbesondere dort, wo Feedback der Dozentin einen Studenten zum Weiterlernen anregen soll. Peer-Feedback und Selbsteinschätzungen (siehe Kap. 8.4) sind auch deshalb wertvolle Feedback-Alternativen, weil sie dieser Antinomie weniger stark ausgesetzt sind.

Vertrauensantinomie

Vertrauen ist in Lehr-Lern-Verhältnissen, die immer auch die Möglichkeit beinhalten, dass Lernende „an Aufgaben scheitern und Fehler machen, deswegen ungemein wichtig, weil nur dann die Bereitschaft besteht, offen mit Verstehensgrenzen und Lernproblemen umzugehen, was wiederum eine Voraussetzung dafür ist, dass Bildungsprozesse nicht stagnieren“ (Helsper 2016, 113). Die Bildung von gegenseitigem Vertrauen wird aber erschwert, wenn ein Machtgefälle vorhanden ist, was in institutionalisierten Lehr-Lern-Beziehungen konstitutiv der Fall ist (siehe Symmetrieantinomie). So ist es durchaus rational, dass Lernende sich gegenüber Lehrenden (auch) taktierend verhalten und nicht allzu viel von sich Preis gegeben möchten, während Lehrende im Gegenzug den Lernenden mit einem gewissen Misstrauen genau solche Verhaltensweisen unterstellen. Die Beziehung zwischen Lehrenden und Lernenden „bedarf daher für die Ausgestaltung von inhaltlichen Bil-

dungsprozessen in besonderem Maße einer Vertrauensgrundlage bei Konstellationen, die die Herstellung des Vertrauens erschweren“ (Helsper 2016, 113).

Die Vertrauensantinomie kommt bei hochschulischen Beurteilungssituationen wiederum vor allem in längerfristigen Betreuungsverhältnissen zum Tragen, besonders bei schriftlichen Arbeiten und Projektarbeiten. Das in Kapitel 3.3.1 (Interrollenkonflikt: Beraten versus Beurteilen) erwähnte Vorgehen einer Kontraktierung der Beratungs- und Beurteilungsbeziehung ist deshalb auch hilfreich, um gegenseitiges Vertrauen zu schaffen.

Differenzierungsantinomie

In unseren westlich geprägten Gesellschaften existiert eine grundsätzliche Übereinkunft, dass Bildungschancen für alle gleich sein sollen (dass dem trotz der ständigen Beschwörung von Chancengleichheit nicht so ist, wurde in Kapitel 2 in aller Deutlichkeit gezeigt). Daraus, aber auch aus organisatorischen Gründen folgt, dass vom Kindergarten bis zur Hochschule viele Abläufe im Bildungssystem homogenisierend gestaltet sind. Lehrende sollen keine Lernenden bevor- oder benachteiligen, sie möglichst gleich behandeln und von persönlichen Besonderheiten absehen. Paradoxerweise erzeugt genau diese (an der Oberfläche als solche erscheinende) Gleichbehandlung „Unterschiede an Wissen, Leistungen und schulischen Statusrängen – die den (...) einzelnen [Lernenden, tz] aufgrund ihrer Leistungs- und Anstrengungsbereitschaft zugeschrieben werden“ (Helsper 2016, 115). Da Lernende aber unterschiedlich und Lerngruppen immer heterogen sind, sind differenzierte didaktisch-lernbezogene Unterstützungshandlungen nötig, damit die Lernenden tatsächlich gleiche Lernchancen haben.

Lehrhandeln sieht sich also mit dem Widerspruch konfrontiert, sowohl individuelle Lernprozesse zu initiieren als auch die äußerliche Gleichförmigkeit zu bewahren (vgl. Helsper 2016, 115; Böing 2009, 52). An Hochschulen ist zwar der Anspruch von Studierenden auf persönliche Unterstützungs- und Fördermaßnahmen gegenüber den vorangehenden Schulstufen geringer. Dennoch begegnet uns die Differenzierungsantinomie in der Hochschullehre als Frage danach, wie viel und welche zusätzliche Unterstützung wir einzelnen Studierenden gewähren sollen. Das Schaffen von möglichst zahlreichen Feedbackgelegenheiten entschärft diese Antinomie zumindest teilweise (siehe Kap. 8 zur formativen Leistungsrückmeldung).

Autonomieantinomie

Diese Antinomie ist schon lange bekannt und wurde bereits von Immanuel Kant formuliert: „Wie kultivire ich die Freyheit bey dem Zwange? Ich soll meinen Zögling gewöhnen, einen Zwang seiner Freyheit zu dulden, und soll

ihn selbst zugleich anführen, seine Freyheit gut zu gebrauchen“ (Kant 1803, 32). Aus diesem Paradox resultiert für Lehrende vor allem die Herausforderung, eine Balance zu finden zwischen der Zuschreibung von genügend Autonomie und der Überforderung durch zu viel davon (vgl. Stangl 2023).

Das Lehrhandeln „muss somit in asymmetrischen Rahmungen Selbstständigkeit (...) fordern, was als Zwang zur Eigenverantwortlichkeit erscheinen und Gefahr laufen kann, noch vorhandene Unselbstständigkeit und Hilfsbedürftigkeit“ der Lernenden zu übersehen (Helsper 2016, 116). Die Autonomieantinomie wird in Hochschulen oft ausgeblendet, indem die Studierenden als grundsätzlich selbständig betrachtet werden. Dabei wird gelegentlich übersehen, dass sie fachlich durchaus unterstützungsbedürftig sein können. Dieser Aspekt kann besonders gut adressiert werden, wenn Studierende persönliches Feedback erhalten und geben (siehe Kap. 8).

Näheantinomie

Diese Antinomie wurde bereits anhand der Spannung zwischen Nähe und Distanz im Interrollenkonflikt von Beraten und Beurteilen gezeigt (siehe Kap. 3.3.1). Sie resultiert daraus, dass die rollenförmige Beziehung und das anhand der Differenzierungsantinomie aufgezeigte Gleichbehandlungsgebot die strukturelle Gleichbehandlung der Studierenden fordern, „während in der sozialen Beziehung die Individualität des Einzelnen in den Vordergrund rückt“ (Böing 2009, 51). Dieser Spannung können sich auch Hochschuldozierende nicht entziehen, sie erfordert ständige Abwägungsprozesse.

Epistemologische Antinomie

Diese Antinomie betrifft spezifisch das Lehren und Lernen an Hochschulen. Didaktik als Wissenschaft vom Lehren und Lernen beschäftigt sich in ihrer allgemeinsten Form mit der Frage, wie Menschen gezielt das Lernen von anderen Menschen ermöglichen können, in der Regel im Rahmen eines institutionellen Kontexts (vgl. z. B. Terhart 2019). Dabei ist das zu erlernende Wissen und Können in der Regel gesetzt – fokussiert wird die Frage, wie es am besten gelernt werden kann. Die Hochschulen und die durch sie und in ihnen geführten wissenschaftlichen Diskurse sind aber gerade jene Orte, in denen die wissenschaftliche Erkenntnisgewinnung laufend stattfindet und etabliertes Wissen ständig hinterfragt wird: „In einer Reflexion auf sich selbst haben Studierende (und im Sinne der Unabschließbarkeit wissenschaftlicher Qualifizierung auch die Lehrenden) an Hochschulen zu lernen, *dass* – und *wie* – das, was sie lernen, kontinuierlich hinterfragt und verändert werden muss, um das sein zu können, was wissenschaftlich gelernt werden kann und gelernt zu werden Sinn ergibt“ (Eugster u. Tremp 2018, 78, Hervorhebungen von Eugster und Tremp).

Diese Herausforderung stellt sich in allen Wissenschaftsrichtungen. Um zwei Beispiele zu nennen: Im Bachelorstudium der Physik wird gegenwärtig eine Vorstellung von Elementarteilchen gelehrt, die auf dem Standardmodell der Physik basiert, während jüngst neue Teilchen entdeckt wurden, die nahelegen, dass das Standardmodell nicht vollständig ist. Und in vielen soziologischen Einführungsveranstaltungen werden der Strukturfunktionalismus von Durkheim oder der symbolische Interaktionismus von Simmel gelehrt, obwohl diese Ansätze im gegenwärtigen soziologischen Diskurs kritisch hinterfragt und teilweise zurückgewiesen werden.

Die epistemologische Antinomie besteht also darin, dass an der Hochschule wissenschaftliche Wissensbestände zugleich gelehrt, gelernt und als erkenntnistheoretisch kritikbedürftig ausgewiesen werden. Diese Spannung wird nicht nur in der Gestaltung hochschulischer Curricula spürbar, sondern auch bei der Beurteilung von Leistungen: Wie beurteilen wir den Erwerb von diskursiv breit akzeptierten Wissens- und Könnensbeständen im Vergleich zu ihrer kritischen Hinterfragung? Weshalb unterrichten wir bestimmte Inhalte zu Studienbeginn als eine Art „fachliche Wahrheit“ und erwarten von fortgeschrittenen Studierenden, dass sie just diese Wissensbestände kritisch hinterfragen? Inwieweit sind wir als Teilhabende des wissenschaftlichen Diskurses überhaupt in der Lage, diese Aspekte „objektiv“ einzuschätzen?

3.3.3 *Das Entlastungspotenzial der Dilemmata*

Zum Abschluss dieses Kapitels über die Funktionen von Leistungsbeurteilung sei nochmals betont: Die Darlegung der strukturtheoretischen Antinomien der geschilderten Dilemmata und der mit ihnen verbundenen Rollenkonflikte dient der Entlastung der Lehrenden. Der Blick auf die Antinomien und die Rollenkonflikte erlaubt uns, krisenhafte Situationen im Lehr-Lern-Alltag anders zu interpretieren: Die sich in ihnen zeigenden Dilemmata sind nicht in individuellen Unzulänglichkeiten begründet, sondern liegen in der Struktur unseres Bildungssystems begründet (vgl. Böing 2009, 53).³³

In diesem Sinne sollen die obigen Ausführungen Ihnen helfen, Ihren eigenen – bewussten – Umgang mit den geschilderten Herausforderungen zu finden und darüber auch transparent mit Studierenden, Arbeitskolleginnen und Vorgesetzten zu kommunizieren. Die folgenden Fragen können Ihnen behilflich sein, diesen bewussten Umgang zu finden und zu pflegen:

- 33 Selbstverständlich sollen problematische Zustände bearbeitet werden, sei es auf institutioneller oder bildungspolitischer Ebene (vgl. Breidenstein 2018). Uns unserer Grenzen im individuellen Lehr- und Beurteilungshandeln bewusst zu sein, ist aber für einen realistischen Umgang mit den eigenen Ressourcen wichtig.

- In welcher Rolle fühlen Sie sich wohler: in jener des Beraters oder jener der Beurteilerin? Und aus welchen Gründen?
- Welche für Sie subjektiv schwierigen Situationen tauchen regelmäßig im Umgang mit Leistungsbeurteilungen auf? Welche Dilemmata, Rollenkonflikte oder Antinomien zeigen sich darin?
- Wie können Sie bezüglich der anhand der obigen Fragen identifizierten strukturellen Herausforderungen Transparenz herstellen gegenüber Studierenden oder anderen Beteiligten?
- Welche weiteren Entlastungsmöglichkeiten können Sie angesichts der Einsicht in die strukturelle Unauflösbarkeit der Herausforderungen identifizieren, allenfalls anhand von Ansätzen Ihrer Kolleginnen oder eigenen gelungenen Vorgehensweisen?

Beispiel

Antworten einer jungen Hochschuldozentin auf diese Fragen:

- Ich fühle mich persönlich wohler in der Rolle der Beraterin. Dies liegt daran, dass ich Studierende sehr gerne in ihrem Lernprozess unterstütze. Die Möglichkeit, ihnen durch Feedback bei der Entwicklung ihrer Fähigkeiten zu helfen, ist für mich äußerst erfüllend. Zugleich erkenne ich jedoch die Notwendigkeit einer summativen Beurteilung, um den Lernfortschritt der Studierenden zu messen.
- Ein regelmäßig auftretender Rollenkonflikt ist jener zwischen der Rolle als Lehrender und als Forschender. Da ich viel Zeit in meine Dissertation investieren muss, bleibt oft weniger Zeit für die Lehre. Dies führt zu einem Dilemma zwischen meiner Leidenschaft für die Lehre und dem Erfolgsdruck in der Forschung.
- Um Transparenz gegenüber Studierenden und anderen Beteiligten herzustellen, zielen ich auf eine klare Kommunikation. Ich werde den Studierenden meine Situation erläutern und betonen, dass mein Hauptziel darin besteht, ihnen zu helfen und ihre Leistungen zu fördern. Dabei werde ich auch meine begrenzte Verfügbarkeit für individuelle Betreuung ansprechen, jedoch alternative Ressourcen zur Unterstützung erläutern. Darüber hinaus plane ich, klare Beurteilungskriterien und Erwartungen zu kommunizieren, um die Transparenz in meinem formativen und summativen Beurteilungsprozess zu gewährleisten.

- Meine zeitliche Belastung kann ich reduzieren, indem ich bewährte Verfahren von Kollegen übernehme. Zum Beispiel könnte ich Peer-Feedback einsetzen, um den Studierenden mehr Verantwortung für ihren eigenen Lernprozess zu übertragen. Ich werde auch Möglichkeiten erkunden, um die Bewertung effizienter zu gestalten, wie beispielsweise klar formulierte Lernziele und Beurteilungsraster.

4 Was ist Lernen und wie funktioniert es?

Um Lernleistungen einschätzen und somit bewerten zu können, ist zumindest ein rudimentäres Verständnis davon notwendig, was Lernen eigentlich ausmacht. In vielen Publikationen zum Themenfeld Leistungsnachweise und Leistungsbewertung bleibt dieses Verständnis eine Blackbox. Dadurch müssen Lesende ohne größere lernpsychologische Vorkenntnisse ihr alltägliches Verständnis von Lehren und Lernen an die Ausführungen zur Leistungsbewertung herantragen, was der Gestaltung oder Fortführung von ineffizienten und kontraproduktiven Praktiken Vorschub leistet.

In diesem Sinn möchte ich Ihnen in diesem Kapitel zentrales Grundwissen über den aktuellen Kenntnisstand zum menschlichen Lernen anbieten. Zugleich wird mit Verweis auf diese Erkenntnisse in den nachfolgenden Kapiteln für Sie hoffentlich greifbar, weshalb gewisse Vor- und Nachteile bestimmter Vorgehensweisen ausgewiesen oder bestimmte Handlungsempfehlungen gegeben werden.

4.1 Zwei Lernexperimente

Bevor wir uns einigen Kerncharakteristika des menschlichen Lernens widmen, möchte ich Sie ermuntern, Ihr eigenes Lernen anhand von zwei Aufgaben zu erkunden. Bitte nehmen Sie sich die Zeit, die Aufgaben in der beschriebenen Weise zu bearbeiten – es wird Ihnen das Verständnis der weiteren Ausführungen wesentlich erleichtern.

Aufgabe 1a: Wortlisten merken

Bitte studieren Sie die folgenden Wortlisten und versuchen Sie, sich die 12 Wörter pro Liste zu merken. Bitte nehmen Sie sich maximal eine Minute pro Liste Zeit. Bearbeiten Sie direkt anschließend Aufgabe 2a.

Tab. 2: Wortlisten in Anlehnung an Roediger und McDermott (1995, 814)

Liste 1	Liste 2	Liste 3
Tisch	fein	Faden
sitzen	rau	Öse
Beine	Sandpapier	nähen
Platz	zerklüftet	scharf
Sofa	körnig	Spitze
Sessel	uneben	stechen
Holz	gestrickt	Fingerhut
Kissen	schroff	Heuhaufen
drehbar	Sand	piksen
Hocker	Planung	Spritze
schaukeln	Schotter	Stoff
Sitzbank	Rüpel	stricken
Liste 4	Liste 5	Liste 6
sauer	Bett	Hügel
Bonbon	ausruhen	Tal
Zucker	wach	klettern
bitter	träumen	Gipfel
gut	Nickerchen	Spitze
Geschmack	aufwachen	Kuppe
Holz	dösen	Ebene
schnuckelig	schlummern	Gletscher
Honig	müde	Ziege
Limonade	Bettdecke	Alpinist
Schokolade	schnarchen	steil
Kuchen	gähnen	Ski

Aufgabe 2a: Sätze lesen

Bitte lesen Sie die folgenden Sätze ein (einziges) Mal aufmerksam durch (es handelt sich um die Rekonstruktion eines Experiments aus Bransford u. Franks 1971):

Die warme Brise bewegte die Luft.
Der Mann las die Geschichte vor.
Die Ameisen aßen die süße Marmelade.
Der Felsen zerstörte die kleine Hütte am Rande des Waldes.
Die Hütte stand am Rande des Waldes.
Der Felsen, der den Berg hinunterrollte, zerstörte die kleine Hütte.
Die Brise bewegte die schwere Abendluft.
Die warme Brise, die vom Meer herüberwehte, bewegte die Luft.
Der Bauer saß auf seinem Traktor.
Die Ameisen in der Küche aßen die Marmelade.
Der Felsen zerstörte die kleine Hütte.
Die Brise bewegte die Abendluft.
Die Ameisen in der Küche aßen die süße Marmelade auf dem Tisch.
Der Felsen rollte den Berg hinunter.
Der Mann saß lesend auf dem Sofa.
Die Chefin schaltete ihren Computer aus und öffnete das Fenster.
Der Mann, der sich auf der Couch ausruhte, las die Geschichte in der Zeitung.
Die Ameisen aßen die süße Marmelade, die auf dem Tisch stand.
Der Stein rollte den Berg hinunter und zerstörte die Hütte.
Die Marmelade stand auf dem Tisch.
Die warme Brise bewegte die schwere Abendluft.
Die Ameisen waren in der Küche.
Die Frau stieg am Hauptbahnhof aus der Straßenbahn.
Der alte Mann las die Geschichte.
Der Wind wehte vom Meer her.
Der alte Mann las die Geschichte in der Zeitung.
Die Abendluft war schwer.

Gehen Sie nun weiter zu Aufgabe 1b.

Aufgabe 1b

Bitte geben Sie – ohne vorne nachzuschauen – für die folgenden Wörter an, ob sie in einer der sechs Listen von Aufgabe 1a enthalten waren. Verwenden Sie dafür folgende Skala:

- 4 = Ich bin sicher, dass das Wort in einer der Listen enthalten ist.
- 3 = Das Wort ist vermutlich in einer der Listen enthalten.
- 2 = Das Wort ist vermutlich nicht in einer der Listen enthalten.
- 1 = Ich bin sicher, dass das Wort nicht in einer der Listen enthalten ist.

fein	_____	Sitzbank	_____	Gipfel	_____
Haus	_____	Bonbon	_____	grün	_____
Seil	_____	raten	_____	Spritze	_____
backen	_____	Hase	_____	salzig	_____
grob	_____	zackig	_____	süß	_____
Pullover	_____	Berg	_____	Holz	_____
Stimme	_____	fit	_____	steil	_____
Bettdecke	_____	flach	_____	Wagen	_____
Kerze	_____	Büro	_____	Pflanze	_____
Gras	_____	schlafen	_____	Sand	_____
stehen	_____	lernen	_____	taumeln	_____
Nadel	_____	stechen	_____	dauernd	_____
hohl	_____	Zwerg	_____	klein	_____
schnuckelig	_____	Stuhl	_____	ausruhen	_____

Aufgabe 2b

Bitte geben Sie – wiederum ohne nachzuschauen :-)) – für jeden der folgenden acht Sätze eine Einschätzung ab, ob der Satz in diesem Wortlaut im Rahmen von Aufgabe 1b aufgeführt war oder nicht.

4 = Ich bin sicher, dass ich den Satz in dieser Form gelesen habe.

3 = Vermutlich habe ich den Satz in dieser Form gelesen.

2 = Vermutlich habe ich den Satz in dieser Form *nicht* gelesen.

1 = Ich bin sicher, dass ich den Satz *nicht* in dieser Form gelesen habe.

- Der alte Mann ruhte sich auf der Couch aus. _____
- Die warme Brise wehte vom Meer her. _____
- Das Kind rannte mit wehenden Haaren zum Spielplatz. _____
- Die Ameisen in der Küche haben die süße Marmelade gegessen.
- Der Mann las die Zeitung. _____
- Der alte Mann, der sich auf der Couch ausruhte, las die Geschichte.
- _____
- Die Brise, die vom Meer her wehte, bewegte die schwere Abendluft.
- _____
- Der Felsen, der den Berg hinunterrollte, zerstörte die Hütte am Waldrand. _____
- Die Ameisen in der Küche aßen die süße Marmelade, die auf dem Tisch stand. _____

Bitte blättern Sie nun zu Kapitel 16, wo Sie die Auflösungen zu Ihren Einschätzungen sowie Erläuterungen zum Hintergrund der beiden Lernexperi-

mente finden. Wenn Sie die Fragen beantwortet und die Erläuterungen in Kapitel 16 gelesen haben, sind Sie optimal auf die weitere Lektüre vorbereitet.

4.2 Lernen ist Konstruktion in einem Netzwerk

„Lernen ist ein aktiver und konstruktiver Prozess.“ Diese Aussage hat seit der Wende zum 21. Jahrhundert viele Studierende der Pädagogik und Didaktik gleichsam als Mantra begleitet. Heute wird sie in Fachtexten dutzendfach in genau der zitierten Formulierung verwendet, wie eine Googlesuche zeigt – manchmal mit, häufiger ohne Quellenangabe.³⁴ Die Aussage betont, dass Lernende ihr Wissen selbst konstruieren müssen und Lehrende es ihnen nicht wie ein physisches Objekt weiterreichen können. Damit erklärt sie, weshalb „nicht jedes Lehren *auch tatsächlich* zum Lernen auf Seiten der Belehrteten führt“ (Terhart 2019, 16f., Hervorhebung von Terhart). Sie verdeutlicht insofern auch die Wichtigkeit von Vorgängen der Leistungskontrolle und -rückmeldung: Sie dienen als Rückkopplungsschleufe für die gegenseitige Verständigung zwischen Lehrenden und Lernenden und helfen dadurch, Lehren und Lernen fortlaufend aufeinander abzustimmen.

Der konstruktive Charakter des Wissensaufbaus hat aber noch weiterreichende Folgen, die erst sichtbar werden, wenn wir uns die Mechanismen der Wissenskonstruktion genauer anschauen. Zentral ist dafür die Erkenntnis aus der Hirnforschung, dass menschliches Lernen in neuronalen Netzwerken stattfindet. Erworbenes Wissen und Können sind demzufolge in neuronalen Netzwerken repräsentiert. Die zugrundeliegenden physiologischen Vorgänge sind enorm komplex (siehe Kap. 4.3). Hilfreich ist deshalb ein Bild: Lernen bedeutet, Spuren zu legen. Der Neurowissenschaftler David Eagleman illustriert dies anhand einer Geschichte von zwei Dörfern:

Stellen Sie sich zwei Städte vor, die nur wenige Kilometer voneinander entfernt sind. Die Menschen, die sich von einer Siedlung zur anderen begeben möchten, nehmen alle möglichen Wege: Einige Reisende gehen die landschaftlich reizvolle Strecke entlang der Hügelkämme, andere bevorzugen den Schatten der Felsen, einige bewegen sich auf glitschigem Gestein am Fluss und wieder andere nehmen den riskanteren, aber schnelleren Weg durch den Wald. Im Lauf der Zeit und indem die Reisenden Erfahrungen sammeln, erweist sich eine Route als immer beliebter. Schließlich wird der meistbenutzte Weg immer ausgetretener und schließlich zum Standard. Nach einigen Jahren legt die lokale Regierung Straßen an. Nach einigen Jahrzehnten werden diese zu Autobahnen ausgebaut. (Eagleman 2020, 194, Übers. tz)

34 Die älteste Verwendung, die ich eruieren konnte, findet sich in Gräsel et al. (1997, 5), was auf eine Herkunft aus dem Umfeld von Heinz Mandl deutet, der zur damaligen Zeit einer der einflussreichsten deutschsprachigen Vertreter eines „gemäßigten Konstruktivismus“ war.

So ähnlich funktioniert auch das Lernen im Hirn und im neuronalen Netzwerk des restlichen menschlichen Körpers. Begegnet man einem Phänomen zum ersten Mal, etwa einem neuen Wort, einem sachlichen Zusammenhang, einer Regel etc., so wird diese Begegnung in unseren Nervenbahnen lediglich leichte Spuren hinterlassen. Im Bild von Eagleman wäre das vielleicht vergleichbar mit niedergetretenem Gras auf einer Wiese oder Fußabdrücken durch nasse Schuhe in der Nähe des Flusses – Spuren, die schon bald wieder verschwunden sind. Nur wenn die Wege immer wieder begangen werden, verfestigen sie sich – und zwar umso mehr, je häufiger sie begangen werden. Hilfreich ist dabei auch, wenn sie nicht immer in die gleiche Richtung, aber mit unterschiedlichen Schuhen begangen werden, wenn zwischendurch auch Mountainbikes oder Motorräder zum Einsatz kommen und wenn ab und zu Seitenwege betreten werden oder hinter umliegende Bäume und Felsen geschaut wird. All diese Aktivitäten führen dazu, dass der Weg immer einfacher zu begehen wird und möglichst viele Verbindungen zu anderen Wegen etabliert werden können.

Die Kehrseite dieses Bildes bedeutet auch: Wege, die nur sehr selten begangen werden, können irgendwann verschwinden. Vielleicht verwaldet eine nicht mehr überquerte Wiese oder der Fluss häuft Geschiebe an, das eine Stelle mit der Zeit unbegehrbar macht. Auf ähnliche Weise werden Nervenverbindungen, die nicht gebraucht werden, irgendwann ausgedünnt, und Nervenzellen, die überhaupt nicht gebraucht werden, können sogar absterben. Das bedeutet auch, dass es mit der Zeit schwieriger wird, neue Wege zu finden, also gänzlich Neues zu lernen – die alten Pfade sind so befestigt und gut bekannt, dass es schwerfällt, sie zu verlassen (vgl. Eagleman 2020, 194). Noch schwieriger wird es, bekannte Wege abzubauen, also umzulernen – man stelle sich den Aufwand für den Abbruch einer Autobahn vor. Das ist ein Grund, weshalb sich mentale Konzepte auch dann lange halten können, wenn man durch bewusstes Lernen eigentlich schon zu anderweitigen Ansichten gelangt ist. Eine besondere Bedeutung hat dieser Umstand im Umgang mit sogenannten Prä- oder Fehlkonzepten im naturwissenschaftlichen Unterricht (vgl. z. B. Deslauriers et al. 2019; Krist 1999). Ganz einfach erleben können Sie das Phänomen, falls Sie plötzlich Ihren Gewürz- oder Kleiderschrank völlig neu ordnen. Sie werden Wochen oder Monate benötigen, bis Sie nicht mehr regelmäßig ohne nachzudenken an den falschen Ort greifen.

Das Bild vom Lernen als Spurenlegen ist noch in einer weiteren Weise hilfreich: Es veranschaulicht, dass das Hirn kein klassischer Speicher ist. Gelerntes ist im Hirn nicht wie in einem Buch oder auf einem Chip abgespeichert. Lernen ist also nicht das Einspeichern von Aussagen oder Abläufen, die dann unverändert wieder abgerufen werden. Vielmehr geht es um Wahrscheinlichkeiten: Je stärker ich eine bestimmte Verbindung gelernt habe, desto höher ist die Wahrscheinlichkeit, dass ich sie beim Auftreten bestimmter

Auslöser auch tatsächlich abrufe, und dass meine Verhaltensweise jedes Mal sehr ähnlich ausfällt.

Dies erklärt nun Ihre mutmaßlichen Erfahrungen in den beiden Aufgaben zu Beginn des Kapitels: Die Listen in Aufgabe 1 enthalten Wörter, deren Bedeutungen starke Assoziationen zu einem bestimmten Wort haben, das aber selbst nicht explizit aufgeführt ist, etwa Stuhl, süß, schlafen. Sollten Sie bei mehreren dieser in Kapitel 4.1 aufgeführten Wörter eine hohe Sicherheit empfunden haben, diese tatsächlich in den Listen „gelernt“ zu haben, so ist dies mit dem Bild von Eagleman zu erklären: Die Listen haben Sie gleichsam dazu gebracht, auf einem bestimmten Weg hin und herzufahren, der sie an der Abzweigung zum Wort Stuhl vorbeiführt – ohne dass Sie diese Abzweigung tatsächlich genommen hätten. Bei der nachträglichen Abfrage – der Re-Konstruktion – sind Sie unbewusst wieder ähnliche Wegstrecken abgefahren und haben irgendwo die Abzweigung zum Wort Stuhl erwischt.

Daran zeigt sich auch der konstruktive Charakter unseres Gedächtnisses: Gelerntes Wissen ist nicht als feste Einheit gespeichert. Vielmehr wird es bei jedem Abrufvorgang neu rekonstruiert, und dabei kann es eben aufgrund von starken Verknüpfungen – „Autobahnen“ im Bild von Eagleman – zu subjektiv sehr angemessen scheinenden Abweichungen gegenüber dem ursprünglichen Input kommen.³⁵ Im Gegenzug bewirkt jedes Abrufen von Gedächtnisinhalten die Festigung der entsprechenden Verknüpfungen.

Die Illusion, die durch Aufgabe 2 hervorgerufen wird, beruht ebenfalls auf dem Lernen durch neuronale Netzwerke: Die verschiedenen, redundanten Sätze bauen insgesamt vier komplexe Aussagen auf. Diese bestehen aus jeweils 4 einfachen Aussagen, zum Beispiel „die Ameisen waren in der Küche“, „die Marmelade war auf dem Tisch“, „die Marmelade war süß“ und „die Ameisen aßen die Marmelade“. Diese 4 Aussagen lassen sich zur vollständigen komplexen Aussage kombinieren „Die Ameisen in der Küche aßen die süße Marmelade, die auf dem Tisch stand“. Durch das Lesen der verschiedenen Sätze, die sich inhaltlich teilweise überlagern, baute Ihr Hirn eine semantische Repräsentation der komplexen Aussagen auf: Es merkte sich die Inhalte, aber nicht den Wortlaut der einzelnen Sätze, also so etwas wie „Ameisen in Küche aßen süße Marmelade auf Tisch“. Diese Gesamtaussage kann es relativ zuverlässig rekonstruieren, die teilweise sehr ähnlichen Sätze, die zu ihrem Aufbau geführt haben, hingegen nicht (vgl. dazu Bransford u. Franks 1971).

35 Dies ist übrigens nicht als Fehlfunktion des Hirns zu interpretieren, auch wenn Roediger u. McDermott (1995) von „memory illusions“ sprechen. Denn Personen mit Gedächtnisstörungen wie z. B. Demenzpatienten erkennen nicht nur deutlich weniger der tatsächlich gelernten Wörter, sondern produzieren auch deutlich weniger Falscherkennungen der Lockwörter. Das bedeutet, dass die Falscherkennungen/Gedächtnisillusionen Anzeichen einer intakten Gedächtnisfunktion sind (vgl. Schacter u. Addis 2007, 775f.).

Konsequenzen für das Lehren und Lernen an Hochschulen

Das Fazit aus den bisherigen Ausführungen lautet: Unser Hirn merkt sich Kernaussagen, aber nicht wörtliche Formulierungen. Der zugrundeliegende neurologische Prozess ist wiederum durch das Bild von Eagleman zu erklären: Je mehr die einzelnen Sätze sich überlappende Aussagen formulieren, desto mehr wird dadurch eine bestimmte Wegverbindung „eingeschliffen“. Das ist übrigens für Lernen im eigentlichen Sinn kein Problem, sondern sehr effizient. Problematisch wird es im Lehrzusammenhang dort, wo unter Lernen Auswendiglernen von Formulierungen verstanden wird. Dafür ist unser Hirn nicht gemacht, und auswendig gekonnte Formulierungen haben in aller Regel wenig zu tun mit tatsächlichem Wissen und Können. So könnte ich mir beispielsweise mit großem Aufwand einen Text in Ungarisch merken, einer Sprache, die ich nicht im Geringsten beherrsche. Mit Verstehen oder Können hätte das aber nichts gemein.

Leider bewirken viele schulische Lernvorgänge nicht das gewünschte Verstehen, sondern nur das oberflächliche Kennen von Begriffen. Der Hauptgrund dafür liegt meist in unklaren Lernzielen und in Leistungsnachweisen, die einseitig auswendig gelerntes Wissen prüfen, wie in Kapitel 5 näher erläutert wird. Der Nobelpreisträger Richard Feynman sprach in diesem Zusammenhang vom Unterschied zwischen „knowing the name of something and knowing something“ (Feynman 1988, 14).

Besonders gut illustriert diesen Unterschied eine Untersuchung, die Physik-Studierende einer australischen Universität im ersten Studienjahr zu ihrem Verständnis von Aspekten der Gravitation befragte. Die Studierenden sollten vorhersagen, mit welcher Geschwindigkeit zwei gleich große, aber unterschiedlich schwere Bälle zu Boden fallen würden – und wie sie etwaige Unterschiede erklären. Die Studienautoren kamen zur Schlussfolgerung, „dass die Studierenden viel über Physik wissen, dies aber nicht auf die Alltagswelt übertragen können. (...) In vielen Fällen benutzten die Studierenden mathematische Gleichungen, um Vorhersagen zu erklären, wenn auch oft in unangemessener Weise, was darauf hindeutet, dass sie viel Physikwissen zur Hand hatten, aber nicht wussten, welcher Teil davon auf die gegebene Situation zutrifft“ (Gunstone u. White 1981, 298f., Übers. tz).

Die zitierte Studie liefert damit eine konkrete Beschreibung der Art von Wissen, die Feynman als „knowing the name of something“ bezeichnete und für die im Deutschen oft der Begriff „träges Wissen“ verwendet wird.³⁶ Lernpsychologisch verkürzt könnte man sagen, dass das Lernen von Namen der Dinge oberflächliches Lernen ist, während das Verstehen der Dinge tieferorientiertes Lernen ist. Diese Unterscheidung ist wesentlich für den Um-

36 In die jüngere deutschsprachige Diskussion eingeführt wurde der Begriff von Renkl (1994). Er scheint aber auf Whitehead (1929) zurückzugehen, der das Lernen von „inert ideas“ im Schulunterricht anprangerte.

gang mit Lernzielen (siehe Kap. 5) und die Frage, wie Leistungsnachweise ihr Erreichen zuverlässig zu überprüfen helfen (siehe Kap. 6.3 und 9.2).

4.3 Neuronale Netzwerke

Das neurologische Prinzip, mit dem Gedächtnispfade eingeprägt werden, wird durch den englischen Merkspruch veranschaulicht: „Neurons that fire together, wire together.“³⁷ Die genaue Funktionsweise von neuronalen Netzwerken ist freilich enorm komplex und kann hier nicht im Detail erläutert werden – dazu wäre mindestens ein ganzes Buchkapitel nötig.³⁸ Entscheidend ist, dass über die ganze Lebensspanne Nervenverbindungen aufgebaut und verändert werden können.

Entgegen früheren Annahmen bleibt das Hirn bis ins hohe Alter plastisch und damit lernfähig. In bestimmten Hirnregionen können sogar neue Neuronen gebildet werden. Diese Plastizität ist aber begrenzt:

- Zunächst einmal führen verschiedene Faktoren dazu, dass Lernen mit fortschreitendem Alter etwas langsamer erfolgt. Dies wird aber in vielen Fällen dadurch ausgeglichen, dass es zugleich effizienter erfolgt, weil es an ein breites Vorwissen anschließen kann. Dies ist in der Regel bei beruflichem Fachwissen der Fall.
- Eine zweite Begrenzung besteht darin, dass bis zum Erwachsenenalter für verschiedene Fähigkeiten bestimmte Entwicklungszeitfenster bestehen. Werden diese verpasst, kann die betreffende Fähigkeit lebenslang nicht mehr oder nur noch bruchstückhaft gelernt werden. Für viele grundlegenden Fähigkeiten wie Sprachenlernen oder das stereoskopische Sehen scheint dieses Zeitfenster von der Geburt bis etwa 6 oder 7 Jahren zu dauern. Ein Kind, das bis 7 Jahre keine Gelegenheit hatte, sprechen zu lernen, wird kaum mehr sprechen lernen (vgl. Eagleman 2020, 195f.).

37 Der Merkspruch wird gemeinhin dem Psychologen Donald Hebb zugesprochen, dessen Werk über das Lernen in neuronalen Netzwerken und synaptische Plastizität als Klassiker der Hirnforschung gilt. Vermutlich stammt der Satz nicht direkt von Hebb, wohl aber diese differenzierte Formulierung des Prinzips: „Any two cells or systems of cells that are repeatedly active at the same time will tend to become ‚associated‘, so that activity in one facilitates activity in the other“ (Hebb 1949, 70).

38 Für interessierte Lesende: Eine der besseren – ebenso verständlichen wie kurzweiligen – Erklärungen ist das Kapitel „Neuroscience 101“ von Sapolsky (2018, 679–706; das Buch ist auch auf Deutsch erhältlich).

- Plastizität ist nicht einseitig positiv zu werten. Der Gedanke, dass das Hirn und damit unser Lernen und Denken stets formbar bleiben, fasziniert uns und verbreitet Optimismus. Das zeigt sich in zahlreichen Büchern mit Titeln im Stil von „Das Gehirn optimiert sich selbst“ oder „Bau dein persönliches Netzwerk: Das erfolgreiche Gehirn“.³⁹ Sapolsky (2018, 152) weist demgegenüber darauf hin, dass die Fähigkeit des Hirns, sich als Reaktion auf Erfahrungen anzupassen, grundsätzlich wertneutral ist: Wenn das Hirn erblindeter Personen beim ertasten von Braille-Schrift den visuellen Kortex nutzt und dadurch zusätzliche Kapazität für die Verarbeitung von taktilen Reizen aufbaut, finden wir das positiv. Ebenso auf die neuronale Plastizität zurückzuführen sind aber Hirnveränderungen infolge von traumatischen Erfahrungen, die große Qualen auslösen können. Auf der gesellschaftlichen Ebene ist es nützlich, dass der motorische Kortex für die Fingersteuerung bei Chirurgen ausgebaut wird – bei Tresorknackern ist das hingegen weniger erwünscht.

Um nun doch noch einen etwas genaueren Blick auf die Funktionsweise des menschlichen Hirns zu werfen, schauen wir uns zuerst an, wie eine Nervenzelle aufgebaut ist (siehe Abb. 8). Ein Neuron besteht aus drei Komponenten (vgl. Zhang 2019, Jäncke 2018):

1. Dem Zellkörper inklusive Zellkern, der sich verzweigt in
2. viele Dendriten („Dendritenäste“), die Signale anderer Nervenzellen empfangen, und
3. einen besonders langen Fortsatz namens Axon, der wie eine Antenne Signale seines Neurons an die Dendriten anderer Neuronen weiterleitet.

Das Axon verzweigt sich gegen sein Ende hin wie Äste eines Baumes. Die Enden dieser Äste liegen jeweils nahe an Enden von Dendritenästen anderer Neuronen. So gibt ein Neuron Signale an bis zu 10.000 andere Neuronen weiter. Entscheidend für die Geschwindigkeit der Signalübertragung durch das Axon ist die Stärke der Myelinschicht um das Axon: Je besser das Axon isoliert ist, desto schneller die Übertragung.

Die Übertragung der Signale findet an so genannten Synapsen statt (siehe Abb. 8). Die genauen physiologischen Vorgänge würden zu weit ins Detail

39 Es handelt sich um von mir erfundene Buchtitel. Es erwies sich freilich als schwierig, Titel zu erfinden, die es nicht schon gibt. Insofern ist nicht auszuschließen, dass nach Publikation dieses Buches Bücher mit den von mir erfundenen Titeln erscheinen. In diesem Fall wären die Übereinstimmungen „weder beabsichtigt noch zufällig, sondern unvermeidlich“, wie Heinrich Böll (1974, 5) es formulierte.

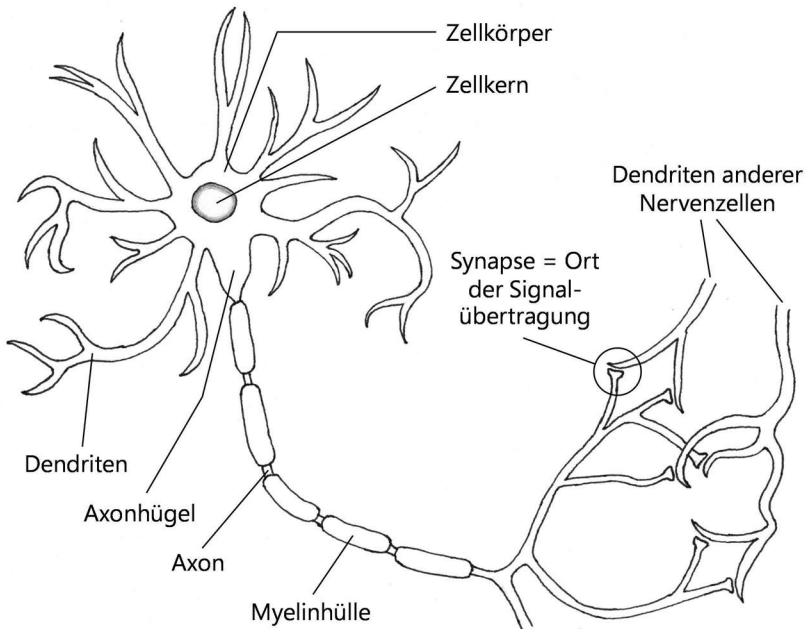


Abb. 8: Aufbau einer Nervenzelle. Quelle: Eigene Handzeichnung

führen.⁴⁰ Entscheidend ist: Um zu lernen, müssen keine neuen Synapsen oder gar neue Neuronen gebildet werden, maßgeblich ist die Verstärkung bestehender Verbindungen. Zudem wird durch zunehmende Myelinisierung von oft gebrauchten Axonen insbesondere der Austausch zwischen weiter auseinanderliegenden Hirnregionen verbessert (vgl. Sapolsky 2018, 138f.; 175).

Die Größenordnung der Vernetzungskapazität im menschlichen Hirn ist überwältigend: Ein durchschnittliches menschliches Hirn umfasst ungefähr 100 Milliarden Neuronen, die Anzahl der Verbindungen wird auf 1 Billiarde (10^{15} oder eine Million Milliarden!) geschätzt. Aufgrund der andersartigen Konstruktionsweise ist die Verarbeitungskapazität schwierig mit Computern zu vergleichen. Geschätzt wird, dass das Hirn pro Sekunde 1 Trillion Bits übertragen kann – also rund 50-mal mehr als eine moderne USB-Verbindung (USB 3.2). Die Gedächtniskapazität scheint noch schwieriger zu schätzen zu sein, die Angaben variieren zwischen einem und tausend Terabytes – die

40 Wer sich gerne vertiefen möchte, dem sei empfohlen, im Internet nach dem Stichwort „action potential“ zu suchen. Bereits der englische Wikipedia-Artikel dazu führt tief in die Materie ein.

tiefste Schätzung entspricht damit immer noch der Speicherkapazität einer im Jahr 2023 zeitgemäßen Laptop-Harddisk (vgl. zu diesen Zahlen Zhang 2019 und Jäncke 2018).

4.4 Drei zentrale Lernmechanismen in neuronalen Netzwerken

Der konstruktive Charakter von Lernen und Gedächtnis ist verbunden mit drei Prozessen, die für das Lernen zentral sind. Alle drei erklären unter anderem die zentrale Bedeutung von Feedback für Lernleistungen. Der erste ist das Lernen durch Vorhersagefehler und der zweite die Aktivierung von Vorwissen als Grundlage für nachhaltiges Lernen. Wesentlich für den Aufbau von komplexem Wissen und Können ist der dritte Prozess, der im Bilden und Anwenden von Schemata besteht.

4.4.1 Lernen durch Vorhersagefehler

Wie bereits sichtbar wurde, sind das Hirn und die menschlichen Sinne keine schlichte Reiz-Reaktions-Maschine, die lediglich auf Stimuli der Außenwelt reagiert. Vielmehr beinhaltet das menschliche Hirn zahlreiche Vorhersageschleifen (siehe Abb. 9). Dazu gehören sinnliche Vorhersagen etwa zur visuellen oder auditiven Wahrnehmung, einschließlich Vorhersagen zur Steuerung der eigenen oder Einschätzung fremder Bewegungen. Die durch diese Schleifen erstellten gehirninternen Vorhersagen werden zu sensorischen Simulationen kombiniert, und diese werden mit sensorischem Input abgeglichen. Stimmen interne Simulation und sensorischer Input überein, so wird die Simulation zur erlebten Erfahrung. Werden Abweichungen festgestellt, muss das Hirn seine Vorhersagefehler korrigieren (vgl. Barrett 2017, 62–64).

Das, was wir als sensorischen Input der Außenwelt wahrnehmen, wird neurologisch betrachtet zu mindestens 90 % durch sensorische Gedächtnisinhalte produziert (Gregory 1998).⁴¹ Unsere Aufmerksamkeit wird vor allem dorthin gerichtet, wo der sensorische Input von unseren Vorhersagen abweicht – das geschieht weitgehend unbewusst oder reflexartig. Nehmen wir zum Beispiel eine überraschende Bewegung aus dem Augenwinkel war, so liegt ein Vorhersagefehler vor: Eigentlich erwartete die visuelle Vorhersage

41 Dies zeigt sich auch physiologisch: So kommen ca. 80 % der zum Nucleus geniculatus lateralis (Teil der Sehbahn) führenden Nervenfasern aus der Hirnrinde und nur ca. 20 % aus den Netzhäuten (Gregory 1998, 1694).

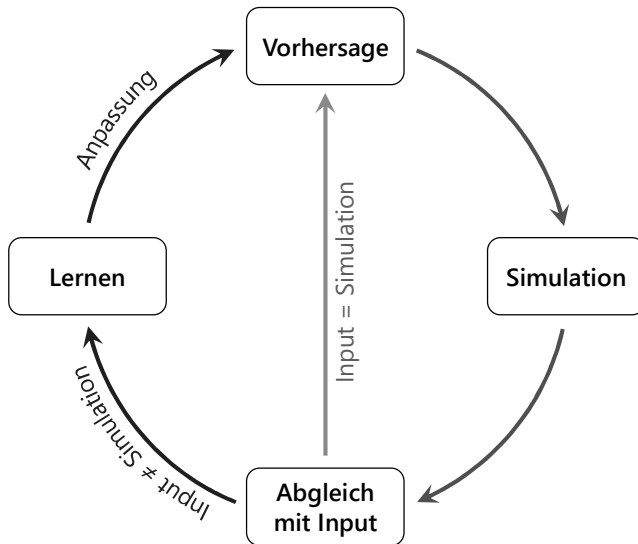


Abb. 9: Lernen durch Vorhersageschleifen. Quelle: Eigene Abbildung, basierend auf Barrett 2017, 63

keine Bewegung, nun scheint aber eine stattzufinden. Deshalb verspüren wir einen starken Impuls (Reflex), unsere Augen dieser Bewegung zuzuwenden, sie so genauer einschätzen, mit unserer ursprünglichen Vorhersage abgleichen und diese schließlich korrigieren zu können.

Ähnliches geschieht auch auf der kognitiven Ebene. Wenn Sie beispielsweise Zeitung lesen und Timor-Leste als asiatisches Land bezeichnet wird, während sie bisher davon ausgingen, es liege in Afrika, so liegt ein Vorhersagefehler vor. Sie können nun entweder Ihr kognitives Wissen anpassen, also Timor-Leste künftig in Asien verorten – dieser Vorgang ist umso wahrscheinlicher, wenn Sie nachschauen, wo das Land genau liegt, denn damit legen Sie mehr Spuren in Ihrem Gedächtnis. Oder Sie gehen von einem Fehler in der Zeitschrift aus und verzichten auf eine Anpassung Ihrer inneren Weltkarte (vgl. Ouden, Kok u. Lange 2012, 3f.).

Interessanterweise wird umso mehr gelernt, desto größer der Vorhersagefehler ist. So zeigte eine Studie, dass Teilnehmende aus dem Feedback zu ihren Falschantworten in Wissenstests dann am meisten lernten, wenn sie sich ursprünglich sehr sicher waren, ihre Antwort sei richtig (vgl. Pine et al. 2018, 3f.). Dies entspricht dem Modell der Vorhersageschleife, demzufolge ein größerer Unterschied zwischen Simulation und Input einen stärkeren Impuls zur Anpassung der Vorhersage bewirkt. Diesen Umstand können wir uns auch didaktisch zunutze machen im Umgang mit sogenannten Prä- oder

Fehlkonzepten im naturwissenschaftlichen Unterricht (vgl. z. B. Deslauriers et al. 2019; Krist 1999): Indem wir Studierende absichtlich Falschlösungen aufgrund ihrer Misskonzepte produzieren lassen und sie danach mit den entstandenen Fehlern konfrontieren, versetzen wir sie am ehesten in die Lage, problematische Konzepte in kurzer Zeit anzupassen.

Die unzähligen Vorhersageschleifen sind auf verschiedenen Ebenen vernetzt: Neuronen sind Teil neuronaler Vorhersageschleifen; Hirnregionen sind Teil von Vorhersageschleifen, die aus verschiedenen Hirnregionen aufgebaut sind. Diese Schleifen laufen in einem Prozess gigantischer Datenverarbeitung parallel und vernetzt ab und bringen unsere Sinneseindrücke hervor, aus denen wiederum unsere Kognitionen, Emotionen und Motivationen hervorgehen. Dabei ist zu unterscheiden zwischen Zuständen, bei denen die generierte Erfahrung primär hirnintern generiert wird wie Tagträumen, Erinnerungen oder optischen Illusionen, und Zuständen, bei denen die sensorische Erfahrung dominiert wie meditative Zustände.

Lernen findet vor allem dann statt, wenn hirninterne Simulation und sensorische Wahrnehmung miteinander abgeglichen werden: Wenn Vorhersageschleifen auf höheren Ebenen (Kognition, Motivation, Psychomotorik) zur Korrektur und damit zum Ausbau von Wissen und Können genutzt werden (vgl. Barrett 2017, 62–65). Mit anderen Worten: Lernen durch Vorhersageschleifen „kann als der Prozess gesehen werden, der uns befähigt, vergangene und aktuelle Ereignisse zu nutzen, um vorauszusagen, was die Zukunft bringt“ (Niv u. Schoenbaum 2008, 265, Übers. tz). Feedback spielt dabei eine zentrale Rolle (siehe Kap. 8).

4.4.2 Aktivierung von Vorwissen

Wie bereits erläutert, ist das Gedächtnis kein Speicher, in den Erinnerungen, Wissen und Können abgelegt werden, um später unverändert wieder abgerufen zu werden. Vielmehr ist jeder Akt der Erinnerung, jede Aktualisierung von Wissen und Können, ein rekonstruktiver Akt: Die erinnerten und gelernten Bestände werden neu konstruiert, wobei auch Veränderungen gegenüber der ursprünglichen Einspeicherung auftreten können, wie Ihnen die beiden Lernexperimente aus Kapitel 4.1 bewusst gemacht haben dürften. Dies ist unter anderem darin begründet, dass ein großer Teil unserer Wahrnehmung aus Simulationen besteht, die zwar wahrscheinlich sind, aber nicht auf sensorischer Erfahrung des de-facto-Geschehens beruhen.

Der wichtigste Grund für den rekonstruktiven Charakter des Rememberns dürfte aber seine biologische Hauptfunktion sein. Sie besteht vermutlich darin, mentale Simulationen künftiger Ereignisse zu konstruieren. Das Gedächtnis ist deshalb nicht in erster Linie darauf ausgerichtet, vergangene Sinneseindrücke oder Kognitionen exakt zu reproduzieren, sondern sie für mög-

lichst wahrscheinliche Annahmen künftiger Vorgänge zu verwenden. Es steht somit im Kontext von Vorhersageschleifen, wie wir sie oben diskutiert haben (vgl. Brigard 2014; Schacter u. Addis 2007).

Wenn wir uns an etwas erinnern, konstruieren wir es also gleichsam erneut. Das gilt sowohl für episodische Erinnerungen (Ereignisse) wie auch für Faktenwissen oder prozedurales Wissen (das ich hier in der Regel als Können bezeichne): Das Wissen wird im Moment seines Abrufs bearbeitbar. Dieser Effekt führt insbesondere im Umgang mit kognitivem Wissen und Können dazu, dass bestehendes Wissen ausgebaut oder angepasst (korrigiert) wird. Der gleiche Effekt wirkt aber auch im affektiven oder psychomotorischen Bereich (vgl. Schacter 2021, 351f.).⁴² Daraus resultiert ein scheinbares Paradoxon: Das Hirn kann umso schneller und umso mehr dazu lernen, je mehr es in einem bestimmten Bereich schon weiß oder kann.⁴³

Konsequenzen für das Lehren und Lernen an Hochschulen

In der Didaktik werden unabhängig von den Erkenntnissen der Hirnforschung schon lange Methoden empfohlen, die den Lernenden das Abrufen ihres Vorwissens erleichtern sollen, namentlich der Advance Organizer (Ausubel 1960; eine gute Erläuterung für den hochschuldidaktischen Einsatz findet sich in Wahl 2011). Die Erkenntnisse zum rekonstruktiven Charakter des Gedächtnisses liefern nun eine Erklärung dafür, weshalb es sinnvoll ist, das Vorwissen zu einem bestimmten Thema vor der Auseinandersetzung mit neuen Informationen zu diesem explizit zu aktivieren: Es wird damit für tiefgreifendere Anpassungen zugänglich. Dabei spielen vermutlich auch die Mechanismen von Vorhersageschleifen eine wesentliche Rolle.

Erhalten Lernende unmittelbar anschließend an eine Lernleistung ein Feedback, sind deshalb die Bedingungen besonders günstig, die angesprochenen Korrekturen oder Ergänzungen in das bestehende Wissen einzubauen. Feedback zu Lernleistungen ist somit nicht einfach eine Form von Korrektur oder Bestandsaufnahme, sondern eine ganz zentrale Lerngelegenheit – die leider im Lehr- und Lernalltag an Hochschulen oft ungenutzt bleibt (siehe Kap. 8).

42 Die hier geschilderte Veränderbarkeit von Wissen durch erneutes Aufrufen ist zwar grundsätzlich nützlich. Sie kann aber auch missbraucht werden, etwa indem Falscherinnerungen suggeriert werden, siehe dazu Schacter (2021).

43 Dennoch ist die Kapazität des Hirns grundsätzlich beschränkt. Siehe die exemplarische Aussage von Itzhak Perlman in Kapitel 4.6.

4.4.3 Schemata

Der Begriff „Schema“ wurde vor fast hundert Jahren durch Bartlett (1932) in die kognitive Psychologie eingeführt. Er bezeichnet einen mentalen Rahmen, der einen bestimmten Ausschnitt von Wissen oder Können ordnet. So haben in unserer Gesellschaft fast alle Menschen ein prozedurales Schema für den Besuch eines Restaurants, das Handlungen wie das Platznehmen, das Bestellen von Essen und Trinken oder das Bezahlen der Rechnung umfasst. Deklarative Schemata beinhalten demgegenüber abstraktere mentale Strukturen zur Organisation und Interpretation von Informationen. In einem Schema für Häuser könnten einige Attribute (und ihre möglichen Werte) wie folgt lauten: Material (Holz, Ziegel), Inhalt (Räume) und Funktion (menschliche Behausung). Schemata sind hierarchisch aufgebaut; sie sind mit übergeordneten Ideen (z. B. Gebäude) und untergeordneten Ideen (z. B. Dach) verbunden (vgl. Schunk 2020, 200).

Generell helfen uns Schemata, Informationen zu interpretieren und zu kategorisieren, indem sie einen mentalen Rahmen für ihre Einordnung bieten. „Durch die starke Vernetzung löst die Aktivierung eines Aspektes eines Schemas nämlich alle anderen Aspekte mit aus, so dass das Schema insgesamt zum Einsatz kommt. Durch Schemata wird schnelles Handeln und Urteilen möglich“ (Brendel, Hanke u. Macke 2019, 87). Dies ist im Alltag generell nützlich, kann aber auch zu Verzerrungen und Irrtümern führen, wie ein Experiment von Brewer u. Treyens (1981) zeigt. Dabei mussten die Teilnehmenden unter einem Vorwand kurz in einem Büro warten und wurden später gefragt, wie das Büro eingerichtet war. Die Ergebnisse zeigten, dass die Teilnehmenden bei der Erinnerung vom Schema einer typischen Büroeinrichtung ausgingen. So neigten sie dazu, sich an Gegenstände zu „erinnern“, die in Wirklichkeit nicht im Raum waren, wie etwa Bücher. Oder sie erinnerten sich nicht an untypische Gegenstände wie eine Bierdose. Dies zeigt, dass Schemata für den bereits in Kapitel 4.4.2 erwähnten konstruktiven Charakter des Gedächtnisses eine wesentliche Rolle spielen – als Grundlage für Vorhersageschleifen. Die Anwendung von Schemata ist ebenfalls ein Hauptgrund dafür, weshalb die Wortlisten von Aufgabe 1 in Kapitel 4.1 Falscherinnerungen provozieren können.

Konsequenzen für das Lehren und Lernen an Hochschulen

Schemata spielen auch eine wesentliche Rolle für akademisches Lernen. Beispielsweise ist betriebswirtschaftliches Fachwissen über das Konzept der Transaktionskosten als Schema abgespeichert, ebenso das sozialstatistische Wissen darüber, bei welcher Art von Daten und Fragestellungen man einen T-Test durchführen kann.

Fachwissenschaftliche Schemata sind nicht nur als Ziel des Lernens wichtig, sondern Voraussetzung für weitergehendes Lernen. Nur wer die Konzepte der Transaktionskosten und der begrenzten Rationalität grundsätzlich verstanden hat, kann den Zusammenhang zwischen den beiden Konzepten bedenken und in sein vertieftes Verständnis integrieren. Dieses Beispiel ist insofern typisch für Lernen an Hochschulen, als es dabei viel häufiger um die Weiterentwicklung und Kombination bestehender als um das Anlegen gänzlich neuer Schemata geht. Eine wesentliche Rolle spielt in diesem Zusammenhang auch die Metakognition, also das Bewusstsein für die eigenen Denkprozesse und dahinterliegende Muster. Zu unterscheiden ist einerseits das deklarative Wissen über das Denken und Lernen und andererseits die handlungsbezogene Regulation des eigenen Denkens und Lernens (vgl. Schunk 2020, 258–264).

Aspekte der Metakognition sind zentral für die Steuerung des Lernens und sollten deshalb insbesondere bei Feedbackprozessen angesprochen werden. Ein Arsenal an metakognitiven Prompts kann die Studierenden beim Aufbau ihres metakognitiven Fachwissens unterstützen, etwa: „Woher wissen Sie X? Anhand welcher Informationen/Daten stellen Sie einen Zusammenhang zwischen X und Y her?“ (in Kap. 8.3.2 und 8.3.3 finden Sie ausführlichere Erläuterungen und Beispiele zu metakognitiven Fragen). Auch Gruppenarbeiten von Studierenden und Peer-Feedback können zum Bewusstmachen und Vertiefen von metakognitivem Wissen beitragen (siehe Kap. 4.7 zum Argumentieren und zur Transaktivität).

4.5 Gefühle, Motivation und Bedeutsamkeit

Die bisherigen Ausführungen haben auf kognitive Aspekte von Lernen, Wissen und Können fokussiert. Dies ist aber in verschiedener Hinsicht einseitig, denn Kognitionen sind untrennbar mit Emotionen, Psychomotorik und weiteren durch das Gehirn gesteuerten Prozessen verbunden: „Sie können Emotionen nicht durch rationales Denken überwinden, denn der Zustand Ihres Körperhaushalts ist die Grundlage für jeden Gedanken und jede Wahrnehmung, die Sie haben (...). Selbst wenn Sie sich als rational erleben, sind Ihr Körperhaushalt und seine Verbindungen zum Affekt vorhanden und lauern unter der Oberfläche“ (Barrett 2017, 81, Übers. tz).

4.5.1 *Motivation und die Steuerung des Lernens*

Die Steuerung von Emotionen und Motivation funktioniert nach dem gleichen Prinzip, wie in Kapitel 4.4 für Sinneswahrnehmung und Kognition beschrieben: durch Vorhersageschleifen. Das ist sogar bei Gefühlen der Fall, auch wenn dies unserer Intuition widerspricht: Auch Gefühle werden zuerst durch Simulationen im Hirn erzeugt und anschließend mit sensorischen Körpersignalen abgeglichen und allenfalls korrigiert (vgl. Barrett u. Simmons 2015).⁴⁴ Im Zusammenhang mit Leistungsnachweisen und Leistungsbeurteilung besonders relevant sind Gefühlsaspekte in Bezug auf das Phänomen der Prüfungsangst. Entscheidend scheint für unseren Zusammenhang, dass Prüfungsangst erstens die Leistungen von Studierenden verschlechtert und zweitens durch Situationen mit hohem Druck verstärkt wird. Insbesondere Zeitdruck und reglementarische Vorgaben wie die beschränkte Wiederholbarkeit von Leistungsnachweisen verstärken die Prüfungsangst und behindern somit das Lernen von Studierenden. Diese komplexe Thematik wird in Kapitel 11 näher beleuchtet.

Auch die Motivation scheint durch Vorhersageschleifen gesteuert zu werden. Dies lässt sich an der zeitlichen Abfolge illustrieren, in der das Motivationshormon Dopamin ausgeschüttet wird. Mit diesem Hormon steuert unser Hirn maßgeblich, welche Ziele und Zustände wir anstreben oder eben nicht. Früher ging man davon aus, dass Dopamin als Belohnung für das Erreichen angestrebter Ziele ausgeschüttet wird. Inzwischen ist klar, dass die Dopaminausschüttung bereits dann einsetzt, wenn das Hirn das Erreichen des erwünschten Zustands annimmt, also simuliert (vgl. Sapolsky 2004, 337f.; Phillips et al. 2003; Schultz, Tremblay u. Hollerman 2000).

Dies erklärt auch sehr gut die Rolle, die Motivation beim Lernen spielt. In den Worten von Sapolsky ausgedrückt: „Wenn man weiß, dass der Appetit gestillt werden wird, dann hat das Vergnügen mehr mit dem Appetit als mit der Sättigung zu tun“ (Sapolsky 2004, 338, Übers. tz). Die eigentlichen Lernvorgänge sind oft anstrengend und nüchtern betrachtet nicht direkt vergnüglich. Was Menschen beim Lernen antreibt, ist also die erwartete Zielerreichung. Diese kann innerhalb des Lernens selbst liegen oder, mehr oder weni-

44 Wie stark die verschiedenen „Systeme“ zusammenhängen, zeigt sich z. B. darin, dass beim (scheinbar rein kognitiven) Lesen von Schilderungen taktiler Eigenschaften die gleichen Hirnareale aktiviert werden wie beim sensorischen Fühlen dieser Eigenschaften (vgl. Lacey, Stilla u. Sathian 2012). Gleiches gilt für das Lesen von Bewegungsbeschreibungen und die Aktivierung von motorischen Hirnarealen, (vgl. Pulvermüller et al. 2005). Man bezeichnet dies als „embodied cognition“ (Shapiro u. Spaulding 2021).

ger stark, außerhalb – je nachdem spricht man von intrinsischer oder extrinsischer Motivation:⁴⁵

- **Intrinsische Motivation:** Dieses Konstrukt beschreibt die natürliche Neigung von Menschen, sich für Aspekte ihrer Umwelt um ihrer selbst willen zu interessieren. Dazu gehört, die als interessant empfundenen Aspekte zu erkunden und weitgehende Fähigkeiten im Umgang mit ihnen zu entwickeln. Diese Neigung ist nicht nur für die kognitive und soziale Entwicklung der Menschen zentral, sondern auch für ihr körperliches und geistiges Wohlbefinden (vgl. Ryan u. Deci 2000, 70; Wild, Hofer u. Pekrun 2001, 221).
- **Extrinsische Motivation:** Dieser Begriff steht für eine Motivation, bei welcher der Antrieb zur Lernhandlung klar außerhalb dieser liegt. Das können materielle und immaterielle Anreize sein wie Geschenke und Lob, aber auch Druck und Zwang wie etwa die Androhung von Strafen (vgl. Ryan u. Deci 2000, 71; Wild, Hofer u. Pekrun 2001, 221). Noten können einerseits als Anreiz verstanden werden, wenn jemand gute Noten erreichen möchte. Sie wirken aber auch als Druckmittel und Zwang – wer sich Leistungsbewertung entzieht oder schlechte Noten erhält, dem werden Berechtigungen verweigert.

Dieser Unterschied erklärt auch, weshalb intrinsische Motivation in vielen Studien sich als stärkerer Prädiktor von tiefenorientiertem, also nachhaltigem Lernen erweist als extrinsische Motivation (siehe zur Unterscheidung zwischen oberflächlichem und tiefenorientiertem Lernen Kapitel 4.2): Wenn der Lernvorgang selbst mit positiven Zielen verbunden ist – nämlich mehr zu verstehen und seinen eigenen kognitiven, affektiven oder psychomotorischen Handlungsspielraum zu erweitern –, dann sind auch langwierige Lernprozesse mit der steten Aussicht auf die Zielbefriedigung verbunden. So können intrinsisch Motivierte ihr zielgerichtetes Handeln auch dann aufrechterhalten, wenn sie beim Verfassen schriftlicher Arbeiten oder Durchführen aufwändiger Experimente mit Durststrecken und Umwegen konfrontiert sind. Insbesondere dürfte jeder festgestellte Zuwachs an Erkenntnis oder Können eine innere Befriedigung, also eine Belohnungsschleife sowie weiteren Antrieb auslösen: „das kann ich“, „das habe ich verstanden“, „sehr interessant, darüber möchte ich mehr wissen“ etc. Verwandt mit intrinsischer Lernmotivation dürfte deshalb das Interesse am betreffenden Lerngegenstand sein (vgl. Schiefele 2000).

45 Die beiden Begriffe der intrinsischen und extrinsischen Motivation werden in der Fachliteratur nicht überall gleich verwendet und gelegentlich auch kritisiert. Sie werden deshalb hier als Pole auf einer Skala und nicht als zwei klar trennbare Kategorien von Motivation verstanden (vgl. Ryan u. Deci 2000, 71–73).

Demgegenüber ist der Anreiz bei extrinsischer Motivation weiter entfernt: Weil der Anreiz außerhalb des eigentlichen Lernvorgangs liegt, fällt es wesentlich schwerer, aufwändige Lernprozesse konsequent zu verfolgen – die inhaltlichen Lernerfolge werden nicht wie bei der intrinsischen Lernmotivation per se als Erfolgserlebnis kodiert. Insbesondere dort, wo summative Leistungsnachweise mit reglementarischen Verschärfungen verbunden sind (z. B. bei wiederholtem Nichtbestehen erfolgt Studienausschluss), zielen sie nicht nur auf extrinsische Motivation, sondern korrumpieren durch den übermäßigen Druck auch die intrinsische Lernmotivation. Das Problem ist in wissenschaftlichen Kreisen längst bekannt, wird aber in der Praxis eher weiter verschärft, denn zu lösen versucht. Der Grund dafür dürfte in der in Kapitel 2 geschilderten, in unserer Gesellschaft vorherrschenden Orientierung an Leistung und ihrer vermeintlichen metrischen Messbarkeit zu suchen sein. So schrieb Walter Edelman bereits vor Jahrzehnten in seinem Lehrbuch zur Lernpsychologie:

Nicht wenige Lehrer und Studierende (...) glauben nicht an die Bedeutung der intrinsischen Lernmotivation. (...) Krause (1995) untersuchte, welche Ziele Studierende beim Besuch von Lehrveranstaltungen verfolgen. In erster Linie sind dies: „...um dort ‚Scheine‘ [heute: ECTS-Punkte, tz] zu erwerben, sich mit prüfungsrelevanten Lehrinhalten zu beschäftigen, allgemein schneller das Studium beenden zu können (...)“ (S. 133). (...) Hier zeigt sich das Ergebnis einer dreizehnjährigen schulischen Sozialisationsgeschichte. Irgendwann begehen Elternhaus und Grundschule einen tragischen Fehler, der von den weiterführenden Schulen noch potenziert wird. Intrinsische Motivation, d. h. an Sachen interessiert sein, dabei Spaß haben, sich in Leistungssituationen anstrengen und über eine optimistische Erwartung verfügen, ist eine (...) kostbare Gabe (...). Das Gegenmodell ist die Außensteuerung des Lernalters durch Belohnung oder Zwang. (Edelman 2005, 258)

Dennoch ist es wichtig, intrinsische und extrinsische Motivation nicht als unvereinbare Gegensätze zu verstehen, sondern als Abstufungen.

4.5.2 Die Selbstbestimmungstheorie der Motivation

Im Rahmen ihrer Selbstbestimmungstheorie betonen Ryan u. Deci (2017, 180–82), dass extrinsische Formen von Motivation über Integrationsprozesse schrittweise verinnerlicht werden können. Entscheidend für diese Integration sind drei psychologische Grundbedürfnisse (vgl. Ryan u. Deci 2017, 94–98; 2000, 67; 73f.):

- **Wirksamkeit:** Bei diesem Bedürfnis geht es um das Gefühl, wirksam mit der Umwelt interagieren zu können und seine eigene Entwicklung aktiv beeinflussen zu können. Damit Menschen dieses Gefühl entwickeln, müssen sie sich für Handlungen verantwortlich fühlen, bei denen sie erfolgreich sind. So zeigen empirische Studien: Wenn Menschen eine Aufgabe erfolgreich bewältigen, für die sie kein Gefühl der Selbstinitiative em-

pfinden, dann trägt dies nicht zu ihrem Wirksamkeitsgefühl oder zur Erhöhung intrinsischer Motivation bei.

- **Soziale Verbundenheit:** Menschen benötigen ein Zugehörigkeitsgefühl zu anderen Menschen. In den sozialen Interaktionen mit ihren Mitmenschen ist es ihnen wichtig, von ihren Gegenübern Wertschätzung und Sensibilität für ihre Bedürfnisse zu erfahren. Viele menschliche Verhaltensweisen sind auf dieses Grundbedürfnis zurückzuführen, von der Kleidung bis zum Demonstrieren von Leistungsfähigkeit. Aus dem Bedürfnis nach Verbundenheit heraus interessieren sich Menschen auch für fremde Ansichten und Haltungen und sind bereit, diese in ihre Sichtweise aufzunehmen. Entscheidend dafür, ob Menschen echte soziale Verbundenheit empfinden, ist das Gefühl, so angenommen zu werden, wie sie sind.
- **Eigenständigkeit:** Dieses Grundbedürfnis bezieht sich auf die Selbstregulation des eigenen Verhaltens und damit eine Erfahrung von Ganzheitlichkeit im Umgang mit der Umwelt und sich selbst. Damit die beiden zuvor genannten Grundbedürfnisse erlebt werden können, ist das Gefühl der Eigenständigkeit notwendig. Dieses Gefühl stellt sich ein, wenn Menschen ihr Verhalten als selbstbestimmt erleben und nicht als von außen gesteuert durch Druck, Zwang oder Anreize. Wenn Menschen mit ihrem vollen Willen handeln, dann bringen sie ihre gesamten Ressourcen in die Handlung ein. Im Gegensatz dazu ermöglichen stärker als außengesteuert erlebte Motivationszustände weniger Zugang zu den kognitiven, affektiven und physischen Möglichkeiten.

Eine wichtige Ergänzung betrifft den Unterschied zwischen Eigenständigkeit und Unabhängigkeit (Englisch: *autonomy* versus *independence*): Auch Personen, die sich als autonom erleben, befinden sich in menschlichen Beziehungen, in denen sie von anderen abhängig sind, und sind mit Gruppen von anderen Menschen vernetzt. Eigenständigkeit ist deshalb nicht mit Unabhängigkeit zu verwechseln, denn diese Vorstellung schließt die soziale Verbundenheit weitgehend aus.⁴⁶

Nachhaltige, selbstbestimmte Motivation entsteht folglich dann, wenn eine Handlung im Sinne einer Befriedigung der drei Grundbedürfnisse Wirksamkeit, soziale Verbundenheit und Eigenständigkeit empfunden wird. So können auch extrinsisch gesetzte Zielhorizonte schrittweise in die internalisierte Motivation überführt werden.

Wichtig für institutionell organisiertes Lernen ist aber auch der Umkehrschluss: Die Durchführung einer Handlung genügt noch nicht, um etwas daraus zu lernen – auch nicht, wenn sie wiederholt durchgeführt wird (vgl. dazu

46 Ähnliches gilt für unreflektierte Vorstellungen von „freiem Willen“, den es weder neurologisch noch philosophisch betrachtet in absolutem Sinne geben kann (vgl. Krüger 2007).

Kap. 4.6 zum Einfluss von Zeit und Erreichen von Expertise). Entscheidend ist, dass die Tätigkeit und/oder ihr Resultat als bedeutsam empfunden werden. Dann erfolgen neurologische Vorgänge, welche die betroffenen Stellen im neuronalen Netzwerk des Hirns plastisch und somit offen für Veränderung machen (vgl. Eagleman 2020, 149–53).⁴⁷ Um noch einmal das Bild von Eagleman aus Kapitel 4.2 zu verwenden: Wenn wir einen bestimmten Wegabschnitt zwischen den beiden Dörfern als besonders bedeutsam wahrnehmen, sorgt das Hirn dafür, dass dieser Abschnitt sehr schnell ausgebaut wird. Unter Umständen kann ein einmaliges Zurücklegen der Strecke eine Asphaltierung bewirken, z. B. wenn ein Kind lernt, dass bestimmte Bonbons lecker schmecken. Im kognitiven Bereich dürften so genannte „Aha-Erlebnisse“ am ehesten eine solche schnelle Konsolidierung einer neuen Erkenntnis bewirken.

Konsequenzen für das Lehren und Lernen an Hochschulen

Zentral für den Aufbau von Expertenwissen, wie es in einem Hochschulstudium angestrebt wird, sind aber längerfristige Vorgänge, bei denen Schritt für Schritt immer mehr und immer stärker vernetztes Wissen und Können aufgebaut wird. Dazu müssen die notwendigen Handlungen als grundsätzlich bedeutsam erlebt werden. Am ehesten ist dies der Fall, wenn im Rahmen des Studiums die drei menschlichen Grundbedürfnisse der Wirksamkeit, der sozialen Verbundenheit und der Eigenständigkeit als erfüllt empfunden werden (vgl. Ryan u. Deci 2000, 73f.).

In diesem Zusammenhang erweist sich das Konstrukt der Selbstwirksamkeit (entwickelt von Bandura 1977) als wichtig. Es bezieht sich auf den Glauben einer Person an ihre eigene Fähigkeit, bestimmte Aufgaben zu bewältigen oder Ziele zu erreichen. Selbstwirksamkeitsüberzeugungen können sehr spezifisch sein und sich auf bestimmte Themen- und Fachgebiete beziehen. Sie sind Teil der fachbezogenen Motivationsstruktur von Lernenden, und insbesondere die Überzeugungen von Studierenden bezüglich ihrer fachlichen Leistungsselbstwirksamkeit haben einen sehr großen Einfluss auf ihren Lernerfolg (vgl. Schneider, Michael u. Preckel 2017, 568; 591). Durch meta-kognitives Feedback können Selbstwirksamkeitsüberzeugungen gefördert werden (Theobald 2021, 16; siehe Kap. 8.3.2 und 8.3.3).

47 Die genauen neurobiologischen Zusammenhänge werden immer noch erforscht. Eine zentrale Rolle spielt sicher der Neurotransmitter Acetylcholin. Dieser wird ausgeschüttet als Teil von Vorhersageschleifen, welche die Aktivitäten des Hirns regulieren, indem sie Relevanz/Bedeutsamkeit erkennen. Acetylcholin bewirkt in Kombination mit anderen Neuromodulatoren die beschriebene verstärkte Formbarkeit der betroffenen Hirngebiete (vgl. Eagleman 2020, 149–153).

Auf der negativen Seite ermöglicht die Bedeutsamkeit auch eine Erklärung des Phänomens „Gaming the System“, das in Kapitel 2.4 im Zusammenhang mit der metrischen Fixierung der Leistungsorientierung erläutert wurde: Zahlenbasierte Anreizsysteme, die außerhalb der eigentlich durch sie zu steuernden Tätigkeit liegen, schaffen extrinsische Motivation. Dies gilt für Fallzahlen in Krankenhäusern, Abschlussquoten in Schulen oder den h-Index von Forschenden. Zum Ziel wird schließlich das gute Abschneiden im Anreizsystem, das meistens auch relativ unabhängig von der Qualität der eigentlich zu steuernden Tätigkeit erreicht werden kann (wobei die verwendeten Taktiken legal, grenzwertig oder betrügerisch sein können).⁴⁸

Gaming the System kann auch im Zusammenhang mit (hoch-)schulischen Bewertungen und Noten auftreten. Im Zusammenhang mit der Bewertung von Lernleistungen ist deshalb die Frage entscheidend, ob das Leistungsnachweis-Setting und die Bewertung tatsächlich die erwünschten Lernvorgänge bewirken. Diese Frage spielt für das Alignment von Lernzielen und Leistungsbeurteilung (Kap. 5 und 7–9) sowie die Gütekriterien von Leistungsbewertungen (Kap. 6.3) eine wesentliche Rolle.

4.6 Aufbau komplexer Wissensstrukturen: Zeit und Expertise

Wie bereits angesprochen benötigt der Aufbau von komplexen Wissensstrukturen Zeit und damit substanzielle Energie, die in ihn investiert werden muss. Verschiedene Aspekte des Zusammenhangs von Lernen und Zeit sind für die Gestaltung von Leistungsnachweisen und den Umgang mit Beurteilungssituationen relevant. Dies gilt für die zeitliche Verteilung von Leistungsnachweisen im Curriculum, zeitliche Vorgaben für die Bearbeitung von einzelnen Leistungsnachweisen oder auch die Wirkung von Feedbackprozessen.

4.6.1 Lernen braucht Zeit

Lernen ist kein Vorgang, der in dem Moment abgeschlossen ist, indem ein Mensch zum ersten Mal eine neue Information oder Erkenntnis kognitiv repräsentiert (also „denkt“). Vielmehr sind insbesondere bei komplexeren Lernvorgängen, wie sie beim an Hochschulen angestrebten Lernen stattfinden

48 Es dürfte auch Fälle geben, in denen „Gaming the System“ mit der Absicht betrieben wird, sich mehr Spielraum für die Erledigung der eigentlich als sinnvoll empfundenen Aufgaben zu verschaffen. Das ändert aber nichts daran, dass die metrischen Kontrollsysteme nicht ihre vorgebliche Wirkung erzeugen.

den, verschiedene Prozesse im Spiel, die sich zeitlich verteilen. So werden wie bereits erwähnt neue Informationen oder Erkenntnisse besser in das bestehende Wissen und Können eingebaut, wenn die vorhandene Struktur zuvor bewusst aktiviert wurde. Ebenfalls besser eingebaut werden können Elemente, die mehrfach und dabei mindestens teilweise auf verschiedene Weise verarbeitet wurden. Und wenn ein Lernvorgang stattgefunden hat, spielen auch Pausen und Interferenzen – also konkurrierende Lernvorgänge – eine wichtige Rolle: Pausen sind wichtig, damit sich das Gelernte verfestigen kann, und Interferenzen behindern diese Verfestigung. Die folgenden Prozesse und ihre Bedeutung sind empirisch gut untersucht:

- **Wichtigkeit von Wiederholungen:** Schon der Begründer der experimentellen Gedächtnisforschung, Hermann Ebbinghaus, stellte Ende des 19. Jahrhunderts den erheblichen Einfluss zeitlicher Abläufe auf das Lernen fest: Bei seinen Selbstexperimenten mit bedeutungslosen Silben machte er desto weniger Fehler, je häufiger er die Listen lernte. Oder er vergaß die Silben einer bestimmten Liste am Anfang relativ schnell, mit wiederholten Lernvorgängen aber immer langsamer (vgl. Schunk 2020, 8).
- **Verarbeitungstiefe:** Wird etwas nur oberflächlich zur Kenntnis genommen, wird es weniger tief verarbeitet, als wenn es zum Beispiel in Bezug auf seine Bedeutung reflektiert wird. Dies kann man beim Lehren und Lernen gezielt nutzen und ist „ein hervorragendes Beispiel dafür, wie man aktiv die Netzwerkstruktur ausnutzen kann. Durch das ‚vertiefte Verarbeiten‘ erzwingt man praktisch, dass ein weit verzweigtes Netzwerk beim Lernen aktiv wird. Man verkoppelt verschiedene Informationen miteinander, wobei insbesondere bereits gespeicherte Informationen als Anknüpfungspunkte für das Gedächtnis genutzt werden“ (Jäncke 2018, 139f.). Die Verarbeitungstiefe ist also proportional zu den erstellten oder aktivierten Verknüpfungen im Gehirn. So ist es beim Vokabellernen von Vorteil, nicht nur die muttersprachliche Übersetzung eines neuen Wortes auswendig zu lernen, sondern sich auch zu überlegen, mit welchen anderen Wörtern der Zielsprache es verwandt ist und sogar Sätze damit zu bilden. Das Erhöhen der Verarbeitungstiefe ist in der Regel kurzfristig mit größerem Zeitaufwand verbunden, wobei dieser durch den größeren Lerngewinn langfristig aber wieder wettgemacht wird.⁴⁹

Bezogen auf Unterricht ist zu beachten, dass die Menge an neuen Informationen, die tiefenorientiert verarbeitet werden können, beschränkt ist. Unterricht, der die Studierenden innerhalb von kurzer Zeit mit vielen

49 Ein früher in der Mediendidaktik oft besprochener Spezialfall der Verarbeitungstiefe ist die Theorie der dualen Codierung nach Paivio (1971). Diese geht davon aus, dass Begriffe und Konzepte besser behalten werden können, wenn sie sowohl sprachlich als auch bildlich verarbeitet werden.

neuen Informationen konfrontiert, führt deshalb zu oberflächlichem Lernen. Entscheidend ist somit ein Austarieren zwischen der Menge neuer Informationen und der für ihre Verarbeitung zur Verfügung stehenden Zeit (siehe Abb. 10).

Die Verarbeitungstiefe ist auch wesentlich für die Wirkung von Leistungsnachweisen. Bestehen Leistungsnachweise aus Aufgaben, die sich ohne große Verarbeitungstiefe bewältigen lassen, so werden sich die Studierenden auch nur mit oberflächlichem Lernen darauf vorbereiten – insbesondere dann, wenn sie extrinsisch motiviert sind und in kurzer Zeit große Informationsmengen verarbeiten müssen (vgl. Biggs u. Tang 2007, 204; siehe dazu ausführlicher Kap. 5.1). So wurde bereits in den 1960er Jahren festgestellt, dass Studierende in einer derartigen Leistungsnachweiskultur einen großen Teil des Gelernten wieder vergaßen – wohl-gemerkt auch jene, welche gute Bewertungen erhalten hatten. So bemerkte Shulman (vgl. 1970, 95) zur Lernleistung von Medizinstudierenden in Makroanatomie, dass sie bezüglich des Gelernten eine Vergessenskurve aufwiesen, die jener entsprach, die Ebbinghaus in den 1880er Jahren als Vergessenskurve für bedeutungslose Silben identifiziert hatte.

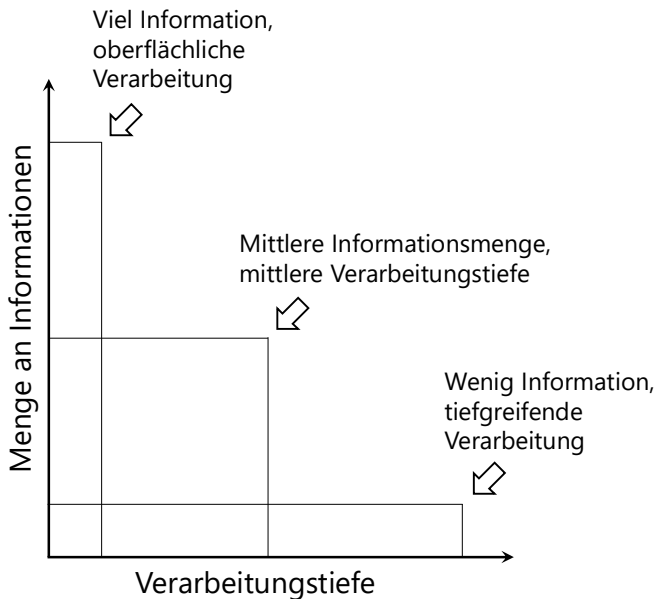


Abb. 10: Zusammenhang von Informationsmenge und Verarbeitungstiefe bei gleichbleibender Lernzeit. Quelle: Eigene Abbildung

- **Konsolidierung:** Wenn einmal etwas gelernt worden ist, braucht das Hirn Zeit, um das Gelernte zu verfestigen. Am besten wird das Gelernte verfestigt, während keine weiteren Lernprozesse stattfinden. Dann kann ein Hirnareal namens Hippocampus die neuen neuronalen Erregungsmuster wiederholt aktivieren. Dieser Mechanismus scheint am besten im Schlaf zu funktionieren; vermutlich stellt die Konsolidierung von Gedächtnisinhalten sogar eine biologische Funktion von Schlaf dar (vgl. Eagleman 2020, 215f.; Gottselig et al. 2004). Die Erhöhung von Lernzeit auf Kosten von Schlaf lohnt sich deshalb nicht, sondern verschlechtert das Lernen.
- **Interferenzen:** Wenn das Hirn neue Informationen verarbeitet und mit weiteren neuen Informationen konfrontiert wird, bevor es pausieren und die neu angelegten Erregungsmuster konsolidieren kann, dann behindern sich diese beiden Informationen gegenseitig. Man bezeichnet dies als Interferenzen (vgl. Gottselig et al. 2004).
- **Verteilte Repetition:** Die Wichtigkeit von Wiederholungen für das Lernen wurde bereits erwähnt. Zentral ist dabei die zeitliche Verteilung der Wiederholungen. Verteilte Wiederholungen bewirken deutlich höhere Behaltensleistungen als massierte Wiederholungen. Dies liegt in einem neurologischen Mechanismus begründet, der eigentlich nützlich ist, nämlich der Automatisierung (siehe dazu den nächsten Abschnitt zur Expertise): Wird neues Wissen und Können in kurzem Abstand mehrfach wiederholt, ist es so griffbereit, dass das Hirn wenig bewusster Anstrengung bedarf, um es abzurufen (vgl. Schacter 2021, 85f.; Wagner, Maril u. Schacter 2000). Etwas zugespitzt könnte man sagen, dass die Wirkung von unmittelbar erfolgenden Wiederholungen beim Lernen exponentiell abnimmt.

Das bedeutet, dass die Repetition von Lerneinheiten mehr bringt, wenn sie zeitlich verteilt erfolgt. Der renommierte Gedächtnisforscher Daniel Schacter empfiehlt etwa: „Wenn Sie zum Beispiel für eine Prüfung lernen wollen, die Sie in einer Woche ablegen werden, und den Stoff zehnmal durchgehen können, ist es besser, die zehn Wiederholungen über die Woche zu verteilen, als sie alle direkt nacheinander durchzuführen“ (Schacter 2021, 86, Übers. tz). Dieses Prinzip ist empirisch bestens belegt: für das Lernen von Kindern und Erwachsenen, für kognitive und psychomotorische Lernprozesse, für einfache und komplexe Lerninhalte (vgl. z. B. Cepeda et al. 2008; Spruit, Band u. Hamming 2015; Vlach u. Sandhofer 2012).

Der Effekt des verteilten Wiederholens dürfte einer von diversen Gründen sein, weshalb formative Leistungsbewertung mit höheren Lerngewinnen einhergeht: Das wiederholte Aufgreifen von Lernresultaten in einem Feedbackprozess führt zu wiederkehrenden Aktualisierungen der Lerninhalte und somit zu nachhaltigerem Lernen. Summative Leistungsnachweise, insbesondere gewichtige Prüfungen, bewirken hingegen oft

sogenanntes Bulimielernen: Die Studierenden versuchen in kürzester Zeit möglichst viel Faktenwissen aufzunehmen (vgl. Gast 2009, 160), das sie nicht nur oberflächlich verarbeiten, sondern mangels verteilter Wiederholung auch schnell wieder vergessen. Wie in Kapitel 10 erläutert wird, ist deshalb bei der curricularen Planung von Leistungsnachweisen gut zu bedenken, welches Lernverhalten bei den Studierenden bewirkt werden soll.

4.6.2 Langfristiger Aufbau von Expertise

Studiengänge an Hochschulen zielen auf ein umfassendes Verständnis des jeweiligen Fachgebiets (mit zunehmender Spezialisierung vom Bachelor bis zum Doktorat). In diesem Sinne sollen Studierende Expertise erwerben, also ein tiefgreifendes Verständnis der studierten Fachgebiete, das auch kritisches Hinterfragen und kreatives Weiterentwickeln dieser Wissensbestände umfasst. So zeichnen sich Expertinnen dadurch aus, dass sie im Vergleich zu Novizen fachliche Probleme aufgrund ihrer Tiefenstruktur einschätzen können, während Novizen sich oft an oberflächlichen Merkmalen orientieren. Dies ist möglich, weil Expertinnen wesentlich umfassendere Schemata für das Fachgebiet aufgebaut haben als Novizen. Mit diesem Unterschied hängt auch zusammen, dass Expertinnen solche Probleme durch vorwärts orientierte Strategien lösen können, indem sie Zwischenziele setzen und erreichen. Diese Strategien laufen teilweise automatisiert ab und sind somit nur durch gezielte Reflexion dem Bewusstsein wieder zugänglich zu machen (siehe das Phänomen der Expertenblindheit in Kap. 8.4.1). Diese Automatisierung dient der kognitiven Entlastung und ist ein wesentliches Merkmal von Expertise in einem Gebiet. Demgegenüber versuchen Novizen häufig, das Problem rückwärts von der Lösung her zu bearbeiten, wobei sie oft stecken bleiben (vgl. Schunk 2020, 257f.; 284f; 307f.).

Der Weg von einem (hoffentlich) breiten Allgemeinwissen, das Studienanfänger mitbringen, bis zur Expertise in einem Gebiet ist wesentlich durch folgende Merkmale gekennzeichnet:

- Viel Zeit: Die Aneignung von Expertenwissen erfordert einen erheblichen Zeitaufwand und viel Übung. Untersuchungen haben gezeigt, dass es tausende von Stunden bewusster Übung bedarf, um Expertin in einem etablierten Expertisegebiet zu werden – allerdings reichen reine Wiederholungen bei weitem nicht aus, vielmehr ist das gezielte Üben entscheidend.⁵⁰ Eine einfache Überlegung: Wenn die reine Wiederholung genü-

50 In der Populärwissenschaft zirkuliert die 10.000-Stunden-Regel, die von Malcolm Gladwell (2008) aufgestellt wurde: Wer in einem Gebiet so lange übe, werde zum Experten darin. Er bezog sich dabei auf die Forschung von Ericsson

gen würde, dann könnte es jede Hobbypianistin zur Meisterschaft bringen, nur weil sie im Laufe des Lebens tausende von Stunden des Piano-spielens anhäuft. Spielt sie aber immer die gleichen Stücke oder Stücke des gleichen Schwierigkeitsgrades und auf die gleiche Weise, wird sie ihr Spiel kaum weiterentwickeln.

- Wie erwähnt basiert Expertenwissen auf komplexen mentalen Repräsentationen, die sowohl deklaratives als auch prozedurales Wissen umfassen (siehe Kap. 4.4.3). Diese Repräsentationen versetzen Expertinnen in die Lage, Probleme in ihrem Fachgebiet schnell zu erkennen und zu lösen. Möglich macht dies ein Vorgang namens *Chunking*: Damit wird die Gruppierung von Informationen in größere, sinnvolle Einheiten bezeichnet. Dadurch können Expertinnen leichter große Wissensstrukturen im Arbeitsgedächtnis bearbeiten, wodurch kognitive Ressourcen für komplexere Aufgaben frei werden (vgl. Anderson 1982, 1992; Sweller 1988). Ein einfaches Alltagsbeispiel ist das Folgende: Versuchen Sie, sich diesen Pincode zu merken: 178919542014. In der Fußnote finden Sie einen Hinweis, wie das relativ einfach geht.⁵¹ Ein Expertenbeispiel für *Chunking* sind Schachmeisterinnen, die sich ganze Schachkonstellationen auf einen Blick merken können. Und Weltklassemenschachspieler können zeitgleich mehrere Partien gegen verschiedene Gegner rein in Gedanken spielen, ohne auf ein Schachbrett zu blicken (Ericsson u. Pool 2016, 88–98).
- Automatisierung und Routinisierung von Handlungen: Expertise beinhaltet die Fähigkeit, komplexe Aufgaben automatisch und effizient auszuführen, ohne dass dies bewusste Aufmerksamkeit oder Anstrengung erfordert, um dadurch noch komplexere, übergeordnete, eventuell auch neuartige Probleme zu lösen. Diese Automatisierung wird durch gezieltes Üben erreicht, das auf die oben erwähnten Prozesse der tiefgehenden Verarbeitung und zeitlich verteilte Wiederholungen setzt sowie auf die Verarbeitung von internem Feedback (Selbstregulation) und externem Feedback (siehe Kap. 8). Mit dem Bild von Eagleman können wir sagen: Es werden Autobahnen gebaut, also neuronale Verbindungen, die so stark sind und so schnell aktiviert werden können, dass die Aktivierung ohne

(v. a. Ericsson, Krampe u. Tesch-Römer 1993). Dieser selbst weist die Stundenzahl als beliebig und die Regel als vereinfacht zurück (Ericsson u. Pool 2016, 161–67). Das bestätigt auch eine Meta-Analyse, der zufolge gezieltes Üben nur einer von mehreren für die Expertisebildung verantwortlichen Faktoren ist (Macnamara, Hambrick u. Oswald 2014).

- 51 Die 12 Ziffern lassen sich in drei Chunks aus vierstelligen Jahreszahlen unterteilen: Französische Revolution (Jahr 1789), erster (1954) und bisher letzter Fußballweltmeistertitel (2014) Deutschlands. Natürlich ist das Beispiel von den Chunks ausgehend konstruiert. Umgekehrt versuchen wir uns z. B. Telefonnummern durch die Bildung von besser einprägsamen Chunks zu merken.

bewusste Anstrengung erfolgen kann. Dies entspricht einer Art von mentalem Programm, das automatisch und ohne bewusste Anstrengung ausgeführt werden kann (vgl. Anderson, John R. 1992; Schunk 2020, 202f.). Aus prozeduralem wird somit automatisches Wissen.

- **Intrinsische Motivation:** Sie ist ein Schlüsselfaktor bei der Entwicklung von Expertenwissen (siehe Kap. 4.5). Wenn die Lernenden die Lern- und Übungsaktivitäten als persönlich sinnvoll und lohnend im Sinne der Selbstbestimmungstheorie empfinden, ist es weitaus wahrscheinlicher, dass sie mit der nötigen Konsequenz und dem erforderlichen Zeitaufwand bewusst üben und ihr Fachwissen weiter ausbauen. Auftretende Schwierigkeiten und Durststrecken werden dann als Herausforderungen und Weiterentwicklungsgelegenheiten empfunden statt als Hindernisse. Eagleman hat dazu ein illustratives Beispiel entwickelt: Stellen wir uns vor, die erfolgreichen Tennisspielerinnen Serena und Venus Williams hätten noch einen Bruder namens Fred. Auch er hätte die gleiche, gänzlich auf Tennis ausgerichtete Kindheit durchlebt wie seine beiden Schwestern. Im Gegensatz zu ihnen hätte er aber Tennis gehasst. Entsprechend empfand er Feedback von Eltern, Trainern oder anderen Spielern nicht als bedeutsam. Fred gewann nie Kinderturniere und die Eltern überhäufte ihn nicht mit Lob. Von außen gesehen hat er somit das gleiche Trainingsprogramm durchlaufen wie die beiden Schwestern, aber es stimmte nicht mit seiner intrinsischen Motivation und seinen Interessen überein. Da Fred keine Bedeutsamkeit im Tennisspielen erkannte, wurde seine Neuromodulation nicht aktiviert und in seinem Hirn und Zentralnervensystem fanden kaum Anpassungsleistungen statt, wie wir sie in den Kapiteln 4.2 bis 4.4 und in diesem Unterkapitel diskutiert haben (vgl. Eagleman 2020, 147 und 152). Gleichmaßen ist ein Humorpsychologe unvorstellbar, der nicht von Humor fasziniert ist. Die heißt übrigens nicht, dass er ein besonders lustiger Zeitgenosse sein muss, so wie eine professionelle Fußballtrainern keine brillante Fußballerin zu sein braucht – aber von Fußball fasziniert sein muss. Kurzum: Die Entwicklung von Expertenwissen ist nur mit intrinsischer Motivation denkbar.

Abschließend seien einige Einschränkungen und Ergänzungen angebracht: Das Spektrum an Fähigkeiten, die sich Menschen aneignen können, ist beeindruckend. Dennoch heißt das in mehrfacher Hinsicht nicht, dass Menschen alles lernen können, was sie wollen:

Einerseits ist Expertisebildung rein physisch bzw. physiologisch nicht in zig völlig verschiedenen Gebieten möglich. Die Lebenszeit reicht nicht aus, um sowohl die nötige Übung zur Berufsgeigerin, zur professionellen Tennisspielerin und zur doktorierten Teilchenphysikerin anzuhäufen – zumindest nicht völlig zeitgleich. Am ehesten ergänzen sich psychomotorische Exper-

tise auf der einen und kognitive Expertise auf der anderen Seite.⁵² Für die meisten von uns bedeutet aber Expertisebildung, sich auf ein Interessengebiet festzulegen, in dem wir uns spezialisieren und weiterentwickeln. Eagleman (vgl. 2020, 140) berichtet dazu eine Geschichte, die über den berühmten Geiger Itzhak Perlman erzählt wird: Nach einem Konzert habe ein Bewunderer zu ihm gesagt: „Ich würde mein Leben dafür geben, so spielen zu können.“ Darauf habe Perlman erwidert: „Das habe ich getan“.

Andererseits haben wie bereits erwähnt nicht alle Menschen die gleichen Möglichkeiten, um Expertise zu erwerben. Dafür spielen vor allem äußere Faktoren eine Rolle: Tausende Stunden Geige üben, Tennis trainieren oder ein Hochschulstudium absolvieren können nur Kinder/Jugendliche/junge Erwachsene, deren Familien über die entsprechenden finanziellen Mittel verfügen oder die gesellschaftliche Unterstützung durch Sportverbände, Stipendien etc. erhalten (siehe die Erläuterungen in Kap. 2 zum Einfluss des sozioökonomischen Status auf den Bildungserfolg). Daneben gibt es auch noch schwieriger zu erfassende Umstände wie die psychosoziale Unterstützung, die junge Menschen durch Familie und Umfeld erhalten, wenn sie sich in einem Interessengebiet vertiefen möchten. Dabei werden oft auch vermeintlich angeborene Fähigkeiten wie Intelligenz und Talente als Gründe angeführt – wissenschaftliche Forschungen zeigen freilich, dass der Einfluss von Genen allein ziemlich gering ist.⁵³ In den Worten von Sapolsky (2018, 248, Übers. tz): „Es ist nicht sinnvoll zu fragen, was ein Gen bewirkt, sondern nur, was es in einer bestimmten Umgebung bewirkt“. Für die Bildung von Expertise, die verhältnismäßig spät im Leben geschieht – wenn das Hirn ziemlich ausgereift ist –, gilt dies umso mehr. Denn je später eine Hirnregion heranreift, desto weniger wird sie von Genen und umso mehr durch Wirkungen der Umwelt geprägt (vgl. Sapolsky 2018, 175).⁵⁴

52 Dies zeigt sich bei gewissen Spitzensportlerinnen, die ihre Sportkarriere mit einem Studium kombinieren. Ein erstaunliches Beispiel stellt etwa die Schweizerin Marlen Reusser dar, die 2021 im Alter von 30 Jahren bereits promovierte Ärztin, professionelle Radsportlerin, Silbermedaillengewinnerin bei den Olympischen Spielen und ehemalige Präsidentin einer Kantonalpartei war (vgl. Wikipedia).

53 Ein illustratives Beispiel ist die GWAS-Studie von Rietveld et al. (2013), die nach Genen suchte, die mit Bildungserfolg korrelieren. Eine einzelne Genvariante erklärt ungefähr 0.02 % (also 2 Tausendstel!) des Bildungserfolgs, alle identifizierten Genvarianten gemeinsam erklären rund 2 % des Bildungserfolgs. Mit anderen Worten: 98 % des Bildungserfolgs einzelner Menschen sind gemäß dieser Studie nicht genetisch erklärbar.

54 Wesentliche Begründungen liefert einerseits das Gebiet der Epigenetik, das die Regulation („Ein- und Ausschalten“) von Genen untersucht. Es liefert eine weitere Bestätigung dafür, dass Umwelteinflüsse während der kindlichen Entwicklung (bereits pränatal) entscheidend sind für das spätere Verhalten im Erwachse-

4.7 Argumentatives Schlussfolgern als sozialer Prozess

Mit Ausnahme der Bedeutung von externem Feedback hat sich das vorliegende Kapitel bisher auf das Lernen des Individuums konzentriert. Damit wurde ein zentraler Faktor ausgeblendet: Menschen sind soziale Wesen, und wenig überraschend ist auch ihr Lernen ganz wesentlich durch soziale Prozesse geprägt – sowohl im engeren als auch im weiteren Sinne.

Grundsätzlich hat jedes Lernen, das auch nur ansatzweise in einem institutionellen Kontext verortet ist, soziale Anteile. Bereits wer ein Buch liest, nimmt an einer – sehr indirekten – Form von sozialem Austausch teil: Bücher werden von Menschen für Menschen geschrieben, auch wenn Autorinnen und Leser in der Regel keinen direkten Kontakt zueinander haben. Selbstlerneinheiten werden von Dozierenden für Studierende gestaltet, Ähnliches gilt für relativ anonyme Verfahren wie Aufnahmeprüfungen. Im Folgenden geht es um soziales Lernen im engeren Sinne. Es spielt besonders im Rahmen der bereits mehrfach erwähnten Feedbackprozesse eine Rolle. Soziale Komponenten sind daneben wie in Kapitel 4.5 angedeutet auch für motivationale Prozesse bedeutsam.

Wir beginnen mit zwei bekannten Aufgabenstellungen aus Experimenten der kognitiven Psychologie (zu Aufgabe 1 vgl. Mercier u. Sperber 2017, 146; Frederick 2005, 26; zu Aufgabe 2 vgl. Mercier u. Sperber 2017, 260; Simonson u. Nye 1992, 442; Thaler 1980). Versuchen Sie diese kurz zu beantworten:

1. Ein Schläger und ein Ball kosten zusammen 1,10 Euro. Der Schläger kostet einen Euro mehr als der Ball. Wie viel kostet der Ball?
2. Sie haben eine Eintrittskarte für ein Basketballspiel in einer Stadt, die sechzig Meilen von Ihrem Wohnort entfernt ist. Am Tag des Spiels gibt es einen heftigen Schneesturm, und die Straßen sind sehr schlecht. Würden Sie eher zu dem Spiel gehen, wenn:

nenalter. Solchermaßen erfolgreiche genetische Regulationen können sogar weitervererbt werden (vgl. Sapolsky 2018, 221f.). Andererseits finden Lernvorgänge postnatal vor allem durch die Elimination von sich als überflüssig erweisenden Synapsen statt – ein Prozess, der durch die Interaktion mit der Umwelt gesteuert ist und von der Geburt bis ca. zum dreißigsten Lebensjahr dauert, vgl. ebd. 156–159. Eine komplementäre Erklärung besteht darin, dass es eine vom Kontext unabhängige Intelligenz gar nicht gibt. So geht Howard Gardner von einer Mehrzahl von Intelligenzen aus, etwa sprachlicher, logisch-mathematischer, musikalischer oder kinästhetischer Intelligenz, die in spezifischen Kontexten ausgebildet werden: „Intelligenzen drücken sich immer im Kontext spezifischer Aufgaben, Domänen und Disziplinen aus“ (Gardner 1993, S. XVI).

- a) Sie 50 Dollar für die Eintrittskarte bezahlt haben.
- b) Sie die Karte umsonst bekommen haben.
- c) a und b spielen für Ihre Entscheidung keine Rolle.

Beiden Aufgaben ist gemeinsam, dass die meisten Menschen sie intuitiv falsch (Aufgabe 1) bzw. unlogisch (Aufgabe 2) beantworten. Aufgabe 1 führt zur Intuition, der Ball koste 10 Cent. Dann wäre aber der Unterschied zwischen Ball und Schläger nur 90 Cent – richtig ist die Antwort, dass der Ball 5 Cent und der Schläger 1.05 Euro kosten. Bei Aufgabe 2 antworten die meisten Menschen, dass sie sich eher dem Schneesturm aussetzen würden, wenn sie selbst für das Ticket bezahlt haben. Ökonomisch betrachtet handelt es sich allerdings um versunkene Kosten – das Geld ist unwiederbringlich ausgegeben. Bei Entscheidungen geht es um die Zukunft, die im Gegensatz zur Vergangenheit geändert werden kann. Die einzig relevante Frage ist deshalb: Finde ich es lohnenswert, mich durch den Schneesturm zum Spiel zu begeben oder soll ich etwas anderes tun (Mercier u. Sperber 2017, 260)?⁵⁵

Die kognitive Psychologie, Wirtschaftspsychologen und die Entscheidungspsychologie haben sich seit dem zweiten Weltkrieg daran abgearbeitet, weshalb Menschen irrationale Entscheidungen fällen.⁵⁶ Mercier u. Sperber (2011; 2017) haben darauf inzwischen mit ihrer Theorie der argumentativen Schlussfolgerung eine überzeugende Antwort geliefert. Diese besagt, dass logisches/schlussfolgerndes Denken zu wesentlichen Teilen auf einem sozialen Prozess basiert und dass rein individuelles Denken deshalb oft zu Fehlern führt. Gemäß Mercier u. Sperber besteht die Hauptfunktion menschlichen Schlussfolgerns nicht darin, Menschen durch individuelle Reflexion zu besseren Überlegungen zu verhelfen. Vielmehr verstehen sie die Herstellung von Schlussfolgerungen als sozialen Prozess, in dem die Einzelnen durch Argumentation und sozialen Gedankenaustausch zu genaueren Begründungen sowie besseren Argumenten und Entscheidungen finden. Entscheidend sind dabei zwei Prozesse: Einzelne Menschen versuchen andere von ihren meist intuitiv gewonnen Ansichten und Entscheidungen zu überzeugen und die Erklärungen und Argumente anderer Menschen kritisch zu überprüfen. Mercier u. Sperber (vgl. 2017, 334) verstehen diese soziale Produktion von Vernunft als evolutionäre Anpassung der Menschen an ihr komplexes Sozialle-

55 Mercier u. Sperber (2017, 260f.) weisen darauf hin, dass die sogenannte „sunk cost fallacy“ so weit verbreitet ist, dass es sozial akzeptabler sein kann, ihr zu folgen, als ihr zu widerstehen. Der Effekt ist ökonomisch bedeutsam und führt zu vielen Fehlentscheidungen, etwa bei betrieblichen Investitionen in erfolglose Projekte (vgl. Kahneman 2012, 344–346).

56 Berühmt geworden ist etwa die Begründung durch eine Unterscheidung zwischen einem schnellen, ungenauen, intuitiven „System 1“ und einem langsameren, genaueren, rationalen „System 2“ (vgl. Kahneman 2012).

ben in Gruppen, in denen Vertrauen fragil bleibt und immer wieder verdient werden muss.

Wenn Denken und Argumentieren primär dem Verteidigen und Rechtfertigen des eigenen Standpunktes dienen, dann ist ihre Einseitigkeit kein Fehler, sondern eine funktionale Eigenschaft. Die oft festgestellte und beklagte Bestätigungstendenz („confirmation bias“, siehe Kap 6.4.1) ist gemäß Mercier u. Sperber daher eine sinnvolle Eigenschaft des menschlichen Denkens – für die sie den Begriff „myside bias“ einführen (vgl. Mercier u. Sperber 2017, 218f.). Damit verbunden ist auch ein ökonomischer Umgang mit den eigenen kognitiven Ressourcen, weshalb sich Menschen fürs Erste in der Regel auf Heuristiken oder mentale Abkürzungen, also auf Intuitionen verlassen. Diese Intuitionen hängen durchaus vom Vorwissen ab – eine Meteorologin kann sich bei einem Blick in den Himmel wohl eher auf ihr Bauchgefühl zur Wetterentwicklung verlassen als ein Laie. Aber im Alltag begegnen wir vielen Themen und Fragen, die wir intuitiv bearbeiten, was zu kognitiven Verzerrungen führt. Dabei ist es ökonomisch, die Überprüfung und etwaige Korrektur unserer intuitiven Schlüsse in sozialen Prozessen vorzunehmen.

Entscheidend für die Konzeption von menschlichem Schlussfolgern als sozialem Prozess ist deshalb die Annahme, dass Menschen die Aussagen und Argumente anderer Menschen wesentlich kritischer prüfen als ihre eigenen. Durch argumentativen Austausch können Einzelne ihre Intuitionen überprüfen, insofern sie bereit sind, diese anzupassen. Der Mechanismus, der dabei zum Tragen kommt, dürfte einer Vorhersageschleife entsprechen, bei der die eigenen Argumente (Vorhersagen) anhand der Konfrontation mit fremden Argumenten und Kritik (=Input) überprüft und falls nötig angepasst werden. Dadurch lassen sich intuitive Vorhersagen in rationale Entscheidungen und Einschätzungen überführen.⁵⁷

Selbstverständlich bedeutet die Theorie der argumentativen Schlussfolgerung nicht, dass Gruppendiskussionen die beste Lösung für jedes Problem darstellen. Vielmehr werden die dargelegten Prozesse vor allem dann bedeutsam, wenn Teilnehmende in Gruppen verschiedene Ansichten und Perspektiven vertreten, aber ein gemeinsames Ziel haben – hier kommen auch motivationale und emotionale Prozesse zum Tragen (Carter u. Phillips 2017). Der

57 Mercier u. Sperber gehen davon aus, dass die menschlichen Intuitionen durch eine Vielzahl an Inferenzmechanismen zustande kommen, die auf neuronalen Netzwerken beruhen, wie sie in Kapitel 4.3 erklärt wurden. Diese Mechanismen produzieren mentale Repräsentationen von physischen und psychischen Objekten und Vorgängen. Demgegenüber basiert der Mechanismus des Schlussfolgerns/logischen Denkens auf Metarepräsentationen, also auf einer Vorhersageschleife zweiter Ordnung: Inferenzen über Folgerungen erster Ordnung. Im Gegensatz zu den primären Inferenzen ist der Mechanismus des Schlussfolgerns allgemein und nicht auf bestimmte Teilbereiche von Wahrnehmung und Denken beschränkt (vgl. Mercier u. Sperber 2017, 90–105).

Gewinn dient dabei zunächst einmal den einzelnen Menschen: „Die kollektiven Vorteile, die in solchen Fällen erzielt werden, sollten als Nebeneffekt eines Mechanismus betrachtet werden, der individuellen Interessen dient“ (Mercier u. Sperber 2017, 334, Übers. tz).

Die Theorie der argumentativen Schlussfolgerung lässt sich anhand klassischer Denkaufgaben wie den beiden am Kapitelanfang aufgeführten überprüfen. Mercier u. Sperber haben eigene Experimente durchgeführt, können aber auch diverse weitere Experimente anführen, die ihre Theorie bestätigen: Eine überwiegende Mehrheit der Teilnehmenden löst die Aufgaben in Einzelarbeit falsch, der Anteil richtiger Antworten liegt dabei oft bei 10–20 %. Werden die Aufgaben dagegen in Gruppen bearbeitet, liegt der Anteil der richtigen Lösungen meist bei 70–80 % (vgl. Mercier u. Sperber 2017, 263–265; Mercier u. Sperber 2011; Laughlin u. Ellis 1986).

Eines der angeführten Experimente ist für unseren Zusammenhang besonders interessant, da es eine wissenschaftliche Begründung für das Setting der zweistufigen Prüfung (siehe Kap. 7.7.2) liefert:

Der Psychologe David Moshman ließ seine Studierenden während vieler Jahre die berühmte Wason-Auswahlaufgabe lösen – jeweils zuerst individuell, danach in Gruppen (vgl. Wason 1966; im Internet finden sich diverse Darstellungen davon). Das Muster entspricht genau der Theorie der argumentativen Schlussfolgerung: In Einzelarbeit kamen nur wenige Studierende auf die richtige Lösung, aber in den Gruppendiskussionen und im Plenum wurde schließlich fast immer die richtige Lösung gewählt. Moshman u. Geil (1998) überprüften deshalb die Eindrücke aus dem Hochschulunterricht in einem experimentellen Setting anhand von 143 Studierenden und konnten sie bestätigen: Nur 9,4 % der Studierenden wählten individuell die richtige Lösung. Wurde die Aufgabe in Gruppen von 5–6 Studierenden bearbeitet, wählten 70 % der Gruppen die richtige Lösung. Wenn die Studierenden zuerst die Aufgabe allein bearbeiteten und danach ihre schriftlichen Lösungen in Gruppen von 5–6 Studierenden diskutierten, wählten 80 % der Gruppen die richtige Lösung. Das Prinzip, das hier zur Anwendung kommt, wurde bereits von Laughlin u. Ellis (1986) als „truth wins“ bezeichnet – etwas bescheidener übersetzt: Wenn die Mitglieder einer Gruppe unterschiedliche Sichtweisen einbringen, um gemeinsam eine intellektuelle Herausforderung zu bewältigen, dann setzt sich die Kraft des besseren Argumentes in der Regel durch. Dieses Prinzip ist generell oft in Feedbackprozessen wirksam, insbesondere unter Peers (vgl. z. B. Crouch u. Mazur 2001).

Die Theorie der argumentativen Schlussfolgerung erklärt damit auch die Forschungsergebnisse zur Transaktivität. „In einer transaktiven Äußerung verarbeitet ein Sprecher den Gedanken eines Gegenübers weiter“ (Zimmermann 2014, 35). In transaktiven Gesprächsbeiträgen wird somit ein kritisches Weiterdenken bisheriger Argumente und Positionen sichtbar. Untersuchungen zeigen, dass einerseits transaktivere Diskussionen zu besseren Aufgaben-

lösungen führen. Andererseits geht transaktiveres Gesprächsverhalten auch mit größerem individuellen Lernerfolg einher, von Kleinkindern bis zu Studierenden (vgl. Zimmermann 2014, 52–55). Das Konstrukt der Transaktivität hilft folglich sichtbar zu machen, ob eine Diskussion Elemente des sozialen Schlussfolgerns enthält. Es kann als Denkwerkzeug hilfreich sein, um bei interaktiven Leistungsnachweisformaten den Studierenden aufzuzeigen, an welchen Stellen und auf welche Weise sie Überlegungen von anderen erfolgreich weiterentwickelt haben (vgl. Zimmermann u. Rickert 2015 zu schriftlichen Online-Diskussionen als Leistungsnachweis).

4.8 Überleitung zum zweiten Teil des Buches

Das vorliegende Kapitel erlaubt die Schlussfolgerung, dass menschliches Denken auf zwei Seiten in Netzwerken stattfindet: Intrapyschisch als mentale Prozesse in einem Netzwerk aus neuronalen Netzwerken und interpsychisch/sozial als Netzwerk denkender Individuen. Das bewusste Selbst befindet sich quasi als Schaltzentrale in der Mitte dieser Netzwerke.⁵⁸ Das gilt für Lernen im engeren Sinne, wie es in diesem Kapitel geschildert wurde,⁵⁹ aber auch für die Selbst- und Fremdverortung von Menschen in der Gesellschaft, wie sie mit Bezug auf Fragen der Leistung in Kapitel 2 behandelt wurde.

Damit sind wir an einem Wendepunkt des Buches angelangt. Die bisherigen Kapitel haben versucht, eine Landkarte durch den Dschungel der sozialen und individuellen Voraussetzungen des menschlichen Lernens, des Leistungsbegriffs und des Leistungs- und Beurteilungsverständnisses im Hochschulbereich zu legen. In den weiteren Kapiteln wird diese Karte nun benutzt, um im komplexen Feld der Leistungsbewertung an Hochschulen Orientierung zu schaffen. Sie dient somit dazu, bewährte Praktiken zu erklären und zu bestätigen, fragliche Praktiken zu hinterfragen und neue Vorschläge für eine lernwirksame Beurteilung und Bewertung von Studierendenleistungen zu machen.

58 Das externe Netzwerk muss nicht nur aus Menschen (und anderen Lebewesen) bestehen, sondern kann auch den – natürlichen oder von Menschen gebauten – physischen Raum umfassen (vgl. Paul 2021).

59 Die Theorie der „distributed cognition“ geht sogar davon aus, dass Wissen grundsätzlich eine soziale Kategorie ist: Ihr zufolge sind mentale Repräsentationen nicht einfach in individuellen Gehirnen angesiedelt, sondern in soziokulturellen Systemen verteilt. Diese gemeinschaftliche Verteilung von Wissen ist die Grundlage für das individuelle Denken und die persönliche Wahrnehmung der Welt (vgl. Hutchins 1995).

Teil II: Anwendung

5 Was sollen Studierende eigentlich können? Lernziele im Constructive Alignment

Die Abstimmung zwischen dem, was die Studierenden lernen, den Leistungsnachweisen, in denen sie das Gelernte unter Beweis stellen, und den in Lehrveranstaltungen oder Selbstlerneinheiten erfolgenden Lehr- und Lern-Aktivitäten ist für gelingende Hochschullehre essenziell. Diese Anforderung wird unter dem Begriff des Constructive Alignments in Kapitel 5.1 näher vorgestellt. Daraus geht hervor, dass die Definition von übergeordneten (Kap. 5.3) und spezifischen (Kap. 5.4) Lernzielen eine zentrale Voraussetzung dafür ist, um gültige Leistungsbeurteilungen anhand von lernwirksamen Leistungsnachweisformaten durchführen zu können. Dies gilt auch dann, wenn einige der im Constructive Alignment abzustimmenden Elemente bereits durch das Curriculum vorgegeben sind. So sind übergeordnete Lernziele oft festgelegt, manchmal auch die spezifischen Lernziele und gelegentlich sogar Lehrmaterialien oder bestimmte Lehr-Lern-Aktivitäten. Und natürlich sind nicht selten auch die Leistungsnachweisformate in bestimmten Aspekten oder sogar gänzlich vorgegeben. Trotz allem lassen sich aber fast immer gewisse Freiheitsgrade ausmachen.

Dieses Kapitel kann also aus zwei Perspektiven gelesen werden: Entweder kann es dabei helfen, die Freiheitsgrade für kleinräumigere Optimierungen von bestehenden Leistungsnachweisen und Leistungsbeurteilungen zu nutzen. Oder es kann als Blaupause für das Neugestalten oder Überarbeiten von Curricula oder einzelnen Elementen (Modulbeschreibungen, Leistungsnachweisen, Lehr-Lern-Methoden) dienen, um Lernziele und Leistungsnachweise optimal aufeinander abzustimmen. Natürlich können sich diese Perspektiven auch ergänzen oder abwechseln. Für beide Blickrichtungen gedacht ist die Planungsvorlage, die in Kapitel 5.2 vorgestellt wird. Abschließend wird ein kritischer Blick auf den Kompetenzbegriff und sein Verhältnis zum Thema Lernziele geworfen (Kap. 5.5).

5.1 Lehre vom Ziel her denken: Constructive Alignment

5.1.1 Dissonanzen rund um Leistungsnachweise

Beginnen wir mit einem konkreten Fall:

Fallvignette: Einführung in die Quellenanalyse

Andreas Breitenstein ist Assistent für Geschichte der Neuzeit an einer Universität. Er hält eine Modulveranstaltung „Einführung in die Quellenanalyse“ für Geschichtsstudierende im ersten Semester (eine Doppelstunde pro Woche). Die Studierenden sollen in dieser Veranstaltung historische Quellen der Neuzeit identifizieren und beschreiben lernen. Sie sollen die Quellen entsprechend ihrem aktuellen Wissen und Können auch analysieren lernen, um aus ihnen Erkenntnisse über historische Ereignisse und Prozesse zu gewinnen. Zu diesem Zweck sollen die Studierenden verschiedene Arten von Quellen kennen lernen und sie in Bezug auf ihre Eigenschaften analysieren können, z. B. Medialität, Urheberschaft, Intentionalität. Zudem sollen sie unterschiedliche Methoden der Quellenanalyse kennen und einige davon im Rahmen der Veranstaltung auf konkrete Quellen anwenden lernen. Übergeordnetes Ziel ist, dass die Studierenden ein grundlegendes Verständnis für die Arbeit mit historischen Quellen entwickeln, wozu sie diese in Bezug auf verschiedene Aspekte beschreiben, historisch kontextualisieren, aber auch kritisch hinterfragen können.

Als Leistungsnachweis schreiben die Studierenden eine zweistündige schriftliche Prüfung, in der sie zwei verschiedene historische Quellen (z. B. ein Text und ein Bild) in Bezug auf grundlegende Aspekte beschreiben und anhand jeweils zweier verschiedener Methoden der Quellenanalyse analysieren und interpretieren sollen. Dabei legen alle Studierenden der insgesamt vier parallel durchgeführten Modulveranstaltungen „Einführung in die Quellenanalyse“ die gleiche Prüfung ab.

Leider wird Herr Breitenstein kurz nach Beginn des Semesters krank. Seine Veranstaltung fällt zweimal krankheitsbedingt und zudem zweimal wegen Feiertagen aus. Sieben Wochen vor Semesterende sieht sich er sich deshalb mit der Herausforderung konfrontiert, den Unterrichtsinhalt von ursprünglich zehn Präsenzterminen in noch sechs verbleibenden Terminen unterbringen zu müssen. Damit er alle geplanten Inputreferate und die entsprechenden Inhalte durchbringen kann, streicht er die meisten Übungen in den Präsenzsitzungen.

In der Prüfung fällt die Hälfte von Herrn Breitensteins Studierenden durch, während die Studierenden der drei Parallelveranstaltungen zu 75–85 % bestehen – wie dies in den beiden Vorjahren auch bei seinen Studierenden der Fall war.

Wie erklären Sie sich das schlechte Abschneiden der Studierenden von Herrn Breitensteins diesjähriger Einführung in die Quellenanalyse? Und wie hätte es verhindert oder zumindest gemildert werden können?

Um diese Fragen zu beantworten, ist das Konzept des Constructive Alignments⁶⁰ hilfreich. Es geht davon aus, dass bei der Planung von Lehre drei zentrale Ebenen aufeinander abgestimmt werden müssen: Die Lernziele, die Leistungsnachweise und die unterrichtsmethodische Umsetzung:

Constructive Alignment ist ein Lehrkonzept, bei dem vor Beginn des Unterrichts [die Lehre, tz] klar festgelegt wird, was die Studierenden lernen sollen [Lernziele, tz] und wie sie ihr Lernen ausdrücken sollen [Leistungsnachweis, tz]. Der Unterricht wird dann so gestaltet, dass die Studierenden in Lernaktivitäten eingebunden werden, die ihre Chancen auf das Erreichen der Lernziele optimieren, und die Leistungsnachweise sind so konzipiert, dass sie eine klare Beurteilung ermöglichen, wie gut diese Lernziele erreicht wurden. (Biggs 2014, 5f., Übers. tz)

Das klingt trivial, ist es aber in mehrfacher Hinsicht nicht. Zunächst ist die Abstimmung gar nicht so einfach, wie es auf den ersten Blick scheint – sie bedingt einiges an Denk- und Planungsleistung, die sich freilich lohnt.

Vor allem aber sind Lehrende aller Stufen häufig einen anderen Ablauf gewohnt: Das Planen vom Inhalt her (vgl. Wiggins u. McTighe 2005, 13–34). Da Lehrende an Hochschulen in der Regel ausgewiesene Experten für die Themen sind, die sie unterrichten, ist die Orientierung an den Inhalten an Hochschulen besonders ausgeprägt – und diese Tendenz ist für die didaktische Planung problematisch. Dies liegt nicht an den Themen und schon gar nicht an der Begeisterung der Lehrenden für ihre Themen. Vielmehr liegt das Problem im daraus folgenden didaktischen Ablauf:

Wird die Lehre mit primärem Fokus auf die zu behandelnden Inhalte geplant, läuft sie in der Regel so ab: Lehrende sprechen vor dem Plenum über ein Thema nach dem anderen.⁶¹ Was vorgetragen wurde, gilt als erledigt. Ob die Studierenden das Vorgestellte verstehen, anwenden, analysieren oder für

60 Das Adjektiv „constructive“ wurde von Biggs (1996, 348f.) gewählt, weil er seinen Ansatz auf dem pädagogischen Paradigma des Konstruktivismus aufbaut, in dem das aktive Konstruieren von Wissen und Können durch die Lernenden im Vordergrund steht. Dies entspricht auch den lernpsychologischen Erkenntnissen, die in Kapitel 4 dargestellt sind.

61 In Seminar-Settings handelt es sich oft um eine Serie von durch Studierende gehaltene Präsentationen, was nur wenig an der Problematik ändert – die Studierenden lernen höchstens über ihr eigenes Vortragsthema mehr.

künftige Eigentätigkeit weiterverwenden können, spielt dabei keine Rolle. Man könnte diesen Ansatz als „Lehren durch Erwähnen“ bezeichnen (vgl. Wiggins u. McTighe 2005, 21). Gegen Ende des Semesters müssen die Lehrenden dann oft noch überlegen, mit welchen Aufgaben und Fragen sie die Studierenden prüfen wollen. Leistungsnachweise bleiben somit eine Art Wurmfortsatz der Lehre; meist werden sie so gestaltet, dass die Studierenden Wissen reproduzieren müssen. Daneben sind Lehrende gelegentlich auch verpflichtet, Lernziele zu formulieren. Diese werden dann in Formulare oder Datenbanken eingetragen und geraten anschließend oft in Vergessenheit, selbst wenn sie für die Studierenden sichtbar im Vorlesungsverzeichnis publiziert werden. Ein Alignment im von Biggs dargelegten Sinn findet dabei nicht statt.

Ein solches intuitives Vorgehen basiert meist auf der – nicht immer bewussten – Annahme, Lernen erfolge wie der Bau eines Hauses. Kleinteilige Lernelemente werden durch die Lehrenden wie Backsteine zu Mauern aufgeschichtet, bis das Haus des Verstehens fertig gebaut ist (vgl. Wiggins u. McTighe 2005, 294). Wie Kapitel 4 gezeigt hat, funktioniert Lernen nicht im Geringsten auf diese Weise, sondern durch dynamische Verknüpfungen und aktive Tätigkeit der Lernenden. Ein inhaltabdeckender Unterricht führt deshalb nicht zu einer tiefgehenden Verarbeitung der Lerninhalte und leistet oberflächlichem Lernen Vorschub, bei dem zu viel Material in zu wenig Zeit verarbeitet werden muss. Diese Problematik wurde in Kapitel 4.6.1 bezüglich des Zusammenhangs von Informationsmenge und Verarbeitungstiefe erläutert. In der Tendenz führt ein inhaltabdeckender Unterricht somit dazu, dass die Studierenden „the name of something“ kennen, ohne es wirklich zu verstehen (vgl. dazu Zitat von Richard Feynman am Ende von Kap. 4.2):

- Der Aufbau von Netzwerken, komplexen mentalen Schemata und intrinsischer Motivation ist durch die mangelnde Tiefenverarbeitung, die fehlende Aktivierung von Vorwissen sowie die erschwerte Realisierung der drei psychologischen Grundbedürfnisse der Selbstbestimmung massiv erschwert (siehe Kap. 4.2 bis 4.5).
- Das Potenzial des sozialen Schlussfolgerns (Kap. 4.7) bleibt weitgehend ungenutzt, da kaum argumentativer Austausch stattfindet.
- Expertise im Sinne von Kapitel 4.6 wird höchstens außerhalb des Studiums aufgebaut: durch fachorientierte Freizeitaktivitäten, Praktika oder die Berufstätigkeit nach dem Studium.

Dass man durch einen inhaltabdeckenden Unterricht mehr Inhalt abhandeln kann, ist somit nur richtig, wenn man einzig die Lehrhandlungen betrachtet. Lehren sollte aber nicht mit dem tatsächlichen Lernen verwechselt werden: Lernen erfolgt nur, wenn Studierende die präsentierten Inhalte auch durch möglichst tiefenorientierte kognitive Prozesse verarbeiten und sie als bedeut-

sam einschätzen (vgl. Wiggins u. McTighe 2005, 44f.). Oder um es in den Worten von Ralph Tyler zu sagen: „Das Lernen erfolgt durch das aktive Verhalten des Lernalers: Er lernt durch das, was er tut, und nicht durch das, was die Lehrperson tut“ (Tyler 1949, 63, Übers. tz).

Ziehen die Inhalte hingegen einfach in Form von Worten der Lehrenden oder Lehrbüchern an den Studierenden vorbei, wird das Haus des Verstehens zum Kartenhaus, das zusammenfällt.

5.1.2 *Constructive Alignment und Backward Design*

Das Constructive Alignment geht gegenüber einer inhaltsabdeckenden Lehre von einer anderen Reihenfolge der Planung aus (siehe Abb. 11): Wenn die Studierenden etwas lernen sollen, müssen wir zuerst überlegen, *was* sie denn lernen sollen: Was sollen sie nach einer Lehrveranstaltung wissen und können? Um dies zu klären, müssen wir übergeordnete und spezifische Lernziele formulieren (siehe Kap. 5.3 und 5.4). Anschließend ist zu überlegen, woran wir erkennen können, ob Studierende das gelernt haben, was sie lernen sollten. Anhand dieser Frage können wir Leistungsnachweise gestalten, also Settings, in denen Studierende ihr neu erworbenes Wissen und Können zeigen können (siehe Kap. 7). Wenn wir beides kennen, können wir schließlich die Lehr- und Lernaktivitäten so gestalten, dass die Studierenden möglichst solche Lernhandlungen vornehmen, die der Erreichung der Lernziele dienen (vgl. Biggs u. Tang 2007, 50–63; Biggs 2014; Biggs 1996).

Damit können wir zurückkehren zur Fallvignette und Herrn Breitenstein. Das Constructive Alignment erlaubt folgende Analyse: Indem er die Übungen gestrichen hat und weitgehend nur noch Vorträge hielt, hatten die Studierenden in den Präsenzveranstaltungen keine Gelegenheit mehr, die Theorien und Methoden anzuwenden und sie damit tiefergreifend zu verarbeiten – das Alignment fehlte. Damit fehlten vielen von ihnen auch sinnvolle Vorgehensweisen, um sich auf die Prüfung vorzubereiten – nur durch das Lesen von Skripten oder Theorietexten sind anspruchsvolle, handlungsorientierte Lernziele wie das Analysieren von Quellen nicht zu erreichen (siehe Kapitel 5.4 zu spezifischen Lernzielen). Viele Studierende dürften also bei diversen Aspekten lediglich „the name of something“ gekannt haben, was für das in der Vignette skizzierte anwendungsorientierte Prüfungsformat nicht ausreichte.

Die Fallvignette ist freilich in einem Punkt etwas speziell gewählt: Im Hochschulalltag entstehen die meisten Brüche im Alignment nicht nachträglich durch Krankheit oder Fehlüberlegungen von Dozierenden, sondern das Alignment wird gar nicht erst hergestellt. Besonders groß ist oft der Bruch zwischen den von den Studierenden erwarteten Lernleistungen und der Art und Weise, wie deren Lernen in Leistungsnachweisen gemessen wird. Als

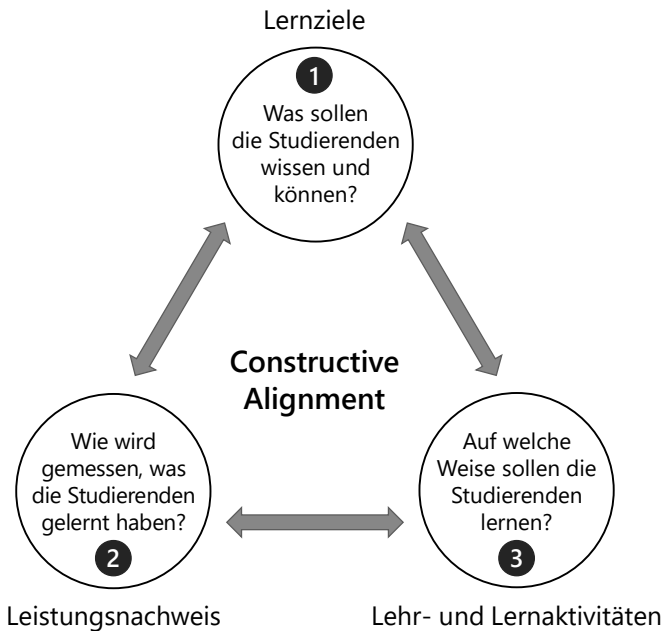


Abb. 11: Constructive Alignment nach Biggs; die Ziffern zeigen die Planungsreihenfolge gemäß Backward Design. Quelle: Eigene Abbildung, in Anlehnung an Institute for Teaching and Learning Innovation, The University of Queensland 2023

typisches Beispiel für einen solchen Bruch führen Wiggins u. McTighe (vgl. 2005, 148) eine Geschichtsprofessorin an, die ein Seminar mit dem Lernziel „doing history“ einzig durch eine summative Multiple-Choice-Prüfung mit 100 Fragen überprüft. Solche Brüche führen zu einer erheblichen Diskrepanz zwischen dem von den meisten Studierenden tatsächlich gezeigten Lernverhalten und jenem, das sich die Lehrenden wünschen.

Der Grund dafür liegt darin, dass die Studierenden ihren Lernerfolg primär am Abschneiden in summativen Leistungsnachweisen bemessen und deshalb so lernen, dass sie in diesen gut bewertet werden. Leistungsnachweise und die Art ihrer Bewertung bestimmen also das Lernverhalten der Studierenden stärker als der Lehrplan, die offiziellen Lernziele oder die Unterrichtsmethoden. Dieser Effekt wird als *Backwash* bezeichnet (vgl. Biggs u. Tang 2007, 169).

Gelegentlich klagen Lehrende über die Bewertungsorientierung der Studierenden und interpretieren den Backwash als eine Form von „Gaming the

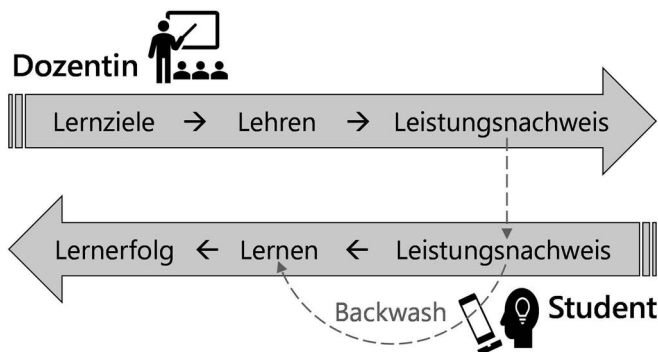


Abb. 12: Die unterschiedlichen Perspektiven der Dozierenden und Studierenden auf Leistungsnachweise und der sich daraus ergebende Backwash. Quelle: Eigene Abbildung, in Anlehnung an Biggs und Tang 2007, 169

system“. Dies ist wenig zielführend, da indikatorenbasierten Systemen, wie summative Leistungsbewertungen sie darstellen, eine verhaltenssteuernde Wirkung grundsätzlich eigen ist, wie in Kapitel 2.4 gezeigt wurde.⁶² Solange studentische Leistungen summativ bewertet werden, ist es deshalb sinnvoll, den Backwash von Leistungsnachweisen anzuerkennen und ihn in die didaktische Planung miteinzubeziehen. Dazu gehört das Bewusstsein des Perspektivenunterschieds zwischen Dozierenden und Studierenden, der in Abbildung 12 dargestellt ist.

Der Backwash-Effekt liefert somit einen weiteren Grund, weshalb es sinnvoll ist, die didaktische Planung im Constructive Alignment vom Ende des Lernprozesses her rückwärtszudenken: Lernziele → Leistungsnachweis → Lehr- und Lernaktivitäten. Dieses Vorgehen wird deshalb als Backward Design bezeichnet (vgl. Wiggins u. McTighe 2005, 13–34).⁶³

Auch falls im laufenden Lehrbetrieb unerwünschte Backwash-Effekte beobachtet werden, empfiehlt es sich, das Alignment der betroffenen Lehrveranstaltungen zu überprüfen. Bruchstellen sollten identifiziert und korrigiert werden. Die resultierenden Anpassungen sollten ebenfalls beim Alignment in der oben genannten Reihenfolge ansetzen: 1) Lernziele, 2) Leistungsnachweis, 3) Lehr- und Lernaktivitäten. Man könnte auch sagen: Im Constructive

62 Wiggins u. McTighe (2005, 49) sprechen von einem Katz-und-Maus-Spiel, bei dem das Bewertungssystem den Studierenden Anreize gibt, Lernen bloß vorzutäuschen. Grundsätzlich ist zu bedenken, dass Studierende wie wir alle durch die in Kap. 2 geschilderten Werte unserer Leistungsgesellschaft geprägt sind.

63 Wiggins u. McTighe haben diesen Begriff geprägt, die Grundidee ist aber bereits Mitte des 20. Jahrhunderts von Tyler (1949) eingeführt worden. Auf Tyler verweisen auch Biggs u. Tang (2007, 30).

Alignment werden Backwash-Effekte als Ressource genutzt – wenn die Leistungsnachweise geschickt gestaltet sind, führen sie die Studierenden dazu, ein für die Erreichung der Lernziele sinnvolles Lernverhalten an den Tag zu legen.

Wenn man einen anderen Planungsablauf gewohnt ist, kann die Umstellung auf ein Backward Design aufwendig scheinen. Das ist aber primär eine Frage der Gewöhnung. Ein Constructive Alignment lohnt sich spätestens mittelfristig, weil dadurch die Lehr- und Lernaktivitäten mit großer Wahrscheinlichkeit erfolgreicher ablaufen. Missverständnisse und Konflikte zwischen Lehrenden und Studierenden können reduziert werden, und der Lernerfolg der Studierenden nimmt zu. Ein gutes Alignment hilft damit auch, die Zahl von Studierenden zu reduzieren, die Module oder Leistungsnachweise wiederholen müssen; im Idealfall sinkt sogar die Drop-Out-Quote. Alle genannten Effekte reduzieren unter dem Strich den Aufwand der Lehrenden.

5.1.3 Sinnvolle Leistungsnachweise führen nicht zu Normalverteilungen

An dieser Stelle ist mit einem tief verankerten Vorurteil aufzuräumen: Dass Leistungsnachweise zu einer möglichst großen Streuung oder sogar einer Normalverteilung (gaußsche Glockenkurve) der Leistungsbewertungen führen sollen.⁶⁴ Das ist eine pädagogisch ebenso inadäquate wie verheerende Vorstellung (siehe Kapitel 9.2.1; vgl. Sacher 2014, 119–122; Biggs u. Tang 2007, 170f.). Zufällig auftretende Größen verteilen sich zwar tatsächlich oft normal. Entsprechend liegt bei vielen psychologischen Eigenschaften wie Persönlichkeitsmerkmalen, Interessen oder Werthaltungen in der Regel eine Normalverteilung vor.

Die testtheoretische Annahme einer Normalverteilung ist bei Leistungsnachweisen im Bildungskontext aber irreführend: Eine Lehre, die davon ausgeht, dass an ihrem Ende die Lernleistungen normalverteilt sind, unterstellt ihre eigene Erfolglosigkeit (vgl. Sacher 2014, 121). Oder in den Worten eines der bekanntesten Lernpsychologen des 20. Jahrhunderts: „Wir können sogar betonen, dass unsere Lehranstrengungen in dem Maße *erfolglos* waren, in dem sich die Verteilung der Leistungen der Normalverteilung annähert“ (Bloom, Madaus u. Hastings 1981, 53, Hervorhebung von Bloom et al.). Der Lernstand von Studierenden nach erfolgter Lehre sollte eben gerade nicht mehr zufällig sein, sondern schief in Richtung erfolgreich erreichter Lern-

64 In der Praxis zeigt sich das oft darin, dass bestimmte Notendurchschnitte erwartet oder sogar explizit eingefordert werden und sich Lehrende zumindest rechtfertigen müssen, wenn sie ihre Studierenden deutlich besser bewerten (vgl. Nölte u. Wampfler 2021, 35; 41f.).

ziele (vgl. Biggs u. Tang 2007, 172) – sonst wäre jegliche didaktische Maßnahme, und sei es nur das Definieren von Studienlektüre, überflüssig.

An der Unbrauchbarkeit der Normalverteilung ändert auch der Umstand nichts, dass gerade die Verwaltung und das Management von Hochschulen die Verwendung von Normalverteilungen mögen, weil sie den Eindruck vermitteln, die Studienleistungen fielen Jahr für Jahr ungefähr gleich aus – ein Trugschluss, da die Noten an der Normalverteilung geeicht werden und somit stets ähnlich verteilt sind (Biggs u. Tang 2007, 174).

Dementsprechend unterläuft eine Benotung entlang der Normalverteilung das Constructive Alignment, das auf eine Bewertung nach fachlich-inhaltlichen Kriterien abzielt, für die der Leistungsvergleich unter den Studierenden irrelevant ist (siehe Kap. 6.2 zu den unterschiedlichen Beurteilungsnormen und 9.2 zu den damit verbundenen Bewertungskonzeptionen). Wenn ein verbessertes Alignment von Lernzielen, Leistungsnachweisen und Lehr-Lern-Settings bessere Leistungsbewertungen hervorbringt, spricht das deshalb nicht gegen das Alignment, sondern ist ein Zeichen für seine erfolgreiche Umsetzung.

5.2 Planungsvorlage für Lernziele

Durch die Vorstellung des Constructive Alignments als grundlegendes Modell für die Planung von Hochschullehre wurde klar, dass zuerst Lernziele definiert werden müssen, bevor konkrete Leistungsnachweise oder Bewertungskriterien konzipiert werden können. Das Vorgehen zum Formulieren wird im verbleibenden Kapitel näher erläutert. Zuvor wird an dieser Stelle eine Planungsvorlage für Lernziele vorgestellt (siehe Abb. 13). Sie soll Ihnen einerseits die praktische Arbeit beim Formulieren von Lernzielen erleichtern. Andererseits bietet sie eine Übersicht über die wichtigsten Aspekte, die in den Kapiteln 5.3 (übergeordnete Lernziele) und 5.4 (spezifische Lernziele) erläutert werden. Sie können diese Planungsvorlage im Internet bei den Materialien zum Buch als Word- und PDF-Dokument herunterladen (Zugang per QR-Code, s. vorne am Ende des Inhaltsverzeichnisses).

Die Planungsvorlage soll nicht als mechanistisch abzuarbeitendes Schema verstanden werden: Je nach Kontext kann es gute Gründe geben, weshalb man einen Schritt überspringt, zwei Aspekte zusammenfasst oder einen zusätzlichen Schritt unternimmt.

Gleiches gilt für die Reihenfolge: Die Darstellung folgt der Reihenfolge, die nach dem Abschluss der Erarbeitung logisch sinnvoll ist. Das heißt aber nicht, dass man zwingend bei den Leitkonzepten beginnen muss – vielleicht drängen sich zuerst Kernfragen auf, oder man füllt zuerst die Teile ein, die

Planungsraster zur Formulierung von Lernzielen

Übergeordnete Lernziele definieren	
Modultitel: ...	
Kurzbeschreibung des Moduls ... Formale Vorgaben - ...	
Etablierte Ziele - ... - ...	
Leitkonzepte - ... - ...	
Übergreifende Kernfragen - ... - ...	Lokale Kernfragen - ... - ...
Nachhaltige Erkenntnisse - ... - ...	
Spezifische Lernziele definieren	
Kognitive Lernziele - ... - ... - ...	
Affektive Lernziele - ... - ...	Psychomotorische Lernziele - ... - ...

Abb. 13: Planungsraster zur Formulierung von Lernzielen. Quelle: Eigene Abbildung, als Download erhältliches Dokument

bereits aus einem Curriculum oder einem Modulhandbuch vorgegeben sind, und vervollständigt und personalisiert die eigene Lernzielplanung darum herum. Zudem wird das Vorgehen häufig iterativ sein, das heißt, gewisse Schritte werden mehrfach durchlaufen. So ist eine Überarbeitung der Lernziele spätestens nach der ersten Durchführung einer Lehrveranstaltung oft sinnvoll.

5.3 Übergeordnete Lernziele identifizieren und formulieren

Durch die Vorstellung des Constructive Alignments als grundlegendes Modell für die Planung von Leistungsnachweisen (und Hochschullehre insgesamt) wurde klar, dass zuerst Lernziele definiert werden müssen, bevor konkrete Leistungsnachweise oder Bewertungskriterien konzipiert werden. In hochschuldidaktischen Handreichungen und Kursen wird dazu meistens die revidierte bloomsche Taxonomie der kognitiven Lernziele als Instrument vorgestellt. Dies greift meines Erachtens etwas zu kurz und birgt die Gefahr einer mechanistisch-dogmatischen Auslegung der Taxonomie (die Taxonomie wird aber in Kap. 5.4.2 kurz erläutert).

Denn diese vermittelt zumindest in ihrer Kurzfassung eine Eindeutigkeit, die es so nicht gibt. Es beginnt schon damit, dass in allgemein gefassten, von Fach und Bildungsstufe unabhängigen Taxonomien die erkenntnistheoretische Komplexität der gelernten Inhalte offenbleibt. So kann ein Begriff wie „Verstehen“ in einem Handlungsbereich wie Lesen von der Primarschule bis zum Studium der Literaturwissenschaft ganz unterschiedlich verstanden werden (vgl. Yorke 2003, 495). Die Taxonomiestufen müssen folglich stets fach- und situationsspezifisch zugewiesen und ausformuliert werden. Hinzu kommt, dass in Lernzieltaxonomien vorgenommene Einteilungen mitnichten irgendwelchen Naturgesetzen entsprechen. Vielmehr handelt es sich um Heuristiken (siehe Kap. 5.4.1).

Um die Taxonomiestufen fach- und situationsspezifisch formulieren zu können, ist es hilfreich, sich vor ihrer Ausformulierung den übergeordneten Kontext zu vergegenwärtigen. Daraus lassen sich dann die spezifischen Lernziele für einzelne Module, Lehrveranstaltungen oder Lektionen ableiten. Für das Identifizieren und Herleiten übergeordneter Zielsetzungen gibt es verschiedene Ansätze, wobei die Vorgehensweise „Understanding by Design“ von Wiggins u. McTighe (2005) besonders hilfreich scheint und nachfolgend in vereinfachter und auf Hochschulen zugeschnittener Form vorgestellt wird.⁶⁵

Dabei ist zu beachten, dass die beschriebenen Elemente im konkreten Fall bereits vollumfänglich, teilweise oder gar nicht vorgegeben sein können. Die Freiheitsgrade bei der Bestimmung der übergeordneten Lernziele können also sehr unterschiedlich sein. Im einen Extrem sind Sie beim Lehren weitgehend

65 Eine etwas technizistische Vorgehensweise in Form eines Abgleichs zwischen Studienzielen auf den drei Ebenen der Hochschule, Lernzielen auf Ebene des Studiengangs und Lernzielen auf Ebene der einzelnen Module/Lehrveranstaltungen empfehlen im Vergleich dazu Biggs u. Tang (2007, 85–88) Dies leistet freilich büro- und numerokratischen Auslegungen des Alignments Vorschub (vgl. Loughlin, Lygo-Baker u. Lindberg-Sand 2021).

frei in der Zielsetzung für Ihre Lehrveranstaltung (etwa in Wahlmodulen), im anderen Extrem sind Sie weitestgehend gebunden. In beiden Fällen ist es aber wichtig, sich diese übergeordneten Aspekte der Zielsetzung bewusst zu machen und in eine übersichtliche Struktur zu bringen. Dies gilt auch für den Fall, dass die herrschenden Vorgaben einer kritischen Analyse unterzogen und dadurch zur Diskussion gestellt werden sollen. Der häufigste Fall dürfte freilich darin bestehen, dass ein Teil der Elemente definiert ist und ein anderer Teil nicht, oder dass Raum besteht, die definierten Elemente um weitere Aspekte zu ergänzen und damit eigene Nuancierungen zu setzen. Zwecks einfacher Lesbarkeit sind die folgenden Abschnitte so geschrieben, als würde es um eine vollständige Neuformulierung der Ziele gehen.

Das Vorgehen zum Identifizieren und Formulieren übergeordneter Lernziele besteht aus folgenden Elementen, die anschließend erläutert werden:

- Titel und Kurzbeschreibung der Lehrveranstaltung/des Moduls
- etablierte Ziele
- Leitkonzepte
- Kernfragen
- nachhaltige Erkenntnisse

Wie Sie in Kapitel 5.4 sehen werden, fällt das Formulieren konkreter und anhand von studentischen Leistungen beobachtbaren Lernzielen wesentlich einfacher, wenn diese übergeordneten Elemente vorgängig erarbeitet werden. Sie stellen somit einen wertvollen Zwischenschritt zur Herstellung des Constructive Alignment dar.

5.3.1 *Titel und Kurzbeschreibung der Lehrveranstaltung/des Moduls*

In der Regel ist es sinnvoll, nicht nur den Titel der Veranstaltung zu notieren, für welche die übergeordneten Lernziele festgelegt werden sollen. Auch eine Kurzbeschreibung der Veranstaltung, wie sie etwa in einem Vorlesungsverzeichnis veröffentlicht wird, hilft, um sich in aller Kürze die wichtigsten Zusammenhänge zu vergegenwärtigen. Sollte es formale Vorgaben zur Veranstaltung geben, wie sie oft in Vorlesungsverzeichnissen oder Studienprogrammen veröffentlicht werden, können diese am Ende der Kursbeschreibung ebenfalls aufgeführt werden: etwa die Zahl der ECTS-Punkte, die Art des Leistungsnachweises oder Module, die zum Besuch dieser Veranstaltung bereits absolviert sein müssen.

Als Beispiel wird hier die Beschreibung eines Moduls für Masterstudierende in Erziehungswissenschaft gewählt, die eine Vertiefung in Hochschuldidaktik gewählt haben. Dieses Beispiel wird im restlichen Kapitel 5 weitergeführt.

Beispiel*Modultitel*

Planen, Durchführen und Beurteilen von Leistungsnachweisen in der Hochschullehre

Kurzbeschreibung des Moduls

In diesem Modul lernen die Studierenden, wie Leistungsnachweise und Leistungsbeurteilungen in der Hochschullehre geplant und durchgeführt werden. Wesentliche Elemente bilden dabei die Grundlagen der Leistungsmessung wie Gütekriterien, kognitive Verzerrungen, Bezugsnormen der Beurteilung, aber auch Fragen des gesellschaftlichen Umgangs mit Leistung in Schule und Arbeitswelt. Bezogen auf das Constructive Alignment wird die Abstimmung zwischen Lernzielen und Leistungsnachweisen sowie zwischen Leistungsnachweisen und Lehr-Lern-Aktivitäten näher thematisiert. Der Hauptfokus liegt schließlich auf unterschiedlichen Formaten von Leistungsnachweisen, der Unterscheidung von formativen und summativen Beurteilungen und dem Erteilen von Leistungsrückmeldungen (Feedback) durch Lehrende und Peers. Die herrschende Praxis an Hochschulen soll dabei kritisch hinterfragt werden mit dem Ziel, Vorgehensweisen zu entwickeln, die Studierende zu möglichst hochwertigem Lernen anregen.

Formale Vorgaben:

- 9 ECTS-Punkte
- Der erfolgreiche Besuch des Moduls „Einführung in die Hochschuldidaktik“ ist vorausgesetzt.
- Leistungsnachweis: Schriftliche Arbeit (Entwicklung eines zweiteiligen Leistungsnachweis-Settings mit formativer und summativer Komponente)

5.3.2 *Etablierte Ziele*

Hier geht es um Ziele auf übergeordneter Ebene.⁶⁶ Damit sind Ziele angesprochen, die etwa in Curricula/Lehrplänen, Studienprogrammen oder Modulzielen festgelegt sind. Je nach Hochschultyp, Fachrichtung oder landesspezifischen Regulierungen lauten die Bezeichnungen dafür unterschiedlich oder es gibt noch weitere Abstufungen wie etwa Modulgruppen. Auch wenn deshalb die Wortwahl hier einigermaßen allgemein bleiben muss: Das Wesentliche besteht darin, dass wir formale, etablierte Ziele identifizieren, die für die von uns geplante Lehrveranstaltung relevant sind.

Diese Ziele können sich explizit auf unsere Lehrveranstaltung oder unsere Thematik beziehen. Es ist aber auch möglich und hilfreich, allgemeinere Ziele zu identifizieren, die einen wesentlichen Bezug zur Thematik der geplanten Veranstaltung haben und deshalb helfen, sie im Curriculum und dessen Verlauf einzuordnen (zu Aspekten der Curriculumentwicklung siehe Kap. 10). Diese Art von Zielen ist in der Regel noch nicht für die praktische Anwendung im Rahmen der Planung von Lehrveranstaltungen oder gar einzelnen Sitzungen gedacht. Sie vermischen deshalb in der Regel auch verschiedene Ebenen oder Aspekte, die wir im Rahmen einer didaktischen Planung analytisch zu trennen versuchen. Dazu gehört die Mischung verschiedener Lernzielarten und -ebenen, aber auch von Leistungsansprüchen gegenüber den Studierenden und von Ansprüchen an die Qualität des eigenen Studienangebots (vgl. Wiggins u. McTighe 2005, 58).

Die Beziehung von Lehrenden zu dieser Art von Lernzielen ist gelegentlich ambivalent, da sie einer Form von Fremdsteuerung entsprechen. Dennoch ist es essenziell, sich damit auseinanderzusetzen und sich damit auf reflektierte Weise zu identifizieren. Dies stellt sicher, dass die verschiedenen Elemente eines Studiengangs nicht unkoordiniert nebeneinanderstehen. Wiggins u. McTighe (2005, 58) mahnen deshalb eindringlich:

Gerechtfertigte Entscheidungen darüber, was gelehrt, was weggelassen, was betont und was abgeschwächt werden soll, können nur dann getroffen werden, wenn es vereinbarte Prioritäten gibt, die sich auf die abschließenden Lernziele [des Studiengangs, tz] beziehen. (...) Der größte Mangel in den Unterrichtsplanungen von Lehrenden besteht insgesamt darin, dass die zentralen intellektuellen Prioritäten (...) durch die Maschen der einzelnen Lektionen, Einheiten und Kurse fallen, die sich auf das Entwickeln tausender einzelner, nicht priorisierter und unverbundener Wissens- und Könnens-Elemente fokussieren. (Wiggins u. McTighe 2005, 58, Übers. tz)

Mit Blick auf die Planung einer konkreten Lehrveranstaltung ist es sinnvoll, diese Ziele bereits etwas handlungsnäher formuliert zu fassen, wobei sie aber noch nicht den praktischen Ansprüchen an spezifische Lernziele genügen müssen, wie sie in Kapitel 5.4 erläutert werden.

66 Diese Ziele können, müssen aber nicht als Lernziele im eigentlichen Sinn formuliert sein (siehe Kap. 5.3.5 und 5.4).

Beispiel (Fortsetzung von Kap. 5.3.1)

Etablierte Ziele

- Die Studierenden verstehen die Bedeutung von Leistungsnachweisen und Leistungsbeurteilung in der Hochschullehre inklusive ihres enormen Einflusses auf das Lernverhalten der Studierenden.
- Sie können selbständig angemessene Leistungsnachweis-Settings planen und bis zur Beurteilung und Bewertung umsetzen.
- Die Studierenden verstehen die zentrale Bedeutung von Feedback für Lehr- und Lernprozesse, können Studierenden konstruktive Leistungsrückmeldungen geben und didaktische Settings konzipieren, in denen Studierende einander Peer-Feedback geben.

5.3.3 Leitkonzepte

Gleich, welchen Maßstab wir heranziehen: Jedes akademische Thema umfasst mehr Inhalte, als wir in einer Lehrveranstaltung und in den zugehörigen Leistungsnachweisen adressieren können. Entsprechend sind Curricula in der Regel überfüllt und können durch die Fülle der Lernziele, die in der gegebenen Zeit erreicht werden sollen, überfordernd wirken. Auf tieferer Ebene ist dies nicht selten auch bei Modulvorgaben der Fall, insbesondere bei Pflichtmodulen. Aber unsere eigenen Vorstellungen, was in einem Themengebiet unbedingt adressiert werden sollte, führen meistens ebenfalls über den zur Verfügung stehenden Rahmen hinaus. Unsere Aufgabe als Lehrende ist es deshalb, Prioritäten zu setzen und diese gegenüber den Studierenden transparent zu machen (vgl. Wiggins u. McTighe 2005, 66).

Priorisieren können wir am besten, indem wir die Leitkonzepte (Wiggins und McTighe sprechen von „big ideas“) identifizieren, die im Kern der adressierten fachlichen Zusammenhänge stehen. Dies sind übergreifende Konzepte oder Prinzipien, die einzelne Fakten, Prozesse und ihre Zusammenhänge in einem Themen- oder Fachgebiet miteinander verbinden und ihnen einen Sinn geben. Die Studierenden sollten diese Leitkonzepte im Laufe der Auseinandersetzung mit dem behandelten Thema erfassen, um ein tiefes und für Erweiterungen offenes Verständnis (nachhaltige Erkenntnisse, Kap. 5.3.5) für dieses zu entwickeln. Leitkonzepte zielen oft über die traditionellen Fachgrenzen hinaus und helfen damit, Verbindungen zwischen verschiedenen Wissens- und Könnensbereichen herzustellen.

Zur Veranschaulichung seien hier vier Beispiele aus verschiedenen akademischen Disziplinen aufgeführt:

- Informatik: Die Grenzen zwischen Mensch und Maschine werden unscharf, Konzepte wie Persönlichkeit, Bewusstsein und Intelligenz müssen immer wieder neu gedacht werden – inklusive ethisch-moralischer Aspekte.
- Literaturwissenschaft: Literatur ist wie ein Kaleidoskop, das menschliche Erfahrungen verarbeitet und zur Sprache bringt, während es zugleich das Weltverständnis der Lesenden und ihre Werte beeinflusst.
- Volkswirtschaftslehre: Wirtschaftspolitik ist ein zweiseitiges Schwert, das entweder einen Weg zu mehr Chancengleichheit bahnen oder bestehende Ungleichheiten in der Ressourcenverteilung vertiefen kann – mit weitreichenden Folgen für Einzelne, Gruppen und Gesellschaften.
- Gesundheitswesen: Die Bekämpfung von Infektionskrankheiten erfordert ein gemeinsames Orchester von Disziplinen, in dem Epidemiologie, Mikrobiologie, Sozialwissenschaften, Wirtschaft, Politik und Kommunikation alle ihre Stimmen beisteuern. Sie funktioniert nur, wenn diese Stimmen als öffentliche Gesundheitsbemühungen harmonisieren, die Grenzen überschreiten und auf globaler Ebene wirken.

Zusammen mit den in Kapitel 5.3.4 erläuterten Kernfragen helfen Leitkonzepte den Studierenden, die behandelten Themen als bedeutsam zu erleben und in ihr Vorwissen einzubauen. Ein Leitkonzept weist dabei folgende zentralen Eigenschaften auf (vgl. Wiggins u. McTighe 2005, 69):

- Es verweist auf zentrale Ideen des Expertenverständnisses.
- Es hilft, das fachliche Verständnis der Thematik als Dreh- und Angelpunkt zu organisieren.
- Es erfordert „Aufdeckung“ – d. h. seine Bedeutung ist für Lernende nicht auf den ersten Blick offensichtlich oder es kann gar kontraintuitiv sein.
- Es hat einen hohen Transferwert, d. h. es lässt sich auf viele andere Themen anwenden – „horizontal“ (themen- oder fächerübergreifend) und „vertikal“ (im Laufe der Semester in späteren Modulen).

Dank einem Leitkonzept können Studierende das Gelernte auch eher auf ungewohnte Situationen übertragen, anstatt unverbundene Fakten auswendig zu lernen oder isolierte Fähigkeiten einzuüben („träges Wissen“, vgl. die in Kap. 4.2 geschilderten Probleme von Physikstudierenden mit der Anwendung des Gravitationsgesetzes). Leitkonzepte sind deshalb zentral, um Lernziele zu entwickeln, die auf den Aufbau von Wissen und Können zielen, das in der Forschung und/oder der Berufswelt anwendbar ist.

Beispiel (Fortsetzung von Kap. 5.3.1)*Leitkonzepte*

- Zielgerichtete Beurteilung: Der Wert von Leistungsnachweisen wird erhöht durch die Verwendung verschiedener Ansätze, um effektiv auf die Kenntnisse, Fähigkeiten und Einstellungen der Lernenden zu schließen und Leistungsrückmeldungen durch Lehrende und Peers zu ermöglichen.
- Dynamische Beurteilung in dynamischen Diskursen: Bewertungskriterien müssen Offenheit für sich entwickelnde Forschungstendenzen, Theorien und Ansätze in den jeweiligen akademischen Disziplinen zulassen.

5.3.4 Kernfragen

Die Kernfragen stehen in enger Beziehung zu den Leitkonzepten. Vielleicht fällt es Ihnen auch leichter, mit den Kernfragen zu beginnen, bevor Sie die Leitkonzepte eruieren; der Prozess kann wie die ganze Entwicklung der übergreifenden Lernziele auch iterativ sein.

Die Kernfragen weisen auf die Leitkonzepte hin und dienen als mentale Türöffner. Sie unterstützen die Studierenden dabei, zentrale Fragestellungen und Probleme zu identifizieren und zu erforschen (vgl. Wiggins u. McTighe 2005, 106). Hier vier Beispiele, die an die in Kapitel 5.3.3 angeführten Leitkonzepte anschließen:

- Informatik: Worin bestehen die ethischen Risiken im Umgang mit maschinellem Lernen und künstlicher Intelligenz? Werden Maschinen jemals eine mit Menschen vergleichbare eigene Persönlichkeit entwickeln?
- Literaturwissenschaft: Wie formen Literatur und Erzählungen unser Verständnis von Gesellschaft, Kultur und menschlicher Identität – und inwiefern sind sie Ausdruck dieses Verständnisses?
- Volkswirtschaftslehre: Welchen Einfluss haben wirtschaftspolitische Regelungen und Maßnahmen auf die Verteilung von Ressourcen und Reichtum in einer Gesellschaft?
- Gesundheitswesen: Welche Strategien können zur Vorbeugung und Bekämpfung von Infektionskrankheiten in der globalisierten Welt eingesetzt werden?

Als „mentale Türöffner“ können Kernfragen gemeinsam mit Leitkonzepten aufgrund zweier zentraler lernpsychologischer Prozesse wirken, der Aktivierung von Vorwissen (Kap. 4.4.2) und dem Motivieren durch Erleben von Bedeutsamkeit (Kap. 4.5.2). Beide Prozesse begünstigen tiefenorientiertes Lernen (vgl. Kap. 4.2) und damit den Aufbau komplexer Wissensstrukturen (Kap. 4.6).

So wirken Kernfragen einerseits motivierend, indem sie durch provokative Formulierungen den Studierenden aufzeigen, inwiefern die adressierten Fachthemen für sie relevant sein können. Dies können fachliche Herausforderungen sein – aber auch der Lebensweltbezug kann die Studierenden die mit den Kernfragen verbundenen fachlichen Konzepte als bedeutsam erleben lassen. Fragen wie die obigen Beispielformulierungen beschäftigen viele Menschen auch im lebensweltlichen Alltag und ermöglichen zudem interessante Diskussionen im Freundeskreis oder in der Familie. So führen Kernfragen idealerweise zur Erkenntnis: „Das hat ganz konkret etwas mit mir und meinem Leben zu tun.“

Kernfragen sollten deshalb möglichst zeitlos sein und sich in verschiedenen Lebensphasen, auf verschiedenen Komplexitätsstufen oder im Laufe der wissenschaftlichen Diskursentwicklung wiederholt stellen (lassen). Sie zeigen somit auch auf, dass es in der Wissenschaft und beim Lernen nicht nur darum geht, die Antwort auf etwas zu finden, sondern darum, die richtigen Fragen zu stellen und sie iterativ zu erkunden (vgl. Wiggins u. McTighe 2005, 108). Andererseits stellen Kernfragen in aller Regel Bezüge zwischen verschiedenen Wissens- und Handlungsdimensionen her und aktivieren somit Vorwissen aus relevanten Wissens- und Könnensbereichen. Diese Aktivierung erleichtert den Einbau neuer Erkenntnisse und Prozeduren in das bestehende Wissen und Können. Zudem soll die Offenheit der Kernfragen auch in die Zukunft wirken und eine künftige Anwendung von Leitkonzepten und damit verbundenen Theorien, Modellen etc. auf neue Kontexte begünstigen: „Die kraftvollsten dieser disziplingebundenen Fragen öffnen das Denken und die Möglichkeiten für alle – für Neulinge und Experten gleichermaßen“ (Wiggins u. McTighe 2005, 108, Übers. tz).

Zu unterscheiden ist weiter zwischen übergreifenden und lokalen Kernfragen. Die vier oben genannten Kernfragen führen über den thematischen Inhalt im engeren Sinne hinaus auf die Leitkonzepte hin, die über einzelne Lehrveranstaltungen oder Module hinausweisen. Es ist aber sinnvoll, diese sehr grundsätzlichen und oft nicht abschließend beantwortbaren Fragen thematisch zu „erden“ mit Kernfragen, die direkter auf die Probleme zielen, mit denen sich die zu planende Lehrveranstaltung beschäftigt. So ließen sich die oben aufgeführten übergreifenden Kernfragen durch folgende lokale Kernfragen ergänzen:

Literaturwissenschaft:

- Wie stellen literarische Texte aus verschiedenen historischen Epochen und kulturellen Kontexten soziale und kulturelle Normen, Werte und Identitäten dar?
- Wie prägen literarische Techniken wie Charakterisierung, Handlung und Sichtweise die Wahrnehmung der Leser von Gesellschaft, Kultur und Identität in einem Text?
- Wie beeinflussen die Überschneidungen von Rasse, Geschlecht, Bildung, Klasse und anderen sozialen Kategorien die Darstellung von Gesellschaft, Kultur und Identität in literarischen Texten?

Volkswirtschaftslehre:

- Wie wirkt sich die Einführung von Mindestlöhnen auf Beschäftigung, Einkommen und Armutsquote aus?
- Welche Auswirkungen haben staatliche Wohlfahrtsprogramme auf die soziale Mobilität?
- Welche sozialen und wirtschaftlichen Auswirkungen hat die staatliche Bildungspolitik auf die Ressourcenverteilung und die Vermögensungleichheit?

Beispiel (Fortsetzung von Kap. 5.3.1)*Kernfragen (Auswahl)*

Übergreifende Kernfragen:

- Wie schließen wir von äußerlich gezeigtem Handeln (Performanz) auf nicht direkt sichtbare Eigenschaften wie Wissen, Können oder Haltungen? Woran kann man eine Lernleistung überhaupt festmachen?
- Woher nehmen wir Beurteilungsmaßstäbe in akademischen Disziplinen, wenn sich der akademische Diskurs ständig weiterentwickelt und sich Antworten auf wesentliche Fragen immer wieder verschieben?⁶⁷

67 Bei dieser Frage handelt es sich um eine Umformulierung der am Ende von Kap. 3.3.2 erläuterten hochschuldidaktisch zentralen epistemologischen Antinomie.

Lokale Kernfragen:

- Wie können wir mit den widersprüchlichen Ansprüchen von formativer und summativer Beurteilung konstruktiv umgehen (beurteilen, um ein- und auszuschließen vs. beurteilen, um Lernen zu fördern)?
- Wie können wir angesichts des heterogenen Vorwissens und der Diversität von Studierenden sowie unterschiedlichen Bezugsnormen eine faire und transparente summative Beurteilung gewährleisten?
- Wie können Beurteilungskriterien akademischer Leistungsnachweise die im Fluss befindliche Entwicklung aktueller Forschungsergebnisse, Theorien und Ansätze der jeweiligen Fachgebiete angemessen berücksichtigen?

5.3.5 Nachhaltige Erkenntnisse

Als Expertinnen und Experten in unseren Fachgebieten haben wir in der Regel einen blinden Fleck in Bezug auf den Wissenserwerb. Es fällt uns schwer, uns in die Perspektive von Novizen auf unserem Gebiet einzudenken. Wir haben unser Wissen und Können so sehr verinnerlicht, dass wir uns meist nicht mehr gut an die Schwierigkeiten erinnern können, die wir bei der ersten Begegnung mit den fachlichen Fragen überwinden mussten (siehe das Phänomen der Expertenblindheit in Kap. 8.4.1). Dies kann zu einer Verständnis- lücke zwischen Expertinnen und Novizen führen, bei der es Ersteren schwerfällt, Letzteren Konzepte mit einfachen Worten zu erklären oder bestehende Wissenslücken zu erkennen (vgl. Collins, Brown u. Newman 1989, 454–460; Wiggins u. McTighe 2005, 138f.).

Umso wichtiger ist es, dass wir uns gut überlegen, was Studierende am Ende verstanden haben sollten. Wir sprechen hier von nachhaltigen Erkenntnissen. Sie sind dauerhaft erworben, aber offen für Anpassungen und Erweiterungen (Wiggins u. McTighe sprechen von „enduring understandings“). Angestrebt wird also eine Form von Wissen und Können, die dem Aufbau komplexer Wissensstrukturen entspricht, wie sie in Kapitel 4.6 beschrieben sind. Basierend auf den vorangehenden Elementen – etablierte Ziele, Leitkonzept und Kernfragen – weist eine nachhaltige Erkenntnis folgende Merkmale auf (vgl. Wiggins u. McTighe 2005, 128f.):

- Sie ist eine wichtige fachliche Schlussfolgerung, die aus der Erfahrung, dem Wissen und Können von Expertinnen abgeleitet ist.
- Sie bezieht sich auf ein Leitkonzept (oder mehrere).
- Sie beinhaltet abstrakte Ideen, die für fachliche Laien oft kontraintuitiv oder missverständlich sind.

- Sie wird am besten durch „Aufdecken“ erarbeitet, das zugrundeliegende Verständnis muss also durch komplexe Lernvorgänge anhand möglichst authentischer Lernaufgaben durch die Lernenden aufgebaut werden.
- Sie fasst wichtige strategische Prinzipien in einem Könnensbereich zusammen.

Zum Aufdecken gehört auch der Erwerb von Fachbegriffen. Ihre Verwendung muss zuerst erworben werden, andernfalls wirken sie auf Novizen schnell als Barrieren, die das Verständnis behindern. Der letzte Punkt zielt darauf, dass einzuübende Fertigkeiten („Skills“) erst dann wirklich erfolgreich beherrscht und automatisiert werden können, wenn sie auf einer fachlichen Einsicht beruhen. Zu dieser gehört auch das Verständnis, wann welche Fertigkeit eingesetzt werden soll. Erst so setzen sich einzelne Fertigkeiten zu einem umfassenden Könnensbereich zusammen (vgl. Wiggins u. McTighe 2005, 129; 139).

Zu beachten ist, dass die Formulierungen von nachhaltigen Erkenntnissen noch nicht in erster Linie als Leitlinien für Studierende gedacht sind. Wiggins u. McTighe (vgl. 2005, 141) vergleichen sie mit dem Bauplan eines Architekten, der noch nicht die Materialien für den fertigen Bau definiert. Nachhaltige Erkenntnisse sind in diesem Sinne übergeordnete Lernziele, die sich auf verschiedene Ebenen der Lehrplanung beziehen können: Modulgruppen, Module oder einzelne Lehrveranstaltungen als Teile von Modulen – je nachdem, wie ein Studiengang strukturiert ist. Sie sind aber noch zu komplex und zu abstrakt, als dass sie direkt verwendet werden könnten, um konkrete Lernleistungen von Studierenden auf ihrer Basis einzuschätzen. Dazu ist der in Kapitel 5.4 beschriebene Schritt nötig, der im Formulieren spezifischer Lernziele besteht.

Nachfolgend seien zur Illustration noch je zwei nachhaltige Erkenntnisse im Anschluss an die oben angeführten Leitkonzepte und Kernfragen aus Literaturwissenschaft und Volkswirtschaftslehre aufgeführt:

Literaturwissenschaft: Die Studierenden werden verstehen, dass...

- ... Literatur die Gesellschaft, die Kultur und die menschliche Identität widerspiegelt und im Gegenzug das Welt-, Selbstbild und die Werte der Lesenden prägen kann.
- ... literarische Techniken von Rhythmik über Tropen wie Metapher oder Ironie bis zur narrativen Gestaltung prägen können, wie Lesende Kultur, Gesellschaft, Identität und Sprache wahrnehmen.

Volkswirtschaftslehre: Die Studierenden werden verstehen, dass...

- ... wirtschaftspolitische Maßnahmen wie Regulierungen und Interventionen erhebliche Auswirkungen auf die Verteilung von Ressourcen und Wohlstand in einer Gesellschaft haben können.

- ... Mindestlöhne, staatliche Wohlfahrtsprogramme und bildungspolitische Maßnahmen Möglichkeiten wirtschaftspolitischer Maßnahmen sind, die unterschiedliche Auswirkungen auf Beschäftigung, Einkommen, Armutsraten, soziale Mobilität, Ressourcenverteilung und Vermögensungleichheit haben können.

Beispiel (Fortsetzung von Kap. 5.3.1)*Nachhaltige Erkenntnisse (Auswahl)*

Die Studierenden werden verstehen, dass...

- ... eine effektive Leistungsbeurteilung verschiedene Ansätze umfasst, die Rückschlüsse auf das Wissen, die Fähigkeiten und die Einstellungen der Lernenden zulassen.
- ... Feedback ein wesentlicher Bestandteil des Beurteilungsprozesses ist und die Lehr- und Lernergebnisse stark beeinflussen kann.
- ... Beurteilungskriterien in akademischen Leistungsnachweisen die fließende Entwicklung von aktuellen Forschungsergebnissen, Theorien und Ansätzen in den jeweiligen Fachbereichen berücksichtigen müssen.

5.4 Spezifische Lernziele identifizieren und formulieren

Mit der Erarbeitung von Leitkonzepten, Kernfragen und nachhaltigen Erkenntnissen ist die Vorarbeit geleistet, um spezifische Lernziele für konkrete Lehrveranstaltungen zu formulieren. Auf deren Basis wiederum können wir dann formative und summative Leistungsnachweise entwickeln, Rückmeldungen geben und die gezeigten Lernleistungen beurteilen. Das Vorgehen wird hier wiederum als Neuformulierung der Ziele beschrieben – im Bewusstsein, dass es in vielen Fällen auch um die Überprüfung, Überarbeitung oder Ergänzung bestehender Zielformulierungen geht.

Dabei ist zu erwähnen, dass heute als Folge der Bologna-Reform Lernziele oft bereits vorformuliert sind – zumindest für Pflichtmodule und ihre Bestandteile. Zudem sind viele Lehrende darauf sensibilisiert, dass für Lehrveranstaltungen Lernziele nötig sind. Beides ist grundsätzlich begrüßenswert. Es kann allerdings dazu verleiten, von den Lernzielen aus direkt die Lehr-

Lernaktivitäten zu planen.⁶⁸ Im Sinne des Constructive Alignments ist aber zu beachten, dass zuerst die formativen und summativen Leistungsnachweise geplant werden sollen und erst dann die methodisch-didaktische Gestaltung der Lehr-Lernaktivitäten (siehe Kap. 5.1).⁶⁹

5.4.1 Grundsätzliche Hinweise zum Formulieren spezifischer Lernziele

Spezifische Lernziele beschreiben gemäß Kennedy (2006, 21), was die Lernenden am Ende einer Lernaktivität als Wissen und Können demonstrieren können. Sie sind damit eine Beschreibung dessen, was Lehrende von ihren Studierenden erwarten.⁷⁰

Bevor wir uns verschiedenen Lernzieldimensionen und -taxonomien widmen, werden in diesem Kapitel vier Aspekte erläutert, die für alle spezifischen Lernzielformulierungen wesentlich sind: 1) die Unterscheidung von verschiedenen Anforderungsniveaus, 2) handlungsorientierte Formulierungen, 3) verschiedene Lernzielbereiche und 4) das Verhältnis zwischen dem Ganzen und seinen Teilen.

Anforderungsniveaus von Lernzielen

Unabhängig von der konkreten fachlichen oder lernpsychologischen Situierung der Lernaktivitäten können empirisch zwei verschiedene Arten von Lernverhalten unterschieden werden: tiefen- und oberflächenorientiertes Lernen. Diese beiden Herangehensweisen an Lernaktivitäten korrelieren mit durch das Lernen verfolgten Zielen, dabei angewendeten Lernstrategien und

68 Abdeckungsorientierter Unterricht stützt seine Planung dagegen oft auf ein Skript oder Lehrbuch, das den Inhalt und die Reihenfolge des Unterrichts vorgibt. Dies ist wie in Kap. 5.1.1 erläutert problematisch. Lernmaterialien wie Skripts, Lehrbücher oder Online-Materialien sollten als Ressourcen genutzt werden, aber nicht aber als Lehrplan (vgl. Wiggins u. McTighe 2005, 21).

69 Die methodisch-didaktische Planung, Umsetzung und Evaluation von Lehr-Lern-Aktivitäten wird in diesem Buch bewusst ausgeklammert, da es eine Lücke zu schließen versucht, die dadurch entstanden ist, dass viele andere hochschuldidaktische Publikationen stärker auf die Lehr-Lern-Aktivitäten fokussieren. Dabei findet die Leistungsbewertung als „Elefant im Raum“ oft nur am Rande Beachtung.

70 Im englischen Sprachgebrauch wird häufig von learning outcomes („Lernergebnisse“) gesprochen. Dieser Begriff legt aus meiner Sicht die Illusion nahe, dass die Resultate von Lernprozessen von außen direkt steuerbar sind. Deshalb finde ich es angemessener, von Zielen zu sprechen. Die Ergebnisse sind das, was tatsächlich beim Lernen herauskommt – und bei jeder Lernenden unterschiedlich.

den Einstellungen zum Lernen. Bei tiefenorientiertem Lernen versuchen Studierende die Lerninhalte zu verstehen sowie unterschiedliche Aspekte der zu lernenden Thematik zueinander in Beziehung zu setzen. Zudem verknüpfen sie Gelerntes mit ihrem bereits vorhandenen fachlichen Wissen und Können sowie ihrem Allgemeinwissen über die reale Welt. Bei oberflächenorientiertem Lernen versuchen Studierende hingegen lediglich, Lernstoff zu reproduzieren. Typische Strategien sind dabei das Auswendiglernen von Lerninhalten und von schablonenartigen Problemlösungen (vgl. Trigwell u. Prosser 2020, 7). Nur tiefenorientiertes Lernen führt zuverlässig zum Aufbau von komplexen Wissensstrukturen. Zudem korreliert es mit positiveren Gefühlen gegenüber den Lerninhalten, einer größeren Überzeugung, erfolgreich lernen zu können und einer höheren Erfolgsmotivation. Demgegenüber korreliert oberflächenorientiertes Lernen mit dem Gefühl, die Dinge nicht zu verstehen, geringerem Vertrauen in die eigenen Lernfähigkeiten, geringerer Erfolgsmotivation – und folgerichtig mit geringerem Lernerfolg (vgl. Trigwell u. Prosser 2020, 19–24).

Der Sinn von Lernzieltaxonomien besteht darin, verschiedene Anforderungsniveaus von Lernzielen zu definieren. Es geht also um Abstufungen in Bezug auf die Durchdringung und Beherrschung der zu lernenden Wissensinhalte und Könnensbereiche. So wissenschaftsbasiert sie auch sein mögen, solche Taxonomien sind stets als Heuristiken zu verstehen. Denn Wissen und Können sind immer kontextspezifisch und können per se nicht abstrahiert und in allgemeine Kategorien eingeteilt werden. Die Berechtigung solcher heuristischen Taxonomien ergibt sich deshalb aus ihrem Nutzen für die Bildungspraxis (vgl. Krathwohl, Bloom u. Masia 1975, 59).

Die wesentlichste Unterscheidung, die Taxonomien zu treffen erlauben, ist folglich jene zwischen komplexeren und weniger komplexen Wissens- und Könnensstufen. Auf der motivationalen Ebene bieten sie damit den Studierenden – je nach ihrer Einstellung zum Lernen intrinsisch wie extrinsisch – Anreize zu tiefenorientiertem Lernen, wo dieses erwünscht ist.

Operationalisierung: Handlungsorientiert formulieren

Lernziele sollten stets handlungsorientiert formuliert werden: Sie beschreiben das beobachtbare Handeln oder dessen Produkt. Dadurch ermöglichen sie einen Rückschluss darauf, inwieweit das dem beschriebenen Handeln zugrunde liegende, aber nicht direkt beobachtbare Wissen und/oder Können erworben wurde (siehe Kapitel 6.1 zum Unterschied zwischen Kompetenz und Performanz).

Dabei haben sich folgende Empfehlungen bewährt, die aber aufgrund der heuristischen Natur von Lernzieltaxonomien nicht mechanistisch⁷¹, sondern reflektiert umgesetzt werden sollten – immer mit dem Ziel, die wichtigsten angestrebten Folgen des Lehrens und Lernens nachvollziehbar zu formulieren (vgl. Bachmann 2018a, 41; Biggs u. Tang 2007, 83f.):

- Lernziele sollten beobachtbares Verhalten bzw. beobachtbare Verhaltensresultate beschreiben.
- Sie sollten aktiv formuliert sein, also beschreiben, was die Lernenden wissen und können sollen (es geht primär um das Handeln der Lernenden als Subjekte und erst sekundär um die Beschreibung von fachlichen Inhalten als Objekte dieser Handlungen).
- Sie sollten auf eine bestimmte Stufe einer Lernzieltaxonomie zielen.
- In der Regel hilft es, wenn nur ein aktives Verb pro Lernziel verwendet wird, und dieses Verb möglichst spezifisch das angestrebte Handeln beschreibt.

Gelegentlich werden in hochschuldidaktischen Publikationen und Handreichungen Verbenlisten für die Formulierung von Lernzielen angeboten (vgl. z. B. Bachmann 2018a, 42f. mit Vorschlägen in Deutsch und Englisch). Das mag zur Überwindung von Formulierungsblockaden hilfreich sein, birgt aber wiederum die Gefahr formelhafter, mechanistischer Formulierungen. Insofern empfehle ich, auf der Basis der erarbeiteten übergeordneten Lernziele (siehe Kap. 5.3) möglichst eigenständige Formulierungen zu suchen.

Verschiedene Lernzielbereiche

Die überwiegende Mehrzahl aller Lernziele an Hochschulen zielt auf den kognitiven Bereich. Es gibt aber auch andere Bereiche des Wissens und Könnens: Bereits die Autorengruppe der ältesten kognitiven Lernzieltaxonomie (Bloom et al. 1956) unterschied zwischen dem kognitiven, affektiven und psychomotorischen Bereich (vgl. Anderson u. Krathwohl 2001, 258). So enthalten letztlich alle kognitiven (und weiteren) Lernziele immer auch affek-

71 Gelegentlich werden Lernzielformulierungen durch verwaltungsnahe Vorgaben eingefordert. In solchen Kontexten wird manchmal die Lernzielformulierung sehr dogmatisch interpretiert, so dass etwa Formulierungen aus rein formalen Gründen zurückgewiesen werden, wenn sie z. B. bestimmte Verben (nicht) enthalten. Bei solchen Anwendungen von Lernzieltaxonomien geht der ursprüngliche Sinn des Formulierens von Lernzielen verloren: Wenn wir Lernziele lediglich formulieren, „um etwa formale Anforderungen (wie z. B. in einem Modulhandbuch) zu befriedigen, ist diese Arbeit nicht mehr als eine Art von Fleißübung“ (Brendel, Hanke u. Macke 2019, 47).

tive Komponenten, und in diversen Fachgebieten spielen auch psychomotorische Lernziele eine wesentliche Rolle. Unmittelbar einsichtig ist dies bei vielen künstlerischen Studienfächern (Theater, Tanz, Musik) oder im sportlichen Bereich, es gilt aber beispielsweise auch für chirurgische Fähigkeiten in der Medizin (vgl. Moulton et al. 2006). Deshalb werden diese drei Lernzielbereiche nachfolgend näher vorgestellt.⁷²

Das Ganze ist mehr als seine Teile

Sinn und Zweck von spezifischen Lernzielen ist es, übergeordnete Lernziele in beobachtbare Komponenten zu unterteilen und somit einer Analyse zugänglich zu machen. Die daraus folgende analytische Aufteilung darf aber nicht darüber hinwegtäuschen, dass sinnhaftes Lernen eben gerade mehr ist als das Aneinanderreihen einzelner Wissens Elemente oder Techniken (Skills). Erst der in Kapitel 4.6 beschriebene Aufbau komplexer Wissensstrukturen führt dazu, dass Studierende nachhaltige Erkenntnisse und darauf basierende Handlungs- und Reflexionsmöglichkeiten aufbauen. Aus diesem Grund ist die gründliche Formulierung übergeordneter Lernziele für die Entwicklung von Leistungsnachweissettings und von lernzielbasierten Bewertungskriterien ebenso wichtig wie die hier adressierten spezifischen Lernziele (vgl. Wiggins u. McTighe 2005, 59).

5.4.2 Kognitive Lernziele

Vorgestellt werden zwei Taxonomien für kognitive Lernziele: Die SOLO-Taxonomie (Biggs u. Collis 1982) und die revidierte Version von Blooms Taxonomie kognitiver Lernziele (Anderson u. Krathwohl 2001). Vorgängig sind zwei Aspekte zu betonen:

- Besonderes Augenmerk sollte auf die Unterscheidung zwischen deklarativem und prozeduralem Wissen gelegt werden: Wo sollen die Studierenden Wissensstrukturen erwerben (Wissen, knowing that) und wo sollen sie Fertigkeiten erwerben (Können, knowing how)? Um übergeordnete Lernziele im Sinne von Kapitel 5.3 zu erreichen, sollten sich in aller Regel deklarative und prozedurale Wissens Elemente, also Wissen und Können, ergänzen.

72 Ob wirklich nur diese drei Bereiche zu unterscheiden seien, ist nicht abschließend geklärt. So unterscheidet etwa Gardner (1993) sieben verschiedene Arten von Intelligenzen, was nahelegen würde, dass es auch mehr als drei Taxonomiebereiche für Lernziele geben sollte. Diese Fragestellung führt aber über unseren Kontext hinaus.

- In gewissen Fachgebieten hält sich hartnäckig die lernpsychologisch nicht haltbare Überzeugung, zu Beginn eines Studiums müssten zunächst einmal „fachliche Grundlagen“ in Form der Anhäufung von möglichst viel deklarativem Wissen auf den untersten Lernzielstufen gelegt werden. Es stimmt zwar, dass oberflächliches Verstehen wie etwa die Kenntnis wesentlicher Begriffe die Grundlage für tiefergehendes Verstehen darstellt (vgl. Hattie u. Zierer 2017, 68). Aber wird in einer Lehrveranstaltung oder gar einem ganzen Grundlagenjahr vorrangig deklaratives Begriffswissen ohne Anwendungen oder Einbettung in authentische Situationen vermittelt, entsteht in erster Linie träges Wissen oder „knowing the name of something“ (siehe Kap. 4.2). Es ist deshalb wichtig, bereits vom ersten Semester an stets eine Kombination deklarativer und prozeduraler Ziele auf der ganzen Bandbreite der Lernzieltaxonomien anzustreben (vgl. Bach u. Zellweger 2022, 83f.); siehe dazu auch Kapitel 10 zur Curriculumentwicklung.

SOLO-Taxonomie

Die SOLO-Taxonomie (Biggs u. Collis 1982) ist eine weltweit vielerorts in der Hochschulbildung verwendete Taxonomie kognitiver Lernziele. Im deutschen Sprachraum ist sie bislang bedauerlicherweise weniger verbreitet als jene von Anderson u. Krathwohl (2001). Ihre Vorteile gegenüber der letztgenannten Taxonomie liegen in der einfacheren Handhabung und der größeren Klarheit (vgl. Koskinen 2007; Whalley et al. 2006; Hazel, Prosser u. Trigwell 2002). Sie kann zudem sehr gut zur Konstruktion von Erfolgskriterien und Beurteilungsrastern verwendet werden (vgl. Hattie u. Zierer 2017, 121–124; Beurteilungsraster werden in Kap. 9.3 vorgestellt). Auch zur Priorisierung der Lernziele kann die SOLO-Taxonomie genutzt werden (Biggs u. Tang 2007, 83): Wichtige Themen sollten auf einer höheren Taxonomiestufe erworben werden als weniger zentrale Themen.

Die SOLO-Taxonomie besteht aus fünf hierarchischen Stufen, die von Unkenntnis bis zu umfassendem Wissen und Können reichen (siehe Abb. 14). Wichtig ist, dass die Stufen nicht einen linearen Fortschritt darstellen – der Lernfortschritt zwischen Stufe 1 und 2 ist nicht unbedingt gleich groß wie jener zwischen Stufe 3 und 4 (vgl. Biggs u. Tang 2007, 76–80; Biggs u. Collis 1982, 17–31; 185–205):

1. *Unstrukturiert*: Auf dieser Stufe hat der Lernende das zu vermittelnde Konzept oder die zu vermittelnde Fähigkeit noch nicht verstanden. Der Lernende zeigt Missverständnisse, mangelndes Verständnis oder unvollständiges Wissen in Bezug auf das Lernziel.

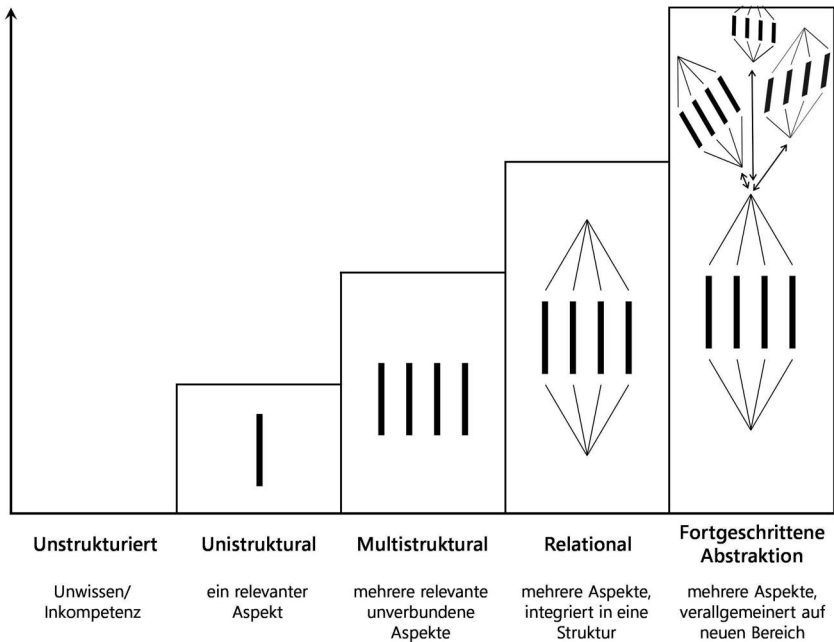


Abb. 14: Die SOLO-Taxonomie für kognitive Lernziele nach Biggs u. Collis (1982), Biggs u. Tang (2007) und Biggs (2023). Quelle: Eigene Abbildung, angelehnt an Biggs u. Tang (2007, 79) und Biggs (2023)

2. *Unistruktural:* Die Lernende hat ein grundlegendes oder oberflächliches Verständnis des Konzepts oder der Fertigkeit. Sie kann sich an einzelne Informationen erinnern oder sie identifizieren, aber erkennt noch nicht die Beziehungen zwischen ihnen. Das Verständnis oder die Fertigkeit beschränkt sich auf einen oder sehr wenige Aspekte des Lernziels.
3. *Multistruktural:* Der Lernende kann auf dieser Stufe mehrere Aspekte oder Teile eines Konzepts oder einer Fertigkeit identifizieren, beschreiben oder auflisten, aber es fehlt noch an der Integration oder Synthese der Informationen – der Lernende versteht noch nicht näher, wie sie miteinander verbunden sind oder wie sie zu einem größeren Ganzen beitragen.
4. *Relational:* Hier ist die Lernende zu induktiven Operationen fähig. Sie kann Verbindungen herstellen, Beziehungen zwischen verschiedenen Aspekten des Konzepts oder der Fertigkeit erkennen und ihre Zusammenhänge oder ihr Zusammenwirken erklären (Mustererkennung). Die Lernende integriert und synthetisiert Informationen; sie kann Schlussfolge-

rungen ziehen, die auf erkannten Beziehungen zwischen den durch das Lernziel adressierten Elementen beruhen.

5. *Fortgeschrittene Abstraktion:* Der Lernende ist zu induktiven und deduktiven Denkopoperationen in der Lage. Er kann über das Lernziel hinausgehen, das durch dieses adressierte Wissen und Können auf neue Kontexte übertragen und es in neuer und/oder kreativer Weise anwenden. Der Lernende analysiert und bewertet Informationen in Bezug auf Aspekte des Lernziels kritisch. Er zeigt ein hohes Maß an Selbstständigkeit und Unabhängigkeit bei der Bearbeitung von Fragestellungen im Zusammenhang mit dem Lernziel.

Es folgen Beispiele für spezifische kognitive Lernziele, die zu den in Kapitel 5.3 aufgeführten Beispielen aus der Literaturwissenschaft und der Volkswirtschaftslehre passen. Die Studierenden werden dabei mit der Sie-Form direkt als Adressaten angesprochen.

Volkswirtschaftslehre:

- Sie beschreiben wirtschaftspolitische Maßnahmen, welche die Verteilung von Ressourcen und Wohlstand in einer Gesellschaft beeinflussen können (multistruktural, deklarativ).
- Sie beschreiben die mutmaßlichen Auswirkungen einer wirtschaftspolitisch motivierten Intervention in das Bildungssystem auf die gesellschaftliche Ressourcenverteilung und die Ungleichheit des Wohlstands (relational, deklarativ).
- Sie entwickeln einen eigenen wirtschaftspolitischen Maßnahmenplan, um bestimmte Auswirkungen auf die gesellschaftliche Ressourcenverteilung und die Ungleichheit des Wohlstands zu erzielen (relational, prozedural).

Literaturwissenschaft:

- Sie verfassen ein Textbeispiel, das eine bestimmte literarische Technik einsetzt, z. B. Rhythmik, Tropen, narratologische Aspekte (multistruktural, prozedural).
- Sie vergleichen literarische Texte aus verschiedenen literaturhistorischen Epochen, können Unterschiede bezüglich des Gebrauchs literarischer Techniken herausarbeiten und diese anhand Ihres literaturhistorischen Wissens erklären (relational, prozedural).

Beispiel (Fortsetzung von Kap. 5.3.1)*Spezifische Lernziele basierend auf der SOLO-Taxonomie (Auswahl)*

Welche der folgenden Ziele tatsächlich eingesetzt werden, hängt davon ab, welches Niveau an Durchdringung und Beherrschung mit Bezug auf welche Aspekte der Lerninhalte erwartet wird:

Deklaratives Wissen auf der multistrukturellen Stufe:

- Sie identifizieren verschiedene Formen von Leistungsnachweisen zur Beurteilung von Wissen und Können der Lernenden.
- Sie benennen die wesentlichen Bestandteile von Feedback, das Lehr- und Lernergebnisse erfolgreich beeinflusst.
- Prozedurales Wissen auf der multistrukturellen Stufe:
- Sie entwickeln Formen von Leistungsnachweisen, die das Wissen und Können von Lernenden zu fördern und beurteilen ermöglichen.
- Deklaratives Wissen auf der relationalen Stufe:
- Sie vergleichen verschiedene Formen von Leistungsnachweisen und beurteilen ihre Stärken und Schwächen.
- Sie beurteilen Feedback an Lernende daraufhin, ob es zur Verbesserung von deren Lernleistung beiträgt.
- Prozedurales Wissen auf der relationalen Stufe:
- Sie entwickeln ein Leistungsnachweis-Setting, das verschiedene Formen von Leistungsnachweisen und -beurteilung umfasst, um Wissen und Können von Lernenden zu fördern und zu beurteilen.
- Sie erteilen Feedback an Lernende, das die Bedingungen erfüllt, um deren Lernen zu fördern.

Revidierte Taxonomie kognitiver Lernziele nach Bloom

Die revidierte Fassung der kognitiven Lernzieltaxonomie nach Bloom von Anderson u. Krathwohl (2001) besteht aus einer hierarchischen⁷³ Abfolge von Kategorien vom Erinnern von Informationen bis zum Schaffen von neuem Wissen oder neuen Artefakten. Die hierarchischen Taxonomiestufen wurden durch eine Wissensdimension ergänzt, so dass die Taxonomie zu einer Matrix wurde (siehe Abb. 15). Wie die SOLO-Taxonomie ist auch die Taxonomie nach Anderson u. Krathwohl (2001) geeignet, um spezifische Lernziele für Wissen und Können auf der Ebene konkreter Lehrveranstaltungen zu formulieren. Zur Vermeidung von Redundanz wird hier auf die Darstellung von Beispielformulierungen verzichtet.

Dimensionen von Wissen	Stufen kognitiver Prozesse					
	Erinnern	Verstehen	Anwenden	Analysieren	Beurteilen	Gestalten
Deklaratives Wissen						
Konzeptionelles Wissen						
Prozedurales Wissen						
Metakognitives Wissen						

Abb. 15: Revidierte Taxonomie kognitiver Lernziele nach Bloom. Quelle: Eigene Abbildung, in Anlehnung an Anderson u. Krathwohl 2001, 28

5.4.3 Affektive Lernziele

Bloom und seine Kollegen haben auch eine Taxonomie für den affektiven Bereich entwickelt, die sich mit emotionalen Aspekten des Lernens befasst (Krathwohl, Bloom u. Masia 1975). Ihre Stufen reichen von der grundsätzlichen Bereitschaft, Informationen aufzunehmen, bis hin zur Integration von neuen Überzeugungen, Ideen und Haltungen in das eigene Wertesystem. Die fünf Stufen umfassen (vgl. Krathwohl, Bloom u. Masia 1975, 87–181; Kennedy 2006, 35f.; Dubs 1995, 319):

73 Im Gegensatz zur ursprünglichen Taxonomie wird die revidierte Fassung zwar als hierarchisch, aber nicht als kumulative Hierarchie verstanden. Eine höhere Stufe muss folglich nicht zwingend alle Aspekte der vorangehenden Stufen umfassen. Dabei gibt es Evidenz, dass die Stufen 2 bis 4 (Verstehen, Anwenden, Analysieren) kumulativ hierarchisch sind – für die oberen Stufen 4 bis 6 lässt sich dies hingegen nicht belegen (vgl. Anderson u. Krathwohl 2001, 267).

1. Aufmerksam werden: Lernende sind bereit, Informationen aufzunehmen und auf andere mit Respekt einzugehen.
2. Reagieren: Lernende gestalten ihren Lernprozess aktiv, entwickeln Interesse am Thema oder sind bereit, an Diskussionen teilzunehmen.
3. Bewerten: Lernende akzeptieren und übernehmen (neue) Werte, etwa (als Teil der) Wertschätzung demokratischer Prozesse oder Sensibilität für kulturelle und individuelle Unterschiede.
4. Einordnung: Lernende lösen anhand von Lernvorgängen Konflikte zwischen verschiedenen Werten, führen Werte zu einer Wertordnung zusammen und richten ihr Verhalten nach dieser aus.
5. Bestimmtsein durch Werte: Lernende verfügen über ein kohärentes Wertesystem („Weltanschauung“), das ihr Verhalten prägt. Dazu gehört auch die Anpassung der eigenen Weltanschauung und des darauf basierenden Verhaltens aufgrund von Lernvorgängen.

Wesentlich ist, dass sich kognitive und affektive Lernvorgänge gegenseitig durchdringen – die eine Dimension ist ohne die andere nicht denkbar. Dieser Zusammenhang zeigte sich bereits in Kapitel 4.5.2 anhand der Bedingtheit von Lernvorgängen durch subjektive Bedeutsamkeit.

So ist Stufe 1 der affektiven Taxonomie, das bloße Aufmerksamwerden, Voraussetzung dafür, sich an etwas zu erinnern, also kognitiv die unilaterale Stufe der SOLO-Taxonomie bzw. die Stufe 1 „Erinnern“ der revidierten Bloom-Taxonomie zu erreichen. Im Gegenzug müssen Werte (vom ethisch-moralischen bis zum künstlerischen Bereich) kognitiv auf höherer Stufe verarbeitet werden, um auf reflektierte Weise in eine Wertordnung eingefügt zu werden. Stufe 4 „Einordnen“ der affektiven Taxonomie setzt somit das Erreichen der relationalen SOLO-Stufe bzw. der Stufe 4 „Analysieren“ der revidierten Bloom-Taxonomie voraus (vgl. Krathwohl, Bloom u. Masia 1975, 46–59; Dubs 1995, 317–321).

Literaturwissenschaft:

- Reagieren: Sie äußern Ihre Meinung über die sozialen, kulturellen und ethischen Implikationen literarischer Texte, indem Sie an Gruppendiskussionen teilnehmen.
- Bewerten: Sie erkennen und würdigen die Rolle der Literatur bei der Gestaltung Ihrer eigenen persönlichen Werte und Überzeugungen, indem Sie darüber nachdenken, wie Ihre eigenen Werte durch literarische Texte beeinflusst wurden.

Volkswirtschaftslehre:

- Bewerten: Sie erkennen die Komplexität wirtschaftspolitischer Entscheidungen und die Notwendigkeit, konkurrierende Interessen und Werte auszugleichen, an.
- Einordnung: Sie bewerten wirtschaftspolitische Vorschläge auf der Basis Ihres eigenen wirtschaftspolitischen Wertsystems anhand Ihres Fachwissens kritisch.

Beispiel (Fortsetzung von Kap. 5.3.1)

Spezifische affektive Lernziele

- Reagieren: Sie entwickeln Interesse an der Erkundung verschiedener Beurteilungsansätze und beteiligen sich aktiv an Diskussionen in der Seminargruppe.
- Bewerten: Sie wissen die Bedeutung von effektivem Feedback zu schätzen und zeigen die Bereitschaft, Feedback zum Lernen von anderen zu geben und solches zum eigenen Lernen aufzunehmen und umzusetzen.
- Einordnen: Sie entwickeln und beschreiben eine persönliche Beurteilungsphilosophie, die mit dem aktuellen Stand von Theorie und Forschung in Lernpsychologie und Hochschuldidaktik übereinstimmt. Sie beschreiben, wie Sie auf dieser Basis mit Lernleistungen von Studierenden umgehen und wie sich das auf das Lernen der Studierenden und ihre Motivation auswirkt.

5.4.4 Psychomotorische Lernziele

Im psychomotorischen Bereich geht es hauptsächlich um körperliche Fähigkeiten, die die Koordination von Gehirn- und Muskeltätigkeit beinhalten. Für diesen Bereich gibt es zwar diverse Taxonomien, von denen aber keine empirisch so gut fundiert ist wie SOLO oder die beiden Taxonomien der Gruppe um Bloom und Krathwohl. Der psychomotorische Bereich wird häufig in Bereichen wie Laborwissenschaften, Gesundheitswissenschaften, Kunst, Musik, Technik, Theater und Sportunterricht verwendet (vgl. Kennedy 2006, 37).

Eine der meistverwendeten Taxonomien psychomotorischer Lernziele ist jene von Elizabeth Simpson (1972), im Folgenden zitiert nach Kennedy (2006, 39). Sie umfasst die folgenden sieben hierarchischen Stufen:

1. Wahrnehmung: Die Lernenden haben die Fähigkeit, sensorische Hinweise zur Steuerung der körperlichen Aktivität zu nutzen.
2. Einstellung: Die Lernenden sind bereit, auf eine bestimmte Weise zu handeln (mentale, physische oder emotionale Disposition).
3. Geführte Reaktion: Die Lernenden können eine bestimmte körperliche Fertigkeit imitieren. Durch Wiederholung führt dies zu einem zunehmend besseren motorischen Ablauf.
4. Mechanismus: Die Lernenden gewöhnen sich die imitierten Fertigkeiten an und können sie zunehmend selbstgesteuert ausführen.
5. Komplexe, offene Reaktion: Die Lernenden beherrschen komplexe Bewegungsmuster und können verschiedene Abläufe präzise koordinieren. Ihre Reaktionen erfolgen automatisch.
6. Anpassungsfähigkeit: Die Lernenden können ihre motorischen Fertigkeiten modifizieren, um mit Problemsituationen umzugehen oder speziellen Anforderungen gerecht zu werden.
7. Erzeugung: Die motorischen Fertigkeiten der Lernenden sind so weit entwickelt, dass sie in besonderen Situationen kreativ Bewegungsmuster entwickeln können.

Hier einige ausgewählte psychomotorische Lernziele aus verschiedenen Studienfächern:

- Wahrnehmung (Chirurgie): Erkennen der verschiedenen Arten von chirurgischen Instrumenten durch Sehen und Tasten
- Einstellung (Tanz): Vorbereitung der Tanzfläche mit den erforderlichen Hilfsmitteln und Geräten vor der Aufführung
- Geführte Reaktion (Aussprache einer Fremdsprache): richtige Aussprache von Wörtern und Sätzen, nachdem man einem Muttersprachler zugehört und die richtige Aussprache geübt hat
- Mechanismus (Chirurgie): eine laparoskopische Cholezystektomie fachgerecht durchführen
- Komplexe, offene Reaktion (Tanz): Choreografieren und Aufführen einer Tanznummer, die mehrere Tanzstile umfasst
- Anpassungsfähigkeit (Aussprache einer Fremdsprache): Änderung der Aussprache, um in einer Fremdsprache mit verschiedenen regionalen Akzenten effektiv zu kommunizieren

- Erzeugung (Chirurgie): Entwicklung einer neuen chirurgischen Technik zur Verbesserung der Behandlungsergebnisse bei einem bestimmten Verfahren

Auf die Weiterführung des Beispiels aus Kapitel 5.3.1 wird an dieser Stelle verzichtet, da psychomotorische Lernziele dafür eine untergeordnete Rolle spielen.

5.5 Epilog: Von Lernzielen und Kompetenzen

Wie Brendel, Hanke u. Macke (2019, 45) feststellen, verzichten „international prominente Autoren und Autorinnen (...) darauf, den Kompetenzbegriff zu verwenden“. Das hat aus meiner Sicht gute Gründe, die nachfolgend erläutert werden. Deshalb verwende ich in diesem Buch anstelle des Kompetenzbegriffs die Formulierung „Wissen und Können“. Diese Begrifflichkeit ist auch hilfreicher für die Beurteilung von Lernzielen, während die Unschärfe des Kompetenzbegriffs oft den Blick auf die Lernziele vernebelt.

5.5.1 Kompetenzen und das Formulieren von Lernzielen

Auch wenn der Kompetenzbegriff in der deutschsprachigen Pädagogik und Didaktik breit etabliert ist, wird er in unseren Breitengraden durchaus hinterfragt. So vertritt Reinmann (2015, 30) die These: „Ein Verzicht auf den Kompetenzbegriff wäre hochschuldidaktisch vorteilhaft.“ In der Pädagogik wird der Begriff dafür kritisiert, dass er die Komplexität von Lernvorgängen in institutionellen Lernsituationen zu wenig berücksichtigt.⁷⁴ Dies lässt sich exemplarisch an der am weitesten verbreiteten Kompetenzdefinition zeigen:

Kompetenzen sind die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können. (Weinert 2001, 27f.)

Eine solche Definition legt, zumindest wenn sie ohne Kontext verwendet wird, schnell eine Art Steuerungslogik nahe. Resultate von Lernprozessen

74 Diese Schwierigkeiten sah bereits Faulstich (2002, 15) kommen: „Schon der Qualifikationsbegriff hat nicht gehalten, was er verspricht: Nämlich eine gegenüber dem (...) hochbelasteten Bildungsbegriff gesteigerte (...) empirische Fundierbarkeit. (...) [D]er Begriff Kompetenz droht ebenfalls zunehmend hohl zu werden.“

sind aber höchstens bedingt vorhersehbar und steuerbar, wie in Kapitel 3.3.2 anhand von Antinomien des Lehrhandelns gezeigt wurde (vgl. Reinmann 2015, 19–21).⁷⁵

Im Kontext des Kompetenzbegriffs wird auch häufig Wissen gegen Können ausgespielt. Kompetenz steht dann für die Überwindung trägen Wissens, wie es in Kapitel 4.2 beschrieben wurde – basierend auf dem Fehlschluss, wer für das Können sei, müsse gegen das Wissen sein (vgl. Reinmann 2015, 22). Das ist freilich sowohl lernpsychologisch (siehe Kap. 4) wie auch pädagogisch (siehe Kap. 3.3.2) schlicht unsinnig. Ironischerweise vertrat auch Weinert selbst diese Auffassung. Können als Ableitung und Generalisierung von Regeln komme laut ihm nicht zustande, „wenn sich die Regeln nicht auf Inhalte beziehen, in engem Zusammenhang mit dem Aufbau von Wissenssystemen flexibel erworben werden und auf diese Weise ihren operativen Stellenwert für die Lösung inhaltlicher Probleme erhalten“ (Weinert 1994, 192). Können ohne Wissen ist vergleichbar mit dem Stricken ohne Wolle (vgl. Schiefner u. Weil 2010, 38), weshalb es auch Schlüsselkompetenzen nicht wirklich geben kann – jedes Können ist immer in erheblichem Maß kontextgebunden, nämlich durch Wissensstrukturen. Deshalb spricht man heute nicht mehr von Schlüssel-, sondern von überfachlichen Kompetenzen (vgl. Bachmann 2018b, 23–25; Schiefner u. Weil 2010).⁷⁶

Mit überfachlichen Kompetenzen sind in der Regel Methoden-, Sozial- und Selbstkompetenzen gemeint (vgl. Bachmann 2018b, 24). Ich vertrete die Ansicht, dass durch eine umsichtige Lernzielplanung, wie sie oben beschrieben ist, diese Aspekte ebenfalls angesprochen sein sollten. So umfassen Selbst- und Sozialkompetenzen oft auch affektive Aspekte, und methodische Aspekte sind in einem wissenschaftlichen Studium so oder so untrennbar mit fachlichen Aspekten verbunden. Insofern kann man das Denken in diesen drei Kategorien überfachlicher Kompetenzen aus meiner Sicht zur Überprüfung verwenden, ob die Lernziele einseitig auf (deklarative) fachliche Wissensinhalte ausgerichtet sind. Gezielt solche Kompetenzen zusätzlich zu „fachlichen“ Lernzielen zu adressieren, führt meiner Meinung nach allerdings gerade zu der Spaltung, die sie eigentlich überwinden helfen sollen – mit den Worten von Bachmann (2018b, 23): „Fachlicher und überfachlicher Kompetenzerwerb bedingen sich (...) gegenseitig.“

Sinnvoll ist der Begriff der Kompetenz für unseren Kontext hingegen zur Unterscheidung zwischen dem, was eine Person grundsätzlich kann, und ihrem tatsächlich gezeigten, beobachtbaren Verhalten. Kompetenz meint in

75 Im lesenswerten Text von Reinmann (2015) finden sich diverse Kritikpunkte zum Kompetenzbegriff und weiterführende Literaturhinweise dazu.

76 Aus den gleichen Gründen machen z. B. Einführungen ins wissenschaftliche Schreiben erst zu dem Zeitpunkt Sinn, wenn die Studierenden tatsächlich eine erste Arbeit verfassen.

dieser Gegenüberstellung das nicht direkt beobachtbare grundsätzliche Können, während das konkret gezeigte Verhalten als Performanz bezeichnet wird (siehe Kap. 6.1).

5.5.2 Von den Lernzielen zum Beurteilen

Lernziele definieren das Wissen und Können, das als Ergebnis des Lernprozesses erwartet wird, den Studierende durchlaufen. Im Constructive Alignment (siehe Kap. 5.1.2) bilden Lernziele die Grundlage, um zu überlegen, wie Studierende zeigen können, ob sie das gelernt haben, was sie lernen sollten. Anhand dieser Überlegung können wir schließlich Leistungsnachweise gestalten, also Settings, in denen Studierende ihr neu erworbenes Wissen und Können zeigen können. Das Gestalten verschiedener Formate von Leistungsnachweisen wird in Kapitel 7 ausführlich erläutert. Zuvor sind allerdings einige grundlegende Fragen zu klären, wie von gezeigten Leistungen auf das dahinterliegende Können geschlossen werden kann. Kapitel 6 erläutert deshalb nicht nur, an welchen Maßstäben das Erreichen von Lernzielen gemessen werden kann. Es thematisiert auch die Gütekriterien für angemessene und zuverlässige Beurteilungen – also Beurteilungen, die sich an den Lernzielen orientieren.

6 Voraussetzungen einer hochwertigen Leistungsbeurteilung

Um sinnvolle Formate von Leistungsnachweisen auszuwählen, neu zu gestalten oder zu kombinieren, sollten wir uns einiger wesentlicher Voraussetzungen bewusst sein. Dies beginnt beim erkenntnistheoretischen Umstand, dass Wissen und Können nur mittelbar beobachtet werden können (Kap. 6.1). Aufgrund dieses erschwerenden Umstands ist es wichtig, sich der Vergleichsebene bewusst zu sein, anhand deren eine Leistung beurteilt werden soll – wir sprechen von so genannten Bezugsnormen (Kap. 6.2). Zudem ist es zentral, dass Leistungserhebungen und -beurteilungen wesentlichen Gütekriterien genügen (Kap. 6.3), weil aufgrund der nur mittelbaren Zugänglichkeit von Wissen und Können das Beurteilen von Leistungen anfällig für die Verzerrung von Wahrnehmungen und deren Interpretationen ist (Kap. 6.4). Die Kenntnis dieser Grundlagen ermöglicht das Ergreifen von Maßnahmen zur Gestaltung qualitativ hochwertiger, also aussagekräftiger Leistungsbeurteilungen (Kap. 6.5).

6.1 Wissen und Können als latente Konstrukte

Die Hauptschwierigkeit beim Umgang mit Lernleistungen besteht darin, dass das hinter gezeigten Leistungen liegende Wissen und Können keiner direkten Beobachtung zugänglich ist. Wie in Kapitel 2.5.1 erläutert, sprechen wir deshalb von latenten Konstrukten. Wir können beobachten, wie jemand bestimmte Aufgabenstellungen bearbeitet, haben aber keinen Zugriff auf die neurologische Basis gezeigter Handlungen und erstellter Produkte. Mit Bezug auf die Sprachproduktion unterschied deshalb Chomsky (vgl. 1965, 4) zwischen Kompetenz und Performanz: Unter Kompetenz verstand er das sprachliche Wissen und Können, über das eine Person bewusst und unbewusst verfügt, unter Performanz hingegen die konkrete Produktion sprachlicher Äußerungen. Beobachtbar ist nur die Performanz, auf die Kompetenz muss anhand der beobachteten Performanz mittelbar geschlossen werden (Hinweis: Dieser Kompetenz-Begriff ist nur partiell verwandt mit dem in Kap. 5.5 kritisch diskutierten pädagogischen Kompetenzbegriff).

Mit Bezug auf Leistungsnachweise bedeutet dies, dass wir immer nur gezeigtes Verhalten und erstellte Produkte beobachten, interpretieren, beurteilen oder bewerten können. Was uns aber eigentlich am meisten interessiert, ist das – kognitive, affektive, psychomotorische – Wissen und Können, das als

Kompetenz hinter der gezeigten Performanz liegt. Aus diesem Grund ist der in den Kapiteln 5.2 bis 5.4 beschriebene Prozess des Formulierens von Lernzielen so wichtig: Operationalisierte Lernziele sind quasi von uns Dozierenden eruierte Signale, anhand deren wir von gezeigtem Verhalten auf das dahinterliegende Wissen und Können als latentes Konstrukt schließen können (siehe den Abschnitt „Operationalisierung“ in Kap. 5.4.1).

Ein Problem besteht darin, dass aus Performanz zwar auf Kompetenz geschlossen werden kann – wenn Performanz ausbleibt oder wir sie nicht beobachten können, bedeutet das aber nicht zwingend, dass die angestrebte Kompetenz nicht vorhanden ist (vgl. Thomann 2017c, 149). Ursachen für trotz vorhandener Kompetenz nicht gezeigte oder nicht beobachtete Performanz gibt es viele. Typische Beispiele dafür sind im Hochschulalltag:

- „Ungeschickt“ formulierte Aufgabenstellungen: Der Klassiker aus dem Primarschulbereich sind Textaufgaben in Mathematik, die unangemessen hohe Anforderungen an das Leseverständnis stellen. Sie erlauben nicht, zwischen solchen Lernenden zu unterscheiden, die über das angestrebte mathematische Können (z. B. Bruchrechnen) verfügen, und solchen, die nicht darüber verfügen. Ihre Trennschärfe ist somit mangelhaft, da falsche Lösungen auch durch das Leseverständnis verursacht sein können (vgl. Ingenkamp u. Lissmann 2008, 57). Solche Elemente kommen auch in Leistungsnachweisen an Hochschulen regelmäßig vor, etwa wenn in Prüfungsaufgaben fachliche Konzepte verwendet werden, denen die Studierenden noch nicht begegnet sind. Während man im akademischen Kontext grundsätzlich erwarten kann und soll, dass Studierende selbständig Informationen zu noch nicht verstandenen Elementen einholen (z. B. durch Konsultation von Fachliteratur, Wikipedia, ChatGPT, Studienkolleginnen, Dozierenden), ist dies vor allem in Leistungsnachweis-Settings mit Zeitbeschränkung und/oder eingeschränkten Hilfsmitteln problematisch. Dann ist die Wirkung ähnlich wie beim Beispiel der mathematischen Textaufgabe: Die Aufgabe verliert an Trennschärfe und der Leistungsnachweis somit an Gültigkeit (siehe Kap. 6.3).
- Lernende zeigen das im Lernziel formulierte Verhalten mehr oder weniger absichtlich nicht, obschon sie über die zugrundeliegende Kompetenz verfügen würden.
 - Eine eher triviale Ursache können hierbei Missverständnisse sein. Diese können in interaktiven Settings wie mündlichen Gesprächen leichter behoben werden als in asynchronen, vereinzelnden Settings wie Klausuren.
 - Eine weitere Ursache kann mangelnde Motivation der Studierenden sein, gewünschtes Verhalten zu zeigen. Problematisch ist, dass die Motivation ebenfalls ein latentes Konstrukt ist, das nur mittelbar erschlossen werden kann – und bezüglich dessen Fehlschlüsse durch

Beurteilungsverzerrung besonders gerne auftreten (siehe Kap. 6.4). In eher persönlichen Settings kann es deshalb angezeigt sein, entsprechende Eindrücke in Form von Fragen anzusprechen: „Ich habe den Eindruck, dass Sie Verhalten X eigentlich beherrschen. Können Sie mir erklären, weshalb Sie Verhalten X nicht zeigen/gezeigt haben?“⁷⁷

- Die wohl problematischste Ursache sind Prüfungsängste, etwa wenn ein Student in einer Prüfungssituation aufgrund von Angst eine Blockade erlebt. Das Bedauerliche in diesem Fall ist, dass die Studierenden lernen und Leistung zeigen wollen, aber nicht können. Die Problematik der Prüfungsangst und was wir als Dozierende im Umgang mit betroffenen Studierenden tun können, wird in Kapitel 11 näher erläutert.
- Anders liegt das Problem, wenn Performanz zwar eigentlich gezeigt wird, aber aufgrund von Beobachtungs- und Beurteilungsverzerrungen durch Dozierende unangemessen interpretiert oder schlicht übersehen wird. Dieser Problematik ist das ganze Kapitel 6.4 gewidmet.

Leistungsnachweise sind institutionell durch Lehrende und ggf. weitere Beteiligte wie Mitarbeitende der Studienadministration konstruierte Settings (siehe Kap. 2.5.2). Wenn Performanz über eine ganze Studierendengruppe hinweg nicht wie angestrebt sichtbar wird, ist deshalb auch über Konstruktionsmängel von Aufgabenstellungen hinaus stets das Setting zu hinterfragen: Können die Bedingungen, unter denen die Studierenden die Leistungen zeigen sollen, für die Performanz hinderlich sein? Zudem muss in solchen Fällen natürlich das Constructive Alignment überprüft werden: Erlauben die Aufgabenstellungen tatsächlich, das in den Lernzielen definierte Verhalten zu beobachten (diese Frage überschneidet sich mit jenen nach der qualitativen Güte des Leistungsnachweises, siehe Kap. 6.3)? Erlaubten die in der Veranstaltung oder der Selbstlerneinheit angestrebten Lehr- und Lernaktivitäten den Studierenden tatsächlich, das den Lernzielen zugrunde liegende Wissen und Können zu erwerben?

Falls tatsächlich Performanz im gewünschten Sinne beobachtbar ist, so ist es dieses Verhalten, das wir beurteilen. Wir ziehen also anhand von fachlichen Kriterien und darauf basierenden Leistungsurteilen Schlüsse darüber, in welchem Maß und in welcher Qualität Studierende gesetzte Lernziele erreicht

77 Grundsätzlich ist es natürlich Aufgabe der einzelnen Lernenden, sich zu motivieren. Wenn allerdings viele Lernende auf Sie angesichts des Leistungsnachweises einen unmotivierten Eindruck machen, empfiehlt es sich, mit den Studierenden das Gespräch zu suchen. Vielleicht können Sie dadurch „Störungen“ identifizieren, welche die Studierenden von einem größeren Engagement abhalten (z. B. zu viele Prüfungen in zu kurzer Zeit oder Brüche im Constructive Alignment, sei es auf der Ebene Ihrer Lehrveranstaltung oder jener des Curriculums).

haben. Die fachlichen Kriterien wiederum können unterschiedlich komplex gegliedert sein (Granularität) und in unterschiedlichen Formen angelegt werden, etwa als ordinale Beurteilungsraster oder als metrisierende Punktesysteme (siehe Kap. 9.3 und 9.4).

6.2 Bezugsnormen der Beurteilung

Um etwas beurteilen zu können, braucht es eine Normvorstellung, an der sich das Urteil ausrichtet:

Wir brauchen einen Maßstab, um einen Leistungsnachweis beurteilen zu können. Selbst wenn wir so weit kommen, dass wir sagen können, jemand habe einen Leistungsnachweis so und so gut bearbeitet und dies sogar mit Punkten quantifizieren können: Wir können erst eine Aussage über die Qualität der Leistung machen, wenn wir wissen, wie schwierig der Leistungsnachweis war. (Zimmermann, Tobias 2018, 57)

Dabei spielen auf jeden Fall die fachlichen Schwierigkeiten eine Rolle, aber auch der Lernstand der Studierenden und der Vergleich der Studierendenleistungen untereinander sind relevant. Deshalb kann sich der Beurteilungsmaßstab einer Beurteilung an drei verschiedenen Bezugsebenen orientieren, die man als Bezugsnormen bezeichnet: Kriteriale, individuelle und kollektive Bezugsnorm (vgl. Zimmermann, Tobias 2018, 57–59; Sacher 2014, 84–94; Ingenkamp u. Lissmann 2008, 63).

6.2.1 Kriteriale Bezugsnorm

Bei dieser Bezugsnorm werden Lernleistungen anhand von Kriterien gemessen, die auf den Lernzielen basieren. Bei der kriterialen Bezugsnorm steht also das Erreichen der Lernziele im Fokus, unabhängig vom ursprünglichen Wissensstand einzelner Studierender oder vom Leistungsstand, den die beurteilte Studierendengruppe insgesamt erreicht.

Die kriteriale Bezugsnorm ermöglicht eine transparente und gültige Beurteilung, sofern sie auf transparent kommunizierten Lernzielen beruht. Die Studierenden können somit die an sie gerichteten Erwartungen einschätzen und sich darüber mit Dozierenden und Peers verständigen.

Einschränkungen und Hinweise:

- Die Kriterien müssen genau definiert und transparent kommuniziert sein, um eine einheitliche und faire Bewertung sicherzustellen (siehe auch Kap. 9.3).

- Auch ein ausführlich ausgearbeitetes kriteriales Beurteilungsraster (siehe Kap. 9.3) kann nicht alle Aspekte der Leistungserbringung erfassen. Insbesondere bei Lernzielen hoher Taxonomiestufen und offenen Aufgabenstellungen kann es Aspekte geben, die schwer im Voraus zu definieren sind. Es ist wichtig, dies zu berücksichtigen und gegebenenfalls ergänzende Beurteilungsmethoden oder -instrumente einzusetzen, um der gezeigten Leistung gerecht zu werden.

6.2.2 Individuelle Bezugsnorm

Werden die Leistungen der Studierenden im Vergleich zu ihrem vorherigen Lernstand beurteilt, wird anhand der individuellen Bezugsnorm gemessen. Sie zielt also auf die persönliche Lernentwicklung der Studierenden.

Diese Sichtweise ermöglicht, den persönlichen Lernfortschritt von Studierenden zu berücksichtigen. Dabei können individuelle Unterschiede, Vorkenntnisse und Lernwege einfließen. Dies ist besonders wertvoll, wenn es darum geht, Studierenden Feedback zu ihrer persönlichen Lernentwicklung zu geben und ihnen Möglichkeiten für ihr weiteres Lernen aufzuzeigen. Durch die individuelle Bezugsnorm können auch persönliche Umstände und Herausforderungen der Studierenden berücksichtigt werden (etwa bei Studierenden mit körperlichen oder lernbezogenen Beeinträchtigungen wie Lese- und Rechtschreibschwäche oder ADHS).

Einschränkungen und Hinweise:

- Auch Beurteilungen und Feedbacks, die sich an der individuellen Bezugsnorm ausrichten, sollen sich an Lernzielen und transparenten Kriterien orientieren, um gültig und fair zu sein. Denkbar ist es, je nach Setting (v. a. bei relativ persönlichen Betreuungssituationen) mit Studierenden auch individuelle Lernziele zu vereinbaren.
- So wertvoll die Betonung individuellen Fortschritts ist, muss die Erreichung der Lernziele im Auge behalten werden. Wenn Studierende zwar individuelle Fortschritte machen, aber die Lernziele kriterial nicht erreichen, ist dies mittel- und langfristig dennoch problematisch; allerdings führen kontinuierliche individuelle Fortschritte in aller Regel just zur Erreichung der Lernziele. Insgesamt ist hier Augenmaß der Dozierenden und eine Balance zwischen Fördern und Fordern angebracht.

6.2.3 Kollektive Bezugsnorm

Hier werden die Leistungen der Studierenden untereinander verglichen. Das Leistungsniveau der beurteilten Gruppe bildet somit den Leistungsmaßstab, die Beurteilung basiert auf dem individuellen Rang einer Person innerhalb der gesamten Rangfolge der Gruppe. Die kollektive⁷⁸ Bezugsnorm ermöglicht es, die Leistungen von Studierenden im Kontext der Leistungen der Mitstudierenden zu betrachten – eine Leistung ist hier dann gut, wenn sie über dem Durchschnitt der Gruppe liegt, und dann schlecht, wenn sie darunter liegt.

Die kollektive Bezugsnorm ist für schulische und akademische Belange problematisch, weshalb die Einschränkungen und Hinweise hier ausführlicher ausfallen:

- Die kollektive Bezugsnorm macht eine zufällige Verteilung (die Zusammensetzung der jeweiligen Studierendengruppe) zur Grundlage der Bewertung. Das wäre nur zulässig, würde man die Leistungsverteilung in der betreffenden Gruppe bereits vorher kennen. Dazu wäre aber eine Voruntersuchung nötig, was weder praktisch machbar noch methodisch leistbar ist (vgl. Sacher 2014, 120).
- Auch bei der kollektiven Bezugsnorm „braucht es eine inhaltliche Norm, nur wird die Skala nicht durch die Abschätzung des Schwierigkeitsgrades geeicht, sondern alleine durch die Bildung einer Rangreihe und die willkürliche Festlegung, welcher Prozentsatz der Studierenden mit welchem Leistungsniveau etikettiert werden soll“ (Zimmermann, Tobias 2018, 58).⁷⁹ Die Problematik der Bewertung anhand der Normalverteilung und anderer willkürlich gesetzter Verteilungen wird in den Kapiteln 5.1.3 und 9.2.1 näher erläutert.
- Sacher spricht bezüglich der kollektiven Bezugsnorm von einem stillschweigenden Imperativ, „so zu sein und so zu handeln wie der Durchschnitt der Alters- und Zeitgenossen. Maßgeblich ist, was die andern tun, wie ‚man‘ handelt und ist“ (Sacher 2014, 87). Wenn unsere Studiengänge Fachexpertinnen und -experten hervorbringen sollen, die zu selbständigen

78 Diese Bezugsnorm wird oft auch soziale Bezugsnorm genannt, was aber irreführend ist, da sie sich eher antisozial auswirkt (vgl. Sacher 2014, 84; Vierlinger 1999, 80).

79 „Um ein konkretes Beispiel zu geben: Es ist wahrscheinlich, dass die Studierenden verschiedener Jahrgänge nicht immer genau gleich gut sind – wenn jedes Jahr aufgrund einer festgelegten Quote 60 Prozent bestehen und 40 Prozent nicht, bestehen Personen mit der lernzielbezogen betrachtet identischen Leistung im einen Jahr die Prüfung und im anderen nicht“ (Zimmermann, Tobias 2018, 58).

und kreativen Problemlösungen imstande sind, untergräbt die kollektive Bezugsnorm folglich unsere Zielsetzung.

- Die Anwendung der kollektiven Bezugsnorm kann sich auf Selbstwertgefühl und Motivation der Studierenden auswirken. Wenn sie sich ständig mit anderen vergleichen, kann dies zu Wettbewerb, Angst oder dem Gefühl der Unterlegenheit führen. Dies sind Auswirkungen, die psychologisch, gesellschaftlich und in der heutigen Arbeitswelt als problematisch erachtet werden.

6.2.4 Vergleich der drei Bezugsnormen: Fazit

Bei Leistungsbeurteilungen wird aufgrund der genannten Vor- und Nachteile in aller Regel der Fokus auf der kriterialen Bezugsnorm liegen – insbesondere bei summativen Leistungsnachweisen.⁸⁰ Auch bei formativen Leistungsnachweisen spielt sie eine wesentliche Rolle als fachbezogener Beurteilungsmaßstab und Entwicklungshorizont (zur Unterscheidung zwischen formativen und summativen Beurteilungen siehe Kap. 3.1 und 3.2).

Die individuelle Bezugsnorm soll vor allem bei formativen Beurteilungen zum Einsatz kommen, da bei der Rückmeldung zum aktuellen Leistungsstand und zu Möglichkeiten für das weitere Lernen die individuelle Entwicklung der Studierenden eine relevante Bezugsebene ist (siehe Kap. 8). Die kollektive Bezugsnorm ist hingegen wann immer möglich zu vermeiden (siehe dazu auch Kap. 9.2.1).⁸¹

80 Dieser pädagogischen Einschätzung folgen übrigens auch die schulrechtlichen Vorgaben in Deutschland: „Die jeweils einschlägigen Vorschriften aller Bundesländer statuieren ein Primat der kriteriumsorientierten Bezugsnorm bei der Zensurenvergabe. Dies geschieht zum Teil – wie etwa in Bayern oder Sachsen – in Form eines generellen Normbefehls, der keine Ausnahmen zulässt; in diesem Fall muss die Notenfindung kriteriumsorientiert erfolgen. In anderen Bundesländern werden hiervon mehr oder weniger weitreichende Ausnahmen gemacht, auch wenn die Berücksichtigung der kriteriumsorientierten Bezugsnorm der Grundsatz bleibt“ (Kostorz 2016, 288).

81 Eine verständliche Ausnahme stellt der folgende Fall dar: „In gewissen Hochschulen und Studiengängen stehen für eine bestimmte Studienstufe nur x Studienplätze zur Verfügung. In diesen Fällen wird man pragmatischerweise nicht mehr als x Studierende den Übertritt in diese Stufe bestehen lassen (unter Anwendung der Lernzielorientierung ist es aber denkbar, dass weniger als x Studierende zu dieser Stufe zugelassen werden, falls weniger als x Studierende das gewünschte Leistungsniveau erreichen). Allerdings sollte man sich bewusst sein, dass eine solche Kontingentierung aus der Sicht der Leistungsdiagnostik problematisch ist“ (Zimmermann, Tobias 2018, 58, Fußnote 2).

6.3 Gütekriterien der Leistungsbeurteilung

Wenn man Leistungen beurteilen möchte, dann müssen diese Beurteilungen bestimmten Qualitätsansprüchen genügen. Dabei besteht ein Spannungsverhältnis zwischen quantitativ und qualitativ orientierten Ansätzen der Beurteilung. Die Qualitätskriterien der beiden Ansätze werden nachfolgend dargestellt, um anschließend praktische Folgerungen für den Umgang mit Qualitätsansprüchen in der Leistungsbeurteilung an Hochschulen zu diskutieren.

6.3.1 *Quantitative Gütekriterien der Testtheorie*

Die Testtheorie ist ein quantitativ orientierter Ansatz in der psychologischen Forschung und bildet einen wichtigen Teilbereich der Psychometrie. Sie beschäftigt sich mit der Entwicklung und Bewertung von psychologischen Tests und zielt darauf, die wissenschaftliche Aussagekraft von Tests sicherzustellen. Ihre breit etablierten Hauptgütekriterien lauten Objektivität, Reliabilität und Validität (vgl. Pospeschill 2022, 16f.). Quantitativ orientierte Leistungsbeurteilung übernimmt diese drei Hauptgütekriterien in der Regel unverändert (so z. B. Sacher 2014, 35–41; Ingenkamp u. Lissmann 2008, 51–60; Metzger u. Nüesch 2004, 6–9).

Objektivität

Das testtheoretische Kriterium der Objektivität bezieht sich im Kontext der Leistungsbeurteilung auf das Ausmaß, in dem die Beurteilung unabhängig von der Person ist, welche die Leistung beurteilt. Etwas spezifischer formuliert bedeutet Objektivität, dass die Leistungsbeurteilungen nicht durch persönliche Vorurteile, Meinungen oder subjektive Einschätzungen der Prüfenden beeinflusst werden sollten.

Die testtheoretische Methode zur Überprüfung der Objektivität besteht darin, verschiedene Personen die gleichen Leistungsnachweise beurteilen zu lassen – je mehr sie übereinstimmen, desto objektiver ist die Leistungsbeurteilung. Dieses Kriterium wird allerdings in der qualitativen Forschung aus erkenntnistheoretischen Gründen abgelehnt, wie im folgenden Unterkapitel dargelegt wird. Die Übereinstimmung von Beurteilenden lässt sich auch quantifizieren und mit statistischen Prüfverfahren überprüfen („Beurteilerreliabilität“, vgl. Wirtz u. Caspar 2002 oder die kurze Darstellung in Zimmermann 2014, 109–111) – was zeigt, dass sie keinesfalls ein absolutes Kriterium darstellt. Auch die quantitative Forschung muss somit eingestehen, dass in einem gewissen – möglichst zu minimierenden – Ausmaß kognitive Verzerrungen bei jeder Beurteilung eine Rolle spielen, siehe Kapitel 6.4.

Reliabilität

Das Kriterium der Reliabilität (Zuverlässigkeit) bezieht sich auf das Ausmaß, in dem die Erhebung einer Leistung konsistent, stabil und frei von Zufällen oder Messfehlern ist: „Je genauer und sicherer eine Messung ein bestimmtes Merkmal misst, desto höher ist ihre Reliabilität“ (Zimmermann, Tobias 2018, 59). Die Reliabilität gibt somit Aufschluss darüber, wie genau die von Studierenden anhand eines Leistungsnachweises gezeigten Leistungen ihr tatsächliches Wissen und Können (ihre Kompetenz als latentes Konstrukt) widerspiegeln.

Wie die Objektivität ist auch die Reliabilität kein absolutes Kriterium, denn eine Messung kann nicht zuverlässiger sein als die Stabilität des gemessenen Merkmals (vgl. Ingenkamp u. Lissmann 2008, 55). So können studentische Leistungen Schwankungen unterliegen, etwa bedingt durch die Tageszeit oder aktuelles Stressempfinden. Aber auch die Art der Messung kann zu Leistungsschwankungen führen: „So kann sich eine Studentin oder ein Student vielleicht mündlich sehr gewandt ausdrücken und eine Frage problemlos beantworten, aber Schwierigkeiten mit dem Schreiben haben. Deshalb kann sie oder er die gleiche Frage schriftlich deutlich weniger angemessen beantworten als mündlich“ (Zimmermann, Tobias 2018, 60). Daher ist zu bedenken, dass Leistungsnachweise ein sozial konstruiertes Setting darstellen, unter dessen spezifischen Bedingungen die Studierenden eine Handlung vollführen, die als Leistung interpretiert wird (Kap. 2.5.2). Ihre Konstruktion beeinflusst wesentlich, welche Art von Leistungen die Studierenden in welchem Ausmaß zeigen können.

Die Reliabilität besagt somit in erster Linie, in welchem Maße sich Messfehler auf die Leistungsbeurteilung auswirken. Sie kann ebenfalls durch verschiedene Verfahren quantifiziert werden, wodurch ihre Güte statistisch überprüfbar wird (vgl. Lienert u. Raatz 1998, 9f.). Ein Problem dabei ist, dass diese statistischen Verfahren nur funktionieren, wenn die studentischen Leistungen in ihrer Qualität stark variieren, während das Ziel jeglicher Lehrbemühungen eigentlich darin liegt, möglichst viele Studierende zu möglichst großem Lerngewinn anzuregen, also diese Varianz zu reduzieren (siehe dazu Kap. 9.2).

Validität (Gültigkeit)

Angemessene Objektivität und Reliabilität sind notwendige, aber keineswegs hinreichende Bedingungen für eine annehmbare Güte der Leistungsbeurteilung. Denn das wichtigste testtheoretische Gütekriterium ist die Validität. Sie „sagt aus, ob tatsächlich das gemessen wird, was man messen will, und nicht irgendetwas anderes“ (Ingenkamp u. Lissmann 2008, 57).

So zentral dieses Kriterium ist – es ist ungleich schwieriger zu erfassen und beurteilen als die beiden vorher genannten. In der Fachliteratur werden zudem verschiedene und sich teilweise überlagernde Aspekte von Validität postuliert. Für unseren Zusammenhang besonders relevant sind die Inhalts- und die Konstruktvalidität (vgl. Sacher 2014, 37–41; Ingenkamp u. Lissmann 2008, 57–59; Lienert u. Raatz 1998, 7–11):

- Die Inhaltsvalidität einer Leistungsbeurteilung bezieht sich auf das Ausmaß, in dem diese den zu beurteilenden Inhaltsbereich angemessen und repräsentativ abdeckt. Um diesem Kriterium zu genügen, muss ein Leistungsnachweis sicherstellen, dass er das relevante Wissen und Können des betreffenden Themas abdeckt – er muss also die gesetzten Lernziele überprüfen (siehe Kap. 5.2 bis 5.4). Zu diesem Zweck sollten die Aufgabestellungen oder Fragen (Items) darauf überprüft werden, ob sie tatsächlich die zu beurteilenden Wissensinhalte und Fähigkeiten adressieren. Ein hohes Maß an Inhaltsvalidität ermöglicht ein umfassendes und genaues Urteil über das Wissen und Können der Studierenden im zu beurteilenden Themengebiet. Dies ist weit weniger trivial als es klingt – insbesondere neigen wir intuitiv dazu, eher simples Faktenwissen denn komplexe Zusammenhänge und einfache Fertigkeiten statt integrierter Fähigkeiten zu überprüfen. Zudem kann Inhaltsvalidität nur durch Fachexpertinnen überprüft werden, so dass Dozierende diesbezüglich kaum Unterstützung bei Prüfungsfachstellen oder hochschuldidaktischen Zentren einholen können.
- Die Konstruktvalidität verweist darauf, dass die zu überprüfende Leistung bzw. das dahinter liegende Wissen und Können ein latentes Konstrukt darstellt. Eine Leistungsmessung ist desto eher konstruktvalide, je mehr sie tatsächlich Rückschlüsse auf das zu ermittelnde Wissen und Können erlaubt. Es geht darum, den Nachweis zu erbringen, dass die Bewertung mit dem zugrunde liegenden theoretischen Konstrukt übereinstimmt und sinnvolle Beziehungen zu anderen relevanten Variablen oder Maßnahmen aufweist. Die Konstruktvalidität erfordert einen klaren theoretischen Rahmen und eine konzeptionelle Definition des zu bewertenden Konstrukts. Dazu gehört das Sammeln von Belegen aus verschiedenen Quellen.

Auch Validität kann quantifiziert und durch den Einsatz statistischer Methoden überprüft werden, etwa indem die Beziehungen zwischen den Beurteilungsergebnissen und externen Beurteilungskriterien oder die interne Struktur der Beurteilung analysiert werden (vgl. Lienert u. Raatz 1998, 10f.).

6.3.2 Gütekriterien qualitativer Forschung

Die oben vorgestellten testtheoretischen Gütekriterien entstammen der quantitativen sozialwissenschaftlichen, insbesondere psychologischen Forschung. Einige Bedingungen, die auf dieses Forschungsparadigma zutreffen, sind bei Leistungsnachweisen im hochschulischen Kontext in der Regel nicht gegeben. Beispielsweise können wir aus rein praktischen Gründen gewöhnlich nicht den gleichen verfahrenstechnischen Aufwand leisten wie bei der Erhebung und Auswertung von sozialwissenschaftlichen Daten mit vielen Studienteilnehmenden (siehe unten das Kriterium der Ökonomie). Und auf der methodischen Ebene sind viele statistische Verfahren nur bei höheren Studierendenzahlen (hundert und mehr Personen pro Leistungsnachweis) sinnvoll anwendbar.⁸²

Über solche eher praktischen Einwände hinaus wird in der Hochschuldidaktik die Orientierung an quantitativen Messmodellen aber auch grundsätzlich hinterfragt. So argumentieren Biggs u. Tang (vgl. 2007, 192f.), dass einige der in der differentiellen Psychologie sinnvollen Grundannahmen im Bildungskontext problematische Auswirkungen haben:

- Aufgrund der Orientierung an differentialpsychologischen Messmodellen führen sie zur Fehlvorstellung, Lernleistungen müssten normalverteilt sein (diese wird in Kap. 5.1.3 und 9.2.1 widerlegt).
- Sie verführen Lehrende zu übertriebenen Genauigkeits- und Objektivitätsvorstellungen und veranlassen sie, sich übermäßig an scheinengenauen Zahlen zu orientieren, statt inhaltlich begründete Urteile anzustreben (vgl. auch Moss 1994). Dies stellt den sinnvollen Zusammenhang zwischen Messungen und Urteilen auf den Kopf, wie bereits in Kapitel 2.4 gezeigt wurde: „Es geht nicht um die Frage von Metrik versus Urteilsvermögen, sondern um Messwerte als eine Informationsquelle für das Fällen von Urteilen, wozu auch gehört, (...) dass man ihre charakteristischen Verzerrungen erkennt und auch das zu schätzen weiß, was nicht gemessen werden kann“ (Muller 2018, 183, Übers. tz). Diese Orientierung an einer Scheinobjektivität geht in der Regel zulasten der Gültigkeit der Beurteilung (vgl. Sacher 2014, 46f.).
- Sie begünstigen eine übermäßige Trennung von Lehr-Lern-Aktivitäten bzw. Unterricht und Beurteilungen des Lernens, was tendenziell extrinsischer Lernmotivation (siehe Kap. 4.5) und oberflächlichen Lernstrategien (siehe Kap. 4.2) Vorschub leistet.

82 In der probabilistischen Testtheorie (item response theory) werden akzeptable Anpassungsgüten für Modelle sogar oft erst ab einer Teilnehmendenzahl von $n > 500$ –1000 erreicht (vgl. Brown u. Abdulnabi 2017; Embretson u. Reise 2015).

Deshalb ist es sinnvoll, für die Beurteilung von Lernleistungen auch die nachfolgend erläuterten Qualitätskriterien aus dem Paradigma der qualitativen Forschung in Betracht zu ziehen. In Kapitel 6.5 wird schließlich diskutiert, in welchen Fällen welche Kriterien eher angemessen sind.

Intersubjektivität

In der qualitativen Forschung und generell in konstruktivistischen Ansätzen wird der Glaube an die Fähigkeit, die Dinge so zu erkennen, wie sie wirklich sind, als Objektivismus abgelehnt. Vielmehr wird davon ausgegangen, dass alle Wahrnehmungen und Darstellungen der Welt relativ zu begrifflichen Konzepten und bestimmten Weltanschauungen sind. Etwas weniger philosophisch, aber nicht minder grundsätzlich ist die Kritik an der Vorstellung, Forschende könnten quasi-objektiv das soziale Leben unbelastet von Vorurteilen und persönlichen Eigenschaften betrachten (vgl. Schwandt 2007, 210f.). Was Handelnde – in unserem Kontext sowohl Studierende als auch Lehrende – als fachliche, soziale, emotionale Tatsachen, also als Realität wahrnehmen, bringen sie somit letztlich in ihren Wahrnehmungen und Handlungen selbst erst hervor (vgl. Bergmann 2000, 122).

Der Begriff der Intersubjektivität (vgl. Eberle 1984, 14–38) bezieht sich deshalb auf den Prozess oder das Ziel, innerhalb einer spezialisierten disziplinären Gemeinschaft durch Dialog, Debatte und begründete Argumente einen Konsens zu erreichen.⁸³ Er gemahnt zu etwas mehr Bescheidenheit bei der Beurteilung von Leistungen anderer Personen als jener der „Objektivität“, da er von einer empirisch weniger gesicherten Beurteilungsgrundlage ausgeht, die diskursiv zu verhandeln ist.

Authentizität

Um die Qualität von Aushandlungen, Resultaten und Produkten einzuschätzen werden in der qualitativen Forschung oft die von Guba u. Lincoln (vgl. 1989, 245–250) entwickelten Authentizitätskriterien angeführt. Diese umfassen folgende Aspekte (vgl. Schwandt 2007, 14f.):

Der Untersuchungsprozess soll unparteiisch, gerecht und ausgewogen sein, indem er die unterschiedlichen Perspektiven/Sinnkonstruktionen der Beteiligten berücksichtigt. Anschließend an das Kriterium der Intersubjektivität beschreibt Authentizität die Qualitätsansprüche an das intersubjektive

83 Mit der Betonung der situativen Aushandlung von Bedeutungen und Interpretationen wird auch das testtheoretische Kriterium der Reliabilität hinfällig: Bei einer erneuten Messung wäre die Situation nicht mehr identisch und die Bedeutungen müssten neu ausgehandelt werden (vgl. Schwandt 2007, 262f.; Moss 1994).

Aushandeln von Untersuchungsergebnissen: Die Untersuchung soll die Welt- und Erlebniskonstruktionen der Beteiligten fassbar machen, anerkennen und dadurch kritische Reflexion, Dialog und Veränderungsprozesse anregen. Sie soll ein breiteres Selbst- und Fremdverständnis der Beteiligten und in der Gesellschaft ermöglichen, dadurch Lernprozesse, Einsicht und letztlich die Handlungsfähigkeit verschiedener Interessensgruppen fördern.

Der Anspruch der Authentizität verdeutlicht den konstruktiven Charakter von Leistungsnachweisen, wie er in Kapitel 2.5.2 herausgestellt wurde: Die zu beurteilenden Leistungen werden durch das Beurteilungs-Setting erheblich beeinflusst. Authentizität verlangt deshalb eine transparente Beschreibung und multiperspektivische Orientierung des Beurteilungsprozesses und seiner Grundlagen. Zudem macht das Kriterium deutlich, dass Leistungsbeurteilungen für alle Beteiligten Lerngelegenheiten sind: Nicht nur für Studierende, sondern auch für Dozierende und weitere Hochschulangehörige.

Validität

Die testtheoretische Konzeption der Gültigkeit fordert, gemessene Daten sollten direkt mit einer beobachterunabhängigen Realität übereinstimmen. Lehnt man eine solche objektivistische Sichtweise ab, kann Gültigkeit stets nur relativ zu den aktuellen Standards einer bestimmten Gemeinschaft bestimmt werden. Der Schwerpunkt verlagert sich somit von der Wahrheitsfindung auf die Untersuchung der Art und Weise, wie Erklärungen in Diskursen soziokulturell entwickelt und beeinflusst werden (vgl. Schwandt 2007, 309–311).

Mit Blick auf Leistungsnachweise und Leistungsbeurteilung sind aus qualitativer Perspektive also solche Verfahren gültig, die durchgängig transparent und entsprechend kritisierbar machen, wie die Erfassung und Beurteilung der fraglichen Lernleistung mit Bezug auf die Lernziele zustande kommen. Die Validität eines Urteils, also dass es das beurteilt, was beurteilt werden soll, ist somit eine Folge von Authentizität.

Generalisierbarkeit

Dieses Kriterium zielt auf die Übertragbarkeit oder Anwendbarkeit von Untersuchungsergebnissen auf andere Kontexte oder Umgebungen. Qualitative Analysen sollen reichhaltige, kontextspezifische Beschreibungen und Erkenntnisse liefern, die für ähnliche Situationen genutzt werden können. Dieser Transfer liegt aber in der interpretativen Verantwortung der Personen, welche qualitative Studienresultate nutzen. Eine solche Übertragung erleichtern und intersubjektiv absichern soll die detaillierte Dokumentation des Untersuchungsprozesses, der Merkmale der Teilnehmenden und des Umfelds sowie der Interpretationen der Forschenden – die Kenntnis der Situativität

gefundener Resultate soll ihre Generalisierung auf andere Kontexte ermöglichen (vgl. Schwandt 2007, 126–128).

Mit Bezug auf Leistungsnachweise und ihre Beurteilung legt auch dieses Kriterium den Fokus auf eine transparente Beschreibung des Leistungsnachweis-Settings, der Beurteilungskriterien und der Leistungsbeurteilung. Dieses Kriterium ist insbesondere für die Rechenschaftslegung von Leistungsbeurteilungen relevant: So kann gezeigt werden, dass sie reglementarischen Anforderungen genügen und die Beurteilungen fachlich und prozessbezogen angemessen sind. Dies ist auch im Fall von Widersprüchen und Klagen gegen Beurteilungen und Benotungen relevant (vgl. Kostorz 2016).

6.3.3 *Nebengütekriterien*

In der Fachliteratur werden oft auch verschiedene so genannte Nebengütekriterien genannt. Für unseren Kontext relevant scheinen die Folgenden (vgl. Sacher 2014, 41; Metzger u. Nüesch 2004, 6–10):

- **Ökonomie:** Der Aufwand für die Planung, Durchführung und Auswertung von Leistungsbeurteilungen soll so gering wie möglich sein, um hinreichend gültige Aussagen zu erhalten: „Eine Prüfung ist umso ökonomischer, je weniger Konstruktionsaufwand sie erfordert (Konstruktionsökonomie), je kürzer die Prüfungszeit ist, je mehr Prüflinge gleichzeitig geprüft werden (Durchführungsökonomie) und je schneller und bequemer die Prüfungsergebnisse ermittelt und bewertet werden können (Auswertungsökonomie)“ (Metzger u. Nüesch 2004, 9).
- **Fairness:** Ein Leistungsnachweis soll von allen Studierenden gleichermaßen bearbeitet werden können, sofern sie auf die transparent kommunizierten Lernziele hingearbeitet und an den vorgesehenen Lehr-Lern-Aktivitäten teilgenommen haben. Werden beispielsweise in einer Selbstlerneinheit zwei Texte als alternative Lektüren angegeben, würde es der Fairness widersprechen, wenn einer der beiden Texte wesentlich besser zur Bearbeitung des Leistungsnachweises befähigen würde.
- **Chancengerechtigkeit:** Dieses Kriterium zielt auf grundsätzlichere Aspekte der Fairness⁸⁴, namentlich auf die „externen Lernbedingungen der Prüflinge (z. B. sozioökonomischer Status, Kulturzugehörigkeit, fachunabhängige Vorkenntnisse, Geschlecht)“ (Metzger u. Nüesch 2004, 10). So wäre z. B. eine Aufgabenstellung, die stark von implizitem Vorwissen über nationale Gegebenheiten ausgeht, nicht chancengerecht gegenüber ausländischen Studierenden. Insgesamt dürfte die implizite Bevorzugung

84 Chancengerechtigkeit und Fairness werden oft als ein gemeinsames Kriterium genannt, so auch von Metzger u. Nüesch (2004, 6–10).

von Studierenden bildungsnaher Herkunft die Chancengerechtigkeit von Leistungsbeurteilungen am meisten beeinträchtigen (siehe Kap. 2.2 zum Einfluss des sozioökonomischen Status auf den Bildungserfolg).

6.3.4 Zielkonflikte

Zwischen verschiedenen Gütekriterien kann es zu Zielkonflikten kommen, das heißt, die Erfüllung des einen Kriteriums steht im Widerspruch zur Erfüllung des anderen. Der schwerwiegendste Zielkonflikt besteht bei den quantitativen Gütekriterien zwischen Objektivität und Validität. So

kann halbwegs brauchbare Objektivität (...) nur in wenigen, eng begrenzten und rigide geregelten Situationen, am ehesten noch in schriftlichen Prüfungen hergestellt werden. (...) Qualitäten des Lernprozesses und Leistungsdaten aus Beobachtungen fallen größtenteils unter den Tisch. Das Bemühen um die Sicherung von Objektivität geht somit ersichtlich auf Kosten der Validität: Das, was einigermaßen objektiv erhoben und beurteilt werden kann, repräsentiert die Schülerleistung nämlich nur noch sehr ausschnitthaft und unvollständig (Sacher 2014, 47).

Ein weiterer typischer Zielkonflikt besteht zwischen der Ökonomie auf der einen und dem Anspruch auf Validität/Authentizität auf der anderen Seite: Je umfassender und transparenter ein Leistungsnachweis und seine Beurteilung das Erreichen von Lernzielen dokumentieren, desto größer wird der Aufwand für die Planung, Durchführung und Beurteilung des Leistungsnachweises. In der Regel müssen deshalb Kompromisse zwischen diesen beiden Kriterien gefunden werden (vgl. Metzger u. Nüesch 2004, 10).

Dieser und weitere mögliche Zielkonflikte zeigen auf, dass die enorm hohen Qualitätsansprüche, die insbesondere an summative Leistungsbeurteilungen von verschiedenen Seiten – angesichts ihrer großen individuellen und gesellschaftlichen Bedeutung verständlicherweise – gestellt werden, in keinem Fall vollständig erfüllt werden können. In der Praxis müssen verschiedene Qualitätsansprüche gegeneinander abgewogen werden und pragmatische Kompromisse gefunden werden. Hilfreich ist dabei die Perspektive auf das Curriculum: Indem unterschiedliche Formen von Leistungsnachweisen im Verlauf des Studiums zum Tragen kommen, können unterschiedliche Stärken dieser Verfahren genutzt und ihre Schwächen teilweise kompensiert werden (siehe Kap. 10 zur Curriculumentwicklung).

6.4 Kognitive Verzerrungen und Beurteilungsfehler

Um schnelle Einschätzungen und Entscheidungen zu treffen, benutzen wir im Alltag meist kognitive Heuristiken. Dabei handelt es sich um eine Art mentale Abkürzungen oder Hilfsstrategien, die uns helfen, in komplexen Situationen schnell Orientierung zu finden. Diese Mechanismen sind in alltäglichen Situationen meist nützlich – ja wir könnten unseren Alltag ohne sie kaum bewältigen.⁸⁵ In einigen Fällen können solche Heuristiken jedoch zu falschen Einschätzungen und schlechten Entscheidungen führen, wie in Kapitel 4.7 aufgezeigt wurde. Man spricht in diesem Fall von kognitiven Verzerrungen (Englisch: cognitive bias). Dabei kommt es oft zu einer Verwischung von Wahrnehmung und Interpretation.

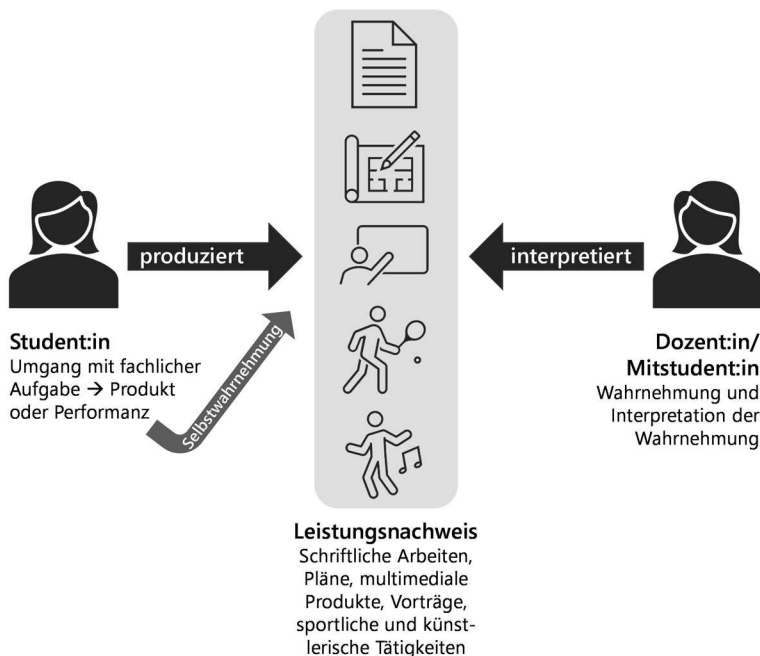


Abb. 16: Mittelbarkeit der Beurteilung studentischer Leistungen. Quelle: Eigene Abbildung, in Anlehnung an Zimmermann, Tobias 2018, 60

85 So definieren Gigerenzer u. Gaissmaier (2006, 333): „Eine Heuristik ist eine Regel, die den Prozess – nicht nur das Ergebnis – einer Problemlösung beschreibt. Sie ist einfach, weil sie auf evolvierte und erlernte Fähigkeiten zugreifen kann, und sie ist intelligent, weil sie Umweltstrukturen nutzen kann.“

Neben den eher systematischen Verzerrungen, die durch überdauernde kognitive Heuristiken seitens der Interpretierenden entstehen, gibt es auch noch stärker situativ bedingte Verzerrungen. Diese hängen etwa mit dem Wetter oder der Müdigkeit der Beurteilenden zusammen. Der Effekt ist allerdings der Gleiche: Die Beurteilungen verändern sich durch Faktoren, die unabhängig sind von der zu beurteilenden Leistung. Nachfolgend werden die wichtigsten derzeit bekannten kognitiven Verzerrungen (Kap. 6.4.1) und situativen Verzerrungen (6.4.2) aufgeführt.

Wenn wir Leistungen anderer Menschen beurteilen möchten, sollten wir uns dieser Verzerrungen bewusst sein und ihre wichtigsten Arten und Ursachen kennen. Dies ermöglicht auch, ihr Auftreten oder ihre unerwünschten Auswirkungen zu reduzieren – ganz ausgeschlossen werden können sie nie, da auf Leistungen immer nur mittelbar geschlossen werden kann, wie Abbildung 16 veranschaulicht.⁸⁶ Die Verfahren zur Reduktion von kognitiven und situativ bedingten Verzerrungen sind dabei grundsätzlich die gleichen, die auch zur Optimierung der qualitativen Gütekriterien wie Authentizität, Validität oder Generalisierbarkeit beitragen (siehe Kap. 6.5).

6.4.1 Kognitive Verzerrungen bei der Leistungsbeurteilung

Die folgende Aufzählung basiert auf verschiedenen Quellen und ist alphabetisch geordnet (vgl. Zimmermann, Tobias 2018, 61–63; Sacher 2014, 47–52; Greve u. Wentura 1997, 56–74):

- **Bestätigungstendenz:** In diesem Fall suchen, wählen und interpretieren wir Informationen so, dass die eigenen bereits bestehenden Erwartungen bestätigt werden, während widersprüchliche Belege ignoriert oder relativiert werden. Bei Leistungsbeurteilungen kann dies dazu führen, dass übermäßig solche Anteile des studentischen Verhaltens oder Produkts beachtet werden, die die bereits existierende Einschätzung über die Leistungsfähigkeit der betroffenen Person bestätigen.

86 Aus hermeneutischer Sicht wird sogar argumentiert, unser Verständnis von uns selbst und der Welt hänge von solchen Heuristiken und damit verbundenen Vorurteilen ab. Nur so könnten wir unsere Wahrnehmungen strukturieren und interpretieren. Das Nachdenken über kognitive Mechanismen und Vorurteile muss deshalb Bestandteil eines begründeten, tieferen Verständnisses sein (vgl. Schwandt 2007, 20f.). Auch deshalb sind soziale Verständigungsmechanismen, wie sie in Kap. 4.7 diskutiert wurden, für die Absicherung von Argumenten und Urteilen so wichtig und gehören zu den wichtigsten Methoden, um Beurteilungen authentischer, gültiger und besser verallgemeinerbar zu machen, siehe Kap. 6.5.

- Erwartungseffekt der Beobachtenden: Hier beeinflusst die Erwartung einer beobachtenden Person ihre Einschätzung der beobachteten Leistung, z. B. wenn eine Dozentin von einem Studenten eine schlechte Leistung erwartet, wird sie diese schlechter beurteilen als sachlich gerechtfertigt. Über subtile Signale wie Gestik, Mimik oder die Art der Begrüßung können solche Erwartungseffekte auch das Verhalten der beobachteten Person beeinflussen.
- Erwartungseffekt der Beobachteten: Dies ist die Ergänzung zum vorangehenden Effekt. Hier versucht sich die beobachtete Person in der Tendenz so zu benehmen, wie es ihrer Vermutung nach die beobachtende Person von ihr erwartet. Dieser Effekt ist in Situationen der Leistungsbeurteilung besonders ausgeprägt, da Studierende die Erwartungen der Beurteilenden möglichst erfüllen möchten. „Gerade in mündlichen Prüfungssituationen kann dieser Effekt die Beurteilung beeinflussen (wenn Studierende das Vokabular benutzen, das die oder der Dozierende mag, oder sich so kleiden, wie es die oder der Dozierende als seriös empfindet etc.)“ (Zimmermann, Tobias 2018, 61f.).
- Geschlechtsspezifische Vorurteile: Annahmen über bestimmte mit einem Geschlecht verbundene Eigenschaften können die Beurteilung einer Leistung auf sachlich nicht gerechtfertigte Weise beeinflussen. Solche Annahmen sind oft unbewusst, was es aber noch schwieriger macht, ihren Einfluss zu begrenzen. Mögliche Beispiele: „Frauen können besser kommunizieren als Männer“ oder „Männer können besser mit Zahlen umgehen“.
- Halo-Effekt (Heiligenschein-Effekt): Aufgrund bestimmter Merkmale wird in wertender Weise auf andere Eigenschaften geschlossen, z. B. „wer oft abwesend ist, ist faul oder unmotiviert“.
- Kontrasteffekt: Wenn sich zwei aufeinanderfolgende Reize unterscheiden, kann der Unterschied deutlicher eingeschätzt werden, als es sachlich gerechtfertigt ist. Bei Leistungsbeurteilungen ist dieser Effekt bedeutsam, weil die Leistungen meistens in einer zeitlichen Abfolge beurteilt werden und diese durch den Kontrasteffekt Wahrnehmung und Beurteilung der Beurteilenden beeinflusst. Dies spricht zum Beispiel dafür, bei wiederholter Beurteilung der gleichen Gruppe die Abfolge jeweils zu ändern, und bei schriftlichen Leistungsnachweisen verschiedene Aufgaben nicht nur pro Aufgabe, sondern auch in jeweils anderer Personenreihenfolge zu beurteilen.
- Milde-/Strenge-Effekt: Leistungen werden systematisch milder/strenger beurteilt, als es sachlich angemessen wäre.

- **Primacy-Recency-Effekt:** Menschen können sich gewöhnlich aus einer Reihe von Eindrücken den ersten und letzten besonders gut merken (ein Prinzip, das in der Rhetorik seit der Antike bekannt ist). Dadurch haben Eindrücke vom Anfang und Ende einer Beobachtung oft einen größeren Einfluss auf die Gesamtbeurteilung. So kann bei einer mündlichen Prüfung ein sachlich gesehen besonders guter/schlechter erster oder letzter Teil die Gesamtbeurteilung in sachlich ungerechtfertigtem Maß prägen.
- **Projektionsfehler:** Ein solcher liegt vor, wenn Beobachtende ihre eigenen Gedanken, Überzeugungen, Werte oder Verhaltensweisen auf die beobachteten Personen projizieren, auch wenn deren beobachtbares Verhalten dafür keine oder gar gegenteilige Belege liefert. In Beurteilungssituationen ist es insbesondere heikel, dass Dozierende Aussagen und Lösungsverhalten von Studierenden vom Stand des Expertenwissens aus beurteilen müssen. Was sie als Expertinnen stillschweigend voraussetzen, konnten Studierende möglicherweise noch gar nicht lernen.
- **Referenzfehler:** Es wird eine unpassende Bezugsnorm oder ein ungeeigneter Maßstab gewählt, z. B. die Leistungsangabe innerhalb einer Gruppe von Studierenden statt eines lernzielorientierten Kriteriums. Dies ist einer der Hauptgründe für den in Kapitel 2.3 berichteten Befund, dass Noten den fachlichen Wissens- und Könnensstand von Lernenden nur sehr unzuverlässig wiedergeben (gleiche Noten lassen nicht auf gleiche Fähigkeiten schließen).
- **Selbsterfüllende Prophezeiung:** Eine solche liegt vor, wenn (oft unbewusste) Erwartungen oder Überzeugungen einer Person ihr Verhalten in einer Weise beeinflussen, die das Eintreten der Erwartungen begünstigt. Dies ist quasi die Verlängerung der oben geschilderten Erwartungseffekte der Beobachtenden und der Beobachteten auf die Ebene des tatsächlichen Verhaltens. Bei Situationen der Leistungsbeurteilung wirken sich zum Beispiel Erwartungen der Studierenden bezüglich ihrer zu erbringenden Leistung oft in die Richtung der Erwartung aus (Studierende, die von sich eine gute Leistung erwarten, erbringen in der Tendenz intersubjektiv betrachtet in der Tat bessere Leistungen).
- **Tendenz zu extremen Urteilen:** Diese Neigung zeigt sich darin, dass vor allem die Extreme einer Skala verwendet werden. Ein Beispiel sind Dozierende, die Leistungen radikal in gute und schlechte einteilen.
- **Tendenz zur Mitte:** Extreme Urteilen werden vermieden. Ein Beispiel sind Dozierende, welche die Maximal- und Minimalbewertung nie oder kaum vergeben.

Verschiedene Arten kognitiver Verzerrung können sich auch überlagern und gegenseitig verstärken: Beispielsweise kann in einer Serie von mündlichen

Prüfungen auf eine sachlich betrachtet schwache Leistung eine sehr gute folgen (Kontrast-Effekt), bei der die Studentin zudem eine äußerst geistreiche Antwort auf die Einstiegsfrage gibt, die sowohl einen Primacy-Effekt, einen Halo-Effekt als auch Erwartungseffekte der Beobachter wie der Beobachteten zur Folge haben kann. Die vorangegangene Leistung sowie ein kurzer Ausschnitt der eigentlich als Ganzes zu beurteilenden Leistung können somit die Beurteilung massiv beeinflussen.

6.4.2 *Situativ bedingte Verzerrungen bei der Leistungsbeurteilung*

Die oben genannten Verzerrungstendenzen beruhen auf Tendenzen, die entweder typisch für die menschliche Wahrnehmung im Allgemeinen sind (z. B. Primacy-Recency-Effekt) oder aber zumindest die Tendenz einer bestimmten Person beschreiben. Die folgenden Verzerrungen sind demgegenüber stärker situativ bedingt und können auch innerhalb des Verhaltens der gleichen Person variieren.⁸⁷ Sie sind folglich auch weniger gut systematisierbar und die folgende Aufzählung stellt eher eine geordnete Beispielsammlung dar denn eine Systematisierung im engeren Sinne.

Äußere Bedingungen der Situation

Äußere Situationsbedingungen können sich in überraschendem Ausmaß und teilweise in überraschender Richtung auf Beurteilungen auswirken:

- Wetter: Eine Untersuchung an der University of Toronto zeigte, dass Bewerbende für das Medizinstudium, deren Zulassungsgespräch an einem regnerischen Tag erfolgte, schlechter bewertet wurden als Studierende, deren Gespräch an einem sonnigen Tag erfolgte. Die Schwankungen waren nicht enorm groß, aber statistisch klar messbar und unabhängig von den Beurteilenden. Sie führten dazu, dass die Aufnahme bzw. Ablehnung von fast 20 % der Bewerbenden durch das Wetter beeinflusst wurde (vgl. Redelmeier u. Baxter 2009).
- Auch die Außentemperatur kann einen erheblichen Einfluss auf folgenreiche Entscheidungen haben (Heyes u. Saberian 2019).
- Gerichte in Frankreich und den USA fällen mildere Urteile, wenn die Gerichtsverhandlung am Geburtstag der Angeklagten stattfindet (vgl. Chen u. Philippe 2019).

87 In einer quantifizierten Logik ausgedrückt, produzieren systematische Verzerrungen in der Tendenz eher schiefe Datenverteilungen, während situative Verzerrungen eher das Rauschen der Daten erhöhen.

- Sportresultate können Gerichtsentscheidungen erheblich beeinflussen (vgl. Eren u. Mocan 2018).

Persönliche situative Bedingungen

Auch persönliche Situationsbedingungen haben nachweislich einen Einfluss auf wichtige Einschätzungen und Entscheidungen, wie die folgenden Beispiele zeigen:

- Ermüdung, Stress: Ärzte verschreiben am Ende eines Arbeitstags mehr Opiode, der Krankheit nicht angebrachte Antibiotika und impfen weniger als zu Beginn des Tages (vgl. Philpot et al. 2018; Linder et al. 2014; Kim et al. 2018). Müdigkeit und Zeitdruck begünstigen die Wahl einfacher, aber falscher Lösungen (vgl. Kahneman, Sibony u. Sunstein 2021, 89).
- Hunger: Eine Analyse von Entscheidungen in israelischen Bewährungsausschüssen zeigte: Je hungriger die Richterinnen waren, desto eher fällten sie vorsichtige Urteile, gewährten also keine Bewährung – unabhängig von der jeweiligen Sachlage (vgl. Danziger, Levav u. Avnaim-Pesso 2011).⁸⁸

Die referierten Studienergebnisse belegen alle „den Einfluss von Emotionen auf Entscheidungen in einem völlig anderen Bereich bei einer einheitlich hoch gebildeten Gruppe von Personen (...), die nach Abwägung Entscheidungen treffen, bei denen viel auf dem Spiel steht“ (Eren u. Mocan 2018, 171, Übers. tz). Es ist naheliegend, dass diese Effekte ebenfalls bei der Beurteilung von Leistungen auftreten, auch wenn unter den genannten Studien nur eine sich ganz konkret auf diese Art von Urteilen bezieht. So vermuten Danziger, Levav u. Avnaim-Pesso (2011, 6892), „dass Experten bei anderen wichtigen sequenziellen (...) Urteilen andere Formen der Vereinfachung von Entscheidungen anwenden, z. B. bei (...) Entscheidungen über die Zulassung zu Universitäten“. Und sie ergänzen: „Unsere Ergebnisse ergänzen die Literatur, die dokumentiert, dass Experten nicht immun gegen den Einfluss irrelevanter Fremdinformationen sind“ (ebd.).

Wie die kognitiven können sich auch die situativ bedingten Verzerrungen überlagern – und zusätzlich mit kognitiven Verzerrungen vermischen. So kann die sequenzielle Platzierung einer Prüfung die Beurteilung der in ihr gezeigten Leistung doppelt beeinflussen – sowohl über systematische kogni-

88 Weil es sich bei den Pausen in dieser Untersuchung um Kaffee- und Essenspausen handelte, wurde gefolgert, dass der Blutzucker und damit das Hungergefühl eine entscheidende Rolle gespielt haben dürften. Es ist aber denkbar, dass eher allgemeine Ermüdung die entscheidende Rolle spielte (vgl. Cominelli 2018, 162).

tive Verzerrungen (etwa den Kontrasteffekt) als auch über situative Einflüsse auf Stimmung oder Müdigkeit der Beurteilenden.⁸⁹

Die Vielzahl an Verzerrungsquellen erklärt, weshalb Leistungsbeurteilungen niemals eine mechanische Präzision erreichen werden und Maßnahmen zur Erhöhung der Transparenz und Gültigkeit von Leistungsbeurteilungen so bedeutsam sind – von der Lernzielformulierung (Kap. 5) über die Orientierung an einer angemessenen Bezugsnorm (Kap. 6.2) bis hin zur Sicherstellung der Qualität von Leistungsbeurteilungen.

6.5 Qualität der Leistungsbeurteilung sicherstellen

Die Widersprüche zwischen quantitativen und qualitativen Gütekriterien und den ihnen zugrundeliegenden Orientierungen, aber auch die angeführten Zielkonflikte zeigen: Die Qualität von Leistungsbeurteilungen sicherzustellen, ist keine leichte Aufgabe und erfordert einen gewissen Pragmatismus.

Zentral ist mit Sicherheit, der Inhaltsvalidität große Aufmerksamkeit zu schenken: Nur Leistungsbeurteilungen, die tatsächlich jene Leistungen messen, die durch spezifische Lernziele (siehe Kap. 5.4) vorgegeben wurden, können gültig sein – unabhängig davon, ob man von qualitativen oder quantitativen Gütekriterien ausgeht (vgl. Ingenkamp u. Lissmann 2008, 166).

Als Entscheidungshilfe dafür, ob man sich am quantitativen oder am qualitativen Paradigma orientieren möchte, stellt Tabelle 3 die wichtigsten Eigenschaften und Eignungen von quantitativen und qualitativen Gütekriterien einander gegenüber. Dabei ist zu beachten, dass die meisten Leistungsnachweissettings eher die Voraussetzungen für das qualitative Paradigma erfüllen. Denn sobald es um kognitiv hochrangige Lernziele geht, was sich meist an einem höheren Anteil an divergenten Denkaspekten (siehe Kap. 7.1.6) und Aufgaben mit komplexen Problemstellungen zeigt, und wenn die Zahl der beurteilten Studierenden deutlich unter 100 sinkt, sind wesentliche Grundbedingungen für ein quantitativ orientiertes Vorgehen nicht mehr gegeben. Mit anderen Worten: Quantitativ orientierte Settings sind letztlich nur bei hohen Studierendenzahlen, relativ niederrangigen Lernzielen und einer hohen Anzahl geschlossener Aufgaben sinnvoll.⁹⁰

89 Beispielsweise fand Baumann (1977), dass die Reihenfolge die Aufsatzbewertung beeinflusst: Später bewertete Aufsätze wurden besser bewertet.

90 Aber selbst dann stellen sich noch wesentliche Fragen des Constructive Alignments. So zeigte eine Untersuchung von Scouller (1998), dass Studierende oberflächlichere Lernstrategien anwenden und dadurch weniger tiefreichendes Wissen und Können erwerben, wenn die gleiche Veranstaltung mit einem Multiple-Choice-Test geprüft wird statt mit einem Essay.

Tab. 3: Gegenüberstellung der quantitativen und qualitativen Gütekriterien für Leistungsnachweise

	Quantitative Kriterien	Qualitative Kriterien
Orientierung	Objektivismus	Intersubjektivismus/ Hermeneutik
Leitmetapher	Messen	Verstehen
Art der Leistung	Nur konvergente Leistungen	Konvergente und divergente Leistungen, komplexe Problemlösungen
Geeignete Aufgabentypen	Geschlossene und ggf. halb-offene Aufgaben Hohe Aufgabenzahl nötig	V.a. offene Aufgaben
Anspruchsniveau	V.a. tiefe Stufen der Lernzieltaxonomien	Alle Stufen der Lernzieltaxonomien
Anzahl Studierende	Mathematische Verfahren funktionieren nur bei Studierendenzahlen > 100	Studierendenzahl spielt keine Rolle

Die Grundfrage beim Beurteilen von Leistungen lautet bekanntlich: Wie können wir möglichst zuverlässig und gültig von Performanz auf Kompetenz, also von gezeigtem Verhalten und erarbeiteten Artefakten auf das latente Wissen und Können der Studierenden schließen? Dazu müssen wir das Erfüllen der Gütekriterien optimieren.

6.5.1 *Optimierung von Objektivität, Intersubjektivität und Reliabilität*

Die drei Gütekriterien, bei denen es sich gemäß der jeweiligen Orientierung (quantitativ/qualitativ) eher um notwendige Voraussetzungen denn um hinreichende Kriterien handelt, werden hier gemeinsam behandelt. Ihre Erfüllung kann namentlich durch folgende eher formale Maßnahmen optimiert werden und spielt vor allem für summative Bewertungen eine Rolle (vgl. Sacher 2014, 42–44; 47–52):

- **Explication und Präzisierung:** Die Prozesse der Durchführung und Beurteilung von Leistungsnachweisen sollten möglichst präzise beschrieben sein. Dies ist umso wichtiger, wenn mehrere Personen (verschiedene Dozierende, Assistierende, Mitarbeitende der Studienadministration, externe Expertinnen) an der Durchführung und Beurteilung des Leistungsnachweises beteiligt sind.

- **Gemeinsame Eichung der Beurteilenden:** Sind mehrere Personen an der Beurteilung beteiligt, sind Beurteiler-Trainings oder Beurteiler-Werkstätten nötig, um eine ausreichende Kalibrierung der Beurteilungen sicherzustellen. Diese Verständigung unter den Beurteilenden ist sowohl bei qualitativer wie quantitativer Ausrichtung der Beurteilung essenziell. Sehr oft werden bei solchen Gelegenheiten auch nochmals Optimierungsmöglichkeiten im Leistungsnachweis-Setting, in den Beurteilungskriterien oder weiteren Aspekten entdeckt.
- **Vielfalt der Aufgaben:** In der Regel haben Leistungsnachweise eine umso geringere Reliabilität, je weniger Aufgaben sie umfassen. Deshalb sollten entweder mehrere kürzere/kleinere Leistungsnachweis-Teile über eine längere Zeitdauer (z. B. während eines Semesters oder Studienjahrs) zu einem Sammel-Leistungsnachweis kombiniert werden, oder ein größerer Leistungsnachweis sollte mehrere und verschiedenartige Aufgaben umfassen. Diese Maßnahme erhöht übrigens auch die Validität und Authentizität. Eine Ausnahme stellen hier Leistungsnachweise dar, die per se schon ein sehr umfassendes Leistungsspektrum anvisieren, wie etwa längere schriftliche Arbeiten oder Portfolios.
- **Reihenfolge variieren:** Eine einfache, aber sehr wirksame Maßnahme ist das Variieren der Reihenfolge, in der Leistungsnachweise oder einzelne Aufgaben korrigiert werden. Umfasst eine schriftliche Prüfung beispielsweise 10 Aufgaben, so sollte jede Aufgabe in einer möglichst unterschiedlichen Reihenfolge beurteilt werden (auf Papier kann man Prüfungen mischen wie Spielkarten, online gibt es Zufallsgeneratoren für solche Anlässe). So lässt sich der Kontrasteffekt reduzieren (siehe Kap. 6.4.1).
- **Anonymisierung:** Gerade elektronisch eingereichte schriftliche Leistungsnachweise lassen sich auf einfache Weise anonymisieren. So können weitere Verzerrungsquellen reduziert werden (z. B. Halo-Effekt, Erwartungseffekt der Beobachtenden, geschlechtsspezifische Vorurteile, siehe Kap. 6.4.1).
- **Schwankungsbereich von Noten:** Quantitative Auswertungen zeigen, dass sich Messfehler bei keiner Art von Leistungsbeurteilung verhindern lassen. Nach Ingenkamp u. Lissmann (2008, 56f.) beträgt der Messfehler im Notensystem Deutschlands (und damit auch der Schweiz) in der Regel eine ganze Notenstufe nach oben und unten. Eine 3 könnte also eine 2 oder eine 4 sein; eine 4 könnte eine 3 oder eine 5 sein. Ziffernbenotungen haben somit einen enorm hohen Schwankungsbereich, was sie besonders in dem Bereich problematisch macht, in dem es um die Frage geht, ob eine Leistung noch als genügend eingestuft wird oder nicht. Umso wichtiger ist, dass Beurteilende und Personen, die über Zulassungen etc. entscheiden, über diesen Umstand Bescheid wissen und seine Folgen ein-

schätzen können. Der Begriff des Konfidenzintervalls (d. h. Schwankungsbereich) und seine Bedeutung für die Notengebung werden in Kap. 9.4.6 ausführlich erläutert.

- Kritisch ist bei summativen Beurteilungen die Unterscheidung zwischen genügenden und ungenügenden Leistungen: Am zuverlässigsten werden extrem gute und extrem schlechte Leistungen beurteilt. Dagegen werden mittlere Leistungen in der Regel durch verschiedene Beurteilende oder durch die gleiche Beurteilerin zu verschiedenen Zeitpunkten ziemlich unterschiedlich beurteilt. Wie das Notenbeispiel im vorangehenden Aufzählungspunkt zeigt, geht es aber gerade in diesem Bereich jeweils um die Frage, ob eine Leistung noch genügend ist oder nicht – und damit an hochschulischen Prüfungen schnell einmal um die Frage, wer weiterstudieren kann, wer ein Modul wiederholen muss oder ganz aus dem Studium ausgeschlossen wird. Dass sogar das Wetter bei 20 % der Studierenden mitentscheidet, ob sie ein Aufnahmegespräch bestehen oder nicht (siehe Kap. 6.4.2), verdeutlicht die Wichtigkeit, bei mittleren Leistungen genauer hinzuschauen. Geht es um folgenreiche Entscheidungen, also summative Bewertungen mit berechtigender oder qualifizierender Wirkung, sollten bei Bewertungen, die sich im Konfidenzintervall zwischen Bestehen und Nichtbestehen befinden, zusätzliche Beurteilende herangezogen oder ergänzende Nachprüfungen durchgeführt werden (siehe Kap. 9.4.6 zur Bedeutung und Berechnung des Konfidenzintervalls).

6.5.2 Optimierung von Authentizität und Validität

Bei den Gütekriterien der Authentizität und Validität handelt es sich um die zentralen Kriterien dafür, ob in einem Leistungsnachweis wirklich die Leistung erhoben wird, die beurteilt werden soll. Die Maßnahmen zu ihrer Erhöhung sind bei der quantitativen und qualitativen Orientierung ähnlich und lauten (vgl. Sacher 2014, 46–52):

- Lernzielbezug: Die Aufgabenstellungen der Leistungsnachweise und die Beurteilung der gezeigten Leistungen sollten sich so direkt wie möglich auf die spezifischen Lernziele (siehe Kap. 5.4) beziehen und infolge des Constructive Alignments (siehe Kap. 5.1) auch auf die in Unterricht und Selbstlerneinheiten erfolgten Lehr-Lern-Aktivitäten. Dazu gehören insbesondere lernzielbezogene Beurteilungsinstrumente wie Beurteilungsraster, die mit den Studierenden frühzeitig besprochen werden (siehe Kap. 9.3).
- Korrespondenz von Aufgabenart und Lernziel: Nicht alle Arten von Aufgaben sind gleich geeignet für die Überprüfung eines spezifischen Lernziels. Mit einem Multiple-Choice-Test lässt sich beispielsweise kaum er-

fassen, ob eine Ärztin oder ein Krankenpfleger gut mit Patienten kommunizieren kann.

- Trennung unterschiedlicher Fachgebiete: Verschiedene Wissens- und Könnensgebiete sind möglichst getrennt zu erfassen und beurteilen. So sollte z. B. die Rechtschreibung – ggf. mit Ausnahme verwechslungsanfälliger Fachbegriffe, sofern die Wichtigkeit ihrer Schreibweise vorgängig thematisiert wurde – nur in Sprachfächern in die Beurteilung einbezogen werden.
- Bezugsnorm: Eine Orientierung an der kollektiven Bezugsnorm wird vermieden (siehe Kap. 6.2 und 5.1.3).
- Variation: Die Durchführungsbedingungen von Leistungsnachweisen sollten variieren, so dass es z. B. nicht immer um Gedächtnisleistungen oder um möglichst hohe Bearbeitungsgeschwindigkeit geht (zwei Facetten von Wissen und Können, die in vielen gebräuchlichen Prüfungsformaten gegenüber mindestens so wichtigen Facetten wie Übersicht über ein breites Themenspektrum oder tiefenorientierte Bearbeitung komplexer Problemstellungen übermäßig fokussiert werden). Verschiedene Facetten von Wissen und Können kann man erheben, indem man Formate von Leistungsnachweisen (schriftlich, mündlich, Zeitdauer, Medialität etc.), verschiedene Beurteilungssituationen (Bewertung von Einzelleistungen, von Tandem- oder Gruppenleistungen, Hausarbeiten, Klausuren/mündliche Prüfungen etc.) und Aufgabentypen (offene und geschlossene Aufgaben, divergente und konvergente Leistungen) variiert. Die Variation sollte einerseits im Längsschnitt über das Curriculum hinweg erfolgen (siehe Kap. 10), andererseits aber auch im Querschnitt über verschiedene Aufgabentypen in einem größeren Leistungsnachweis oder in Sammel-Nachweisen zu einem Modul oder einer Modulgruppe (siehe Kap. 7).

Bezüglich aller Gütekriterien ist zu betonen, dass alle Formen der Planung, Durchführung und Beurteilung von Leistungsnachweisen, bei denen mehr als eine Person beteiligt sind, ihre Güte erhöhen. Denn durch den Austausch verschiedener Perspektiven werden typische Mängel der menschlichen Kognition mit wesentlich höherer Wahrscheinlichkeit überbrückt oder sogar eliminiert (siehe Kap. 6.4 zu kognitiven Verzerrungen und Kap. 4.7 zum argumentativen Schlussfolgern als sozialem Prozess). Nach der Klärung dieser grundlegenden Voraussetzungen können wir uns im nächsten Kapitel mit den verschiedenen Formaten von Leistungsnachweisen auseinandersetzen.

7 Formate von Leistungsnachweisen

Dieses Kapitel soll Ihnen praktische Einsichten vermitteln und Anregungen dazu geben, wie sich Leistungsnachweise so gestalten lassen, dass sie die erwünschten Auswirkungen haben. Es geht also um die Gestaltung von Leistungsnachweisen, welche die Studierenden zu den erwünschten Lernhandlungen motivieren, ihnen die Demonstration des angestrebten Wissens und Könnens erlauben und eine angemessene Beurteilung und Bewertung des gezeigten Wissens und Könnens ermöglichen. Zu diesem Zweck werden zunächst wesentliche Grundeigenschaften verschiedener Leistungsnachweisformate vorgestellt (Kap. 7.1). Dazu gehören etwa die Funktion des Leistungsnachweises oder die Realisierungsform: Dient der Leistungsnachweis primär der Förderung des Lernens oder der Berechtigung/Qualifikation? Wird die Leistung in mündlicher oder schriftlicher Sprache erbracht und/oder kommen weitere Realisierungsformen zum Einsatz wie Bilder, körperliche Aktivitäten (Tanz, Sport, Schauspiel, physische Arbeit an Patienten etc.)?

Ein weiterer wichtiger Aspekt ist die Aufgabenstellung (Kap. 7.2): Wie ist sie konstruiert, auf welche Arten von Leistung zielt sie, wie stehen sie im Verhältnis zu späteren Tätigkeitsfeldern der Studierenden? Und wie lassen sich die verschiedenen Aufgaben oder Leistungsnachweisformate kombinieren (Kap. 7.3) und klassifizieren (7.4)? Auf diesen grundlegenden Überlegungen basiert die Beschreibungsstruktur für Leistungsnachweise, die in Kapitel 7.5 vorgestellt wird. Sie wird genutzt, um vier „klassische“ Formate von Leistungsnachweisen (Kap. 7.6) und drei innovativere Formate (Kap. 7.7) näher vorzustellen und zu analysieren. Eine abschließende Aufstellung gibt Hinweise zu weiteren Gestaltungsmöglichkeiten für innovative- und lerngerechtere Leistungsnachweise (Kap. 7.8).

7.1 Grundeigenschaften von Leistungsnachweisen

Es gibt aus lerntheoretischer und didaktischer Sicht keine grundsätzlichen Beschränkungen, durch welche Nachweisformate Leistungen erhoben und beurteilt werden können. Übergeordnete gesetzliche Vorgaben und institutionsspezifische Reglemente können aber Einschränkungen vorsehen, was vor allem in letzteren gelegentlich der Fall ist. Dabei sind gerade die Beschränkungen durch institutionelle Reglemente der Hochschulen oder Studiengänge oft weniger einengend, als ihnen unterstellt wird: Bei genauerem Hinschauen lassen sie oft mehr an Vielfalt in Formaten und Beurteilungs-

formen zu als gedacht, und als einengend erweisen sich eher überkommene Vorstellungen („in unserem Institut wird das so gemacht“).

Zu bedenken ist zudem, was in Kapitel 2.5.2 herausgearbeitet wurde: Das Leistungsnachweisformat konstruiert ein Setting, in dessen spezifischen Bedingungen die Studierenden eine Handlung vollführen, die als Leistung interpretiert wird. Leistungsnachweise als institutionelle Settings sind immer sozial konstruiert, und diese Konstruktion beeinflusst wesentlich, welche Art von Leistungen die Studierenden in welchem Ausmaß zeigen können! Dieses Kapitel bietet deshalb einen Überblick über die wichtigsten Eigenschaften von Leistungsnachweisen, die variiert werden können.

7.1.1 Funktion: *formativ versus summativ*

Die Unterscheidung zwischen formativer und summativer Leistungsbeurteilung wurde in Kapitel 3 ausführlich beschrieben. Sie bezieht sich auf unterschiedliche Zwecke und Zeitpunkte der Beurteilung und Rückmeldung von Lernfortschritten und -ergebnissen: Formative Leistungsbeurteilungen zielen darauf, den Lernprozess zu fördern. Wesentlicher Bestandteil sind kontinuierliche Rückmeldungen, die den Studierenden helfen, ihr künftiges Lernen zu steuern. Summative Leistungsbeurteilungen zielen demgegenüber darauf, die Lernleistung abschließend zu beurteilen und daraus eine Form von Qualifikation oder Berechtigung abzuleiten (Noten, Diplome, Zugangsberechtigungen).

In der Regel kann ein Leistungsnachweis zu einem bestimmten Zeitpunkt nur eine der beiden Funktionen haben. Falls aber länger an einem Produkt gearbeitet wird, können verschiedene Etappen oder Aspekte des gleichen Arbeitsprozesses respektive seines Resultates zu unterschiedlichen Zeitpunkten mit unterschiedlicher Zielsetzung beurteilt werden. Üblicherweise erfolgt dann zunächst eine formative und zum Abschluss eine summative Beurteilung (siehe Kap. 7.7.1 zur mehrphasigen Beurteilung).

7.1.2 Prozess und Produkt

Eine Unterscheidung, die gelegentlich mit jener zwischen formativer und summativer Beurteilung gleichgesetzt wird, aber durchaus eine andere Ebene betrifft, ist jene zwischen Lernprozess und Lernprodukt. Insbesondere summative Beurteilungen sind stark am Endresultat von Lernprozessen orientiert, das meist anhand produktartiger Erzeugnisse wie schriftlichen Arbeiten, Klausurlösungen, Programmiercodes, künstlerischen Artefakten etc. oder dann anhand performativer Tätigkeiten wie Präsentationen oder Aufführungen abgelesen wird. Entsprechend wird in der Beurteilungsliteratur oft gefordert, dass dem Lernprozess bei der Beurteilung und vor allem bei der forma-

tiven Leistungsrückmeldung mehr Beachtung geschenkt werden solle (vgl. Sacher 2014, 201; Winter 2015, 63f., Ziegenspeck 1999, 56f.).

So moniert Sacher (ebd.), dass in produktartigen Leistungsnachweisen Aspekte wie Anstrengungsbereitschaft, durchdachtes Vorgehen, Selbststeuerung und Metakognition, Methodenbewusstsein oder Kooperationsfähigkeit kaum sichtbar werden. Um Einblick in diese Aspekte zu erhalten, müssen auch Elemente des Lernprozesses betrachtet werden – Rückmeldungen dazu sind zentral, um das Lernen der Studierenden zu fördern.

Leistungsnachweisformate, die besonders gut Einblicke in den Lernprozess ermöglichen, sind Portfolios und Lerntagebücher (inklusive Blogs). Aber auch in Gruppenarbeiten geben sich die Studierenden bei sinnvoller Aufgabenkonstruktion (klares Ziel und klare Rollenverteilung zwischen den Gruppenmitgliedern) in der Regel aufgabenbasiertes Peer-Feedback zu Aspekten des Lernprozesses, was man durch entsprechende Aufforderungen und Leitfragen systematisieren kann. Zudem kann bei der Erarbeitung von längeren Produkten ganz bewusst ein zwei- oder mehrphasiges Beurteilungsverfahren gewählt werden, bei dem die Studierenden zu Zwischenständen ihrer Arbeit ein Feedback von Lehrenden und/oder Peers erhalten und ihr Produkt auf dieser Basis weiterentwickeln (siehe Kap. 7.7.1).

7.1.3 Sprachliche und weitere Realisierungsformen

Eine Unterscheidung, die in der Literatur zur Leistungsbeurteilung stets vorgenommen wird, ist jene zwischen mündlichen und schriftlichen Leistungsnachweisen (vgl. z. B. Walzik 2012, 51–64; Metzger u. Nüesch 2004, 41–123); teilweise wird auch die Trias mündlich – schriftlich – praktisch unterschieden (siehe Kap. 7.4.1; vgl. Sacher 2014, 56). Diese Unterscheidung greift aber zu kurz – umso mehr seit dem Aufkommen digitaler Medien. So sind neben mündlichen und schriftlichen auch bildliche oder gar physische Leistungen oder Produkte möglich, z. B. sachliche oder künstlerische Bild Darstellungen, Baupläne, Skulpturen oder architektonische Modelle. Bei den „praktischen“ Leistungen ist zu unterscheiden zwischen Tätigkeiten „in vivo“ wie Prüfungslektionen von angehenden Lehrpersonen in echten Schulklassen oder Behandlungen von Modellpatientinnen durch angehende Gesundheitsfachleute auf der einen Seite und eher auftrittsartigen Aktivitäten wie künstlerischen Darbietungen oder sportlichen Leistungen. Zudem können beinahe alle genannten Tätigkeiten auch im (teil-)digitalen Raum und/oder mit Hilfe von Maschinen – von Large Language Models wie ChatGPT bis zu Robotern – ausgeführt werden. Und natürlich gibt es Kombinationsmöglichkeiten, z. B. von Text und Bild, analogen und digitalen Anteilen etc. Jegliche Dichotomien werden angesichts dieses Möglichkeitsraums schwierig (siehe Kap. 15).

7.1.4 *Zeit, Raum und Medium*

Auch bezüglich Zeit, Raum und Medium sind die Unterscheidungsmöglichkeiten alles andere als dichotom. Einerseits sind schon schriftlich und mündlich zwei verschiedene Medialitäten, auch wenn wir sie oben unter die Realisierungsformen gezählt haben. Andererseits sind insbesondere im digitalen Raum zahlreiche Kombinationsmöglichkeiten von raumzeitlichen und medialen Anordnungen vorhanden. So gibt es zu Settings, die auf dem menschlichen Kontakt von Angesicht zu Angesicht basieren (z. B. vor Ort erfolgende schriftliche Klausuren und mündliche Prüfungen) nicht ein klares Gegenstück, sondern einen Fächer an Alternativen. Dieser beginnt bereits bei Hausarbeiten, bei denen keine digitale Vermittlung nötig ist. Bei den digitalen Möglichkeiten ist schließlich im Minimum zwischen synchronen und asynchronen, zwischen text-, audio- und audiovisuellen Möglichkeiten zu unterscheiden (vgl. Digital Learning und Zentrum für Hochschuldidaktik und -entwicklung 2020, 2f.). Hinzu kommen spezifische Möglichkeiten wie CAD, Fernsteuerung von Maschinen wie digitalen chirurgischen Operationssystemen etc. und natürlich angereicherte und virtuelle Realität – nicht nur in technischen Gebieten, sondern auch in sozialen und kommunikativen Studiengängen wie der Lehrerbildung (vgl. Gerholz, Maidanjuk u. Schlottmann 2022).

7.1.5 *Sozialform der Leistungserbringung*

Wie in Kapitel 2.5 herausgearbeitet wurde, hat menschliches Lernen immer eine soziale Komponente, und auch Wissenschaft wäre eine sinnlose Vorstellung ohne eine Gruppe von Menschen, welche das geschaffene Wissen nutzen. Insofern sind Leistungsnachweis-Settings, in denen viel Mühe darauf verwendet wird, dass einzelne Menschen ohne Hilfe von außen eine Leistung erbringen, in verschiedener Hinsicht künstlich. Einmal sind die Lernenden vor der Beurteilungssituation nicht solipsistisch unterwegs, sondern Teil einer Gesellschaft und von Peer- und familiären Systemen, so dass ihr Lernen selbst unter Ausklammerung der Lehrveranstaltungen stets in einem sozialen Kontext stattfindet.⁹¹ Des Weiteren macht das beurteilte Wissen nur in einem sozialen Kontext Sinn – sei dieser die wissenschaftliche Gemeinschaft der Academia oder eine berufliche Tätigkeit. Und schließlich wird Wissen und Können im realen Leben – gleich ob Academia oder andere Berufsfelder – nur ganz selten ohne die Möglichkeit sozialer Unterstützung oder anderer Hilfsmittel verwendet.

91 Dass dieser soziale Kontext einen erheblichen Einfluss auf die Leistung(sfähigkeit) der Lernenden hat, wurde in Kap. 2.2 und 2.3 herausgearbeitet.

Insofern ist die große Betonung von Leistungsnachweisen, in denen „Prüflinge“ möglichst systematisch vereinzelt werden, in der angestammten Beurteilungspraxis weder dem epistemologischen Zustandekommen noch dem Erwerb oder dem Gebrauch von Wissen und Können sonderlich angemessen. Umso mehr macht es Sinn, neben so genannten „Einzelleistungen“ auch Leistungsnachweise zu gestalten, in denen Studierende Leistungen ganz offiziell in Tandems oder Gruppen erbringen und bei denen die Beurteilung auf diesen Umstand Rücksicht nimmt. So können in kollektiven Leistungsnachweis-Settings gezielt auch Aspekte des Sozialverhaltens adressiert und beurteilt werden (vgl. Walzik 2012, 65–68).⁹²

7.1.6 Lösungsrichtungen: Konvergentes und divergentes Denken

Konvergentes und divergentes Denken sind die beiden Pole einer Skala der kognitiven Herangehensweise an die Beantwortung von Fragen und das Lösen von Problemen.⁹³ Auf der konvergenten Seite geht das Denken davon aus, dass es auf eine Frage nur eine richtige Antwort und für ein Problem nur eine Lösung gibt und sucht konsequent nach dieser richtigen Antwort oder Lösung. Am anderen Ende der Skala sucht das divergente Denken nach mehreren Perspektiven und mehreren möglichen Antworten auf Fragen und Probleme; dabei erkundet es Alternativen zu den akzeptierten Vorgehensweisen (vgl. Kim u. Pierce 2013, 245f.).

Konvergentes Denken ist nützlich und sinnvoll, wenn eine Aufgabe abgeschlossen, ein Problem gelöst werden soll. Divergentes Denken ist im Gegensatz dazu dann nützlich, wenn neue Ideen, alternative Lösungen und Kreativität

92 Gemäß Hirdina (2019, 56f.), dürfen in Deutschland aus prüfungsrechtlicher Sicht Studierende bei Gruppenleistungen nur für einen klar abgrenzbaren Teil der Gruppenarbeit beurteilt werden (in der Schweiz und Österreich scheint die Rechtslage weniger klar zu sein). Eine derart individualistische Rechtssetzung (oder -auslegung) beschränkt die Beurteilung von Kooperationsfähigkeit aus juristischer Sicht auf den formativen (= unbewerteten) Bereich. Aus pädagogischer und lernpsychologischer Perspektive wäre eine diesbezügliche Anpassung von Gesetzgebung und/oder Rechtspraxis an den wissenschaftlichen Erkenntnisstand Not.

93 Zu beachten ist freilich, dass Extremausprägungen sowohl konvergenten wie divergenten Denkens kaum mehr hilfreich sind: Extrem konvergentes Denken ist zu rigide und schränkt die Lösungsfindung ein, extrem divergentes Denken produziert zu viele Möglichkeiten und ist unfähig, zum Abschluss zu kommen. Das produktivste Denken findet deshalb dann statt, wenn sich die beiden Arten des Denkens gegenseitig ergänzen und zu sinnvollen Ergebnissen führen (vgl. Kim u. Pierce 2013, 247f.).

tät gefragt sind. Allerdings würde es zu kurz greifen, divergentes Denken mit Kreativität gleichzusetzen. Vielmehr ist bei kreativen Problemlösungen zum Abschluss der Problemlösung nach einer Phase divergenter Erkundung des Problemraums auch konvergentes Denken nötig (vgl. Kim u. Pierce 2013, 246).

Konvergente Leistungen lassen sich an einem klaren kriterialen Maßstab messen. Sie können in einfacheren Fällen als richtig oder falsch beurteilt werden, in komplexeren Fällen kann zumindest ihre Güte klar beurteilt und in der Regel auch quantifiziert werden. Das umfasst einen großen Bereich „in den verschiedensten Fachgebieten und Niveaustufen, die vom Faktenwissen bis zur begründeten Wertung reichen“ (Ingenkamp u. Lissmann 2008, 133).

Divergente Leistungen können zu vielen gleichwertigen Ergebnissen führen oder zumindest zu vielen Ergebnissen, deren Güte nicht mit eindeutigen Maßstäben in eine Rangreihe gebracht werden kann. Typischerweise denkt man hier an Fachgebiete, in denen kreative Leistungen eine große Rolle spielen, etwa alle Künste oder auch gewisse geisteswissenschaftliche Gebiete wie die Interpretation von Literatur (vgl. Zimmermann, Tobias 2018, 56).

Aber es gibt in allen Fächern divergente Leistungen, vor allem dort, wo soziale Elemente zum Tragen kommen. Damit sind divergente Leistungen vor allem für das praktische Handeln eines Fachexperten relevant: Es mag vielleicht von der Statik aus betrachtet nur eine beste Lösung geben, wie man an einem bestimmten Ort eine Brücke mit einer bestimmten Tragfähigkeit bauen kann – aber je nachdem, welche Anforderungen die Gesellschaft bezüglich landschaftlicher Gestaltung, Lärmschutz, Schattenwurf oder Finanzrahmen stellt, muss auch der Baustatiker verschiedene Aspekte gegeneinander abwägen und nach einer möglichst situationsgerechten Lösung suchen. (Zimmermann, Tobias 2018, 56)

Gelegentlich wird die Ansicht vertreten, divergente Leistungen könnten nicht gültig beurteilt werden. Mit Bezug auf quantitative Qualitätskriterien mag dies stimmen: Einer quantifizierenden Beurteilung sind divergente Leistungen nur schwer zuzuführen. Aber mit Bezug auf die qualitativen Qualitätskriterien intersubjektivität und Authentizität ist eine Beurteilung sehr wohl möglich. Wie bei jeder fairen Beurteilung ist es dazu wichtig, dass die Qualitätskriterien im Voraus bekannt gegeben und mit den Studierenden besprochen werden. Beurteilungsraster (siehe Kap. 9.3) eignen sich für die Beurteilung von Leistungen mit Anteilen divergenten Denkens besonders gut, wozu in der Regel auch komplexe Problemlösungen gehören, wie sie zum Beispiel anhand von Fallstudien vorgenommen werden müssen. Würde eine Aufgabenstellung im Sinne des oben zitierten Beispiels nicht nur auf die statische Belastbarkeit einer Brücke zielen, sondern die Erfüllung und Abwägung weiterer Ansprüche umfassen, dann gäbe es nicht eine einzige richtige Lösung – aber sehr wohl besser gelungene Lösungen, die fachlich angemessen oder sogar innovativ sind (hier ergänzen sich divergentes und konvergentes Denken) und solche, bei denen wesentliche Überlegungen oder Aspekte fehlen.

7.2 Aufgabenstellungen

Bisher wurden fast ausschließlich Leistungsnachweise und Leistungsbeurteilungen auf übergeordneter Ebene behandelt. Allerdings ist zu beachten, dass je nach Format des Leistungsnachweises ein einzelner Nachweis aus einer mehr oder weniger großen Anzahl verschiedener Aufgabenstellungen bestehen kann. Deshalb ist es wichtig, verschiedene Arten von Aufgaben zu trennen und zu wissen, welche Aufgabenarten sich zur Überprüfung welcher Arten von Lernzielen eignen. Dabei sind Aspekte der Aufgabenkonstruktion ebenso zu beachten wie das inhaltliche Anspruchsniveau und der Bezug zu späteren Tätigkeitsfeldern der Studierenden.

7.2.1 Geschlossene, halboffene und offene Aufgaben

Bezüglich der Konstruktionsweise der Aufgaben ist zwischen geschlossenen, halboffenen und offenen Aufgaben zu unterscheiden (vgl. Sacher 2014, 59–66).

Geschlossene Aufgaben

In der Formulierung der Aufgabenstellungen sind die Antwortmöglichkeiten vollständig enthalten, darunter auch die richtige(n). Diese Konstruktionsweise setzt voraus, dass es auf die gestellte Frage eine oder mehrere richtige Antworten oder zumindest eine klar angemessenste Antwort gibt. Geschlossene Aufgaben sind ausschließlich geeignet, um konvergentes Denken zu erheben, sie lassen keine Alternativlösungen zu. Zudem eignen sie sich fast ausschließlich für eine schriftliche Darstellung.

Der auf Hochschulstufe am häufigsten verwendete Typ geschlossener Aufgaben dürfte die so genannte Multiple-Choice-Aufgabe sein (wobei es sich streng genommen meistens um Single-Choice-Aufgaben handelt). Dieser Aufgabentyp wird in Kapitel 13 näher erläutert. Weitere typische Beispiele sind Umordnungs- und Zuordnungsaufgaben:

- Umordnungsaufgabe: Bringen Sie die Stilrichtungen in die richtige chronologische Reihenfolge: Expressionismus – Impressionismus – Kubismus – Surrealismus (vgl. Sacher 2014, 63).
- Zuordnungsaufgabe: Welche der folgenden Begriffe gehören paarweise zusammen?

Vogel	Tracheen
Krebs	Lungen
Insekt	Kiemien

(vgl. Graf 2001, 424)

Die Beispiele sind bewusst einfach gehalten, die inhaltliche Komplexität kann durch mehr Antwortoptionen erhöht werden. Auch ist es beispielsweise möglich, Bilder und Textelemente zu- oder umordnen zu lassen, etwa für Stilelemente in der Kunst, für Diagnosen in der Medizin oder Abläufe in naturwissenschaftlichen und technischen Gebieten.

Zu beachten ist, dass die Aufgabenstellung keine impliziten Hinweise auf die richtige Lösung enthalten sollte. Diese können etwa durch Wiederholung von Begriffen oder Wortstämmen erfolgen, auch über mehrere Aufgabenstellungen hinweg. Solche unbeabsichtigten Hinweise beeinflussen den Lösungsprozess und reduzieren die Trennschärfe der betroffenen Aufgaben (das Problem der Lösungshinweise wird in Kap. 13 zu Multiple-Choice-Prüfungen näher besprochen). Deshalb lohnt es sich, erarbeitete Aufgabestellungen anderen Personen zur Durchsicht zu geben. Dazu eignen sich fachfremde Personen, denen solche Aspekte oft eher auffallen, da sie sich im Vergleich zu Expertinnen stärker an der „Oberfläche“ der Aufgaben orientieren müssen und damit vergleichbare Voraussetzungen haben wie nicht gut vorbereitete Studierende.

Halboffene Aufgaben

Im Gegensatz zu offenen Aufgaben sind die Lösungen hier nicht in der Aufgabenstellung enthalten, den Lehrenden aber bekannt. Wie die geschlossenen Aufgaben setzen halboffene Aufgaben voraus, dass es auf die gestellte Frage eine oder mehrere richtige Antworten oder zumindest eine klar angemessenste Antwort gibt. Weil die Antworten nicht explizit vorgegeben sind, besteht die Gefahr, dass Studierende sie irrtümlich als offene Aufgaben interpretieren. Es ist deshalb in der Aufgabenstellung explizit klarzustellen, dass eine ganz bestimmte Antwort bzw. Antworten aus einem ganz bestimmten Spektrum erwartet werden. Wie geschlossene erlauben halboffene Aufgaben ausschließlich die Erfassung konvergenten Denkens. Sie sind aber besser als geschlossene Aufgaben zur Verwendung in mündlichen Situationen geeignet. Der eingeschränkte Antwortraum muss dabei ebenso klar deklariert werden wie im schriftlichen Kontext, wodurch halboffene Aufgaben auch in mündlichen Prüfungssituationen den kommunikativen Austausch deutlich einschränken.

Typische halboffene Aufgaben sind Freiantworten, Ergänzungsaufgaben (meist Lückentexte), Substitutions-, Assoziations- und Aufbauaufgaben:

- Freantwort: Welches Konzil im 4. Jahrhundert hat die Trinitätslehre formuliert und welches Hauptproblem wurde dabei diskutiert?

- **Ergänzungsaufgaben**
 - Lückentext: Das Verfahren der _____ ermöglicht es, die Anzahl der Dimensionen eines Merkmalsvektors zu reduzieren, indem irrelevante Informationen entfernt werden. Dies trägt zur effizienteren _____ von Daten in der Data Science bei.
 - Andere Formen wie Ergänzung von Tabellen, Diagrammen, Grafiken, mathematischen Formeln oder Gleichungen. Beispiel: Vervollständigen Sie das Diagramm, indem Sie die fehlenden Werte eintragen (+ Abbildung eines Diagramms mit Achsenbeschriftungen, vorhandenen und fehlenden Datenpunkten).
- **Substitutionsaufgabe:** „Streichen Sie den falschen Teil der Aussage durch und schreiben Sie auf, wie es richtig heißen muss: Karbonsäuren enthalten die Karboxylgruppe-COOH und dissoziieren in wässriger Lösung in Hydroxonium-Ionen und Säure-Kationen R-COO“ (Sacher 2014, 62).
- **Assoziationsaufgabe:** Nennen Sie fünf Methoden und Techniken, die im Maschinenbau verwendet werden, um die Effizienz und Leistung von Verbrennungsmotoren zu verbessern.
- **Aufbauaufgabe:** Erstellen Sie aus den nachfolgenden Begriffen ein hierarchisches Begriffsnetz (vgl. Graf 2001, 423).

Wiederum sind die Beispiele um der allgemeinen Verständlichkeit willen einfach gehalten. Die fachliche Komplexität kann gesteigert werden, wobei sich auch hier bildliche oder multimediale Elemente anbieten.

Bei halboffenen Aufgaben ist ebenfalls das Risiko von impliziten Lösungshinweisen zu beachten. Zusätzlich zu den bereits bei den geschlossenen Aufgaben besprochenen Aspekten kommt insbesondere bei Lückentexten die Größe der Lücke hinzu: Variiert diese beispielsweise in Korrelation mit der Länge des gesuchten Wortes, gibt dies unnötige Hinweise auf die gesuchte Antwort. Darüber hinaus können grammatikalische Aspekte wie das Geschlecht von Artikeln oder die Subjekt-Verb-Kongruenz unerwünschte Hinweise auf das in der Lücke gesuchte Wort geben. Deshalb empfiehlt sich auch hier eine Kontrollbearbeitung der erstellten Aufgaben durch eine fachfremde Person.

Offene Aufgaben

Offene Aufgabenstellungen unterscheiden sich von geschlossenen und halboffenen Aufgaben durch die wesentlich größere Freiheit, wie die Aufgabe bearbeitet respektive die Fragestellung beantwortet werden kann. Diese Freiheit muss sich nicht auf den Inhalt der Lösung beschränken, sondern kann zudem Lösungsstrategie und Lösungswege umfassen. Offene Aufgaben sind

dann sinnvoll, wenn kreative Lösungswege zugelassen werden sollen oder wenn es keine absolut richtige oder genau definierbare Lösungsmenge gibt, wie das bei wissenschaftlichen Problemstellungen oft der Fall ist. Anders als bei den anderen Konstruktionsweisen haben hier also auch die Dozierenden keine abschließende Vorstellung davon, welche Antworten oder Lösungswege möglich oder ideal sind.

Offene Aufgaben sind somit im Gegensatz zu geschlossenen und halb-offenen Aufgaben geeignet, um divergentes Denken zu erheben. Dabei ist zu beachten, dass nicht alle offenen Aufgaben auf eine im eigentlichen Sinn kreative Problemlösung zielen. Gerade fallartige Aufgaben, wie sie im wissenschaftlichen Kontext oft sinnvoll sind, zielen in der Regel auf eine Abfolge von divergenter Möglichkeitserwägung und konvergenter Ausarbeitung spezifischer Lösungsvorschläge (siehe das Beispiel zum Brückenbau in Kap. 7.1.6). Entsprechend kann die Offenheit des Lösungsraums bei offenen Aufgaben variieren – von völliger Offenheit etwa bei klassischen Aufsatzthemen im kreativen Schreiben bis zu einer sehr relativen Offenheit wie im Beispiel der Brückenbauaufgabe, die vermutlich auf eine Kombination von Themengebieten zielt, die in einer oder mehreren Veranstaltungen oder in vorgegebener Lektüre behandelt wurden und für die durchaus klare Vorstellungen von fachlicher Angemessenheit vorhanden sind.

Die Offenheit des Lösungsraums soll sich auch in den Beurteilungskriterien zeigen, die in einer hochwertigen Leistungsbeurteilungspraxis frühzeitig bekannt gemacht und diskutiert werden (siehe Kap. 9.3.4). Solche Kriterien, namentlich in Form von Beurteilungsrastern, geben Studierenden wesentliche Hinweise, welches Ausmaß an Offenheit und somit divergentem Denken und welches Maß an Orientierung an fachlich eindeutigen Lösungsvorgaben und somit an konvergentem Denken bei der Bearbeitung eines Leistungsnachweises gefragt ist.

Das bekannteste Beispiel für offene Aufgaben sind Essay-Fragen. Es gibt aber viele weitere Formen offener Aufgaben, etwa Fallbearbeitungen, Projektaufgaben, Reflexionsaufgaben oder Forschungsaufgaben. Den Formen und dem Bearbeitungsraum sind hier viel weniger Grenzen gesetzt als bei den geschlossenen und halboffenen Aufgaben. Zudem ist es bei offenen Aufgaben möglich, dass Studierende innerhalb eines enger oder weiter gesetzten Rahmens ihre Aufgabenstellung selbst (mit-)bestimmen. Die folgenden Beispiele geben deshalb nur einen kleinen Einblick in die Möglichkeiten offener Aufgaben:

- Essay-Fragen: „Wie könnte sich der Einsatz künstlicher Intelligenz bei der Spracherzeugung auf die menschliche Kommunikation auswirken?“; „Diskutieren Sie die Herausforderungen und Chancen der Arbeit mit vielfältigen und multikulturellen Gemeinschaften in der sozialen Arbeit.“; „Eine bekannte Content-Sharing-Plattform bittet Sie um ein Rechtsgutachten mit der Fragestellung: ‚Die zentralen Grundsätze und Herausforde-

rungen hinsichtlich Rechten an geistigem Eigentum im digitalen Zeitalter, mit Fokus auf Plattformen zur gemeinsamen Nutzung von Online-Inhalten.“

- **Fallbearbeitung:** An der California State University Los Angeles wurde als abschließender Leistungsnachweis in einem Masterstudiengang in Erziehungswissenschaft mit Schwerpunkt Leseförderung ein fallbasierter Leistungsnachweis eingeführt (vgl. Ulanoff, Fingon u. Beltrán 2009, 128f., für ein illustratives Beurteilungsraster zum Leistungsnachweis vgl. 130). Dieser simuliert eine klinische Erfahrung, bei der die Studierenden als Lesespezialisten agieren und ihr Wissen und Können anwenden. Dazu mussten sie die Lese- und Schreibfertigkeiten eines echten Schülers mit Leseschwierigkeiten analysieren und basierend auf der Analyse und ihren Kenntnissen über Lese- und Schreibförderung einen Interventionsplan zur Förderung des Schülers erarbeiten. Dieser Ansatz zielte darauf ab, dass die Master-Absolvierenden so gut als möglich auf die tatsächliche Arbeit mit Schülern vorbereitet sind. Dieser Leistungsnachweis erfüllt übrigens auch die Kriterien für transferorientierte und authentische Aufgabenstellungen (siehe Kap. 7.2.2 und 7.2.3).
- **Projektaufgabe:** „Entwickeln Sie eine mobile Anwendung für eine bestimmte Zielgruppe. Wählen Sie ein relevantes Thema, wie beispielsweise Gesundheit, Bildung oder Unterhaltung. Planen Sie die Funktionalitäten, die Benutzeroberfläche und die technischen Anforderungen. Implementieren Sie die Anwendung und dokumentieren Sie Ihren Entwicklungsprozess.“ Die Übergänge zwischen Projektaufgaben und Fallbearbeitungen sind fließend, etwa wenn im Anschluss an eine Problemanalyse (Fallbearbeitung) Lösungsvorschläge erarbeitet oder sogar umgesetzt werden sollen. Dabei haben Projekte in der Regel eher Entwicklungscharakter, während Fallbearbeitungen eher analytisch orientiert sind. Entsprechend spielen kreative Lösungen und somit divergentes Denken in Projekten in der Regel eine größere Rolle als in Fallbearbeitungen.
- **Reflexionsaufgaben:** Dies sind in aller Regel nicht einzeln gestellte Aufgaben, sondern sie werden als Zusatzaufgabe zu einer vorangehenden Aufgabe gestellt. Sie eignen sich besonders bei Aufgaben, in denen Studierende mit anderen Personen interagiert haben. Beispielsweise könnte die Aufgabenstellung zum oben vorgestellten fallbasierten Leistungsnachweis auch eine Aufforderung zur Reflexion beinhalten wie: „Wie erlebten Sie Ihre Rolle als Fachexpert:in im Rahmen der Fallstudie? Wie gelang Ihnen der Beziehungsaufbau zum untersuchten Schüler? In welcher Hinsicht möchten Sie Ihr Wissen und Können bezüglich Diagnose und Förderung von Lesekompetenzen nach dem Studienabschluss weiterentwickeln?“ Reflexionsaufgaben können aber auch mündlich gestellt werden, etwa in der Nachbesprechung eines Leistungsnachweises wie einer

schriftlichen Arbeit, einer Prüfungslektion, der Arbeit mit Patienten oder einer künstlerischen Darbietung (vgl. Ryan 2012, 207f.).

- Forschungsarbeiten: Hier sind einerseits natürlich alle Fragestellungen zu studentischen Forschungsarbeiten zu nennen. Wie sinnvolle Fragestellungen aussehen, hängt stark von fachlichen Kontexten und von der Studienstufe ab, deshalb wird hier auf Beispiele verzichtet. Daneben sind auch Aufgaben möglich, die auf Teilaspekte der Forschungstätigkeit zielen, beispielsweise:
 - Durchführung einer systematischen Literaturrecherche inklusive Kommentar und/oder einer zusammenfassenden Übersicht
 - Auswertung von vorgegebenen Daten, etwa in der Sozialforschung oder den Naturwissenschaften
 - Entwicklung eines Forschungsdesigns für eine bestimmte Fragestellung (die vorgegeben oder ebenfalls durch die Studierenden zu entwickeln sein kann)
 - Verfassen einer Rezension zu einem wissenschaftlichen Artikel

Wie die Beispiele illustrieren, bieten offene Aufgaben ein beinahe unerschöpfliches Potenzial für anspruchsvolle wissenschaftliche oder wissenschaftsbasierte Fragestellungen, deren Bearbeitung eine umfassende Denkleistung erfordert. So kann mit offenen Fragen erhoben werden, ob Studierenden komplexe Zusammenhänge erfassen und analysieren sowie Lösungsvorschläge für komplexe Problemstellungen erarbeiten können. Durch die Verwendung offener Aufgabenstellungen werden kritisches Denken, Problemlösungsfähigkeiten und kreative Vorgehensweisen gefördert. Die Lernenden werden ermutigt, aktiv zu denken, ihre Ideen zu präsentieren und ihre Argumentation zu entwickeln, was zu einem tieferen Verständnis und einer aktiven Auseinandersetzung mit dem Lernmaterial führen kann. Mit offenen Aufgaben ist es somit einfacher, höhere Stufen der Lernzieltaxonomien zu adressieren, auch wenn dies mit geschlossenen und halboffenen Aufgaben theoretisch ebenfalls möglich ist. Letzteres erfordert allerdings sehr viel Aufwand in der Aufgabenentwicklung und lohnt sich nur bei äußerst hohen Studierendenzahlen und einer guten Aussicht auf wiederholte Einsetzbarkeit der entwickelten Aufgaben.

7.2.2 Reproduktions- und Transferaufgaben

Wenn Lernziele auf höhere Stufen der kognitiven Lernzieltaxonomien zielen, sollte sich die Leistungsbeurteilung nicht auf die Kontrolle beschränken, ob die Studierenden Lerninhalte auswendig gelernt haben und möglichst genau wiedergeben können. Solche Gedächtnistests sind leider immer noch vieler-

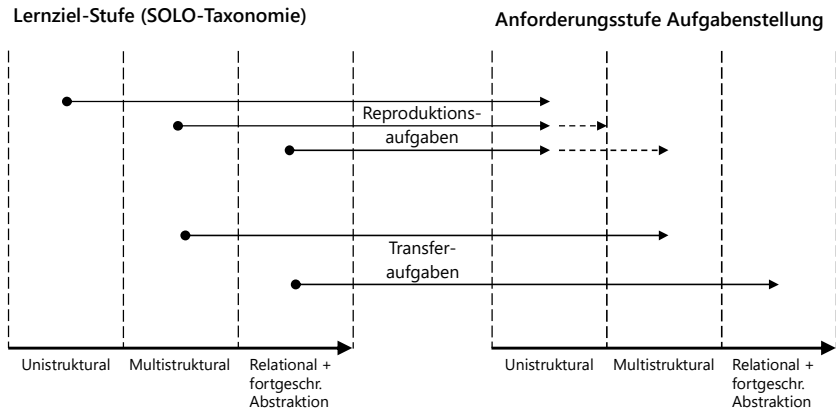


Abb. 17: Reproduktions- und Transferaufgaben. Quelle: Eigene Abbildung, in Anlehnung an Metzger u. Nüesch 2004, 17f.

orts üblich, decken aber nur die Stufen zwei und drei der SOLO-Taxonomie respektive die beiden untersten Stufen der bloomschen Taxonomie ab (siehe Kap. 5.4.2 zu den kognitiven Lernzieltaxonomien).

Damit Beurteilungen die Taxonomiestufen erfassen können, die durch die Lernziele definiert sind, ist es wichtig, dass die gestellten Aufgaben nicht nur die Reproduktion von Wissensinhalten erfordern und eine klare Unterscheidung zwischen Reproduktions- und Transferaufgaben erfolgt (siehe Abb. 17). Transferaufgaben verlangen von den Studierenden, dass sie sich mit einem Problem auseinandersetzen, das nicht bereits in der gleichen Weise behandelt oder gelöst wurde, sei es im Rahmen von Lehr-Lern-Aktivitäten oder Studienmaterialien (vgl. Feixas u. Zimmermann 2018, 86f.; Metzger u. Nüesch 2004, 17–21; Maier et al. 2010, 86f.).

Beispiel

Reproduktions- versus Transferaufgabe

Lernziel: Die Studierenden können Beurteilungsaufgaben und -formate entwickeln, die für das angestrebte Kompetenzniveau angemessen sind.

Reproduktionsaufgabe: Nennen Sie die Bezugsnormen für die Beurteilung. Wofür sind sie geeignet und was sind ihre Vor- und Nachteile?

(Die Aufgabe zielt auf die Reproduktion des vermittelten Wissens; im besten Fall adressiert sie die multistrukturale Stufe der SOLO-Taxonomie.)

Transferaufgabe: Eine Kollegin, die zum ersten Mal einen Kurs an der Universität unterrichtet, bittet Sie um Rat: Ihre Studierenden sollen die Grundlagen der ethischen Argumentation lernen. Wie kann sie am besten beurteilen, ob die Studierenden in der Lage sind, auf der Grundlage ethischer Prinzipien zu argumentieren? Erklären Sie ihr, vor welchen Herausforderungen sie steht, schlagen Sie ein Leistungsnachweisformat vor, das Ihrer Meinung nach für die Situation Ihrer Kollegin geeignet ist, und begründen Sie diesen Vorschlag.

(Die Aufgabe zielt darauf, Wissen über Beurteilung auf eine neue Situation zu übertragen und adressiert die relationale Stufe der SOLO-Taxonomie.)

7.2.3 Lebensweltbezug

Verschiedene Systematisierungen unterscheiden Aufgabenstellungen bezüglich des Grades, in dem sie sich auf die Lebenswelt der Lernenden beziehen (vgl. Maier et al. 2010; Shaffer u. Resnick 1999; Wiggins 1993, 206–55). Dabei werden Aufgabenstellungen mit hohem Lebensweltbezug als „authentisch“ bezeichnet, wobei dieser Begriff der angelsächsischen Diskussion entstammt und stark von Wiggins geprägt wurde. Allerdings zeigt sich bei genauerer Betrachtung der Literatur, dass sowohl der Lebensweltbezug als auch die Merkmale, an denen ein solcher festgemacht wird, erheblich variieren. So identifizieren Shaffer u. Resnick (vgl. 1999, 195–201) anhand einer Literaturrecherche vier unterschiedliche Verstehensweisen von authentischem Lernen, die sich teilweise überlagern:

- für die Lernenden persönlich bedeutsames Lernen
- Lernmaterialien und -aktivitäten, die sich auf die reale Welt außerhalb der Lernumgebung beziehen
- Lernen, das direkt die Praktiken bestimmter wissenschaftlicher Disziplinen anwendet
- Lernen, bei dem die Leistungsbewertung nahe an Aufgabenstellungen erfolgt, wie sie im Rahmen der vorgegebenen Lehr- und Lernaktivitäten bewältigt werden müssen

Insbesondere im durch Wiggins (vgl. 1993; Wiggins u. McTighe 2005) geprägten Verständnis von authentischer Leistungsbeurteilung („authentic assessment“) überlagern sich diese Aspekte stark. In Bezug auf konkrete Aufgabenstellungen scheint es aber sinnvoll, diese Aspekte voneinander zu unterscheiden. Je nach Lernzielen wird man den einen oder anderen Aspekt der Authentizität vielleicht stärker in den Vordergrund stellen wollen.

Eine andere Unterscheidung machen demgegenüber Maier et al. (2010, 88f.), indem sie eine graduelle Abstufung des Lebensweltbezugs vorschlagen:

1. „Aufgaben ohne Lebensweltbezug“: Sie enthalten keine Verknüpfung zwischen Fachwissen und Erfahrungswelt der Lernenden.
2. „Aufgaben mit konstruiertem Lebensweltbezug“: Sie enthalten eine Verknüpfung zwischen Fachwissen und einer Lebenswelt, die stark konstruiert wirkt und in der die Lernenden wenig Analogien zur eigenen Erfahrung erkennen dürften.
3. „Aufgaben mit konstruiertem, aber authentisch wirkendem Lebensweltbezug“: Sie stellen den Bezug zur Lebenswelt zwar künstlich her, dieser ist aber hinsichtlich der künftigen Tätigkeit der Lernenden sinnvoll. Hier eröffnet sich wieder ein Spektrum, das von eher methodischer Authentizität⁹⁴ (z. B. Analyse einer echten historischen Quelle, Laboraufgaben in naturwissenschaftlichen Fächern) bis hin zu die reale Arbeitswelt durchdringenden Aufgaben reicht (z. B. Bearbeitung echter, durch Firmen vorgelegter Problemstellungen durch Engineering- oder Wirtschafts-Studierende; Prüfungslektionen von Lehramtsstudierenden in realen Schulklassen).
4. „Aufgaben mit realem Lebensweltbezug“: Bei diesen Aufgaben tendiert die Differenz zwischen Aufgabe und Lebenswelt gegen null. Die Lernenden beschäftigen sich mit einer Problemstellung, die tatsächlich auch „im echten Leben“ gelöst werden muss. Ein typisches Beispiel wäre die Vorbereitung von Bewerbungsschreiben.

Aus der Unterscheidung wird sichtbar, dass für nachhaltige Leistungsbeurteilungen vor allem Aufgabenstellungen des dritten Typs geeignet sind. Aufgaben des zweiten Typs dürften gegenüber Aufgaben ohne Lebensweltbezug wenig Zusatznutzen bezüglich der Beurteilung des Erreichens von Lernzielen bringen, und Aufgaben des vierten Typs dürften nur in vereinzelten Fällen zu finden sein und eignen sich außerdem nur für formative Beurteilungen (es erscheint wenig sinnvoll, etwa ein reales Bewerbungsschreiben auch noch summativ beurteilen zu wollen).

Mit Bezug auf die motivationale Wirkung sind Aufgaben des dritten und vierten Typs solchen Aufgaben überlegen, die keinen deutlichen Bezug zur Erfahrungswelt der Studierenden herstellen können. Denn der Lebensweltbezug erhöht das Ausmaß, in dem Studierende die zu bewältigenden Aufgaben – und damit aufgrund des Constructive Alignments (siehe Kap. 5.1) die zu erreichenden Lernziele – als bedeutsam und somit als motivierend erleben. Zudem sprechen Aufgaben des dritten und vierten Typs, sofern sie Aspekte

94 Man könnte in diesen Fällen auch von Kontextualisierung sprechen (vgl. Villarreal et al. 2018, 841).

kooperativen Lernens oder anderen sozialen Austauschs enthalten, die ganze Breite der Selbstbestimmungstheorie nach Deci und Ryan an: Sie geben den Studierenden die Möglichkeit, sich als wirksam/erfolgreich, sozial eingebunden und eigenständig zu erleben (siehe Kap. 4.5.2 zur Selbstbestimmungstheorie).

7.3 Aufgabenkombinationen und Sammelnachweise

Aufgaben und Leistungsnachweise können in zweifacher Weise kombiniert werden. Einerseits ist zu unterscheiden zwischen Leistungsnachweisen, die aus einer einzigen Aufgabenstellung bestehen, und solchen, die mehrere Aufgaben umfassen. Handelt es sich um geschlossene Aufgabenformate wie Einfach- und Mehrfachantwortaufgaben („Multiple Choice“, siehe Kap. 13), kann ein Leistungsnachweis sehr viele Aufgaben umfassen. Handelt es sich um Fragen, die ausführlich zu beantworten sind wie etwa offene Fragen des Essay-Typs, wird ein Leistungsnachweis in der Regel deutlich weniger Aufgaben umfassen – ganz unabhängig von der konkreten Konstruktionsweise (mündlich/schriftlich, analog/digital, vor Ort/Distanz etc.).

Andererseits können verschiedene Leistungsnachweise zu einem Sammelnachweis kombiniert werden. Dies ist etwa der Fall, wenn während eines Semesters Übungsaufgaben eingereicht und beurteilt werden, deren Beurteilung dann mit jener aus einem abschließenden Leistungsnachweis wie einer Klausur, einer mündlichen Prüfung oder einem innovativeren Format (siehe Kap. 7.7) verrechnet werden. Soweit dies die Prüfungsordnung zulässt und es aufgrund von Vielseitigkeit und Umfang der Lernziele sinnvoll ist, können auch mehr als zwei Teilnachweise zu einem Sammel-Leistungsnachweis kombiniert werden.

Zu beiden Kombinationsmöglichkeiten werden im Folgenden einige Gestaltungs- und Konstruktionshinweise gegeben.

7.3.1 Kombination von Aufgaben

Werden in einem Leistungsnachweis mehrere Aufgaben gestellt, sind insbesondere bei summativen Beurteilungen folgende Aspekte zu beachten (siehe Kap. 9):

- Wird erwartet, dass die Studierenden alle Aufgaben bearbeiten oder bestehen Wahlmöglichkeiten? (Diese müssen möglichst klar signalisiert werden, damit die Studierenden im Stress der Beurteilungssituation keinen Missverständnissen unterliegen.)
- Wie viele Aufgaben müssen für die Maximalbewertung vollständig bearbeitet respektive korrekt gelöst werden?

- Werden alle Aufgaben gleich gewichtet? Falls nicht, sollte dies signalisiert werden, damit die Studierenden ihren Zeitaufwand der Gewichtung entsprechend einteilen können.
- Zeitrahmen: Können die Studierenden die Aufgaben in der Zeit lösen, die ihnen zur Verfügung steht? Wichtig: Wenn der Leistungsnachweis in erster Linie das Erreichen der Lernziele und nicht die Belastbarkeit oder Stressresistenz prüfen soll, dann empfiehlt es sich, bezüglich des Zeitbedarfs eher zurückhaltend zu sein – also genügend Zeit oder nicht zu viele Aufgaben zu geben.
- Um den Zeitbedarf abschätzen zu können, empfehlen sich insbesondere bei erstmaliger Durchführung von Leistungsnachweisen oder bei Umgestaltung bestehender Formate Testdurchläufe mit freiwilligen Studierenden (z. B. Angehörige des vorangehenden Studienjahres, Tutoren oder studentischen Mitarbeiterinnen). Dies ermöglicht eine realistischere Einschätzung des Zeitbedarfs, als wenn Dozierende als Expertinnen und Experten von ihren Kenntnissen aus schließen.
- Bei geschlossenen und halboffenen Aufgaben, die sich für eine quantitative Auswertung eignen, können mit statistischen Verfahren Eigenschaften wie die Schwierigkeit oder die Trennschärfe der einzelnen Aufgaben berechnet werden. Je nachdem, was die Prüfungsordnung zulässt, können Aufgaben, die sich als zu schwer/einfach oder zu wenig trennscharf erweisen, nachträglich aus der Prüfung gestrichen werden. Oder falls langfristige Fragedatenbanken geführt werden, aus denen jährlich nach dem Zufallsprinzip neue Prüfungen zusammengesetzt werden, so können diese Fragen entfernt oder in einen Pool mit nicht funktionierenden Fragen verschoben werden (mehr zu diesen Fragen in Kap. 13 zum Thema Multiple-Choice-Prüfungen).

Abschließend sei mit Nachdruck gesagt: Falls gewisse Aspekte, die Sie gerne ändern würden, in Prüfungsordnungen festgelegt scheinen, prüfen Sie diese genau und nehmen Sie unter Umständen Rücksprache mit den zuständigen Fachpersonen Ihres Instituts/Ihrer Hochschule etc. Oft erweisen sich Prüfungsordnungen als flexibler als zunächst gedacht. Und falls nicht, müssen deren Vorgaben keinesfalls schicksalsergeben hingenommen werden, sondern können im Rahmen rechtlicher Vorgaben auch geändert werden. In der Schweiz lassen die Vorgaben von Gesetzgebung und die Rechtsprechung dabei mehr Freiheitsgrade zu als in Deutschland.

7.3.2 *Sammel-Leistungsnachweise*

Wie einzelne Aufgaben zu einem Leistungsnachweis kombiniert werden können, so können auch mehrere Leistungsnachweise einen Sammel-Leistungsnachweis bilden.

nachweis bilden. Der Unterschied besteht darin, dass die verschiedenen Beurteilungsanlässe hier voneinander zeitlich und/oder räumlich getrennt sind. Solche Sammelnachweise haben verschiedene Vorteile:

- Es können sehr unterschiedliche Stufen und Dimensionen von Lernzielen beurteilt werden.
- Durch eine zeitliche Staffelung der Beurteilungsanlässe kann die Gefahr reduziert werden, dass Studierende nur für einen abschließenden Leistungsnachweis lernen. Die zeitliche Verteilung der Beurteilungsanlässe führt zu nachhaltigerem Lernen („verteiltes Lernen“, siehe Kap. 4.6.1).
- Die Studierenden können unterschiedliche Stärken zum Tragen bringen beziehungsweise Schwächen kompensieren. Gemäß der Selbstbestimmungstheorie nach Deci u. Ryan (siehe Kap. 4.5.2) wirkt sich das positiv auf die Motivation und somit das Lernen der Studierenden aus.

Zu beachten ist im Gegenzug, dass nicht zu viele kleinteilige Beurteilungsanlässe geschaffen werden, welche die Studierenden unter einen dauernden Beurteilungsdruck setzen. Mittelwege können darin bestehen, einen Teil der Beurteilungsanlässe ausschließlich formativ zu beurteilen oder die einen Bestandteile deutlich geringer zu gewichten, so dass das Hauptgewicht auf einem oder zwei Bestandteilen des Sammelnachweises liegt. In Tabelle 4

Tab. 4: Beispiel für einen Sammel-Leistungsnachweis (leicht abgewandelt aus Sorensen-Unruh 2020, 142–144)

Sammelmodulnachweis in der Vorlesung „Organische Chemie II“			
Leistung	Zeitpunkt	Beurteilung	Gewichtung
5 Einträge im Online-Lerntagebuch (veranstaltungsinterner Blog)	Verteilt über das Semester	Summativ: pass/fail (ausreichende Anstrengung = pass)	15 % ⁹⁵
Übungen und Quizze (teils Einzel- teils Gruppenarbeit)	Verteilt über das Semester	Summativ: pass/fail (ausreichende Anstrengung = pass)	15 %
3 Zwischenprüfungen (schriftliche Klausuren)	Wochen 4, 8 und 11, je 45 Minuten.	Summativ mit Lehrenden- und Selbstbeurteilung (Kap. 9.6.2)	45 %
Schlussklausur	Ca. 1 Woche nach Semesterende	Summativ mit Lehrendenbeurteilung	25 %

95 Studierenden, die mindestens zweimal einer Kommilitonin ausführliches Online-Feedback zu einem Lerntagebuch-Eintrag geben, kann dies zu 5 % an die Übungen und Quizze angerechnet werden (sie können also ein Drittel davon weglassen).

wird eine leicht vereinfachte Fassung des von Sorensen-Unruh (2020, 142–144) entwickelten Sammel-Leistungsnachweises dargestellt. Das Vorgehen bei der summativen Selbstbeurteilung durch die Studierenden wird in Kapitel 9.6.2 näher erläutert. Für die Berechnung der Abschlussnote werden nur die Zwischenprüfungen und die Schlussklausur berücksichtigt, die beiden anderen Leistungen müssen bestanden sein, um zur Schlussklausur zugelassen zu werden.

7.4 Strukturierungsansätze

Es gibt verschiedene Versuche, Leistungsnachweise in verschiedene Formate oder Typen zu strukturieren. Aufgrund der vielen in den Kapiteln 7.1 bis 7.3 beschriebenen Merkmale können diese Versuche bestenfalls als Heuristiken verstanden werden und vernachlässigen jeweils wichtige Aspekte. Dennoch sollen hier zwei bekannte Strukturierungsansätze kurz vorgestellt werden. So können sie als heuristische Entscheidungshilfe herangezogen werden bei der Suche nach einem geeigneten Leistungsnachweisformat – oder bei der Begründung, weshalb bestimmte Formate eben nicht optimal sind und angepasst werden sollten.

7.4.1 *Mündlich – schriftlich – praktisch*

Die Unterscheidung mündlich – schriftlich – praktisch nehmen beispielsweise Frölich-Steffen u. Ouden (2019, 20) und Sacher (2014, 56) vor, siehe Tabelle 5.

Bereits aus der Lektüre von Tabelle 5 wird ersichtlich, dass die Einteilung in die drei Kategorien nicht distinktiv ist: So können Poster entweder als schriftliche Leistungsnachweise oder ihre Präsentation als mündlicher Leistungsnachweis betrachtet werden – am sinnvollsten ist wohl eine Kombination beider Aspekte, womit aber der Leistungsnachweis zugleich mündlich und schriftlich erfolgt. Gleichermaßen können Fallbearbeitungen je nach Aufgabenstellung auch praktisch situiert sein (etwa als Service-Learning im Auftrag externer Partner wie Schulen, Ämter oder Firmen), und bei einem Programmierprojekt mit stark konstruiertem Lebensweltbezug (siehe die Differenzierung in Kap. 7.2.3) kann es sich de facto um einen schriftlichen Leistungsnachweis handeln. Vor allem aber vermischen sich in den meisten hier als praktisch klassifizierten Leistungsnachweisen schriftliche und mündliche Aspekte (siehe Kap. 7.1.3). Aus meiner Sicht ist deshalb der folgende Systematisierungsversuch hilfreicher.

Tab. 5: Strukturierung schriftliche – mündliche – praktische Leistungsnachweise
(in Anlehnung an Frölich-Steffen u. Ouden 2019, 20)

Leistungsnachweisformat		
schriftlich	mündlich	praktisch
Essayklausuren Multiple-/Single-Choice-Klausuren Wissenschaftliche Arbeiten Wissenschaftliche Poster Fallbearbeitung Portfolio	– Referate/Präsentationen – Mündliche Prüfungsgespräche – Disputationen – Posterpräsentationen – Debatten(-beiträge) ⁹⁶	– Experimente und Laboraktivitäten – Prüfungslektionen – Programmierprojekte – OSCE-Prüfungen ⁹⁷ – Feldforschung – Praktika

7.4.2 *Passung von Lernzielstufen, Lehr-Lernaktivitäten und Leistungsnachweisen*

Gemäß der Idee des Constructive Alignment ist die Kongruenz von Lernzielen, Lehr- und Lernaktivitäten sowie der Beurteilungsformate zentral (siehe Kap. 5.1). Die mangelnde Passung des angestrebten Lernzielniveaus mit dem gewählten Beurteilungsformat ist freilich einer der häufigsten Mängel in der Leistungsbeurteilung an Hochschulen, denn nicht jede Aufgabenart ist für jedes Lernziel geeignet. Beispielsweise ist es wenig sinnvoll, eine Lehrveranstaltung, deren Lernziele mehrheitlich auf die SOLO-Taxonomiestufen „relational“ oder „fortgeschrittene Abstraktion“ zielen, mit einer closed-book Multiple-Choice-Klausur „abzuprüfen“ und als Frontalvorlesung zu halten. Ebenso inkongruent wäre es, wenn eine Veranstaltung auf den Erwerb von grundlegenden Begriffen und Konzepten zielt, aber der Leistungsnachweis dazu in einer anspruchsvollen Projektarbeit besteht. Fälle des ersten Typs sind freilich wesentlich häufiger, also dass anspruchsvolle Lernziele bestehen, die verwendeten Leistungsnachweisformate aber deren Erreichen gar nicht zu überprüfen erlauben.

Die Zusammenstellung in Tabelle 6 soll deshalb helfen, die zum Lernziel-Niveau passenden Leistungsnachweisformate und Lehr-Lern-Aktivitäten zu wählen. Zur Erinnerung: Gemäß dem Backward Course Design sollte bei der Planung einer Lehrveranstaltung nach Festlegung der Lernziele zuerst ent-

⁹⁶ Vgl. Gutmiedl-Schumann 2018.

⁹⁷ Objective Structured Clinical Examinations (OSCE) dienen der Überprüfung klinischen Wissens und Könnens in der Medizin (vgl. Dilly 2018; Khan et al. 2013).

Tab. 6: Passung von Lernzielstufen, Lehr-Lern-Aktivitäten und Leistungsnachweisen (in Anlehnung an Feixas u. Zimmermann 2018, 87⁹⁸)

Taxonomiestufe	Lehr-Lern-Aktivität	Leistungsnachweisformat	
SOLO: Fort-geschrittene Abstraktion Bloom: Gestalten und Beurteilen	Etwas Neues schreiben, bauen, entwerfen (z.B. Aufsatz, Forschungsprojekt, Businessplan, Ausführungen, Designs); Komplexe Inhalte wie Produkte, Texte, Aufführungen anhand von Kriterien überprüfen, beurteilen, evaluieren	Komplexe Fallstudie/Projektarbeit, Abschlussarbeit (v.a. Master- oder Doktorarbeit), explorative Laborarbeit Längere schriftliche Arbeit, Portfolio, umfassende Projektdokumentation/-evaluation	Herangehensweise bei Bearbeitung ↑ ↓ Divergentes Denken
SOLO: relational Bloom: Analysieren und Anwenden	Relevante Aspekte eines Falls/Projekts/ Konzepts etc. unterscheiden; bestimmen, wie verschiedene Teile/Aspekte/ Elemente miteinander verbunden sind; ein Verfahren auf mehr oder weniger strukturierte Aufgaben anwenden; die geeignetsten Verfahrensweisen für eine Aufgabenstellung finden	Kurzer Aufsatz, kleines Projekt, einfache Fallstudie, Concept Map, SWOT-Analyse, Lerntagebuch Mündliche Prüfung, schriftliche Online-Diskussion, einfache Fallstudie oder Laboraufgabe	
SOLO: multistruktural Bloom: Verstehen	Zusammenfassen, paraphrasieren, klassifizieren, Elemente vergleichen (Theorien/Texte/ Fälle etc.)	Vortrag/Präsentation, Zusammenfassung eines Themas (z.B. durch offene Essay-Fragen), Gruppendiskussion, Protokoll einer Veranstaltung	
Unistruktural Bloom: Erinnern	Definitionen und Fakten aus den Lernmaterialien abrufen oder erkennen; Auswendiglernen	Wissensüberprüfungen wie SC-/MC-Fragen, Lückentexte, Richtig-falsch-Fragen oder Kurzantwortfragen («wie lautet die Definition von X?»).	

98 Die Darstellung von Feixas und Zimmermann beruht auf Iller u. Wick (2009) sowie Eberly Center (2008/2015).

schieden werden, mit welcher Aktivität bzw. mit welcher Kombination an Aktivitäten das Erreichen der Lernziele überprüft werden soll (siehe Kap. 5.1.2). Zuerst ist also der Leistungsnachweis (ggf. ein Sammelformat) festzulegen, bevor die Lehr-Lern-Aktivitäten geplant werden. Nur so ist letztlich die optimale Passung zwischen Leistungsnachweis und Lehr-Lern-Aktivitäten zu gewährleisten, da die Studierenden ihr Lernverhalten an den durch den Leistungsnachweis und seine Beurteilung de facto gestellten Anforderungen ausrichten.

Auch bei dieser Darstellung ist der heuristische Charakter zu betonen: Die Stufen sind nicht strikt voneinander trennbar. Eine Concept Map kann auch Teil einer komplexen Fallstudie sein, und ein Multiple-Choice-Test kann auch die relationale Lernzielstufe adressieren, wenn die Fragen sehr gut konstruiert sind. Allerdings gibt es Grenzen: Selbst der beste MC-Test ermöglicht keine umfassende Beurteilung der Lernzielstufe „fortgeschrittene Abstraktion“ (vgl. Feixas u. Zimmermann 2018, 87f.).

7.5 Beschreibungsschema für Leistungsnachweisformate

Anhand der im Folgenden erläuterten Beschreibungsstruktur werden in Kapitel 7.6 vier „klassische“ Leistungsnachweisformate und in Kapitel 7.7 drei innovativere Formate vorgestellt. Kapitel 7.8 listet zudem in knapper Darstellung zwölf weitere innovative Formate auf. Dabei ist zu beachten, dass potenziell enorm viele Leistungsnachweisformate denkbar sind, die ihrer „Erfindung“ noch harren.⁹⁹

Das immer gleiche Beschreibungsschema sorgt für eine bessere Vergleichbarkeit der Beispiele. Sie können es auch verwenden, um eigene For-

99 Einer Überlegung von Sacher (2014, 224–227) folgend gibt es allein anhand der Kapitel 7.1 und 7.2 folgende Kombinationsmöglichkeiten: Summativ/formativ x Prozess/Produkt x schriftlich/mündlich/bildlich/körperlich/gegenständlich x kurzzeitig/langfristig x vor Ort/in Distanz x analog/digital x Vereinzelung/Gruppenarbeit x konvergent/divergent x geschlossene/halboffene/offene Aufgaben x kein/konstruierter/authentischer/realer Lebensweltbezug. Das sind $2 \times 2 \times 5 \times 2 \times 2 \times 2 \times 2 \times 3 \times 4 = 7680$ Möglichkeiten. Dabei sind die verschiedenen Unterarten von geschlossenen bis offenen Aufgaben ebenso wenig berücksichtigt wie eine Abstufung inhaltlicher Komplexität, von Kombinationsmöglichkeiten verschiedener Aufgaben innerhalb eines Leistungsnachweises nicht zu reden. „Dabei muss eingeräumt werden, dass natürlich nicht alle Kombinationen gleich sinnvoll sind. In jedem Fall aber dürfte die Zahl der Möglichkeiten in die Millionen gehen“ (Sacher 2014, 227).

mate zu entwickeln oder die wesentlichen Aspekte eines bereits entwickelten Formates zu erfassen. Das leere Beschreibungsschema kann im Internet bei den Materialien zum Buch als Word- und PDF-Dokument heruntergeladen werden (Zugang per QR-Code, s. vorne am Ende des Inhaltsverzeichnisses). Es enthält folgende Elemente:

Allgemeine Beschreibung

Hier werden die Grundeigenschaften des Leistungsnachweises beschrieben: Wird er summativ bewertet oder formativ beurteilt/rückgemeldet (Kap. 7.1.1, Kap. 3.1/3.2)? Inwiefern werden die Dimensionen des Lernprozesses oder der Lernprodukte adressiert (Kap. 7.1.2)? Welche Realisierungsformen werden verwendet (Kap. 7.1.3)? Wie werden Zeit, Raum und Medien eingesetzt (Kap. 7.1.4)? Welche Sozialformen werden verwendet (Kap. 7.1.5)? Inwiefern geht es bei der Bearbeitung um konvergentes und/oder divergentes Denken (Kap. 7.1.6)?

Backwash – bewirktes Lernverhalten der Studierenden

Durch welche Art von Lernverhalten bereiten sich die Studierenden aus ihrer eigenen Perspektive möglichst effizient und effektiv (hohe Erfolgswahrscheinlichkeit bei ökonomischem Zeiteinsatz) auf den Leistungsnachweis vor? Beziehungsweise bei zeitlich überdauernden Aufgabenstellungen wie Projekten, wissenschaftlichen Arbeiten, Portfolios etc.: Welche Art von Bearbeitung ermöglicht ein effizientes und effektives Abschließen der Aufgabe? Mit anderen Worten: Welches Lernverhalten löst der Leistungsnachweis realistischerweise aus (Kap. 7.4.2, das Phänomen des Backwash wird in Kap. 5.1.2 näher erläutert)?

Lebensweltbezug

Welchen Lebensweltbezug schafft der Leistungsnachweis (Kap. 7.2.3)? Passt dies zum Anforderungsniveau, das die Lernziele vorgeben (Kap. 5.4)?

Geeignet für welche Art der Leistungsrückmeldung

Welche Bezugsnormen kommen zur Anwendung (Kap. 6.2)?

Formative Beurteilung: Welche Leistungsrückmeldung können Sie oder Peers zum Lernprodukt oder Lernprozess geben, und welche Art von Selbsteinschätzung der Studierenden ist möglich (siehe Kap. 8)?

Summative Beurteilung: Welche Art von Bewertung ist vorgesehen (Noten, Prädikat)? Anhand welcher Kriterien wird das Erreichen der Lernziele

beurteilt? Wie stellen Sie die Güte der Beurteilung sicher (siehe Kap. 9 zur Bewertung und Kap. 6.3/6.4 zur Güte der Beurteilung)?

Beachten Sie, dass formative Rückmeldungen nur einen Sinn ergeben, wenn die kommentierte Leistung (noch) nicht summativ bewertet wurde (siehe Kap. 9.1: Sobald Noten erteilt werden, beachten Studierende inhaltliche Rückmeldungen nicht mehr).

Wer hat's getan?

In dieser etwas flapsig überschriebenen Rubrik¹⁰⁰ wird beschrieben, mit welcher Sicherheit das Leistungsnachweisformat Rückschlüsse auf die Urheberchaft der zu beurteilenden Leistung erlaubt, was vor allem bei summativen Beurteilungen relevant ist. Hier können auch vorgesehene Kontrollmaßnahmen beschrieben oder Gründe angeführt werden, weshalb auf sie verzichtet wird.

Ökonomie

Hier erfolgen Überlegungen dazu, wie sich der Zeitaufwand für Lehrende wie auch Studierende gestaltet: Wofür muss wie viel Aufwand betrieben werden?

Größte Vorteile/größte Nachteile

In diesen beiden Rubriken werden die größten Vorteile oder Stärken bzw. die größten Nachteile oder Schwächen des Formates angesprochen. Hier in diesem Buch müssen diese Beschreibungen generisch bleiben, bei der Beschreibung eines Leistungsnachweises für ein ganz spezifisches Setting können an dieser Stelle auch kontextbedingte/situative Vor- und Nachteile aufgeführt werden.

7.6 Beispiele für „klassische“ Formate

Die folgenden Beschreibungen unterscheiden sich bezüglich ihres Konkretisierungsgrades, da sie in den Rubriken des Beschreibungsschemas jeweils unterschiedliche Freiheitsgrade aufweisen. Um die Beschreibungen kurz und handlich zu halten, werden nicht alle Gestaltungsvarianten aufgeführt. Die beschriebenen Formate können und sollen je nach den kontextspezifischen Anforderungen konkretisiert und weiterentwickelt werden.

100 Der Titel entspricht der aus dem Vokabular der Kriminalliteratur stammenden Bezeichnung „whodunit“, die Race (2014) für diesen Aspekt verwendet hat.

7.6.1 *Handschriftliche Klausur*

Allgemeine Beschreibung: Eine schriftliche Klausur (von Lateinisch *clausura*, Einschließung) findet in einem klar definierten Zeitraum statt, der 4 Stunden in der Regel nicht überschreitet. Sie dient fast immer der summativen Beurteilung und ist produktbezogen. Obschon inzwischen fast alle Studierenden über elektronische Geräte verfügen (in der Schweiz wird der Besitz eines Laptops an den meisten Hochschulen inzwischen sogar reglementarisch vorausgesetzt), werden aufgrund verschiedener technischer Restriktionen und von Sicherheitsbedenken die meisten schriftlichen Klausuren immer noch vor Ort und von Hand durchgeführt. Traditionell zielen Klausuren auf eine vereinzelnde Bearbeitung, wobei inzwischen mit der zweistufigen Prüfung eine Variante entwickelt wurde, die auch den Vorteil sozialer Bearbeitung nutzt (dieses Format wird in Kap. 7.7.2 beschrieben). Oft stehen konvergente Lösungen im Vordergrund, was angesichts des meist herrschenden Zeitdrucks grundsätzlich sinnvoll scheint.

Zu unterscheiden ist zwischen Klausuren, in denen (alle oder klar definierte) Hilfsmittel erlaubt sind und solchen, in denen keinerlei Hilfsmittel erlaubt sind. Aufgabenformate sind theoretisch viele möglich, wobei eine Tendenz zu geschlossenen und halboffenen Fragen besteht. Werden offene Fragen verwendet, werden zu ihnen meist eher kurze Antworten erwartet.

Backwash – bewirktes Lernverhalten der Studierenden

Klausuren, in denen keine nennenswerten Hilfsmittel erlaubt sind, zielen durch die meist sehr engen Aufgabenstellungen meist auf das Abrufen von Wissensinhalten. Sie übertreffen deshalb selten die beiden unteren Lernzielstufen der SOLO-Taxonomie (falls doch, dann am ehesten durch Anwendungsaufgaben im Sinne der Bloom-Taxonomie) – auch wenn die Prüfenden oder Curriculumverantwortlichen gelegentlich Hoffnungen auf höhere Ansprüche hegen.

Klausuren führen deshalb häufig zu einem Lernverhalten, das sich stark auf das Auswendiglernen von Inhalten fokussiert: Es werden Definitionen, Listen, Tabellen etc. so gut als möglich memorisiert. Dies wiederum führt zu einem oberflächlichen Lernen, das einem nachhaltigen Aufbau von Expertise nicht dienlich ist (siehe Kap. 4.2 und 4.6). Soll ein tiefenorientiertes Lernen bewirkt werden, müssen im gleichen Themengebiet parallel auch andere Formate der Leistungsüberprüfung eingesetzt werden. Eine Ausnahme stellen Klausuren mit bewusst komplexen Aufgabenstellungen wie einer Fallbearbeitung dar – müssen die Studierenden darauf vorbereitet sein, einen relativ komplexen Fall in verschiedener Hinsicht zu bearbeiten, löst die Klausur mit deutlich größerer Wahrscheinlichkeit tiefenorientiertes Lernen aus, da dieses für den Prüfungserfolg nötig ist.

Lebensweltbezug

In Klausuren sind alle Abstufungen von Lebensweltbezug möglich. In der angestammten Praxis besteht gewiss eine Tendenz zu Aufgaben ohne oder mit lediglich konstruiertem Lebensweltbezug.

Geeignet für welche Art der Leistungsrückmeldung

In Klausuren kommt – zumindest offiziell – in der Regel die Kriterialnorm zur Anwendung. Entsprechend sollten die Bewertungskriterien vor der Prüfungsdurchführung erarbeitet werden, ausformuliert vorliegen und idealerweise den Studierenden vorher bekannt gemacht worden sein. Oft spielt allerdings – bewusst oder unbewusst – auch die Kollektivnorm eine Rolle, was problematisch ist (siehe Kap. 5.1.3 und Kap. 9.2).

Da in aller Regel summativ beurteilt wird, empfiehlt sich eine knappe Form der Leistungsrückmeldung in Form der Bekanntgabe einer Note oder eines Prädikates (ausführlichere Rückmeldungen haben zu diesem Zeitpunkt keinen Einfluss mehr auf das Lernen der Studierenden, siehe Kap. 9.1). Sinnvoll ist, interessierten Studierenden eine Nachbesprechung und/oder die Einsicht in die Bewertung ihrer Prüfung anzubieten. Bei hohen Studierendenzahlen kann man dieses Angebot aus Ressourcengründen auf Studierende mit einer ungenügenden Bewertung beschränken.

Wer hat's getan?

Klausuren finden unter Bedingungen der Überwachung statt, so dass in der Regel mit hoher Sicherheit gesagt werden kann, dass die Studierenden ihre Klausur allein ausgefüllt haben. Typische Überwachungsbedingungen sind: Aufsichtspersonen in den Prüfungsräumen, Verbot/Einzug von mobilen Kommunikationsgeräten, 2–3 Serien mit unterschiedlicher Aufgabenreihenfolge, sofern allen Prüflingen die gleichen Fragen gestellt werden. Die hohe Gewissheit über die Urheberschaft wird erkaufte mit einem Klima der Kontrolle, was motivational ungünstige Auswirkungen hat, da Kontrolle das Bedürfnis nach Selbstbestimmung untergräbt (siehe Kap. 4.5.2). Klausuren gehören zusammen mit mündlichen Prüfungen wohl auch zu den Formaten, die am meisten Prüfungsangst (siehe Kap. 11) auslösen.

Ökonomie

Für die Lehrenden kann das Format durch seine hohe Skalierbarkeit bei hohen Studierendenzahlen relativ ökonomisch sein. Dies gilt insbesondere, wenn die Auswertung durch geschlossene Fragen (oder künftig künstliche Intelligenz) automatisiert werden kann, wie das seit Längerem bei Multiple-

Choice-Klausuren der Fall ist (siehe Kap. 13). Allerdings können Kontrollmaßnahmen wie Aufsichtspersonen, das Erstellen verschiedener Serien, das Buchen mehrerer Prüfungsräume zur gleichen Zeit etc. erheblichen Zusatzaufwand verursachen.

Aus Sicht der Lernenden ist das Format wenig ökonomisch: Mit viel Aufwand müssen sie Inhalte auswendig lernen, die ohne zusätzliche Anwendung/Vertiefung größtenteils nicht Eingang in ihre Expertise finden werden. Das Format ist somit auch bildungsökonomisch zu hinterfragen – es ist zwar auf Seiten der Institution relativ kostengünstig, hat aber wenig Lernwirkung und viele unerwünschte motivationale und emotionale Nebeneffekte. Um die Anwendung des Formats zurückzudrängen, müssten ggf. durch die Politik bis hin zur Gesetzgebung entsprechende Anreize geschaffen werden.

Größte Vorteile

- sehr gut skalierbar, bei hohen Studierendenzahlen relativ kostengünstig
- erreicht unter dem Gesichtspunkt quantitativ orientierter Gütekriterien oft gute Objektivitäts- und Reliabilitätswerte
- prüft, ob Studierenden unter Zeitdruck wesentliche Informationen abrufen können

Größte Nachteile

- erhebliche unerwünschte Wirkungen, einerseits auf die Lernmotivation (Reduktion der intrinsischen Motivation), andererseits durch Begünstigung von Prüfungsängsten
- erreicht unter dem Gesichtspunkt der Gütekriterien (quantitativ wie qualitativ) oft keine gute Validität – prüft also nicht in der Tiefe das, was die Studierenden lernen sollen
- erzeugt Auswendiglernen durch die Studierenden (eine wenig nachhaltige Form des Lernens)

7.6.2 Mündliche Prüfung

Allgemeine Beschreibung: Eine mündliche Prüfung findet in direkter Interaktion zwischen Prüfenden und Studierenden statt, in aller Regel vor Ort, wobei heutzutage auch Videoformate denkbar sind und während der Coronapandemie auch vielerorts durchgeführt wurden. Ihre zeitliche Dauer ist im Vergleich zu schriftlichen Prüfungsformen meist relativ kurz und überschreitet selten eine halbe Stunde. Manchmal wird zusätzlich noch eine (ebenfalls

meist kurze) Vorbereitungszeit vor Ort zum Studium der Fragen oder Materialien zur Verfügung gestellt. Mündliche Prüfungen dienen traditionell fast immer der summativen Beurteilung (obschon sich mündliche Situationen auch gut für formatives Feedback eignen würden). Im Vergleich zu schriftlichen Formen erlaubt die direkte Interaktion Rückfragen, was es ermöglicht, den Denkprozessen der Prüflinge näher auf den Grund zu gehen. So können in der Tendenz höherrangige Lernziele und mehr Prozessaspekte des Lernens erfasst werden als bei schriftlichen Klausuren (aber immer noch weniger als bei manch anderem Format).

Da mündliche Äußerungen flüchtig sind, werden solche Prüfungen in der Regel von mindestens zwei Prüfenden durchgeführt, von denen oft eine Person eine Beisitzfunktion hat und ein Protokoll¹⁰¹ führt. In der Regel zielen mündliche Prüfungen eher auf vereinzelnde Leistungen, es sind aber auch Gruppenformate möglich oder Formate, in denen mehrere anwesende Prüflinge seriell geprüft werden.¹⁰² Häufig stehen, auch aufgrund der knappen Zeit, konvergente Lösungen im Vordergrund, jedoch können insbesondere durch Rückfragen der Prüfenden auch divergente Denkleistungen angeregt werden.

Backwash – bewirktes Lernverhalten der Studierenden

In der Vorbereitung auf mündliche Prüfungen konzentrieren sich Studierende in der Regel auf eine umfassende Kenntnis des Prüfungsgebiets. Studierende,

101 Die Aufzeichnung von Prüfungsgesprächen ist heute technisch problemlos möglich, aber aus rechtlicher Sicht problematisch (z. B. hinsichtlich Einhaltung von Datenschutz und Prüfungsgeheimnis sowie Wahrung des Rechts auf informationelle Selbstbestimmung). Zudem dürften Aufzeichnungen Prüflinge wie Prüfende zusätzlich unter Druck setzen, was einer angenehmen Prüfungsatmosphäre (die gemäß Hirdina (2019, 41) in Deutschland ebenfalls rechtlich verankert ist) nicht förderlich ist (vgl. Koller 2002, 110).

102 In den Rechtswissenschaften scheint es in Deutschland sogar mündliche Prüfungen zu geben, in denen mehrere Prüflinge in direkten Wettbewerb zueinander treten, inkl. Beurteilung anhand der Kollektivnorm (vgl. Neumann 2019). Diese Praxis in – ausgerechnet rechtswissenschaftlichen! – Prüfungen an Hochschulen steht aus meiner Sicht im Widerspruch zu den Vorgaben des Schulrechts, wie sie Kostorz (2016) beschreibt. Auch wenn Schulrecht und Hochschulrecht zwei verschiedene juristische Bereiche sind: Weshalb an Hochschulen andere Beurteilungsnormen angemessen sein sollen als in der vorangehenden Schulbildung – und dies entgegen zentralen Erkenntnissen aus Pädagogik und Psychologie (siehe bes. Kap. 2, 4.5 und 6) –, scheint mir erklärungsbedürftig. Generell wäre es aus meiner Sicht wünschenswert, dass insbesondere die Rechtspraxis bei ihren Urteilen pädagogische und lernpsychologische Erkenntnisse stärker berücksichtigt.

die sich der Anforderungen dieses Formats bewusst sind, legen auch Wert auf die Fähigkeit, das Gelernte mündlich ausdrücken und für Argumentationen verwenden zu können. Im Vergleich zu schriftlichen Klausuren erfordert dies ein tieferes Verständnis der Materie, da die Studierenden in der Lage sein müssen, komplexe Zusammenhänge zu erklären, Fragen zu beantworten, ihre Argumentation zu verteidigen und dabei verschiedene Perspektiven zu berücksichtigen. Ist der Umfang des geprüften Themengebiets und die Art der Aufgabenstellung, die im Rahmen der mündlichen Prüfung zu erwarten ist, nicht hinreichend klar (was leider nicht selten der Fall ist), kann dieses Format Studierende auch überfordern und bewirken, dass sie sich entweder kaum oder nur sehr unspezifisch auf die Prüfung vorbereiten. Es ist deshalb hier besonders wichtig, die Erwartungen an das Wissen und Können sowie die Prüfungsbedingungen frühzeitig transparent zu kommunizieren.

Lebensweltbezug

In mündlichen Prüfungen sind alle Abstufungen von Lebensweltbezug möglich. Im Vergleich zu schriftlichen Klausuren es hier in der herrschenden Praxis gebräuchlicher, Bezüge zur Lebenswelt der Studierenden herzustellen – nicht nur in der Aufgabenstellung, sondern auch durch Rückfragen oder in gesprächsartigen Prüfungen. Einen besonderen Fall stellen Prüfungsgespräche etwa im Anschluss an praktische Prüfungsformate wie Prüfungslektionen, Laborprüfungen, Projektpräsentationen sowie Disputationen dar – diese Situationen sorgen für einen zumindest authentischen, wenn nicht sogar realen Lebensweltbezug.

Geeignet für welche Art der Leistungsrückmeldung

In mündlichen Prüfungen kommt offiziell in der Regel die Kriterialnorm zur Anwendung. Entsprechend sollten die Bewertungskriterien vor der Prüfungsdurchführung erarbeitet werden, ausformuliert vorliegen und idealerweise den Studierenden vorab bekannt gemacht worden sein. Oft spielt allerdings – bewusst oder unbewusst – auch die Kollektivnorm eine Rolle, was problematisch und möglichst zu verhindern ist. Falls die Prüfenden die Prüflinge schon früher betreut oder begleitet haben, ist – zu formativen Zwecken – auch das Hinzuziehen der Individualnorm möglich.

Da in aller Regel summativ beurteilt wird, empfiehlt sich eine knappe Form der Leistungsrückmeldung in Form der Bekanntgabe einer Note oder eines Prädikates. Wird die Bewertung direkt am Ende der Prüfung bekanntgegeben, ist eine kurze Begründung möglich oder man kann die Studierenden fragen, ob sie Fragen zur Beurteilung haben. Noch besser ist es, die Studierenden vor Bekanntgabe der Bewertung um eine Selbsteinschätzung zu bitten. Viele Prüfende schrecken vor solchen Selbsteinschätzungen zurück, aber

die meisten Studierenden können ihre Leistung erstaunlich gut einschätzen. Die Mehrheit schätzt ihre Leistung eher tiefer ein als die Prüfenden, wodurch selten problematische Situationen entstehen.

Darf die Prüfungsbewertung aufgrund der Prüfungsordnung erst später bekanntgegeben werden, so empfiehlt es sich, zumindest Studierenden mit einer ungenügenden Bewertung die Einsicht in das Prüfungsprotokoll und/oder eine Nachbesprechung anzubieten.

Unabhängig davon, ob formativ rückgemeldet oder summativ bewertet wird, ist zu beachten, dass die Unmittelbarkeit der Kommunikationssituation mündliche Prüfungen stärker anfällig für Verzerrungen macht (siehe Kap. 6.4 zur Vielzahl von Verzerrungseffekten). Kapitel 8.5.2 zu mündlichem Feedback enthält Hinweise zum Umgang mit dieser Herausforderung.

Wer hat's getan?

Aufgrund der direkten Interaktion zwischen Prüfenden und Prüflingen ist bei mündlichen Prüfungen unmittelbar erkennbar, welche Studierenden welche Beiträge leisten. Wenn die Prüfenden die Prüflinge nicht per definitionem persönlich kennen (wie bei Disputationen), sind Maßnahmen zur eindeutigen Identifikation der Studierenden zu treffen (Vorlage eines Ausweisdokuments). Zudem hilft die Anwesenheit einer zweiten Prüfungsperson/eines Beisitzers sicherzustellen, dass die Prüfenden mit den Prüflingen nach dem Gebot der Gleichbehandlung (vgl. Hirdina 2019, 41) umgehen.

Ökonomie

Für Lehrende ist der Aufwand bei diesem Format ziemlich proportional zur Studierendenzahl. Der Zeitaufwand für die Planung und Durchführung ist deshalb ab einer bestimmten Studierendenzahl größer als z. B. bei schriftlichen Klausuren.

Für die Studierenden hängt insbesondere von der Transparenz der gestellten Anforderungen ab, ob eine Vorbereitung möglich ist, bei der Aufwand und Ertrag in einem guten Verhältnis stehen. Da gut durchgeführte mündliche Prüfungen eher tiefenorientierte Lernstrategien begünstigen, bewirken sie tendenziell nachhaltigeres Lernen als schriftliche Klausuren. Zu beachten ist allerdings, dass mündliche Prüfungen im Vergleich zu anderen Leistungsnachweisformaten wie wissenschaftlichen Arbeiten (Kap. 7.6.4) und vielen der neueren Formate (Kap. 7.7) mehr Prüfungsängste auslösen, was dem Lerngewinn wiederum abträglich ist (siehe Kap. 11 zum Thema Prüfungsangst).

Größte Vorteile

- ermöglicht direkte Interaktion mit Studierenden und somit eine Vertiefung ihrer Antworten durch Rückfragen, Folgefragen u.Ä.
- bewirkt bei sinnvoller Gestaltung im Vergleich zum anderen zentralen „klassischen“ Prüfungsformat, den schriftlichen Klausuren, tiefenorientierteres Lernen
- verursacht kaum Fragen nach Betrug oder Urheberschaft

Größte Nachteile

- jede mündliche Prüfung verläuft anders, was zwar die Validität erhöht, aber die Vergleich- und Generalisierbarkeit beeinträchtigt
- bei hohen Studierendenzahlen sehr aufwändig für die Lehrenden (kann auch die Qualität beeinträchtigen)
- hoher Leistungsdruck in kurzer Zeit und unvorhersehbarer Ablauf lösen bei vielen Studierenden Prüfungsängste aus

7.6.3 Vortrag/Referat

Allgemeine Beschreibung: Ein Vortrag oder Referat findet in direkter Interaktion unter Beisein eines Publikums statt, bei dem es sich in der Regel um Studierende der gleichen Lehrveranstaltung handelt. Denkbar sind auch externe Expertinnen oder eine breitere Öffentlichkeit als Publikum. Studierende stellen in diesem Format ihre Arbeitsergebnisse oder ein durch sie erarbeitetes Thema vor. Vorträge können sowohl formativ als auch summativ beurteilt werden. Der Vortrag hat Produktcharakter, in seinem Rahmen können die Studierenden aber auch den Lernprozess thematisieren, den sie bei seiner Vorbereitung oder der Erarbeitung der präsentierten Resultate durchlaufen haben. Vorträge sind sowohl in Einzel- als auch in Partner- und Gruppenform gebräuchlich. Das Verhältnis von konvergentem zu divergentem Denken hängt von der Aufgabenstellung ab.

Auch bei den medialen, zeitlichen und räumlichen Realisierungsformen gibt es zahlreiche Optionen, von denen wohl heute die Bildschirmpräsentation unter Gebrauch eines Videoprojektors (Beamers) die häufigste, aber nicht immer sinnvollste ist. Bezüglich Medialität sind Posterpräsentationen (siehe Kap. 7.7.3), praktische Demonstrationen von Vorgehensweisen oder die Verwendung von Requisiten oder Hilfsmitteln wie Modellen gute Alternativen. Auch die zeitliche Strukturierung kann konkreter sein als „30 Minuten Vortrag plus 15 Minuten Diskussion“, etwa beim Elevator Pitch oder

beim Pecha Kucha (20 Bilder/Folien werden je 20 Sekunden angezeigt, Präsentation dauert 400 Sekunden; vgl. Zimmermann, Germs 2018). Möglich sind auch Online-Vorträge und Realisierungen als Videos, die dann asynchron von Mitstudierenden und Lehrenden betrachtet und diskutiert werden können. Das Format erlaubt die Adressierung einer Vielzahl von Lernzielen, wobei leider in der Hochschulpraxis das Potenzial oft nicht ausgeschöpft wird (siehe Backwash).

Backwash - bewirktes Lernverhalten der Studierenden

Der Backwash hängt wesentlich von der Aufgabenstellung und den kommunizierten Erwartungen ab. Nicht spezifizierte Erwartungen und reine Themenvorgaben wie „Malerei im Expressionismus“ bewirken oft eintönige, inhaltswiedergebende Vorträge, bei deren Erarbeitung die Studierenden oberflächensorientiert Inhalte kompilieren und die auch bei der Zuhörerschaft keine nachhaltigen Lernprozesse auslösen. Die oben aufgeführten Gestaltungsmöglichkeiten sowie das transparente Bekanntgeben, Besprechen oder gar gemeinsame Definieren von Qualitätskriterien interessanter und lehrreicher Vorträge regt die Studierenden zu stärker tiefenorientierten Lernstrategien an. Dabei können die Mitstudierenden als Feedbackgebende aktiv einbezogen werden. Dieser Prozess erfolgt am besten strukturiert, indem z. B. alle Studierenden ein- oder zweimal im Semester einer anderen Gruppe Peer-Feedback erteilen.

Zu beachten ist, dass auch Vortragssituationen Ängste auslösen: Das Auftreten vor Publikum kann für Studierende belastend sein – wobei es mit Blick auf spätere Tätigkeitsfelder je nach Studienrichtung sinnvollerweise geübt werden sollte (siehe Lebensweltbezug).

Lebensweltbezug

In Vorträgen sind alle Abstufungen von Lebensweltbezug möglich. Es hängt sowohl von der Aufgabenstellung, der Kontextualisierung wie auch von der Eigeninitiative der Studierenden ab, in welchem Ausmaß sie im Vortrag Bezüge zu ihrer Lebenswelt herstellen können. Werden im Vortrag die Resultate und Erfahrungen aus einer umfassenden Projektarbeit oder Fallstudie erläutert, sind andere Lebensweltbezüge möglich, als wenn ein theoretisches Thema einführend erläutert werden soll (letzteres erfolgt in aller Regel effizienter durch einen Lektüreauftrag). Aber auch Formate wie ein Elevator Pitch erzeugen Lebensweltbezug (vgl. Saunders 2022; Morgan u. Wright 2021). An einen Vortrag anschließende Diskussionen bieten eine weitere Möglichkeit für das Herstellen von Lebensweltbezug.

Geeignet für welche Art der Leistungsrückmeldung

Bei der Beurteilung von Vorträgen kommt in der Regel die Kriteriennorm zur Anwendung. Die Kollektivnorm ist generell zu vermeiden, und die Anwendung einer Individualnorm ist in den meisten Situationen schwierig, da weder Lehrende noch Mitstudierende systematische Kenntnisse früherer Vortragsleistungen haben dürften.

Vorträge können sowohl formativ als auch summativ beurteilt werden, was beides in der Hochschulpraxis gebräuchlich ist. Bei formativen Beurteilungen lohnt es sich, die Peers einzubeziehen. Ihr Feedback zu Verständlichkeit, Argumentation, visueller Gestaltung oder Handouts ist für die Vortragenden oft sogar hilfreicher, da sie auf einem vergleichbaren Wissens- und Könnensstand sind. Zudem können die Vortragenden auch eine Selbsteinschätzung ihrer eigenen Leistung anhand der gleichen Kriterien vornehmen. Diese Selbsteinschätzung erfolgt idealerweise, bevor sie das Feedback der Lehrenden und/oder Kommilitoninnen erhalten.

Bei summativen Bewertungen gelten die gleichen Empfehlungen wie bei den vorangehenden Formaten.

Wer hat's getan?

Die Feststellung der Urheberschaft ist bei Vorträgen zwar ähnlich schwierig wie bei schriftlichen Arbeiten. Da der Vortrag aber vor Ort durch die verantwortliche Person oder Personengruppe gehalten wird und diese auch Fragen von Mitstudierenden oder der Lehrenden beantworten müssen, dürfte es sich de facto um ein relativ unproblematisches Format halten. Verstärkt werden kann dies, indem die Diskussion der Vortragsinhalte angekündigter Bestandteil des Leistungsnachweises ist. Das Diskussionsverhalten der Vortragenden kann zudem sowohl formativ wie summativ beurteilt werden (wozu natürlich Kriterien vorliegen sollten).

Ökonomie

Vorträge können für Lehrende aus verschiedenen Gründen ein ökonomisches Format sein. Sie können sich durch den Einbezug von Peer-Feedback bei der formativen Leistungsrückmeldung entlasten und zugleich die Lernwirksamkeit ihrer Veranstaltung erhöhen. Kritischer einzuschätzen ist der Umstand, dass sich Lehrende durch das Vortragsformat in der Vorbereitung und Durchführung der Lehre entlasten können – dies geht schnell auf Kosten der Lernwirksamkeit: Werden keine klaren Qualitätsansprüche an die Vorträge formuliert und spielen die Mitstudierenden als Zuhörende keine definierte Rolle, etwa als Feedbackgebende oder als Diskussionsteilnehmende, droht die beim Backwash beschriebene Serie eintöniger Vorträge, deren Lernwirksamkeit

äußerst gering ist. Bei sehr hohen Studierendenzahlen eignen sich Vorträge als Format zudem nicht mehr, weil die Studierenden nur noch Kleinstbeiträge zu Gruppenvorträgen leisten könnten, was weder eine sinnvolle formative Leistungsrückmeldung noch eine gültige summative Bewertung ermöglicht.

Für die Studierenden hängt es von der sinnvollen didaktischen Gestaltung und Einbettung der Vorträge ab, ob ihr Aufwand in einem guten Verhältnis zum Ertrag steht. Je klarer und dem Lernkontext angemessener die Qualitätsansprüche¹⁰³ an die Vorträge und je interaktiver das Verhältnis von vortragenden und zuhörenden Studierenden gestaltet wird, desto tiefenorientierter gestalten sich die Erarbeitung der Vorträge und ihre Diskussion im Plenum – und desto mehr lernen die Studierenden im Verhältnis zu ihrem Zeitaufwand.

Größte Vorteile

- ermöglicht fachlichen Austausch zwischen den Studierenden, was tiefenorientiertes Lernen begünstigt
- ist für die Dozierenden ein ökonomisches Format
- ermöglicht diverse Möglichkeiten für formative Leistungsrückmeldungen, auch durch Peers

Größte Nachteile

- bei mangelnder didaktischer Gestaltung langweilige Vorträge mit geringer Lernwirksamkeit für Vortragende wie Zuhörende
- nur für kleinere und mittelgroße Veranstaltungen geeignet
- wird in vielen Fachkulturen so häufig verwendet, dass das Format die Studierenden nicht mehr zu motivieren vermag

103 Die Qualitätskriterien für studentische Vorträge können je nach Lernkontext stark variieren. So dürften in einem forschungsorientierten Vortrag die wissenschaftliche Genauigkeit oder die Fähigkeit zur kritischen Analyse von Daten im Vordergrund stehen. Bei der Präsentation von Geschäftsideen dürften hingegen Kriterien wie das Identifizieren geschäftlicher Herausforderungen, die Entwicklung von Lösungsansätzen oder die Überzeugungskraft gegenüber potenziellen Investoren im Vordergrund stehen. Um die je spezifischen Ansprüche transparent zu machen, eignen sich Beurteilungsraster (siehe Kap. 9.3).

7.6.4 Wissenschaftliche (schriftliche) Arbeit

Allgemeine Beschreibung: Eine wissenschaftliche Arbeit ist eine schriftliche Leistung, die sich an typischen Publikationsformen der jeweiligen Fachkultur orientiert (v. a. wissenschaftliche Artikel/Aufsätze, bei Abschlussarbeiten auch Buchpublikationen). In aller Regel wird den Studierenden für die Erarbeitung ein längerer Zeitraum zur Verfügung gestellt, der mehrere Wochen oder Monate umfassen kann.¹⁰⁴ Wissenschaftliche Arbeiten werden gewöhnlich summativ beurteilt, wobei oft die Gelegenheit für formative Leistungsrückmeldungen verpasst wird. Sie sind produktorientiert. Aspekte des Lernprozesses können implizit als Erkenntnisgewinn im Rahmen des in der Arbeit geschilderten Forschungsprozesses angesprochen werden.

Meist werden wissenschaftliche Arbeiten einzeln verfasst, sie können aber auch als Partner- oder Gruppenarbeiten erfolgen, wie es in der akademischen und sonstigen Berufswelt häufig der Fall ist. Bei der Erarbeitung können eine Vielzahl an Medien wie gedruckte und elektronische wissenschaftliche Publikationen, aber auch Laborarbeit oder sozialwissenschaftliche Erhebungen zum Einsatz kommen – oft werden verschiedene Elemente kombiniert. Je nach Fach, Thema oder Zielsetzung der Lehrenden sind die Aufgabenstellungen mehr oder weniger offen; in den Geistes- und Sozialwissenschaften werden sie oft innerhalb eines gesetzten Rahmens von Studierenden vorgeschlagen. Je größer die Offenheit der Fragestellung ist, desto mehr Raum für divergentes Denken bietet eine wissenschaftliche Arbeit.

Backwash - bewirktes Lernverhalten der Studierenden

Die offizielle Zielsetzung von wissenschaftlichen Arbeiten lautet meist, mit ihnen werde „die Fähigkeit zur eigenständigen wissenschaftlichen Arbeit nachgewiesen“ (Gießmann 2019, 121). Diese Zielsetzung ist äußerst anspruchsvoll, aber die hohen Ansprüche bewirken in der Regel, dass die Studierenden durch das Verfassen von wissenschaftlichen Arbeiten tatsächlich tiefgreifende Lernprozesse durchlaufen. Denn das Format erfordert komplexe Tätigkeiten wie eine eigenständige Literaturrecherche, die Analyse von wissenschaftlicher Literatur und/oder eigenen Forschungsdaten, die kritische Bewertung der eigenen Ergebnisse sowie eine verständliche schriftliche Darstellung. Besonders an die Schreibfähigkeiten stellt das Format hohe Ansprüche, die im Idealfall auch einen erheblichen Lernzuwachs zur Folge haben. Studierende, die aktiv den Austausch mit Peers und/oder Expertinnen über

104 Es sind aber auch kürzere Zeiträume möglich, so gab es früher an der Universität Zürich das Format der dreitägigen Hausarbeit im Rahmen der Abschlussprüfungen zum Lizentiat (heute Masterstudium); vgl. auch das Format der 24-Stunden-Hausarbeit (Reinmann 2018).

ihren Forschungs- und Schreibprozess suchen, haben den größten Lerngewinn.

Wenn sie mit mangelnder Vorbereitung (Propädeutik) und fehlender Begleitung der Studierenden einhergehen, können die hohen Ansprüche die Studierenden allerdings auch überfordern. Während sich die einen dann sinnvollerweise Unterstützung bei fortgeschritteneren Peers oder institutionellen Angeboten wie einem Schreibzentrum holen, können andere Grenzüberschreitungen wie Plagiate oder die Nutzung von Ghostwriting begehen (siehe Abschnitt „Wer hat’s getan?“).

Lebensweltbezug

In wissenschaftlichen Arbeiten sind alle Abstufungen von Lebensweltbezug möglich. Es hängt sowohl von der Aufgabenstellung, der Kontextualisierung wie auch von der Eigeninitiative der Studierenden ab, in welchem Ausmaß sie in der schriftlichen wissenschaftlichen Arbeit Bezüge zu ihrer Lebenswelt herstellen können. Bei Arbeiten mit eigenem Forschungsanteil ist in der Regel mehr Lebensweltbezug möglich als bei reinen Literaturarbeiten oder theoretischen Auseinandersetzungen. Mehr Mitsprache bei der Gestaltung der Fragestellung eröffnet den Studierenden größere Möglichkeiten, die Arbeit auf ihre Lebenswelt zu beziehen,

Geeignet für welche Art der Leistungsrückmeldung

Bei der Beurteilung wissenschaftlicher Arbeiten sollte die Kriteriálnorm zur Anwendung kommen. Die Kollektivnorm ist generell zu vermeiden. Die Individualnorm kann formativ zur Anwendung kommen, wenn das Verfassen der Arbeit gut begleitet wird.

In aller Regel werden wissenschaftliche Arbeiten summativ bewertet, meist durch Noten. Aufgrund der hohen Validität des Formats ist dies zwar sinnvoll. Leider wird aber oft die Gelegenheit verpasst, den Studierenden während ihres Forschungs-, Analyse- und Schreibprozesses formatives Feedback zu ermöglichen. Formative Rückmeldungen während des Erarbeitungsprozesses bieten allerdings ein enormes Lernpotenzial: Studierende erhalten dadurch Hinweise zum Forschungs- und Schreibprozess, die sie ganz konkret im Rahmen der jeweiligen wissenschaftlichen Arbeit umsetzen können, die ihnen oft aber auch weit über diesen Rahmen hinaus nützlich sind.

Empfehlenswert ist deshalb folgendes Vorgehen (das Ähnlichkeiten hat mit der in Kap. 7.7.1 beschriebenen mehrphasigen Beurteilung eines Endprodukts; für weitere Hinweise zur Begleitung und Bewertung von längeren schriftlichen Arbeiten siehe auch Kap. 12):

1. Die Lehrenden sollten während des Erarbeitungsprozesses mindestens eine formative Leistungsrückmeldung organisieren. Diese können sie selbst oder ihre Mitarbeitenden geben oder es können Verfahren des Peer-Feedback/Peer-Review organisiert werden. Die beiden Möglichkeiten können auch kombiniert werden, z. B. kann zum Zeitpunkt 1 ein Feedback durch Hochschulangehörige und zum Zeitpunkt 2 ein Peer-Feedback erfolgen.
2. Die summative Bewertung des fertigen Produkts sollte dafür nur noch das Prädikat/die Note mit einer kurzen Begründung umfassen. Denn zu diesem Zeitpunkt können die Studierenden Hinweise nicht mehr umsetzen, zudem beachten sie nach Bekanntgabe der Note inhaltliches Feedback kaum mehr (siehe Kap. 9.1). Sinnvoll ist dagegen eine mündliche Schlussbesprechung der Arbeit.

Wer hat's getan?

Die Feststellung der Urheberschaft ist bei allen Formen von Hausarbeiten schwierig – letztlich können die Studierenden bei Recherche und Schreiben beliebig Unterstützung in Anspruch nehmen, von Peers über Software (inkl. KI-artigen Programmen wie DeepL oder ChatGPT) bis zu Plagiaten und Ghostwriting (durch Menschen oder künftig Maschinen). Das Problem ist aber kleiner als es zunächst scheint. Denn einerseits kann mit Ausnahme von Plagiaten und Ghostwriting die erhaltene Unterstützung durchaus lernwirksam sein und sogar als Unterstützung/Entlastung der Lehrenden betrachtet werden. Andererseits kann die Gefahr von Plagiaten und Ghostwriting massiv reduziert werden, wenn wissenschaftliche Arbeiten wie im obigen Abschnitt vorgeschlagen durch Lehrende formativ begleitet werden und eine Schlussbesprechung stattfindet. Vor allem wenn die formative Leistungsrückmeldung im direkten Gespräch erfolgt, wird es schnell offensichtlich, wenn Studierende ihre Arbeit nicht selbst verfasst haben. Trotzdem kann zusätzlich das Unterzeichnen einer Redlichkeitserklärung durch die Studierenden sinnvoll sein – zur Sensibilisierung der Studierenden und auch aus rechtlichen Gründen.

Ökonomie

Gemessen am Zeitaufwand der Studierenden ist jener der Lehrenden relativ klein, so dass wissenschaftliche Arbeiten grundsätzlich eine gute Ökonomie aufweisen. Das gilt auch dann noch, wenn Lehrende sich Zeit für die Begleitung nehmen, etwa durch das Erteilen von formativem Feedback. Unnötigen Zusatzaufwand ersparen können sie sich, indem die inhaltlichen und formalen Ansprüche an den Erarbeitungsprozess und das Schlussprodukt frühzeitig

und transparent kommuniziert werden. Das erhöht die Qualität der Produkte und reduziert das Risiko von Beschwerden und Rekursen.

Für die Studierenden weisen wissenschaftliche Arbeiten in der Regel ein gutes Verhältnis von Aufwand und Lerngewinn auf: Der Aufwand ist zwar in aller Regel groß, aber durch den tiefenorientierten Lernprozess ist das Gelernte in der Regel auch relevant und nachhaltig erworben. Sofern sie die Fragestellung mitbestimmen konnten, sind die Studierenden oft auch intrinsisch motiviert, was dem Lerngewinn förderlich ist.

Größte Vorteile

- kann umfassende Lernprozesse bei den Studierenden bewirken
- im Verhältnis zum Aufwand der Dozierenden viel Lerngewinn der Studierenden möglich
- hohe inhaltliche Validität

Größte Nachteile

- Vergleichbarkeit der studentischen Leistungen ist schwierig
- Anfälligkeit für unredliches Verhalten (allerdings vor allem bei fehlender Begleitung)
- hoher Zeitaufwand der Studierenden kann zu hohen Belastungen und Konflikten mit anderen Studienelementen führen

7.7 Beispiele für innovativere Formate

Die folgenden Beispiele unterscheiden sich hinsichtlich ihres Konkretisierungsgrades, und es werden nicht alle Gestaltungsvarianten aufgeführt. So handelt es sich bei der mehrphasigen Beurteilung eines Endprodukts (Kap. 7.7.1) eher um eine Grundstruktur, während die zweistufige Prüfung nach Wieman (Kap. 7.7.2) ein konkreter Gestaltungsvorschlag ist. Natürlich kann und soll auch dieser an die jeweilige Situation angepasst werden, was ebenfalls für Poster und Posterkonferenz (Kap. 7.7.3) gilt: Alle beschriebenen Formate können und sollen je nach den kontextspezifischen Anforderungen ausgearbeitet und weiterentwickelt werden.

7.7.1 *Mehrphasige Beurteilung eines Endprodukts*

Allgemeine Beschreibung: Es handelt sich um ein allgemeines Konzept, das auf alle Leistungsnachweisformate angewendet werden kann, die am Ende als Produkt vorliegen. Entsprechend ist es bezüglich des Raums, der Medien oder der Sozialform weitgehend offen. In aller Regel werden der Erarbeitungsprozess und/oder das Produkt eher komplex sein, so dass auch Anteile divergenten Denkens erforderlich sein dürften.

Entscheidend ist bei diesem Vorgehen, dass man das Produkt nicht nur nach seinem Abschluss summativ bewertet, sondern es (und ggf. den Erarbeitungsprozess) während seiner Erarbeitung einmal oder mehrmals formativ beurteilt. Das bedeutet, dass die Studierenden während der Erarbeitung des Produkts Hinweise zur Weiterarbeit oder Verbesserung erhalten.

Backwash - bewirktes Lernverhalten der Studierenden

Durch die formativen Beurteilungsanlässe wissen die Studierenden genauer, wo sie bezüglich des zu erarbeitenden Produkts stehen. Diese Gewissheit ist hilfreich für die weitere Arbeit: Gelungenes kann belassen werden, Fehlendes kann ergänzt werden, methodische Mängel können behoben werden, vielleicht werden im Austausch mit den Lehrenden auch weitere Ideen generiert. In jedem Fall kann die weitere Arbeit am Produkt durch die zwischenzeitliche Leistungsrückmeldung fokussierter erfolgen und die Studierenden können gezielt an den Aspekten arbeiten, die ihnen noch Mühe bereiten. Nachteile sind eigentlich nur zu befürchten, wenn die formative Leistungsrückmeldung nicht auf konstruktive Weise erfolgt (siehe Kap. 8 zum Geben von Leistungsrückmeldungen).

Lebensweltbezug

Der Lebensweltbezug hängt in erster Linie von der Art des Endprodukts ab.

Geeignet für welche Art der Leistungsrückmeldung

Bei der Beurteilung des Produkts sollte die Kriteriálnorm zur Anwendung kommen. Die Kollektivnorm ist generell zu vermeiden. Die Individualnorm kann formativ zur Anwendung kommen, wenn die Weiterentwicklung des Produkts anhand des erhaltenen Zwischenfeedbacks beurteilt wird.

Das Produkt wird bei diesem Vorgehen in aller Regel summativ beurteilt. Die während der Erarbeitung des Produkts vorgenommenen Beurteilungen sollten in jedem Fall formativ sein, also konkrete Hinweise umfassen, welche Teile des Produkts oder seines Erarbeitungsprozesses wie gut gelungen sind, was noch verbessert werden kann und was noch fehlt bzw. welche Schritte

als nächstes anstehen. Zusätzlich handelt es sich bei diesem Fall um eine der ganz wenigen Ausnahmen, bei denen inhaltliche Leistungsrückmeldung und Bewertung durch Noten oder Prädikate kombiniert werden können. Denn dadurch, dass das erhaltene Feedback für die Fertigstellung des Produkts relevant ist und auch das fertige Produkt wieder bewertet wird, ist für die Studierenden ein starker (freilich extrinsischer) Anreiz gegeben, der erhaltenen Rückmeldung Beachtung zu schenken und erhaltene Hinweise umzusetzen.

Wer hat's getan?

Insbesondere wenn die zwischenzeitliche Leistungsrückmeldung durch direkte Gespräche erfolgt, schafft das mehrphasige Vorgehen eine zusätzliche Sicherheit, dass die Studierenden ihre Produkte selbst erarbeiten (siehe „Wer hat's getan“ in Kap. 7.6.4 zur wissenschaftlichen Arbeit).

Ökonomie

Für die Lehrenden ist das Erteilen von Zwischen-Rückmeldungen auf den ersten Blick ein Zusatzaufwand. Da dadurch aber bessere Endprodukte erzielt und unlautere Verhaltensweisen reduziert werden, dürfte mindestens ein Teil der investierten Zeit nach Abgabe der Endprodukte wieder eingespart werden. Die verbesserten Kontakte zu den Studierenden und die besseren Endprodukte dürften zudem auch mehr Freude an der Arbeit bewirken.

Für die Studierenden verbessert sich das Verhältnis zwischen Aufwand und Ertrag klar, da sie fokussierter an ihren Produkten arbeiten und ihre Zeit vor allem in Wissens- und Könnensbereiche investieren können, in denen sie noch Lernbedarf haben.

Größte Vorteile

- größere Lernwirksamkeit
- verbessert Kontakt zwischen Lehrenden und Studierenden und erhöht beiderseits Motivation
- reduziert Wahrscheinlichkeit unredlichen Verhaltens durch positive Anreize

Größte Nachteile

- kann zu einem Interrollenkonflikt bei den Lehrenden führen (Beraten vs. Beurteilen, siehe Kap. 3.3.1)

7.7.2 Zweistufige Prüfung nach Wieman

Allgemeine Beschreibung: Die zweistufige Prüfung ist ein Leistungsnachweisformat, das vom Physik-Nobelpreisträger Carl Wieman entwickelt wurde (vgl. Wieman, Rieger u. Heiner 2014, für eine deutsche Erläuterung siehe Kautz u. Billerbeck 2018). Es integriert formative und summative Beurteilung, kombiniert Einzelarbeit mit sozialem Austausch und adressiert sowohl die Produkt- als auch die Prozessebene. Da das Format Aufgaben bedingt, die auf hohe Lernzielstufen zielen, aber eine klare richtige Lösung haben, zielt es primär auf konvergentes Denken. Das konkrete Vorgehen sieht folgendermaßen aus:

Die erste Phase entspricht einer schriftlichen Klausur, in der die Studierenden einzeln die Prüfungsaufgaben bearbeiten und ihre Lösungen anschließend abgeben. In der zweiten Phase bearbeiten die Studierenden zu dritt oder viert in Kleingruppen die gleichen Probleme wie in der ersten Phase und verständigen sich auf die richtige Lösung. Bei Wieman reichen die Gruppen ihren Konsens dann gemeinsam unter Nennung aller Studierendennamen inklusive Matrikelnummer ein. Prüfungsrechtliche Vorgaben, etwa in Deutschland (vgl. Hirdina 2019, 56f.), können aber verlangen, dass die Studierenden ihre Lösungen nach der Gruppendiskussion wiederum einzeln einreichen und dabei auch vom Gruppenkonsens abweichen können (vgl. Kautz u. Billerbeck 2018, 106). Wichtig ist zudem, dass die Studierenden vorab in der Lehrveranstaltung „ausreichend Gelegenheit hatten, das fachliche Diskutieren und gemeinsame Erarbeiten von Lösungen zu üben“ (Kautz u. Billerbeck 2018, 108).

Backwash – bewirktes Lernverhalten der Studierenden

Die erste Stufe des Formats hat grundsätzlich den gleichen Backwash wie andere schriftliche Klausuren. Die zweite Stufe allerdings nutzt die Vorteile sozialen Lernens, wie sie in Kapitel 4.7 geschildert wurden – insbesondere den Umstand, dass Menschen die Argumente ihrer Mitmenschen kritischer überprüfen als ihre eigenen und das dadurch resultierende Prinzip, dass sich in fachlich begründetem Austausch in der Regel das bessere Argument durchsetzt. Die zweite Stufe des Formats ist also nicht nur eine Prüfung, sondern führt zugleich zu Abruf- und Lernprozessen, die den nachhaltigen Erwerb der Inhalte begünstigen.

Lebensweltbezug

Der Lebensweltbezug hängt einerseits von den Aufgabenstellungen ab. Andererseits ist das gemeinsame Bearbeiten und Lösen von Problemen sowie der kritisch-konstruktive Austausch von Argumenten eine zentrale Fähigkeit in

der modernen Arbeitswelt, was das Format sehr authentisch macht (vgl. Wieman, Rieger u. Heiner 2014, 51).

Geeignet für welche Art der Leistungsrückmeldung

Zur Bezugsnormorientierung gilt das Gleiche wie für schriftliche Klausuren (Kap. 7.6.1): Es soll die Kriterialnorm gelten. Hinzu kommt eine etwaige Berücksichtigung der Individualnorm:

Das Format wurde ursprünglich für summative Bewertungen erfunden, kann natürlich auch formativ eingesetzt werden – in diesem Fall werden weder die Einzel- noch die Gruppenlösungen summativ bewertet. Wird summativ bewertet, so wird in der Regel die Gruppenbearbeitung bei den Studierenden berücksichtigt, bei denen diese besser ausfiel als ihre Einzelbearbeitung. Wieman, Rieger u. Heiner (vgl. 2014, 51) empfehlen dabei eine Gewichtung des Verhältnisses von Einzel- zu Gruppenbearbeitung von 3 zu 1. Wichtig ist, dass die Gruppenbearbeitung nur dann für die summative Bewertung berücksichtigt wird, wenn sie besser ausgefallen ist als die Einzelbearbeitung.¹⁰⁵

Wer hat's getan?

Für die erste Stufe gilt das Gleiche wie für schriftliche Klausuren (Kap. 7.6.1), für die zweite das Gleiche wie bei mündlichen Prüfungen (Kap. 7.6.2).

Ökonomie

Nobelpreisträger Wieman meint: „Die gesamte Bewertungszeit erhöht sich durch die Gruppenprüfung nur geringfügig, da ein großer Teil der Lösungen vollständig richtig ist, was die Bewertung einfach und schnell macht“ (Wieman, Rieger u. Heiner 2014, 51).

Für Studierende erhöht sich bei gleichem Aufwand der Ertrag gegenüber herkömmlichen Klausuren, da sie ihre Bewertung durch die zweite Stufe verbessern und dabei zugleich noch einen Lerngewinn erzielen können. Zudem hilft das Wissen um die zweite Stufe, Prüfungssängste zu reduzieren.

Größte Vorteile

- verbesserter Lerngewinn gegenüber herkömmlichen schriftlichen Klausuren

105 Dies gilt umso mehr, falls die Studierenden nach der Gruppendiskussion wiederum Einzelbearbeitungen einreichen (müssen) – in diesem Fall könnten sonst Anreize entstehen, die anderen Gruppenmitglieder zu täuschen, um sie zu falschen Lösungen zu verleiten.

- Betonung der Wichtigkeit sozialen, kritischen Austauschs von Argumenten
- hohe Akzeptanz bei den Studierenden

Größte Nachteile

- ggf. einmaliger Aufwand, um die rechtlichen Bedingungen abzuklären

7.7.3 Poster und Posterkonferenz

Allgemeine Beschreibung: Die Studierenden erstellen Poster, die ihre Arbeitsergebnisse zu einer bestimmten Aufgabenstellung oder ein spezifisches, zugeteiltes Thema darstellen. Die Poster werden im Rahmen einer Art Konferenz präsentiert, bei der andere Studierende, aber auch externe Expertinnen oder ein breiteres Publikum teilnehmen können.

Die Bewertung kann formativ oder summativ erfolgen oder die beiden Beurteilungsziele in einem mehrphasigen Vorgehen kombinieren (siehe Kap. 7.7.1). Sofern mehrphasig beurteilt wird, kann neben dem Poster und seiner Vorstellung als Produkt auch der Prozess der Erarbeitung beurteilt werden. Poster können sowohl in Einzel- als auch Gruppenarbeit erarbeitet und präsentiert werden. In Bezug auf Raum und Medien gibt es diverse Optionen. Ursprünglich fanden Posterkonferenzen vor Ort in einem oder mehreren Räumen statt, in denen die Studierenden sich gegenseitig ihre Poster präsentierten (für Organisationsmöglichkeiten vgl. Decker 2023, 9–12). Die Konferenz kann aber auch online erfolgen, wobei sowohl asynchrone als auch synchrone Varianten denkbar sind (vgl. Arcila Hernández, Chodkowski u. Treibergs 2022). Das Format kann neben konvergentem Denken auch divergentes Denken fördern, das nicht nur bei der Erwägung diverser Möglichkeiten zur grafisch konzisen Darstellung der Inhalte, sondern auch bei der Diskussion der Poster zum Tragen kommen kann. Es ist sinnvoll, im Umgang mit Postern ungeübten Studierenden Einführungen anzubieten wie einen Poster-Workshop oder ein Online-Tutorial (vgl. Decker 2018, 186).

Backwash – bewirktes Lernverhalten der Studierenden

Ist das Format integraler Bestandteil der Lehrveranstaltung, bewirkt es bei der Erarbeitung der Poster das Anwenden tiefenorientierter Lernstrategien (ähnlich wie bei didaktisch gut gestalteten Vorträgen, siehe Kap. 7.6.3). Als Vorteil kommt hier aber dazu, dass die Studierenden für das Erarbeiten einer komprimierten Darstellung erhebliche kognitive Umformungsprozesse durchlaufen müssen, was die Verarbeitungstiefe deutlich erhöht (siehe Kap. 4.6.1). Ein weiterer Lerngewinn erfolgt, werden die Poster im Rahmen

von Konferenzen kritisch diskutiert: Menschen überprüfen die Argumente ihrer Mitmenschen kritischer als ihre eigenen, wodurch sich wie bei der zweistufigen Prüfung nach Wieman (Kap. 7.7.2) in fachlich begründetem Austausch in der Regel das bessere Argument durchsetzt (siehe Kap. 4.7). Die Posterbesprechung ist somit nicht nur ein Leistungsnachweis, sondern führt zugleich bei den Präsentierenden wie dem Publikum zu Abruf- und Lernprozessen, die den nachhaltigen Erwerb der Inhalte begünstigen. Die Wirkung der Posterkonferenz ist somit vergleichbar mit jener der zweiten Stufe in der zweistufigen Prüfung nach Wieman.

In der Tendenz dürften Posterpräsentationen zudem weniger Prüfungsängste auslösen als klassische Vorträge, da hier mit dem Poster ein getrennt beurteilbares Produkt vorliegt, was die Präsentationssituation entlastet.

Lebensweltbezug

In Postern und ihrer Besprechung sind alle Abstufungen von Lebensweltbezug möglich, siehe die diesbezüglichen Ausführungen zum Format Vortrag in Kapitel 7.6.3. Hinzu kommt aber, dass das Format an sich einen hohen Lebensweltbezug hat: In fast allen akademischen Disziplinen wie auch in vielen anderen anspruchsvollen Berufsfeldern ist der fachliche Austausch anhand von Postern gebräuchlich – hier wird also etwas geübt, das die Studierenden später gebrauchen können. Darüber hinaus ist das kondensierte, auch grafische Darstellen von Kernaussagen eine Fähigkeit, die ebenfalls in fast allen akademischen Disziplinen und Berufsfeldern wesentlich ist. Zusätzlich gesteigert werden kann der Lebensweltbezug, wenn für die Posterkonferenz Fachexperten aus der Akademie oder dem Berufsfeld und/oder ein breiteres Publikum eingeladen werden.

Geeignet für welche Art der Leistungsrückmeldung

Grundsätzlich gilt Ähnliches wie hinsichtlich der Leistungsrückmeldung bei der mehrphasigen Beurteilung eines Endprodukts und der zweistufigen Prüfung nach Wieman (Kap. 7.7.1 und 7.7.2). Wird das Poster als Endprodukt betrachtet, kann es durch ein mehrphasiges Format formativ begleitet werden. Wird das Poster als Stufe 1 betrachtet, kann die Besprechung des Posters als Stufe 2 betrachtet werden: Durch ein überzeugendes Diskussionsverhalten bei der Präsentation des eigenen Posters können die Studierenden ihre Bewertung verbessern. Hier ist ebenso eine stärkere Gewichtung der Diskussionsphase bis hin zu einer Gleichgewichtung von Poster und Diskussion denkbar. Damit für das Diskussionsverhalten des studentischen Publikums keine kontraproduktiven sozialen Anreize und nicht zu viel Druck entstehen, sollte auch hier die Bewertung des Diskussionsverhaltens nur berücksichtigt werden, wenn sie besser ausfällt als jene des Posters.

Wer hat's getan?

Grundsätzlich gilt das Gleiche wie bei Vorträgen (Kap. 7.6.3), wobei die in der Konferenz erfolgende kritische Besprechung das Risiko noch reduziert, dass Lernende Poster vorstellen, die sie nicht selbst erarbeitet haben.

Ökonomie

Poster und Posterkonferenzen können für Lehrende ökonomisch sein. Der Einbezug der Studierenden oder externer Expertinnen als Feedbackinstanzen können sie zeitlich entlasten sowie Feedbackqualität und Lernwirksamkeit erhöhen. Das Format ist zudem im Gegensatz zu klassischen Vorträgen auch bei größeren Teilnehmendenzahlen gut einsetzbar. Die Posterkonferenz muss dann vielleicht auf mehr Termine verteilt werden oder es müssen zusätzliche Lehrende für die Durchführung und Beurteilung herangezogen werden – aber dieser Zusatzaufwand fällt nicht höher aus als bei der Steigerung der Teilnehmerzahlen in anderen Leistungsnachweisformaten.

Für die Studierenden weist das Format grundsätzlich eine gute Ökonomie auf, da alle ihre Aktivitäten sich direkt auf den Lernzuwachs und/oder die Bewertung auswirken.

Größte Vorteile

- hohe Lernwirksamkeit und Authentizität
- gute Ökonomie für Lehrende und Studierende
- ohne extreme Kontrollmaßnahmen wenig anfällig für unredliche Verhaltensweisen

Größte Nachteile

- Gestalten von Postern muss ggf. zuerst eingeübt werden¹⁰⁶
- ggf. Anpassung von Prüfungsordnungen nötig, um das Format summativ einsetzen zu können (vgl. Decker 2018, 188)

106 Allerdings wäre dies auch hinsichtlich Bildschirmpräsentationen dringend erforderlich, wie sich an textlastigen Folien, zu kleinen Schriftgrößen, unnötigen Animationen, ungünstigen Farbkontrasten etc. von Vorträgen im ersten Studienjahr bis zu solchen bei wissenschaftlichen Konferenzen zeigt.

7.8 Auflistung weiterer innovativer Formate

Grundsätzlich sind der Gestaltung von Leistungsnachweisformaten wenige Grenzen gesetzt, die Zahl der Gestaltungsmöglichkeiten ist kaum benennbar (siehe Kap. 7.5). Deshalb werden nachfolgend 12 weitere erprobte innovative Formate in aller Kürze vorgestellt. Tabelle 7 enthält jeweils den Namen des Formates, eine Kurzbeschreibung und einen Literaturhinweis zu einer ausführlicheren Beschreibung des Formats. Auch für diese Formate gilt, dass es sich primär um Anregungen handelt und sie jeweils an den konkreten fachlichen, curricularen und institutionellen Kontext angepasst werden müssen. Überdies können Elemente verschiedener Formate zu einem neuen Format kombiniert werden. In jedem Fall erweitert es den eigenen Denkraum, sich mit Formaten aus ganz anderen Fachbereichen, Hochschultypen oder Sprachräumen auseinanderzusetzen. Als besonders anregend seien die Sammelbände von Gerick, Sommer u. Zimmermann (2018) und Betts u. Oprandi (2022) empfohlen, in denen viele weitere Leistungsnachweisformate und zahlreiche fachliche Kontexte beschrieben werden.

Tab. 7: Liste von innovativen Leistungsnachweis-Formaten

Leistungsnachweis-Format	Literaturhinweis
<i>24h-Hausarbeit</i> Schriftliche Hausarbeit mit klar begrenztem Zeitabstand zwischen Aufgabenstellung und Abgabe des Produkts.	Reinmann 2018
<i>Blogbeitrag</i> Studierende verfassen einen oder mehrere Blogbeiträge in einem passwortgeschützten oder öffentlichen Blog. Auch Kommentare anderer Beiträge sind möglich.	Kahrs u. Kohl 2018
<i>Challenge und Challenge X</i> Studierendengruppen arbeiten an realitätsnahen, komplexen Problemstellungen und entwickeln praxisrelevante Handlungsfähigkeiten; bei der Challenge X entwickeln die Gruppen die Problemstellung selbst. Die Challenges werden durch Coaches begleitet.	Studiengangsteam Data Science 2022; Challenge X: Benites et al. 2022
<i>Elevator Pitch</i> Studierende präsentieren eine Arbeit in einem (max.) fünfminütigen mündlichen Format; Fokus auf der Fähigkeit, zentrale Argumente hervorzuheben und einem bestimmten Publikum zu vermitteln.	Saunders 2022
<i>Haikus und andere Gedichte verfassen</i> Studierende bringen Aussagen in Gedichtform. Dies fokussiert die Fähigkeit, innerhalb von Wortgrenzen und sprachformalen Vorgaben formulieren zu können. Geht auch mit Sonetten, Limericks etc.	Hancock 2022

<i>Hospitation</i> Studierende beobachten ein (potenzielles) Berufsfeld unter Einnahme einer beruflichen Perspektive und erstellen dazu einen Reflexionsbericht. Spezialvariante in der Lehrerbildung: Studierende hospitieren sich gegenseitig in Unterrichtspraktika und verfassen Hospitationsberichte.	Karcher 2018 (allgemein); Lehrerbildung: Kreis 2015¹⁰⁷
<i>Moot Court</i> Juristisches Format, bei dem Studierende in simulierten Gerichtsverhandlungen auftreten, um Fähigkeiten zur juristischen Argumentation usw. zu demonstrieren. Strukturierte Debatten eignen sich auch generell als Leistungsnachweise.	Finke 2019; Debatten allgemein: Gutsmiedl-Schümann 2018
<i>Objektive strukturierte praktische/klinische Prüfung</i> Studierende demonstrieren praktische Fähigkeiten und klinische Kompetenzen in realitätsnahen Situationen. (Ursprung: Medizin und weitere Gesundheitswissenschaften)	Dilly 2018
<i>Pecha Kucha</i> Bei Pecha-Kucha-Präsentationen werden 20 Folien mit Bildern und höchstens minimalem Text mit einer Dauer von jeweils 20 Sekunden präsentiert, um prägnante und gut vorbereitete Präsentationen zu fördern. Geht mit Gruppen- und Einzelpräsentationen.	Zimmermann, Germo 2018
<i>Unessay</i> Studierende können das Format der Themenbearbeitung frei wählen (z. B. Podcasts, Modelle, Comics oder Essays); Studierende erarbeiten Bewertungskriterien für Unessays gemeinsam durch Peer Review.	Walden 2022
<i>Videoclips erstellen</i> Studierende produzieren einen Videoclip zu einem Thema. Dieses Format liegt heute vielen Studierenden näher als schriftliche Texte und stellt damit eine niederschwelligere Alternative z. B. für Studienanfängerinnen dar. Zudem wird die Arbeit mit Videos in vielen künftigen Tätigkeitsfeldern immer wichtiger.	Bourke 2022
<i>Virtual Reality (VR), Augmented Reality (AR)</i> Studierende bearbeiten Aufgaben in einer virtuellen oder virtuell angereicherten Umgebung. So können sie Handlungsfähigkeiten in einer sicheren Umgebung demonstrieren, z. B. beim Bedienen komplexer Geräte wie Flugzeugen oder bei medizinischen Behandlungen.	VR: Sun, Chikwem u. Nyingifa 2019 AR: Sanchez et al. 2014

107 Kollegiale Hospitationen sind in der Weiterbildung von Lehrpersonen schon länger etabliert, in der Ausbildung/im Studium hingegen wesentlich weniger. Der Beitrag von Kreis (2015) fokussiert entsprechend auf bereits berufstätige Lehrpersonen, das Konzept lässt sich aber auch auf Studierende übertragen.

8 Formative Leistungsrückmeldung

Wenn Leistungsnachweise eine didaktisch-lernbezogene Funktion erfüllen sollen, ist die formative Leistungsrückmeldung die zentrale Gelenkstelle zwischen der erbrachten Lernleistung und dem zu fördernden weiteren Lernen der Studierenden. Die Wichtigkeit von Feedback ist heute unter Lehrpersonen aller Stufen weitgehend unbestritten.¹⁰⁸ Das Erteilen und Einfordern von lernwirksamem Feedback ist allerdings äußerst anspruchsvoll. Wie eine Rückmeldung verarbeitet wird, ist zudem ebenso wichtig wie die Art und Weise, in der sie gegeben wird.

In diesem Kapitel wird deshalb zuerst erläutert, was der Begriff der Leistungsrückmeldung genau meint (Kap. 8.1). Auf dieser Basis werden verschiedene Eigenschaften von Leistungsrückmeldungen unterschieden. Das Kernstück des Kapitels bilden dabei die Hinweise zur motivationalen Wirkung von Leistungsrückmeldungen – sie sind für alle Arten, Zeitpunkte und Kontexte relevant (Kap. 8.2).

Eine Leistungsrückmeldung kann sich auf die zurückliegende Entwicklung, den aktuellen Stand oder die zukünftigen Entwicklungsmöglichkeiten richten, sie kann sich auf verschiedene Ebenen beziehen wie die Aufgabe oder den Lernprozess (Kap. 8.3). Auch sind verschiedene Instanzen zu unterscheiden, die Feedback geben können (Kap. 8.4).

Je nach der Medialität der Rückmeldung sind bestimmte Aspekte besonders zu beachten (Kap. 8.5). Ein weiterer zentraler Punkt der didaktisch-lernbezogenen Leistungsbeurteilung ist zudem, dass die Studierenden die erhaltenen Leistungsrückmeldungen auch nutzen, um ihr zukünftiges Lernverhalten zu steuern – dieses Schließen des Feedbackkreislaufs ist das Ziel von formativen Beurteilungen (Kap. 8.6).

8.1 Was ist eine formative Leistungsrückmeldung?

Eine formative Leistungsrückmeldung ist eine bestimmte Form von kommunikativem Feedback. Deshalb lohnt sich eine kurze Verortung von Feedback in der menschlichen Kommunikation (Kap. 8.1.1), bevor wir genauer be-

108 Ich gebrauche die Begriffe Feedback und Rückmeldung übereinstimmend mit dem Alltagssprachlichen Gebrauch (vgl. Dudenredaktion o. J.) weitgehend synonym, während andere Autoren einen Unterschied konstruieren (z. B. Müller u. Ditton 2014, 13–16).

stimmen, was eine formative Rückmeldung zu einer Leistung ausmacht (Kap. 8.1.2).

8.1.1 *Feedback als kommunikativer Prozess*

Feedback ist eine spezifische kommunikative Aktivität, in der einer Person durch eine Instanz Informationen über ihr Verhalten oder ihre Äußerungen angeboten werden. Bei der Instanz kann es sich um andere Menschen handeln wie Lehrende, Peers, Freunde, Familienmitglieder, aber auch um unbelebte Instanzen wie Bücher und Computerprogramme oder um die eigene Person oder Erfahrung (vgl. Hattie u. Timperley 2007, 81).¹⁰⁹ Feedback kann dabei ganz verschiedene Aspekte adressieren. Es kann korrigieren oder bestätigen, es kann alternative Vorgehensweisen vorschlagen, es kann ermutigen usw.

Feedbackprozesse spielen eine zentrale Rolle in der menschlichen Kommunikation. Um ihre Elemente und deren Wirkung genauer analysieren zu können, lohnt sich ein Blick auf alltägliche Verständigungsprozesse. Zunächst erscheinen diese als Abfolge von Informationen, die mündlich und/oder schriftlich ausgetauscht werden, wobei die verschiedenen Beteiligten abwechselnd Sender oder Empfänger sind. Zudem hat menschliche Kommunikation stets mehr als eine Ebene – so unterscheidet etwa Schulz von Thun (1981, 25–30) vier Ebenen einer Äußerung: Neben einer Sach- und einer Beziehungsebene umfasst sie jeweils auch eine mehr oder weniger bewusste Selbstdarstellung der Senderin sowie einen Appell an den Empfänger, also ein Wirkungsziel. Es hängt auch vom Empfänger ab, auf welche Ebene einer Äußerung er in welcher Weise reagiert. Weiter sind Kommunizierende immer parallel sowohl Sender als auch Empfänger von Botschaften. Dies ist besonders auffällig in mündlichen Gesprächen, wo ein Zuhörer durch Gestik und Mimik seine Reaktion auf das Gehörte rückmelden kann.

Das Transaktionsmodell der Kommunikation nach Barnlund (1970) zeigt auf, dass Kommunikationsprozesse noch deutlich umfassender sind. Ihm zufolge hilft Kommunikation uns nicht nur, die soziale Welt um uns zu verstehen, sondern sie bringt sie zu wesentlichen Teilen hervor und kann sie folglich auch verändern. So kommunizieren wir nicht nur, um Nachrichten auszutauschen, sondern auch, um Beziehungen und Gemeinschaften zu schaffen und zu bearbeiten oder unser Selbstkonzept zu konstruieren. Dabei spielen mehrere Kontextebenen eine Rolle. Dazu gehört der soziale Kontext mit offiziellen Regeln oder informellen Normen, welche die Kommunikation betreffen, etwa „unterbrich andere nicht“ oder „bedanke dich für Komplimente“.

109 Bei Selbsteinschätzungen oder dem Lernen aus eigenen Erfahrungen kann man eigentlich nur dann von einer kommunikativen Aktivität im engeren Sinne sprechen, wenn diese Selbstreflexionen explizit verbalisiert werden.

mente“. Auch Aspekte der Beziehung spielen eine Rolle, etwa Hierarchieunterschiede oder wie gut man das Gegenüber kennt. Kulturelle Aspekte kommen ebenfalls zum Tragen und werden kommunikativ ausgehandelt, etwa Aspekte der Identität wie Nationalität, Ethnie, Geschlechtszugehörigkeit oder sozialer Status. Wie diese kurzen Erläuterungen nahelegen, ist Kommunikation in vielen Fällen für die einen Beteiligten heikler oder anspruchsvoller als für die anderen. Dies ist etwa der Fall, wenn jemand mit den sozialen Normen in einem neuen Umfeld noch nicht gut vertraut ist (z. B. bei Studienbeginn), wenn mit hierarchisch höher gestellten Personen (z. B. Lehrenden) kommuniziert wird oder wenn Angehörige einer Minderheit sich mit Angehörigen einer Mehrheit verständigen (z. B. Angehörige einer anderen Ethnie mit Einheimischen). Insgesamt dürften somit viele Studierende in der Kommunikation mit Lehrenden mehr Unsicherheiten verspüren als umgekehrt die Lehrenden ihnen gegenüber. Dies spielt auch eine Rolle, wenn es um die Verständigung über Lernleistungen und Lernstand geht.¹¹⁰

8.1.2 Förderorientierte Rückmeldung zu einer Lernleistung

Hattie u. Timperley beginnen ihre einflussreiche Publikation „The Power of Feedback“ mit den Worten „Feedback ist eine der mächtigsten Einflussgrößen auf Lernen und Leistung, aber dieser Einfluss kann entweder positiv oder negativ sein“ (Hattie u. Timperley 2007, 81, Übers. tz). Dies ist lernpsychologisch gut erklärbar, wie in Kapitel 4 gezeigt wurde: Einerseits können Lernende Feedback nutzen, um das eigene Wissen und Können zu überprüfen und gegebenenfalls anzupassen: Dadurch können sie mittels Vorhergesageschleifen lernen (Kap. 4.4.1 und 4.7), ihr Vorwissen nachhaltig erweitern (Kap. 4.4.2) und ihr metakognitives Wissen und ihre metakognitive Steuerung von Denken und Lernen weiterentwickeln (Kap. 4.4.3). Andererseits hat Feedback in der Regel erhebliche motivationale Wirkungen, die positiv oder negativ ausfallen können. Die eben skizzierten Wirkungen treten nur ein, wenn das Feedback motivational positiv oder neutral wirkt. Wirkt es demotivierend, dann werden Lernprozesse reduziert oder gar abgebrochen, was der von Hattie u. Timperley erwähnten negativen Wirkung entspricht. Die tritt insbesondere ein, wenn das Feedback die drei grundlegenden psychologi-

110 Neben diesen systematischen Ebenen wird Kommunikation auch situativ durch äußere Einflüsse auf Wahrnehmung und Verhalten beeinflusst. Dies führt zu Schwankungen, die weder mit den Persönlichkeiten noch den sozialen Konstellationen direkt zusammenhängen. Zu den äußeren Einflüssen gehören Aspekte wie der physikalische Kontext (ist die Umgebung warm/kalt, laut/leise etc.), aber auch der körperliche Zustand der Kommunizierenden (Wachheit/Müdigkeit, Wohlbefinden/Krankheit etc.) oder ihre psychische Befindlichkeit (z. B. schlechte Laune wegen eines vorangehenden Ereignisses) (vgl. Barnlund 1970).

schen Bedürfnisse der eigenen Wirksamkeit, der sozialen Eingebundenheit und der Eigenständigkeit unterminiert und dadurch die intrinsische Motivation abschwächt (siehe Kap. 4.5.2).

Formatives Feedback zu einer Lernleistung ist somit eine Rückmeldung, die eine förderliche Wirkung auf die Entwicklung der Lernleistung beabsichtigt, zu der sie erfolgt. Entscheidend ist dabei, dass die Studierenden die Feedbackinformationen sinnvoll verarbeiten und zur Verbesserung ihrer Leistungen oder ihres Lernens nutzen (vgl. Carless u. Boud 2018, 1315). Man bezeichnet das als Schließen des Feedbackkreislaufs (Kap. 8.6). Dies beschreibt freilich erst den Anspruch an die Wirkung der Leistungsrückmeldung. Im Folgenden wird deshalb geklärt, was zu berücksichtigen ist, damit eine solche Wirkung erzielt werden kann:

- Zu unterscheiden ist zwischen den inhaltlichen Eigenschaften von Rückmeldungen und ihrer Wirkung auf das inhaltliche Lernen einerseits sowie ihrer motivationalen Wirkung andererseits.
- Die Art und Weise, in der Feedback erteilt wird, hat entscheidenden Einfluss auf die motivationale Wirkung des Feedbacks. Dies ist ausschlaggebend dafür, ob das Feedback bei den Studierenden eine lernförderliche Wirkung entfaltet und wird deshalb noch vor den inhaltlichen Aspekten behandelt (Kap. 8.2).
- Auf der inhaltlichen Seite kann Feedback verschiedene Aspekte und Ebenen des Lernens adressieren (Kap. 8.3).
- Feedback kann durch unterschiedliche Instanzen erfolgen, welche ebenfalls einen Einfluss auf seine Wirkung haben (Kap. 8.4).
- Die Art der medialen Vermittlung des Feedbacks geht mit bestimmten Möglichkeiten und Einschränkungen einher (Kap. 8.5).

8.2 Leistungsrückmeldungen und Motivation

Die Art und Weise, in der Feedback zu Lernleistungen erfolgt, ist entscheidend für seine motivationale Wirkung. Sie ist stark damit verbunden, welche Einstellungen die Feedbackgebenden zu Lernprozessen wie auch gegenüber den Lernenden haben.

8.2.1 *Erwartungen an die Lernenden*

Die Erwartungen von Feedbackgebenden an Lernende beeinflussen die Lernentwicklung letzterer erheblich. Dies ist insbesondere durch zahlreiche Un-

tersuchungen empirisch belegt, welche die Wirkung von Lehrpersonen als Feedbackinstanzen untersucht haben.¹¹¹

Eines der bekanntesten Experimente ist jenes von Rosenthal u. Jacobson (1968), das unter dem Begriff des Pygmalion-Effekts berühmt wurde:¹¹² Die Forschenden führten mit Schülerinnen und Schülern Tests durch und teilten anschließend deren Lehrpersonen mit, welche Schülerinnen und Schüler im kommenden Jahr „aufblühen“ werden (es waren jeweils ca. 20 % einer Klasse). Diese Einteilung basierte aber auf einer zufälligen Ziehung und nicht auf den Testresultaten. Das Resultat zeigte, dass der IQ (nicht die konkreten Schulleistungen!) der „Aufblüher“ (bloomers) sich im Lauf dieses Jahres deutlich stärker steigerte als jener der Kontrollgruppe. Da die Lehrpersonen keine anderen Informationen als jene einer zu erwartenden besonders großen Leistungssteigerung erhalten hatten, kann der Unterschied nur mit den höheren Erwartungen der Lehrpersonen erklärt werden, die von den Lernenden internalisiert wurden.

Rubie-Davies (vgl. 2016, 148f.) konnte zeigen, dass sich die Erwartungen grundsätzlich von Lehrperson zu Lehrperson unterschieden. Mit anderen Worten, die einen Lehrpersonen haben für alle ihre Lernenden höhere Erwartungen, die anderen für alle tiefere. Zentral für unseren Zusammenhang ist: Je höher die Erwartungen einer Lehrperson sind, desto größer sind sowohl der Lernzuwachs ihrer Lernenden als auch deren lernbezogenes Selbstkonzept (ebd.).¹¹³

Aus diesen Befunden lässt sich ableiten, dass es äußerst bedeutsam ist, allen Studierenden ein großes Lernpotenzial zuzutrauen und Rückmeldungen in der Erwartung eines großen Lerngewinns zu formulieren. Dazu passen auch die Forschungsbefunde, die zeigen, dass inhaltlich hohe Erwartungen und ihre transparente Kommunikation an die Lernenden entscheidend sind für deren Lerngewinn (vgl. Hattie u. Zierer 2017, 60–71). Die Maxime für Lehrende und Feedbackgebende sollte deshalb lauten: Erwarte viel und fordere viel!

111 Untersuchungen zur Wirkung des Erziehungsverhaltens von Eltern weisen auf ähnliche Wirkungen hin (vgl. Lamborn et al. 1991); ob die Erwartungen von feedbackgebenden Peers gleich viel Einfluss haben, ist nicht geklärt und empirisch schwierig zu untersuchen.

112 Pygmalion, eine Figur der griechischen Mythologie, verliebt sich so sehr in eine von ihm selbst geschaffene Statue, dass er sich wünscht, sie würde zum Leben erwachen. Sein Wunsch wird erfüllt, sie heiraten und leben glücklich zusammen.

113 Auch das von unabhängigen Beobachtenden erhobene Lehr- und Kommunikationsverhalten von Lehrpersonen mit hohen und tiefen Erwartungen unterschied sich markant und dürfte eine wesentliche Moderatorvariable für die Wirkung ihrer Erwartungen auf den Lernerfolg der Lernenden sein (vgl. ebd.).

8.2.2 Die Wichtigkeit von Entwicklungsorientierung

In der Lernpsychologie geht man davon aus, dass Menschen generalisierte Überzeugungen bezüglich ihrer Lernfähigkeiten haben. Diese müssen nicht unbedingt jegliches Lernen betreffen, sondern können je nach Gebieten unterschiedlich sein. So kann sich jemand für sprachlich begabt halten und für mathematisch unbegabt. Oder jemand kann sich für mathematisch gebildet halten, aber davon überzeugt sein, Stochastik nicht zu beherrschen. Derartige Selbstkonzepte haben erhebliche Auswirkungen auf das Lernverhalten, die vor allem durch Interesse und Motivation vermittelt werden. Besonders interessant sind in unserem Zusammenhang die so genannten Kausalattributionen bei Erfolg und Misserfolg (Wild, Hofer u. Pekrun 2001, 229f.). Diese in Tabelle 8 zusammengefasste Taxonomie geht davon aus, dass sich Menschen darin unterscheiden, ob sie Lernergebnisse

- eher durch in ihrer eigenen Person liegende Ursachen oder durch äußerliche Umstände erklären (internale vs. externale Ursachen),
- auf zeitlich veränderliche oder überdauernde Faktoren zurückführen (variable vs. stabile Ursachen) und
- ob sie die Bedingungen der Leistungserbringung als durch sie selbst kontrollierbar erachten oder nicht.

Tab. 8: Kausalattributionen bei lernbezogenem Erfolg und Misserfolg (nach Wild, Hofer u. Pekrun 2001, 229)

	internal		external	
	stabil	variabel	stabil	variabel
kontrollierbar	Wissen	Anstrengung	Lern- umgebung	Aufgabenwahl
nicht kontrollierbar	Begabung	Krankheit	Schwierigkeit des Faches	Zufall

Es ist gut belegt, dass eine internale, kontrollierbare und variable Attribution, also die Erklärung des eigenen Lernerfolgs durch Anstrengung, sich positiv auf den Lernerfolg auswirkt (Wild, Hofer u. Pekrun 2001, 230). Besonders ungünstig ist in der Regel die Erklärung von Misserfolg durch mangelnde Begabung, da sich die Person damit eine gute Leistung gar nicht zutraut. Generell sind variable Ursachen und kontrollierbare Bedingungen günstigere Attributionen, da sie Veränderungsmöglichkeiten implizieren und ermöglichen, sich als wirksam und eigenständig zu erleben. Demgegenüber wird jemand, der Erfolg oder Misserfolg seines Lernens als fremdgesteuert und unkontrollierbar erlebt, wenig Wirksamkeit und Eigenständigkeit verspü-

ren und somit seinen Lernerfolg als wenig selbstbestimmt erleben (vgl. Kap. 4.5.2).

Diese Modellannahmen werden durch die Forschungsergebnisse von Carol Dweck et al. bestätigt, die sich um Einstellungen von Menschen zu ihren Entwicklungsmöglichkeiten drehen. Diese zeigen unter anderem:

- Verschiedene Studien von Mueller u. Dweck (1998) zeigten: Werden Lernende nach Erledigen einer Aufgabe mit Bezug auf ihre Begabung gelobt, so sinkt nicht nur ihre Anstrengungsbereitschaft, sondern auch ihre intrinsische Motivation, also die Freude am Lernen um seiner selbst willen. Demgegenüber steigen Anstrengungsbereitschaft und intrinsische Motivation, wenn das Lob mit Bezug auf die erfolgte Anstrengung erfolgt („du hast gewiss lange und intensiv an dieser Lösung gearbeitet“!).

Diese Befunde zeigen, dass Lob für nicht veränderbare Eigenschaften wie Begabung Menschen dazu führt, ihren Selbstwert an die Bedingung der eigenen Begabung zu binden. Misserfolge bedrohen dadurch schnell das eigene Selbstwertgefühl, da sie die Bedingung für das Selbstwertgefühl in Frage stellen (Burhans u. Dweck 1995). Auch Anstrengung wird als Folge eines bedingten Selbstwertgefühls problematisch – der Erfolg soll ja Resultat der Begabung und nicht von harter Arbeit sein. Entsprechend suchten für ihre Intelligenz gelobte Lernende auch leichtere Aufgaben, um die Erfolgswahrscheinlichkeit zu steigern. Diese bieten aber weniger Gelegenheit zum Lernen als herausfordernde Aufgaben. Umgekehrt steigert die Überzeugung, der Erfolg des eigenen Lernens sei durch das eigene Verhalten steuerbar, die Bereitschaft, anhand herausfordernder Aufgaben dazuzulernen. Letzteres gilt insbesondere in Gebieten, in denen hohes persönliches Interesse besteht (Stavrova, Pronk u. Kokkoris 2019).

- In eine ähnliche Richtung gehen Untersuchungen zum Feedbackverhalten von Mathematiklehrpersonen (Rattan, Good u. Dweck 2012). Diese zeigen, dass Lehrpersonen, welche die mathematische Intelligenz für unveränderbar halten, Lernende eher als mathematisch unbegabt einschätzen als Lehrpersonen, welche die mathematische Intelligenz für entwicklungsfähig halten. Die Lehrpersonen mit fixem Intelligenzbild trösten die Lernenden bei schlechten Resultaten auch eher für ihre tiefe Begabung und geben ihnen weniger oder anspruchslosere Hausaufgaben. Studierende wiederum nehmen die tieferen Erwartungen wahr, die mit tröstendem Feedback einhergehen, und reagieren darauf mit geringerer Motivation und reduzierten Erwartungen an den eigenen Lernerfolg. Demotivierender Trost kann z. B. lauten: „Nicht jeder kann gut sein in Mathe“, oder „viele Lernende haben mit diesem Fach Mühe, sind aber dafür in anderen Gebieten erfolgreich“ (Rattan, Good u. Dweck 2012, 734).

- In der Studie von Robins u. Pals (2002) reagieren College-Studierende, die ihre Intelligenz als unveränderbar betrachten, ganz anders auf Herausforderungen und Misserfolge als solche, die Intelligenz als entwicklungsfähig erachten. Jene mit einem fixen Intelligenzbild begegnen Schwierigkeiten und Rückschlägen mit einem hilflosen Reaktionsmuster und einer Tendenz zum Aufgeben. Demgegenüber reagieren Studierende mit einem entwicklungsorientierten Intelligenzbild mit einem lernorientierten Reaktionsmuster und der Tendenz, sich umso mehr anzustrengen. Interessanterweise hinterfragen sie dabei nicht nur das Ausmaß ihrer Anstrengung, sondern auch ihre Lernstrategien und sind bereit, neue Vorgehensweisen auszuprobieren. Dies entspricht exakt den oben formulierten Modellannahmen. Zudem sinkt das Selbstwertgefühl der Studierenden mit fixem Intelligenzbild im Laufe des Studiums, während sich jenes der Studierenden mit entwicklungsorientiertem Intelligenzbild steigert.¹¹⁴ Dieser letzte Befund lässt vermuten, dass ein fixes Intelligenzbild ebenso wie oben für Lob beschrieben zu einem bedingten Selbstwertgefühl führt, das durch Misserfolge bedroht wird.

All diese Befunde beziehen sich auf die Begründung von Erfolgen oder Misserfolgen. Sie dürfen nicht dahingehend missverstanden werden, dass nur Positives oder Gutes rückgemeldet werden soll. Wir sollten Lernende auch nur dann für ihr Engagement loben, wenn wir uns sicher sind, dass sie sich angestrengt haben.

Entscheidend ist die Haltung als rückmeldende Person: Sehe ich Fehler als etwas Negatives, Beschämendes und/oder schließe ich aus vielen Fehlern auf stabile Persönlichkeitsmerkmale wie Faulheit oder Dummheit? Oder sehe ich Fehler als Chance, sich über fachliche Herausforderungen auszutauschen, und als Lerngelegenheit? Traue ich den Studierenden das Lernen aus Fehlern zu? Wenn für Studierende aus Rückmeldungen die zweitgenannte Haltung ersichtlich wird, werden sie wesentlich positivere Auswirkungen auf ihr weiteres Lernverhalten haben.

Zudem ist diese Haltung auch lernpsychologisch angemessen: Wie in den Kapiteln 4.4 und 4.7 gezeigt wurde, lernen Menschen ganz wesentlich durch die Korrektur von Fehlern, während nur rund 2 % der individuellen Unter-

114 Interessanterweise hatten in der untersuchten Kohorte jene Studierenden mit einem fixen Intelligenzbild beim Übertritt ins Studium zwar die besseren Werte in einer unabhängigen Leistungsmessung (SAT-Score) als jene mit einem entwicklungsorientierten Bild, erzielten aber weder in der High School noch im College bessere Abschlussnoten. Die Studienautoren folgern daraus, dass sie ihre größeren schulischen Grundfähigkeiten nicht in bessere Schul- oder Studierleistungen umsetzen konnten – wohl aufgrund ihres hilflosen Reaktionsmusters auf Rückschläge und des geringeren Selbstwertgefühls (vgl. Robins u. Pals 2002, 322f.).

schiede im Bildungserfolg mit genetischen Ursachen erklärt werden können (Rietveld et al. 2013). Oder wie es Hattie u. Zierer (2017, 147) ausdrücken: Fehler sind die Grundlage für Wachstum!

Abschließend sei ein Beispiel dafür gegeben, wie man ein positives Feedback ohne Lob formulieren kann. Es basiert auf dem Schema Beobachtung – Interpretation – Wirkung – Wunsch/Anliegen von Thomann u. Pawelleck (2013, 100–102): Beobachtung: „Sie haben in Ihrer Fallbearbeitung nicht nur die häufigsten Auswirkungen von Gender Bias in Feedbackgesprächen erwähnt, sondern auch die vermuteten psychologischen und sozialen Ursachen inklusive Forschungsstand erläutert. Dabei haben sie auch Quellen erwähnt, die wir in der Vorlesung nicht behandelt haben.“ Interpretation: „Dies zeigt, dass sie sich nicht nur intensiv mit dem Thema des Gender Bias auseinandergesetzt haben, sondern auch an den wissenschaftlichen Erklärungen sowie am aktuellen wissenschaftlichen Diskurs interessiert sind.“ Wirkung: „Deshalb habe ich Ihnen für diese Aufgabe die volle Punktzahl gegeben.“ Wunsch/Anliegen: „Behalten Sie Ihre wissenschaftliche Neugier und führen Sie Ihre vernetzende Arbeitsweise fort.“

8.2.3 Beziehungsebene

In Rückmeldungen geht es auch immer um die Beziehungsebene. Es ist wichtig, dass Lehrende die Anstrengungen der Studierenden wertschätzen. Dies gilt auch für Peers, die einander Feedback geben. Diese Wertschätzung stärkt die Beziehungsebene, mangelnde Wertschätzung schwächt sie (vgl. Hattie u. Zierer 2017, 135).

Positiv auf die Beziehungsebene wirken sich auch Freundlichkeit und ein gesunder Humor aus. So wirkt sich etwa Humor am Arbeitsplatz substantiell positiv auf die Arbeitsleistung aus – er erklärt etwas mehr als 10 % der Unterschiede in der Arbeitsleistung (vgl. Hattie u. Zierer 2017, 136).

Dass Freundlichkeit und positive Einstellung ansteckend wirken, hat unter anderem mit dem Chamäleon-Effekt zu tun, der neurowissenschaftlich mit den so genannten Spiegelneuronen erklärt wird: Wir Menschen passen unser Verhalten unbewusst an das unserer sozialen Umgebung an (vgl. Hattie u. Zierer 2017, 137). Dieser Effekt gilt gleichermaßen für negative Verhaltensweisen wie die Tendenzen, andere bloßzustellen oder Wutausbrüche zu haben. Umso wichtiger ist es, als Lehrperson im Beziehungsverhalten mit gutem Beispiel voranzugehen und so die Grundlage für ein positive, lernförderliche Stimmung im Hörsaal oder Seminarraum zu schaffen.

8.2.4 *Eigenschaften einer motivationsförderlichen Rückmeldung*

Aus den berichteten Befunden lassen sich Elemente von Leistungsrückmeldungen ableiten, die sich positiv auf die Lernmotivation, das Selbstkonzept und somit auf das weitere Lernverhalten auswirken. Wenn mit Peer-Feedback gearbeitet wird oder wenn Feedback durch externe Expertinnen, Praktikumsleitende etc. erfolgt, kann es sich lohnen, diese Aspekte mit den Studierenden oder den externen Beteiligten zu besprechen:

- Lob ist keinesfalls ein zwingender Bestandteil von Leistungsrückmeldungen. Erstens sollte nur etwas gelobt werden, was tatsächlich vorhanden ist (z. B. eine sichtbare Anstrengung). Zweitens ist eine intensive inhaltliche Auseinandersetzung mit den fachlichen Leistungen der Studierenden für diese meist in sich motivierend genug. Denn dabei werden mit der sozialen Eingebundenheit und der Wirksamkeit zwei von drei zentralen Dimensionen der Selbstbestimmung angesprochen (vgl. Kap. 4.5.2).
- Falls Lob erteilt wird, sollte sich dieses nicht auf die Richtigkeit einer Lösung und schon gar nicht auf unveränderbare Eigenschaften wie Begabung und Intelligenz richten, sondern wenn schon auf die Anstrengung oder den geschickten Einsatz von Lernstrategien.
- Aufgrund des Pygmalioneffekts neigen Studierende dazu, Attribuierungen von Lehrenden zu übernehmen (Kap. 8.2.1). Deshalb ist es wichtig, möglichst auf Signale zu verzichten, die Studierende schließen lassen können, man würde sie im jeweiligen Fachgebiet für nicht entwicklungsfähig halten. Noch besser ist es, tatsächlich als Lehrende an die Entwicklungsfähigkeit aller Studierenden zu glauben!
- Formative Rückmeldungen sollten stets erfolgen, bevor die Aktivität, zu der die Rückmeldung erfolgt, bewertet wird. Denn sobald Studierende Noten oder Prädikate wie bestanden/nicht bestanden erhalten haben, sind sie am inhaltlichen Feedback nicht mehr interessiert und der Aufwand der Lehrenden ist vergeblich (dieser Aspekt wird in Kapitel 9.1 näher begründet).
- Entscheidend ist, durch Rückmeldungen den Studierenden Entwicklungsmöglichkeiten aufzeigen – unabhängig von ihrem Leistungsstand und dem Tempo ihres Lernfortschritts.

8.3 Inhaltliche Eigenschaften von Leistungsrückmeldungen

Die positive motivationale Wirkung einer Rückmeldung ist nur die notwendige, aber noch keine hinreichende Voraussetzung, um das weitere Lernen zu fördern. Dazu sind selbstverständlich inhaltliche Informationen nötig: Feedback muss Informationen bieten, welche die „Lücke“ schließen helfen zwischen dem Ort, wo ein Lernender steht, und dem Ort, wo er hin soll (vgl. Hattie u. Clarke 2019, 4). Mit der zeitlichen Perspektive der Rückmeldung (Kap. 8.3.2), den verschiedenen inhaltlichen Bezugsebenen (Kap. 8.3.3) und der Orientierung an Leistungskriterien (Kap. 8.3.4) werden drei zentrale inhaltliche Gesichtspunkte erläutert. Darüber hinaus spielen auch die Ausführlichkeit und die sprachliche Komplexität, in der die Inhalte der Rückmeldung formuliert werden, eine wesentliche Rolle für die Wirkung von Feedback (Kap. 8.3.5).

8.3.1 Vorbedingungen für inhaltliches Feedback

Um eine studentische Leistung sinnvoll – ob formativ oder summativ – beurteilen zu können, müssen im Minimum drei Voraussetzungen gegeben sein (vgl. Biggs u. Tang 2007, 187):

- Es müssen spezifische Lernziele definiert sein (siehe Kap. 5.4).
- Es muss entschieden werden, wie Studierende zeigen können, dass sie diese Lernziele erreicht haben (siehe Kap. 7).
- Es müssen auf die Lernziele bezogene Qualitätskriterien definiert sein, die es ermöglichen, gezeigte Leistungen daraufhin zu beurteilen, inwieweit die adressierten Lernziele erreicht wurden (siehe Kap. 8.3.4).

Traditionell sind die Studierenden an der Ausarbeitung dieser drei Voraussetzungen nicht beteiligt. Lernziele sind oft auch den Lehrenden vom Curriculum zumindest in Teilen vorgegeben. Im Rahmen der vorhandenen Spielräume sollen Studierende aber in die Erarbeitung aller drei Voraussetzungen so weit als möglich eingebunden werden (siehe Kap. 9.3.4). Zudem ist es für den Lernprozess wesentlich, Studierende auch als Feedbackinstanzen einzubeziehen (siehe Kap. 8.4).

8.3.2 Zeitliche Perspektive der Rückmeldung

Die kommentierte Lernleistung liegt zwar stets in der Vergangenheit. Aber die zeitliche Perspektivierung der Rückmeldung kann unterschiedlich sein – und sie ist entscheidend (vgl. für die folgende Aufzählung Hattie u. Zierer 2017, 87):

- **Vergangenheitsorientierte Rückmeldung:** Hier wird der aktuelle mit dem früheren Lernstand der Studierenden verglichen. Eine solche Rückmeldung zu einem formativen Leistungstest in Form einer Problemanalyse würde etwa die angemessenen Analyseansätze, korrekt durchgeführte Analyseschritte und noch fehlende oder inkorrekte Aspekte abgleichen mit der letzten Problemanalyse der gleichen Person: Was hat sich verbessert, was nicht?
- **Gegenwartsorientierte Rückmeldung:** In diesem Fall wird der aktuelle Lernstand der Studierenden mit dem angestrebten Zielzustand verglichen, also mit den durch Kriterien definierten Elementen der angestrebten Lernziele (siehe Kap. 9.3). Eine solche Rückmeldung zu einem formativen Leistungstest in Form einer Problemanalyse würde etwa die angemessenen Analyseansätze, korrekt durchgeführte Analyseschritte und noch fehlende oder inkorrekte Aspekte abgleichen mit der Kriterial in Bezug auf die Lernziele erwarteten Leistung: Was ist schon da, was fehlt noch?
- **Zukunftsorientierte Rückmeldung:** Darin wird der angestrebte Zielzustand basierend auf dem gegenwärtigen Lernstand der Studierenden illustriert. Eine zukunftsorientierte Rückmeldung zu einem formativen Leistungstest in Form einer Problemanalyse würde etwa die am Ende erwarteten Analyseansätze und korrekt durchgeführten Analyseschritte erläutern und erklären, was noch zu tun ist, um dorthin zu gelangen. Dies könnten etwa Hinweise auf nicht berücksichtigte Analysekonzepte, falsch oder unvollständig verstandene Elemente von verwendeten Analysekonzepten oder Analyseschritten sein, verbunden mit Aussagen dazu, wo diese nachgeschlagen, wie sie geübt oder eingeordnet werden können.¹¹⁵

Wenn Sie Feedback zum Stand einer Arbeit erhalten: Welche dieser Perspektivierungen würde Ihnen am meisten nützen für Ihr weiteres Vorgehen? Studierende finden in der Regel vor allem Feedback wertvoll, das ihnen möglichst konkrete Hinweise darauf gibt, wie sie ihr weiteres Lernen gestalten können – also zukunftsorientierte Rückmeldungen (Hattie u. Zierer 2017, 81).

115 Diese Perspektivierung von Feedback wird gelegentlich auch Feedforward genannt. Auf diesen oft verwirrenden Begriff wird hier aber verzichtet.

Black u. Wiliam (1998, 53) kommen in ihrer einflussreichen Analyse der Forschungsliteratur zu formativer Beurteilung sogar zum Schluss, dass einzig zukunftsorientierte Rückmeldungen formativ wirken. Hattie u. Clarke (2019, 135) gelangen angesichts der Visible-Learning-Studie zum gleichen Ergebnis. Ihnen zufolge sind Rückmeldungen, die keine Hinweise zum weiteren Vorgehen enthalten, ebenso nutzlos wie ein didaktisches Setting, in dem die Studierenden keine Gelegenheit erhalten, die in der Rückmeldung erhaltenen Hinweise umzusetzen.

Ein Grund für die Wichtigkeit von zukunftsorientiertem Feedback könnte in einer anderen zeitlichen Bedingung liegen: Verschiedene Studien zeigen, dass viele Lernende, auch auf der Hochschulstufe, erst für instruktionale Unterweisung zugänglich sind, wenn sie allein (oder auch mit Peers) nicht mehr weiterkommen (Kapur 2012; Westermann u. Rummel 2012; VanLehn et al. 2003). Erst in solchen Momenten sind sie wirklich offen für Feedback, das ihnen Hinweise gibt, wie sie beim Problemlösen oder Lernen vorankommen. Oft genügen dabei metakognitive Anregungen, und nur in einigen Fällen bewirken inhaltliche Erklärungen tatsächlichen Lerngewinn (Chi et al. 2001, 508–17; VanLehn et al. 2003, 245f.). Die von VanLehn et al. (ebd.) aus diesen Befunden abgeleitete Strategie ist deshalb auch für zukunftsorientiertes Feedback zu empfehlen:

1. Im ersten Schritt fordern wir Lernende, die an einer Stelle nicht mehr weiterkommen (oder einen wesentlichen, nicht oberflächlichen Fehler gemacht haben) anhand metakognitiver Fragen auf, den angemessenen nächsten Schritt oder die richtige Lösung zu finden, umzusetzen und anschließend zu erklären, weshalb das Vorgehen funktioniert hat. Metakognitive Fragen können z. B. lauten:
 - „Können Sie diesen Lösungsschritt in Verbindung bringen mit Vorgehensweisen, die Sie in den vorangehenden Wochen gelernt haben?“ (Monitoring des eigenen Lernens, Verknüpfung von Schemata)
 - „Welche Lösungsschritte oder welche Elemente des Lösungsschrittes sind Ihnen unklar?“ (Bewusstmachen des eigenen Vorgehens)
 - „Welche Strategien haben Sie bei ähnlichen Problemen in früheren Situationen benutzt?“ (Strategien identifizieren)
2. Eine inhaltliche Erklärung sollten wir nur dann geben, wenn die Lernenden den richtigen Schritt partout nicht finden oder erfolglos versucht haben, eine eigene Erklärung für die richtige Lösung zu geben.

Dieses Vorgehen ist alles andere als trivial. So zeigen diverse Untersuchungen, dass Lehrende dazu tendieren, viel zu früh und zu viele inhaltliche Erklärungen zu geben – zu einem Zeitpunkt, an dem die Studierenden nicht bereit sind, diese aufzunehmen und zu verarbeiten (VanLehn et al. 2003, 245; Chi et al. 2001, 512–17).

8.3.3 Inhaltliche Bezugsebene der Rückmeldung

Rückmeldungen können sich auf verschiedene Ebenen beziehen: Das Selbst, also die Person des Adressaten, die bearbeitete Aufgabe, den Lösungsprozess und die Selbstregulation der Lernenden (vgl. Hattie u. Timperley 2007, 90–97). Sie sollten sich auf allen Ebenen explizit auf die Lernziele und die Qualitätskriterien für deren Erreichen beziehen:

- Selbst-Ebene: Diese zielt auf die Person der Lernenden. Hier wird häufig Lob (und seltener Kritik) an der Person ausgedrückt, wie die oft gehörten Quittierungen „gut gemacht“ oder „super“.

Rückmeldungen auf der Selbst-Ebene unterminieren meist die Motivation, außer sie beziehen sich auf eine tatsächlich erfolgte Anstrengung oder eine geschickte Selbstregulation des Lernens (siehe Kap. 8.2.2).

- Aufgabenebene: Diese Ebene wird in Rückmeldungen von Lehrenden am häufigsten adressiert (leider oft vermischt mit Lob auf der Selbst-Ebene, was durch die ungünstige motivationale Wirkung das Feedback ineffektiv macht). Wirkungsvoll ist Feedback auf dieser Ebene vor allem, wenn es fehlerhafte Anwendungen, Interpretationen u. Ä. adressiert.

Fehlt nötiges Wissen, ist weitere Instruktion hilfreicher als lange Rückmeldungen. Auf Hochschulstufe eignen sich dazu oft Hinweise auf zu studierende Materialien; sollten viele Studierende etwas nicht wissen, was eigentlich behandelt wurde, empfiehlt es sich, in der Lehrveranstaltung nach den Gründen zu fragen. Je nachdem kann das Thema dann nochmals adressiert werden.

- Prozessebene: Hier geht es um Rückmeldungen dazu, wie Studierende die Aufgabe bearbeitet und gelöst haben. Dies sind z. B. Hinweise zu fachspezifischen Lösungsprozessen oder zu Strategien für das Entdecken und Korrigieren von Fehlern.

Feedback auf der Prozessebene zeigt oft mehr Wirkung in der Förderung von tiefenorientierten Lernstrategien als aufgabenbezogene Rückmeldungen. Am wirkungsvollsten scheint freilich die Kombination beider Ebenen zu sein.

- Selbstregulationsebene: Selbstregulierung ist ein Zusammenspiel aus Selbstbeobachtung, Selbsteinschätzung und darauf basierender Steuerung der eigenen Lernhandlungen. Sie hängt mit Vertrauen in die eigenen Fähigkeiten und einem darauf basierenden Autonomieerleben zusammen und befähigt unter anderem auch zu Selbstdisziplin (vgl. Perez, Huber u. Geissler 2001, 384f.). Der Wert von Feedback zur Selbstregulation liegt darin, dass es die Studierenden befähigt, mehr in die Handlungsrolle zu kommen und sich als Subjekt der nötigen Lernhandlungen zu verstehen (Hattie u. Zierer 2017, 83).

Um die Selbstregulation zu adressieren, eignen sich metakognitive Hinweise auf das Lern- und Studieverhalten. Scheint etwa ein Aspekt oberflächlich bearbeitet, empfiehlt sich ein Hinweis im Sinne von „Schauen Sie sich die Analyseschritte von Prozess XY noch einmal genauer an. Sind in Ihrer Analyse alle Schritte enthalten?“. Auf einer allgemeinen Ebene können auch Hinweise zur Überprüfung/Überarbeitung einer Aufgabenbearbeitung vor ihrer Einreichung oder die Ermutigung zu eigenen Beurteilungen und Einschätzungen des Lernprozesses oder Produkts sinnvoll sein, etwa: „Ich schlage Ihnen vor, am Ende Ihrer nächsten Analyse eine Einschätzung der gefundenen Resultate und ihrer Bedeutung vorzunehmen.“

Auch wenn jede Rückmeldungsebene ihre spezifische Wirkung hat, sind Rückmeldungen dann am wirksamsten, wenn sie darauf zielen, die Studierenden von der Aufgaben- auf die Prozessebene und von dort zur Selbstregulation zu bringen. Zu viel oder gar ausschließliche Betonung einer Ebene kann sich hingegen sogar negativ auf die Lernentwicklung auswirken (Hattie u. Timperley 2007, 91). Wie die in Kap. 8.3.2 berichteten Befunde zeigen, sind oft bewusste Anstrengungen nötig, um Feedback zu geben, das auf die Prozess- und Selbstregulationsebene zielt und somit am meisten Wirkung hat.

8.3.4 Orientierung an Kriterien

Es ist essenziell, dass sich formative Leistungsrückmeldungen stets auf die Qualitätskriterien zur Erreichung der Lernziele beziehen (zum Formulieren dieser Kriterien siehe Kap. 9.3). Dies ist im Sinne der Zukunftsorientierung zentral für die Studierenden, um ihr Lernverhalten auf die noch zu erreichenden Ziele auszurichten. Denn die spezifischen Lernziele (siehe Kap. 5.4) bilden zusammen mit den Qualitätskriterien für ihr Erreichen den Bezugsrahmen für die Lernentwicklung. Empirische Studien zeigen zudem, dass Studierende die Kriterien oft erst verstehen, wenn sie Gelegenheit erhalten, sich näher mit ihnen auseinanderzusetzen und sich über sie zu verständigen. Wie man dazu vorgehen kann, wird in Kapitel 9.3.4 aufgezeigt.

Hattie u. Timperley (2007, 89) weisen darauf hin, dass Rückmeldungen in der Praxis häufig keinen Bezug zu den entscheidenden Dimensionen eines Lernziels aufweisen. Ihr Beispiel ist prototypisch: Besteht das Lernziel darin, in einer Geschichte eine bestimmte Stimmung zu erzeugen, ist eine Rückmeldung nicht lernwirksam, die sich auf die Textstruktur, die Rechtschreibung und die Textlänge bezieht. Solches Feedback hilft schlicht nicht, die oben erwähnte Lücke zwischen dem Lernstand zu schließen, wo eine Lernende steht, und dem, wo sie hinsoll. Hilfreich ist in diesem Zusammenhang die Aufforderung an die Studierenden, gemeinsam mit der Abgabe ihrer

Arbeit (an Lehrende oder Peers) wesentliche Punkte zu formulieren, zu denen sie sich Feedback wünschen (siehe Kap. 8.5.1).

Die empirische Forschung zeigt zudem, dass Aufgaben und das folgende Feedback dann am lernwirksamsten sind, wenn sie herausfordernd, aber nicht überfordernd sind. Angesichts des oft heterogenen Vorwissens der Studierenden ist dies aber kaum in einer Weise umzusetzen, dass die Aufgabe für alle Teilnehmenden einer Veranstaltung das optimale Anspruchsniveau erreicht. Eine Lösung kann sein, Aufgabenstellungen mit verschiedenen Schwierigkeitsstufen zu formulieren, so dass für alle eine herausfordernde Stufe dabei ist.¹¹⁶

8.3.5 Ausführlichkeit und sprachliche Komplexität der Rückmeldung

Die zahlreichen Hinweise zur Gestaltung von formativen Rückmeldungen bedeuten nicht, dass diese lang zu sein brauchen. Im Gegenteil: Forschungsergebnisse wie praktische Erfahrungen zeigen, dass zu ausführliches Feedback die Studierenden oft überfordert oder demotiviert. Zu ausführliche Rückmeldungen sind also sehr unökonomisch: Sie kosten die Lehrenden viel Zeit und fördern das Lernen der Studierenden kaum. Folgende Formulierungstipps helfen, den Wirkungsgrad der Rückmeldung zu erhöhen. Diese Hinweise gelten für gesprochene wie geschriebene Sprache und für alle Medien; auf mediale Eigenheiten wird in Kapitel 8.5 näher eingegangen:

- Weniger wirkt mehr: Fachleute empfehlen, sich in einer einzelnen Rückmeldung auf zwei bis drei Aspekte zu beschränken, zu denen nähere Hinweise gegeben werden (z.B. Hassler 2022, 155; Nicol 2010, 512; Elbow u. Sorcinelli 2006, 202). Mehr wirkt auf die meisten Studierenden überfordernd oder demotivierend und reduziert den Wirkungsgrad. Wichtig ist, dass diese Aspekte sich konkret auf bestimmte Stellen oder Eigenschaften der kommentierten Arbeit bzw. des kommentierten Produkts beziehen (Nicol 2010, 512).
- Die ausgewählten Aspekte sollten in einer dem Lernstand und den Studierenden angemessenen Ausführlichkeit rückgemeldet werden. Das bedeutet ausreichend, aber nicht zu viele Hinweise darauf, was schon gelungen ist und wo noch optimiert werden kann. Wirksam sind diese Hinweise, wenn sie an Stellen der Denkstruktur ansetzen, die für die Studierenden

116 Das kann aber wieder auf jene demotivierend wirken, die mit den anspruchsvolleren Stufen Mühe haben, während durch die Bearbeitung komplexer Aufgaben just jene gefördert werden, die sowieso schon am weitesten fortgeschritten sind.

anschlussfähig sind. Zum Beispiel weist man Studierende, die ihre erste wissenschaftliche Arbeit schreiben, besser auf übergeordnete Aspekte der Struktur wissenschaftlicher Arbeiten hin als auf Feinheiten einzelner Textkomponenten. Am besten eignen sich zudem Stellen/Aspekte der studentischen Arbeit, in denen sich klar zeigt, dass die Studierenden nicht mehr weiterkommen oder einen ungeeigneten/falschen Ansatz gewählt haben – hier ist es am wahrscheinlichsten, dass Studierende das Feedback auch für ihre weiteren Lernaktivitäten berücksichtigen (siehe Kap. 8.3.2).

- Eine für die Studierenden verständliche Sprache ist wichtig. Fachausdrücke wirken distanzierend und sollten sparsam verwendet werden. Angebracht sind sie dann, wenn ihr Verständnis vorausgesetzt werden kann und sie für das Verstehen der Rückmeldung notwendig sind (vgl. Elbow u. Sorcinelli 2006, 203).
- Die Rückmeldung sollte möglichst deskriptiv und nicht wertend formuliert werden. Hilfreich ist der fachliche Bezug auf Lernziele, während Bewertungen wie Noten oder Prädikate in formativen Rückmeldungen zu vermeiden sind (vgl. Nicol 2010, 512; Gibbs u. Simpson 2004, 18).
- Idealerweise stellt die Rückmeldung eine Balance zwischen Hinweisen auf Gelungenes und auf Verbesserungswürdiges her. Und Vorsicht: Bei Hinweisen auf Gelungenes ist auf Lob zu verzichten (siehe Kap. 8.2.2)!

8.3.6 *Inhaltliche Qualitätskriterien für formative Rückmeldungen*

Die berichteten Befunde lassen sich zu folgenden inhaltlichen Qualitätskriterien für formative Leistungsrückmeldungen zusammenfassen, die das weitere Lernen und die Selbstregulation des Lernens fördern. Wenn mit Peer-Feedback gearbeitet wird oder wenn Feedback durch externe Expertinnen, Praktikumsleitende etc. erfolgt, kann es sich lohnen, diese Aspekte mit den Studierenden oder den externen Beteiligten zu besprechen. Zudem ist zu beachten, dass die inhaltlichen Qualitäten von Rückmeldungen nur zum Tragen kommen, wenn die motivationalen Aspekte berücksichtigt sind (siehe Kap. 8.2):

- Formative Rückmeldungen sollen zukunftsorientiert sein: Sie sollen den Studierenden möglichst konkrete Hinweise darauf geben, wie sie ihr weiteres Lernen gestalten können, um die Lernziele zu erreichen. Am erfolgversprechendsten scheinen dabei metakognitive Hinweise zu sein.
- Sie sollten die Aufgabenebene nur so weit als nötig adressieren und wenn immer möglich die Prozess- und die Selbstregulationsebene ansprechen, um die Studierenden dabei zu unterstützen, von der einen zur nächsten

Ebene zu gelangen. Die Selbst-Ebene sollte möglichst wenig angesprochen werden.

- Auch formative Rückmeldungen sollten sich auf transparent im Voraus kommunizierte Lernziele und Erfolgskriterien beziehen (die dann u.U. auch in späteren summativen Beurteilungen zur Anwendung kommen).
- Sie sollten sich auf zwei bis drei spezifisch, präzise und in möglichst einfacher Sprache formulierte Aspekte beschränken. Die Formulierungen sollten möglichst deskriptiv und nicht wertend sein.

8.4 Feedbackinstanzen

Leistungsrückmeldungen können durch verschiedene Instanzen erteilt werden. In der Regel denken wir aber zuerst an Rückmeldungen, die Lehrende Lernenden geben (Kap. 8.4.1). Allerdings müssen und sollen Leistungen nicht nur durch Lehrende formativ beurteilt werden. Es gibt sehr gute lernpsychologische Gründe, weshalb Peer-Feedback und Selbstbeurteilungen für zusätzlichen Lerngewinn sorgen (Kap. 8.4.2. und 8.4.3). Außerdem können weitere Feedbackinstanzen wie externe Experten (Kap. 8.4.4) und elektronische Tools inklusive künstliche Intelligenz wie ChatGPT einbezogen werden (Kap. 8.4.5). Wertvoll ist auch die Kombination von Rückmeldungen verschiedener Instanzen zu einem Mehrquellen-Feedback (Kap. 8.4.6).

8.4.1 Lehrende

Die bisherigen Ausführungen lassen sich gut als Hinweise für das Erteilen von Feedback durch Lehrende lesen, obschon alle Aussagen auch für solches durch andere Instanzen gelten. Bezüglich Lehrender als Feedbackinstanzen seien deshalb noch zwei Hinweise angebracht, die für den Einsatz anderer Feedbackinstanzen sprechen:

- Expertenblindheit: Mit zunehmender Expertise werden Wissen und Können automatisiert. Durch diesen Prozess nimmt das Bewusstsein dafür ab, wie sich einzelne Wissens- und Könnenselemente zusammensetzen (siehe Abb. 18). Vielleicht haben Sie schon einmal einem Kind erklärt, wie man einen Schnürsenkel bindet, und dabei gemerkt, dass Sie sich die einzelnen Schritte zuerst wieder bewusst machen mussten, bevor Sie sie erklären konnten? Dies wäre ein einfacher Fall von Expertenblindheit. Das Phänomen wird umso ausgeprägter, je komplexer Expertenwissen und

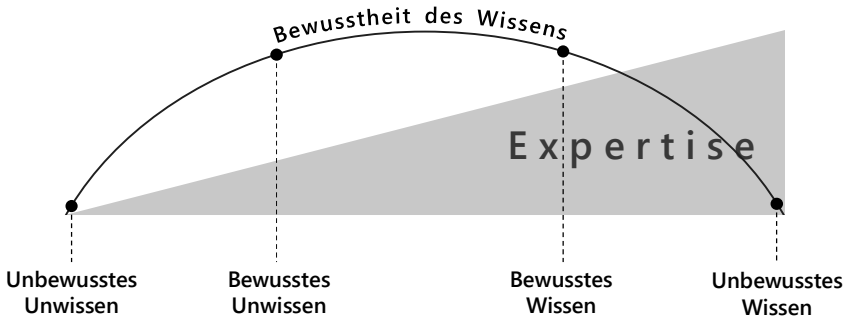


Abb. 18: Inverses Verhältnis von Expertise und Bewusstheit des eigenen Wissens.
 Quelle: Eigene Abbildung, in Anlehnung an Ambrose et al. 2010, 97

-können sind (vgl. Ambrose et al. 2010, 95–99; Nathan u. Petrosino 2003; Hinds 1999).

Für Lehrende, die Themen unterrichten, in denen sie über mehrjährige Forschungserfahrung verfügen, ist die Unbewusstheit ihres Expertenkönnens besonders herausfordernd. Peer-Feedback kann deshalb eine sinnvolle Ergänzung oder Alternative sein. Sinnvoll kann darüber hinaus der Einsatz von fortgeschrittenen Studierenden als Tutorinnen und Tutoren sein, da sie sich oft im Stadium der Wissensentwicklung befinden, das in Abbildung 18 als „bewusstes Wissen“ bezeichnet ist.

- **Ökonomie:** Auch zur Entlastung der Lehrenden ist es sinnvoll, alternative Feedbackinstanzen einzusetzen. Besonders effektiv ist dies, wenn sich Lehrende durch den Einsatz von anderen Feedbackinstanzen Zeitressourcen verschaffen können, um Studierenden ihr Feedback in solchen Momenten zu geben, in denen dieses besonders wirkungsvoll ist. Dabei sollten sie sich auf das Rückmelden weniger zentraler Punkte fokussieren – das erhöht die Wahrscheinlichkeit, dass die Studierenden diese erkennen und für ihr weiteres Lernen berücksichtigen (siehe Kap. 8.6.2).

8.4.2 Peers

Die oben berichtete Expertenblindheit ist ein Argument, das ex negativo für den Einsatz von Peer-Feedback spricht. Mindestens so wichtig sind aber zwei direkt positive Wirkungsweisen von Peer-Feedback: Die Wirkung auf die Feedbackempfangenden und die Wirkung auf die Feedbackgebenden.

Wirkung auf Feedbackempfangende

Peers können als Feedbackinstanz das Lernen ihrer Mitstudierenden fördern. Um dieses Potenzial nutzen zu können, erscheint es sinnvoll, mit den Studierenden nicht nur die wichtigsten Aspekte der bereits erläuterten motivationalen und inhaltlichen Eigenschaften und Wirkungen von Rückmeldungen zu besprechen, sondern auch Aspekte des Umgangs von Studierenden mit Peer-Feedback:

- Art der gesuchten Unterstützung: Ryan u. Shin (2011, 254) unterscheiden zwischen adaptiver und behelfsmäßiger Suche nach Unterstützung. Die adaptive Unterstützungssuche zielt auf die Unterstützung bei der Bearbeitung einer Aufgabe, etwa durch Erklärungen oder Beispiele (Hattie u. Clarke 2019, 97). Behelfsmäßige Unterstützungssuche zielt hingegen darauf, die „richtige“ Antwort zu erhalten oder gar abzuschreiben. Lernförderlich ist nur das Suchen von adaptiver Unterstützung.
- Der optimale Zeitpunkt für eine Rückmeldung durch Peers ist dann gegeben, wenn die feedbackempfangende Person die ersten beiden Stufen der SOLO-Taxonomie großenteils erreicht, also wesentliche Grundbegriffe und Konzepte verstanden hat (siehe Kap. 5.4.2): Dann ist sie bereit dafür, (weitere) Zusammenhänge zu erkennen und Verknüpfungen zwischen den neuen Wissensinhalten zu erstellen (dritte, relationale SOLO-Stufe). Vorher hat Feedback wie in Kapitel 8.3 erläutert in der Regel wenig Wirkung, und die Studierenden sollten sich besser nochmals mit dem deklarativen Grundlagenwissen auseinandersetzen.

Unter diesen Bedingungen kann Peer-Feedback oder auch Peer-Coaching sehr wirkungsvoll sein, indem es das tiefenorientierte Lernen unterstützt: „Eine wichtige Strategie besteht darin, dass die Lernenden zu Lehrenden für andere werden und von Peers lernen, da dies ein hohes Maß an Regulierung, Überwachung, Antizipation sowie ein Zuhören in Bezug auf die Auswirkungen [der eigenen Rückmeldungen] auf das Gegenüber erfordert“ (Hattie u. Donoghue 2016, 7, Übers. und Erg. tz).

Wirkung auf Feedbackgebende

Anderen Studierenden Rückmeldungen zu geben, fördert aber auch das Lernen der Feedbackgebenden. Dafür gibt es mehrere lernpsychologische Erklärungen:

- Anderen eine Rückmeldung zu einem Thema zu geben, an dem man selbst gearbeitet hat, bedingt eine neuerliche Auseinandersetzung mit den Inhalten – sie müssen erneut aus dem Gedächtnis abgerufen werden. Aus der Gedächtnisforschung ist bekannt, dass das Abrufen von Gedächtnis-

inhalten ihre „Abspeicherung“ im Langzeitgedächtnis deutlich verstärkt (Roediger u. Karpicke 2006, siehe auch Kap. 8.4.5).

- Um ihren Peers Rückmeldungen geben zu können, müssen sich die Studierenden vertieft mit den Lernzielen und vor allem den Qualitätskriterien für deren Erreichen auseinandersetzen. Durch die aktivere Beschäftigung mit diesen Kriterien und die darauf basierenden Beurteilungsprozesse verarbeiten sie diese tiefer. Dies wirkt sich auch auf ihre Fähigkeit zur Einschätzung eigener Leistungen positiv aus. Damit verbunden ist ein weiterreichendes Verständnis dafür, was gute Belege für das Erreichen von Lernzielen und Qualitätskriterien ausmacht (Biggs u. Tang 2007, 187).
- Durch das Einschätzen von Leistungen ihrer Peers kommen die Studierenden stärker in eine handelnde Rolle, als wenn Beurteilungen – auch formative – nur durch Lehrende erfolgen (Biggs u. Tang 2007, 188). Im Sinne der Selbstbestimmungstheorie stärkt dies die intrinsische Motivation, da sich die Studierenden selbst als wirksam, sozial eingebunden und autonom erleben können (siehe Kapitel. 4.5.2).

Ganz ähnliche Vorteile gelten auch für Selbstbeurteilungen, wie sie im nachfolgenden Kapitel erläutert werden.

8.4.3 Selbstbeurteilungen

„Bei Selbstbeurteilungen werden die eigenen Prozesse und Produkte überprüft, um Anpassungen vorzunehmen, die das Lernen vertiefen und Leistung verbessern“ (Andrade 2019, 10, Übers. tz). Umfangreiche empirische Belege zeigen, dass Selbstbeurteilungen das Lernen am besten fördern, wenn sie formativ eingesetzt werden. Sie fördern dann sowohl das Lernen an sich als auch die Selbstregulation des Lernens (vgl. ebd.). Dennoch können Selbstbeurteilungen auch im Rahmen von summativen Beurteilungen zum Einsatz kommen, wie in Kapitel 9.6.2 gezeigt wird.

Konfrontiert mit der Idee, dass Studierende ihr Lernen selbst beurteilen sollten, reagieren viele Lehrende mit Skepsis: Können Studierende ihren Lernfortschritt überhaupt zuverlässig beurteilen? Missbrauchen sie Selbstbeurteilungen nicht, um ihre Noten aufzubessern? Es gibt viel empirische Forschung, die diese Fragen zu einem großen Teil beantwortet. Ein Problem liegt zunächst einmal darin, dass, wie in Kapitel 2.3 aufgezeigt, bereits die Beurteilungen und insbesondere die Notenbewertungen von Lehrenden sehr unzuverlässig sind und Urteile von Lehrenden untereinander schlecht korrelieren.

Dennoch ist empirisch belegt, dass die Konsistenz zwischen Selbstbeurteilungen der Lernenden und jenen der Lehrenden mit zunehmendem Alter steigt. Selbstbeurteilungen sind somit an Hochschulen am einfachsten einzu-

führen, was auch damit zusammenhängt, dass Studierende eher als jüngere Lernende ihren Nutzen und ihre Wirkung verstehen (vgl. Andrade 2019, 6–8). Die Konsistenz der Selbstbeurteilungen kann zudem erhöht werden durch den Einsatz von weiterem (z. B. Peer) Feedback, die Formulierung und Besprechung von Leitlinien für das Vorgehen und klaren, möglichst konkret auf die jeweilige Aufgabe anwendbaren Beurteilungskriterien (vgl. Andrade 2019, 6; Fastré, van der Klink u. van Merriënboer 2010).

Entscheidend ist allerdings: Die lernförderliche Wirkung von Selbstbeurteilung liegt nicht primär in der Frage nach Konsistenz der Selbsteinschätzungen zu Fremdeinschätzung – sowieso nicht bei formativem Einsatz. Vielmehr zeigt die vorhandene Evidenz klar auf, dass formative Selbstbeurteilungen mit einem erhöhten Lernzuwachs einhergehen (Andrade 2019, 8). Dies ist mit vielen bereits erläuterten kognitiven und motivationalen Prozessen erklärbar. Dazu gehören etwa motivationale Aspekte: In den Beurteilungsprozess einbezogen zu werden, erhöht das Selbstbestimmungsgefühl der Studierenden (Kap. 4.5.2). Durch den Abgleich von Selbst- und Fremdbeurteilungen lernen sie, ihr Wissen und Können sowie mittelbar auch die Wirkung ihrer Lernhandlungen und Lernstrategien besser einzuschätzen, was der Selbstregulation des Lernens förderlich ist (Kap. 8.3.3). Durch die intensivere Auseinandersetzung mit dem eigenen Wissen und Können fördern Selbstbeurteilungen zudem die Metakognition, die ebenfalls entscheidend ist für die Steuerung des Lernens sowie für das Herstellen von Zusammenhängen zwischen verschiedenen kognitiven Schemata (Kap. 4.4.3).¹¹⁷

So kommt Boud zum Schluss: „Bei der Selbstbeurteilung geht es darum, dass Studierende ihre Lernfähigkeiten weiterentwickeln. Sie ist nicht einfach ein weiteres Beurteilungsformat neben anderen. Es geht darum, dass sich die Studierenden mit Kriterien für gute Leistungen und mit dem Fällen komplexer Urteile auseinandersetzen“ (1995, 17, Übers. tz). Dabei brauchen Studienanfänger in der Regel mehr Unterstützung als fortgeschrittene Studierende. So kann es sinnvoll sein, Studierenden in den ersten Semestern insbesondere des Bachelor-Studiums aufgabenspezifische Fragen zur Selbstbeurteilung bereitzustellen, während weiter fortgeschrittene Studierende eher mit generischen Hinweisen arbeiten können. Für beide Optionen werden nachfolgend Beispiele gegeben.

117 Das genaue Zusammenspiel und der unterschiedliche Einfluss der verschiedenen Faktoren, die Selbstbeurteilungen so lernwirksam machen, sind allerdings noch nicht erforscht, ihre Klärung ist ein wichtiges Forschungsdesiderat (Andrade 2019, 10).

Beispiele für aufgabenspezifische Selbstbeurteilungsfragen

Hinweis: Die Anschlussfragen nach den Auswahllisten sind zentral. Aus Platzgründen sind sie nur beim ersten Beispiel aufgeführt. Die Struktur der Fragen folgt den Beispielen in Ambrose et al. (2010, 126f.):

Erziehungswissenschaft/Didaktik: Können Sie erklären, was formatives Feedback ist, und könnten Sie es in Ihrem Unterricht anwenden?

- a. Ich habe noch nie davon gehört und/oder weiß nicht, was es bedeutet.
 - b. Ich habe eine Vorstellung davon, was es bedeutet, weiß aber nicht, wann oder wie ich es anwenden soll.
 - c. Ich habe eine klare Vorstellung davon, was formatives Feedback ist, habe es aber noch nicht eingesetzt.
 - d. Ich kann erklären, was formatives Feedback ist, und habe es in meiner pädagogischen Praxis eingesetzt.
- Welche Option entspricht am ehesten Ihrem Kenntnisstand? Erläutern Sie die Gründe.
 - Wie zufrieden sind Sie mit Ihrem Kenntnisstand? Erläutern Sie ggf., was Sie noch dazulernen möchten und wie Sie vorgehen wollen.

Medizin: Sind Sie mit dem Karpaltunnelsyndrom vertraut und können Sie die häufigsten Symptome nennen?

- a. Ich habe noch nie vom Karpaltunnelsyndrom gehört.
- b. Ich habe davon gehört, aber ich bin mir über die Symptome nicht sicher.
- c. Ich habe eine allgemeine Vorstellung von den Symptomen, aber ich könnte sie nicht genau aufzählen.
- d. Ich bin mit dem Karpaltunnelsyndrom vertraut und kann die häufigsten Symptome aufzählen.

Theologie: Können sie das Konzept der Soteriologie kritisch diskutieren?

- a. Der Begriff Soteriologie ist mir noch nie begegnet.
- b. Ich habe schon von Soteriologie gehört, kann mich aber nicht an die genaue Bedeutung oder die Implikationen erinnern.
- c. Ich habe ein grundlegendes Verständnis der Soteriologie und ihrer Bedeutung in theologischen Diskussionen, aber für eine kritische Diskussion des Konzepts müsste ich mir noch vertiefere Kenntnisse aneignen.
- d. Ich verfüge über ein umfassendes Verständnis der Soteriologie, einschließlich verschiedener Theorien und ihrer Bedeutung innerhalb verschiedener religiöser Traditionen.

Beispiele für generische Aufforderungen/Fragen zur Selbstbeurteilung

Wenn mit generischen Fragen zur Selbstbeurteilung gearbeitet wird, ist es umso wichtiger, dass den Studierenden klare Kriterien zum Erreichen der mit

der Aufgabenstellung verbundenen Lernziele vorliegen (siehe Kap. 9.3). Bei Selbstbeurteilungen ist es zudem besonders sinnvoll, die Beurteilungskriterien gemeinsam mit den Studierenden zu entwickeln oder zumindest zu besprechen und ggf. anzupassen (siehe Kap. 9.3.4). Dies erleichtert ihnen einerseits die anstehende Selbsteinschätzung und ist andererseits motivational kongruent mit der Selbstbestimmungsorientierung von Selbstbeurteilungen.

Fragen zur inhaltlichen Bearbeitung der Aufgabe (diese vier Fragen basieren auf Race 2001, 15):

- Welcher Aspekt/welcher Teil ist mir in dieser Aufgabe (dieser Arbeit, diesem Projekt etc.) am besten gelungen?
- Welcher Aspekt/welcher Teil ist mir in dieser Aufgabe am wenigsten gut gelungen?
- Welches war für mich der schwierigste Teil/Aspekt dieser Aufgabe?
- Was ist das Wichtigste, was ich aus dem Bearbeiten dieser Aufgabe gelernt habe?

Fragen zum Lösungsprozess bei der Bearbeitung der Aufgabe:

- Welche Vorgehensweisen, die ich zur Bearbeitung/Lösung der Aufgabe eingesetzt habe, beherrsche ich schon gut?
- Bei welchen Vorgehensweisen habe ich Unsicherheiten verspürt, musste nachschauen/nachfragen oder bin ich unsicher, ob ich sie richtig anwenden konnte?
- Was muss ich tun, um die benötigten Vorgehensweisen noch besser verstehen oder anwenden zu können?

Fragen zur Selbstregulation bei der Bearbeitung der Aufgabe:

- War die Zeit, die ich in die Bearbeitung der Aufgabe und ihrer Elemente investiert habe, angemessen? Welche Elemente hätte ich genauer bearbeiten sollen, wo hätte auch weniger Aufwand genügt?
- Wie hilfreich waren die Lernstrategien, mit denen ich mir das Wissen zur Bearbeitung der Aufgaben angeeignet habe? Würde ich sie bei einer ähnlichen Aufgabe wieder einsetzen oder andere wählen?
- Welche Lernstrategien scheinen mir am hilfreichsten, um die noch nötigen weiteren Lernschritte in Angriff zu nehmen?

8.4.4 Externe Expertinnen

Lehrende können auch externe Expertinnen heranziehen, um den Studierenden Rückmeldungen zu geben. Das ist beispielsweise bei projekt- und fall-basiertem Lernen sinnvoll, da diese externen Personen Fach- und Professionswissen aus betroffenen Bezugsfeldern einbringen können. Da externe Expertinnen in den meisten Fällen keine Lehrenden oder Pädagoginnen sind, ist es vorteilhaft, ihr Feedback mit Rückmeldungen weiterer Instanzen zu einem Mehrquellen-Feedback (Kap. 8.4.6) zu kombinieren. Dies gilt besonders, wenn ihr Feedback fachlich angemessen, aber motivational problematisch ist (siehe die Hinweise in Kap. 8.2).

Je größer der Kontakt zwischen den externen Expertinnen und den Studierenden ist, je größer die Motivation der Expertinnen ist und je größer die materielle oder immaterielle Aufwandsentschädigung ausfällt, desto eher macht es Sinn, sie im Feedbackgeben zu schulen und zu begleiten.

8.4.5 Elektronisches Feedback und künstliche Intelligenz

Neben menschlichen gibt es auch computerbasierte Instanzen, die auf die eine oder andere Weise formative Rückmeldungen zum Lernstand geben können. Dabei ist zu unterscheiden zwischen bisher üblicher tutorieller Software und den derzeit rasant aufkommenden Tools, die auf künstlicher Intelligenz basieren.

Tutorielle Software

Viele Quiz-Softwares und die Quiz- und Test-Funktionen vieler gängiger Learning-Management-Systeme erlauben es, Lösungshinweise zu Fragen zu hinterlegen oder ein automatisches Feedback zu einem ausgefüllten Selbsttest zu geben. In den meisten Fällen ist damit aber nur ziemlich triviales Feedback auf der Aufgabenebene, gelegentlich auch der Prozessebene möglich. Formatives Feedback, welches das weitere Lernen auf allen drei in Kapitel 8.3.3 erwähnten Ebenen – Aufgabe, Prozess und Selbststeuerung – fördert und jeweils an der situativ passenden Stärke oder Schwäche ansetzt, ist hingegen enorm anspruchsvoll vorausschauend zu gestalten und zu programmieren.¹¹⁸

118 Einen bereits erfolgreich angewandten Ansatz hat Narciss (2020; 2006) mit ihrem „Interactive-Two-Feedback-Loops-Model“ entwickelt. Dieser berücksichtigt sowohl Elemente der Lernenden als auch der Lehr-Lern-Situation und ermöglicht die Implementierung verschiedener Feedbackstrategien.

Weiter gibt es auch Ansätze, Software zu nutzen, um metakognitive Hinweise zu geben, aber auch hier stellen sich in der Praxis noch diverse Herausforderungen (vgl. Engelmann, Bannert u. Melzner 2021). Möglicherweise werden tutorielle Systeme bald durch die Verwendung von KI-Instrumenten überholt.

Künstliche Intelligenz (KI)

Eine Feedbackinstanz, die vermutlich in den kommenden Jahren immer umfassender genutzt werden kann, ist KI. Im Moment sind in dieser Kategorie vor allem so genannte Large-Language-Models zu nennen (siehe Kap. 15 zur Bedeutung von KI/LLM für die Leistungsbeurteilung an Hochschulen). In aller Munde ist zum Zeitpunkt der Abfassung dieses Buchs der Chatbot ChatGPT der Firma OpenAI. In den Massenmedien wird derzeit zwar viel über Betrugsmöglichkeiten mithilfe von KI spekuliert (siehe Kap. 15). Die Verwendung angemessener, integrierter Leistungsnachweise (Kap. 7) lässt dieses Problem aber als kaum größer als die aktuell mit Plagiaten und Ghostwriting bestehenden erscheinen. Expertinnen für digitales Lernen und für Leistungsbeurteilung sehen dagegen KI eher als Ressource für das Lernen. Eine Möglichkeit besteht unter anderem darin, gefundene Lösungen oder Formulierungen durch eine KI überprüfen und sich von ihr Feedback geben zu lassen.

Dabei kann man den Studierenden entweder weitgehende Freiheiten lassen und sie bitten, selbst Wege zu finden, sinnvolles Feedback einzuholen. Das kann von der Wahl des geeigneten KI-Tools bis zur Formulierung der Aufforderung (Prompt) an die Software reichen. Oder man kann mehr oder weniger stark vorgeben, auf welchem Weg das Feedback generiert werden soll, indem man das KI-Tool und/oder den Rückmeldungsprompt (bzw. Teile davon) vorgibt. Damit Sie sich eine Vorstellung davon machen können, wie das ganz konkret aussehen kann, drucke ich im Folgenden einen Dialog mit ChatGPT ab.¹¹⁹ Bei der durch ChatGPT kommentierten Frage handelt es sich um die erste Beispielfrage für aufgabenspezifische Selbstbeurteilungen aus Kapitel 8.4.3:

119 Der Dialog ist am 12. Juni 2023 zwischen Tobias Zimmermann und ChatGPT in der kostenlosen Version vom 24. Mai 2023 erfolgt und wird hier verbatim wiedergegeben. Es soll nicht der Eindruck erweckt werden, der Prompt sei optimal gestaltet – er basiert auf guten Anwenderkenntnissen, wurde aber nicht überarbeitet. Vielmehr geht es darum zu zeigen, wie bereits heute KI-Tools mit relativ bescheidenem Aufwand als Feedbackinstanz nutzbar sind.

Eingabeaufforderung von Tobias Zimmermann:

„Ich möchte gerne schauen, ob du mir hilfreiches Feedback geben kannst. Bitte schaue dir den folgenden Entwurf einer Frage an, die Studierende zur Selbstbeurteilung ihres Lernfortschritts verwenden können sollen.

Fragen: Ist die Fragestellung sinnvoll formuliert? Was könnte man ggf. an ihr verbessern? Sind die Antwortoptionen sinnvoll abgestuft? Was könnte man an ihnen ggf. verbessern?

Bitte nenne mir mindestens 3 gelungene und 3 verbesserungswürdige Eigenschaften meines Entwurfs.

Entwurf:

„Erziehungswissenschaft/Didaktik: Können Sie erklären, was formatives Feedback ist, und könnten Sie es in Ihrem Unterricht anwenden?

- a. Ich habe noch nie davon gehört oder ich habe davon gehört, weiß aber nicht, was es bedeutet.
- b. Ich habe eine Vorstellung davon, was es bedeutet, weiß aber nicht, wann oder wie ich es anwenden soll.
- c. Ich habe eine klare Vorstellung davon, was formatives Feedback ist, habe es aber noch nicht eingesetzt.
- d. Ich kann erklären, was formatives Feedback ist, und habe es in meiner pädagogischen Praxis effektiv eingesetzt.“

Antwort von ChatGPT:

„Der Entwurf der Fragestellung ist insgesamt sinnvoll formuliert. Es wird nach dem Verständnis von formativem Feedback gefragt und ob die Studierenden es in ihrem Unterricht anwenden können. Es gibt klare Antwortoptionen, die verschiedene Grade des Verständnisses und der Anwendungsfähigkeit abbilden.

Gelungene Eigenschaften des Entwurfs:

Klarheit: Die Fragestellung ist klar und verständlich formuliert. Es wird deutlich nach dem Wissen über formatives Feedback und der Anwendungsfähigkeit gefragt.

Abstufung der Antwortoptionen: Die Antwortoptionen sind gut abgestuft und ermöglichen den Studierenden, ihren Kenntnisstand und ihre Erfahrung mit formativem Feedback genau einzuschätzen.

Bezug zur Praxis: Die Frage bezieht sich auf die Anwendung von formativem Feedback im Unterricht, was einen praktischen Bezug herstellt und den

Lernfortschritt der Studierenden in Bezug auf ihre pädagogische Praxis erfasst.

Verbesserungswürdige Eigenschaften des Entwurfs:

Differenzierung der Antwortoptionen: Es könnte hilfreich sein, zwischen verschiedenen Stufen des Anwendungsgrads zu unterscheiden. Zum Beispiel könnten weitere Antwortoptionen hinzugefügt werden, um zwischen einer begrenzten Anwendungserfahrung und einer umfassenden Anwendungserfahrung zu differenzieren.

Präzisierung der Antwortoptionen: Die Antwortoptionen könnten präziser formuliert werden, um Missverständnisse zu vermeiden. Insbesondere die Option „Ich habe eine Vorstellung davon“ könnte weiter konkretisiert werden, um klarer zu machen, ob es sich um ein oberflächliches Verständnis oder ein tiefergehendes Wissen handelt.

Einbeziehung des Lernprozesses: Die Fragestellung könnte auch den Lernprozess berücksichtigen, indem sie nach dem Verständnis des formativen Feedbacks und den Bemühungen der Studierenden fragt, dieses Wissen zu erweitern oder anzuwenden. Dadurch würde der Fokus nicht nur auf dem aktuellen Kenntnisstand liegen, sondern auch auf dem Streben nach weiterer Entwicklung.

Insgesamt ist der Entwurf jedoch gut strukturiert und bietet den Studierenden die Möglichkeit, ihren Kenntnisstand und ihre Anwendungsfähigkeit von formativem Feedback einzuschätzen. Durch die Berücksichtigung der oben genannten Verbesserungsvorschläge kann die Fragestellung noch präziser und aussagekräftiger gestaltet werden.“

Die Rückmeldung von ChatGPT ist durchaus angemessen und hilfreich. Insbesondere die drei Verbesserungsvorschläge sind inhaltlich erstaunlich akkurat. ChatGPT wendet darin mit den Unterscheidungen zwischen oberflächlichem und tiefergehendem Wissen (Kap. 4.2) und zwischen Lernprodukt und Lernprozess (Kap. 7.1.2) Konzepte an, die auch in diesem Buch bereits behandelt wurden.

Entscheidend sind bei der Nutzung von KI-Tools als Feedbackinstanzen die Prompts, also die Aufforderungen an die Instanz – von ihnen hängt die Qualität der Antwort ab (vgl. Fleischmann 2023, 7).¹²⁰ Bereits beim heutigen Entwicklungsstand von Large-Language-Models wie ChatGPT liegt es je nach Fachinhalt eher an einem suboptimalen Prompt als am Unvermögen der KI, wenn diese keine angemessene Rückmeldung gibt.

120 Fleischmann (ebd.) weist darauf hin, dass bereits Marktplätze existieren, bei denen bewährte Prompts für KI-Tools wie ChatGPT oder DALL-E verkauft werden (z. B. <https://promptbase.com>), und dass Software-Firmen wie Google „Prompt-Engineers“ für hohe Gehälter einstellen (vgl. Brien, 29. März 2023).

8.4.6 Mehrquellen-Feedback

Rückmeldungen einer Instanz unterliegen immer gewissen Einschränkungen. Lehrende haben beschränkte Zeitressourcen und können von Expertenblindheit betroffen sein, Studierende haben zu Beginn ein eingeschränktes fachliches Gesamtverständnis und technologische Tools können kaum auf emotional-motivationale Aspekte eingehen. Deshalb und zur Schonung insbesondere der menschlichen Zeitressourcen kann es sinnvoll sein, Feedbacks der verschiedenen Instanzen zu kombinieren. Zur Beurteilung von Fach- und Führungskräften werden mithin so genannte 360-Grad-Feedbacks eingesetzt (vgl. z. B. Lepsinger u. Lucia 2009).¹²¹

Dabei geht es weniger darum, dass das Feedback „genauer“ wird – Rückmeldungen unterschiedlicher Instanzen weichen gewöhnlich erheblich voneinander ab, wie Untersuchungen zu Mehrquellen-Feedback im betrieblichen Kontext zeigen (Müller u. Ditton 2014, 19).¹²² Vielmehr sollen sich die unterschiedlichen Perspektiven ergänzen und der Feedback-Empfängerin gestatten, sich ein möglichst umfassendes Bild ihres Standes und ihrer Entwicklungsmöglichkeiten zu machen. Solche Variation im Feedback ist insbesondere dann wertvoll, wenn es um komplexe, offene Aufgabenstellungen geht – und sie bereitet Studierende auf die Art und Weise vor, wie sie Feedback in vielen Berufsfeldern oder beim Publizieren wissenschaftlicher Texte erhalten werden (vgl. Nicol 2010, 510). In der Hochschullehre wird es dabei in vielen Fällen am wirkungsvollsten sein, die Rückmeldungen aus den verschiedenen Quellen auch zeitlich zu staffeln, so dass die Studierenden zu verschiedenen Zeitpunkten und je nach Lehrveranstaltung zu verschiedenen Themen oder Entwicklungsständen eine Rückmeldung erhalten. Dadurch können die Studierenden die einzelnen Rückmeldungen besser berücksichtigen und umsetzen. Zudem ist eine solche Staffelung in der Regel für Lehrende wie Lernende auch organisatorisch einfacher zu bewältigen.

121 Der Begriff ist freilich unklar und irreführend (vgl. Müller u. Ditton 2014, 18): Wessen Feedback ist für welchen Kreisabschnitt zuständig, mit wie vielen Feedbackinstanzen schließt sich der Kreis?

122 Andere Befunde wären angesichts der ungenügenden Konsistenz in der schulischen Notenvergabe (siehe Kap. 2.3) und der beschränkten Konsistenz zwischen Selbst- und Fremdbeurteilung (siehe Kap. 8.4.3) auch überraschend.

8.5 Medialität von Feedback

Die bisherigen Teile dieses Kapitels sind so gehalten, dass sie weitestgehend für alle medialen Realisierungsformen von Feedback gelten. Es gibt aber auch noch einige Aspekte, die es spezifisch für die Verwendung bestimmter medialer Formen zu berücksichtigen gilt. Dank der digitalen Medien können Rückmeldungen heute auf vielfältige Weise realisiert werden. Nachfolgend werden die beiden traditionellen Varianten des schriftlichen und mündlichen Feedbacks (Kap. 8.5.1 und 8.5.2) behandelt, und als eine modernere Alternative darüber hinaus das multimediale Feedback (Kap. 8.5.3). Zu beachten ist dabei: Die Medialität der Rückmeldung muss nicht mit der Medialität der kommentierten Leistung übereinstimmen, multimediales Feedback zu schriftlichen Texten kann zum Beispiel sehr wirkungsvoll sein.

8.5.1 Schriftliches Feedback

Die größte Herausforderung bei schriftlich erteiltem Feedback besteht in der Regel im erheblichen zeitlichen Abstand zwischen dem Erbringen der Leistung und dem Erhalt der Rückmeldung und in der als groß empfundenen kommunikativen Distanz zwischen Absender und Adressat der Rückmeldung. Beide Aspekte machen es zu einer Herausforderung, schriftliches Feedback als Teil eines (iterativen) Feedbackreislaufs zu konzipieren und wahrzunehmen. So zeigen zahlreiche Umfragen und Studien, dass viele Studierende schriftliches Feedback von Lehrenden als mangelhaft und wenig relevant für ihr weiteres Lernhandeln einschätzen (vgl. Nicol 2010, 501f.; Gibbs u. Simpson 2004, 10).

Es gibt aber Maßnahmen, um schriftliches Feedback dialogischer zu gestalten:

- Bewährt sind Aufforderungen an die Studierenden, gemeinsam mit der Abgabe ihrer Arbeit (an Lehrende oder Peers) wesentliche Punkte zu formulieren, zu denen sie sich Feedback wünschen. Das unterstützt den Aufbau eines iterativen Feedbackprozesses (siehe Kap. 8.6). Das Feedback ist dann nicht der Start des Dialogs, sondern bereits dessen Fortsetzung (Elbow u. Sorcinelli 2006, 202). Dieses Vorgehen ist vor allem dann erfolgsversprechend, wenn wie oben erwähnt sichergestellt wurde, dass die Studierenden die Erfolgskriterien für die Aufgabenstellungen verstanden haben (Bloxham u. Campbell 2010). Wie bei der folgenden Maßnahme ist es zentral, dass die Fragen der Studierenden in der Rückmeldung beantwortet werden! Mit zunehmender Übung der Studierenden sollten die benannten Punkte auch die Stellen sein, bei denen sie Schwierigkeiten haben und an denen sie deshalb offen für zukunftsorientierte Hinweise sind (siehe Kap. 8.3.2).

- Eine stärker geführte, für die Studierenden niederschwelligere Möglichkeit besteht darin, dass sie einen Steckbrief zu der Arbeit einreichen, zu welcher die Rückmeldung erfolgt (dieses Vorgehen und die folgenden Fragen beruhen auf Elbow u. Sorcinelli 2006, 201f.). Der Steckbrief sollte Antworten geben auf Fragen wie: Welches ist der zentrale Punkt Ihrer Arbeit/Ihres Produkt(entwurf)s? Wie erlebten Sie Ihren Schreibprozess/Erarbeitungsprozess? Welche Teile/Aspekte Ihrer Arbeit scheinen Ihnen gut, welche weniger gut gelungen? Bei Überarbeitungen ist auch die Frage wichtig: Welche Änderungen haben Sie gegenüber der letzten Version vorgenommen und weshalb?

Durch das Beantworten dieser Fragen sollten die Studierenden eine genauere Vorstellung entwickeln, wozu ihnen eine Rückmeldung dienlich sein könnte. So kann der Steckbrief abgeschlossen werden mit der Frage: Welche Fragen haben Sie an mich als Lesende/n?

- Soweit machbar, hilft die zeitnahe Übermittlung der Rückmeldung – der Engpass liegt in der Regel bei der Zeit der Dozierenden. Deshalb kann es sich lohnen, auch bei schriftlichen Rückmeldungen mit Peer-Feedback zu arbeiten und dafür klare zeitliche Vorgaben zu machen. Auch der didaktische Einsatz von Online-Diskussionsforen kann sinnvoll sein (vgl. Zimmermann u. Rickert 2015; Zimmermann, Bucher u. Hurtado 2010).
- Auch wenn es schon mehrmals gesagt wurde: Wenn (schriftliches) Feedback gleichzeitig mit Noten mitgeteilt wird, beachten es die Studierenden kaum mehr. Formatives Feedback sollte insofern immer ohne Noten mitgeteilt werden (siehe Kap. 9.1). Eine Möglichkeit, die formatives Feedback und Noten seriell kombiniert, sind mehrphasige Leistungsnachweise (siehe Kap. 7.7.1).

Je nach didaktischem Setting sind noch diverse weitere Möglichkeiten denkbar, die schriftliches Feedback in einen umfassenden Feedback-Dialog integrieren. Zentral ist dabei die Haltung der Lehrenden und ggf. Peers, dass schriftliches Feedback keine Einwegkommunikation ist. Vielmehr sollen die Studierenden auch im Umgang mit schriftlichen Rückmeldungen als aktive Akteure ihres eigenen Lernens verstanden werden, welche die Bedeutung von Rückmeldungen aktiv konstruieren (Nicol 2010, 502f.).

8.5.2 Mündliches Feedbackgespräch

Mündliches Feedback unterscheidet sich in einigen Punkten wesentlich von schriftlichem Feedback. Dabei ist hier die Rede von Feedbackgesprächen, die synchron von Angesicht zu Angesicht oder ggf. über Videokonferenzen erfolgen. Explizit nicht gemeint sind Rückmeldungen, die im Rahmen von Publikumssituationen im Seminarraum oder Vorlesungssaal erfolgen – sie

erfüllen die Voraussetzungen für ein formatives Feedback nicht. Ein Vorteil der Mündlichkeit besteht darin, dass die Beteiligten unmittelbar auf Gesagtes reagieren können, was bei schriftlichem Feedback nicht möglich ist. Diese Unmittelbarkeit birgt aber auch einige Risiken in sich, die mit der Flüchtigkeit von Gesprochenem zusammenhängen (Dainton 2018, 41). Besonders wichtig bei mündlichem Feedback ist die Beziehungsebene. Diese kommt hier aufgrund der Unmittelbarkeit der Kommunikationssituation stärker zum Tragen (vgl. Thomann u. Pawelleck 2013, 100–102).

Insgesamt ergeben sich folgende zentrale Hinweise für den Umgang mit mündlichem Feedback, das in Gesprächssituationen erfolgt. Wie aus den Erläuterungen sichtbar wird, sind die Übergänge zu Beratungsgesprächen teilweise fließend:

- Eine wertschätzende Grundhaltung ist zentral. Auch wenn es um eine fachliche Verständigung geht, wird die wertschätzende Haltung primär über die Beziehungsebene etabliert. Als Person in der institutionell definierten Machtposition ist hierbei das Verhalten der Lehrenden entscheidend. Abgesehen von einem freundlichen Gesprächsverhalten ist es deshalb hilfreich, wenn Sie das Gespräch mit einer Frage nach der aktuellen Befindlichkeit des Gegenübers eröffnen („Wie geht es Ihnen?“, „Sind Sie gerade im Prüfungsstress?“ etc.). Dies hilft, einen vertrauensvollen Gesprächsrahmen zu etablieren, was gerade auch dann wichtig ist, wenn wenig Zeit zur Verfügung steht (Thomann u. Pawelleck 2013, 59).
- Wichtig, um eine wertschätzende Grundhaltung im Gesprächsverlauf zu vermitteln, sind das Stellen von (möglichst offenen, keinesfalls rhetorischen) Fragen und das aktive Zuhören. Dies signalisiert, dass Sie an den Gedanken, Fragen und allenfalls auch Schwierigkeiten der Studierenden interessiert sind.
 - Mit offenen Fragen ermöglichen Sie den Studierenden besser, ihr Denken und ihre Anliegen einzubringen. Eine offene Frage wäre zum Beispiel: „Wie ist Ihre Formulierung ‚XYZ‘ gemeint?“ Demgegenüber wäre die geschlossene Frage „Beziehen Sie sich mit Ihrer Formulierung ‚XYZ‘ auf die Theorie A?“ ungünstig, da sie suggestiv wirken kann. Heikel sind zudem Warum-Fragen, da sie schnell Rechtfertigungsdruck erzeugen (Thomann u. Pawelleck 2013, 98).
 - Aktives Zuhören hilft, diese Anliegen zu verstehen und auf sie eingehen zu können. Gemeint ist ein Vorgehen, bei dem die zuhörende Person jeweils kurz zusammenfasst, wie sie den vorangehenden Gesprächsbeitrag des Gegenübers verstanden hat, bevor sie mit ihrem eigenen Beitrag weiterfährt. Dies hilft, eine vertrauensvolle und auf gegenseitige Verständigung zielende Gesprächsatmosphäre aufrechtzuerhalten (vgl. Thomann u. Pawelleck 2013, 94).

- Die Trennung von Wahrnehmung und Wertung ist in mündlichem Feedback noch wichtiger als bei asynchronem Feedback, da kognitive Verzerrungen (siehe Kap. 6.4) aufgrund der Unmittelbarkeit der Kommunikationssituation stärker zum Tragen kommen. Falls es die Situation nötig macht, dass problematische Aspekte angesprochen werden müssen, ist es bei mündlichem Feedback besonders zentral, Wahrnehmung und Interpretation zu trennen.

Ein Beispiel: Zwischenbesprechung zu einer Masterarbeit, Ihr Student zitiert im eingereichten Textausschnitt Quellen auf uneinheitliche Weise. Ein ungünstiges Feedback wäre nun: „Sie müssen doch im Lauf des Studiums gelernt haben, wie man wissenschaftlich korrekt zitiert.“ Stattdessen könnte ein günstiges Feedback sich am Schema Beobachtung – Interpretation – Wirkung – Wunsch/Anliegen von Thomann u. Pawelleck (2013, 100–102) orientieren. Beobachtung: „Sie zitieren wissenschaftliche Quellen teilweise mit Fußnoten, teilweise in Klammern. In einigen Fällen nennen Sie Vornamen der Autorinnen, in anderen nicht.“ Interpretation: „Ich vermute deshalb, dass Sie mit den Richtlinien wissenschaftlichen Zitierens wenig vertraut sind.“ Wirkung: „Das macht auf mich den Eindruck, als würden Sie wenig Wert auf eine einheitlich Zitierweise legen.“ Wunsch/Anliegen: „Bitte schauen Sie sich das Manual zum wissenschaftlichen Zitieren unseres Instituts an und vereinheitlichen Sie die Zitate in Ihrer Arbeit basierend auf dem in unserem Fach üblichen Chicago Manual of Style.“

- Nonverbale Kommunikation wie Gestik, Mimik oder Distanzverhalten und ihre Interpretation sind hoch individuell, zudem gibt es kulturspezifische Unterschiede; Ähnliches gilt für das Empfinden von Räumen (Thomann 2019, 276–79). In Gesprächssituationen in räumlicher Ko-Präsenz spielen diese Aspekte eine wichtige Rolle, aber aufgrund der großen individuellen Unterschiede können kaum generelle Verhaltenstipps gegeben werden. Hilfreich kann sein, dem studentischen Gegenüber eine räumliche Wahlmöglichkeit anzubieten, etwa bei der Wahl des Sitzplatzes oder bei längeren Gesprächen sogar des Treffpunktes. Auf jeden Fall sollte die Sitzordnung „die Gleichwertigkeit der Gesprächsteilnehmer (...) verdeutlichen“ (Rosenbusch 2004, 199).¹²³

Mimik und Gestik können insbesondere Hinweise auf Gefühlszustände der Gesprächspartner geben: Positive und negative Gefühle, Erregung versus Entspannung lassen sich manchmal aus ihnen erschließen. Aufgrund der hohen Individualität nonverbaler Signale lohnt es sich aber, entsprechende Eindrücke in Form von Rückfragen anzusprechen und kommunikativ zu validieren (vgl. Rosenbusch 2004, 198). Eine entspre-

123 Rosenbusch (2004, 199) empfiehlt als Anordnung: „möglichst über Eck, auf gleicher Höhe, etwa 1,50–2,00 m Abstand, bei mehreren Personen im Kreis“.

chende Rückfrage könnte lauten: „Sie wirken auf mich ziemlich enttäuscht, wie geht es Ihnen?“

- Weil mündliche Gespräche besonders flüchtig sind, lohnt es sich, spätestens zum Gesprächsende wesentliche Punkte schriftlich festzuhalten. Dies hilft nicht nur, noch bestehende Missverständnisse und Unklarheiten zu beseitigen, sondern schafft vor allem Verbindlichkeit (vgl. Thomann u. Pawelleck 2013, 71). Diese ist wichtig dafür, dass die Studierenden aus Feedbackgesprächen konkrete Handlungsvorsätze ableiten und in die Tat umsetzen (siehe Kap. 8.6).

Weitere hilfreiche Hinweise zur Gestaltung von Feedbackgesprächen finden sich in der Literatur zum Beraten von Studierenden, etwa Thomann u. Pawelleck (2013) oder Thomann, Honegger u. Suter (2017).

8.5.3 *Multimediales Feedback: Rückmeldungsvideos*

Mit multimedialem Feedback sind Rückmeldungen gemeint, in denen die klassischen Medien der Schrift und der gesprochenen Sprachen ersetzt oder durch zusätzliche Kanäle ergänzt werden (Hassler 2022, 143). Aus den vielen dadurch entstehenden Möglichkeiten wird hier eine Variante fokussiert: Das Erteilen von Rückmeldungen durch das Aufzeichnen und Übermitteln eines Videos.¹²⁴ Eine Vielzahl von Studien deutet darauf hin, dass dieses Feedback Stärken von schriftlichem und mündlichem Feedback kombiniert und bei Studierenden auf hohe Akzeptanz stößt (Mahoney, Macfarlane u. Ajjawi 2019).

Realisiert werden können Rückmeldungsvideos in verschiedenen Formaten (vgl. Hassler 2022, 143f., Mahoney, Macfarlane u. Ajjawi 2019, 158f.):

- als Talking-Head-Video, bei dem eine Person ihr Gesicht filmt, während sie ein Feedback in die Kamera spricht.
- als Screencast, bei dem der Inhalt des Computerbildschirms oder eines einzelnen Fensters zusammen mit Mausbewegungen und einem Audio-kommentar aufgezeichnet wird. Hier wird in der Regel das studentische Erzeugnis angezeigt – ein Text, eine Grafik, eine Berechnung etc. –, das ggf. mit Annotierungen wie farbigen Markierungen oder Stichwörtern versehen werden kann.

124 Dieses Format ist nicht zu verwechseln mit Feedback, bei dem eine mit einer Kamera aufgezeichnete Tätigkeit nachträglich betrachtet und kommentiert wird. Dieses oft als Video-Feedback bezeichnete Vorgehen wird z. B. von Thomann (2019, 298–301) beschrieben.

- Diese beiden Formen können kombiniert werden, indem sich die Dozentin bei einem Screencast zusätzlich mit der Kamera aufzeichnet. Screencast-Software bietet in der Regel eine Option, diese Aufzeichnung als kleinen Talking Head in einer Ecke des Screencasts anzuzeigen.

Reine Talking-Head-Videos erschweren es den Studierenden oft, das Feedback einer genauen Stelle ihres Erzeugnisses zuzuordnen. Deshalb sind insbesondere bei Rückmeldungen zu Texten Screencasts zu bevorzugen. Mittels Screencasts gegebenes Feedback hat unter anderem folgende Vorteile:

- Im Vergleich zu schriftlichem Feedback wirkt die soziale Präsenz der rückmeldenden Person stärker, was vor allem Vorteile auf der emotionalen Ebene bewirkt. Dies gilt sowohl, wenn nur die Stimme zu hören ist, als auch bei der Variante mit einem kleinen Talking-Head-Video. Grund dafür dürfte die Möglichkeit sein, mit der Stimme paraverbale Signale zu geben und somit Humor, Freundlichkeit, Ernsthaftigkeit etc. auszudrücken. Vielen Lehrenden fällt es leichter, Emotionen mit Screencasts auszudrücken, und Studierende können den emotionalen Gehalt in Screencasts leichter erkennen. Entsprechend finden beide Seiten, ein solches Feedback ermögliche mehr soziale Nähe als schriftliches Feedback (Borup et al. 2014, 239–45), was sich wiederum positiv auf die weiteren Kontakte auswirken kann (Hassler 2022, 145).
- Der Fokus des Feedbacks verlagert sich in Screencast-Videos von oberflächlichen Merkmalen wie Tippfehlern oder Stilistik zu tiefergehenden Aspekten wie inhaltlichen Zusammenhängen oder Argumentation. Zudem scheinen Screencasts eine Konzentration des Feedbacks auf wenige, zentrale Punkte zu begünstigen, was es wirksamer macht (Hassler 2022, 146, siehe Kap. 8.3.5).
- Viele Dozierende berichten, dass sie für Screencast-Rückmeldungen nach der anfänglichen Einübung deutlich weniger Zeit benötigen; Berichte über Mehraufwand gibt es keine (Mahoney, Macfarlane u. Ajjawawi 2019, 162f.). Zusammen mit den inhaltlichen Vorteilen ergibt dies einen deutlich höheren Wirkungsgrad als bei konventionellem schriftlichem Feedback.

Als Nachteile nennen Lehrende technische Probleme sowie den Umstand, dass für die Aufzeichnung ein ruhiger Raum nötig ist und Rückmeldungen z. B. nicht bei einer Zugfahrt oder in einem Großraumbüro erstellt werden können (Hassler 2022, 147; Borup et al. 2014, 242).

Was die Wirkung auf das Lernen betrifft, liegen die Herausforderungen von Screencast-Feedback am ehesten darin, dass dieses Format wie schriftliche Rückmeldungen für sich genommen eine Einwegkommunikation darstellt. Die in Kapitel 8.6.1 gegebenen Hinweise zur Einbindung des Feed-

backs in einen Feedbackkreislauf sind deshalb auch für den Umgang mit Rückmeldungsvideos zielführend (vgl. die Empfehlungen in Hassler 2022, 149–55). Zusätzlich sind folgende Aspekte zu beachten (vgl. Hassler 2022, 153–55):

- Die Übermittlung der Screencast-Videos an die Studierenden ist technisch heutzutage kein Problem mehr: Sie können auf ein Learning-Management-System hochgeladen oder cloudbasiert zur Verfügung gestellt werden. Die Herausforderung liegt vielmehr im Datenschutzrecht. Hier sind die jeweils spezifischen rechtlichen Vorgaben zu beachten, die sich je nach nationaler Gesetzgebung unterscheiden können; zudem kann es verschärfende institutionelle Vorgaben geben. Problematisch sind diesbezüglich in der Regel cloudbasierte Systeme, vor allem dann, wenn dadurch Daten auf ausländischen Servern gespeichert werden. Es ist folglich in der Regel empfehlenswert, die technologischen Systeme der eigenen Institution zu verwenden, um die Videos den Studierenden zugänglich zu machen (vgl. Hassler 2022, 153f.).
- Fast alle Screencast-Programme erlauben es, die Aufnahme zu pausieren. „Dies erlaubt, einen Teil des Feedbacks aufzuzeichnen, die Aufzeichnung zu pausieren und zunächst die Gedanken zu sammeln, bevor der nächste Aspekt des Feedbacks aufgezeichnet wird. (...) Mein Feedback wird knapper und zielgerichteter“ (Hassler 2022, 154).
- Das Video muss nicht TV-Qualität haben, und das aufgezeichnete mündliche Feedback muss nicht sprachlich perfekt formuliert sein wie bei geschriebener Sprache. Ein Streben nach formaler Perfektion ist deshalb wenig zielführend, Zeit und Energie sollten eher in die inhaltliche Substanz der Rückmeldung investiert werden. „Ein Räuspern, ein kleiner Grammatikfehler oder Abbruch mitten im Satz sind allesamt verzeihbar. Die Erwartungen an die formale Korrektheit eines multimedialen Feedbacks sind geringer als bei schriftlichem Feedback“ (Hassler 2022, 155).

Durch die technische Entwicklung wird sich die Erstellung von Screencast-Rückmeldungen weiter vereinfachen, und durch Social Media und die fortschreitende Digitalisierung gewöhnen wir uns gesellschaftlich auch immer besser an solche medialen Formate. Hinweise auf bestimmte Software erfolgen aufgrund der rasanten technologischen Entwicklung hier nicht.¹²⁵

125 Zum Verfassungszeitpunkt dieses Buchs aktuelle und hilfreiche Hinweise finden sich in Hassler (2022, 153f.).

8.6 Feedbackbasiertes Handeln: Schließen des Kreislaufs

Nachdem geklärt wurde, was gutes Feedback ausmacht und welche Instanzen es erteilen können, bleibt ein letzter Aspekt zu behandeln: Erst wenn Studierende erhaltenes Feedback für ihre nächsten Lernhandlungen berücksichtigen, wird der Feedbackkreislauf geschlossen. Auch bei hochwertigem Feedback ist es keineswegs selbstverständlich, dass Studierende dies tun. Darum wird besprochen, wie das Aufgreifen erhaltener Rückmeldungen gefördert werden kann (Kap. 8.6.1). Welche weitreichenden lernförderlichen Wirkungen eine formative Feedbackkultur haben kann, wird zum Abschluss dieses Kapitels aufgezeigt (8.6.2).

8.6.1 *Das Berücksichtigen von Feedback fördern*

Auch wenn Studierende formatives Feedback schätzen, und sogar, wenn sie wissen, wie wertvoll dieses für ihr Lernen sein kann, berücksichtigen sie es oft nicht für ihr weiteres Lernen. Die Herausforderung, dass viele Studierende das erhaltene Feedback nicht umsetzen, ist durch zahlreiche Studien belegt (vgl. die Aufzählungen in Jonsson u. Panadero 2018, 531 und Winstone et al. 2017, 2027). Die Studie von Brookhart (2001) zeigt, dass erfolgreiche Lernende das erhaltene Feedback intensiv nutzen und als Steuerungsinformation für ihr Lernverhalten nutzen. Sie zeichnen sich dadurch aus, dass sie Lernen als einen fortlaufenden Prozess erleben und sogar summative Beurteilungen als Zwischenschritte auf einem lebenslangen Lernweg betrachten (Brookhart 2001, 167). Die Annahme liegt nahe, dass viele der von Brookhart untersuchten Studierenden ein entwicklungsorientiertes Bild von Begabung besitzen (siehe Kap. 8.2.2).

Ein erfolgversprechender Weg liegt somit sicher darin, bei möglichst allen Studierenden die Überzeugung wachsen zu lassen, dass formative Leistungsrückmeldungen eine wertvolle Gelegenheit für die Optimierung des eigenen Lernens sind. Dies erhöht ihre intrinsische Motivation, Rückmeldungen tatsächlich zu berücksichtigen und den Feedbackkreislauf zu schließen (siehe Abb. 19). Die in den Kapiteln 8.2 bis 8.5 besprochenen Qualitätsmerkmale sind dafür wesentlich. Daneben gibt es aber auch zwei konkretere Bündel von Maßnahmen, die das Umsetzen von Feedback fördern: Formate von Leistungsnachweisen, die formativem Feedback mehr Gewicht geben, und das eingehende Besprechen oder gar Aushandeln von Beurteilungskriterien mit den Studierenden.

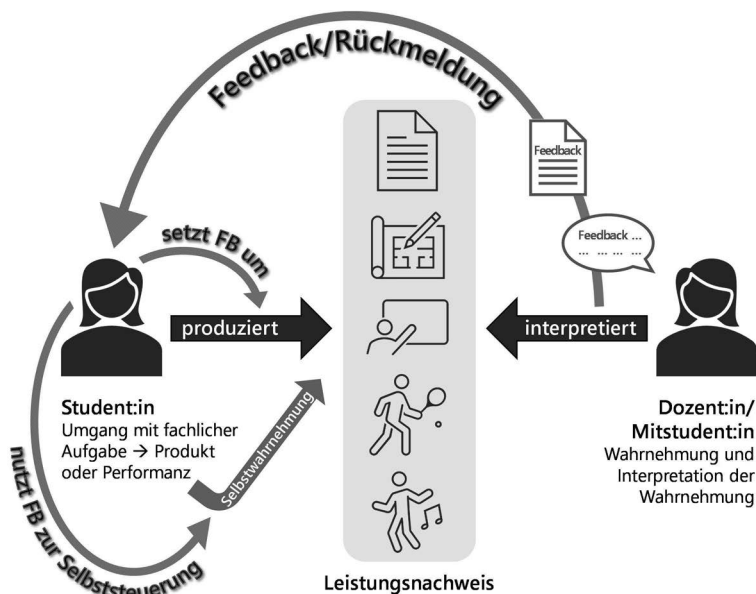


Abb. 19: Schließen des Feedbackkreises durch Umsetzung von Anregungen aus Rückmeldungen. Quelle: Eigene Abbildung, in Anlehnung an Zimmermann 2018, 60

Formativen Rückmeldungen mehr Gewicht geben

Sind Leistungsnachweisformate so strukturiert, dass die formativen Rückmeldungen im weiteren Verlauf des Settings nützlich sind oder sogar kommunikativ weiterverwendet werden, schafft dies Anreize, erhaltene Anregungen umzusetzen. Dazu gehören Formate wie die mehrphasige Beurteilung eines Endprodukts (Kap. 7.7.1), zweistufige Prüfungen (Kap. 7.7.2) oder Posterkonferenzen (Kap. 7.7.3). Solche Formate helfen, Feedbackprozesse als Interaktion zweier oder mehrerer Personen zu gestalten und dadurch ein gemeinsames Verständnis von Aufgabe und Inhalten sowie zwischenmenschliche Verbindlichkeit zu schaffen (Nicol 2010).

Insbesondere schwächere¹²⁶ und wenig intrinsisch motivierte Studierende neigen dazu, von Rückmeldungen kleinschrittige Hinweise zu erwarten, wie sie ihre Arbeit verbessern können. So zeigt die Studie von Orsmond u. Merry

126 Das Adjektiv „schwächer“ verweist nicht auf eine etwaige Begabung, sondern meint rein deskriptiv Studierende, die zu einem bestimmten Zeitpunkt keine guten Leistungen erbringen, unabhängig von möglichen Ursachen.

(2013), dass Studierende mit guten Leistungen viel aktiver mit Feedback umgehen als solche mit schwächeren Leistungen: Insbesondere leiten bessere Studierende aus dem Feedback auch Strategien für das weitere Lernen ab, während die schwächeren dies nicht tun, sondern sich auf kleinschrittige Verbesserungen beschränken. Es ist sicherlich hilfreich, wenn Lehrende den Studierenden regelmäßig aufzeigen, wie sie erhaltenes Feedback (gleichgültig, von welcher Instanz) auch strategisch zum Weiterlernen nutzen können. Dieser Aspekt ist in Abbildung 19 durch den Pfeil „nutzt FB zur Selbststeuerung“ repräsentiert. Entscheidend, um diese Selbstregulation zu stärken, sind Hinweise auf die Wichtigkeit der Metakognition (siehe v.a. Kap. 8.3.2 und 8.3.3). Dazu sind Fragen förderlich wie: Welche Lösungsstrategie ist wofür (nicht) geeignet? Wie können die Lernstrategien für ein noch wirkungsvolleres Lernen oder für eine andere Art von Aufgabe angepasst werden (vgl. Theobald 2021, 12)?

Eine interaktive Integration von Rückmeldungen in ein übergeordnetes Setting kann zudem insbesondere schwächeren Studierenden ein implizites Gerüst an die Hand geben, Feedback für mehr als kleinräumige Korrekturen zu nutzen (vgl. Winstone et al. 2017, 2038).

Besprechen oder Aushandeln von Beurteilungskriterien

Ein großes Problem ist die Konzentration vieler Studierender auf die (abschließenden) Noten. Solange in fast jedem Modul Noten erteilt werden, die für das Schlusszeugnis (Bachelor, Master) eine Rolle spielen, unterminieren sie formatives Feedback (Winstone et al. 2017, 2039; zur demotivierenden Wirkung von Noten siehe Kap. 9.1). Eine hilfreiche Strategie im Umgang mit dieser Problematik besteht darin, mit den Studierenden die Qualitätskriterien für gute Leistungen intensiv zu besprechen. Dies ist nicht trivial: Es ist zwar recht gebräuchlich, die formalen Ansprüche von Leistungsnachweisen mit den Studierenden zu besprechen. Teilweise werden dabei sogar Beurteilungskriterien schriftlich zur Verfügung stellt – eingehend erörtert werden sie aber selten. Dabei zeigt sich, dass Studierende diese Kriterien und damit auch darauf basierende Rückmeldungen oft nur teilweise verstehen (Orsmond u. Merry 2013, 746–50).

Abhilfe können folgende Vorgehensweisen schaffen, bei denen sich die Studierenden intensiver mit den fachlichen Ansprüchen auseinandersetzen und auch aktiver in den Beurteilungsprozess eingebunden werden, was ihre wahrgenommene Selbstbestimmung erhöht (Nicol 2010, 505f.; zur Selbstbestimmung siehe Kap. 4.5.2).

- Neuformulierung der Kriterien in eigenen Worten: Die Studierenden werden gebeten, die Beurteilungskriterien in Gruppen zu diskutieren und sie in eigenen Worten neu zu formulieren (Nicol 2010, 505). In den Gruppen offen gebliebene Fragen können im Plenum geklärt werden.

- Direkte Beurteilungserfahrung: Die Studierenden erhalten gelungene und schwache Bearbeitungen einer Aufgabe und formulieren auf dieser Basis in Gruppen selbst die Kriterien, die gute Bearbeitungen/Lösungen auszeichnen. Idealerweise handelt es sich bei den besprochenen Beispielen um anonymisierte, echte Artefakte früherer Studierender (Nicol 2010, 505f.). Sadler (1989, 134) bezeichnet dieses Vorgehen als direkte Beurteilungserfahrung (“direct evaluative experience”), da die Studierenden hier selbst Arbeiten von anderen beurteilen und anhand dieser Erfahrung Qualitätskriterien konstruieren. Dies ist umso wichtiger bei Beurteilungskriterien, die von Lehrenden formuliert werden. Selbst wenn sie schriftlich festgehalten werden, setzen sie wesentliche Aspekte oder Voraussetzungen oft unausgesprochen voraus. Die Beispiele machen den Studierenden deutlich, was fachlich gut ist und was nicht genügt, zeigen zugleich aber auf, dass Qualität keine unveränderliche Eigenschaft ist, sondern verschiedene Ausprägungen aufweisen kann (Nicol 2010, 506).
- Bei fortgeschrittenen Studierenden kann die direkte Beurteilungserfahrung auch dahingehend erweitert werden, dass gute und schlechte Beispiele selbst gesucht oder konstruiert werden müssen. Das stellt nochmals deutlich höhere kognitive Anforderungen an das fachliche Verständnis und ist nur dann sinnvoll, wenn die Studierenden schon über ein relativ umfassendes Verständnis des Wissens und Könnens im entsprechenden Fachgebiet verfügen.

Dass in allen drei Vorgehensweisen die Kriterien in Gruppen besprochen werden, begünstigt durch soziale Aushandlung das Bewusstmachen unbewusster Voraussetzungen und das Klären von Missverständnissen. Innerhalb der Varianten der direkten Beurteilungserfahrung können die Lehrenden beim abschließenden Festhalten der Beurteilungskriterien mehr oder weniger steuernd eingreifen. Für das Selbstbestimmungsempfinden und damit die intrinsische Motivation der Studierenden ist es optimal, wenn diese Steuerung so gering wie möglich gehalten wird. Werden die Beurteilungskriterien auch für summative Beurteilungen herangezogen, geben rechtliche Rahmenbedingungen manchmal vor, sie vor Beginn der Veranstaltung zu publizieren. In diesem Fall können die beschriebenen Vorgehensweisen auch so gerahmt werden, dass die resultierenden Kriterienformulierungen als „Übersetzungshilfe“ oder „Glossar“ zu den offiziellen Kriterien fungieren (siehe Kap. 9.3.4). Sämtliche Vorteile der intensiven Auseinandersetzung mit den Qualitätsansprüchen an fachliche Arbeiten bleiben dabei erhalten, nur die zusätzliche motivationale Wirkung, Einfluss auf die beurteilungswirksame Kriterienformulierung zu haben, fällt weg.

Die vertiefte Auseinandersetzung mit den Beurteilungskriterien als Ansprüche an fachlich gute Leistungen hilft insbesondere auch (noch) weniger leistungstarken Studierenden, formative Rückmeldungen besser zu verste-

hen und ihr Lernhandeln insgesamt als Weg zum Erbringen fachlich guter Leistungen zu begreifen.

8.6.2 Formative Feedbackkultur: Die Selbstwirksamkeit stärken

Formativen Rückmeldungen durch das Leistungsnachweis-Setting mehr Gewicht zu geben und das Besprechen oder Aushandeln von Beurteilungskriterien können als Schritte auf dem Weg zur Etablierung einer umfassenden Feedbackkultur verstanden werden. In einer solchen würden Studierende aktiver in die Beurteilung ihrer Leistungen einbezogen (was nicht mit einer Selbstbewertung zu verwechseln ist!). Dies motiviert sie wie gezeigt, mehr Selbstverantwortung und eine aktivere Rolle im Beobachten und Beurteilen ihres eigenen Lernfortschritts zu übernehmen. Das übergeordnete Ziel einer solchen Kulturveränderung besteht darin, die Studierenden in ihrer Selbstbestimmung und Selbstregulation als Lernende zu stärken – im hochschulischen Kontext, aber auch für das lebenslange Lernen (Nicol 2010, 515; vgl. auch Carless u. Boud 2018, 1322f.). Die Studierenden sollten deshalb darin unterstützt werden, gegenüber Feedback eine proaktive Rezeptionshaltung zu entwickeln („proactive recipience“, Winstone et al. 2017, 2039). Diese Wirkung wird in Abbildung 19 durch die Pfeile symbolisiert, die über die Selbststeuerung und Selbstwahrnehmung auf das Lernen und die künftige Lernleistung wirken. Während die direkte Anpassung von Wissenselementen oder Produkten aufgrund von Rückmeldungen einen Feedbackeffekt erster Ordnung darstellt, ist die Wirkung auf die Selbstregulation und die Selbstwahrnehmung als Feedbackeffekt zweiter Ordnung zu betrachten: Dadurch wird auch das künftige Lernen und Erbringen von Leistungen weiterentwickelt.

Peer-Feedback, metakognitives Feedback und kooperatives Lernen steigern sowohl direkt als auch vermittelt über das Stärken der Selbstregulation den Lernerfolg (vgl. Schneider u. Preckel 2017; Vermunt u. Donche 2017, 272–82; Theobald 2021, 16). Zudem steigert der stärkere Einbezug der Studierenden in die Praxis der Leistungsbeurteilung ihre wahrgenommene Selbstbestimmung und dadurch die intrinsische Motivation, die sich ebenfalls positiv auf den Lernerfolg auswirkt (Schiefele u. Schreyer 1994): Die Studierenden können sich durch den Feedbackkreislauf gegenüber frontaleren Lehr-Lern-Settings als wirksamer und eigenständiger in der Steuerung ihres Lernens und zugleich als sozial stärker eingebunden erleben (siehe Kap. 4.5.2). Deshalb wirkt das Schließen des Feedbackkreislaufts über Feedbackprozesse erster und zweiter Ordnung mehrfach positiv auf den Lernerfolg.

Zum Schließen des Feedbackkreislaufts gehört notabene auch das Entgegennehmen, Verarbeiten und Aufgreifen von Studierendenfeedback durch Lehrende. Es ist für die Optimierung von Lehren und Lernen noch wirkungs-

voller als die Rückmeldungen, die Studierende erhalten (vgl. Hattie u. Zierer 2017, 77)!¹²⁷ Aber das wäre ein Thema für ein anderes Buch. Das folgende Kapitel widmet sich hingegen dem Bewerten und Benoten von Lernleistungen und der Frage, wie man dieses trotz verschiedener Problematiken so lernförderlich wie möglich gestalten kann.

127 Die Meta-Metastudie (Metastudie zu Metastudien) „Visible Learning“ von Hattie (Hattie u. Zierer 2017, 4–6) zeigt: Lassen sich Lehrende formatives Feedback von den Lernenden geben, hat das eine größere Wirkung auf den Lernerfolg der Lernenden (!) als alle Formen von Feedback an die Lernenden. An Hochschulen haben sich zu diesem Zweck dialogische Feedback- und Evaluationsverfahren bewährt (vgl. Weitzel et al. 2022).

9 Summativ beurteilen: Bewerten und Benoten

Dass ich die Thematik des Bewertens und Benotens erst gegen Ende dieses Buches behandle, liegt nicht daran, dass ich sie für den krönenden Abschluss handle. Im Gegenteil, wie bereits mehrfach angesprochen wurde, sind das Bewerten und besonders das Vergeben von quantifizierenden Noten der problematischste Teil des Umgangs mit Lernleistungen (siehe Kap. 3, 4.5, 5.1.3, 6.2 und 8). Das Thema ist deshalb am Ende des Buches platziert, weil das Verständnis der vorgängig erläuterten Aspekte eine wesentliche Voraussetzung dafür darstellt, Leistungen umsichtig und verantwortungsvoll zu bewerten – im Bewusstsein dessen, was Bewertungen und Benotungen bewirken und was nicht. Dabei wird mit diversen Vorurteilen bezüglich Noten aufzuräumen sein: Noten helfen weder die Qualität des Lernens zu sichern noch motivieren sie Studierende (oder jüngere Lernende) zu echtem Lernen.

Dies bedeutet nicht, dass Lernleistungen nicht bewertet werden sollen – aber zum Setzen von quantifizierenden Noten durch Lehrende gibt es zahlreiche Alternativen der Bewertung, die sich auf das Lernen positiver auswirken und die Qualität des Lernens mindestens so gut sichern helfen. Sie sind auch praktisch umsetzbar und ihre Wirkung ist empirisch dokumentiert. Entgegen stehen ihnen allenfalls unsere kulturell eingeschliffenen Denkmuster. Zu diesen zähle ich auch Rechtsnormen und institutionelle Vorgaben, die oft als Gründe angeführt werden, weshalb modernere Formen von Bewertungen nicht umsetzbar seien. Dazu ist allerdings anzumerken, dass erfahrungsgemäß viele Rechtsnormen bei umsichtiger Planung des Vorgehens auch durch alternative Vorgehensweisen erfüllt werden können; Ähnliches gilt für institutionelle Regelungen, die zudem schneller angepasst werden können. Außerdem scheint es mir wichtig, von Seiten der Pädagogik, (Hochschul-)Didaktik und Lernpsychologie mittels gesellschaftlichen und politischen Dialogs auf die rechtlichen Praktiken bis hin zur Gesetzgebung einzuwirken – mit dem Ziel, diese dort, wo es noch nicht erfolgt ist, an den heutigen wissenschaftlichen Kenntnisstand anzupassen.

Als erstes werden in diesem Kapitel die Wirkungen besprochen, die Bewertungen und insbesondere Noten auf die Lernmotivation und die Selbstregulation des Lernens haben (Kap. 9.1). Anschließend werden zwei verschiedene Vorstellungen der Messung von Lernleistungen kontrastiert (Kap. 9.2). Darauf basierend wird gezeigt, wie Bewertungskriterien definiert und Bewertungsraster konstruiert werden können (Kap. 9.3). Damit die Studierenden die Beurteilungskriterien und -raster verstehen und sie auf selbstbestimmte Weise als Richtschnur ihres Lernhandelns verwenden können, müssen diese zumindest intensiv mit ihnen besprochen werden. Noch positiver auf die Lernmotivation und die Selbstregulation des Lernens wirkt es sich

freilich aus, wenn die Beurteilungskriterien gemeinsam mit den Studierenden ausgehandelt und definiert werden (Kap. 9.3.4). Im Bewusstsein schwerwiegender mathematischer Unzulänglichkeiten können schließlich Aufgaben gewichtet, auf der Basis von Beurteilungskriterien und -rastern Notenskalen gebildet und Noten berechnet werden (Kap. 9.4). Sinnvoller als Noten sind allerdings grob abgestufte Basisbewertungen (Kap. 9.5). Zum Abschluss werden noch zwei grundlegendere Alternativen zur herkömmlichen Bewertungs- und Benotungspraxis vorgestellt (Kap. 9.6).

9.1 Die Wirkung von Bewertungen auf Motivation und Selbstregulation

In den Kapiteln 3.1 und 3.2 haben wir unterschieden zwischen Bewertungen, also wertenden Einschätzungen einer Leistung, und nicht wertenden Beurteilungen, die eine Leistung einschätzen, ohne eine formale Bewertung mit Noten oder Wortprädikaten („sehr gut“, „gut“ etc.) vorzunehmen. Bewertungen wirken sich im Unterschied zu nicht wertenden Beurteilungen eindeutig negativ auf die Motivation und den Lernfortschritt von Lernenden jeglicher Stufen aus. Dies ist durch zahlreiche empirische Untersuchungen seit den Experimenten von Ruth Butler Ende der 1980er Jahre bestens belegt (vgl. z.B. Marx 2020, 80–82; Hattie u. Clarke 2019, 2–7; Brookhart 2018b, 62–64; Butler 1988). Zudem gibt es zahlreiche Studien, die positive Effekte von formativem Feedback auf die Lernmotivation und den Lernfortschritt belegen (siehe Kap. 8), wobei es sich bei formativem Feedback um eine spezifische Form von Beurteilung handelt (vgl. Brookhart 2018b, 58–62). Insofern spricht, wie bereits durch dieses ganze Buch in verschiedenen Zusammenhängen erläutert, viel dafür, den Einsatz von Noten im Rahmen jeglicher Bildungsformen zumindest einzudämmen.¹²⁸

Ein gewichtiges Argument zugunsten einer Eindämmung besteht darin, dass Lernende aller Stufen die inhaltlichen Rückmeldungen der Lehrpersonen

128 Ob Noten als notwendiges Übel hinzunehmen sind oder eine Schule ohne Noten mehr und besseres Lernen und damit wohl auch einen größeren gesellschaftlichen Nutzen im wirtschaftlichen Bereich bis zu größerer Mündigkeit und Selbstverantwortung der einzelnen Menschen bewirken würde, ist eine spätestens seit der Bewegung der Reformpädagogik an der Wende vom 19. zum 20. Jh. anhaltende Debatte (vgl. Bohl 2005, 225f.). Sie wird in der Pädagogik und Didaktik, aber auch der breiten Gesellschaft mehr oder minder anhaltend geführt. Über politische Prozesse wirken sich neue Erkenntnisse aus dieser Debatte mit teilweise erheblicher Verzögerung (oder auch gar nicht) in Form von Anpassungen der rechtlichen und regulatorischen Vorgaben auf das Bildungswesen aus.

weniger bis gar nicht mehr beachten, sobald sie zugleich Noten erhalten. Exemplarisch sei hier von zwei Studien berichtet, in denen die Wirkung von Noten und formativen Rückmeldungen auf jüngere Lernende und auf Studierende untersucht wurde:

- In der Studie von Butler (1988) erhielten israelische Schülerinnen und Schüler im Alter von 11–12 Jahren entweder numerische Noten oder aufgabenbezogene individuelle Kommentare oder beides, nachdem sie interessante Aufgaben bearbeitet hatten. Die Gruppe, die nur Kommentare erhalten hatte, machte deutlich mehr Lernfortschritte als die beiden anderen Gruppen. Die Gründe lagen darin, dass nur die Lernenden, die ausschließlich Kommentare erhalten hatten, diese sowohl beachteten als auch ernst nahmen. Die Lernenden, die sowohl Note als auch Kommentar erhalten hatten, beachteten fast nur die Note. Sie verstanden den Kommentar als ein Signal der Wertschätzung durch die Lehrperson, aber schenkten ihm oft nur wenig Beachtung. Für sie enthielt die Note die wesentliche Aussage darüber, was ihre Aufgabenlösung „wert“ war.
- In der rund 20 Jahre später durchgeführten Studie von Lipnevich u. Smith (2009) erhielten Studierende von Einführungsveranstaltungen in Psychologie an zwei US-Colleges¹²⁹ unterschiedliche Kombinationen von Feedback, Noten und Lob zu einem von ihnen eingereichten Entwurf eines Essays. Dabei zeigte sich eine klar negative Wirkung von Noten: Studierende, die eine Note erhalten hatten, verfassten deutlich schlechtere definitive Essays als jene, die nur Lehrenden-Feedback erhalten hatten. Zudem bewirkten Noten eine schlechtere Stimmung und eine schlechtere Einschätzung der eigenen Selbstwirksamkeit als Feedback ohne Noten. Dass sich Lob negativ auf die Motivation auswirkte, ist überdies konsistent mit den Resultaten von Dweck und anderen (siehe Kap. 8.2.2).

Die beiden Studien zeigen, dass relativ unabhängig von Ort, Zeit und Alter der Lernenden das Erteilen von Noten sich negativ auf den Umgang mit Feedback auswirkt: Gleichzeitig erhaltene inhaltliche Rückmeldungen werden weniger für das weitere Lernen genutzt, ja oft kaum beachtet. Man könnte diesen Effekt von Noten als Rückmeldungsunterdrückung bezeichnen. Wie Lipnevich u. Smith (2009, 330) aufzeigen, kann der Effekt mit der Feedback-Interventionstheorie von Kluger u. DeNisi (1996) erklärt werden. Diese Theorie besagt, dass erfolgreiche Rückmeldungen die Aufmerksamkeit der Lernenden auf die Aufgabe und die für ihre erfolgreiche Bewältigung hilfreichen Strategien richten sollten und nicht auf ihre Persönlichkeit (so auch Hattie u. Timperley 2007, 96f., siehe Kap. 8.3.2). Letzteres tun Noten allerdings: Da sie wertend sind, aber kaum Informationsgehalt haben, lenken sie

129 Dies entspricht dem Beginn eines Bachelor-Studiums im europäischen Hochschulraum.

die Aufmerksamkeit weg von der Aufgabe hin zur eigenen Persönlichkeit, was ihre negative Auswirkung auf intrinsische Motivation, Lernen und Leistung erklärt. Die Tendenz, Noten auf die generelle Leistungsfähigkeit von Lernenden zu beziehen, zeigt sich ebenfalls in den Wortkonstruktionen Einer-, Zweier-Schülerin etc., wie sie gerne auch von Lehrpersonen verwendet werden (vgl. Breidenstein u. Thompson 2014, 93–97). Die Bestätigungstendenz ist hier als kognitive Verzerrung bereits im Begriff angelegt, und sie demotiviert Lernende zuverlässig, da unter dieser Verzerrung erhaltene Bewertungen durch die eigene Leistung wenig zu beeinflussen sind (siehe Kap. 6.4.1 zu den kognitiven Verzerrungen).

9.2 Verschiedene Konzeptionen von Leistungsbeurteilung

Dem praktischen Umgang mit Leistungsbeurteilung wie auch der Forschung darüber liegen zwei grundlegend verschiedene Konzeptionen zugrunde. Sie unterscheiden sich darin, wie Leistung zu erfassen ist, wie sie zu beurteilen ist und letztlich sogar darin, was eine Leistung überhaupt ausmacht. Diese beiden Konzeptionen werden im öffentlichen wie auch im wissenschaftlichen Diskurs oft vermischt, was aber nicht über ihre grundlegenden methodologischen und lernpsychologischen Differenzen hinwegtäuschen darf. Wenn es darum geht, Leistungen zu bewerten, ist es entscheidend, diese Differenzen zu kennen. Nur auf dieser Grundlage können wir erkennen, durch welche Konzeption von Leistungsbewertung ein konkreter Leistungsnachweis, ein Studiengang, eine Hochschule oder ein ganzes Bildungssystem geprägt sind. Und vor allem gibt uns das Wissen um die beiden Konzeptionen die Möglichkeit zu entscheiden, welche Konzeption wir bevorzugen und in welche Richtung wir unsere eigene Bewertungspraxis und jene unserer Institution – im Rahmen der auch formal gegebenen Möglichkeiten und Grenzen – entwickeln wollen.

Die beiden Konzeptionen werden nachfolgend bezeichnet mit „Leistungsbewertung als Messen und Vergleichen“ und „Leistungsbewertung als Überprüfen von Lernzielen“ (vgl. Taylor 1994). Wenn Sie nun denken, dass sich die beiden Konzeptionen doch problemlos kombinieren lassen, dann zeigt dies vor allem, wie sehr sie in den letzten Jahrzehnten vermischt wurden. Einige der dadurch entstandenen Widersprüche werden wir nun genauer betrachten.

9.2.1 Leistungsbewertung als Messen und Vergleichen

Die Konzeption von Leistungsbewertung als Messen und Vergleichen wurde von der differentiellen Psychologie übernommen (Taylor 1994, 236f.; Biggs u. Tang 2007, 170). Diese geht davon aus, dass sich Menschen in psychischen Merkmalen in zeitlich stabiler Weise unterscheiden, und untersucht diese Differenzen sowie ihre Zusammenhänge und Ursachen (vgl. Hagemann, Spinath u. Mueller 2022, 19–21). Dazu werden menschliche Eigenschaften oder Verhaltensweisen mit psychometrischen Methoden gemessen und statistisch ausgewertet. Dieses Vorgehen basiert auf den beiden Grundannahmen, dass

1. der Messwert einer Person für ein Merkmal relativ zur Verteilung der Messwerte anderer Personen angegeben werden kann und
2. Messinstrumente (z. B. Testverfahren, Fragebögen, aber auch beobachtungsorientierte Verfahren wie z. B. die Messung von Reaktionszeiten) entwickelt werden können, welche diese individuellen Unterschiede zuverlässig (reliabel) messen.

Bereits Francis Galton (1889), der heute als Begründer der differentiellen Psychologie betrachtet wird, postulierte in diesem Zusammenhang, so gemessene Eigenschaften verteilten sich entsprechend der gaußschen Normalverteilung. Diese Annahme ist auch heute noch eine wesentliche Grundlage jener psychologischen Richtung (vgl. Hagemann, Spinath u. Mueller 2022, 21–24).¹³⁰ Daraus resultierten die modernen statistischen Verfahren, die heute verwendet werden, um die Reliabilität von psychometrischen Messungen festzustellen (siehe Kap. 6.3.1) und die Ergebnisse auszuwerten. Damit diese mathematischen Verfahren funktionieren, müssen die Messwerte eine erhebliche Varianz aufweisen und verschiedene Messelemente müssen möglichst unabhängig voneinander sein – dies setzen die zugrundeliegenden mathematischen Modelle voraus (Taylor 1994, 237). So ist zum Beispiel eine Aufgabe, die von allen Studierenden richtig gelöst wird, im Rahmen dieser Konzeption von Leistungsbewertung nicht brauchbar, da sie keine Streuung erzeugt. Und eine größere Zahl an Aufgaben erhöht die Reliabilität durch die Steigerung der Streuung.

Während diese Annahmen für psychologische Eigenschaften wie Persönlichkeitsmerkmale oder physiologische Größen wie Körpergröße durchaus angemessen sein mögen, sind sie im Bildungskontext irreführend: Eine Lehrhandlung, die davon ausgeht, dass an ihrem Ende die Lernleistungen normal-

130 Allerdings gibt es durchaus Zweifel an dieser Annahme, vgl. den Artikel von Micceri (1989) mit dem sprechenden Titel „The unicorn, the normal curve, and other improbable creatures“, der für 440 große Stichproben von Leistungsmessungen und psychometrischen Daten zeigt, dass sie nichtnormal verteilt sind.

verteilt sind, unterstellt ihre eigene Erfolglosigkeit (vgl. Sacher 2014, 121). Der Lernstand von Studierenden nach erfolgter Lehre sollte eben gerade nicht mehr zufällig sein, sondern schief in Richtung erfolgreich erreichter Lernziele (Biggs u. Tang 2007, 172) – sonst wäre jegliche didaktische Maßnahme überflüssig (siehe dazu Kap. 5.1.3). Gleichmaßen ist Abschied zu nehmen von der testtheoretischen Forderung nach möglichst hoher Streuung der Messwerte (Varianz) – der Erfolg didaktischer Interventionen bemisst sich vielmehr auch daran, dass die Streuung der Lernleistungen reduziert wird. Dies zeigt ein Gedankenexperiment: Wie viele Betriebe würden eine interne Weiterbildung für ihre Mitarbeitenden als erfolgreich einschätzen, die eine hohe Streuung bzw. eine Normalverteilung der Lernfortschritte erzeugt (vgl. Cohen 1987, 16)?¹³¹

Man kann dieses Gedankenexperiment auch umdrehen: Für eine Lehrveranstaltung in einer Sozialwissenschaft besteht ein spezifisches Lernziel darin, dass die Studierenden mit einem geeigneten Datensatz einen T-Test durchführen können. Nun lösen im Rahmen einer Klausur alle geprüften Studierenden eine Aufgabenstellung korrekt, bei der anhand eines Datensatzes ein solcher Test durchgeführt werden soll. Gemäß der Konzeption von Leistungsbewertung als Messen und Vergleichen müsste diese Aufgabe für die Testbewertung weggelassen werden, da sie keine Streuung erzeugt: Die Aufgabe ist somit „zu leicht“ und sie hat keine Trennschärfe, sie trennt also nicht „gute“ von „schlechten“ Prüflingen. Fänden Sie es in Anbetracht des Alignments von Lernziel und Aufgabenstellung sinnvoll, diese Aufgabe für die Bewertung/Benotung nicht zu berücksichtigen? Wir kommen im Rahmen des folgenden Unterkapitels darauf zurück.

Vergleichsorientierte Leistungsbewertungen funktionieren also nur, wenn die einen Studierenden die anderen übertreffen (Taylor 1994, 242). Dies kann man weltanschaulich heikel finden: Wollen wir, dass die Studierenden derart direkt miteinander konkurrieren müssen? Erwerben sie dadurch sozial erwünschte Einstellungen? Die Vergleichsorientierung ist aber vor allem völlig untauglich, um festzustellen, was Studierende nun konkret können. Salopp formuliert: „Wenn ich daran denke, eine Spritze von einem Krankenpfleger oder einer Ärztin zu erhalten, welche die beste in einer Kohorte von Nichtsnutzen waren, dann bevorzuge ich deutlich, dass alle Gesundheitsfachleute ein definiertes Kriterium für das Spritzengeben erfüllen, vielen Dank!“ (McKeachie u. Svinicki 2006, 133, Übers. tz). Dies ist auch ein weiterer Grund, weshalb die in Kapitel 6.2.3 erläuterte kollektive Bezugsnorm, von

131 Eine hohe Varianz erzeugt auch stets Verlierer bzw. „schlechte“ Studierende. Biggs und Tang (2007, 173) merken dazu an, eine gezielte Selektion von Studierenden mache grundsätzlich nur bei der Aufnahme von Studierenden Sinn; seien sie einmal aufgenommen, müsse das Ziel sein, dass sie möglichst erfolgreich studieren. Auch deshalb ist ein systematisches Produzieren von Varianz problematisch.

der die Konzeption von Leistungsbewertung als Vergleichen zwingend ausgehen muss, problematisch ist.

Ein nächstes zentrales Problem stellt sich in der mess- und vergleichsorientierten Konzeption von Leistungsbewertung bezüglich der Relation verschiedener Aufgabenstellungen untereinander: Um die erwünschte Varianz zu erzeugen, müssen viele verschiedene Aufgaben gestellt werden. Aber die inhaltliche Relation zwischen diesen Aufgaben ist kaum zu quantifizieren. Sind alle Aufgaben genau gleich schwierig? Repräsentieren sie alle je einen Lerninhalt, der genau gleich wichtig ist? Sind alle Aufgaben vollkommen unabhängig voneinander, d. h. keine setzt einen Wissensinhalt voraus, der notwendig ist, um eine andere zu lösen? Keine dieser Fragen kann guten Gewissens mit ja beantwortet werden, und die letzte lässt die Absurdität der messtheoretischen Voraussetzungen aufscheinen: Es macht wenig Sinn, in einer Prüfung zu einem Themenfeld lauter Aufgaben zu stellen, die überhaupt keinen Zusammenhang untereinander aufweisen.

Die Nichtvergleichbarkeit einzelner Aufgabenstellungen ist weit mehr als eine akademische Spitzfindigkeit, sondern mathematisch relevant. Wenn eine Aufgabe zu einer anderen nicht exakt gleichwertig ist, dann folgt daraus, dass keine Äquivalenzrelation zwischen ihnen besteht: Wir können empirisch nicht nachweisen, dass $a = b$ oder $b = c$, und schon gar nicht, dass $a = c$. Das heißt, dass sich die Lösungen von verschiedenen Aufgaben im Rahmen z. B. einer Klausur nicht quantifizieren lassen. So ist es etwa eine unzulässige Verkürzung, alle 30 Aufgaben einer Prüfung mit je einem Punkt zu bewerten und dann damit mathematische Operationen durchzuführen. Diese Problematik führt der Mathematiker und Mathematikdidaktiker Walther L. Fischer in seinem lesenswerten Text „Mathematische Kritik der Ziffernnoten und ihrer Interpretation“ (1991, 233 u. 247) näher aus. Sie „läßt es grundsätzlich fraglich erscheinen, ob wir die geforderten Äquivalenzen formal konstituieren können. (...) Die Anwendung der Testtheorie im und auf den Benotungsprozess ist (...) bestenfalls als approximierende Modellbildung zu verstehen.“¹³²

Hinzu kommt, dass Leistungen im Gegensatz zu Persönlichkeitseigenschaften nicht sehr stabile „Merkmale“ (schon der Begriff ist unpassend) darstellen: Tageszeit, Gesundheitszustand, psychisches Wohlbefinden und vor allem auch die Art, wie die Leistung sozial konstruiert wird (siehe Kap. 2.5), beeinflussen diese erheblich. Die Modellannahme, es handle sich um einigermaßen stabile Merkmale, ist deshalb ebenfalls bereits problematisch.

Die Konzeption von Leistungsbewertung als Messen und Vergleichen ist also weder psychometrisch noch lernpsychologisch noch mathematisch ihrem

132 Wohlmerkt: Diese Problematik besteht bereits vor der Übertragung von Punktwerten oder anderen Wertzuweisungen auf eine Noten- oder andere Bewertungsskala. Auf mathematische Probleme geht Kap. 9.4 näher ein.

Gegenstand angemessen. Dennoch beruht das Bewertungssystem im Bildungswesen seit über hundert Jahren auf dieser Vorstellung.

9.2.2 Leistungsbewertung als Überprüfen von Lernzielen

Die Konzeption der Leistungsbewertung als Überprüfen von Lernzielen wurde hingegen entwickelt, „um Veränderungen in der Leistung als Ergebnis des Lernens zu bewerten mit dem Ziel zu sehen, was gelernt wurde und wie gut“ (Biggs u. Tang 2007, 177, Übers. tz). Diese Vorstellung geht grundsätzlich von der kriterialen Bezugsnorm aus, die in Kap. 6.2.1 dargelegt wurde. Entscheidend ist in dieser Konzeption, wie weit und wie gut Studierende die gesetzten Lernziele erreichen. Es geht nicht darum, Studierende untereinander zu vergleichen, sondern ihre Leistungen mit den Lernzielen abzugleichen (Biggs u. Tang 2007, 177). Dies bedeutet auch, dass sehr unterschiedliche Arten von Leistungen das Erreichen der gleichen Lernziele demonstrieren können. Damit ist nicht nur gemeint, dass beispielsweise wissenschaftliche Arbeiten auf unterschiedliche Weise „gelingen“ sein können. Vielmehr kann das Erreichen eines Lernziels wie „Sie vergleichen verschiedene Formen von Leistungsnachweisen und bewerten ihre Stärken und Schwächen“ auf ganz verschiedene Weise sichtbar gemacht werden, etwa als wissenschaftliche Arbeit, als Essay, als Blogpost, als Podcast, im Rahmen einer Präsentation etc.¹³³

Um Leistungen auf das Erreichen von Lernzielen hin zu bewerten, ist deshalb eine ganzheitliche Bewertung der gezeigten Leistungen notwendig (Biggs u. Tang 2007, 178; Taylor 1994, 245f.). Demgegenüber reduziert die Frage, wie viele Punkte man welchem Abschnitt gibt, die Komplexität der Bewertung auf viele lokal begrenzte Entscheidungen – aber wie soeben gezeigt auf methodisch und inhaltlich problematische, um nicht zuzagen unzulässige Weise. Dies verweist uns zurück auf die in Kapitel 2.4 kritisierte metrische Fixierung, die viele gesellschaftliche Bereiche inklusive des Bildungswesens dominiert. Dort habe ich betont, wie wichtig begründete Urteile

133 Einwänden, dies sei zu beliebig, ist entgegenzuhalten, dass wir daran gewöhnt sind, dass Lehrende (oder der Lehrplan) die arbiträre Entscheidung fällen, in welcher Form eine zu bewertende Leistung erbracht werden soll – und dann müssen alle Studierenden ihre Leistung in dieser Form erbringen. Der Stotterer muss einen Vortrag halten oder die Legasthenikerin eine schriftliche Klausur absolvieren. Ob man dies didaktisch angebracht findet, ist eine normative Entscheidung. Sie steht jedenfalls im Widerspruch zur Forderung von Bloom, Madaus u. Hastings (1981, 53): „Die grundlegende Aufgabe in der Bildung besteht darin, Strategien zu finden, welche die individuellen Unterschiede berücksichtigen, dies aber in einer Weise, die die bestmögliche Entwicklung des Einzelnen fördert.“

sind, für die soziometrische Zahlen letztlich nur eine von verschiedenen Informationsgrundlagen darstellen. Die angemessene Bewertung von Lernleistungen stellt einen konkreten Anwendungsfall dieses Urteilsvermögens dar. Lehrende müssen sich dieses Urteilsvermögen ebenso erarbeiten wie Studierende, um ihre eigenen Leistungen und jene ihrer Peers einschätzen zu können. Es trägt auch bei zu einer akademischen Gemeinschaft, die sich laufend über die Standards wissenschaftlichen Wissens und wissenschaftlicher Methoden verständigt und damit deren kritische Weiterentwicklung ermöglicht (siehe die Erläuterungen zur epistemologischen Antinomie in Kap. 3.3.2).

Ein Instrument, das sich besonders für diese Art von ganzheitlichen Beurteilungen eignet, sind Beurteilungsraster, die in Kapitel 9.3 erläutert werden (Biggs u. Tang 2007, 178–215; Taylor 1994, 250f.). Insbesondere wenn sie gemeinsam mit Studierenden diskutiert und wenn möglich sogar überhaupt erst erarbeitet werden, tragen sie viel zur Entwicklung eines fachlichen Urteilsvermögens und damit der Internalisierung von fachlichen Ansprüchen auch bei den Studierenden bei. In Kapitel 9.3.4 werden diesbezügliche Möglichkeiten erläutert. Hilfreich bei der Arbeit mit Beurteilungsrastern sind auch konkrete Beispiele von Leistungen, die als Ankerbeispiele fungieren. Idealerweise werden diese als Teil des Lernprozesses mit den Studierenden besprochen und zur Eichung verwendet, falls verschiedene Lehrende als summative Bewertende fungieren (Taylor 1994, 243f.).

Von Anhängern der mess- und vergleichsbezogenen Leistungsbewertung wird der lernzielorientierten Bewertung oft vorgeworfen, sie vernachlässige die testtheoretischen Gütekriterien (Biggs u. Tang 2007, 188). Diese sind allerdings mitnichten unproblematisch, wie bereits in Kapitel 6.3.1 herausgearbeitet wurde. So orientiert sich eine lernzielorientierte Bewertung an anderen, nicht minder anspruchsvollen Ansprüchen. Diese richten sich an das Fällen von qualitativ begründeten Urteilen, auf denen die summativen Bewertungen zu beruhen haben (siehe Kap. 6.3.2). Das Ziel der Objektivität wird dabei ersetzt durch jenes der Intersubjektivität, das eine diskursive Aushandlung der Beurteilungen und Bewertungen anstrebt. Dieses Aushandeln wird ermöglicht durch eine transparente Beschreibung und multiperspektivische Orientierung des Beurteilungsprozesses und seiner Grundlagen. Ein solches Vorgehen zielt auf Authentizität: Indem verschiedene subjektive Realitätskonstruktionen berücksichtigt und verhandelt werden, stellt der soziale Aushandlungsprozess sicher, dass das beurteilt wird, was beurteilt werden soll – also die Validität. Die von den mathematischen Modellen des vergleichsorientierten Ansatzes geforderten großen Leistungsunterschiede zwischen den Studierenden sind dadurch in der lernzielorientierten Konzeption von Leistungsbewertung nicht relevant.

9.2.3 *Das Beste beider Welten kombinieren?*

Prallen zwei unterschiedliche Ansätze aufeinander, ist es oft erfolgsversprechend, sie zu einem dritten Paradigma zu integrieren, welches die Stärken der beiden konkurrierenden Paradigmen integriert und ihre Schwächen reduziert. Im Fall der beiden besprochenen Konzeptionen von Leistungsbewertung ist dies aber schwer vorstellbar, da sie sowohl in ihrer Zielsetzung als auch ihrer Grundannahme stark divergieren. Das macht eine Entscheidung nötig:

- Gehen wir davon aus, dass es sich bei Lernleistungen um überdauernde Persönlichkeitseigenschaften handelt und es unsere Aufgabe als Lehrende ist, die guten von den schlechten Studierenden zu trennen? Dann liegt es nahe, dass wir den Sinn der summativen Leistungsbewertung im Messen, Vergleichen und ggf. Aussortieren von Studierenden sehen.
- Gehen wir davon aus, dass Lernleistungen auch kurzfristig schwanken, mit der Lernmotivation und damit auch mit der ganzen sozialen Lernumgebung zusammenhängen, sich von Thema zu Thema (sogar innerhalb eines Studienfachs) unterscheiden können und es unsere Aufgabe ist, durch unsere Aktivitäten als Lehrende möglichst vielen Studierenden möglichst viel Lernzuwachs zu ermöglichen? Dann liegt es nahe, dass wir den Sinn der summativen Leistungsbewertung darin sehen, das Erreichen der Lernziele zu überprüfen.

Welche Vorstellung liegt Ihnen näher? Aus meiner Sicht spricht die große Mehrheit der lernpsychologischen, neurowissenschaftlichen, soziologischen, sozialwissenschaftlichen, mathematischen, pädagogischen und didaktischen Zusammenhänge für die zweite Konzeption der Leistungsbewertung. Denn die Art, wie Lehren und Lernen gestaltet wird – von den gesellschaftlichen Haltungen und institutionellen Setzungen bis hin zum persönlichen Umgang von Lehrenden mit Lernenden –, beeinflusst, was gelernt werden kann, wie es gelernt werden kann und wie nachhaltig diese Lernprozesse sind. Diese Zusammenhänge wurden insbesondere in den Grundlagenkapiteln dargelegt und in den bisherigen Kapiteln des Anwendungsteils weiter vertieft. In den nachfolgenden Abschnitten versuche ich Möglichkeiten aufzuzeigen, wie Leistungsbewertungen lernpsychologisch, pädagogisch und didaktisch so sinnvoll wie möglich gestaltet werden können in einem gesellschaftlichen und institutionellen Umfeld, das einer lernzielorientierten Leistungsbewertung bestenfalls ambivalent gegenübersteht.

9.3 Beurteilungsraster und andere Beurteilungsinstrumente

Beurteilungsraster sind das Instrument der Wahl, um das Erreichen von Lernzielen zu überprüfen sowie Feedback und Diskurse zu Lernleistungen für alle Beteiligten (lern-)wirksam zu gestalten. Ihre Verwendung wird von den Vertretern der lernzielorientierten Konzeption von Leistungsbewertung empfohlen und beschrieben (vgl. Biggs u. Tang 2007, 178–215; Wiggins u. McTighe 2005, 172–90; Taylor 1994, 250f.). Hochschuldidaktikerinnen, die sich nicht oder nicht eindeutig dieser Richtung zurechnen lassen, messen dem Instrument ebenfalls einen hohen Stellenwert zu (vgl. Ambrose et al. 2010, 231–43; McKeachie u. Svinicki 2006, 95–97; Metzger u. Nüesch 2004, 30–32).

Dass Beurteilungsraster auf so breiter Linie empfohlen werden, hängt unter anderem damit zusammen, dass sie sowohl für formative Leistungsrückmeldungen als auch für summative Leistungsbewertungen verwendet werden können. Sie wirken sich empirischen Befunden zufolge in der Hochschullehre in beiden Beurteilungskontexten positiv auf das Lernen der Studierenden aus (vgl. Brookhart 2018a; Panadero u. Jonsson 2013; Reddy u. Andrade 2010). Damit sind sie das ideale Instrument, um das Erreichen von Lernzielen zu überprüfen, können aber auch für das Vergeben von Bewertungen wie Noten und Prädikaten verwendet werden. Eine Ausnahme stellen geschlossene Aufgaben wie die verbreitete Form der Multiple-Choice-Aufgaben dar. Geschlossene Aufgaben werden in der Regel durch andere Instrumente in Bewertungen überführt (Kap. 9.3.5).

Im Folgenden wird einerseits erklärt, was Bewertungsraster genau sind, wie sie konstruiert werden (Kap. 9.3.1) und wie sie zur formativen (Kap. 9.3.2) und summativen Beurteilung (Kap. 9.3.3) eingesetzt werden können. Zudem wird erläutert, wie sie gemeinsam mit Studierenden besprochen oder gar erarbeitet werden können – ein Vorgehen, das sich erwiesenermaßen positiv auf die Lernmotivation und die Lernleistung der Studierenden auswirkt (Kap. 9.3.4).

9.3.1 Beurteilungsraster und ihre Konstruktion

Ein Beurteilungsraster ist ein Instrument, das die Leistungserwartungen für eine Aufgabenstellung in dichter Weise aufführt. Die Qualität der Leistung wird durch eine Skala in mindestens drei Stufen unterteilt. Dabei lassen sich analytische und globale Beurteilungsraster unterscheiden. Globale Beurteilungsraster weisen nur eine Skala auf, anhand derer die gesamte Leistung ganzheitlich beurteilt wird. Demgegenüber umfassen analytische Beurteilungsraster mehrere Leistungsdimensionen und sind in den meisten hoch-

Tab. 9: Beurteilungsraster für einen argumentativen Text (ausgewählte Dimensionen aus Andrade 2000, 17, sinngem. Übers. tz)

Argumentationsqualität				
Aufgabenstellung: Schreiben Sie ein Thesenpapier zum Themengebiet ...				
	Stufe 1	Stufe 2	Stufe 3	Stufe 4
Die These	Ich stelle eine These auf und erkläre, warum sie umstritten ist.	Ich stelle eine These auf, erkläre aber nicht, warum sie umstritten ist.	Meine These ist versteckt, verworren und/oder unklar.	Ich sage nicht, was mein Argument oder meine These ist.
Stützende Argumente	Ich begründe meine Behauptung klar und genau.	Ich gebe Gründe zur Unterstützung meiner Behauptung an, übersehe aber wichtige Gründe.	Ich gebe 1–2 schwache und/ oder irrelevante Gründe an, die meine These nicht ausreichend stützen.	Ich gebe keine Gründe zur Unterstützung meiner These an.
Gegenargumente	Ich erörtere die Gründe gegen meine These und erkläre, warum sie trotzdem gültig ist.	Ich erörtere Gründe gegen meine These, lasse aber einige außer Acht oder erkläre nicht, warum die These trotzdem gültig ist.	Ich sage, dass es Gründe gegen die These gibt, aber ich diskutiere sie nicht oder nur sehr verkürzt.	Ich erkenne die Gründe, die gegen meine These sprechen, nicht an und diskutiere sie nicht.
Organisation	Mein Text hat einen prägnanten Anfang, der Mittelteil ist informativ, der Schluss überzeugend.	Mein Text hat einen Anfang, eine Mitte und ein Ende.	Meine Gliederung ist grob, aber brauchbar. Ich schweife manchmal vom Thema ab.	Mein Schreiben ist ziellos und unorganisiert.

schulischen Kontexten zu bevorzugen (vgl. Ambrose et al. 2010, 231; Metzger u. Nüesch 2004, 30f.; Taylor 1994, 250). Tabelle 9 zeigt ein Beispiel für ein analytisches Beurteilungsraster.

Tab. 10: Formale Struktur eines Beurteilungsrasters (basierend auf Stevens u. Levi 2013, 31)

Titel			
Aufgabenstellung: Hier wird die Aufgabenstellung (ggf. gekürzt) beschrieben.			
	Skalenstufe 1	Skalenstufe 2	Skalenstufe 3
Dimension A	Beschreibung A-S1	Beschreibung A-S2	Beschreibung A-S3
Dimension B	Beschreibung B-S1	Beschreibung B-S2	Beschreibung B-S3
Dimension C	Beschreibung C-S1	Beschreibung C-S2	Beschreibung C-S3
Dimension D	Beschreibung D-S1	Beschreibung D-S2	Beschreibung D-S3

Qualitätsskala und Leistungsdimensionen werden also in Form einer Matrix dargestellt. Ein analytisches Beurteilungsraster umfasst somit vier Bestandteile, die in Tabelle 10 abstrahiert dargestellt sind (Stevens u. Levi 2013, 29–40):

- Beschreibung der Aufgabenstellung
- eine Skala
- verschiedene Leistungsdimensionen
- Beschreibungen der verschiedenen Skalenstufen pro Dimension

Der große Vorteil von Beurteilungsrastern besteht darin, dass sie in knapper Form die wesentlichen Aspekte einer Aufgabenstellung, der mit ihnen verbundenen Lernziele sowie der verschiedenen Abstufungen der Erreichung dieser Lernziele sichtbar machen. Im Falle von summativen Beurteilungen kann auch die unterschiedliche Gewichtung verschiedener Leistungsdimensionen bzw. Lernziele im Raster sichtbar gemacht werden. Die Klarheit von Beurteilungsrastern ist für die Verständigung zwischen Lehrenden und Studierenden hilfreich. Sie erhöht die Wahrscheinlichkeit, dass die Studierenden verstehen, welche Lernziele sie erreichen sollen und wie sie deren Erreichen im Rahmen der gestellten Aufgabe sichtbar machen können. Diese Verständigung wird umso mehr erhöht, je stärker die Studierenden an der Interpretation oder gar Erarbeitung des Beurteilungsrasters beteiligt werden (siehe Kap. 9.3.4).

Beurteilungsraster können für alle Arten von offenen Aufgabenstellungen und somit für die meisten Leistungsnachweisformate verwendet werden: Von Essayklausuren über Vorträge, Posterpräsentationen und mündliche Prüfungen bis zu umfassenden Arbeiten wie künstlerischen Darbietungen, Projekten und Fallbearbeitungen und natürlich wissenschaftlichen Arbeiten (Stevens u. Levi 2013, 29; Ambrose et al. 2010, 231). Die einzige größere Ausnahme stellen Formate dar, die mit geschlossenen Aufgaben arbeiten, wobei es sich im Hochschulkontext meist um Multiple-Choice-Aufgaben handeln dürfte.

Die vier Bestandteile von Beurteilungsrastern werden nachfolgend im Sinne einer Bauanleitung näher erläutert. Wo nicht anders vermerkt, folgen die Ausführungen der Beschreibung von Stevens u. Levi (2013, 29–40).

Beschreibung der Aufgabenstellung

Das Raster sollte mit einer Beschreibung der Aufgabenstellung beginnen. Falls die Aufgabenstellung umfassend ist, wie etwa bei einer Fallarbeit oder größeren Projekten, sollte sie stichwortartig umschrieben werden und einen Verweis auf die Stelle enthalten, an der sie in ihrer Gesamtheit beschrieben ist.

Es gibt auch Beurteilungsraster, die nicht auf eine einzelne Aufgabenstellung zielen, sondern generisch funktionieren, etwa für das Halten von Vorträgen oder das korrekte Verfassen von Laborberichten. In diesem Fall wird die zu beurteilende Tätigkeit generisch beschrieben oder es wird auf die Stelle verwiesen, an der etwaige Richtlinien zu finden sind (z. B. zu Laborberichten).

Skala

Ein Beurteilungsraster sieht eine Skalierung vor, also eine Abstufung der Qualitätsmerkmale. Diese umfasst mindestens drei Stufen; mehr als fünf Stufen sind nicht empfehlenswert. Hierbei ist eine Abwägung vorzunehmen: Mehr Stufen können den Studierenden klarer machen, welche Leistungsaspekte zu berücksichtigen sind. Andererseits wird es mit zunehmender Stufenzahl auch schwieriger, eine konkrete Leistung trennscharf einer bestimmten Stufe zuzuordnen.

Die Stufen sollten mit gut verständlichen Überschriften versehen werden, welche die Leistungsabstufungen verdeutlichen. Basierend auf den Erläuterungen in Kapitel 8.2.2 ist es zielführend, diese Überschriften entwicklungsorientiert zu benennen. Ob dabei eher die Ebene des Produkts oder des Lernprozesses angesprochen werden soll, hängt sowohl von der spezifischen Aufgabenstellung wie von weiteren Faktoren (z. B. fachlichen Konventionen) ab. Beispiele für Skalenbenennungen:

- eher am Lernprozess orientiert:
 - versiert, handlungsfähig, voranschreitend, beginnend
 - Expertin, Fortgeschrittene, fortgeschrittene Novizin, Novizin
- eher produktorientiert:
 - über den Erwartungen, Erwartungen erfüllt, braucht noch mehr Arbeit
 - vorbildlich/ausgezeichnet, fachlich korrekt, genügend mit Verbesserungspotenzial, verbesserungsbedürftig

- etwas härter und traditioneller: sehr gut, gut, genügend, noch nicht genügend

Auch die SOLO-Taxonomie eignet sich gut, um Beurteilungsraster zu strukturieren (siehe Kap. 5.4.2). Die taxonomische Hierarchie von unstrukturiert, unistuktural, multistuktural, relational bis zu fortgeschrittener Abstraktion lässt sich relativ gut auf konkretere Anforderungen hin ausformulieren. Im Abschnitt zur Beschreibung der Skalenstufen pro Dimension findet sich ein konkretes Beispiel dafür.

Wenn das Bewertungsraster für summative Zwecke verwendet wird, kann es sinnvoll sein, bereits von Anfang an auch die den verschiedenen Skalenstufen zugeordneten Noten oder Prädikate aufzuführen (siehe Kap. 9.4.3 zum Ermitteln von Noten mit Beurteilungsrastern).

Leistungsdimensionen

Um die zentralen Leistungsdimensionen einer Aufgabenstellung oder einer generischen Leistungskategorie zu identifizieren, ist eine Analyse erforderlich: Welche Lernziele sollen durch die Aufgabenstellung (oder generische Leistungskategorie) adressiert werden? Wie wird in der Bearbeitung der Aufgabenstellung, ihrer Lösung bzw. dem fertigen Produkt das Erreichen der einzelnen Lernziele sichtbar? Allenfalls kann es sich auch lohnen zu überlegen, was besonders gelungene oder besonders schwache Bearbeitungen ausmacht – falls vorhanden, kann sich dabei ein Blick auf frühere Arbeiten von Studierenden lohnen (Stevens u. Levi 2013, 61).¹³⁴

Das Benennen verschiedener Dimensionen und in einem nächsten Schritt verschiedener Skalenstufen verdeutlicht auf relativ einfache Weise die zentralen Ansprüche an eine Aufgabenstellung. Optimal ist, wenn eine Dimension ein Lernziel anspricht, wobei eine solchermaßen klare Zuteilung nicht immer möglich ist. Die verschiedenen Dimensionen können auch noch gewichtet werden, etwa mit Prozentzahlen. Solche Gewichtungen verdeutlichen die unterschiedliche Bedeutung verschiedener Aspekte für das Bearbeiten und Lösen der Aufgabenstellung zusätzlich und können v. a. auch für die summative Bewertung herangezogen werden. Beispiele für Leistungsdimensionen:

- Indem sie sich an der vorgegebenen Textstruktur für eine schriftliche Argumentationsaufgabe orientieren, unterscheiden Biggs u. Tang (2007,

134 Für das Herausarbeiten der zentralen Leistungsdimensionen kann auch das Verfahren der Primary Trait Analysis wertvoll sein, mit dem zentrale Anforderungen einer Aufgabenstellung herausgearbeitet werden (vgl. Walvoord u. Anderson 1998, 65–92).

210) die folgenden Leistungsdimensionen: Einleitung, Argumentation, Zusammenfassung und Schlussfolgerung, Umgang mit Fachliteratur.

- Für Peer-Rückmeldungen in einem Online-Diskussionsforum zu einer Vorlesung unterscheiden Zimmermann, Bucher u. Hurtado (2010, 325f.) die folgenden Leistungsdimensionen: Qualität des Sachinhalts, kritisches Denken und Intensität des Austauschs.
- In einem Beurteilungsraster für Laborberichte unterscheiden Stevens u. Levi (2013, 160) folgende Leistungsdimensionen, die teils Prozess- und teils Produktaspekte adressieren: Materialien, Verfahren, höfliche Umgangsformen und Sicherheit, Zielsetzung, Datenerfassung, Datenanalyse, Auswertung des Experiments.

Beschreibung der Skalenstufen pro Dimension

In der Regel ist es am einfachsten, mit einer Beschreibung der höchsten Skalenstufe einer Dimension zu beginnen. Die nächsttiefere Stufe beschreibt dann den Unterschied zwischen dem Maximum und dieser Stufe. So kann auch mit etwaigen weiteren Stufen verfahren werden. Die letzte Stufe beschreibt idealerweise das, was hätte geleistet werden sollen, aber nicht geleistet wurde. Sie fokussiert auf diese Weise nicht nur auf das Defizit, sondern auch auf die Entwicklungsmöglichkeiten. Beispiele für Beschreibungen von Skalenstufen pro Dimension:

- Die Anforderungen an Zusammenfassung und Schlussfolgerung einer schriftlichen Argumentationsaufgabe formulieren Biggs u. Tang (2007, 210, Übers. tz) auf Basis der SOLO-Taxonomie folgendermaßen:
 - unstrukturiert: Es gibt keine erkennbare Struktur und keine erkennbaren Schlussfolgerungen.¹³⁵
 - unistruktural: Die Zusammenfassung ist eine Aufzählung von Vor- oder Nachteilen, die zu einer einseitigen Schlussfolgerung führt.
 - multistruktural: Die Zusammenfassung bezeichnet Unterschiede, vermag sie aber nicht aufzulösen; schwache oder voreilige Schlussfolgerung.
 - relational: Die Zusammenfassung ist ausgewogen und führt zu einer gut begründeten Schlussfolgerung.
 - fortgeschrittene Abstraktion: Die Zusammenfassung führt zu einer überraschenden oder originellen Schlussfolgerung, die neue Fragestellungen generiert.

135 Biggs u. Tang (2007, 210) machen zu dieser Stufe keine Angaben, sie wurde hier zur Vervollständigung ergänzt.

- Die Anforderungen an die Datenerfassung in einem Laborbericht formulieren Stevens u. Levi (2013, 160, Übers. tz) wie folgt:
 - Die Rohdaten, einschließlich der Einheiten, werden in angemessener und klarer Weise aufgezeichnet. Der Titel der Datentabelle ist enthalten.
 - Die Rohdaten, einschließlich der Einheiten, werden aufgezeichnet, wenn auch nicht so deutlich und angemessen wie möglich. Der Titel der Datentabelle ist enthalten.
 - Die Rohdaten, einschließlich der Einheiten, werden nicht angemessen und klar aufgezeichnet. Der Titel der Datentabelle ist nicht enthalten.

Eine ausführlichere, sehr hilfreiche Schritt-für-Schritt-Anleitung, wie man Beurteilungsraster konstruieren kann, findet sich im Buch von Stevens u. Levi (2013, 58–80).

9.3.2 *Beurteilungsraster und formative Leistungsrückmeldungen*

Wie im Rahmen von Kapitel 8 zur formativen Leistungsrückmeldung bereits mehrfach erwähnt wurde, sind klare Beurteilungskriterien auch für formatives Feedback von zentraler Bedeutung (siehe v. a. Kap. 8.3.4 und 8.6.1). Sie formulieren die fachlich-inhaltlichen Ansprüche, auf die sich jegliche Art von formativem Feedback beziehen sollte. Beurteilungsraster sind besonders geeignet, um einen für alle Beteiligten klar verständlichen kriterialen Rahmen herzustellen. Dabei unterscheiden sich Beurteilungsraster für die formative und solche für die summative Verwendung höchstens in Nuancen. Grundsätzlich lässt sich für beide Beurteilungsarten das gleiche Raster verwenden (siehe Kap. 9.3.3). Unterschiedlich ist vor allem die Art und Weise, wie das Raster unter den beiden unterschiedlichen Zielsetzungen verwendet wird.

Bei der formativen Leistungsrückmeldung dient die Rückmeldung dazu, den Studierenden klarzumachen, welche Aspekte eines Lernziels erfüllt sind und wo noch mehr Arbeit bzw. mehr Lernen nötig ist. Durch ein Beurteilungsraster kann ein ganzes Bündel von Lernzielen formativ adressiert werden. Hierfür sind verschiedene Vorgehensweisen möglich, aus denen drei bewährte erläutert werden (vgl. Stevens u. Levi 2013, 115–124):

Tab. 11: Beurteilungsraster mit angekreuzten Skalen versus markierten Aussagen

Rückmeldungen in Online-Diskussionsforen				
Aufgabenstellung: ...				
	Versiert	Handlungs- fähig	Voranschreitend	Beginnend
Kriti- sches Denken Ange- kreuzt	☒ Text wirft Fragen auf und beantwortet sie...	☐ Text wirft Fragen auf...	☐ Text wirft Fragen auf...	☐ Text ist rein...
Kriti- sches Denken Markiert	<u>Text wirft Fragen auf und beantwor- tet sie; themati- siert/hinterfragt Voraussetzungen und vergleicht sie mit eigenen Kon- zepten</u> oder de- nen <u>von Mitspu- dierenden</u> .	Text wirft Fra- gen auf und beantwortet sie; themati- siert/ hinter- fragt Voraus- setzungen.	Text wirft Fragen auf und beant- wortet sie, the- matisiert/ hinter- fragt aber keine Voraussetzun- gen.	Text ist rein repro- duktiv, wirft keine Fragen auf.

- Ankreuzen von Skalenwerten: Dies ist die zeitsparendste Möglichkeit, die allerdings auch am wenigsten auf die individuelle Leistung zugeschnitten ist, die beurteilt wird (siehe Tab. 11). Diese Option macht einerseits dann Sinn, wenn die Zeit knapp ist. So kann etwa direkt während eines Vortrags ein entsprechendes Beurteilungsraster ausgefüllt und anschließend den Vortragenden zur Verfügung gestellt werden (wobei Lehrenden- und Peer-Feedback möglich ist). Auch wenn sehr viele Leistungen rückzumelden sind, kann diese Option für Lehrende wertvoll sein. Dabei ist zu bedenken, dass Feedback vor allem dann wirksam ist, wenn es zeitnah zur erbrachten Leistung erfolgt. Abstriche bei der Individualisierung zugunsten der Geschwindigkeit können deshalb sinnvoll sein (Gibbs u. Simpson 2004, 18f.).

Zusätzlich ist diese Option kombinierbar mit individualisiertem Feedback (siehe letzten Punkt in dieser Liste).

- Markieren zutreffender Aussagen: In diesem Fall ist das Beurteilungsraster nicht mit Checkboxes versehen, die angekreuzt werden. Die zutreffenden Passagen werden vielmehr in den verschiedenen Skalenstufen pro Dimension durch Unterstreichen, Leuchtstiftmarkierung, Umkreisen o. Ä. markiert (siehe Tab. 11). Dadurch kann die rückmeldende Person in einer Dimension auch Aspekte von zwei verschiedenen Skalenwerten markieren, wenn eine Leistung in Bezug auf diese Dimension unterschiedliche Qualitäten aufweist. Das ermöglicht eine etwas größere inhaltliche Diffe-

renzung, ohne dass ausführliches individualisiertes Feedback verfasst werden muss. Im Gegenzug wirken markierte Raster allerdings weniger übersichtlich als solche mit angekreuzten Checkboxes. Auch hier ist zusätzlich individualisiertes Feedback möglich.

- Individualisiertes Feedback: Falls es kombiniert ist mit einer der obigen Optionen, kann ein einzelner Satz als Hinweis, wo am ehesten noch Arbeit/Lernzeit investiert werden muss, sehr wirksam sein (Zimmermann, Tobias 2018, 75). Insbesondere bei formativen Rückmeldungen zu Zwischenständen größerer Arbeiten wie wissenschaftlichen Arbeiten, Fallbearbeitungen und Projektarbeiten ist auch ein ausführlicheres Feedback angemessen und wirksam. Ob in einem solchen Fall ein Ankreuzen oder Markieren von Skalenwerten noch als nötig erachtet wird, hängt vom Kontext ab.

Falls die Studierenden (wie in Kap. 8.5 empfohlen) gebeten wurden, zwei oder drei Fragen an die rückmeldende Person zu formulieren, sollte das Raster ebenfalls ein Feld für verbales Feedback enthalten, um auf diese Fragen zu reagieren.

Die oben erwähnten Optionen beziehen sich primär auf schriftliches Feedback. Selbstverständlich können Beurteilungsraster im Rahmen von mündlichen Feedbackgesprächen oder von Rückmeldungsvideos ganz ähnlich genutzt werden (siehe Kap. 8.5.2 und 8.5.3). In Rückmeldungsvideos lassen sich ausgefüllte Beurteilungsraster auch direkt im Video anzeigen (Screencast).

9.3.3 *Beurteilungsraster und summative Leistungsbewertung*

Beurteilungsraster können ebenfalls für summative Leistungsbewertungen verwendet werden. Sie führen dann zu Noten, Basisbewertungen oder Prädikaten wie „bestanden/nicht bestanden“. Der entscheidende Unterschied zur formativen Verwendung besteht darin, dass die verschiedenen Skalenstufen in Bewertungen überführt werden. Dies kann auf diverse Weisen geschehen, von denen nachfolgend drei erläutert werden. Bei den Verfahren, in denen Punkte vergeben werden, braucht es zusätzlich eine Notenskala, welche die Punktzahl umwandelt in eine Ziffernnote (siehe dazu Kap. 9.4.6).

Zentral ist: Wenn Noten vergeben werden müssen, so ist die tiefste genügende Note gleichzusetzen mit dem Erreichen des minimalen Lernziels. Diese Setzung hilft, entsprechende Kalibrierungen vorzunehmen.

- Skalenstufen werden direkt mit Werten versehen: Wie das genau geschieht, hängt sowohl von der Anzahl der Skalenstufen als auch von der Anzahl an Noten, Basisbewertungen oder Prädikaten ab. Zudem können die Leistungsdimensionen unterschiedlich gewichtet werden. In Tabelle 12 sind beide Optionen umgesetzt und für jede Dimension eine Beispielbewertung eingetragen (die Berechnung der Note wird in Kapitel 9.4.3 erläutert).

Tab. 12: Beurteilungsraster mit Notenzuteilung, Gewichtung und Gesamtnote

Titel			
Aufgabenstellung			
	Skalenstufe 1 = Note 1	Skalenstufe 2 = Note 3	Skalenstufe 3 = Note 5
Dimension A Gewichtung: 40 %	X		
Dimension B Gewichtung: 20 %		X	
Dimension C Gewichtung: 30 %	X		
Dimension D Gewichtung: 10 %			X
NOTE (gewichteter Median): 1			

- Alternativ können pro Dimension Punkte vergeben werden, was differenziertere Abstufungen erlaubt (vgl. Biggs u. Tang 2007, 210). Auch hier ermöglicht die maximale Punktzahl pro Dimension eine unterschiedliche Gewichtung der Dimensionen: Können für Dimension A doppelt so viele Punkte vergeben werden wie für Dimension B, so ist ihr Gewicht doppelt so hoch. Tabelle 13 zeigt ein Beispiel für dieses Vorgehen. Um die verschiedenen Skalenstufen gegeneinander abzusetzen, wird jeweils mindestens eine Punktzahl zwischen den Skalenstufen nicht vergeben (z. B. bei Dimension C: Die Punktzahlen 5 und 11). Die Bandbreite der Punkte muss dabei nicht für jede Skalenstufe gleich sein. Wenn es z. B. in einer Dimension schwieriger erscheint als in anderen Dimensionen, die oberste Stufe zu erreichen, dann kann hier die Bandbreite etwas schmaler ausfallen und dafür z. B. in der mittleren Skalenstufe etwas größer sein.

Tab. 13: Beurteilungsraster mit Punktebandbreiten

Titel			
Aufgabenstellung			
	Skalenstufe 1	Skalenstufe 2	Skalenstufe 3
Dimension A Max. Punkte: 20	Beschreibung Punkte: 15–20	Beschreibung Punkte: 8–13	Beschreibung Punkte: 0–5
Dimension B Max. Punkte: 10	Beschreibung Punkte: 8–10	Beschreibung Punkte: 4–6	Beschreibung Punkte: 0–2
Dimension C Max. Punkte: 15	Beschreibung Punkte: 12–15	Beschreibung Punkte: 6–10	Beschreibung Punkte: 0–4
Dimension D Max. Punkte: 5	Beschreibung Punkte: 5	Beschreibung Punkte: 3	Beschreibung Punkte: 0–1
Gesamtpunktzahl: ____			

- Bei der letzten Option handelt es sich um eine Variation der Möglichkeit, Punkte zu vergeben. In diesem Fall erfolgt die Quantifizierung der Punkte durch konkret benannte Aspekte, was mehr Transparenz schafft. So können z. B. die Dimension A fünf Aspekte umfassen, die entweder erfüllt oder nicht erfüllt werden. Für jeden erfüllten Aspekt wird ein Kreuz gesetzt und ein Punkt vergeben (vgl. Stevens u. Levi 2013, 127f.). Dieses Vorgehen ist in Tabelle 14 exemplarisch dargestellt. Es hat den Nachteil, dass hier keine Skalenstufen abgebildet werden, andererseits ist es neutraler, was die Bewertung der einzelnen Aspekte betrifft: Jeder vorhandene Aspekt hat eine bessere Bewertung zur Folge.

Tab. 14: Beurteilungsraster mit bepunkteten Aspekten

Schriftliche Fallanalyse		
Aufgabenstellung: ...		
Textorganisation Max. Punkte: 4	☒ Klare Textstruktur ☒ Klare Leseführung („roter Faden“) ☐ Pointierte Schlussfolgerungen ☒ Literaturverweise in passender Zahl und an passenden Textstellen	Punkte: 3
Dimension B Max. Punkte: 9	☐ Aspekt 1 ☐ Aspekt 2 ☐ Aspekt 3 etc.	Punkte: ____
Dimension C Max. Punkte: 4	etc.	Punkte: ____
Gesamtpunktzahl: ____		

9.3.4 *Beurteilungsraster gemeinsam besprechen und ausarbeiten*

Beurteilungskriterien mit den Studierenden zu besprechen, ist in jedem Fall eine zentrale Maßnahme, um den Studierenden die Bedeutung der Lernziele und den Weg zu ihrem Erreichen zu verdeutlichen. Das Besprechen bereits fertig gesetzter Kriterien mit den Studierenden ist deshalb die Minimalvariante eines adäquaten Umgangs mit Leistungsanforderungen. Die Maximalvariante besteht darin, Studierende so stark als möglich in die Erstellung eines Beurteilungsrasters einzubeziehen – im Extremfall ist praktisch nur noch die Aufgabenstellung durch die Lehrenden vorgegeben, während das Raster weitgehend durch die Studierenden unter Moderation der Lehrenden erarbeitet wird.

In Kapitel 8.6.1 wurden mit Blick auf das Schließen des Feedbackkreislauks bereits einige Möglichkeiten erläutert, wie Beurteilungskriterien mit Studierenden besprochen oder ausgehandelt werden können. Der Fokus lag dabei darauf, dass Studierende erhaltenes Feedback tatsächlich für ihr weiteres Arbeiten bzw. Lernen berücksichtigen. Beim formativen Einsatz von Beurteilungsrastern ist dies wiederum das Hauptziel, während beim summativen Einsatz ein Element der Ermächtigung hinzukommt. In beiden Fällen ist der Einbezug von Studierenden in die Gestaltung von Beurteilungsrastern mit positiven Auswirkungen verbunden. Diese betreffen sowohl die Motivation der Studierenden als auch ein gesteigertes Verständnis für die Lernziele und die Lernhandlungen, die zu ihrem Erreichen führen. Für uns Lehrende bietet der Einbezug der Studierenden ebenfalls Vorteile: Im Vorfeld von Leistungsnachweisen sind wir in der Regel weniger mit Fragen zu deren Ansprüchen („ist das Prüfungsstoff?“) konfrontiert, und im Nachgang reduzieren sich Diskussionen mit Studierenden, die mit der Beurteilung unzufrieden sind. Diese Verbesserungen resultieren daraus, dass die Studierenden die Anforderungen und auf ihnen basierende Beurteilungen und Bewertungen besser verstehen. Überdies bietet die Besprechung oder Erarbeitung von Beurteilungsrastern eine wertvolle Gelegenheit, um von den Studierenden zu erfahren, was sie bereits wissen und können – und was noch nicht. Auch Missverständnisse oder falsche Annahmen bezüglich der Aufgabenstellung oder inhaltlicher Anforderungen können dabei oft geklärt werden (vgl. Stevens u. Levi 2013, 82–87).

Wenn Beurteilungsraster (auch) für summativ Beurteilungen verwendet werden, geben rechtliche Rahmenbedingungen manchmal vor, sie vor Beginn der Veranstaltung zu publizieren. In diesem Fall können die beschriebenen Vorgehensweisen auch so gerahmt werden, dass die resultierenden Formulierungen als „Übersetzungshilfe“ oder „Glossar“ zum offiziellen Beurteilungsraster fungieren. Dieses bleibt formalrechtlich verbindlich, aber die Ausein-

Tab. 15: Mögliche Rollenverteilungen bei der kooperativen Erarbeitung von Beurteilungsrastern (in Anlehnung an Stevens und Levi 2013, 88)

Stufen der kooperativen Erarbeitung von Beurteilungsrastern					
	Stufe	Lernziele abgleichen	Lernziele auflisten	Dimensionen erarbeiten	Raster erstellen
Kontrollschwerpunkt Lehrende Studierende	1	Lehrende	Lehrende	Lehrende	Lehrende; Studierende reflektieren und klären ihr Verständnis
	2	Lehrende	Lehrende	Lehrende	Lehrende; Studierende machen Anpassungsvorschläge
	3	Lehrende	Lehrende, Studierende	Lehrende und Studierende gruppieren Vorschläge d. Studierenden	Lehrende und Studierende erstellen definitives Raster
	4	Lehrende	Studierende (Lehrende moderieren)	Studierende (Lehrende moderieren)	Lehrende und Studierende erstellen definitives Raster
	5	Lehrende + Studierende	Studierende (Lehrende moderieren)	Studierende (Lehrende moderieren)	Studierende (Lehrende moderieren)

andersetzung der Studierenden mit den Beurteilungskriterien ist für sie dennoch lernwirksam und motivierend, da sie dadurch nicht nur besser verstehen, welche Lernziele sie erreichen sollen, sondern auch erfahren, dass es den Lehrenden ein Anliegen ist, sich mit ihnen über die Lernziele und die damit verbundenen Ansprüche zu verständigen. Das erfüllt die Bedürfnisse der Selbstbestimmung (siehe Kap. 4.5.2) und steigert somit die intrinsische Motivation der Studierenden in Bezug auf die zu beurteilende Aufgabenstellung. Noch ausgeprägter wird die intrinsische Motivation gesteigert, wenn die Studierenden das Beurteilungsraster nicht nur „autonom nachvollziehen“, sondern tatsächlich in wesentlichen Teilen mitproduzieren können. Diese positive Wirkung wird durch explorative Studien zu Verfahren bestätigt, in denen Bewertungskriterien (Kearney 2013, 887f.) bzw. Beurteilungsraster (Lewis, Berghoff u. Pheeny 1999, 193–95) gemeinsam mit Studierenden erarbeitet wurden.

Stevens u. Levi (2013, 87–102) unterscheiden fünf Abstufungen, wie intensiv Studierende in die Entwicklung von Beurteilungsrastern einbezogen werden können (siehe Tab. 15). Während die Studierenden in der ersten Stufe das von den Lehrenden definitiv erstellte Beurteilungsraster genau lesen, diskutieren und Fragen dazu stellen, erarbeiten sie in der letzten Stufe fast

alle Bestandteile des Rasters. Je mehr der inhaltliche Kontrollschwerpunkt an die Studierenden übergeht, desto mehr beschränken sich die Lehrenden auf die Moderation des Erarbeitungsprozesses.

Wie die Stufen 1, 3 und 5 aus Tabelle 15 konkret umgesetzt werden können, wird nachfolgend kurz erläutert. Es handelt sich dabei um erfahrungsbasierte Vorschläge, die auch abgewandelt oder kombiniert werden können. Detaillierte Hinweise zu den Möglichkeiten, auch für die Stufen 2 und 4, finden Sie bei Stevens u. Levi (2013, 87–102), auf deren Erläuterungen die nachfolgenden Kurzdarstellungen im Wesentlichen beruhen.¹³⁶

Fertiges Beurteilungsraster gemeinsam besprechen

Bereits festgelegte Beurteilungsraster mit den Studierenden eingehend zu besprechen, ist der Minimalstandard beim Einsatz dieses Beurteilungsinstruments. Wie erläutert hat dieses Vorgehen gegenüber einem reinen Austeilen der Kriterien diverse Vorteile. Die dafür in der Lehrveranstaltung investierte Präsenzzeit lohnt sich deshalb nicht nur mit Blick auf das verbesserte Lernen der Studierenden. Sie führt auch zu einer später erfolgenden Zeitersparnis der Lehrenden, indem die größere Klarheit bezüglich der Lernziele zu weniger Rückfragen der Studierenden führt.

Vorgehen: Das Beurteilungsraster wird auf Papier ausgeteilt oder online zur Verfügung gestellt und die Studierenden erhalten Zeit, um es durchzulesen. Anschließend können sie im Plenum Verständnisfragen stellen. Hierfür sollte einiges an Zeit zur Verfügung gestellt werden, und es lohnt sich, ein bis zwei anregende oder gar provokative Fragen bereit zu haben, falls die Studierenden sich anfänglich nicht zu melden getrauen, etwa aus Angst „blöde“ Fragen zu stellen. Die Fragen der Studierenden können verschiedene Aspekte betreffen. Manchmal kommen Unklarheiten zum Vorschein, deren Klärung wichtig ist, damit alle Studierenden verstehen, welche Anforderungen sie erfüllen sollen. Für diese einfachste Möglichkeit des Vorgehens sollte ungefähr eine halbe Stunde Zeit eingeplant werden.

Wichtig ist, dass wir als Lehrende auch auf scheinbar banale Fragen geduldig reagieren und erklären, was gemeint ist. Falls zum Beispiel bei einem Beurteilungsraster für das Schreiben einer wissenschaftlichen Arbeit Fragen zum im Raster formulierten Anspruch an wissenschaftlich korrektes Zitieren gestellt werden, kann dies ein Indiz sein, dass nicht alle Studierenden das Vorwissen mitbringen, von dem wir ausgegangen sind. Dies gibt uns die Gelegenheit zu überlegen, wie wir mit diesem Umstand umgehen wollen.

Eine zusätzliche Möglichkeit besteht darin, dass die Studierenden das Kriterienraster zuerst in Gruppen von ca. 4 Studierenden diskutieren und ihre

136 Dabei wurden gewisse Aspekte vereinfacht oder an Gepflogenheiten des deutschen Sprachraums angepasst.

Fragen notieren, bevor die Gruppen diese im Plenum stellen. Gewisse Fragen können dadurch schon in Peer-Diskussionen geklärt werden, und es können sich auch Studierende an der Diskussion beteiligen, die im Plenum zurückhaltend sind.

Als weitere Möglichkeit, die kombiniert werden kann mit der Gruppendiskussion, können wir den Studierenden Beispiele von früheren Bearbeitungen zu diesem Leistungsnachweis zur Verfügung stellen und sie auffordern, die Beispiele anhand des Rasters zu beurteilen (Nicol 2010, 505f.; Sadler 1989). Bei den Bearbeitungen kann es sich um Texte, Code, Tabellen, künstlerische Artefakte oder ggf. auch Aufzeichnungen von Vorträgen oder künstlerischen Aufführungen handeln (zu denen natürlich das Einverständnis der Aufgezeichneten eingeholt werden muss). Durch die Auseinandersetzung mit früheren Leistungen müssen die Studierenden einen Perspektivenwechsel von der Situation der beurteilten zu jener der beurteilenden Person vornehmen. Dies bedingt kognitive Umformungsprozesse, die das Verständnis der Kriterien begünstigen, und kann auch das Verständnis für die anspruchsvolle Tätigkeit des Beurteilens fördern (vgl. Zimmermann 2014, 30). Es empfiehlt sich, für dieses Vorgehen ein gelungenes und ein nicht gelungenes Beispiel zur Verfügung zu stellen. Durch die Kontrastierung der Beispiele können die Studierenden besser herausarbeiten, woran sich das Erreichen von Lernzielen erkennen lässt bzw. welche Merkmale Anzeichen für deren Nichterfüllung sind.

Beurteilungsdimensionen gemeinsam erarbeiten

Die bereits angeführten Vorteile des gemeinsamen Besprechens können noch vertieft werden, wenn Elemente des Rasters gemeinsam mit den Studierenden erarbeitet werden. Das bietet eine gute Balance zwischen der Beteiligung von Dozierenden und jener von Studierenden. Dieses Vorgehen funktioniert am einfachsten mit Studierenden, die bereits mit dem Instrument des Beurteilungsrasters vertraut sind, ist aber gerade bei fortgeschrittenen Studierenden oder bei tieferen Teilnehmendenzahlen auch ohne Vorkenntnisse anwendbar.

Vorgehen: Die Lehrenden erläutern zu Beginn sowohl die genaue Aufgabenstellung des Leistungsnachweises als auch den Aufbau und den Verwendungszweck eines Beurteilungsrasters. Dabei kann es sinnvoll sein, Raster aus anderen Veranstaltungen als Beispiel zu zeigen oder die gewählte Skalierung zu erklären. Anschließend fordern die Lehrenden die Studierenden auf, mit Bezug auf die vorliegenden Lernziele die möglichen Erwartungen an die Bearbeitung des Leistungsnachweises zu formulieren. Dies kann je nach Studierendenzahl im Plenum oder in Gruppen mit anschließender Sammlung im Plenum erfolgen. Die Lehrenden gruppieren dann diese Erwartungen zu Dimensionen und suchen gemeinsam mit den Studierenden nach möglichst prägnanten Benennungen für diese. Danach erhalten alle Studie-

renden 3 bis 5 Notizzettel (oder eine äquivalente digitale Möglichkeit). Sie notieren nun auf jeden Zettel ein Merkmal, anhand dessen sich die Erfüllung der obersten Anspruchsstufe einer Dimension erkennen lässt.

Diese Vorschläge können anschließend entweder direkt in der Veranstaltung ausgewertet und zu Beschreibungen der Skalenstufen in jeder Dimension verarbeitet werden. Oder die Lehrenden nehmen die Vorschläge mit und erarbeiten das Beurteilungsraster auf ihrer Basis außerhalb der Veranstaltung und bringen es in der nächsten Präsenzsitzung zurück. Besonders für den Fall, dass die Lehrenden das Raster in Einzelarbeit erstellen, empfehlen Stevens u. Levi (2013, 95f.), möglichst viele Formulierungen der Studierenden einzubauen. So erkennen die Studierenden, dass ihr Engagement tatsächlich berücksichtigt wird, was die Legitimität des Rasters und der Vorgehensweise erhöht. An gleicher Stelle betonen Stevens u. Levi auch, dass entgegen gelegentlichen Befürchtungen von Lehrenden die Studierenden bei diesem Vorgehen kaum je wesentliche Beurteilungsdimensionen übersehen. Und falls es doch einmal vorkommt, können die Dozierenden die entsprechende Dimension ergänzen und dies begründen.

Ab der Besprechung des „fertigen“ Rasters wird so verfahren wie im Abschnitt zum Besprechen von fertigen Beurteilungsrastern erläutert. Besonders lohnend ist es, das gemeinsam erstellte Beurteilungsraster wie dort vorgeschlagen auf Beispielleistungen früherer Studierenden anzuwenden und damit auch praktisch zu validieren.

Dieses Vorgehen ist zeitlich aufwändiger als das reine Besprechen eines vorgegebenen Rasters. Die zur Erarbeitung der Dimensionen und ihrer verschiedenen Stufen erforderlichen kognitiven Prozesse und der intensive soziale Austausch machen dieses Vorgehen aber zugleich zu einem Teil des inhaltlichen Lernprozesses und bringen somit inhaltliches Lernen und fachliche Beurteilung von Lernleistungen und Produkten näher zusammen. Insofern ist diese Zeit in der Regel gut investiert, indem sie nachhaltigeres Lernen bewirkt.

Gesamtes Beurteilungsraster wird durch Studierende erarbeitet

Die beiden entscheidenden Unterschiede zur vorangehenden Variante bestehen darin, dass bei dieser Option die gesamte Erarbeitung des Rasters durch die Studierenden erfolgt – und damit auch die Skalierung des Beurteilungsrasters. Die Lehrenden übergeben die inhaltliche Kontrolle zu einem großen Teil an die Studierenden. Sie haben aber natürlich immer noch erheblichen Einfluss auf den Leistungsnachweis und damit auf die Beurteilung, da sie die Aufgabenstellung und in aller Regel ebenfalls die Lernziele vorgeben. Zudem beinhaltet auch die Moderation die eine oder andere Möglichkeit, Anregungen einzubringen, falls dies dringend nötig scheint. Diese Variante zielt darauf, die motivationalen und lernpsychologischen Vorteile einer starken

Selbstbestimmung der Studierenden zu nutzen. Dafür muss noch etwas mehr Zeit investiert werden als beim zuvor erläuterten Vorgehen, allerdings fallen auch Lerngewinn und Motivationszuwachs der Studierenden noch höher aus.

Zu betonen ist, dass dieses Vorgehen ebenfalls weit entfernt ist von einem „Anything Goes“, da sich die Studierendengruppe gemeinsam auf Kriterien einigen muss, die dann für alle gelten. Nicht selten erweisen sich die Studierenden dabei sogar als anspruchsvoller denn die Lehrenden, was für diese eine Rollenentlastung bewirken kann (siehe Kap 3.3.1). Sie müssen plötzlich nicht mehr mahnend auf Ansprüchen beharren, sondern können Studierenden gelegentlich empfehlen, nicht zu hohe Ansprüche zu stellen.

Vorgehen: Stevens u. Levi (2013, 99–102) schlagen ein iteratives Verfahren vor. Dabei erarbeiten die Studierenden zunächst in Gruppen von ungefähr vier Personen jeweils vier Beurteilungsdimensionen. Jene Dimension, die ihnen am wichtigsten scheint, wird von einer Sprecherin im Plenum vorgestellt. Im Anschluss werden in einem demokratischen Verfahren durch das Plenum die wichtigsten Dimensionen bestimmt, die dann definitiv in das Raster kommen. Danach kehren die Studierenden in ihre Gruppen zurück und erarbeiten die Beschreibungen der Skalenstufen pro Dimension. Auch diese Formulierungen werden wieder im Plenum vorgestellt, und durch ein demokratisches Vorgehen werden die besten Beschreibungen pro Rasterfeld bestimmt. Je nachdem können zum Abschluss die Studierendengruppen noch Benennungen für die Skalenstufen vorschlagen und demokratisch bestimmen. Dazu lohnt es sich, den Studierenden Hinweise zur Benennung der Stufen zu geben, wie sie in Kapitel 9.3.1 im Abschnitt „Skala“ erläutert sind.

Stevens u. Levi (2013, 100f.) schlagen vor, bei diesem Vorgehen sowohl die Anzahl der Dimensionen als auch die Skalenstufen von vornherein auf je vier festzulegen. Gleich, ob man genau diese Zahlen übernehmen möchte: Eine Setzung dieser Größen reduziert die Komplexität des Vorgehens, da nicht auch noch eine Verständigung über die Anzahl der Dimensionen und Qualitätsabstufungen erfolgen muss.

9.3.5 Andere Beurteilungsinstrumente

Neben Beurteilungsrastern können auch andere Instrumente verwendet werden, um erhobene Leistungen in Bewertungen zu überführen. Während Beurteilungsraster für alle Formen von offenen Aufgabenstellungen geeignet sind, kommen andere Instrumente primär bei geschlossenen und halboffenen Aufgabenstellungen zur Anwendung. Dazu gehören namentlich die an Hochschulen vor allem in gewissen Fächern verbreiteten Multiple-Choice-Aufgaben (siehe Kap. 13) sowie halboffene Formate wie Lückentexte oder geschlossene Freiantworten (siehe Kap. 7.2.1 zu den verschiedenen Formaten von geschlossenen, halboffenen und offenen Aufgaben).

In diesen Fällen wird man in der Regel Punkte für richtige Lösungen vergeben. Wenn nicht die ganze Lösung korrekt ist, werden gelegentlich auch Punkte (oder Bruchteile von Punkten) für richtige Lösungsteile vergeben. Hier ist jeweils zu entscheiden, für welche Art von richtigen Antworten bzw. Antwortteilen wie viele Punkte zu vergeben sind. Dabei ist insbesondere die Schwierigkeit der Aufgaben zu berücksichtigen. Möglicherweise sind aber auch weitere Kriterien zu beachten wie die zur Lösung benötigte Zeit oder die Wichtigkeit, die dieser Art von Wissen oder Können im Rahmen der Lernziele zugemessen wird. Bei allen diesen Entscheidungen handelt es sich um Ermessensentscheide und nicht um Eigenschaften, die den Aufgabenstellungen oder den in ihnen angesprochenen Sachinhalten innewohnen. Im folgenden Unterkapitel wird ausführlich dargelegt, inwiefern dies wesentliche Folgen für das Vorgehen hat, wie auf der Basis von Punkten Noten erteilt werden.

An dieser Stelle sei auch nochmals darauf hingewiesen, dass sich geschlossene und halboffene Aufgaben wesentlich schlechter zur Überprüfung hochrangiger Lernziele eignen (siehe Kap. 7.2.1).

9.4 „Berechnung“ von Noten

Vielleicht fragen Sie sich, weshalb das Wort Berechnung in der Kapitelüberschrift in Anführungszeichen steht – Noten kann man doch berechnen, es sind doch Zahlen? Weshalb dies ein mathematischer Kurzschluss ist, der zu vielen Missverständnissen und dysfunktionalen Praktiken im Umgang mit Noten führt, erläutert Kapitel 9.4.1. Auf dieser Grundlage wird die Anschlussfrage nach der Gewichtung verschiedener Aufgaben diskutiert (Kap. 9.4.2), um sonach Beurteilungsraster als Bewertungsinstrumente und Notenskalen näher zu betrachten (Kap. 9.4.3. und 9.4.4). Schließlich werden Benotungsmodelle (Kap. 9.4.5) und die Herausforderungen bei der Berechnung von Noten anhand von Notenskalen dargelegt (Kap. 9.4.6).

9.4.1 *Mathematische Probleme bei der „Berechnung“ von Noten*

In seinem lesenswerten Beitrag zur mathematischen Kritik an den Ziffernnoten und ihrer Interpretation geht Fischer (1991, 228–236) auf die Frage ein, welcher Art von Skalentyp Noten angehören. In der Mathematik unterscheidet man vier verschiedene Skalentypen (siehe Tab. 16): Nominalskala, Ordinalskala, Intervallskala und Verhältnisskala (vgl. z. B. Schäfer 2016, 24–29).

Entscheidend ist für unseren Zusammenhang, dass für Werte in nominalen und ordinalen Skalen keine algebraischen Beziehungen existieren. Man kann mit solchen Werten also nicht rechnen wie mit natürlichen, ganzen, gebrochenen oder reellen Zahlen. Ich kann zum Beispiel nicht den Durchschnitt aus zwei Nominalwerten wie Postleitzahlen bilden. Und auch bei ordinalen Ordnungen ist das nicht möglich: Ich kann zwar sagen, dass eine Kundenbewertung mit 4 von 5 Sternen um drei Stufen größer ist als eine mit nur einem Stern – aber nicht, dass sie 4x größer ist. Auch Schulnoten sind ordinalskaliert: Eine 4 ist nicht dreimal schlechter als eine 1 (Fischer 1991, 229f.). Besser sichtbar wird dies etwa im US-Notensystem, das die Stufen von A (beste Bewertung), B, C, D und F umfasst. Dabei wird das E ausgelassen, A–D stehen für genügende Noten und F ist die einzige ungenügende Stufe.¹³⁷ Wenn diese Stufen A, D, E, P, T lauten würden, wäre das genauso passend.

Tab. 16: Eigenschaften von Skalenarten (vgl. Schäfer 2016, 28, und Bortz 2005, 23)

Eigenschaften von Skalenarten				
Skalentyp	Mess-niveau	Mögliche Aussagen	Mögl. Operationen	Beispiele; Maß für Mittelwert
Nominalskala	nominal	Gleichheit Verschiedenheit	= / ≠	Autonummern Familienstand Mittelwert: Modus
Ordinalskala	ordinal	Größer-kleiner-Relationen	= / ≠ < / >	Tabellenplatz in Sport-Ligen, Schulnoten Mittelwert: Median
Intervallskala	metrisch	Gleichheit von Differenzen	= / ≠ < / > + / –	Temperatur in Grad Celsius, Zeitdauer Mittelwert: Arithmetisches Mittel
Verhältnis-skala	metrisch	Gleichheit von Verhältnissen	= / ≠ < / > + / – • / ÷	Alter in Jahren Temperatur in Kelvin Mittelwert: Geometrisches Mittel

137 Allerdings ist es im US-System gebräuchlich, Prozentsätze richtiger Antworten in diese letter grades umzuwandeln. 90–100 % gibt dann oft ein A, 80–89 % ein B, 70–79 % ein C, 60–69 % ein D und alles unter 60 % ist ein F. Auch hier wird also mit Zahlen operiert und es stellt sich ebenfalls ein Gewichtungungsproblem, siehe Kap. 9.2.1 und Kap. 9.4.2.

Obschon Noten also lediglich eine Rangordnung signalisieren – eine 1 ist besser als eine 2, eine 2 ist besser als eine 3 etc. – werden sie häufig „stillschweigend als (natürliche) Zahlen umgedeutet dadurch, dass man mit ihnen rechnet. Man bestimmt aus ihnen z. B. (nicht-ganzzahlige) Mittelwerte. (...) Der rechnerische Umgang mit den Notenwerten („Ziffernnoten“) ist aus verschiedenen skalentheoretischen Gründen unzulässig. Man wechselt dabei stillschweigend (...) vom Ordinalskalenniveau auf ein Skalenniveau, das man der Benotung nicht zugrunde gelegt hatte“ (Fischer 1991, 234f.). So ist es unzulässig, für ordinale Werte ein arithmetisches Mittel zu bilden. Der korrekte Mittelwert für Werte dieses Skalenniveaus ist der Median. „Nicht genug: Nach Berechnung arithmetischer bzw. gewogener Mittel liest und denkt man diese Mittelwerte wieder als Elemente einer (nunmehr verfeinerten) Rangskala, d. h. man interpretiert sie wieder unter ordinalen Aspekten. Hin und her, hinauf, herab in der Skalenniveau-Hierarchie, wie es gerade ins naive Konzept passt“ (Fischer 1991, 235f.).

Fischer kritisiert weitere mathematische Unzulänglichkeiten der gängigen Praktiken, mit denen verschiedene Noten aggregiert oder gewichtet werden. Daraus soll noch eine herausgegriffen werden, die vielen Lehrenden aus ihrer Bewertungserfahrung bekannt ist: Die arithmetische Verrechnung ordinaler Benotungen, als wären es metrische Daten, schaukelt eine Genauigkeit vor, die es so nicht gibt. Vielmehr kommt es genau dadurch zu Verzerrungen: Bekommt Studentin A ganz knapp nicht mehr eine 1, sondern eine 2, und Student B ganz knapp noch eine 2 statt eine 3, so haben beide die gleiche ordinale Bewertung. Wenn diese nun arithmetisch mit weiteren Werten, z. B. einer weiteren 2, gemittelt und dieser Mittelwert als Endnote ausgewiesen wird, bekommt Studentin A eine zu schlechte und Student B eine zu gute Note (Fischer 1991, 241f.).

Trotz der massiven Defizite, die er in der gängigen Benotungspraxis ausmacht, muss Fischer (1991, 244) eingestehen: „Auch wir vermögen keine Alternativen zu nennen, die die angesprochenen Mängel und Unzulänglichkeiten der gängigen Notenpraxis vermeiden, dabei formal fundierbar und zugleich ebenso leicht lesbar und zu handhaben sind, wie die aus der üblichen Notengebung resultierenden Zeugnisse.“ Er formuliert dann einige Alternativen, die allerdings deutlich komplexer ausfallen: Begriffliche Skalentypen, nicht-lineare Rangskalen oder eine nicht-archimedische Skalierung werden wir hier aber nicht näher verfolgen. Noch am ehesten praktikabel scheint der Vorschlag eines dreidimensionalen Notensystems, in dem eine Leistung jeweils anhand der drei Bezugsnormen Individualnorm, Kriterialnorm und Kollektivnorm (vgl. Kap. 6.2) benotet würde. Die drei Werte würden dann an-

hand eines Koordinatensystems mit drei Achsen dreidimensional abgebildet (vgl. Fischer 1991, 244f.).¹³⁸

Man kann somit festhalten: Eine mathematisch überzeugende Alternative zu Notenbewertungen wurde bisher nicht gefunden. Fischer selbst spricht von einem Schulnoten-Paradoxon: „Im Bemühen um Gleichberechtigung, Vergleichbarkeit und Gerechtigkeit verfehlt der Benotungsprozess eben diese Ziele. Hinzukommt, dass durch die weitgehende Operationalisierung des Unterrichts die Entwicklung kreativen Denkens blockiert wird“ (Fischer 1991, 246; zum kreativen Denken siehe Kap. 7.1.6).

Angesichts der Faktenlage würde aus meiner Sicht sehr viel dafür sprechen, mehr auf qualitativ gut begründete Urteile zu setzen, statt sich hinter komplizierten Berechnungen zu verstecken, die eine Pseudogenauigkeit und damit Scheingerechtigkeit schaffen, die letztlich in erster Linie herrschende Ungleichheiten aufrechterhalten (siehe dazu Kap. 2). So lange die herrschende Praxis Noten verlangt, sind allerdings Instrumente zu suchen, welche die besprochenen Schwächen der Benotung wenigstens in Teilen überbrücken oder zumindest lindern können. Vorschläge dazu werden in den folgenden Unterkapiteln erläutert. In Kapitel 9.5 werden Basisbewertungen und Prädikate als alternative Bewertungsmöglichkeiten und in Kapitel 9.6 mit dem Notenvertrag und mit Selbstbewertungen zwei grundlegendere Alternativen für die summative Beurteilung vorgestellt.

9.4.2 Gewichtung von Aufgaben

Die Frage, in welchem Verhältnis verschiedene Aufgaben zueinander stehen, wurde bereits im Rahmen der Konzeption von Leistungsbeurteilung als Messen und Vergleichen angesprochen (Kap. 9.2.1). Sie wird hier nochmals aufgegriffen, da sie relevant ist dafür, wie von der Bearbeitung einzelner Aufgaben eine Gesamtbewertung abgeleitet wird. Ähnliches gilt für verschiedene Leistungsdimensionen einer größeren Aufgabenstellung wie einer wissenschaftlichen Arbeit, einer Fallbearbeitung, Projektarbeit etc.

Wir haben bereits festgehalten, dass eine Aufgabe zu einer anderen nicht exakt gleichwertig ist und damit keine Äquivalenzrelation zwischen ihnen besteht: Wir können empirisch nicht nachweisen, dass $a = b$ oder $b = c$, und schon gar nicht, dass $a = c$. Gleichermaßen kann man nicht sagen, welche

138 Dieser Vorschlag erinnert an die nicht-lineare zweidimensionale Bewertung von Ruf u. Gallin (2003, 85–89). Sie führt eine Note, die sich an der Individualnorm ausrichtet, mit einer klassischen Schulnote zusammen. Eine so genannte „Schmetterlingsformel“ sorgt dafür, dass die bessere der beiden Durchschnittsnoten stärker gewichtet wird, wobei dieser Effekt umso stärker ausfällt, je größer der Abstand zwischen den beiden Noten ist. Allerdings stellt sich hier wieder das Problem, dass ordinale Werte arithmetisch verrechnet werden.

Relation zwischen zwei oder mehr Leistungsdimensionen der gleichen komplexen Aufgabe besteht: Was ist das Verhältnis zwischen der Dimension Textorganisation und der Dimension Umgang mit Fachliteratur? In der Praxis behilft man sich in beiden Fällen in der Regel mit Gewichtungen. In einer Multiple-Choice-Klausur werden dann richtige Lösungen von einfacheren Aufgaben zum Beispiel mit einem Punkt bewertet, während es für solche von komplexeren Aufgaben zwei oder mehr Punkte gibt (was aber bei MC-Aufgaben problematisch ist, siehe Kap. 13.5). Ähnlich sehen die Verfahren bei Beurteilungsrastern aus, indem verschiedene Leistungsdimensionen durch unterschiedliche Prozentzahlen oder Punktzahlen gewichtet werden (siehe Kap. 9.3.3).

In vielen dieser Fälle fehlt allerdings das Bewusstsein dafür, dass es sich bei Quantifizierungen solcher Beziehungen um willkürliche Ermessensentscheidungen handelt, oder es geht im späteren Berechnungsprozess von Punkten und Noten wieder verloren. Im Bereich von Multiple-Choice-Prüfungen werden Beziehungen zwischen Aufgaben durch Konstrukte wie die Aufgabenschwierigkeit oder die Trennschärfe von Aufgaben quantifiziert (für eine knappe Darstellung vgl. Sacher 2014, 132–37). Dabei handelt es sich um einen einfacheren Anwendungsfall der Item-Response-Theory (IRT), die darauf zielt, die latenten Strukturen hinter der Oberfläche von Daten zu erkennen. Dieser theoriegeleitete Ansatz versucht, die zugrunde liegenden Merkmale oder Fähigkeiten aufzudecken, die die Antworten von Personen auf Testaufgaben beeinflussen. Die IRT geht eben gerade davon aus, dass die Antworten zu einzelnen Fragen nicht gleichwertig sind. Ältere Ansätze der IRT zielten primär darauf, externe Störquellen zu eliminieren und dadurch eine Vergleichbarkeit verschiedener Aufgaben herzustellen. Konstrukte wie die Aufgabenschwierigkeit oder die Trennschärfe basieren auf dieser Zielsetzung. Demgegenüber verfolgen neuere IRT-Modelle andere Zielsetzungen und versuchen etwa Problemlösungsprozesse zu modellieren (vgl. zu diesen Ausführungen van der Linden u. Hambleton 1997): „Anstatt die Fähigkeitsmessungen von den Auswirkungen der Störvariablen zu befreien, hat sich das Interesse (...) auf das Verhalten der Störvariablen selbst verlagert“ (van der Linden u. Hambleton 1997, 22, Übers. tz).

Das von Fischer (1991) beschriebene Problem vermögen aber auch die Ansätze der IRT nicht zu lösen: Die Eigenschaften verschiedener Aufgaben lassen sich nicht direkt quantifizieren. Dass anhand der sichtbaren Daten auf latente Eigenschaften der Aufgabenlösenden geschlossen werden muss, um ein tieferes Verständnis für den Messprozess und die zu bewertenden Aufgaben zu erlangen, bestätigt dieses Problem nur. Die IRT ist zwar hilfreich, um die Eigenschaften von Prüfungsaufgaben zu explorieren und dadurch möglichst valide Tests zusammenzustellen, die auf dem Verständnis von Leistungsbewertung als Messen und Vergleichen beruhen (siehe Kap. 9.2.1). Ihre Verfahren sind aber außerordentlich aufwändig und funktionieren nur

zuverlässig mit größeren Teilnehmendenzahlen, je nach Modell zwischen $n > 50$ und $n > 500$ –1000 (Embretson u. Reise 2015). Sie lohnen sich, wenn überhaupt, nur für Aufgabenbatterien als Basis von Klausuren mit großer Tragweite in Massenfächern, in denen jährlich mehrere hundert Studierende geprüft werden.¹³⁹ Für summative Leistungsnachweise in Fächern und Modulen mit tieferen Teilnehmendenzahlen eignen sie sich aus Ressourcengründen nicht. In diesen Fällen ist deshalb wichtig, dass sich die Bewertenden im Rahmen des Bewertungsprozesses bewusst sind, dass Gewichtungen wie Prozentzahlen oder unterschiedliche Punkte bestenfalls heuristische Schätzungen der Bedeutsamkeit der adressierten Lernziele oder der Schwierigkeitsgrade der bewerteten Aufgaben darstellen und ihrerseits wiederum das Lern- und Lösungsverhalten der Studierenden beeinflussen können. Unter Ausklammerung der Ansätze der IRT zielen deshalb die restlichen Ausführungen in diesem Kapitel darauf, Benotungen und Prädikate als inhaltlich begründbare Urteile zu vergeben.

9.4.3 Notengebung anhand von Beurteilungsrastern

Bestehen Leistungsnachweise aus einer oder wenigen (in diesem Fall meist eher komplexen) Aufgabenstellung(en), eignen sich wie in Kapitel 9.3.3 besprochen Beurteilungsraster für summative Bewertungen. In jenem Kapitel wurden bereits drei Möglichkeiten vorgestellt, wie Beurteilungsraster genutzt werden können, um Notenwerte abzuleiten. In der ersten Variante werden den verschiedenen Skalenstufen direkt Notenwerte zugeordnet. Die beiden anderen Varianten funktionieren über die Vergabe von Punkten, die wiederum in Noten umgewandelt werden müssen. Alle drei Varianten werden nun näher im Hinblick auf die Umwandlung der Rasterbewertung in ordinale Werte erläutert.

Mit Notenwerten versehene Skalenstufen

Weil es so zentral ist, sei es hier wiederholt: Wenn Noten vergeben werden müssen, so ist die tiefste genügende Note gleichzusetzen mit dem Erreichen

139 Streng genommen müssten dazu die entwickelten Aufgaben jeweils anhand von Testprüfungen mit Ansätzen der IRT ausgewertet werden und die Aufgabenbatterie aufgrund der Befunde erstellt werden. Methodisch denkbar wäre auch, die Aufgaben einer durchgeführten Klausur auszuwerten und auf dieser Basis problematische Aufgaben auszuschließen. Letzteres empfiehlt auch Krebs (2019, 122; siehe Kap. 13.5), es widerspricht aber möglicherweise zumindest in Deutschland den herrschenden Rechtsgrundsätzen, die etwa für Multiple-Choice-Prüfungen die Festlegung von Bestehensgrenzen im Rahmen der Prüfungsordnungen vorsehen (vgl. Hirdina 2019, 53f.).

des minimalen Lernziels. Diese Setzung hilft, entsprechende Kalibrierungen vorzunehmen. In Tabelle 12 wurde zum Beispiel entschieden, dass die mittlere Stufe im deutschen Notensystem als „befriedigend“ = Ziffernote 3 gewertet werden soll und die gerade noch genügende Note sich in der Mitte zwischen dieser und der untersten Stufe befindet.

Um mithilfe von Skalenstufen direkt eine Benotung vorzunehmen, muss ein Notenmittelwert anhand der durch die Beurteilung pro Dimension vergebenen Skalenstufen gebildet werden. Dafür kommen zwei Mittelwerte in Frage: Mathematisch angemessen wäre es, den Median zu verwenden, da es sich bei Noten wie in Kapitel 9.4.1 ausgeführt um ordinale Werte handelt. Will man eher dem mathematisch unangemessenen, aber gebräuchlichen Vorgehen folgen, bildet man den arithmetischen Mittelwert anhand der Notenzahlen. Aus meiner Sicht spricht einiges dafür, den Median zu verwenden, da dadurch einzelne Ausreißer weniger gewichtet werden. Tabelle 12 veranschaulicht die beiden Vorgehensweisen (siehe Kap. 9.3.3).

Bilden wir anhand der vorgenommenen Gewichtungen aus der dort abgebildeten Beurteilung das arithmetische Mittel, so ist zu rechnen:

$$\text{Arithmetisches Mittel} = (4 \times 1 + 2 \times 3 + 3 \times 1 + 5) / 10 = 1,8$$

Wären die vier Dimensionen nicht gewichtet, so betrüge das arithmetische Mittel hingegen: $(1 + 3 + 1 + 5) / 4 = 2,5$

Der Median wird durch jene ordinale Variable gebildet, welche die Grenze zwischen der oberen und der unteren Hälfte der Stichprobe bildet. Im vorliegenden Fall wären das infolge der Gewichtungen die folgenden Werte: 1, 1, 1, 1, 3, 3, 1, 1, 1, 5. Ordnen wir sie, erhalten wir:

Geordnete Stichprobe: 1, 1, 1, 1, 1, 1, 3, 3, 3, 5.

Ist die Anzahl der Werte gerade, bildet man den Median als arithmetisches Mittel der beiden mittleren Werte, die hier beide 1 sind – also liegt der Median bei 1. Anders wäre es übrigens, wenn die vier Dimensionen gleich gewichtet wären. Dann lautete die geordnete Stichprobe 1, 1, 3, 5 und der Median wäre das arithmetische Mittel aus 1 und 3, läge also bei 2.

Wichtig ist der Hinweis, dass bei anderen Verteilungen durchaus auch der Median schlechter ausfallen kann als das arithmetische Mittel, er ist also keineswegs eine mildere Form der Bewertung, sondern trägt lediglich der mathematischen Tatsache besser Rechnung, dass die Einschätzungen pro Dimensionen keine Messungen auf metrischem Niveau, sondern nur Rangreihen darstellen.

Werden statt Noten die Prädikate „bestanden“/„nicht bestanden“ (pass/fail) vergeben, wird im Prinzip gleich vorgegangen, nur dass in diesem Fall von der besten bis zur tiefsten genügenden Note/Bewertung das Prädikat „bestanden“ vergeben wird und darunter das Prädikat „nicht bestanden“.

Mit Punkten versehene Skalenstufen

Der indirektere Weg von Beurteilungsrastern zu Noten oder Prädikaten führt über die Vergabe von Punkten. Die zwei für dieses Vorgehen gebräuchlichsten Optionen wurden in Kapitel 9.3.3 vorgestellt. In diesem Fall müssen die erreichten Punktzahlen anhand einer Notenskala in Noten überführt werden. Werden nur die Prädikate „bestanden“/„nicht bestanden“ unterschieden, ist die Punktzahl festzulegen, die gerade noch einen ausreichenden Grad der Lernzielerreichung repräsentiert. Bildet ein einzelnes Beurteilungsraster die Grundlage für die Benotung, muss nun so verfahren werden, wie in den Kapiteln 9.4.5 und 9.4.6 beschrieben. Andernfalls müssen die verschiedenen Punktzahlen noch zusammengeführt werden.

Zusammenführen der Bewertungen aus mehreren Beurteilungsrastern

Beim Zusammenführen der Bewertungen aus mehreren Beurteilungsrastern ist zu unterscheiden zwischen Rastern, die zu Direktbenotungen führen und solchen, die zunächst Punkte vergeben. Die beiden Vorgehensweisen gelten auch grundsätzlich für das Zusammenführen von verschiedenen Noten und verschieden bepunkteten Aufgabenstellungen:

- Raster, die zu Direktbenotungen führen: In diesem Fall ist das Vorgehen zur Zusammenführung vergleichbar mit jenem, bei dem aus benoteten Skalenstufen die Note für das gesamte Beurteilungsraster ermittelt wird. Wie oben erläutert ist in diesem Fall das Verfahren zu wählen, mit dem der Mittelwert bestimmt wird. Auch hier ist aus meiner Sicht der Median gegenüber dem arithmetischen Mittel zu bevorzugen, da wir es wie bei allen Noten mit ordinal skalierten Werten zu tun haben. Zudem ist vor der Bildung des Mittelwerts festzulegen, ob die Noten aus allen Rastern gleich oder unterschiedlich gewichtet werden.
- Raster, mit denen Punkte vergeben werden: In diesem Fall müssen die Punkte der verschiedenen Aufgabenstellungen zusammengezählt werden. Das anschließende Vorgehen ist in den Kapiteln 9.4.5 und 9.4.6 beschrieben.

9.4.4 Zusammenführen der Bewertungen aus verschiedenen Leistungsnachweisen

Werden verschiedene summativ bewertete Einzelleistungsnachweise zu einem Sammelleistungsnachweis zusammengeführt, so gelten im Prinzip die beiden gleichen Vorgehensweisen wie oben für das Zusammenführen von Bewertungen aus mehreren Beurteilungsrastern beschrieben:

- Beim Zusammenführen verschiedener Noten ist gleich vorzugehen wie beim Zusammenführen von Benotungen aus verschiedenen Beurteilungsrastern.
- Werden die zusammenzuführenden Leistungsnachweise nur bepunktet und noch nicht benotet, so ist gleich vorzugehen wie oben beim Zusammenführen von bepunkteten Beurteilungsrastern beschrieben.

9.4.5 Benotungsmodelle

Um Noten zu erteilen, ist in einem ersten Schritt ein Benotungsmodell nötig. Dabei handelt es sich um ein System bzw. eine Methode, durch die Leistungen von Studierenden in Noten überführt werden. Es legt also die Kriterien für die Bewertung und Benotung von Leistungen fest. In Kapitel 9.4.3 wurde deshalb ein Benotungsmodell für direkt benotete Beurteilungsraster vorgeschlagen und der erste Teil eines Benotungsmodells für bepunktete Beurteilungsraster. Auch für das in Kapitel 9.4.4 beschriebene Zusammenführen von verschiedenen Leistungsnachweisen ist ein Benotungsmodell erforderlich.

Ein Benotungsmodell muss einige Anforderungen erfüllen, damit es den Ansprüchen einer fairen und transparenten Benotung genügen kann (Sacher 2014, 99–101; Zimmermann, Tobias 2018, 79f.):

- Die Zuweisung von Noten muss eindeutig sein, gleiche Leistungen müssen immer gleich benotet werden. Für unterschiedliche Leistungen können hingegen durchaus gleiche Noten vergeben werden, da Noten immer einen bestimmten Leistungsbereich zusammenfassen. „Werden 100 Studierende bewertet, kann es 100 verschiedene Leistungsausprägungen geben – werden sie mit Ziffernnoten von 1 bis 6 (mit halben Noten) bewertet, kann es aber nur 11 verschiedene Noten geben“ (Zimmermann, Tobias 2018, 79).
- Die Zuweisung muss der gewählten Bezugsnorm entsprechen. Ein Benotungsmodell hat nur die Leistungsmerkmale zu erheben, die für die gewählte Bezugsnorm relevant sind (siehe Kap. 6.2). „Das bedeutet beispielsweise, dass man bei einer lernzielbezogenen Benotung nicht nach der Korrektur die Benotungsskala nochmals abändern darf, nur weil zu viele Studierende genügend sind und man mit Bezug auf die soziale Bezugsnorm das Gefühl hat, es müssten doch mehr durchfallen“ (Zimmermann, Tobias 2018, 79).
- „Die Zuweisung von Noten sollte fehlerkontrolliert erfolgen. Das bedeutet, dass man sich beim Vergabe von Noten über die damit verbundenen Möglichkeiten von Messfehlern bewusst sein sollte – und Anzahl wie Ausmaß von Messfehlern reduzieren sollte“ (Zimmermann, Tobias 2018,

79). Aus diesem Grund ist es wichtig, die in Kapitel 6.3 erläuterten Gütekriterien für die Qualität von Leistungsnachweisen zu beachten. Ob man sich dabei an quantitativen oder qualitativen Kriterien orientiert, hängt von der gewählten Konzeption von Leistungsbeurteilung ab (siehe Kap. 9.2). Zudem unterliegt die Zuweisung von Noten zu erhobenen Leistungen stets einem gewissen Schätzfehler, der zu berücksichtigen ist (siehe Kap. 9.4.6).

9.4.6 Notenskalen

Werden Punkte oder Fehler in Noten umgewandelt, muss mit einer Notenskala gearbeitet werden. In diesem Fall arbeiten wir mit den Grundannahmen der Konzeption von Leistungsbewertung als Messen und Vergleichen, gehen also quantitativ orientiert vor. Bezüglich Notenskalen sind zwei grundsätzliche Unterscheidungen auszumachen: Es werden Punkte- und Fehlerskalen sowie lineare und nichtlineare Skalen unterschieden (Sacher 2014, 101f.).

Punkte- und Fehlerskalen

In Punkteskalen wird für eine bestimmte Bandbreite an Punkten eine bestimmte Note vergeben. Wie die Punkte genau zustande kommen, wird nicht durch die Notenskala festgelegt, sondern durch das Instrument, das zur Bepunktung führt. Dabei handelt es sich entweder um Bewertungsraster (siehe Kap. 9.3.1, 9.3.3 und 9.4.3) oder um andere Kriterien (siehe Kap. 9.3.5). Auch diese Instrumente sind wesentliche Bestandteile des jeweiligen Benotungsmodells. Aber erst die Notenskala überführt die Punkte schließlich in Noten.

Bei Fehlerskalen wird die Note desto schlechter, je mehr Fehler gemacht werden. Diese Form der Bewertung widerspricht mit ihrer Defizitorientierung wesentlichen pädagogischen Grundsätzen und wirkt demotivierend. Zudem passt sie nicht zur in diesem Buch bevorzugten Konzeption von Leistungsbeurteilung als Überprüfung des Erreichens von Lernzielen – deren Erreichen ist auf positive Weise zu beschreiben und nicht ex negativo durch die Abwesenheit von Fehlern zu definieren. Deshalb werden Fehlerskalen in der Folge nicht weiter behandelt.

Lineare und nicht lineare Punkteskalen

In linearen Punkteskalen besteht ein konstanter Abstand zwischen den Noten, wenn die für die jeweilige Note zu erreichende Mindestpunktzahl auf einer Skala von 0 Punkten bis zum Punktemaximum abgetragen wird. Oft wird die Linearität nur näherungsweise erreicht. An einem Beispiel erklärt: Werden

nur ganze Noten erteilt, gibt es in einem Notensystem von 1–6 fünf Notenabstände, die in der Skala durch Punktbänder repräsentiert sind: 1–2, 2–3, 3–4, 4–5, 5–6. Sind z. B. 29 Punkte zu vergeben, ergibt sich in diesem Notensystem folgende lineare Punkteskala (siehe Abb. 20): 0–4 Punkte = Note 6, 5–9 Punkte = Note 5, 10–14 Punkte = Note 4, 15–19 Punkte = Note 3, 20–24 Punkte = Note 2, 25–29 Punkte = Note 1.¹⁴⁰ Sind z. B. 32 Punkte zu vergeben, gibt es keine Möglichkeit, für alle Noten das Punktbänder so zu definieren, dass alle 6 Punktbänder gleich groß sind. Man wird in diesen Fällen, sofern Linearität angestrebt wird, die Abweichungen der Notenbänder so gering wie möglich halten.

Beabsichtigt nicht lineare Punktzahlen liegen dann vor, wenn die Punktbänder für verschiedene Noten deutlich unterschiedliche Größen haben. Am häufigsten werden nicht lineare Punkteskalen gewählt, um am unteren oder oberen Ende der Skala Gewichtungen vorzunehmen. So könnte man etwa im obigen Beispiel mit maximal 29 Punkten entscheiden, dass es auch mit einigen wenigen Punkten noch die Minimalnote gibt. Zugleich könnte man, etwa bei einer schwierigen Prüfung, entscheiden, dass das Punktbänder für die Bestnote etwas breiter sein soll. Gemäß diesen Überlegungen könnte man die Skala z. B. in folgende nicht lineare Punktbänder einteilen (siehe Abb. 20): 0–8 Punkte = Note 6, 9–13 Punkte = Note 5, 14–16 Punkte = Note 4, 17–19 Punkte = Note 3, 20–22 Punkte = Note 2, 23–29 Punkte = Note 1.

Die beiden hier angeführten Skalen scheinen beide suboptimal. Linearität ist ein mathematisches Kriterium und wird in vielen Fällen dem Anspruchsniveau der Lernziele nicht angemessen sein (Sacher 2014, 102f.). Ebenso fällt es schwer, sich eine Bewertungssituation auszudenken, welche die unterschiedlichen Notenbänder der abgebildeten nicht linearen Skala rechtfertigt. Sehr ungünstig erscheinen insbesondere die schmalen Punktbänder für die Noten 4 bis 2: Erhält man mit 16 Punkten noch eine 4, ist es bei 20 Punkten bereits eine 2. Ob eine darauf basierende Zuteilung der drei betroffenen Notenstufen wirklich zuverlässige Unterscheidungen erlaubt, darf bezweifelt werden.

Sinnvoll scheint hingegen, die Skaleneinteilung von der Überlegung her zu beginnen, welche Punkteleistung erforderlich ist, um gerade noch eine genügende Note zu erhalten (Sacher 2014, 103). Dieser Schritt wird unten näher erläutert. Ein weiterer zweckmäßiger Schritt liegt darin, die Punktzahl festzulegen, die für die Maximalnote erforderlich ist, und jene, unterhalb deren die Minimalnote erteilt werden soll. In vielen Fällen dürfte es ange-

140 Sacher (2014, 101, FN 2) weist darauf hin, dass auch der Fall zu berücksichtigen sei, dass 0 Punkte erreicht werden. Deshalb ergibt sich eine lineare Skala für n Notenstufen immer nur für eine Gesamtpunktzahl, die $(x \text{ mal } n) - 1$ entspricht. Im deutschen und Schweizer System entspricht dies $(x \text{ mal } 6) - 1$.

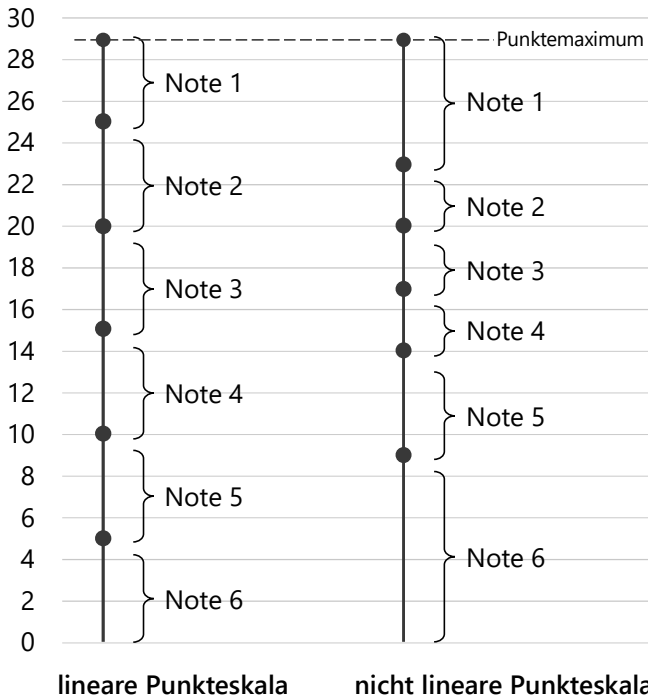


Abb. 20: Lineare und nicht lineare Punkteskalen. Quelle: Eigene Abbildung

messen sein, anschließend den genügenden Skalenbereich gleichmäßig zu unterteilen. Würde man also im Beispiel von Abbildung 20 entscheiden, dass 12 Punkte das Minimum sind, bei dem das Lernziel als erreicht betrachtet werden kann, dass unter 7 Punkten die Minimalnote und ab 24 Punkten die Maximalnote zu vergeben ist, so ergäbe sich folgende nicht lineare Einteilung: 0–6 Punkte = Note 6, 7–11 Punkte = Note 5, 12–15 Punkte = Note 4, 16–19 Punkte = Note 3, 20–23 Punkte = Note 2, 24–29 Punkte = Note 1.

Messfehler und Konfidenzintervall

Die Notengebung anhand von Punkten ist wie bereits mehrfach betont keine genaue Messung, ja eher eine Schätzung. Um in der Konzeption von Leistungsbewertung als Messen und Vergleichen zu bleiben, sprechen wir allerdings von Messungen. Die Messung der Punkte ist nun mit allerlei Mess-

fehlern¹⁴¹ behaftet. Dies beginnt damit, dass trotz aller Bemühungen nicht sicherzustellen ist, dass alle Punkte durch gleichwertige Arten von Wissen und Können zu erwerben sind. Weitere Messfehler rühren von folgenden Ursachen her (Sacher 2014, 104f.):

- Die testtheoretischen Gütekriterien, auf denen Messungen gemäß dieser Konzeption beruhen, können in sozialen Kontexten immer nur approximativ erfüllt werden (siehe Kap. 6.3.1).
- Punkte werden vergeben, wenn bestimmte Leistungsschwellen überschritten werden. Auch wenn für eine Aufgabe keine oder nicht alle Punkte vergeben werden, kann es sein, dass die Studentin durchaus über relevante Aspekte von Wissen und Können verfügt, die sie aber bei der Bearbeitung der konkreten Aufgabe nicht ausreichend sichtbar machen konnte, um (alle) Punkte zu erwerben.
- Ein damit verwandtes Problem kann man als Rundungsproblem bezeichnen: Wenn ein Student gewisse, aber nicht alle für das Erwerben eines Punktes erforderlichen Leistungsmerkmale zeigt, so ist es ein Ermessensentscheid, ob der Punkt nun erteilt werden soll.
- Jede Auswahl von Aufgaben kann das Wissen und Können, das sie überprüfen soll, nur teilweise überprüfen, es handelt sich gewissermaßen um eine Stichprobenziehung. Als Folge davon ergibt sich ein Stichprobenfehler als Differenz zwischen dem tatsächlichen Können und jenem, das anhand der gestellten Aufgaben gezeigt werden konnte.

Im Gegensatz zu den anderen genannten Punkten kann der Stichprobenfehler mathematisch relativ einfach modelliert werden. Ist eine Studentin zum Beispiel in der Lage, 80 % aller Aufgabenstellungen zu einem bestimmten Themengebiet korrekt zu lösen, erhält aber an der Prüfung überproportional solche Aufgaben gestellt, die sie nicht lösen kann, so könnte man etwa anhand der Prüfung fälschlicherweise zum Schluss kommen, sie könne nur 65 % der Aufgaben korrekt lösen. Sacher (2014, 106) vergleicht die Situation einer prüfenden Person deshalb mit einer Wahlforschungsgruppe, die Hochrechnungen zu einer laufenden Wahl vornimmt. So wie bei der Wahlprognose anhand bereits ausgezählter Bezirke auf das Gesamtergebnis zu schließen versucht wird, muss bei einer Prüfung anhand des gezeigten Prüfungs-

141 Ich verwende hier den Begriff „Messfehler“ im Einklang mit den meisten aktuellen Lehrbüchern in Statistik für die Sozialwissenschaften. Unter diesem Begriff sind allerdings auch Schwierigkeiten subsumiert, die eigentlich besser als Ungenauigkeiten denn als Fehler im engeren Sinn zu bezeichnen wären. So werden i. d. R. unter „Messfehlern“ systematische Verzerrungen verstanden, während mit „Messungenauigkeiten“ auch solche Fehler gemeint sind, die durch zufällige Abweichungen entstehen („Rauschen“ der Daten).

könnens das tatsächliche Können von Studierenden hochgerechnet werden. Dies mit dem Unterschied, dass die Hochrechnung im Gegensatz zu Wahlen nicht mit einem definitiv gültigen Ergebnis abgeglichen werden kann.

Der Schätzfehler von Hochrechnungen bei Wahlen wird in den Medien gelegentlich angegeben. Dann steht hinter dem Ergebnis für eine bestimmte Partei ein Schwankungsbereich, z. B. Beispielpartei: 17,4% ($\pm 1,8\%$). Dieser Schwankungsbereich wird Konfidenz- oder Vertrauensintervall genannt. Auch bei bepunkteten Prüfungen kann der Schätzfehler berechnet werden. Für statistisch Interessierte wird ein einfaches Vorgehen nachfolgend erläutert.¹⁴² Die sich daraus ergebenden Folgerungen sind für alle Lesenden relevant und werden in den letzten beiden Absätzen dieses Abschnitts erläutert.

Ab ca. 30 Studierenden ist es statistisch zulässig, die Standardabweichung zur Bildung des Konfidenzintervalls heranzuziehen (gemäß Schäfer 2016, 129; Pospeschill 2022, 107, schlägt allerdings $n \geq 60$ vor). Dazu ist zuerst die Varianz der Datenverteilung zu berechnen. Im Fall einer Punkteverteilung berechnet sich die Varianz s^2 folgendermaßen:

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

n = Anzahl geprüfter Studierender

$\bar{x}$ = arithmetisches Mittel aller erreichten Punktzahlen

x_i = einzelne Punktwerte

In Worten ausgedrückt: Die Varianz einer Datenverteilung entspricht dem Durchschnitt aller quadrierten Abweichungen der Datenpunkte von ihrem Mittelwert. Die Standardabweichung s berechnet sich als Quadratwurzel aus der Varianz der Datenverteilung (vgl. Schäfer 2016, 64).

Als weiterer Wert wird nun die Reliabilität der Prüfung benötigt. Zu deren Berechnung gibt es verschiedene statistische Verfahren. Den meisten Sozialwissenschaftlern dürfte Cronbachs α als Maß für die interne Konsistenz einer Messung bekannt sein. Zu seiner Berechnung werden die Werte der einzelnen Items, in unserem Kontext also der Aufgaben, herangezogen. Allerdings setzt dieses Maß die Erfüllung strenger Annahmen voraus, die bei Prüfungen häufig nicht erfüllt sind. Deshalb werden für die Bestimmung der Konsistenz von Prüfungen andere Maße verwendet (Pospeschill 2022, 23f.). Deren Erläuterung würde an dieser Stelle zu weit führen.

142 Geschildert wird ein statistisch möglichst einfach durchführbares Vorgehen. Allerdings unterschätzt dieses die Komplexität der Datenstruktur tendenziell und liefert deshalb eher zu optimistische Ergebnisse. Aus anspruchsvolleren Vorgehensweisen, wie sie Pospeschill (2022, 97–109), schildert, dürften eher noch größere Konfidenzintervalle resultieren.

Anhand der Standardabweichung und des Wertes für die Reliabilität kann nun der Standardmessfehler s_e der Messung berechnet werden. Er resultiert aus der Multiplikation der Standardabweichung s mit der Wurzel aus 1 minus dem Reliabilitätskoeffizienten r :

$$s_e = s \cdot \sqrt{1-r}$$

Aus dem Standardmessfehler kann nun das Konfidenzintervall gebildet werden. Zur Vereinfachung wählen wir hier ein Konfidenzintervall von 95 %. Das bedeutet, dass der Punktwert, der aus dem tatsächlichen Wissen und Können der Studierenden resultieren würde, mit einer Wahrscheinlichkeit von 95 % innerhalb der berechneten Bandbreite liegt. Dazu muss der Standardmessfehler mit 1,96 multipliziert werden.¹⁴³ Dieser Wert kann zu einer tatsächlich erreichten Punktzahl addiert und subtrahiert werden. Dadurch ergeben sich die Ober- und Untergrenze des Konfidenzintervalls. Dieses beträgt also $\pm 1,96 \cdot s_e$. Das tatsächliche Wissen und Können, das quasi hinter dem Punktwert x_i einer einzelnen Studentin liegt, befindet sich also mit einer Wahrscheinlichkeit von 95 % in einem Punkteband von $x_i \pm 1,96 \cdot s_e$.

Die Bedeutung dieser Berechnungen soll anhand eines einfachen Beispiels erläutert werden. Wir ziehen wieder die oben erwähnte Prüfung mit einem Maximum von 29 Punkten heran, für die in Abbildung 20 eine lineare und eine nicht lineare Punkteskala dargestellt wurde. Wir gehen davon aus, sie sei von 60 Studierenden bearbeitet worden, deren erfundene, aber realitätsnahe Punktwerte in der folgenden Fußnote aufgeführt sind.¹⁴⁴ Aus diesen 60 Punktwerten ergibt sich anhand obiger Formel die Stichprobenvarianz: $s^2 = 28,061$. Die Standardabweichung ist die Wurzel daraus: $s = 5,297$. Da das Beispiel fiktiv ist, müssten zur Berechnung eines Reliabilitätskoeffizienten auch noch Punkteverteilungen für die einzelnen Aufgaben erfunden werden; bei 10 Aufgaben wären dies 600 Werte. Darauf wird verzichtet und stattdessen ein fiktiver Wert von 0,85 gesetzt, der eine sehr hohe Reliabilität der Messung anzeigt. Auf dieser Basis kann nun der Standardmessfehler berechnet werden: $s_e = 5,297 \cdot \sqrt{(1-0,85)} = 5,297 \cdot 0,387 = 2,052$. Damit ergibt sich ein Konfidenzintervall von $x_i \pm 1,96 \cdot 2,052$, also $x_i \pm 4,021$.

Gehen wir nun von der oben als Beispiel vorgeschlagenen nicht linearen Punkteskala aus: 0–6 Punkte = Note 6, 7–11 Punkte = Note 5, 12–15 Punkte

143 Wenn ein Konfidenzintervall vom 95 % gewählt wird, müssen 95 % der Fläche einer Normalverteilungskurve symmetrisch abgedeckt sein. Die ungefähre beidseitige maximale Abweichung vom Mittelwert beträgt in diesem Fall je ca. 1,96 Standardabweichungen.

144 21, 23, 24, 18, 16, 9, 19, 20, 21, 15, 19, 27, 21, 23, 14, 7, 18, 22, 26, 20, 16, 27, 25, 23, 18, 20, 12, 10, 8, 5, 15, 19, 23, 11, 18, 21, 21, 13, 11, 17, 22, 18, 23, 25, 24, 9, 16, 19, 28, 21, 14, 17, 23, 12, 18, 22, 15, 21, 23, 15.

= Note 4, 16–19 Punkte = Note 3, 20–23 Punkte = Note 2, 24–29 Punkte = Note 1.

Anhand dieser Skala befindet sich das tatsächliche Wissen und Können, das hinter einer gemessenen Punktzahl von 21 liegt, mit einer Wahrscheinlichkeit von 95 % im Bereich von 17 bis 25 Punkten. Wir wissen somit nicht, ob wir der betroffenen Studentin anhand unserer Skala eine 3, eine 2 oder eine 1 geben müssten. Immerhin kann als hinreichend sicher gelten, dass diese Studentin ein genügendes Wissen und Können hat: Die gerade noch genügende Punktzahl liegt deutlich außerhalb des Konfidenzintervalls von 17–25 Punkten. Viel problematischer ist deshalb der Bereich um die Punktzahlen 11 (gerade nicht mehr genügend) und 12 (gerade noch genügend) herum: Die Punktzahl 8 könnte auch für eine 12 stehen und die Punktzahl 15 für eine 11 (immer mit einer Wahrscheinlichkeit von 95 %).

Mit anderen Worten: Für den Bereich von 8 bis 15 Punkten können wir mit weniger als 95 % Sicherheit sagen, ob das Resultat für ein genügendes oder ungenügendes Wissen und Können steht! Dieser Bereich umfasst mehr als einen Viertel der möglichen Punktwerte. Für alle Studierenden, die Werte in diesem Bereich erzielten, müsste deshalb durch einen zusätzlichen Leistungsnachweis geklärt werden, ob ihr Wissen und Können nun einer genügenden Lernzielerreichung entspricht oder nicht (vgl. Sacher 2014, 114).

Das Beispiel zeigt zudem auf, weshalb sogar Situationsbedingungen wie Wetterunterschiede für fast 20 % der Studierenden beeinflussen können, ob sie an eine Hochschule aufgenommen werden oder nicht (siehe Kap. 6.4.2; Redelmeier u. Baxter 2009). Dabei sind die obigen Zahlen eher optimistisch gewählt: Fallen die Varianz der Stichprobe größer, z. B. durch mehr Punktwerte unter 10, und die Reliabilität geringer aus, so wächst das Konfidenzintervall. Hinzu kommt, dass dieses lediglich den Unsicherheitsbereich bezeichnet, der durch den Stichprobenfehler entsteht. Alle anderen eingangs dieses Abschnitts erwähnten Messfehler sind dabei noch nicht berücksichtigt. Könnten sie ebenfalls in die Berechnung einbezogen werden, dürfte ein wesentlich größeres Konfidenzintervall entstehen. Die einfachste Maßnahme, um das Konfidenzintervall relativ zur Gesamtpunktzahl zu verringern, besteht im Gegenzug in einer Verlängerung der Prüfung um zusätzliche Aufgaben (siehe Kap. 9.2.1), was aber mit neuen Problemen für die Messgenauigkeit einhergehen kann, etwa mit Konzentrationsproblemen der Studierenden durch Ermüdung.

In der Realität wird somit der Punktebereich, in dem wir nicht sicher zwischen genügenden und ungenügenden Leistungen unterscheiden können, in den meisten Fällen nochmals deutlich größer ausfallen als im hier präsentierten Beispiel.

Mindestanforderungen für eine genügende Note festlegen

Die Mindestanforderungen festzulegen, anhand deren sich ein minimales Erreichen eines oder mehrerer Lernziele ablesen lässt, ist ein Ermessensentscheid. Die Anforderungen lassen sich nicht direkt aus fachlichen Zusammenhängen oder Aufgabenstellungen ablesen. Entscheidend für den Ermessensentscheid sind hingegen zwei Leitfragen:

1. Aufbau weiterer Lernfähigkeit. Diese Frage ist vor allem relevant, wenn im gleichen Studiengang weitere Lehrveranstaltungen zum gleichen Themengebiet folgen: Was müssen Studierende wissen und können, um in den folgenden Lehrveranstaltungen erfolgreich weiter zu lernen?
2. Erwartetes Wissen und Können. Diese Frage ist vor allem relevant, wenn das Thema im Studiengang bis zum Abschluss (Bachelor/Master) nicht weiter behandelt wird: Was wird vom Curriculum, von weiterführenden Studien- und anderen Bildungsgängen oder in der Berufswelt von Studienabgängerinnen in diesem Fachgebiet an Wissen und Können erwartet?

Die Mindestanforderungen sind dementsprechend so festzulegen, dass sie weiteres Lernen im betreffenden Gebiet ermöglichen und/oder festgelegten Anforderungen entsprechen. Interessant ist vor allem die Frage nach dem erfolgreichen weiteren Lernen – das Kriterium „erfolgreich“ ist näher zu definieren. Da Menschen grundsätzlich zu jedem Zeitpunkt weiterlernen können (siehe Kap. 4) lässt sich dieses Kriterium letztlich nur vor dem Hintergrund des summativen Beurteilungssystems beantworten: Erfolgreiches Weiterlernen bemisst sich danach, dass die Studierenden auch weiterhin mindestens genügende Leistungen erzielen können (Sacher 2014, 125). Damit wird die Bewertung aber tautologisch: Die aktuelle Bewertung bemisst sich an der Einschätzung dessen, welche Leistung welche künftige Bewertung erzielen wird. Ein solcher Ermessensentscheid kann vor allem dann einigermaßen zuverlässig gefällt werden, wenn man mit dem System, in dem er gefällt wird, gut vertraut ist. Mit Sacher (2014, 126) können Lehrende dazu folgende heuristische Vorgehensweisen heranziehen:

- Ab welchem durchschnittlichen Beherrschungsgrad würden Sie in einer Lehrveranstaltung mit neuem Stoff weitermachen und auf weitere Wiederholungen oder Vertiefungen verzichten?
- Welche Schlüsse lassen sich aus den Curricula bezüglich der Mindestexpectationen an Wissen und Können für die fragliche Lehrveranstaltung und/oder das weitere Studium ziehen?
- Wie schätzen Ihre lehrenden Kolleginnen und Kollegen die Mindestanforderungen in diesem Fachgebiet ein? Hierzu können informelle Gespräche oder auch offizielle Besprechungen erfolgen.

- Haben Sie Erfahrungswerte, welcher Beherrschungsgrad es einigermaßen wahrscheinlich macht, dass Studierende Zwischen- oder Abschlussprüfungen in diesem Fachgebiet bestehen werden?

Selbstverständlich sind dies alles nur Notbehelfe, die keinen höheren Ansprüchen an Genauigkeit standhalten. Andererseits ist es mit Sicherheit kein brauchbares Kriterium, als Lernzielerreichung nur jene Leistungen einzustufen, in denen sämtliche Aufgaben (ggf. auch noch in einem vorgegebenen Zeitrahmen) vollständig korrekt gelöst wurden. Diese Anforderung ist auch deshalb unrealistisch, weil nicht jeder Fehler eine Folge von mangelndem Wissen oder Können ist, sondern viele Fehler auch aus Konzentrationschwierigkeiten oder Missverständnissen hinsichtlich der Aufgabenstellung resultieren. Zudem ist rein empirisch festzustellen, dass die meisten Studierenden, „die in einem bestimmten Stadium eines schulischen Lernprozesses keine fehlerfreien Leistungen zu Wege bringen, durchaus erfolgreich weiterlernen können. Ganz offensichtlich muss man nicht in allen Gebieten denselben Grad an Perfektion erlangen, um künftigen Lernerfolg zu haben“ (Sacher 2014, 127).

Sacher (ebd.) schlägt deshalb für klassische Prüfungsformen, bei denen viele kleinteilige Aufgaben zu lösen sind, eine Orientierung an Prozentzahlen vor. Auch hier handelt es sich natürlich nur um grobe Schätzwerte, da ihre Aussagekraft von der Aufgabenschwierigkeit und der bereits mehrfach besprochenen, problematischen Prämisse abhängt, alle erteilten Punkte seien gleichwertig (siehe Kap. 9.2.1 und 9.4.2). Dabei können für zentrale Lernziele höhere Prozentwerte als Mindestanforderung vorausgesetzt werden als für weniger zentrale Lernziele (Sacher 2014, 127). So könnte beispielsweise eine Abstufung für eine klassische schriftliche Klausur mit einer Vielzahl an Aufgaben wie folgt aussehen:

Zentrale Lernziele: 60 % richtige Lösungen

Lernziele mittlerer Wichtigkeit: 50 % richtige Lösungen

Weniger zentrale Lernziele: 40 % richtige Lösungen

Diese Abstufung ist quasi die quantifizierende Alternative zum Vorschlag von Biggs u. Tang (2007, 83), dass für wichtige Themen Lernziele auf einer höheren Taxonomiestufe formuliert werden sollen als für weniger bedeutsame Wissens- und Könnensinhalte (siehe Kap. 5.4.2).

9.5 Weniger ist mehr: Basisbewertungen und Prädikate

Notenunterteilungen sind in der Regel so kleinteilig, dass zwei oder gar drei nebeneinander liegende Notenstufen oft nicht mit ausreichender Sicherheit voneinander unterschieden werden können. Wie im Kapitel 9.4.6 erläutert, lässt sich das auch mit statistischen Verfahren zeigen. Im Gegensatz dazu sind einfachere Unterteilungen leichter vorzunehmen. Darüber, ob eine Leistung „gut, schlecht oder mittelmäßig ist, können Kollegen sich gewöhnlich leicht verständigen. Die Schwierigkeiten beginnen dann, wenn auch noch entschieden werden soll, ob die mittelmäßige Leistung mit der Note 3 oder 4, die gute mit 1 oder 2 und die schlechte mit 5 oder 6 zu bewerten ist“ (Sacher 2014, 112). Dieses Problem ist mit kleinteiligeren Notenstufen nicht zu lösen, wie eine Analogie aufzeigt: Wenn eine Personenwaage ungenau misst, nützt es nichts, wenn sie auch noch Milligramm anzeigen kann (Sacher 2014, 112f.). Hinzu kommt, dass in der Regel für Leistungsbewertungen an Hochschulen die Frage entscheidend ist, welche Leistungen noch als genügend bzw. als Beleg für das Erreichen der Lernziele einzustufen sind. Problematisch sind somit in erster Linie Leistungen, bei der diese Frage nicht klar zu beantworten ist, wie anhand des Konfidenzintervalls in Kapitel 9.4.6 veranschaulicht wurde.

Insofern würde es in den meisten Fällen ausreichen und die Ressourcen von Lehrenden wie Lernenden auf die wesentlichen Bereiche und Situationen zu fokussieren helfen, nicht von 6 oder gar 11 (bei halben Noten wie in der Schweiz üblich) Stufen auszugehen, sondern nur von 2 wie bei den Prädikaten „bestanden/nicht bestanden“ oder allenfalls 3 wie in der Abstufung Übertreffen – Erfüllen – Verfehlen des Lernziels. Zumindest im Rahmen von Leistungsnachweisen, die für die Fortsetzung oder den Abschluss des Studiums von großer Bedeutung sind, scheint es zudem sinnvoll, Fälle zu identifizieren, in denen anhand der gezeigten Leistung nicht mit befriedigender Sicherheit entschieden werden kann, ob das dahinter liegende Wissen und Können in genügendem Ausmaß den zu überprüfenden Lernzielen entspricht. Für diese Fälle wäre eine geeignete Form eines zusätzlichen Leistungsnachweises zu bestimmen, die ein besser gesichertes summatives Urteil über den Grad der Lernzielerreichung erlaubt. Für diese zusätzliche Leistungsbeurteilung sollten andere Lehrende als Beurteilende hinzugezogen und/oder andere Aufgabenarten und Medialitäten gewählt werden. Dadurch können kognitive Verzerrungen bei der Beurteilung ebenso reduziert werden wie Stichprobenfehler.

Spielräume für solche Vorgehensweisen sollten dort genutzt werden, wo sie bereits vorhanden sind, und dort geschaffen werden, wo sie es noch nicht sind. Dies kann von der Anpassung von Studien- und Prüfungsordnungen

einzelner Studiengänge bis zur Anpassung gesetzlicher Rahmenbedingungen reichen.

9.6 Grundlegendere Alternativen zur klassischen summativen Beurteilung

Aufgrund der in diesem Kapitel erläuterten diversen Schwächen der herkömmlichen summativen Beurteilung von Lernleistungen, bei der Noten durch Lehrende gesetzt werden, existiert eine Vielzahl alternativer Vorgehensvorschläge. Zwei derzeit vielversprechende Optionen werden nachfolgend vorgestellt: Der Notenkontrakt (9.6.1) und der Ansatz, Studierende in ihre eigene summative Beurteilung einzubeziehen (9.6.2).

9.6.1 *Notenkontrakt*

Im bisherigen Verlauf des Buches wurde mehrfach erläutert, weshalb das summative Beurteilen von Lernleistungen sich negativ auf die Lernmotivation und folglich auch auf die Lernergebnisse auswirkt. Auf einen Satz verkürzt: Eine Bewertung durch andere Menschen erschwert es der bewerteten Person deutlich, ein Gefühl der Selbstbestimmung aufzubauen, welches für die intrinsische Lernmotivation entscheidend ist (siehe Kap. 9.1 und zum lernpsychologischen Hintergrund Kap. 4.5.2).

Das Konzept des Notenkontrakts setzt an diesem Punkt an und gibt den Studierenden mehr Kontrolle über ihre Noten und den Weg zu ihrem Erreichen: Lehrende handeln mit Studierenden individuelle Kontrakte aus. In diesen werden die Lernziele festgehalten, auf deren Erreichen die Studierenden sich verpflichten, und der Umfang der Arbeit, die zu erledigen ist, um im vorliegenden Kontext die vereinbarte Note zu erhalten. Indem die Studierenden mitentscheiden können, wie sehr sie sich für welche Lernziele anstrengen und welche Note sie erreichen wollen, werden die drei psychologischen Grundbedürfnisse der Selbstbestimmung nach Deci u. Ryan adressiert: Wirksamkeit, Eigenständigkeit und soziale Einbindung (vgl. Ryan u. Deci 2017, 94–98). Die Arbeit mit Notenkontrakten steigert deshalb die intrinsische Motivation der Studierenden.

Lehrende, die mit Notenkontrakten arbeiten, berichten unter anderem, dass die Lehr-Lern-Aktivitäten, die Lernziele und fachlichen Herausforderungen wie beabsichtigt stärker im Zentrum stehen als zuvor. Zudem berichten die meisten Fallstudien, dass mit den Studierenden weniger oder keine Diskussionen um die Fairness erhaltener Noten geführt werden mussten. Als Schwierigkeit nennen die Lehrenden Entscheidungen im Umgang mit nicht

erfüllten Anforderungen. Soll man z. B. einen Unterschied machen zwischen der Bewertung der Aktivitäten von Studierenden, die die Anforderungen trotz sichtbarer Bemühungen auch nach Überarbeitung nicht erfüllen, und der Bewertung ungenügender Aktivitäten, die minimalistisch bearbeitet wurden (Parks 2023, 3)?

Unbedingt vorher abzuklären ist, ob und unter welchen Bedingungen Studienordnung und Prüfungsrecht den Einsatz von Notenkontrakten zulassen.¹⁴⁵ Sogar in den USA, wo der Ansatz verbreiteter ist, gibt es Hochschulen, die ihn nicht oder nur restriktiv zulassen. Auch deshalb gibt es nicht einen richtigen Weg, um Notenverträge einzuführen, sondern es muss neben fachlichen und curricularen Aspekten auch der regulatorische Rahmen berücksichtigt werden (Katopodis u. Davidson 2020, 106f.).

Grundprinzip und Varianten

Das Konzept wurde unter dem Begriff „Contract Grading“ durch den Betriebswirtschaftsprofessor Elmer G. Dickson (1974) entwickelt und gewann insbesondere seit der Jahrtausendwende in verschiedenen Fachgebieten an Verbreitung. Der Notenkontrakt ist ein flexibles Konzept, das von den meisten Lehrenden etwas unterschiedlich eingesetzt wird. Wesentliche Faktoren sind neben fachdidaktischen Anforderungen auch die Anzahl Studierender und die jeweiligen Anforderungen, die Studienordnungen und weitere Regelungen an die Bewertung stellen. Je nachdem können die Kontrakte mehr oder weniger Freiheitsgrade zulassen; so ist der ursprüngliche Ansatz von Dickson (1974) ziemlich offen, während z. B. Davidson mit einem relativ engen Katalog an möglichen Lernaktivitäten und -resultaten arbeitet (Katopodis u. Davidson 2020, 110–115).

Das Grundprinzip des Notenkontrakts besteht darin, dass Studierende sich anspruchsvollere oder weniger anspruchsvolle Lernziele setzen und sich damit verbunden zu mehr bzw. aufwändigeren oder weniger bzw. weniger aufwändigen Lernaktivitäten verpflichten. Dabei kann bereits genau vorgegeben sein, welche Kombination von Aktivitäten – sofern erledigt – zu welcher Note führt. In der Regel ergibt bei dieser enger geführten Variante das Erledigen aller Lernaktivitäten die Maximalnote, während das Weglassen von bestimmten Aktivitäten durch die Reduktion der Note „erkauft“ werden kann

145 Auch die gelegentlich benutzten Begriffe „Notenvertrag“ oder „Lernvertrag“ sind problematisch, da es sich nicht um Verträge in einem juristischen Sinne handelt. Deshalb wird hier von Kontrakten gesprochen. Mit Bezug auf das Schulrecht in Deutschland (das sich vom Hochschulrecht unterscheidet) erläutert Richter (2007) rechtliche Fragen rund um einen rechtlich korrekten Einsatz von Notenkontrakten. Teilweise wird auch der Kontraktbegriff vermieden (vgl. Jötten 2007, 274).

(so z. B. Katopodis u. Davidson 2020, 110–15).¹⁴⁶ Demgegenüber schlägt Dickson (1974, 23f.) eine große Palette an möglichen Aktivitäten vor, die mit unterschiedlichen Punktzahlen versehen werden. Für bestimmte Gesamtpunktzahlen werden dann verschiedene Notenniveaus festgesetzt, z. B. wer eine 1 möchte, muss mindestens 60 Punkte erarbeiten, für 50 gibt es eine 2, für 42 eine drei und für 35 eine 4. Die beiden Ansätze werden unten in verkürzter Form erläutert, um die Arbeit mit Notenkontrakten zu veranschaulichen.

Als weiterführende Lektüre seien hier exemplarisch einige Publikationen genannt, die über den Einsatz von Notenkontrakten und deren Gestaltung in verschiedenen Fächern berichten: Betriebswirtschaftslehre (Hiller u. Hieta-pelto 2001), Informatik (Marriott u. Abraham 2023), Literaturwissenschaft und Pädagogik (Katopodis u. Davidson 2020), Pädagogik (Parks 2023) und Schreibdidaktik (Danielewicz u. Elbow 2009).¹⁴⁷

Einführung des Konzepts in der Lehrveranstaltung

Wichtig ist es, das Konzept des Notenkontrakts und die Überlegungen dahinter gleich zu Semesterbeginn einzuführen (vgl. zu den folgenden Erläuterungen Parks 2023, 2). Idealerweise wird bereits in der Modulbeschreibung darauf hingewiesen, dass mit einem Notenkontrakt gearbeitet wird.

Bei der Einführung sollten die Unterschiede zu institutionell übliche(re)n Bewertungsformen herausgestellt und die lernpsychologischen Gründe für den Einsatz des Ansatzes dargelegt werden. Anschließend können der Kontrakt und die vorgesehenen Optionen besprochen werden, wobei sich ein ähnlich interaktives Vorgehen eignet wie in Kapitel 9.3.4 für die Einführung von Kriterienrastern vorgeschlagen. Die Studierenden können dann innerhalb eines bestimmten Zeitraums, meist bis zur folgenden Woche, für sich überlegen, welchen Lernaufwand sie betreiben wollen bzw. welche Note sie anstreben möchten. Auf dieser Basis wählen sie dann die für sie passende Option aus einer beschränkten Auswahl oder erarbeiten bei Varianten mit größeren Freiheitsgraden einen individuellen Kontraktvorschlag (siehe die Erläuterungen der Varianten unten). Die Kontraktierungsphase endet damit, dass jede Studentin und jeder Student den eigenen Kontrakt unterzeichnet und den Lehrenden abgibt. Dabei behalten sie entweder ein physisches Doppel oder laden den unterzeichneten Kontrakt als Scan auf eine Lernplattform hoch.

146 Aus diesem Grund wird der Ansatz im angelsächsischen Raum auch „labor-based grading“ genannt, was man mit Benotung nach Aufwand übersetzen könnte.

147 Leider scheint das Konzept auf Hochschulstufe bisher v. a. im angelsächsischen Raum und besonders den USA verbreitet, aus denen sämtliche aufgezählten Publikationen stammen.

Variante mit wenigen Freiheitsgraden

Eine überschaubare Anzahl an Optionen der Kontraktgestaltung eignet sich, wenn zu viele Wahlmöglichkeiten für die Studierenden überfordernd sein könnten, etwa in Veranstaltungen zu Studienbeginn. Eine hohe Studierendenzahl kann ein weiterer Grund für das Arbeiten mit klar definierten Wahlmöglichkeiten sein, da es für die Lehrenden sonst zu viel Aufwand bedeuten könnte, mit jeder Studentin einen individualisierten Kontrakt abzuschließen.

Das nachfolgende Beispiel ist eine Zusammenfassung der Notenkontrakt-Optionen von Katopodis u. Davidson (2020, 110–15) zu einer erziehungswissenschaftlichen Veranstaltung, die mit einem Seminar vergleichbar ist. Die Studierenden haben folgende Optionen:

1. Teilnahme an den Präsenzveranstaltungen: zulässig sind zwei unentschuldigte Absenzen
2. wöchentliches Verfassen eines Blogbeitrags (400–500 Wörter), wodurch ein Work-in-Progress-Forschungspaper entsteht. Jeder nicht fristgerecht hochgeladene Blogbeitrag führt zur Verringerung der Abschlussnote um 0,5
3. interaktive Gestaltung einer Veranstaltungsstunde zu einem aus einer Liste ausgewählten Thema (in Gruppen von 2–3 Studierenden)
4. zwei substanzielle Beiträge zu einer öffentlichen Online-Publikation (etwa einem Artikel in Wikipedia), eine ausführliche Buchrezension auf Amazon oder ein ausführlicher Kommentar zu einem Artikel in einer seriösen Zeitung oder Zeitschrift
5. Semesterhälfte: Mitarbeit an einem ausführlichen Blogpost, der die Erkenntnisse aus der ersten Semesterhälfte zusammenfasst und auf einer Seite wie hastac.org oder medium.com veröffentlicht wird¹⁴⁸
6. Abschluss: Kollaboratives Erstellen eines dreiminütigen Videos zum Thema aus 3), das in der letzten Semesterwoche besprochen, ggf. angepasst und dann auf einer geeigneten Plattform öffentlich geschaltet wird (z. B. Youtube-Kanal der Hochschule)

148 Die Studierenden schlugen der Dozentin allerdings aus Eigeninitiative vor, statt der Aktivitäten 5) und 6) kollaborativ ein Buch zum Seminarthema zu verfassen. Deshalb wurde der Kontrakt entsprechend angepasst, und alle außer einem Studenten verpflichteten sich zur Maximalnote mit der Mitarbeit am Buch als Hauptaktivität (Katopodis u. Davidson 2020, 115f.). Das entstandene Buch ist im Literaturverzeichnis unter der Autorschaft 21st Century Collective (2013) aufgeführt.

Anhand dieser Optionen lassen sich folgende drei Notenkontrakte abschließen:¹⁴⁹

- Formulierung für Note 1: „Mit der Unterzeichnung dieses Kontrakts für eine Eins in dieser Veranstaltung erkläre ich mich mit allen oben genannten Bedingungen einverstanden.“
- Formulierung für Note 2: „Ich möchte in dieser Veranstaltung die Note 2 erreichen. Um meinen Kontrakt für eine 2 zu erfüllen, werde ich die oben genannten Aktivitäten 1, 2, 3, 4 und 6 zufriedenstellend erfüllen. Ich werde in der Präsenzveranstaltung höchstens dreimal unentschuldigt fehlen. Mir ist klar, dass sich bei weiteren unentschuldigten Abwesenheiten meine Note für die Veranstaltung für jede weitere Abwesenheit um 0,5 verringert.“
- Formulierung für Note 3: „Ich möchte in dieser Veranstaltung die Note 3 erreichen. Um meinen Kontrakt für eine 3 zu erfüllen, werde ich die oben genannten Aktivitäten 1, 2, 3 und 6 zufriedenstellend erfüllen. Ich werde in der Präsenzveranstaltung höchstens viermal unentschuldigt fehlen. Mir ist klar, dass sich bei weiteren unentschuldigten Abwesenheiten meine Note für die Veranstaltung für jede weitere Abwesenheit um 0,5 verringert.“

Die Kontraktvorlage enthält zudem den folgenden Hinweis zu den Noten 4 und tiefer: „Die Dozentin behält sich das Recht vor, die Note 4 oder ungenügende Noten an diejenigen zu vergeben, die einer vertraglichen Verpflichtung nicht systematisch nachkommen. Eine 4 bedeutet eine minimale Erfüllung des Vertrags. Eine 5 oder tiefer bedeutet, dass die vertraglich vereinbarte Leistung nicht ausreicht, um das Bestehen des Kurses zu rechtfertigen. Sowohl die Note 4 als auch ungenügende Noten bedeuten einen Bruch des Kontrakts, der durch die Unterzeichnung einer der oben genannten Optionen zustande gekommen ist“ (Katopodis u. Davidson 2020, 115, sinngem. Übers. tz).

Offenere Variante

Eine größere Anzahl an Optionen der Kontraktgestaltung eignet sich, wenn den Studierenden relativ selbständiges Arbeiten zugetraut werden kann, etwa in fortgeschrittenen Semestern. Es ist in diesem Fall wichtig, die Kontraktentwürfe der Studierenden gut durchzuschauen und ein Feedback dazu zu geben, bevor die Studierenden den finalen Kontrakt erstellen. Auch ein Peer-Feedback kann sinnvoll sein.

149 Dabei wurde das US-amerikanische Notensystem wie folgt ins deutsche Notensystem überführt: A = 1, B = 2, C = 3, D = 4, F = ungenügende Noten (das US-System kennt nur das „F“ für ungenügende Leistungen).

Tab. 17: Liste von Lernaktivitäten für einen Lernkontrakt in einer Veranstaltung in Finanzwirtschaft (aus Dickson 1974, 24, sinngem. Übers. tz)

Liste möglicher Aktivitäten	Punkte
Erstes Gespräch mit dem Dozenten	5
Vertrag in zweifacher Ausfertigung ausfüllen*	5
Lesen der zusammenfassenden Kapitel*	25
Paper über Mittelherkunft und -verwendung*	20
Paper über die Finanzanalyse*	20
Buchbesprechungen von finanzorientierten Büchern	15
Kurzvorträge über Finanzmanagement	5
Plenumspräsentation zu relevanten Finanzmanagement-Themen	10
Diskussionsleitung zu einem Thema des Finanzmanagements	5
Umfassendes Paper zu Themen des Finanzmanagements	30
Lösen von Finanzmanagementaufgaben am Time-Share-Terminal (fünf Aufgaben müssen gelöst werden)	5
Programmierung von Finanzmanagement-Problemen auf dem Time-Share-Terminal oder dem Computer	20
Teilnahme an FINANSIM auf dem Time-Share-Terminal	20
Umfassende Fallanalyse	30
Klausuren: ein Punkt für je zehn Punkte bei einer Prüfung mit 100 Punkten	
Weitere selbst initiierte Aktivitäten (Punktzahl zu verhandeln) *Pflicht für alle	

Der hier geschilderte Notenkontrakt von Dickson (1974, 23f.) ist ein Beispiel für eine relativ offene Realisierung des Konzepts. Er umfasst zwar auch einige obligatorische Elemente, aber darüber hinaus eine lange Liste mit sehr verschiedenen und unterschiedlich aufwändigen Lernaktivitäten (siehe Tab. 17). Die Punkte sind einerseits ein Hinweis auf den durch den Dozenten erwarteten Zeitaufwand, dienen aber den Studierenden andererseits ebenfalls zur Abschätzung, ob die Aktivität für sie in einem günstigen Aufwands-Ertragsverhältnis stehen. Die vorgeschlagenen Aktivitäten könnten heutzutage z. B. um Blogbeiträge, Poster oder multimediale Arbeiten ergänzt werden. Wichtig ist zudem, dass der Kontrakt offen ist für durch die Studierenden vorgeschlagene weitere Aktivitäten, wobei der Punktebetrag mit dem Dozenten auszuhandeln ist.

Der eigentliche Notenkontrakt besteht in diesem Fall aus einer offenen Liste, in der die Studierenden die gewählten Aktivitäten eintragen können.

Tab. 18: Notenkontrakt von Dickson (1974, 23, sinngem. Übers. tz)

Notenkontrakt zur Veranstaltung XYZ		
Ich habe die beigefügte Beschreibung der Anforderungen für die Noten 1, 2 und 3 in der Veranstaltung XYZ gelesen. Auf der Grundlage dieser Anforderungen möchte ich, _____, einen Kontrakt für die Note ____ abschließen. Mir ist bekannt, dass ich diesen Kontrakt nur in Richtung einer höheren Note neu aushandeln kann. Mir ist auch bekannt, dass jede Abweichung vom unten vereinbarten Studienplan mit dem Dozenten abgesprochen und auf diesem Studienplan vermerkt werden muss.		
Lehr-Lern-Aktivitäten	Punkte	Abgabedatum
1. _____	_____	_____
2. _____	_____	_____
3. _____	_____	_____
4. _____	_____	_____
etc. _____	_____	_____
Ich erkläre mich mit dem obigen Studienplan einverstanden und übernehme die Verantwortung für die Erfüllung aller Anforderungen, die für seinen Abschluss erforderlich sind. Mir ist bekannt, dass ich das Prädikat „unvollständig“ oder die Note 4 erhalte, wenn der Kontrakt nicht zum vereinbarten Zeitpunkt erfüllt wird. Wenn diese Note nicht durch die Erfüllung der Kontraktbedingungen bis zum Ende der zweiten Woche des folgenden Semesters aufgehoben wird, wird die Note in eine 5 umgewandelt. Ich bestätige, dass ich die Standards für die Benotung in diesem Kurs verstanden habe.		
Datum _____ Neuverhandlung _____	Unterschrift Student:in _____ Unterschrift Dozent:in _____	

Diese Liste ist gerahmt von den Kontraktformulierungen, siehe Tabelle 18. Zusätzlich legt der Dozent fest, für welche Noten welche Punktzahlen erreicht werden müssen. Dickson (1974, 21) erwähnt für das hier gezeigte Beispiel einen Vorschlag von 125 Punkten = 1, 100 Punkten = 2 und 75 Punkten = 3 (tiefere Noten sieht seine Kontraktierung nicht vor, obschon das US-amerikanische D noch der Note 4 = ausreichend entsprechen würde).

Ebenfalls denkbar ist ein Vorgehen, bei dem die Studierenden in die Erstellung des Kontrakts und der möglichen Lehr-Lern-Aktivitäten einbezogen werden (die Vorschläge in Kap. 9.3.4 zum kollaborativen Erarbeiten von Beurteilungsrastern können Anregungen dazu geben).

9.6.2 Selbstbewertung

Das Einbeziehen von Studierenden in ihre eigene summativ Beurteilung erfolgt unter den gleichen lernpsychologischen Annahmen wie die Arbeit mit Notenkontrakten: Das Grundbedürfnis der Selbstbestimmung sowie die damit verbundene intrinsische Lernmotivation sollen gestärkt werden und dadurch eine stärkere Konzentration auf die Lehr-Lern-Aktivitäten statt auf die Bewertung ermöglichen. Auch bei Selbstbewertungen ist vorher abzuklären, ob und unter welchen Bedingungen Studienordnung und Prüfungsrecht ihren Einsatz zulassen.

Die Befundlage, ob summativ Selbstbeurteilungen wirklich die erwünschte motivationale Wirkung haben, ist weniger klar als für Notenkontrakte (vgl. Andrade 2019, 8). Deshalb wird hier zur Anregung lediglich ein vielversprechender Ansatz vorgestellt, den die Chemie-Dozentin Clarissa Sorensen-Unruh (2020) entwickelt hat. Er besteht neben anderen Komponenten daraus, dass die Studierenden sich selbst Noten für ihre Klausurleistungen geben (die weiteren Komponenten sind in Tabelle 4, Kap. 7.3.2 näher dargestellt). Interessant scheint er mir deshalb, weil er einen Mechanismus beinhaltet, der realistische Selbstbewertungen begünstigt. Das System ist somit abgesichert gegen einen Missbrauch und ermöglicht den Studierenden trotzdem mehr Autonomie und einen stärkeren Einbezug in die Bewertung ihrer Leistungen als eine reine Fremdbewertung. Die Vorgehensweise besteht aus folgenden Schritten (vgl. Sorensen-Unruh 2020, 143f.; 152f.):

1. Dreimal im Semester findet eine herkömmliche Klausur statt.
2. Die Dozentin bewertet die Klausuren anschließend, notiert die Punktzahlen und Noten aber nur in eine elektronische Tabelle, während sie in die Prüfungen lediglich formative Rückmeldungen notiert. Diese enthalten also weder die von der Dozentin vergebenen Punktzahlen noch eine Note.
3. Nach Rückgabe ihrer kommentierten Prüfungen erhalten die Studierenden Gelegenheit, anhand ihrer Aufgabenlösungen, der Rückmeldungen der Dozentin und der bereits in jeder Prüfungsaufgabe mitgeteilten Maximalpunktzahl für jede Aufgabe zu bestimmen, welche Punktzahl sie erreicht haben. Diese reichen sie der Dozentin ein.
4. Die Dozentin trägt die Punktzahlen der Studierenden ebenfalls in ihre Tabelle ein und gleicht sie mit den eigenen Punktzahlen ab. Wenn für eine Aufgabe die Punktzahl der Dozentin höher liegt als jene der Studentin, wird die Zahl der Dozentin gezählt. Andernfalls wird der Durchschnitt aus den beiden Punktzahlen gezählt.
5. Anschließend gleicht die Dozentin die Gesamtpunktzahlen der Selbstbewertung pro Studentin mit ihrer eigenen Bewertung ab. Dabei kommt

folgender Mechanismus zum Tragen: Um zu verhindern, dass Studierende ihre Note ungerechtfertigt aufbessern, erhalten sie zusätzlich zur gemäß 4) berechneten Punktzahl 5 Bonuspunkte, wenn ihre Gesamtpunktzahl innerhalb einer Standardabweichung (berechnet anhand der Bepunktung der Dozentin)¹⁵⁰ liegt. Liegt die Gesamtpunktzahl der Selbstbewertung mehr als 3 Standardabweichungen von jener der Dozentin entfernt, kann die Dozentin nach eigenem Ermessen 5 bis 10 Punkte von der gemäß 4) errechneten Gesamtpunktzahl abziehen.

- Hinweis: Sorensen-Unruh (2020, 144) gibt an, dass die Standardabweichung in der Regel 8–15 Punkte betrage. Die maximale Gesamtpunktzahl gibt sie nicht an; oft werden Klausuren in den USA so konstruiert, dass maximal 100 Punkte erreichbar sind.
6. Bei der finalen Rückgabe der Prüfungen notiert die Dozentin vier Gesamtpunktzahlen und die zugehörige Note:
- die Selbstbewertung der Studierenden
 - die Bewertung der Dozentin
 - die Durchschnittsnote (berechnet nach dem obigen Schritt 4)
 - die definitive Note (unter Berücksichtigung des Mechanismus' in Schritt 5)

Indem dieser Prozess dreimal im Semester erfolgt, wird der Dialog zwischen Dozentin und Studierenden über die Bewertung der Lernleistung etabliert und geübt. Zudem stellen die Schritte 2 und 3 wirksam sicher, dass die Studierenden die erhaltenen formativen Leistungsrückmeldungen tatsächlich umsetzen – ein Schritt, der ohne klare Anreize oder Verbindlichkeiten oft nicht erfolgt (siehe Kap. 8.6.1). Selbstverständlich bringt auch das geschilderte Vorgehen Herausforderungen mit sich – wie mit diesen umgegangen werden kann, beschreibt Sorensen-Unruh (2020) anschaulich.

150 Die Berechnung der Standardabweichung wird in Kapitel 9.4.6 (Abschnitt „Messfehler und Konfidenzintervall“) erläutert. Das marktübliche Tabellenkalkulationsprogramm MS Excel bietet dazu eine Funktion an, so dass die Berechnung auch ohne komplexere Statistiksoftware möglich ist.

Teil III: Spotlights

10 Leistungsnachweise im Curriculum

Die Entwicklung und Gestaltung von Curricula steht in einem engen Zusammenhang zur Frage, wie das Erreichen von Lernzielen überprüft, beurteilt und bewertet wird. Leistungsnachweise und Leistungsbeurteilung sind letztlich Instrumente und Prozesse, um sicherzustellen, dass die Studierenden das lernen, was das Curriculum vorgibt. Wie in den Kapiteln 2 und 3.1 erläutert, hat insbesondere die summative Leistungsbeurteilung, also die abschließende und wertende Beurteilung des Lernerfolgs, auch erhebliche gesellschaftliche Aspekte. Sie dient aus Sicht der Hochschulen dazu, die Qualität ihrer Studienabschlüsse zu gewährleisten. In diesem Kapitel werden einige Aspekte angesprochen, die zu beachten sind, wenn man Leistungsnachweise und Leistungsbeurteilung durch eine gute curriculare Abstimmung möglichst lernförderlich gestalten möchte. Dazu gehört die Problematik, dass summative Beurteilungen als Bewertungen auch Wert zuweisen (Kap. 10.1). Aber auch Fragen der Integration von verschiedenen Lernzielen sind zu diskutieren. In diesem Zusammenhang ist es wichtig, von Beginn eines Studiums an deklaratives und prozedurales Wissen bzw. Wissen und Können zu integrieren (Kap. 10.2). Zudem wird die Unterscheidung zwischen horizontaler und vertikaler Integration eingeführt, bei der es darum geht, verschiedene Lernziele fachlich und zeitlich miteinander zu verknüpfen (Kap. 10.3).

Für weiterführende Hinweise zum Thema Curriculumentwicklung sei verwiesen auf das Buch von Wiggins u. McTighe (2005), das schon fast als Standardwerk zur Curriculumentwicklung bezeichnet werden kann. Spezifischer auf den deutschsprachigen Raum bezogen ist die Einführung in die Studiengangentwicklung von Brinker u. Tresp (2012); ein integratives Modell der Curriculumentwicklung an Hochschulen stellen Jenert et al. (2018) vor. Hilfreiche Überlegungen für professionsorientierte Studiengänge, wie sie an vielen Fachhochschulen üblich sind, finden sich in Bach u. Zellweger (2022).

10.1 Bewertungen weisen Wert zu

Dass summative Leistungsbeurteilungen entscheidend bestimmen, was und wie Studierende lernen, wurde in Kapitel 5.1.2 erläutert. Auf dieser Erkenntnis basieren wesentliche Prinzipien wie das Constructive Alignment und das Backward Design, die grundlegende Modellannahmen für die in den Kapiteln 6 bis 9 geschilderten Prinzipien und Vorgehensweisen darstellen. Das Prin-

zip, das Studierende dazu bewegt, summativen Leistungsbeurteilungen so viel Gewicht zu geben, lässt sich pointiert wie folgt beschreiben: Nur was bewertet wird, hat einen Wert. Dass Leistungsnachweisformate und Beurteilungsansätze deshalb die Lernziele unterlaufen können, wenn sie nicht mit diesen kongruent („aligned“) sind, ist dementsprechend ein wiederkehrendes Thema in diesem Buch. Dass nur das Bewertete einen Wert hat, entfaltet jedoch auch über einzelne Veranstaltungen, Leistungsnachweise und Teilgebiete eines Curriculums hinaus Wirkung. Denn wenn ein Teilgebiet eines Curriculums oder einzelne Veranstaltungen und Leistungsnachweise intensiver oder strenger summativ beurteilt werden, beeinflusst dies auch andere Teilgebiete, Veranstaltungen und Leistungsnachweise. Das gilt ganz besonders für Elemente, die mehr oder weniger zeitgleich stattfinden, etwa im gleichen Semester oder Studienjahr. Studierende werden mehr Lern- und Arbeitsaufwand investieren in Leistungsnachweise, die als besonders anspruchsvoll gelten, und dementsprechend weniger Energie auf andere zur gleichen Zeit zu bearbeitende Leistungsnachweise oder Veranstaltungen aufwenden. Dies behindert das Lernen insbesondere dann, wenn Studierende wertvolle Gelegenheiten für formatives Feedback verpassen, weil sie Prüfungsvorbereitungen oder dem Schreiben von wissenschaftlichen Arbeiten eine höhere Priorität zumessen. Die Wahrscheinlichkeit ist groß, dass solche Effekte mittel- und langfristig, also bis zum Ende des Studiums und darüber hinaus, tiefenorientiertes Lernen hemmen und oberflächenorientiertes Lernen fördern.

An dieser Stelle möchte ich deshalb nochmals betonen, dass im Curriculum ein Ausgleich zwischen formativen und summativen Beurteilungen anzustreben ist. Beide spielen eine wichtige Rolle: Formative Leistungsrückmeldungen begünstigen den Lernfortschritt und die intrinsische Motivation der Studierenden, während summative Beurteilungen überprüfen, ob sie die Lernziele in ausreichendem Ausmaß erreicht haben. Rein quantitativ sollten formative Beurteilungen die summativen deutlich überwiegen. Als Maxime für die Curriculumentwicklung könnte gelten: So viele formative Beurteilungen wie möglich, so viele summative Beurteilungen wie nötig.

Insbesondere größere Klausuren oder Zwischenprüfungen sollten deshalb curricular gut geplant und eingebettet werden – und realistische Anforderungen vorsehen. Aber auch eine zeitgleiche Abgabe zu vieler verschiedener Produkte zwingt die Studierenden zu vermutlich curricular unerwünschten Priorisierungen. Zeigen sich im Betrieb eines Studiengangs Probleme mit bestimmten Leistungsnachweisen und/oder deren Beurteilung, so sind dies in der Regel Hinweise, dass über eine Anpassung des Curriculums nachgedacht werden sollte: Kann das Format gewisser Leistungsnachweise lernförderlicher gestaltet werden? Ist es möglich, summative Beurteilungen zugunsten von formativen Beurteilungen zu reduzieren? Können die Zeitpunkte der

Prüfungen oder die Abgabe/Präsentation von Produkten besser sequenziert werden? Etc.

Zur kontextspezifischen Beantwortung dieser Fragen können die beiden nachfolgend geschilderten Aspekte der Integration hilfreich sein.

10.2 Deklaratives und prozedurales Wissen im Studienverlauf

Die lernpsychologisch nicht haltbare Überzeugung, die ersten Semester eines Studiums dienten der Bildung „fachlicher Grundlagen“ durch Anhäufung von möglichst viel deklarativem Wissen, hält sich hartnäckig. Wie bereits in Kapitel 5.3.2 erläutert, stellt oberflächliches Verstehen wie die Kenntnis wesentlicher Begriffe die Grundlage für tiefergehendes Verstehen dar. Aber damit deklaratives Wissen handlungswirksam wird, muss es stets auch mit Handlungswissen, also prozeduralem Wissen (=Können) gekoppelt werden – wenn jemand ein finnisches Wörterbuch auswendig kann, vermag er noch lange nicht Finnisch zu sprechen, sondern er hat primär träges Wissen erworben (siehe Kap. 4.2). Bach u. Zellweger betonen deshalb, dass bereits ab dem ersten Semester „Fallstudien oder problembasierte Zugänge auf das Lernen im gesamten Studium vorbereiten und auch die soziale und fachliche Integration stärken“ (2022, 83f.). Grundsätzlich sind dafür ebenfalls in der Studieneingangsphase Leistungsnachweise mit offenen Aufgabenstellungen zu empfehlen, z. B. Vorträge (Kap. 7.6.3), kleinere wissenschaftliche Arbeiten (Kap. 7.6.4), mehrphasige Beurteilungen kleinerer Produkte (Kap. 7.7.1) oder Posterpräsentationen (Kap. 7.7.3). Formative Leistungsrückmeldungen spielen zu diesem Zeitpunkt eine noch größere Rolle. Dabei sollte curricular möglichst von Beginn an das Potenzial von Peer- und Selbstbeurteilungen genutzt werden (Kap. 8.4).

10.3 Vertikale und horizontale Integration

Die Konzepte der vertikalen und horizontalen Integration zielen darauf, dass Studierende am besten lernen, wenn sie die Verbindungen zwischen verschiedenen Fachgebieten oder Teilgebieten eines Fachs erkennen können. Ein integriertes Curriculum unterstützt die Studierenden darin, ein tieferes Verständnis für die fachlichen und wenn möglich auch für überfachliche Zusammenhänge zu entwickeln. Dies befördert insbesondere das selbständige und kritische Denken:

- Vertikale Integration bezieht sich auf den Prozess der Verbindung von Lernzielen im zeitlichen Studienverlauf. Sie zielt darauf, dass die Studierenden Wissen und Können zunehmend ausbauen und vertiefen und dabei erkennen, wie verschiedene Aspekte aufeinander aufbauen. Ein Beispiel für eine vertikale Integration wäre etwa, wenn im Studienfach Psychologie im ersten Studienjahr ein Modul zum Teilgebiet Entwicklungspsychologie besucht wird und im zweiten Jahr ein Modul zum Teilgebiet Sozialpsychologie, in dem Konzepte der Entwicklungspsychologie wieder aufgegriffen und im Hinblick darauf behandelt werden, wie soziale Einflüsse die psychologische Entwicklung von Menschen beeinflussen können (siehe auch „verteilte Repetition“ in Kap. 4.6.1).
- Horizontale Integration bedeutet, dass verschiedene Fachgebiete oder Teilgebiete eines Faches miteinander verbunden werden, so dass die Studierenden wesentliche Zusammenhänge herstellen können. Dies ist insbesondere wesentlich, um ein ganzheitlicheres Verständnis von Fachgebieten und ihrem Weltbezug zu entwickeln. Eine horizontale Integration im Rahmen eines Bachelor-Studiengangs in Rechtswissenschaften könnte zum Beispiel folgende Elemente enthalten:
 - Bewusste Thematisierung überlappender Themen: Das Thema „Vertragsabschluss“ spielt in verschiedenen Rechtsgebieten wie Zivilrecht, Handelsrecht oder Arbeitsrecht eine wichtige Rolle. In einem integrierten Curriculum könnten gezielt wesentliche Gemeinsamkeiten und Unterschiede zwischen Vertragsabschlüssen in diesen Rechtsgebieten herausgearbeitet werden.
 - Authentische Situationen: Fallstudien, die verschiedene Rechtsgebiete miteinander verknüpfen, könnten den Studierenden ermöglichen, mit dem Zusammenwirken dieser Rechtsgebiete vertraut zu werden. So könnte ein konkreter Fall das Einbeziehen straf- und zivilrechtlicher Aspekte erfordern.
 - Interdisziplinäre Ansätze: Im Fall der Rechtswissenschaften könnten andere Wissenschaften wie Ethik, Politikwissenschaften oder Wirtschaftswissenschaften im Rahmen interdisziplinärer Ansätze einbezogen werden. So könnten etwa Zusammenhänge zwischen ethischen Prinzipien wie Gerechtigkeit, Gleichheit oder individueller Freiheit und der Ausgestaltung von Gesetzen und Rechtsprechung ausgelotet werden.

Die Kombination von horizontaler und vertikaler Integration trägt wesentlich zur Kohärenz und Konsistenz des Curriculums bei. Man spricht deshalb auch von einem spiralförmigen Curriculum, bei dem Inhalte oder Themen im Verlauf des Studiums auf höherer Komplexitätsstufe wieder aufgegriffen werden (vgl. Ireland u. Mouthaan 2020, 7f.). Eine solche Organisation des Curriculums soll die Studierenden befähigen, ihr Wissen und Können kontinuierlich

zu entwickeln und für künftige Entwicklungen anschlussfähig zu machen, auch über den Studienabschluss hinaus.

In modularisierten Studiengängen, wie sie seit der Bologna-Reform vorgesehen sind, werden Teilleistungen meist nach einzelnen Modulen oder seltener Modulgruppen zertifiziert (vgl. Bach u. Zellweger 2022, 84; Pautsch u. Dillenburger 2011, 67f.). Dadurch haben Schlussprüfungen und Abschlussarbeiten etwas an Bedeutung verloren. Es gibt Befürchtungen, dies begünstige eine Fragmentierung des Lernens, da Studierende seltener ganze Themengebiete auf einmal bearbeiten müssten (vgl. Effinger 2018, 47; Ghisla 2005, 168f.). Leider fehlen empirische Untersuchungen zu Auswirkungen von curricularen Strukturreformen auf das Lernen der Studierenden weitgehend. Dennoch scheint das Konzept von Capstone-Leistungsnachweisen als Instrument zur curricularen Integration interessant: „Capstone-Aktivitäten integrieren eine Reihe von wichtigen Lernprozessen und -ergebnissen“ (Bach u. Zellweger 2022, 84). Sie führen gegen Ende des Studiums Elemente des curricularen Lernprozesses zusammen und bereiten dadurch auch auf den Übergang in die wissenschaftliche oder berufliche Praxis vor (Kinzie 2013). „Integrative Schlussmodule, übergreifende Projekte und Abschlussarbeiten oder Schlusspraktika sind unter diesem Blickwinkel interessante Formen“ (Bach u. Zellweger 2022, 84).

Damit erweisen sich ähnliche Leistungsnachweisformate für die Integration sowohl von Wissen und Können als auch von vertikalen und horizontalen fachlichen Zusammenhängen als vielversprechend. Wenig überraschend sind es Formate, die hochrangige Lernziele stärker betonen, Elemente kooperativen Lernens enthalten und vielfältige Gelegenheiten zu formativen Leistungsrückmeldungen schaffen (siehe v. a. Kap. 7.7 und 8).

11 Leistungsbeurteilung und Prüfungsangst

Viele Menschen empfinden Situationen als unangenehm, in denen ihre Leistung beurteilt wird. Die Grenze zwischen „normalen“ und durchaus auch „gesunden“ Belastungsempfindungen und übermäßigen Ängsten gegenüber Beurteilungssituationen ist nicht einfach zu definieren. Letztere gehören den wichtigsten aktuellen Klassifikationssystemen für psychische Krankheiten zufolge „nicht explizit zu den dort definierten ‚Störungen mit Krankheitswert‘“ (Fehm, Fydrich u. Sommer 2022, 4). Entsprechend gibt es auch keine allgemein anerkannte Definition von Prüfungsangst. Diese hat aber erhebliche Auswirkung sowohl auf das Lernen als auch auf das abrufbare Leistungsvermögen: Sorgen und Selbstzweifel in der Vorbereitungsphase von Prüfungen beeinträchtigen die Verarbeitung von Informationen und ihre Einspeicherung im Gedächtnis. Und akuter, durch Angstempfinden ausgelöster Stress behindert das Abrufen bereits vorhandenen Wissens und Könnens. Die hemmende Wirkung von Ängstlichkeit in Lernphasen ist für das Lernen schädlicher, da sie zu geringerem Lernen führt, während akute Angst in der Leistungssituation „lediglich“ das Zeigen des Gelernten behindert (vgl. Brühwiler u. Helmke 2018, 80). Insgesamt wurde in über 70 Jahren psychologischer Forschung „der Zusammenhang zwischen größerer Prüfungsangst und geringerer Prüfungsleistung (...) in Hunderten von Studien mit vielen Tausend Teilnehmern nachgewiesen“ (von der Embse et al. 2018, 492, Übers. tz).

Nachfolgend wird eine Minimaldefinition von Prüfungsangst gegeben, kombiniert mit Angaben zur Häufigkeit ihres Auftretens (Kap. 11.1). Anschließend werden die wichtigsten Ursachen und Behandlungsmöglichkeiten von Prüfungsangst aufgeführt (Kap. 11.2 und 11.3). Am Ende folgen Hinweise, wie sich angsthemmende Aspekte von Leistungsbeurteilungen verstärken und angstausslösende Aspekte reduzieren lassen (Kap. 11.4). Die aufgeführten Gestaltungsvorschläge entlasten Studierende sowie Lehrende und fördern dadurch tiefenorientiertes Lernen.

11.1 Definition und Häufigkeit von Prüfungsangst

Eine Definition, die viele der wesentlichen Aspekte von Prüfungsangst zu einem Minimalkonsens umfasst, lautet: „Wir definieren Prüfungsangst als einen subjektiven, emotionalen Zustand, der vor oder während einer bestimmten Prüfungssituation erlebt wird und sich auf den Akt der Durchfüh-

rung der Prüfung selbst, die Gefahr des Nichtbestehens und die damit gedanklich verbundenen negativen Folgen bezieht“ (Bonaccio u. Reeve 2010, 618, Übers. tz). Diese Angst kann sich auf vier verschiedenen Ebenen äußern (Fehm, Fydrich u. Sommer 2022, 6–10):

- Emotionale Ebene: Neben dem zentralen Gefühl starker Angst und ggf. akuter Panik können weitere Gefühle wie Besorgnis, Verzweiflung, Hoffnungslosigkeit oder das Empfinden von Minderwertigkeit mit Prüfungsangst verbunden sein. Viele dieser Gefühle haben einen hohen kognitiv-interpretativen Anteil.
- Kognitive Ebene: Prüfungsangst äußert sich kognitiv oft durch intensive Katastrophengedanken, negative Vorstellungen über die Prüfungssituation und befürchtete langfristige Konsequenzen wie Studienabbruch und Versagen. Die Betroffenen zweifeln an ihrer eigenen Kompetenz und erwarten plötzliche „Blockaden“, die mit dem Gefühl des Versagens und der Unfähigkeit einhergehen, die gestellten Fragen zu verstehen. Diese Kognitionen sind stark mit den oben genannten Emotionen verbunden.
- Physiologische Ebene: Auf dieser Ebene ist zu unterscheiden zwischen Symptomen akuter Stresssituationen und solchen, die als langfristige Folge von Prüfungsangst auftreten können. Akute Prüfungsangst äußert sich durch Stressreaktionen wie starkes Herzklopfen, erhöhte Atemfrequenz, Schwitzen, Zittern, Übelkeit und Durchfall. Betroffene können auch sensorische Wahrnehmungsstörungen erleben, etwa die Stimme der Prüferin nur noch von weit entfernt hören. Zudem werden die Blutgefäße verengt, die Extremitäten geringer durchblutet und der Blutdruck erhöht. Insgesamt führt akute Prüfungsangst also zu einer so genannten „Fight-or-Flight-Reaktion“.

Langfristige Prüfungsangst kann demgegenüber zu Konzentrationsproblemen, Schlaflosigkeit, Grübeln, Motivationsmangel und kognitiver Inaktivität führen. Ihre Folgen können somit eher denen einer depressiven Episode gleichen. Dieser längerfristige Stress kann sich bis hin zu einer verringerten Immunantwort auswirken (Borella et al. 1999).

- Verhaltensebene: Im Zusammenhang mit Prüfungsängsten sind zwei Bündel verhaltensbezogener Aspekte zu unterscheiden: Ungünstige Lernstrategien und Vermeidungsverhalten. Zu ungünstigen Lernstrategien gehören etwa exzessives (Auswendig-)Lernen, zu detaillierte Aufarbeitung des Lernstoffs, mangelnde Organisation, fehlende Zusammenfassungen oder eine unzureichende Nutzung von Gedächtnistechniken. Beim Vermeidungsverhalten kann zwischen Vermeiden oder Aufschieben von Prüfungsvorbereitung und anderen Lernhandlungen (Prokrastination) und dem kompletten Vermeiden von Prüfungen unterschieden werden. Vermeidungsverhalten steht im Widerspruch zu den eigentlichen Zielen der

Betroffenen wie dem Abschließen eines Studiums und führt selbst wieder zu schwerwiegenden Dilemmata.

Aufgrund des Fehlens einer breit anerkannten Definition von Prüfungsangst sind Studienergebnisse zur Häufigkeit ihres Auftretens schwer zu vergleichen. Fehm, Fydrich u. Sommer (2022, 14f.) fassen die vorhandenen Studien so zusammen:

- Ca. 25 % der Studierenden erleben Angst vor den meisten Prüfungen.
- Ca. 20 % können in Prüfungssituationen häufig nicht ihr eigentlich vorhandenes Wissen und Können zeigen.
- Starke Prüfungsängste erleben mindestens 5 % der Studierenden.
- Frauen leiden etwas häufiger unter Prüfungsängsten als Männer (vgl. auch von der Embse et al. 2018, 489; für andere Geschlechtsidentitäten sind keine Angaben vorhanden).

Zudem berichten Beratungsstellen, „dass der Anteil Studierender, der sich mit dem Problem Prüfungsangst an eine hochschulnahe Beratungsstelle wendete, in den letzten Jahrzehnten deutlich zugenommen hat“ (Fehm, Fydrich u. Sommer 2022, 16).

11.2 Die wichtigsten Ursachen von Prüfungsangst

Prüfungsangst ist ein komplexes Phänomen, das verschiedene Ursachen haben kann. In der Regel dürfte es sich um ein Zusammenspiel zwischen unterschiedlichen Einflussfaktoren auf verschiedenen Ebenen handeln. Dabei kann zwischen begünstigenden, auslösenden und vorhandene Prüfungsangst aufrechterhaltenden Faktoren unterschieden werden (Fehm, Fydrich u. Sommer 2022, 28–30):

- begünstigende (prädisponierende) Faktoren:
 - personenbezogene Faktoren, z. B. allgemeine Ängstlichkeit und Depressivität, Perfektionismus, Neigung zu katastrophisierendem Denken oder instabiler Selbstwert
 - Misserfolgserfahrungen, subjektiv als ungerecht erlebte Bewertungen oder prüfungsängstliche Personen als „Modelle“ im persönlichen Umfeld
 - biologische Vulnerabilität und Belastungen, z. B. erhöhte physiologische Angstbereitschaft oder dauerhaft erhöhtes Stressniveau

- auslösende Faktoren: z. B. negatives Prüfungserlebnis, Aufkommen von Zweifeln an der eigenen Kompetenz während der Vorbereitung auf eine Prüfungssituation und/oder belastende Lebenssituationen (wie Geldsorgen, kranke Angehörige etc.)
- aufrechterhaltende Faktoren:
 - situative Variablen, z. B. Schwierigkeitsgrad der Prüfung, Zeitdruck, Prüfungsbedingungen inklusive Haltung der Prüfenden und Kommunikation mit Studierenden
 - Kognitionen, z. B. (anhaltendes) Katastrophisieren („wenn ich hier durchfalle, ist mein Leben ruiniert“) oder hoher Leistungsdruck („ich bin nur mit einer 1 gut genug“)
 - Verhalten, z. B. Aufschieben des Lernens/Arbeitens (Prokrastination) oder Absagen von Prüfungen (Vermeidungsverhalten)

Je mehr begünstigende Faktoren vorhanden bzw. je stärker sie ausgeprägt sind, desto geringfügigere Erlebnisse können ausreichen, um Prüfungsangst auszulösen. Je mehr aufrechterhaltende Faktoren aufgebaut werden bzw. je stärker sie ausgeprägt sind, desto mehr wird sich die Prüfungsangst verfestigen und intensivieren.

11.3 Behandlungsmöglichkeiten von Prüfungsangst

Auch wenn die Ursachen von Person zu Person variieren, sind die grundlegenden Muster der Wirkfaktoren oft ähnlich. Das Verständnis ihrer Zusammenhänge kann helfen, Strategien zur Bewältigung und Verringerung der Prüfungsangst zu entwickeln. Allerdings sollten wir Studierenden mit ausgeprägter Prüfungsangst empfehlen, Hilfe bei Fachleuten, also Psychologinnen und Psychologen, zu suchen. An vielen Hochschulen gibt es entsprechende Anlaufstellen, seien es allgemeine Beratungsstellen für Studierende oder spezielle psychologische Beratungseinrichtungen für Studierende mit psychischen Schwierigkeiten. Alternativ können natürlich auch psychologische Fachpersonen außerhalb der Hochschule aufgesucht werden, die sich auf das Behandeln von Personen mit Prüfungsängsten verstehen.

Seit den 1980er Jahren wurden verschiedene Metaanalysen vorgenommen, um die Wirksamkeit unterschiedlicher Ansätze zur Behandlung von Prüfungsängsten zu untersuchen. Trotz methodologischer Unterschiede sind die Befunde grundsätzlich ähnlich. Zentral scheinen dabei zwei Aussagen: Ansätze, in denen Betroffene lediglich ihre Studiertechniken verbessern, zeitigen keine Wirkung. Als besonders wirksam erweisen sich dagegen von Fachleuten durchgeführte Behandlungsinterventionen, bei denen behaviorale oder kognitiv-behaviorale Behandlungstechniken mit einem Training der Stu-

dierfertigkeiten kombiniert wurden (vgl. zu diesen Befunden Huntley et al. 2019, 47; Fehm, Fydrich u. Sommer 2022, 95–97). Welche Behandlungsintervention in welchem individuellen Fall besonders angemessen ist, sollte jedoch den behandelnden Fachleuten überlassen werden. Wenden sich Studierenden mit bereits ausgeprägten Prüfungsängsten an Lehrende, so scheint es am hilfreichsten, ihnen gut zuzuhören, verständnisvoll zu reagieren und sie dann auf geeignete Unterstützungsangebote innerhalb oder außerhalb der Hochschule hinzuweisen. Lehrende können überdies durch die Gestaltung von Prüfungssituationen und ihr Verhalten im Umgang mit Leistungsanforderungen dazu beitragen, dass möglichst wenige Faktoren entstehen, die bei Studierenden zum auslösenden oder aufrechterhaltenden Moment (siehe Kap. 11.2) für Prüfungsangst werden können. Dies wird nachfolgend erläutert.

11.4 Angstausslösende und angsthemmende Aspekte von Leistungsbeurteilung

Die Lehr- und Beurteilungskultur an Hochschulen kann vor allem hinsichtlich der angstausslösenden Faktoren und teilweise hinsichtlich der angstaufrechterhaltenden Faktoren angsthemmend gestaltet werden. Dabei können einerseits angstausslösende Momente reduziert und angsthemmende Aspekte verstärkt werden. Viele entsprechende Vorgehensweisen wurden in den Kapiteln 5 bis 9 vorgestellt. Sie werden nachfolgend mit Verweis auf die jeweiligen Textstellen aufgelistet.

Angsthemmende Elemente im Umgang mit Leistungsbeurteilung

- **Constructive Alignment:** Es wird wirklich das summativ geprüft, was die Lernziele benennen, und die Lehr-Lern-Aktivitäten führen auf das Erreichen der Lernziele hin (Kap. 5). Brüche in dieser Abstimmung führen zu Inkonsistenzen und vor allem zu einer Intransparenz bezüglich der eigentlichen Lernziele und ihrer summativen Überprüfung. Solche Unklarheiten belasten Studierende unnötig und können Ängste auslösen.
- Für die zu beurteilenden Leistungen liegen klare und verständliche Beurteilungskriterien vor, die mit den Studierenden diskutiert und soweit möglich sogar erarbeitet werden (Kap. 9, besonders 9.3).
- **Formative Feedbackkultur:** Eine Beurteilungskultur, in der formative Leistungsrückmeldungen durch verschiedene Instanzen (Lehrende, Peers, Selbstbeurteilungen u. a. m.) erfolgen und für das weitere Lernen berück-

sichtigt werden, wirkt angsthemmend (Kap. 8). Sie erhöht die intrinsische Lernmotivation und die lernbezogene Selbstwirksamkeit der Studierenden. Diese schätzen somit auch summative Beurteilungen positiver ein, da sie das Lernen als sinnvoll erleben und besser einschätzen können, was sie wie gut können (Kap. 4.5). Besonders wichtig ist in diesem Zusammenhang, dass Lehrende ihr Feedback an die Studierenden motivationsförderlich gestalten (Kap. 8.2).

- **Kommunikationskultur:** Im Zusammenhang mit einer formativen Feedbackkultur ist auch wichtig, dass die Kommunikation zwischen Lehrenden und Studierenden möglichst offen und transparent gestaltet wird. Studierende sollten Lehrende hinsichtlich der Lehrinhalte, Lernziele und Beurteilungskriterien alles fragen können, ohne mit unfreundlichen Antworten oder gar einer Auswirkung auf die summative Beurteilung rechnen zu müssen (vgl. Rost, Sparfeldt u. Buch 2018, 431).

Angstausslösende Aspekte der Leistungsbeurteilung reduzieren

- **Kollektive Bezugsnorm vermeiden:** Bei summativen Leistungsnachweisen ist wenn immer möglich zu vermeiden, dass Studierende anhand der kollektiven Bezugsnorm bewertet werden. Dies fördert unnötiges Konkurrenzdenken zwischen den Studierenden, was einer kooperativen und offenen Lernatmosphäre abträglich ist (Kap. 6.2, vgl. auch Rost, Sparfeldt u. Buch 2018, 426). Zwingend zu vermeiden sind in diesem Zusammenhang Bewertungsskalen, die eine Normalverteilung der zu bewertenden Leistungen voraussetzen (Kap. 5.1.3, Kap. 9.2.1).
- **Curriculare Gestaltung:** Es sollte nicht zu viel von einer einzigen Prüfung abhängen. Der erfolgreiche Abschluss eines Studiums sollte von einer größeren Zahl an (summativen) Beurteilungsanlässen abhängen, bei denen unterschiedliche Arten von Wissen und Können anhand unterschiedlicher Leistungsnachweisformate beurteilt werden (Kap. 7 und 10). Zudem sollten verschiedene Beurteilungsanlässe nicht innerhalb eines zu kurzen Zeitraums erfolgen, da auch dies unnötig den psychischen Druck erhöht und zudem nachhaltiges Lernen behindert (Kap. 4, insbes. Kap. 4.5 und 4.6). Ebenfalls ist ein hoher Zeitdruck in Klausuren zu vermeiden, da dadurch mehr die Stressresistenz der Studierenden als ihre fachlichen Fähigkeiten geprüft werden. Eine Ausnahme bilden hier Lernziele, die sich explizit auf Handeln und Zeitdruck richten – typisches Beispiel dafür ist die Notfallmedizin.

Weiterführende Ansätze

- **Vermittlung von Lernstrategien und Arbeitstechniken:** Ein Teil der Studierenden, die starke Prüfungsangst entwickeln, zeigen Defizite bezüglich der verwendeten Lernstrategien und Arbeitstechniken. Frühere Modelle der Prüfungsangst gingen deshalb davon aus, dass solche Defizite die Hauptursache für die Angstentstehung seien. Diese Defizitmodelle wurden inzwischen als zu einseitig verworfen, da auch Studierende mit guten Lernstrategien starke Prüfungsangst entwickeln können. Defizite im Bereich der Lern- und Arbeitstechniken gelten aber weiterhin als ein angstbegünstigender Faktor (vgl. Fehm, Fydrich u. Sommer 2022, 23f. und 29).

Es scheint deshalb sinnvoll, den Studierenden in der Eingangsphase von Studiengängen Lernstrategien und Arbeitstechniken zu vermitteln. Diese sollten für die jeweilige Studienstufe und Fachrichtung besonders relevant und möglichst parallel zu ihrer Vermittlung anwendbar sein, etwa bei der Vorbereitung von Prüfungen oder beim Erarbeiten von Produkten wie Postern, Präsentationen, Blogbeiträgen usw. Dadurch kann bei einer erheblichen Zahl Studierender ein angstbegünstigender Faktor eliminiert werden, und natürlich sollten sich die erworbenen Lernstrategien und Arbeitstechniken generell günstig auf das Lernen auswirken.

- **Adressieren von Heterogenität:** Studierende weisen zum Studienbeginn eine enorme und in den letzten Jahrzehnten stark gewachsene Heterogenität bezüglich ihres Vorwissens, aber auch des sozioökonomischen Hintergrunds auf (vgl. Hanft, Zawacki-Richter u. Gierke 2015; Kiener 2012). Für Lehrende ist es äußerst anspruchsvoll, die Lehr- und Lernaktivitäten so zu gestalten, dass sie für Studierende mit verschiedenem Vorwissen, unterschiedlichen Bildungsbiografien und diversem sozialen Hintergrund gleiche Lerngelegenheiten bieten. Dies stellt besonders Angehörige von Minderheiten vor Herausforderungen, etwa Studierende der ersten Generation, Frauen in MINT-Studiengängen oder auch Männer in einigen Studiengängen (z. B. Studium zur Lehrperson in Grundschule/Primarschule).

Eine vielversprechende Möglichkeit wird derzeit an der ETH Zürich erprobt: Im Rahmen von MINT-Vorlesungen im ersten Studienjahr können sich Studierende als Alternative zu den „gewöhnlichen“ Übungsgruppen in Fokusgruppen eintragen. Diese richten sich insbesondere an Studierende, die ihr Vorwissen bezüglich der Veranstaltungsthematik als zu gering einschätzen (was nicht zwingend mit einem tatsächlich geringen Vorwissen einhergehen muss). Diese Fokusgruppen weisen demzufolge einen gegenüber der Gesamtkohorte erhöhten Anteil an Frauen und an Studierenden der ersten Generation auf. Die Tutoren in diesen Gruppen fokussieren stärker auf Kernkonzepte der Vorlesung und behandeln diese detaillierter. Eine erste Erprobung im Physikstudium mit zwei Fokus-

gruppen im Herbstsemester 2021 zeigte positive Resultate: Die Teilnehmenden der Fokusgruppen waren mit diesem Format zufrieden und wiesen mit 79,6 % sogar eine leicht erhöhte Bestehensquote gegenüber den 74,6 % der Gesamtkohorte auf. Das Format wurde deshalb im Herbst 2022 bereits in zehn Vorlesungen und vier verschiedenen Departementen angeboten (diese Informationen sind der Präsentation von Niggli 2023 entnommen).

12 Wissenschaftliche Arbeiten begleiten und beurteilen

Das Begleiten von (längeren) schriftlichen Arbeiten inklusive Abschlussarbeiten wie Bachelor- und Masterarbeiten oder auch Dissertationen stellt besondere Ansprüche. Grundsätzlich sei dazu auf einschlägige Publikationen verwiesen wie jene von Girgensohn u. Sennewald (2012) oder spezifisch zum Begleiten von Abschlussarbeiten das hilfreiche Büchlein von Buff Keller u. Jörissen (2015). Im Folgenden werden lediglich einige Besonderheiten beim Begleiten von wissenschaftlichen Arbeiten aufgeführt (Kap. 12.1) und die Wichtigkeit der Kontraktierung von Betreuungsverhältnissen dargelegt (Kap. 12.2).

12.1 Besonderheiten beim Begleiten wissenschaftlicher Arbeiten

Hochschullehrende sind oft stark auf den wissenschaftlichen Diskurs und methodische Fragen fokussiert. Selbstverständlich sind diese Aspekte wichtig, dienen doch schriftliche Arbeiten in der Regel auch der Heranführung und Einarbeitung in den fachlichen Diskurs und/oder in relevante Forschungsmethoden.

Zugleich stellt für Studierende – auch noch am Ende des Masterstudiums – das wissenschaftliche Schreiben an sich bereits eine erhebliche Herausforderung dar. Denn akademische Schreibkompetenz muss über mehrere Jahre aufgebaut werden. Dieser Entwicklungsprozess des wissenschaftlichen Schreibens verläuft gemäß Steinhoff (2007, 130–48) über mehrere Stufen, bis ein eigener Umgang mit den domänenspezifischen Anforderungen des schriftlichen wissenschaftlichen Diskurses gefunden wird. Der Erwerb und Ausbau akademischer Schreibfähigkeiten ist so komplex, dass er mit dem Masterabschluss oder Doktorat nicht abgeschlossen ist und lebenslang andauern kann.

Deshalb stellen kürzere und längere wissenschaftliche Arbeiten sowie Abschlussarbeiten für Studierende wichtige Meilensteine auf dem Weg zu wissenschaftlichen Schreibfähigkeiten dar. Eine gute Begleitung durch Lehrende auch im Hinblick auf Aspekte des wissenschaftlichen Schreibens ist dabei von großem Wert.

Während der Betreuung nehmen Lehrende gegenüber den Studierenden unterschiedliche Rollen ein: Den größten Teil der Zeit haben sie die Rolle

von Begleitenden inne, wechseln aber in der Regel am Ende in die beurteilende Rolle. Nachdem sie die Studierenden lange unterstützend begleitet haben, müssen sie nun plötzlich ein abschließendes Urteil über die schriftliche Arbeit fällen. Der sich daraus ergebende Rollenkonflikt sollte reflektiert werden, um einen angemessenen Umgang damit zu finden (siehe Kap. 3.3.1, vgl. auch Buff Keller u. Jörissen 2015, 86–89). Zentral ist, dass sowohl den Studierenden als auch den begleitenden und beurteilenden Lehrenden bewusst ist, wie lange sie formative Leistungsrückmeldungen geben und ab wann sie summativ beurteilen. Das in Kapitel 7.7.1 beschriebene mehrphasige Beurteilen eines Endprodukts kann einen sinnvollen Rahmen zum Umgang mit den beiden Rollen des Begleitens und Beurteilens bieten.

Vertieftes Wissen und ein ausführlicher theoretischer Hintergrund zum Thema des wissenschaftlichen Schreibens und seines Erwerbs finden sich z. B. in den Monografien von Ortner (2000) und Pohl (2007).

12.2 Betreuungsverhältnis kontraktieren

Wenn die Beteiligten sich entscheiden, ein Betreuungsverhältnis einzugehen, sollte ein Contracting erfolgen: In einer gemeinsamen Vereinbarung werden wichtige Punkte der Zusammenarbeit schriftlich festgehalten. Transparente und verbindliche Vereinbarungen vor Beginn der eigentlichen Betreuung schaffen günstige Rahmenbedingungen. Sie zeigen auch, ob die Parteien tatsächlich gegenseitig kompatible Vorstellungen von der konkreten Gestaltung des Ablaufs und der Verteilung der Verantwortlichkeiten haben. Der Kontrakt ermöglicht zudem durch gemeinsames Unterschreiben auch einen ritualisierten Start in die Beratung.¹⁵¹ Die Unterzeichnung kann auch symbolisch erfolgen, etwa im Mailkontakt.

Sind weitere Personen beteiligt, zum Beispiel hochschulexterne Praxispartner oder weitere Bewertende, dann sind sie ebenfalls ins Contracting einzubeziehen. In diesen Fällen ist es sinnvoll, einen Dreieckskontrakt ins Auge zu fassen. Ein solcher sollte aufgrund der inhärenten Komplexität stets schriftlich geregelt werden. Die Seiten eines solchen Dreiecks können umfassen: Vereinbarung zwischen Betreuerin und Student, Vereinbarung zwischen Betreuerin und externem Praxispartner und Vereinbarung zwischen Student

151 Der genaue rechtliche Status eines solchen Dokuments ist vorgängig zu klären und hängt von den rechtlichen Rahmenordnungen der Hochschule ab. Es spannt sich eine Skala auf von Hochschulen oder Studiengängen, die standardmäßig solche Vereinbarungen vorsehen bis zu solchen, wo die Regelungen das formale Unterzeichnen einer Betreuungsvereinbarung ausschließen.

Vereinbarung zur Masterarbeit von ...	
Thema der Arbeit:	
Student	
Name, Kontaktdaten: weitere Angaben (Vorbildung; besuchte, für die Arbeit relevante Veranstaltungen; Interessen etc.):	
Betreuerin	
Name, Kontaktdaten: Erreichbarkeit:	
Beginn:	Abgabetermin:
Auftrag/Ziel:	
Gegenseitige Erwartungen	
an den Studenten:	
an die Betreuerin:	
Organisation der Betreuung	
Struktur/Ablauf: weitere Beteiligte: Kompetenzen und Zuständigkeit (Begleitung, Benotung):	
Ressourcen	
Zeitaufwand Student/Betreuerin:	
Infrastruktur:	
Finanzen:	
Termine (Besprechungen, Meilensteine):	
Ort, Datum:	
Unterschrift Betreuerin	Unterschrift Student

Abb. 21: Beispiel für eine Vereinbarung zur Betreuung einer Masterarbeit.
Quelle: Eigene Abbildung (inhaltlich weitgehend übernommen von Buff
Keller und Jörissen 2015, 65f.)

und externem Praxispartner (vgl. zu diesen Ausführungen Buff Keller u. Jörissen 2015, 65f.; Thomann u. Pawelleck 2013, 42–45; Thomann 2017b). Abbildung 21 zeigt ein konkretes Beispiel für eine schriftliche Vereinbarung zur Betreuung einer Masterarbeit.

13 Multiple-Choice-Prüfungen

Leistungsnachweise sind oft aufwändig in der Gestaltung und Beurteilung. Prüfungsaufgaben mit Mehrfachwahl scheinen effizient und sparen besonders beim Korrigieren viel Zeit. Allerdings sind valide Mehrfachwahl-Aufgaben aufwändig zu konstruieren – also Aufgaben, die wirklich das prüfen, was die Studierenden können sollten. Um auch über die Ökonomie hinaus abschätzen zu können, wofür sich das Prüfen mit MC-Items eignet, spricht dieses Kapitel nach einer Begriffsklärung (Kap. 13.1) einige zentrale Aspekte an: Möglichkeiten und Grenzen von MC-Prüfungen (Kap. 13.2), Struktur und Typen von MC-Aufgaben (Kap. 13.3), das Problem der ungewollten Lösungshinweise (Kap. 13.4.) und Fragen der Auswertung von MC-Prüfungen (Kap. 13.5). Zudem wird abschließend darauf hingewiesen, dass sich MC-Prüfungen auch für formative Selbsttests eignen (Kap. 13.6).

Wer anhand dieser Ausführungen zum Schluss kommt, das Format anwenden zu wollen, sollte sich unbedingt vertieft mit den Anforderungen an eine angemessene Konstruktion und Auswertung auseinandersetzen. Dazu kann ich insbesondere die Arbeiten von René Krebs empfehlen: Bei Krebs (2004) handelt es sich um eine äußerst hilfreiche, im Internet verfügbare Kurzanleitung zur Erstellung von MC-Items, welche die Lesenden innerhalb kürzester Zeit befähigt, brauchbare Items zu erarbeiten. In Krebs (2019) geht der gleiche Autor auf Hintergründe sowie auf die Fragen der Auswertung und Bewertung von Aufgaben und ganzen Prüfungen genauer ein. Auch die Möglichkeiten elektronischer Prüfungen werden im Buch berücksichtigt. Es ist daher besonders auch jenen Lehrenden zu empfehlen, die nicht nur einzelne Items herstellen, sondern ganze Prüfungen konzipieren und auswerten wollen.

13.1 Zum Begriff „Multiple Choice“

Der Begriff „Multiple Choice“ hat sich als Begriff für die Aufgabentypen etabliert, die mit einer vorgegebenen Anzahl von Antworten arbeiten und deshalb korrekterweise als Mehrfachwahl-Aufgaben zu bezeichnen wären. Strenggenommen bezeichnet „Multiple Choice“ jene Aufgabenformate, in denen mehrere Antworten ausgewählt werden können. Die weitaus üblicheren und meistens geeigneteren Formate, bei denen nur eine Antwort zu wählen ist, werden eigentlich als „Single Choice“ bezeichnet. In diesem Buch

wird aber wie im didaktischen Diskurs üblich der Begriff „Multiple Choice“ als Oberbegriff für alle Typen von Mehrfachwahl-Aufgaben verwendet.

13.2 Was leisten MC-Prüfungen – und was nicht?

MC-Prüfungen orientieren sich am quantitativen Verständnis von Leistungsbeurteilung, wie es in den Kapiteln 6.3., 6.5 und 9.2 erläutert wird. Wie bereits erwähnt sind MC-Prüfungen, die wirklich die gemäß den Lernzielen zu erbringenden Lernleistungen überprüfen, anspruchsvoll und aufwändig zu erarbeiten. So meint Krebs (2004, 1): „Auch geübte Autoren rechnen durchschnittlich mit einer Stunde Arbeit, um ein MC-Item herzustellen.“ Da eine MC-Prüfung minimal 30 Items umfassen sollte und oft 40 oder mehr Items für eine statistisch zuverlässige Auswertung nötig sind, ist deshalb jenseits didaktischer Überlegungen abzuwägen, ob das Format zeitökonomisch wirklich sinnvoll ist. Entspricht die Studierendenzahl nicht mindestens dem Doppelten der Anzahl Items, dürfte dies kaum der Fall sein; zudem verlieren die statistischen Verfahren, auf denen eine sachgemäße Auswertung von MC-Prüfungen beruht, bei einer Zahl von deutlich unter 100 beurteilten Personen schnell an Aussagekraft (vgl. Krebs 2019, 123–127). Werden die Items mit geringem Zeit- und Denkaufwand hergestellt, wird der Zeitgewinn mit Abstrichen bei den testtheoretischen Gütekriterien, also der inhaltlichen Qualität, erkaufte (siehe Kap. 6.3 zu den Gütekriterien der Leistungsbeurteilung). MC-Prüfungen sind also nur unter bestimmten Umständen eine für Lehrende zeitsparende Alternative zu anderen Formaten von Leistungsnachweisen.

Aber auch inhaltlich können mit MC-Aufgaben bei weitem nicht alle Arten von Lernzielen gut überprüft werden. Grundsätzlich sind MC-Aufgaben für die Überprüfung der unteren Stufen der kognitiven Lernzieltaxonomien geeignet. In der SOLO-Taxonomie sind dies die Stufen unistruktural und multistruktural, in der revidierten Taxonomie nach Bloom sind es die Stufen Wissen und Verstehen (die genannten Taxonomien werden in Kap. 5.4.2 vorgestellt). Ob MC-Aufgaben auch das Erreichen der relationalen Stufe (SOLO) bzw. der Stufe Anwenden (Bloom revidiert) zu überprüfen erlauben, ist umstritten (vgl. Biggs u. Tang 2007, 203f.; Haladyna 2004, 19–64). Ich vertrete die Ansicht, dass Formate von Leistungsnachweisen, die handlungsnäher sind als MC-Prüfungen, sich für das Überprüfen höherrangiger Lernziele besser eignen.

13.3 Struktur und Typen von MC-Aufgaben

Unter den verschiedenen Arten von Multiple-Choice-Fragen lassen sich zwei Grundtypen unterscheiden: Best-Choice-Typen und Richtig/Falsch-Typen. Best-Choice-Typen erfordern in der Regel die Wahl einer einzigen, passendsten Antwort, während die Distraktoren (Falschantworten) auch richtige Anteile aufweisen können. Dies ermöglicht die Konstruktion komplexerer und realitätsnäherer Problemstellungen, zu deren Beantwortung anspruchsvollere kognitive Prozesse erforderlich sind. Bei Richtig/Falsch-Typen ist für jede einzelne Antwort oder Aussage zu entscheiden, ob sie richtig oder falsch ist. Daher sind nur Fragestellungen/Aussagen geeignet, bei denen diese Entscheidung eindeutig getroffen werden kann. Richtig/Falsch-Typen bergen deshalb noch stärker als Best-Choice-Typen das Risiko, kontextarmes Faktenwissen zu überprüfen (knowing the name of something statt knowing something, siehe Kap. 4.2).

13.3.1 Struktur

Man unterscheidet bei MC-Fragen zwischen dem Fragestamm und den Wahlantworten. Der Stamm kann einfach oder zweiteilig sein (vgl. Krebs 2019, 52–54). Ein einteiliger Fragestamm besteht aus einer einfachen, prägnanten Frage wie z. B. „Was verstand Jean Piaget unter dem Begriff der Äquilibration?“. Mit solchen Fragen lässt sich in der Regel bloß Faktenwissen prüfen.

Deshalb wird oft mit Vignetten gearbeitet, mit denen zuerst eine komplexere Problemstellung dargelegt wird. Auf diese folgt eine explizite Fragestellung, so dass der Fragestamm in diesem Fall zweiteilig ist. Vignetten sind erforderlich, um das Erreichen höherrangiger Lernziele zu überprüfen (SOLO: multistuktural, evtl. relational; Bloom revidiert: Verstehen und evtl. Anwenden). Dabei müssen Vignetten nicht nur aus sprachlichen Beschreibungen bestehen, sondern können auch Bilder, Grafiken, Karten etc. umfassen. Bei elektronischen Prüfungen sind sogar kurze Ton- und Filmdokumente als (Teil von) Vignetten möglich. Einfache Beispiele für Vignetten sind in den nachfolgenden Beispielaufgaben enthalten.

13.3.2 Best-Choice-Typen

Es gibt verschiedene Typen von Best-Choice-Aufgaben, die regelmäßig eingesetzt werden und empirisch untersucht wurden. Tabelle 19 gibt eine Übersicht der wichtigsten Typen.

Tab. 19: Wichtigste Best-Choice-Typen nach Krebs (2019, 73; 2004, 5)

Typ	Aufgabe
A	Die richtige/beste Antwort muss aus (meist) fünf Antwortmöglichkeiten gewählt werden.
B	Mehreren Fragen wird die jeweils richtige/beste aus einer Gruppe von (meist fünf) Wahlantworten zugeordnet.
R	Mindestens zwei Fragen wird die jeweils richtige/beste aus einer langen Gruppe von bis zu 26 Wahlantworten (A–Z) zugeordnet.
PickN	Mehrfachwahl, d. h. gleich konstruiert wie Typ A oder R, aber es muss mehr als eine Antwortmöglichkeit ausgewählt werden für die richtige/beste Lösung (→ Multiple Choice im wörtlichen Sinn).

Nachfolgend wird ein Beispiel für Typ A gegeben, da er international am besten erprobt und bewährt ist (Krebs 2019, 73).

Vignette: Sie nehmen als Beisitz an einer mündlichen Prüfung in Ihrem Fachgebiet teil. Die prüfende Dozentin stellt der vierten Studentin die gleichen drei Fragen wie dem dritten Studenten, den sie unmittelbar zuvor geprüft hat. Studentin B beantwortet zwei der drei Fragen ein wenig besser als Student A, dem die Dozentin die Note 3 erteilt hat. Die dritte Frage beantworten beide Studierenden in vergleichbarer Weise. Nun möchte die Dozentin Studentin B die Note 1 erteilen. Sie sind nicht einverstanden und halten die Note 1 von Studentin B für zu hoch im Vergleich zur Note 3 von Student A.

Frage: Welche kognitive Verzerrung ist die wahrscheinlichste Ursache für die Ihrer Ansicht nach zu hohe Note von Studentin B?

Wahlantworten:

- a) Hier liegt ein Halo-Effekt vor.
- b) Hier liegt ein Kontrasteffekt vor.
- c) Hier liegt ein geschlechtsspezifisches Vorurteil vor.
- d) Hier liegt eine Bestätigungstendenz vor.
- e) Hier liegt eine Tendenz zu extremen Urteilen vor.

Lösung: b ist die beste Antwort.

Hinweis zur Konstruktion von Best-Choice-Aufgaben: Oft wird unterschätzt, wie anspruchsvoll das Formulieren guter Distraktoren ist. Diese sollten einerseits nicht auf den ersten Blick als falsch erkennbar sein, müssen sich aber andererseits doch einigermaßen deutlich in ihrer Korrektheit/Angemessenheit von der besten Antwort abheben. Problematisch ist zudem die Verwendung von Negativformulierungen oder von Frageformaten, bei denen nach falschen Antwortoptionen gesucht werden muss. Sie prüfen in der Regel eher logisches Denken als fachliche Kenntnisse.

13.3.3 Richtig/Falsch-Typen

Es gibt verschiedene Typen von Richtig/Falsch-Aufgaben, die regelmäßig eingesetzt werden und empirisch untersucht wurden. Tabelle 20 gibt eine Übersicht der gebräuchlichsten Typen.

Tab. 20: Wichtigste Richtig/Falsch-Typen nach Krebs (2019, 73f.; 2004, 5)

Typ	Aufgabe
Kprim	Vierfache Entscheidung über richtig oder falsch
R/F	Einzelne Aussagen müssen als richtig oder falsch beurteilt werden.
E	Zwei Aussagen und ihre kausale Verknüpfung müssen als richtig oder falsch beurteilt werden.

Hier wird ein Beispiel für den Typ Kprim gegeben, der hinsichtlich statistischer Messwerte wie Schwierigkeit und Trennschärfe die besten Werte erreicht (Krebs 2019, 73 u. 85–94).

Vignette: Sie wenden in Ihrem Modul seit zwei Jahren ein neu entwickeltes Leistungsnachweisformat an, das der abschließenden summativen Bewertung dient, bei der Noten erteilt werden. Aufgrund der Erfahrungen in den ersten beiden Durchführungen sind Sie der Ansicht, dass sich Authentizität und Validität der anhand des Formats vorgenommenen Leistungsbeurteilung noch verbessern lassen.

Frage: Welche der folgenden Maßnahmen eignen sich dazu? (Jede der vier Optionen ist mit richtig oder falsch zu beantworten)

Wahlantworten:

- Sie überprüfen, ob sich das verwendete Leistungsnachweisformat durch weitere Aufgabentypen ergänzen lässt.
- Sie überprüfen die Bezugsnorm, anhand deren das Leistungsnachweisformat beurteilt wird.
- Sie überprüfen anhand des Konfidenzintervalls der bisherigen Durchführungen die Möglichkeit von Nachprüfungen.
- Sie überprüfen das Constructive Alignment Ihres Moduls, einschließlich des Lernzielbezugs Ihres Leistungsnachweises.

Lösung: a und d sind richtig, b und c falsch.

Hinweis zur Bewertung: Um einen ganzen Punkt für eine richtige Antwort zu erhalten, müssen alle vier Fragen korrekt beantwortet werden. Laut Krebs (2019, 85–90) haben Auswertungen gezeigt, dass es sinnvoll ist, drei richtige Antworten mit einem halben Punkt zu bewerten. Für zwei richtige Antworten oder gar nur eine richtige Antwort sollten hingegen keine Punkte erteilt wer-

den. Diese Empfehlung wird angesichts der Ratewahrscheinlichkeit nachvollziehbar: Die Wahrscheinlichkeit, dass mindestens eine der vier Antworten zufällig korrekt geraten wird, ist mit 93,75 % sehr hoch. Auch zwei Antworten werden immer noch mit einer Wahrscheinlichkeit von 37,5 % korrekt erraten, während dieser Wert für drei korrekte Antworten auf 25 % sinkt (das ist der gleiche Wert, wie wenn aus vier Antwortoptionen die beste/richtige zu wählen ist). Die Wahrscheinlichkeit, dass alle vier Antworten korrekt erraten werden, ist hingegen mit 6,25 % sehr gering. Weitere Hinweise zur Auswertung von MC-Prüfungen erfolgen in Kap. 13.5.

13.4 Ungewollte Lösungshinweise vermeiden

Bei der Konstruktion von MC-Aufgaben gibt es eine beträchtliche Anzahl Stolpersteine. Beispielsweise ist es anspruchsvoll, im Zusammenspiel von Fragestellung und Antwortoptionen jegliche unfreiwilligen Lösungshinweise zu vermeiden, die es erlauben, ohne Sachwissen auf die richtige Antwort zu schließen oder bestimmte Antwortoptionen als falsch auszuschließen. Ein einfaches Beispiel für einen Lösungshinweis wäre die folgende Frage (abgewandelt aus Philipps 1978, 13):

Ein Unternehmensdelikt ist eine

- a) abstraktes Gefährdungsdelikt.
- b) durch ein Organ einer juristischen Person begangenes Delikt.
- c) Strafvorschrift, bei der die Vollendung und der Versuch gleichgestellt sind, so dass die Vollendungsstrafbarkeit vorverlagert ist.
- d) Wirtschaftsverbrechen.
- e) durch einen Wirtschaftsbetrieb begangenes Verbrechen.

Hier passt nur die Option c) grammatikalisch zur Formulierung der Frage. Ein weiterer typischer Hinweis besteht darin, dass Option c) am ausführlichsten und komplexesten formuliert ist – die längste/komplexeste Antwort ist oft die anzuwählende.

Lösungshinweise können sich zudem auch aus der Kombination verschiedener Fragen ergeben, etwa wenn in einer Frage nach einem Umstand gefragt wird, der im Fragestamm einer vorangehenden Frage bereits als gegeben vorausgesetzt worden war.

Insgesamt ergeben sich deshalb folgende Empfehlungen zur Vermeidung von Lösungshinweisen (vgl. Krebs 2019, 60–67; Krebs 2004, 16–21):

- Die Antwortoptionen sind ähnlich lang/komplex.
- Alle Antwortoptionen schließen grammatikalisch korrekt an die Frage an.

- Es kommen keine absoluten Begriffe vor wie nie, immer, einzig, nur. Sie sind fast immer Hinweise auf Falschantworten, da wissenschaftlich korrekte Aussagen meist differenzierter sind.
- Antwortoptionen wie „Alle Antworten sind richtig“ bzw. „...sind falsch“ sind zu vermeiden, da sie zu logischen Problemen führen.¹⁵²
- Die Position der richtigen Antwortoption variiert von Frage zu Frage und über die ganze Prüfung hinweg ungefähr gleichmäßig zwischen allen Positionen (Menschen haben eine Tendenz, die richtige Antwort in der Mitte zu platzieren). Wird die Prüfung online durchgeführt, ist es empfehlenswert, von der Softwareoption zum zufälligen Anordnen der Antwortoptionen Gebrauch zu machen. Studierende sollten darauf hingewiesen werden, dass die Antwortoptionen zufällig angeordnet sind.
- Es ist darauf zu achten, dass Fragen sich nicht aufgrund zuvor gestellter Fragen einfacher beantworten lassen.

Tipp: Lassen Sie Ihre MC-Aufgaben/Ihre Prüfungen von einer fachfremden Person beantworten, die das Beantworten von MC-Aufgaben gewohnt ist. Diese Person kann Sie auf etwaige ungewollte Lösungshinweise aufmerksam machen.

13.5 Auswertung von MC-Prüfungen

Bei der Auswertung von MC-Prüfungen mit dem Ziel einer summativen Beurteilung spielen statistische Überlegungen und Berechnungen eine wesentliche Rolle. So ist beispielsweise von der Gewichtung der Aufgaben abzuraten, da auch bei einer Frage vom Typ A stets eine Wahrscheinlichkeit von 20 % besteht, dass die richtige Antwort rein zufällig geraten wird. Raten Personen bei allen Aufgaben, so werden jene belohnt, die zufällig stärker gewichtete Aufgaben richtig erraten haben.

152 Bei einer Option wie „alle Antworten sind richtig“ oder „Antworten a und d sind richtig“ sind logischerweise alle bzw. mehrere Antworten anzukreuzen, was in der Regel (bei Best-Choice-Fragen zwingend) der Frageformulierung widerspricht. Optionen, die alle anderen Optionen negieren (z. B. „keine der genannten Antworten ist richtig“), taugen nur dann als Falschantwort, wenn die anzuwählende Antwort fachlich absolut korrekt ist – ansonsten wählen Personen mit hoher Fachkenntnis diese Option an, wodurch die Frage keine zuverlässige Unterscheidung zwischen Personen mit hoher und tiefer Sachkenntnis erlaubt. Eine „wasserdichte“ Formulierung der richtigen Antwortoption führt aber meist dazu, dass die Frage sehr einfach zu beantworten ist (vgl. Krebs 2004, 18).

Zudem muss den Prüflingen bei jeder Aufgabe klar sein, wie viele Antwortoptionen auszuwählen sind. Ansonsten ist eine wesentliche Entscheidungsgrundlage unklar. Eine sinnvolle Interpretation der nicht erwünschten Antworten, die nur einen Teil der erwünschten Optionen enthalten, wird dadurch unmöglich.

Statistische Rechenverfahren ermöglichen zudem, die Qualität der einzelnen MC-Aufgaben („Items“) zu berechnen. Bei dieser Itemanalyse sind insbesondere die Schwierigkeit und die Trennschärfe der Aufgaben zu berücksichtigen. Mit Trennschärfe ist gemeint, wie gut ein Item mit dem Gesamtergebn der Prüflinge korreliert. Je höher die Trennschärfe eines Items ausfällt, desto höher ist die Wahrscheinlichkeit, dass seine korrekte Beantwortung mit einem guten Abschneiden in der gesamten Prüfung einhergeht. Eine sehr hohe Aufgabenschwierigkeit und tiefe (v. a. negative) Trennschärfen weisen auf Probleme in der Itemqualität hin. Krebs (2019, 122) empfiehlt deshalb: „In den rechtlichen Grundlagen zur Prüfung sollte festgelegt sein, dass Items, die einen offensichtlichen inhaltlichen oder formalen Fehler enthalten, aus der Bewertung ausgeschlossen werden und als nicht gestellt gelten. Die Itemanalyse macht auf Items aufmerksam, die mit erhöhter Wahrscheinlichkeit einen solchen Mangel aufweisen (...)“¹⁵³

13.6 MC-Tests als formative Selbsttests

Formative Selbsttests können den Studierenden dabei helfen, ihren derzeitigen Lern- bzw. Leistungsstand einzuschätzen und ihr weiteres Lernen anzupassen. Wenn eine abschließende Multiple-Choice-Prüfung ansteht, können Selbsttests dazu beitragen, dass die Studierenden im Voraus mit den Regeln und dem Stil der Prüfungsfragen vertraut werden. Auf diese Weise können für alle Prüfungsteilnehmenden ähnliche Bedingungen geschaffen und Unterschiede in Bezug auf die Erfahrung mit Tests minimiert werden (vgl. Krebs 2019, 19).

153 Diese Empfehlung des Schweizer Experten Krebs widerspricht möglicherweise den in Deutschland herrschenden Rechtsgrundsätzen, die die Festlegung von Bestehensgrenzen für Multiple-Choice-Prüfungen im Rahmen der Prüfungsordnungen vorsehen (vgl. Hirdina 2019, 53f.). Aus wissenschaftlicher Sicht scheint es mir allerdings nicht angemessen, auf die Möglichkeiten der Itemanalyse zu verzichten. Man vergibt damit die Möglichkeit, MC-Prüfungen reliabler und insofern auch gerechter zu machen. Deshalb lohnt sich meines Erachtens eine genaue Abklärung der Bedingungen, unter denen die Itemanalyse für summative Beurteilungen verwendet werden kann.

Solche Selbsttests sind heute insbesondere mit elektronischen Mitteln einfach realisierbar, etwa in gängigen Learning-Management-Systemen oder durch spezialisierte Software. Aus Gründen der Transparenz sollten Selbsttests den Studierenden im Minimum zurückmelden, welche Aufgaben sie richtig/falsch gelöst haben. Im Idealfall gibt eine Auswertung den Studierenden auch Hinweise darauf, in Bezug auf welche Lernziele sie ihr Wissen und Können noch vertiefen können. Allerdings hat allein das Lösen von Testaufgaben einen erheblichen Lerneffekt, da das nochmalige Abrufen des zu Erlernenden anhand von Testfragen die Gedächtnisspuren stärker vertieft als nochmaliges Lesen von Lernmaterialien (vgl. Roediger u. Karpicke 2006; siehe Kap. 4.2). Deshalb ist es für die Verwendung von formativen MC-Selbsttests auch nicht maßgeblich, ob die Studierenden bei späteren summarischen Beurteilungen MC-Aufgaben begegnen werden oder nicht.

14 Bildungsgeschichtliche Vorläufer von Leistung als schulischem Leitkonzept

Wie eingangs von Kapitel 2.1 erläutert wurde, ist der Leistungsbegriff im Bildungskontext noch nicht so lange etabliert, wie man aufgrund seiner heute so prägenden Wirkung vermuten könnte. Interessant ist deshalb zunächst ein Blick in die allgemeine Begriffsgeschichte (Kap. 14.1), um dann eine bildungshistorische Rückschau vorzunehmen (Kap. 14.2 und 14.3).

14.1 Begriffsgeschichte

Das Verb „leisten“ lässt sich auf die vor ungefähr 3000–4000 Jahren existierende indoeuropäische Sprachwurzel **leis-* zurückführen, die vermutlich „Fuß“ oder „Spur“ bedeutete.¹⁵⁴ Im Mittelhochdeutschen (ca. 1050–1350 n. Chr.) hatte „leisten“ eine abstrakte Bedeutung angenommen und wurde vor allem im rechtlichen Kontext verwendet. So wurde etwa eine Pflicht geleistet (vgl. Schlie 1988, 50–52). Bis ins 19. Jahrhundert war diese Bedeutung vorherrschend. Erst in dessen zweiter Hälfte weitete sich das Bedeutungsspektrum von „leisten“ und „Leistung“ zunehmend aus, um dann die vier Bedeutungselemente zu umfassen, die in Kapitel 2.1 erläutert wurden (vgl. Schlie 1988, 50–52; Verheyen 2018, 130f.): Leisten ist zielgerichtetes Handeln, bedingt Anstrengung, zielt auf ein Ergebnis und kann beurteilt werden (vgl. Schlie 1988, 63f.). Dieses Bedeutungsspektrum ist es erst, dass ein Buch wie das vorliegende als relevant erscheinen lässt.

14.2 Ehre und Ehrtrieb als Leitkonzept der Jesuitenschulen

Dass Leistung nicht immer so stark mit Fragen der Bildung gekoppelt war, zeigt ein Blick in die Geschichte des Schulwesens. Dabei dürfen wir freilich das heutige Verständnis von Leistung nicht unreflektiert auf frühere Zeiten

154 Fun Fact: Auch die neuhochdeutschen Verben „lehren“ und „lernen“ gehen auf diese Wurzel zurück – leisten und lernen sind also sprachgeschichtlich eng miteinander verwandt.

übertragen (vgl. Reh, Berdelmann u. Scholz 2015, 37). Zum Beispiel entwickelten die Jesuitenschulen im 17. Jahrhundert ein ausgeklügeltes Prüfungssystem: Nicht nur wurden fast täglich Wettbewerbe und Prüfungen durchgeführt und in den Schulklassen Rangordnungen gebildet; am Jahresende gab es auch pompöse öffentliche Preisverleihungen. Dies scheint auf den ersten Blick gut zu heutigen kompetitiven Bewertungspraktiken etwa an Gymnasien zu passen, und so werden gelegentlich die Jesuiten als Erfinder der modernen schulischen Leistungsbewertung dargestellt (vgl. Keck 1991). Allerdings wurde der Begriff „Leistung“ damals noch nicht verwendet, und als Leitkonzept fungierte vielmehr die „Ehre“. Indem die Schüler Ehre gewinnen oder verlieren konnten, sollten sie zu möglichst intensivem Lernen motiviert werden. Dies lag auch daran, dass den Auszeichnungen der Jesuitenschulen der heute für Schulen so zentrale Berechtigungsaspekt fehlte – sie waren außerhalb der Schule praktisch bedeutungslos. Beispielsweise bestimmten die Schüler oder ihre Familien selbst, wann sie aus der Schule an die Universität übertraten (Reh, Berdelmann u. Scholz 2015, 39–41). Gymnasien und Universitäten wurden damals nur von (männlichen) Sprösslingen adliger und wohlhabender bürgerlicher Familien besucht. Die gesellschaftliche Stellung war in der bis Ende des 18. Jahrhunderts dauernden Feudalzeit weit entscheidender für den Besuch von Gymnasien und Universitäten als der Lernfortschritt der Schüler.

14.3 Meritentafeln und der Übergang zu Leistung als Leitkonzept

An der Wettbewerbsorientierung der Jesuitenschulen übten Vertreter der philanthropischen Pädagogik im 18. Jahrhundert deutliche Kritik. So schrieb Joachim Heinrich Campe (1998 [1788], 50), durch die Konkurrenz würden schädliche Gefühle genährt – „Eitelkeit (...) in dem Herzen der Sieger; Neid, Unwille, Zorn und Rachsucht in denen, welche sich herabgesetzt“ fühlen. Auch die von den Philanthropen entwickelten Meritentafeln setzten allerdings am Konzept der Ehre an: Auf diesen im Schuleingang aufgehängten Tafeln wurden für jeden Schüler nach genau definierten Kriterien weiße Nägel für fleißiges und schwarze für faules Verhalten eingeschlagen. Und da die Meritentafeln für alle Lehrer und Schüler sichtbar waren, fand auch hier durchaus ein Vergleich statt (vgl. Reh, Berdelmann u. Scholz 2015, 42–46).¹⁵⁵ Die quantifizierte Messung des Fleißes erweist sich dabei als Vorläu-

155 Man kann dies aus heutiger Perspektive durchaus als frühen Schritt in Richtung der Datafizierung von Lernverhalten interpretieren (vgl. Jornitz u. Macgilchrist 2021, 100).

fer moderner Beurteilungspraktiken: „Graduierbarkeit und Messbarkeit von Leistung, d. h. deren Differenzierbarkeit, gehören zusammen. Das aber rückt die Beobachtung stärker ab (...) von der Beurteilung moralischen Verhaltens und stellte mehr und mehr auf den Erwerb von Kenntnissen (...) ab“ (Reh, Berdelmann u. Scholz 2015, 45). In diesem Prozess entstanden schließlich die Schulnoten, wie wir sie heute kennen. Dabei rückten zunehmend die fachlichen Kenntnisse der Schülerinnen und Schüler in den Fokus. Demgegenüber wurden Aspekte des Betragens zwar immer noch gewertet, aber von den fachlichen Aspekten getrennt (vgl. Ziegenspeck 1999, 73–80).

Der Übergang war auch geprägt durch einen Leitkonzeptwechsel: Standen im 17. und 18. Jahrhundert noch Ehre und Ehrtrieb im Fokus der pädagogischen Beurteilungspraktiken, so rückte im Lauf des 19. Jahrhunderts zunehmend die auf das fachliche Lernen bezogene „Leistung“ in den Fokus. Diese Entwicklung erfolgte zeitlich parallel zum oben beschriebenen Bedeutungswandel des Leistungsbegriffs und führte bis zur Mitte des 20. Jahrhunderts schrittweise zur Entstehung der heute für uns so selbstverständlichen Notensysteme. Dass es sich bei diesen um willkürlich gesetzte Ordnungen handelt, wird allerdings schnell sichtbar, wenn man die Notensysteme verschiedener Länder vergleicht: Beispielsweise variiert der Anteil der ungenügenden Noten am gesamten Notenband allein in Deutschland, Österreich und der Schweiz zwischen 25 % im österreichischen und 60 % im Deutschschweizer System (siehe Fußnote 2 in Kapitel 2.1). Und während es im deutschen Sprachraum selbstverständlich ist, dass es sich bei Noten um Zahlen handelt, werden in den USA, Kanada und teilweise auch Großbritannien und Australien die Noten durch Buchstaben repräsentiert.¹⁵⁶

156 Die verschiedenen Notensysteme bringen auch unterschiedliche Schwierigkeiten hervor: Die Repräsentation von Noten durch Zahlen begünstigt z. B. mathematische Fehlinterpretationen, wie in Kap. 9.4.1 dargelegt wird. Demgegenüber liegt hinter den Buchstabennoten meist ein Rechnen in Prozentzahlen, das ebenfalls nicht unproblematisch ist, siehe ebd.

15 Künstliche Intelligenz und menschliche Leistung

Das Aufkommen künstlicher Intelligenz (KI), insbesondere von so genannten Large Language Models (LLMs), bietet neue Perspektiven für die Leistungsbeurteilung an Hochschulen, die durch eine erhöhte Objektivität und Effizienz in der Beurteilung charakterisiert sind. Diese technologischen Fortschritte bieten das Potential, individuelles Feedback in bisher unerreichtem Umfang und mit einer Präzision zu liefern, die das Lernen und die persönliche Entwicklung der Studierenden signifikant fördern kann. Gleichzeitig bergen sie jedoch auch Risiken, wie den Verlust der menschlichen Interaktion in der Beurteilung und potenzielle Verzerrungen durch algorithmische Entscheidungen, welche die Fairness und Akzeptanz der Beurteilungsprozesse gefährden könnten.

Die neuen Perspektiven beruhen auf dem Umstand, dass KI bereits heute in der Lage ist, sprachliche Äußerungen und kulturelle Artefakte (z. B. Bilder) zu produzieren, die nicht mehr von solchen zu unterscheiden sind, die von Menschen hervorgebracht wurden. So wurde der vorherige Abschnitt (von „neue Perspektiven“ bis „gefährden könnten“) verbatim durch ChatGPT formuliert.¹⁵⁷ Er zeigt sehr gut einige positive Eigenschaften von durch LLMs formulierten Texten auf: Der Abschnitt klingt inhaltlich sinnvoll und unterscheidet sich in der Art der Formulierung nicht erkennbar von menschengemachter Sprache. Man bezeichnet solche Formen von KI, die anhand von Vorgaben („Prompts“) basierend auf vorhandenen Informationen neue, sinnvolle Inhalte generieren, als „generative KI“.

Zugleich sind die marktüblichen LLMs wie ChatGPT oder Gemini (früher Google Bard) derzeit nicht in der Lage, wirklich innovative Texte zu schreiben – vor allem nicht solche, die mehr als einige Abschnitte umfassen und gut in einen menschlichen Kommunikationskontext eingepasst sind (zu Stärken und Schwächen von durch LLMs produzierten Texten vgl. Wampfler 2021 und Ray 2023, 134–42; zu den grundsätzlichen Grenzen der Technisierung siehe Kap. 15.2).

157 Der Prompt dazu lautete: „Ich schreibe ein Kapitel über Leistungsbeurteilung und Leistungsrückmeldung an Hochschulen im Zeitalter von KI (insbesondere, aber nicht nur Large Language Models). Kannst du bitte drei Sätze für eine Einleitung formulieren, die in meinem Schreibstil gehalten sind? Diese drei Sätze sollen die zwei am häufigsten genannten Chancen sowie die zwei am häufigsten genannten Risiken im Zusammenhang von Leistungsbeurteilung im Zeitalter von KI enthalten.“ Produziert wurde das Zitat von ChatGPT 4 (kostenpflichtige Version) am 19.12.2023.

Der Umstand, dass LLMs bereits heute in der Lage sind, qualitativ hochwertige Texte von mittlerer Länge (grobe Schätzung: je nach Qualität und Ausführlichkeit des Prompts von einigen Abschnitten bis zu mehreren A4-Seiten) zu verfassen, bietet für die Beurteilung und das Rückmelden von Lernleistungen in der Tat neue Möglichkeiten, wirft aber im gleichen Zug auch große Fragen auf. In den Massenmedien werden demgegenüber vor allem Risiken thematisiert (siehe auch Kap. 8.4.5), was zu Schlagzeilen führt wie „ChatGPT bei Prüfungen im Studium: Eine Studentin packt aus“ (Anonym, 26. März 2023), „KI darf nicht das Denken von Studierenden ersetzen“ (Gugler, 17. April 2023) oder „Das Ende der Hausaufgaben?“ (Röhl, 12. April 2023).¹⁵⁸

In diesen kulturpessimistischen Chor will ich nicht ein-, sondern eher diesem provokativen Titel zustimmen: „ChatGPT zeigt uns, wie dumm unsere Prüfungen sind“ (Zachrau, 17. April 2023).¹⁵⁹ Damit will ich sagen: Prüfungen, die bestanden werden können, indem man die Fragestellung durch ChatGPT beantworten lässt, dürften in der Regel den Ansprüchen nicht genügen, die in diesem Buch insbesondere in den Kapiteln 5 bis 7 formuliert worden sind. Solche Beurteilungsanlässe führen auch ohne die Inanspruchnahme von KI nicht zu einem tiefenorientierten, nachhaltigen Lernen im Sinne von Kapitel 4. Insofern vertrete ich die Meinung, dass LLMs und weitere KI-Tools uns zu überdenken helfen, welche Formen von Leistungsnachweisen zuverlässig tiefenorientiertes Lernen bewirken und welche Arten von Lernleistungen wir künftig beurteilen wollen. Das folgende Panorama bietet deshalb kurze Einblicke in die Interaktion zwischen KI, menschlichen Leistungen und deren Beurteilung.

Zunächst wird die Medien- und Denkgeschichte des ‚Extended Mind‘ beleuchtet, die mit der Schrift als Speichermedium beginnt und sich bis zur aktuellen Kooperation von Mensch und KI erstreckt (Kap. 15.1). Aus letzterer ergibt sich ein Wandel in der Arbeitsteilung zwischen Mensch und Maschine, wobei neue Qualitäten menschlicher Arbeit in den Fokus rücken (Kap. 15.2). Damit einhergehend dürften sich Lernprozesse in Schulen und Hochschulen verändern, wobei eine stärkere Personalisierung des Lernens und eine Rollenveränderung von Lehrenden im Zentrum stehen (Kap. 15.3). Als Folge dieser Verschiebungen in Gesellschaft und Bildungswesen dürfte sich auch die Prüfungskultur in Bildungseinrichtungen ver-

158 Allerdings stimmt die letzte Schlagzeile Bildungsfachleute nicht unbedingt pessimistisch, ist doch die Wirkung von Hausaufgaben insbesondere in der Primarschule im besseren Fall gering und im schlechteren Fall problematisch (bessere Schüler profitieren mehr davon), vgl. Hattie (2009, 234–36).

159 Ähnlich äußert sich auch Wampfler (2022, 4): „Etwas polemisch könnte man festhalten, dass Algorithmen nur Schreibaufgaben unterminieren, die an sich schon problematisch sind.“ Die im Text von Zachrau enthaltene Aufforderung zum „Schummeln“ an Studierende unterstütze ich hingegen in keiner Weise.

ändern (Kap. 15.4). Trotz all der Potenziale ergeben sich aus heutiger Sicht durch den Einsatz von KI im Bildungswesen auch gewichtige ethische Fragen, etwa nach Verzerrungen in der Datenbasis von KI, dem Datenschutz oder urheberrechtlichen Aspekten (Kap. 15.5.).

15.1 Kulturelle Evolution oder der Mensch als Cyborg

Sprache ist eine, wenn nicht die entscheidende Grundlage dafür, weshalb sich die kulturelle Entwicklung des Menschen von jener aller anderen Lebewesen auf der Erde, inklusive jener anderer Primaten, abhebt. Dass Schriftsprache eine Erweiterung des menschlichen Gedächtnisses darstellt, ging schon bei den alten Griechen mit Warnungen einher, die uns bekannt vorkommen. So befürchtete Sokrates gemäß Platons Phaidros-Dialog (ca. 400 v.Chr.), dass die Schriftsprache nicht nur das Gedächtnis künftiger Generationen verkümmern lasse, sondern auch die menschliche Fähigkeit zum lebendigen Dialog (vgl. Wolf 2007, 69–78).¹⁶⁰

Der Einfluss der Schrift auf die kulturelle Entwicklung der letzten wenigen Tausend Jahre ist dabei kaum zu überschätzen. Während viele Tiere in der Lage sind, Verhaltensweisen von ihren Artgenossen zu lernen, wurde nur bei Menschen eine kumulative kulturelle Evolution beobachtet, die Tomasello, Kruger u. Ratner (1993, 495) als „Wagenhebereffekt“ (oder Ratschenefekt) bezeichneten. Besonders verstärkenden Einfluss auf diesen Effekt dürfte die Erfindung der Schrift gehabt haben, da sie die Speicherung und Übertragung von Informationen über Zeit und Raum hinweg ermöglicht, und zwar über die Grenzen der direkten mündlichen Kommunikation hinaus. Dies gilt von Tontafeln und Papyrusrollen über Tinte und Papier bis zu Schreibmaschine und Computer. Clark (1997, 218, Übers. tz) bezeichnet Sprache deshalb als „das ultimative Artefakt“.

Evolutionsgeschichtlich gesehen ist die Schriftsprache, gemeinsam mit weiteren kognitiven Artefakten wie komplexen Zahlen- und Rechensystemen, Karten oder Messmethoden und -geräten eine sehr neue Erscheinung (vgl. Heersmink 2022). Das Besondere am menschlichen Verhalten ist dabei nicht die Verwendung von Werkzeugen oder Artefakten an sich, sondern die flexible und komplexe Art, wie sie verwendet werden: „Wir sind von Natur aus Cyborgs, also Wesen, die auf natürliche Weise hybride Systeme mit einer Vielzahl komplexer technologischer Artefakte bilden, die tief mit unseren motorischen, wahrnehmenden, kognitiven und affektiven Systemen verwachsen sind“ (Heersmink 2022). So betrachtet, nutzen Menschen schon seit lan-

160 Sokrates' Bedenken sind freilich nur dank Platons Verschriftlichung überliefert, was seinen Argumenten eine ironische Note verleiht.

ger Zeit – seit einigen Tausend Jahren zunehmend intensiver – Werkzeuge und Artefakte außerhalb des eigenen Hirns für ihr Denken. Und diese Werkzeuge und Artefakte wurden im gleichen Zeitraum immer komplexer: Zunächst wurden vermutlich Finger zum Zählen genutzt oder Zahlen mit Strichen symbolisiert, dann entwickelten sich komplexe und zunehmend automatisierte Schrift- und Bildsysteme (vom Buchdruck oder der Fotografie bis zur elektronischen Datenverarbeitung).¹⁶¹ Diese Entwicklung dauert immer noch an. Ihr neuester Schritt besteht in der Nutzung generativer KI wie LLMs oder Bildgeneratoren – und Diskussionen über die mit ihrer Verbreitung verbundenen kulturpessimistischen Befürchtungen.

15.2 Neue Arbeitsteilung zwischen Mensch und Maschine

Wie Böhle, Huchler u. Neumer (2019, 32–34) feststellen, spielt Digitalisierung vor allem in Bereichen eine große Rolle, in denen objektivierende Handlungen erforderlich sind. Durch die Möglichkeiten generativer KI betrifft dies nicht mehr nur mechanische Bereiche und vergleichsweise einfache geistige Tätigkeiten, die sich durch ein starres Regelset erfassen lassen. Vielmehr sind nun auch Gebiete betroffen, die bisher als höherwertige geistige Arbeiten galten, wie juristische Recherche oder die Auswertung von Röntgenbildern. Das Kriterium der Digitalisierbarkeit ist nun vielmehr „die Möglichkeit der Formalisierung und Formalisierbarkeit von Informationen und Verfahren, so wie dies grundsätzlich bei Rechenoperationen der Fall ist. Auf dieser Basis ist es möglich, auch höherwertige geistige sowie dynamische und flexible Prozesse zu technisieren“ (Böhle, Huchler u. Neumer 2019, 33). Die Grenzen der Technisierung verlaufen damit nicht mehr entlang der Unterscheidung von körperlichen und geistigen Tätigkeiten oder von einfachen und komplexen Denkopoperationen. Vielmehr zeigen sie sich nun bei der Bewältigung von Unwägbarkeiten im Arbeitsprozess, die nur durch subjektivie-

161 Menschen benutzen allerdings nicht nur „klassische“ Kulturtechniken, um ihr Denken über ihr Hirn hinaus zu erweitern (vgl. Paul 2021): Körperliche Bewegungen und Wahrnehmungen gehören ebenso dazu wie räumliche Anordnungen, etwa wenn wir unsere Gedanken oder Äußerungen mit Gesten körperlich untermalen oder wenn wir architektonische Entwürfe als Modelle räumlich erfahrbar machen. Sogar die soziale Dimension nutzen Menschen systematisch, um ihr Denken zu erweitern: Wie in Kap. 4.7 gezeigt wurde, dürfte das argumentative Schlussfolgern als sozialer Prozess und nicht als individuelle Fähigkeit entwickelt worden sein.

rendes Handeln sinnvoll zu bewältigen sind.¹⁶² Beispiele dafür reichen von Montagetarbeiten (z. B. montieren von Möbeln in Wohnungen von Kunden) über den Umgang mit Qualitätsschwankungen in verwendeten Materialien bis zur Abstimmung von Prozessen mit Kundinnen oder Patienten (vgl. Böhle, Huchler u. Neumer 2019, 33). Solche Unwägbarkeiten treten in den meisten Arbeitsprozessen irgendwo auf, so dass es wenig wahrscheinlich erscheint, dass bestimmte Arbeiten in absehbarer Zeit komplett digitalisiert werden können, während es umgekehrt auch kaum Tätigkeiten gibt, von denen nicht wenigstens Teilbereiche digitalisierbar sein dürften (vgl. ebd.). „An die Stelle der Ersetzung menschlicher Arbeit durch Technik oder der Beherrschung und Steuerung der Technik durch den Menschen tritt demnach die ‚Kooperation‘ zwischen Mensch und Technik“ (ebd. S. 34).

Die Arbeitsteilung zwischen Mensch und Maschine dürfte sich also durch die Verbreitung generativer KI insofern verändern, dass die Teilleistungen von Menschen und Maschinen innerhalb komplexer Arbeitsvorgänge noch stärker als bisher miteinander verwoben werden – ein Prozess, der bereits in vollem Gange ist. Menschen werden dabei auch anspruchsvolle (aber objektivierbare) Denkleistungen zunehmend an Maschinen auslagern, so dass komplexe Denkleistungen künftig verstärkt ein Kooperationsprodukt von Mensch und Maschine (bzw. KI) sein dürften.

Wie man diese Entwicklung einschätzt, ist eine normative Frage. So gibt es sowohl eher optimistische theoretische Einschätzungen der geschilderten Prozesse wie jene des Informationsethikers Floridi (2019), während z. B. der Physiker Stephen Hawking vor der Entwicklung einer vollständigen künstlichen Intelligenz warnte, da sie das Ende der Menschheit bedeuten könnte (Cellan-Jones, 2. Dezember 2014). In jedem Fall stimme ich Döbeli (2023, 2) zu, dass auf die gesellschaftliche und wirtschaftliche Bedeutung von generativer KI vermutlich das sogenannte Gesetz von Amara zutrifft. Dieses besagt, dass Menschen dazu neigen, die Wirkung einer neuen Technologie auf kurze Sicht zu überschätzen, auf lange Sicht aber zu unterschätzen.

Für Hochschulen dürfte beispielsweise relevant sein, dass Menschen künftig wissenschaftliche Texte mehr oder weniger stark „in Zusammenarbeit“ mit generativer KI verfassen werden. Das gilt mitnichten nur für Stu-

162 Der von Böhle (2017) geprägte Begriff des subjektivierenden Handelns betont, dass Menschen in aller Regel nicht streng regelbasiert handeln, sondern gleichsam in einer Art holistischem Dialog mit situativen Merkmalen, in die auch die sinnliche Wahrnehmung stark einfließt. Dabei bringen sie subjektive Fähigkeiten wie Gefühle, Empathie, implizites Wissen oder das Bilden von Assoziationen zwischen verschiedenen Sinnesmodalitäten ein, um mit Unwägbarkeiten, Ungewissheiten und komplexen Situationen umzugehen. Im Gegensatz zum objektivierenden Handeln betont der Begriff des subjektivierenden Handelns die Bedeutung des menschlichen Urteilsvermögens (vgl. Böhle, Huchler u. Neumer 2019, 31).

dierende, die unter Zeitdruck wissenschaftliche Arbeiten verfassen müssen, sondern auch für arrivierte Wissenschaftlerinnen und das Verfassen hochwertiger wissenschaftlicher Publikationen. Der menschliche Schreibprozess dürfte sich dadurch erheblich verändern, was auch erkenntnistheoretische Auswirkungen haben wird. So dürften Abwägungen stark an Bedeutung gewinnen, „was warum wahr ist und wie wir sinnvoll vorgehen, um begründbare und ausreichend sichere Antworten auf diese Frage zu finden“ (Philipp 2023b). Damit wird die Fähigkeit zu begründeten Qualitätsurteilen („evaluative judgement“), also auf der Grundlage festgelegter Kriterien fundierte Entscheidungen über die Qualität einer Arbeit oder einer Aussage zu treffen, noch bedeutsamer. Sie beinhaltet ein komplexes Zusammenspiel von kognitiven, metakognitiven und affektiven Prozessen (vgl. Tai et al. 2018), das gemäß dem heutigen Forschungsstand nicht objektivierbar sein dürfte (vgl. Böhle, Huchler u. Neumer 2019, 30–32).

15.3 Veränderungen im institutionell organisierten Lehren und Lernen

Der Einsatz von LLMs in Schulen, Hochschulen und Weiterbildungseinrichtungen führt auch zu Veränderungen in institutionell organisierten Lehr- und Lernprozessen. Betroffen sind insbesondere solche Prozesse, in denen epistemisches Schreiben eine wesentliche Rolle spielt. Mit diesem Begriff wird eine Verwendung des Schreibens bezeichnet, in deren Rahmen bestehende mit neuen Wissenselementen kombiniert werden, wobei in der Regel Lese-, Lern- und Schreibprozess durch transformative und evaluative Prozesse miteinander verbunden werden (vgl. Philipp 2023a, 6). Da kompetente Schreibende in epistemische Schreibprozesse nun die Möglichkeiten von LLMs einbeziehen, erfolgt auch hier eine stärkere Integration menschlicher und maschineller Möglichkeiten: „Die Tools werden wie eine Stahlfeder, ein Radiergummi oder ein Textverarbeitungsprogramm selbstverständlicher Bestandteil von Schreibprozessen (...). Schreiben ist und war nie technologiefrei, sondern ist ein Prozess, der verfügbare Technologie einbezieht“ (Wampfler 2022, 4).

Aber auch über das epistemische Schreiben hinaus bietet das Einbinden von KI-Modellen (die neben den Möglichkeiten von LLMs auch jene von Bildgeneratoren, mathematisch ausgerichteten Modellen etc. integrieren) diverse weitere Möglichkeiten, Lerninhalte und Lernprozesse verstärkt auf die individuellen Bedürfnisse und Fähigkeiten der Lernenden zuzuschneiden. Zu diesen Personalisierungsmöglichkeiten zählen beispielsweise die folgenden, die für Lernen an Hochschulen relevant scheinen:

- Individuelle Lernpfade: Generative KI kann Lerninhalte basierend auf dem Kenntnisstand und den Interessen der Lernenden anpassen, was individuellere und effektivere Lernwege begünstigt.
- Personalisierte Übung inklusive Feedback auf der Aufgabenebene: KI kann Übungen und Aufgaben generieren, die auf die spezifischen Fähigkeiten und Lernfortschritte der Lernenden zugeschnitten sind, und ein auf die individuellen Lösungen abgestimmtes Feedback geben.
- Adaptives Lernen: Die beiden oben genannten Punkte lassen sich kombinieren, so dass KI-Systeme Lerninhalte und Übungen in Echtzeit anpassen, um Herausforderungen oder Wissenslücken gezielt anzugehen. KI-Systeme können dadurch auch als Tutor oder Lernbegleitung eingesetzt werden (dies kann auch selbstgesteuert durch Lernende erfolgen, siehe Kap. 8.4.5).
- Sprachunterstützung: LLMs können in verschiedenen Sprachen interagieren und damit mehrsprachiges Lernen oder Sprachunterstützung für nicht-muttersprachliche Lernende bieten.

Da die Aktivitäten von Lehrenden und Lernenden und ihre Nutzung von Lernmaterialien inklusive Lerntechnologien systematisch zusammenhängen, wird es auch zu Veränderungen im Rollenstrauß von Lehrenden kommen (zu den Rollen von Hochschuldozierenden siehe Kap. 3.3.1). Wenn KI-basierte Systeme die Aufbereitung und Vermittlung von Lerninhalten übernehmen und bis zu einem gewissen Grad sogar die individuellen Aneignungsprozesse der Lernenden durch personalisierte Feedbackprozesse steuern, so werden Lehrende in diesem Bereich entlastet und können sich stärker auf die Kontextualisierung der fachlichen Inhalte konzentrieren. In diesem Bereich ist allerdings zu erwarten, dass den KI-basierten Systemen noch lange deutliche Grenzen gesetzt sein werden, da ihnen die Möglichkeiten der Subjektivierung fehlen. Dies betrifft nicht nur psychologische und soziale Aspekte wie die Perspektivenübernahme oder Empathie im Umgang mit Lernenden. Auch fachliche und arbeitspraktische Aspekte sind davon betroffen, etwa die ganzheitliche Einschätzung von komplexen Situationen. KI-Systeme könnten also Lehrenden künftig die Möglichkeit bieten, im Umgang mit den Lernenden stärker auf psychologische Aspekte des Lernprozesses, den Aufbau von fachbezogenen Fähigkeiten zum Fällen begründeter Qualitätsurteile sowie die Anwendung von Gelerntem in komplexen Anwendungssituationen zu fokussieren.

In einer Rückkopplungsschleife gehört dazu dann auch die Anwendung von generativer KI zu Lern-, Forschungs- und Arbeitszwecken. Hier dürfte das „Prompting“ eine wichtige Rolle spielen, also die Fähigkeit, präzise und effektive Eingabeaufforderungen an generative KI zu formulieren, um eine spezifische Antwort oder einen bestimmten Output zu generieren. Das

Prompting bestimmt letztlich, wie nützlich die Antworten/Ausgaben der KI-Modelle sind. Während gewisse Aspekte von Prompting allgemein gültig sein dürften, ist insbesondere für die Tertiärstufe zu erwarten, dass es auch fachspezifische Aspekte des Promptings geben wird. Zudem hängt es von der Programmierung des jeweiligen Modells ab, wie es auf Eingabeaufforderungen reagiert. ChatGPT ist beispielsweise in der Lage, im Rahmen eines Dialogs zunehmend genauer zu verstehen, was die Nutzerin von ihm möchte.

Ein interessanter Systematisierungsversuch, wie im Kontext institutionellen Lernens mit Prozessen umgegangen werden kann, die durch KI ausgeführt werden können, findet sich bei Spannagel (2023a). Mit Blick auf die Herausforderung, dass KI-Tools Lernende zwar beim Lernen unterstützen, ihnen aber das Lernen bzw. Denken nicht abnehmen sollen, unterscheidet er drei Ansätze. Diese sind nicht trennscharf, sondern können auch kombiniert werden:

1. Man betrachtet bestimmte kognitive (oder gestalterische etc.) Prozesse nicht länger als Lernziel, da künstliche Intelligenz-Systeme diese Aufgaben übernehmen können.

Diesen Ansatz erachtet Spannagel in vielen Fällen als problematisch, worin ich ihm zustimme (siehe nächsten Aufzählungspunkt).

2. Man kommt zum Schluss, dass bestimmte kognitive (oder gestalterische etc.) Prozesse weiterhin relevante Fähigkeiten darstellen, die von den Lernenden erlernt werden sollten.

Denn für die Resultate von „Produkten“, welche mithilfe von KI erstellt wurden, müssen weiterhin Menschen verantwortlich sein. Wirklich Verantwortung übernehmen kann man aber in diesem Zusammenhang nur, wenn man die Qualität eines Produkts beurteilen kann. Ein derart begründetes Qualitätsurteil ist lediglich möglich, wenn man über die Kompetenz, die einem von der KI ausführend abgenommen wurde, grundsätzlich selbst verfügt.¹⁶³

Die Herausforderung bei diesem Ansatz besteht darin, die Lernenden davon zu überzeugen, dass der damit verbundene Aufwand lohnenswert ist. Spannagel (ebd.) nennt als mögliches Argument die Selbstbestimmungstheorie (siehe Kap. 4.5.2): Studierende „fühlen sich autonomer, wenn sie selbst Entscheidungen treffen können und diese nicht an ein KI-System abgeben. Wer sich abhängig von Technologie macht, büßt an Autonomie ein.“

3. Eine dritte Option besteht darin, den Erwerb von Fähigkeiten, die durch eine KI ersetzt werden könnten, in Präsenzveranstaltungen zu integrieren.

163 So betont auch Bubenhofer (2022), es müsse „das Ergebnis korrekt eingeschätzt werden können – es muss angepasst, korrigiert und erweitert werden“ (können).

Hier wird das didaktische Konzept des Inverted oder Flipped Classroom zentral, bei dem traditionelle Lernaktivitäten wie Vorlesungen und Informationsvermittlung als Hausaufgabe durchgeführt werden, während die Anwendung des Gelernten und interaktive Übungen in den Präsenztunterricht verlagert werden (vgl. Werner et al. 2018).

15.4 Veränderungen in der Beurteilungskultur

Der verbreitete Einbezug von Large Language Models und anderen generativen KI-Systemen wird auch zu Änderungen im hochschulischen Umgang mit der Beurteilung von Lernleistungen („Prüfungskultur“) führen. Wie oben erläutert, stehen im öffentlichen Diskurs derzeit noch häufig Betrugsbefürchtungen im Vordergrund. Obwohl die Versuchung zum Betrug durch die Verwendung von LLMs zunimmt, ist sie nicht grundsätzlich neu. Sie ähnelt früheren Herausforderungen wie dem Abschreiben von Sitznachbarn, Ghostwriting oder dem Kopieren von Lösungen aus dem Internet.

Gerade angesichts des Umstands, dass Bildungsorganisationen eine sinnvolle Integration von KI-Technologie in die menschlichen Lern-, Arbeits- und Kommunikationsprozesse fördern sollen (siehe Kap. 15.3), liegt eine konstruktive Lösung nicht in verstärkter Überwachung und Bestrafung, sondern in der Entwicklung einer an den Prinzipien der Selbstbestimmung (siehe Kap. 4.5.2) und der Bedeutung von Feedback (siehe Kap. 8) orientierten Lernkultur. In einer solchen Kultur werden Ehrlichkeit, Eigenständigkeit und kritisches Denken gefördert. Lernende sollen verstehen, dass der Einsatz von LLMs als Werkzeug zur Erweiterung ihres Wissens und nicht zur Umgehung des Lernprozesses gedacht ist (vgl. Wampfler 2022, 5f.).

Die Möglichkeiten generativer KI zeigen also die Problematik einer lern- und motivationspsychologisch nicht effektiven Beurteilungskultur deutlich auf und verstärken die Notwendigkeit von Anpassungen. Dabei sind insbesondere Alternativen der summativen Beurteilung wie der Notenkontrakt vielversprechend, die in Kapitel 9.6 vorgestellt wurden. Diese Ansätze führen dazu, dass Studierende – und in der Folge auch Lehrende – den Fokus stärker auf den Lernprozess statt auf das Erreichen von Noten legen. Da der Anreiz entfällt, gute Noten um jeden Preis zu erzielen, wird auch das Betrugsproblem zumindest teilweise gelöst – das eigenständige Lernen und die persönliche Entwicklung der Studierenden rücken in den Vordergrund.

Darüber hinaus zeichnen sich mindestens zwei weitere Tendenzen ab:

- Stärkere Fokussierung auf höherrangige Lernziele: Die Auswirkungen von KI auf die Beurteilung von Lernleistungen umfassen auch eine Verschiebung des Fokus' von der bloßen Wissensabfrage hin zur Anwen-

dung, zum Transfer und zur kritischen Beurteilung von Informationen (womit Betrugsmöglichkeiten durch KI und andere unerlaubte Hilfen unweigerlich abnehmen). Formative und summative Beurteilungen sollten stärker die Fähigkeiten der Studierenden fokussieren, Informationen und Argumente zu analysieren und zu evaluieren. Dies erfordert ein Umdenken in der Art und Weise, wie Leistungsnachweise konzipiert, durchgeführt und beurteilt werden (vgl. Philipp 2023b; Wampfler 2022).

- **Steigende Bedeutung von Prompting:** Die Bedeutung der Fähigkeit, effektive und präzise Anfragen an KI-Systeme zu stellen, wird ebenfalls zunehmen. Diese Kompetenz sollte Teil von Leistungsbeurteilungen und Feedbackprozessen werden, um sicherzustellen, dass Studierende KI nicht nur für die Reproduktion von (unhinterfragtem) Wissen nutzen, sondern sie auch kreativ und verantwortungsvoll in Lern- und Arbeitsprozessen einsetzen können (vgl. Feixas u. Zimmermann 2023).

Insgesamt geht es angesichts der Möglichkeiten von generativer KI im Umgang mit der Beurteilung von Lernleistungen darum, eine Balance zwischen der Nutzung fortschrittlicher Technologien und der Förderung von kritischem Denken, Eigenständigkeit, Kreativität und ethischem Verhalten zu finden. So dürfte es sinnvoll sein, für verschiedene Lehr-, Lern- und Beurteilungssituationen künftig so genannte „Rules for Tools“ festzulegen, wie das Spannagel (2023b) in meines Erachtens gelungener Weise getan hat. Dort hält er etwa fest: „Alle Hilfsmittel haben ihre Grenzen. (...) KI-Sprachmodelle wie ChatGPT können zwar gut formulieren, sie machen aber inhaltliche Fehler und reproduzieren Vorurteile. (...) Das Werkzeug denkt nicht für Sie, sondern Sie denken mit Hilfe des Werkzeugs“ (Spannagel 2023b, 2).

15.5 Offene ethische Fragen

Auch wenn die Tendenz mit hoher Wahrscheinlichkeit in Richtung einer stärkeren Integration von generativer KI in Lern- und Arbeitsprozesse geht, bleiben nicht zuletzt diverse ethische Fragen offen. Diese können hier nicht beantwortet, sondern lediglich umrissen werden:

- **Bias von KI-Systemen:** Da KI-Modelle auf Datensätzen trainiert werden, die historische Vorurteile widerspiegeln können, besteht das Risiko, dass diese Vorurteile in den Output der KI-Modelle einfließen. Dies könnte auch das Lernen auf verschiedenen Ebenen beeinflussen, und wie damit bei der Beurteilung von Lernleistungen umzugehen ist, wird zu klären sein (vgl. Fleischmann 2023, 13–15).

- **Datenschutz:** KI-Systeme benötigen große Mengen an Daten, um effektiv zu funktionieren, und sie müssen aus den Daten der Nutzerinnen lernen können. Es ist noch unklar, wie sichergestellt werden kann, dass die Privatsphäre der Nutzenden gewahrt bleibt und ihre Daten sicher und verantwortungsvoll behandelt werden. Dies betrifft auch die Nutzung von KI-Systemen durch Lehrende und Lernende, sei es zu Lern- oder zu Forschungszwecken (vgl. Fleischmann 2023, 10f.).
- **Urheberrechtliche Aspekte:** Wenn KI-Systeme verwendet werden, um Inhalte zu generieren oder zu modifizieren, ist zu klären, wem die Rechte an von KI erstellten Werken zustehen und wie diese genutzt werden dürfen. Dies betrifft beispielsweise den Umgang mit KI-generierten Texten, Bildern oder Videos, wie sie in wissenschaftlichen Arbeiten, Präsentationen oder Projektarbeiten verwendet werden (vgl. Fleischmann 2023, 9f.).
- **Chancengerechtigkeit:** Der Zugang zu fortschrittlicher KI-Technologie kann ungleich verteilt sein, wobei wohlhabendere und besser gebildete Personen in der Regel besseren Zugang haben und KI effektiver nutzen können. Dies könnte zu einer Verstärkung bestehender Bildungsungleichheiten führen (vgl. Fleischmann 2023, 16f.).
- **Autonomie der Lernenden:** Werden KI-Systeme auch in der Leistungsbeurteilung eingesetzt, stellt sich die Frage nach der Autonomie der Studierenden als Lernende verstärkt (wobei alle vier bisher aufgeführten Punkte zum Tragen kommen). Dies gilt selbst bei formativem Feedback, da auch dieses die Lernwege und -methoden beeinflusst, welche Studierende in ihrem weiteren Lernprozess wählen. Selbstverständlich stellt sich diese Frage aber auch angesichts der Beeinflussung von Lernenden durch Interaktionen mit Menschen.

Die Integration von KI in die Leistungsbeurteilung erfordert eine gründliche Auseinandersetzung mit diesen und weiteren ethischen Fragen, um sicherzustellen, dass die Technologie fair, transparent und im Einklang mit den grundlegenden gesellschaftlichen Werten eingesetzt wird.

16 Auflösung und Erläuterungen zu den Aufgaben in Kapitel 4.1.1

Nachfolgend finden Sie die Auflösungen zu den beiden in Kapitel 4.1 enthaltenen Lernexperimenten.

16.1 Auflösung Aufgabe 1b

Tatsächlich enthalten waren in den sechs Listen aus Kapitel 4.1.1, Aufgabe 1a die folgenden zwölf Wörter: fein, Bettdecke, schnuckelig, Sitzbank, Bonbon, stechen, Gipfel, Spritze, Holz, steil, Sand, ausruhen.

Schauen Sie nun bitte nochmals Ihre Einschätzungen an. Vermutlich haben Sie bei einem oder mehreren der Wörter grob, Nadel, Berg, schlafen, Stuhl und süß eine hohe Sicherheit angegeben, dass Sie das Wort in einer der Listen gesehen haben. Dies ist aber nicht der Fall. Die Listen waren vielmehr so konstruiert, dass sie jeweils lauter Wörter enthielten, die deutlich mit den genannten Wörtern assoziiert sind. Somit waren die Konzepte „Stuhl“, „grob“ usw. in Ihrem Vorstellungsvermögen präsent, ohne dass Sie sich dies bewusst waren. Wie sich diese „Fehlleistung“ – die eben eigentlich keine ist – mit der Funktionsweise des menschlichen Lernens erklären lässt, wird in Kapitel 4.2 erklärt.

16.2 Auflösung Aufgabe 2b

Bei den 23 Sätzen aus Aufgabe 2a handelt es sich um Teilmengen der folgenden vollständigen Aussagen – mit Ausnahme dreier Sätze, die der Abwechslung dienen. Diese vier komplexen Sätze bestehen aus jeweils vier einfachen Aussagen:

- Die Ameisen in der Küche aßen die süße Marmelade, die auf dem Tisch stand.
- Die warme Brise, die vom Meer her wehte, bewegte die schwere Abendluft.
- Der Felsen, der den Berg hinunterrollte, zerstörte die kleine Hütte am Rande des Waldes.

- Der alte Mann, der sich auf der Couch ausruhte, las die Geschichte in der Zeitung.

Die vier einfachen Aussagen für den ersten Satz würden zum Beispiel lauten:

- Die Ameisen waren in der Küche.
- Die Marmelade war auf dem Tisch.
- Die Marmelade war süß.
- Die Ameisen aßen die Marmelade.

Die Aufgabenstellung 2b lautete, für jeden der dort aufgeführten acht Sätze eine Einschätzung abzugeben, ob er in diesem Wortlaut in Aufgabe 2a enthalten war. Nun wollen Sie sicher wissen, wie viele es tatsächlich waren: Kein einziger. Vielmehr handelt es sich bei den Sätzen in Aufgabe 2b um andere Teilmengen der oben aufgeführten vollständigen Aussagen, von denen keine wörtlich übereinstimmt mit den Sätzen aus 2a.

Sollten Sie einen oder mehrere der Sätze aus Aufgabe 2b als identisch mit Sätzen aus Aufgabe 2a identifiziert haben, ist allerdings auch dies keine Fehlleistung, sondern zeigt, wie sehr unser Hirn auf sinnvolles Lernen hin organisiert ist. Wir merken uns nämlich die Kernideen und nicht den genauen Wortlaut der Sätze. Aus diesem Grund haben Sie vermutlich für die letzten vier Sätze in 2b eine höhere Sicherheit angegeben, sie zuvor bereits gelesen zu haben. Dies liegt daran, dass diese Sätze jeweils drei oder gar alle Elemente der vollständigen Aussage enthalten und somit mehr Erinnerungen auslösen als die einfacheren Sätze in der ersten Hälfte von 2b. Eine genauere Erklärung zu den zugrundeliegenden Mechanismen finden Sie ebenfalls in Kapitel 4.2.

Literatur

- 21st Century Collective. 2013. *Field Notes #21st Century Literacies: A Guide to New Theories, Methods and Practices for Open Peer Teaching and Learning*. Chapel Hill NC.
- Amabile, Teresa M. 1998. „How to Kill Creativity.“ *Harvard Business Review* 76 (5): 76–87. Online verfügbar unter: <https://hbr.org/1998/09/how-to-kill-creativity>.
- Ambrose, Susan A., Michael W. Bridges, Michele DiPietro, Marsha C. Lovett, Marie K. Norman und Richard E. Mayer. 2010. *How Learning Works: Seven Research-Based Principles for Smart Teaching*. 1. Auflage. The Jossey-Bass Higher and Adult Education Series. San Francisco, CA: Jossey-Bass a Wiley Imprint.
- Ambrosius, Gerold. 2018. *Globalisierung*. Wiesbaden: Springer Fachmedien.
- Anderson, John R. 1982. „Acquisition of Cognitive Skill.“ *Psychological Review* 89 (4): 369–406. doi:10.1037/0033-295X.89.4.369.
- Anderson, John R. 1992. „Automaticity and the ACT Theory.“ *The American Journal of Psychology* 105 (2): 165. doi:10.2307/1423026.
- Anderson, Lorin W. und David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Unter Mitarbeit von Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Rath und Merlin C. Wittrock. Gekürzte Ausgabe. New York, München: Longman.
- Andrade, Heidi G. 2000. „Using Rubrics to Promote Thinking and Learning.“ *Educational Leadership* 57 (5): 13–18.
- Andrade, Heidi L. 2019. „A Critical Review of Research on Student Self-Assessment.“ *Frontiers in Education* 4.
- Angermüller, Johannes und Jens Maeße. 2015. „Regieren durch Leistung. Zur Verschulung des Sozialen in der Numerokratie.“ In *Leistung*, hrsg. v. Alfred Schäfer und Christiane Thompson, 60–108. Paderborn: Ferdinand Schöningh.
- Anonym. 2023. „ChatGPT bei Prüfungen im Studium: Eine Studentin packt aus.“ Aufgezeichnet von Sina Alpiger. *Watson*, 26.3.2023. <https://www.watson.ch/schweiz/leben/858406632-chatgpt-bei-pruefungen-im-studium-eine-studentin-packt-aus>. Zugriff 15.2.2024.
- Arcila Hernández, Lina, Nicole Chodkowski und Kira Treibergs. 2022. „A Guide to Implementing Inclusive and Accessible Virtual Poster Sessions.“ *Journal of Microbiology & Biology Education* 23 (1). doi:10.1128/jmbe.00237-21.
- Ausubel, David P. 1960. „The Use of Advance Organizers in the Learning and Retention of Meaningful Verbal Material.“ *Journal of Educational Psychology* 51 (5): 267–72. doi:10.1037/h0046669.
- Bach, Dagmar und Franziska Zellweger. 2022. „Entwicklung berufsorientierter Curricula.“ In *Höhere Fachschulen in der Schweiz: Herausforderungen und Perspektiven*, hrsg. v. Dagmar Bach, Erik Haberzeth und Stefan Osbahr, 70–92. Forum Hochschuldidaktik und Erwachsenenbildung 12. Bern: hep verlag.

- Bachmann, Heinz. 2018a. „Formulieren von Lernergebnissen – learning outcomes.“ In *Kompetenzorientierte Hochschullehre: Die Notwendigkeit von Kohärenz zwischen Lernzielen, Prüfungsformen und Lehr-Lern-Methoden*, hrsg. v. Heinz Bachmann. 3., überarbeitete Auflage, 34–49. Forum Hochschuldidaktik und Erwachsenenbildung 1. Bern: hep.
- Bachmann, Heinz. 2018b. „Hochschullehre neu definiert – shift from teaching to learning.“ In *Kompetenzorientierte Hochschullehre: Die Notwendigkeit von Kohärenz zwischen Lernzielen, Prüfungsformen und Lehr-Lern-Methoden*, hrsg. v. Heinz Bachmann. 3., überarbeitete Auflage, 14–33. Forum Hochschuldidaktik und Erwachsenenbildung 1. Bern: hep.
- Bandura, Albert. 1977. „Self-Efficacy: Toward a Unifying Theory of Behavioral Change.“ *Psychological Review* 84 (2): 191–215. doi:10.1037//0033-295x.84.2.191.
- Barnlund, Dean C. 1970. „A Transactional Model of Communication.“ In *Foundations of Communication Theory*, hrsg. v. Kenneth K. Sereno und C. D. Mortensen, 83–102. New York: Harper & Row.
- Barrett, Lisa F. 2017. *How Emotions Are Made: The Secret Life of the Brain*. London: MacMillan.
- Barrett, Lisa F. und W. K. Simmons. 2015. „Interoceptive Predictions in the Brain.“ *Nature Reviews. Neuroscience* 16 (7): 419–29. doi:10.1038/nrn3950.
- Bartlett, Frederic C. 1932. *Remembering: A Study in Experimental and Social Psychology*. Cambridge, MA: Cambridge University Press.
- Baurmann, Jürgen. 1977. „Untersuchungen zur Bedeutung von Vorinformationen und Reihenfolgeeffekten auf die Aufsatzbeurteilung.“ In *Die Fragwürdigkeit der Zensur Umgebung: Texte und Untersuchungsberichte*, hrsg. v. Karlheinz Ingenkamp. 7. Auflage, 117–30. Weinheim, Basel: Beltz.
- Beck, Ulrich. 1988. *Gegengifte: Die organisierte Unverantwortlichkeit*. Frankfurt a. M.: Suhrkamp.
- Becker, Rolf und Andreas Hadjar. 2009. „Meritokratie – Zur gesellschaftlichen Legitimation ungleicher Bildungs-, Erwerbs- und Einkommenschancen in modernen Gesellschaften.“ In *Lehrbuch der Bildungssoziologie*, hrsg. v. Rolf Becker, 35–59. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Benites, Fernando, Monika Schlatter, Marcel Messerli und Rocco Custer. 2022. „Work-in-progress: Data Science Challenge-X: Self-directed, Competence-based, Project-based Learning.“ Paper presented at 2022 IEEE Global Engineering Education Conference (EDUCON). Tunis, 28.-31. März 2022. Tunis: IEEE Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/EDUCON52537.2022.9766710>. Zugriff 17.2.2024.
- Bergmann, Jörg R. 2000. „Ethnomethodologie.“ In *Qualitative Forschung: Ein Handbuch*, hrsg. v. Uwe Flick, Ernst von Kardorff und Ines Steinke, 118–35. Reinbek bei Hamburg: Rowohlt.
- Betts, Tab und Paolo Oprandi, Hrsg. 2022. *100 Ideas for Active Learning*. The Active Learning Network. <https://openpress.sussex.ac.uk/ideasforactivelearning/open/download?type=pdf>. Zugriff 15.2.2024.

- Biggs, John B. 1996. „Enhancing Teaching through Constructive Alignment.“ *Higher Education* 32 (3): 347–64. doi:10.1007/BF00138871.
- Biggs, John B. 2014. „Constructive Alignment in University Teaching.“ *HERDSA Review of Higher Education* 1: 5–22.
- Biggs, John B. 2023. „SOLO Taxonomy.“ <https://www.johnbiggs.com.au/academic/solo-taxonomy/>. Zugriff 15.2.2024.
- Biggs, John B. und Kevin F. Collis. 1982. *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. Educational Psychology. New York: Academic Press.
- Biggs, John B. und Catherine S. Tang. 2007. *Teaching for Quality Learning at University: What the Student Does*. 3. Auflage. Maidenhead: McGraw-Hill Open University Press.
- Black, Paul und Dylan Wiliam. 1998. „Assessment and Classroom Learning.“ *Assessment in Education: Principles, Policy & Practice* 5 (1): 7–74. doi:10.1080/0969595980050102.
- Bloom, Benjamin S., Max D. Engelhart, Edward J. Furst, Walker H. Hill und David R. Krathwohl. 1956. *Taxonomy of Educational objectives: The Classification of Educational goals: Handbook I: Cognitive domain*. New York: David McKay.
- Bloom, Benjamin S., John T. Hastings und George F. Madaus, Hrsg. 1971. *Handbook on Formative and Summative Evaluation of Student Learning*. New York: McGraw-Hill.
- Bloom, Benjamin S., George F. Madaus und John T. Hastings. 1981. *Evaluation to Improve Learning*. New York, Hamburg: McGraw-Hill.
- Bloxham, Sue und Liz Campbell. 2010. „Generating Dialogue in Assessment Feedback: Exploring the Use of Interactive Cover Sheets.“ *Assessment & Evaluation in Higher Education* 35 (3): 291–300. doi:10.1080/02602931003650045.
- Blühdorn, Hardarik. 2006. „Textverstehen und Intertextualität.“ In *Text – Verstehen: Grammatik und darüber hinaus*, hrsg. v. Hardarik Blühdorn, Eva Breindl und Ulrich H. Wassner, 277–98. Berlin: de Gruyter.
- Blum, Susan D. 2009. *My Word! Plagiarism and College Culture*. Ithaca: Cornell University Press.
- Bohl, Thorsten. 2005. *Leistungsbeurteilung in der Reformpädagogik: Analyse und Gehalt der Beurteilungskonzeptionen*. Weinheim: Beltz.
- Böhle, Fritz, Hrsg. 2017. *Arbeit als Subjektivierendes Handeln*. Wiesbaden: Springer Fachmedien.
- Böhle, Fritz, Norbert Huchler und Judith Neumer. 2019. „Wozu noch menschliche Arbeit – Grenzen der Digitalisierung als neue Herausforderung für die Weiterbildung.“ In *Digitalisierung und Lernen: Gestaltungsperspektiven für das professionelle Handeln in der Erwachsenenbildung und Weiterbildung*, hrsg. v. Erik Haberzeth und Irena Sgier, 21–43. Forum Hochschuldidaktik und Erwachsenenbildung 8. Bern: hep.
- Böing, Ursula. 2009. *Professionelles Handeln von Lehrpersonen im Unterricht mit Schülern und Schülerinnen mit schwerer Behinderung: Die „Forschungswerkstatt“ als Instrument der Lehrerbildung in arbeitsplatzbezogenen Reflexionsprozessen*. Köln: Universität Köln. Dissertation.

- Böll, Heinrich. 1974. *Die verlorene Ehre der Katharina Blum oder: Wie Gewalt entstehen und wohin sie führen kann*. Köln: Kiepenheuer und Witsch.
- Bonaccio, Silvia und Charlie L. Reeve. 2010. „The Nature and Relative Importance of Students' Perceptions of the Sources of Test Anxiety.“ *Learning and Individual Differences* 20 (6): 617–25. doi:10.1016/j.lindif.2010.09.007.
- Borella, P., A. Bargellini, S. Rovesti, M. Pinelli, R. Vivoli, V. Solfrini und G. Vivoli. 1999. „Emotional Stability, Anxiety, and Natural Killer Activity Under Examination Stress.“ *Psychoneuroendocrinology* 24 (6): 613–27. doi:10.1016/S0306-4530(99)00016-5.
- Bortz, Jürgen. 2005. *Statistik für Human- und Sozialwissenschaftler: Mit 242 Tabellen*. 6., vollständig überarbeitete und aktualisierte Auflage Heidelberg: Springer Medizin.
- Borup, Jered, Richard E. West, Rebecca Thomas und Charles R. Graham. 2014. „Examining the Impact of Video Feedback on Instructor Social Presence in Blended Courses.“ *The International Review of Research in Open and Distributed Learning* 15 (3): 232–56. doi:10.19173/irrodl.v15i3.1821.
- Boud, David. 1995. *Enhancing Learning through Self Assessment*. London: RoutledgeFalmer.
- Bourke, Roseanna. 2022. „Normalising Alternative Assessment Approaches for Inclusion.“ In *Assessment for Inclusion in Higher Education*, hrsg. v. Rola Ajjawi, Joanna Tai, David Boud und Trina Jorre de St Jorre, 189–98. London: Routledge.
- Bransford, John D. und Jeffery J. Franks. 1971. „The Abstraction of Linguistic Ideas.“ *Cognitive Psychology* 2 (4): 331–50. doi: 10.1016/0010-0285(71)90019-3.
- Breidenstein, Georg. 2018. „Das Theorem der ‚Selektionsfunktion der Schule‘ und die Praxis der Leistungsbewertung.“ In *Leistung als Paradigma: Zur Entstehung und Transformation eines pädagogischen Konzepts*, hrsg. v. Sabine Reh und Norbert Ricken, 307–27. Wiesbaden: Springer Fachmedien.
- Breidenstein, Georg und Christiane Thompson. 2014. „Schulische Leistungsbewertung als Praxis der Subjektivierung.“ In *Interferenzen: Perspektiven kulturwissenschaftlicher Bildungsforschung*, hrsg. v. Christiane Thompson, Kerstin Jergus und Georg Breidenstein, 89–109. Weilerswist: Velbrück Wissenschaft.
- Brendel, Sabine, Ulrike Hanke und Gerd Macke. 2019. *Kompetenzorientiert lehren an der Hochschule*. UTB Schlüsselkompetenzen, Hochschuldidaktik 5047. Opladen, Toronto: Verlag Barbara Budrich.
- Brewer, William F. und James C. Treyns. 1981. „Role of Schemata in Memory for Places.“ *Cognitive Psychology* 13 (2): 207–30. doi:10.1016/0010-0285(81)90008-6.
- Brien, Jörn. 2023. „335.000 Dollar für Prompt Engineers: ‚KI-Flüsterer‘ in Firmen heiß begehrt.“ *T3N*, 29. März. Online verfügbar unter: <https://t3n.de/news/prompt-engineers-ki-fluesterer-begehrt-1544700/>.
- Brigard, Felipe de. 2014. „Is Memory for Remembering? Recollection as a Form of Episodic Hypothetical Thinking.“ *Synthese* 191 (2): 155–85. doi:10.1007/s11229-013-0247-7.
- Brinker, Tobina und Peter Tremp, Hrsg. 2012. *Einführung in die Studiengangentwicklung*. Blickpunkt Hochschuldidaktik 122. Bielefeld: Wbv Bertelsmann.

- Brookhart, Susan M. 2001. „Successful Students’ Formative and Summative Uses of Assessment Information.“ *Assessment in Education: Principles, Policy & Practice* 8 (2): 153–69. doi:10.1080/09695940123775.
- Brookhart, Susan M. 2018a. „Appropriate Criteria: Key to Effective Rubrics.“ *Frontiers in Education* 3.
- Brookhart, Susan M. 2018b. „Summative and Formative Feedback.“ In *The Cambridge Handbook of Instructional Feedback*, hrsg. v. Anastasiya A. Lipnevich und Jeffrey K. Smith, 52–78. Cambridge, New York, NY: Cambridge University Press.
- Brookhart, Susan M., Thomas R. Guskey, Alex J. Bowers, James H. McMillan, Jeffrey K. Smith, Lisa F. Smith, Michael T. Stevens und Megan E. Welsh. 2016. „A Century of Grading Research.“ *Review of Educational Research* 86 (4): 803–48. doi:10.3102/0034654316672069.
- Brown, Gavin T. L. und Hasan H. A. Abdalnabi. 2017. „Evaluating the Quality of Higher Education Instructor-Constructed Multiple-Choice Tests: Impact on Student Grades.“ *Frontiers in Education* 2. doi: 10.3389/feduc.2017.00024. Zugriff 15.2.2024.
- Brühwiler, Christian und Andreas Helmke. 2018. „Determinanten der Schulleistung.“ In *Handwörterbuch pädagogische Psychologie*, hrsg. v. Detlef H. Rost, Jörn R. Sparfeldt und Susanne Buch. 5., überarbeitete und erweiterte Auflage, 78–91. Weinheim, Basel: Beltz.
- Bubenhof, Noah. 2022. „Wie wir in Zukunft wissenschaftliche Texte schreiben (könnten) – Teil 1.“ <https://www.bubenhof.com/sprechtagel/2022/12/08/wie-wir-in-zukunft-wissenschaftliche-texte-schreiben-koennten-teil-1/>. Zugriff 3.1.2024.
- Buff Keller, Eva und Stefan Jörissen. 2015. *Abschlussarbeiten im Studium anleiten, betreuen und bewerten*. UTB Schlüsselkompetenzen 4345. Opladen: Budrich.
- Bukodi, Erzsébet, Robert Erikson und John H. Goldthorpe. 2014. „The Effects of Social Origins and Cognitive Ability on Educational Attainment.“ *Acta Sociologica* 57 (4): 293–310. doi:10.1177/0001699314543803.
- Bulik-Sullivan, Brendan K. und Patrick F. Sullivan. 2012. „The Authorship Network of Genome-Wide Association Studies.“ *Nature genetics* 44 (2): 113. doi:10.1038/ng.1052.
- Burhans, Karen K. und Carol S. Dweck. 1995. „Helplessness in Early Childhood: The Role of Contingent Worth.“ *Child Development* 66 (6): 1719–38. doi:10.1111/J.1467-8624.1995.TB00961.X.
- Burrell, Jenna und Marion Fourcade. 2021. „The Society of Algorithms.“ *Annual Review of Sociology* 47 (1): 213–37. doi:10.1146/annurev-soc-090820-020800.
- Burton, Nancy W. und Leonard Ramist. 2001. „Predicting Success in College: SAT® Studies of Classes Graduating Since 1980.“ <https://files.eric.ed.gov/fulltext/ED562836.pdf>. Zugriff 15.2.2024.
- Butler, Ruth. 1988. „Enhancing and Undermining Intrinsic Motivation: The Effects of Task-Involving and Ego-Involving Evaluation on Interest and Performance.“ *The British Journal of Educational Psychology* 58 (1): 1–14. doi:10.1111/j.2044-8279.1988.tb00874.x.

- Campbell, Donald T. 1976. *Assessing the Impact of Planned Social Change*. Hanover, NH: The Public Affairs Center, Dartmouth College.
- Campe, Joachim H. 1998 [1788]. *Über das Zweckmäßige und Unzweckmäßige in den Belohnungen und Strafen*, hrsg. v. Birgit Ofenbach. Heinsberg: Dieck.
- Carless, David und David Boud. 2018. „The Development of Student Feedback Literacy: Enabling Uptake of Feedback.“ *Assessment & Evaluation in Higher Education* 43 (8): 1315–25. doi:10.1080/02602938.2018.1463354.
- Carnevale, Anthony P. und Stephen J. Rose. 2004. „Socioeconomic Status, Race/Ethnicity, and Selective College Admissions.“ In *America's Untapped Resource: Low-Income Students in Higher Education*, hrsg. v. Richard D. Kahlenberg, 101–56. New York: Century Foundation Press.
- Carter, Ashli B. und Katherine W. Phillips. 2017. „The Double-Edged Sword of Diversity: Toward a Dual Pathway Model.“ *Social and Personality Psychology Compass* 11 (5): e12313. doi:10.1111/spc3.12313.
- Cellan-Jones, Rory. 2014. „Stephen Hawking Warns Artificial Intelligence Could End Mankind.“ *BBC.com*, 2. Dezember. <https://www.bbc.com/news/technology-30290540>. Zugriff 15.2.2024.
- Cepeda, Nicholas J., Edward Vul, Doug Rohrer, John T. Wixted und Harold Pashler. 2008. „Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention.“ *Psychological Science* 19 (11): 1095–1102. doi:10.1111/j.1467-9280.2008.02209.x.
- Chen, Daniel L. und Arnaud Philippe. 2019. „Clash of Norms: Judicial Leniency on Defendant Birthdays.“ *SSRN Journal*. doi:10.2139/ssrn.3203624. Zugriff 15.2.2024.
- Chi, Michelene T., Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi und Robert G. Hausmann. 2001. „Learning from Human Tutoring.“ *Cognitive Science* 25 (4): 471–533. doi:10.1207/s15516709cog2504_1.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge MA: MIT Press.
- Clark, Andy. 1997. *Being There: Putting Brain, Body, and World Together Again*. Cambridge MA: MIT Press.
- Cohen, S. A. 1987. „Instructional Alignment: Searching for a Magic Bullet.“ *Educational Researcher* 16 (8): 16–20.
- Collins, Allan, John S. Brown und Susan E. Newman. 1989. „Cognitive Apprenticeship: Teaching the Crafts of Reading, Writing, and Mathematics.“ In *Knowing, Learning, and Instruction: Essays in honor of Robert Glaser*, hrsg. v. Lauren B. Resnick. 3. print, 453–94. Hillsdale, NJ: Erlbaum.
- Collins, Randall. 1979. *The Credential Society: An Historical Sociology of Education and Stratification*. New York: Academic Press.
- Cominelli, Luigi. 2018. *Cognition of the Law: Toward a Cognitive Sociology of Law and Behavior*. Cham: Springer International Publishing.
- Crouch, Catherine H. und Eric Mazur. 2001. „Peer Instruction: Ten Years of Experience and Results.“ *American Journal of Physics* 69 (9): 970–77. doi:10.1119/1.1374249.

- Dainton, Nora. 2018. *Feedback in der Hochschullehre*. Stuttgart, Bern: UTB; Haupt Verlag.
- Danielewicz, Jane und Peter Elbow. 2009. „A Unilateral Grading Contract to Improve Learning and Teaching.“ *College Composition and Communication* 61 (2): 244–68.
- Danziger, Shai, Jonathan Levav und Liora Avnaim-Pesso. 2011. „Extraneous Factors in Judicial Decisions.“ *Proceedings of the National Academy of Sciences of the United States of America* 108 (17): 6889–92. doi:10.1073/pnas.1018033108.
- Deci, Edward L., Richard Koestner und Richard M. Ryan. 1999. „A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation.“ *Psychological Bulletin* 125 (6): 627–68; discussion 692–700. doi: 10.1037/0033-2909.125.6.627.
- Decker, Christian. 2018. „Posterkonferenzen.“ In *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*, hrsg. v. Julia Gerick, Angela Sommer und Germo Zimmermann, 185–88. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Decker, Christian. 2023. „Die Posterkonferenz als hochschuldidaktische Methode und Alternative zum Referat.“ In *Neues Handbuch Hochschullehre: Praxisnah – erfahrungsbasiert – wissenschaftsfundiert*. G 1.2 (78. EL), hrsg. v. Brigitte Berendt, Andreas Fleischmann, Gudrun Salmhofer, Niclas Schaper, Birgit Szczyrba, Matthias Wiemer und Johannes Wildt. Berlin: DUZ Verlags- und Medienhaus.
- Deslauriers, Louis, Logan S. McCarty, Kelly Miller, Kristina Callaghan und Greg Kestin. 2019. „Measuring Actual Learning versus Feeling of Learning in Response to Being Actively Engaged in the Classroom.“ *Proceedings of the National Academy of Sciences of the United States of America* 116 (39): 19251–57. doi:10.1073/pnas.1821936116.
- Dickson, Elmer G. 1974. „Contract Grading.“ *Journal of Financial Education* 2 (3): 21–24.
- Digital Learning und Zentrum für Hochschuldidaktik und -entwicklung. 2020. „Lehre gestalten an der PH Zürich: Neue Mischformen zwischen Online- und Präsenzlehre.“ https://tiny.phzh.ch/lehre_hybrid. Zugriff 17.5.2023.
- Dilly, Marc. 2018. „Objektive strukturierte praktische/klinische Prüfung.“ In *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*, hrsg. v. Julia Gerick, Angela Sommer und Germo Zimmermann, 158–62. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Döbeli, Beat. 2023. „ChatGPT & Schule: Einschätzungen der Professur „Digitalisierung und Bildung“ der Pädagogischen Hochschule Schwyz.“. <https://lernentrotz.corona.ch/pub/LLM/WebHome/2023-chat-gpt-und-schule-v128.pdf>. Zugriff 27.12.2023.
- Dröge, Kai, Kira Marrs und Wolfgang Menz. 2008. „Rückkehr der Leistungsfrage. Leistung in Arbeit, Unternehmen und Gesellschaft – zur Einleitung.“ In *Rückkehr der Leistungsfrage: Leistung in Arbeit, Unternehmen und Gesellschaft*, hrsg. v. Kai Dröge, Kira Marrs und Wolfgang Menz, 7–18. Berlin: Sigma.

- Dubs, Rolf. 1995. *Lehrerverhalten: Ein Beitrag zur Interaktion von Lehrenden und Lernenden im Unterricht*. Schriftenreihe für Wirtschaftspädagogik 23. Zürich: Verlag des Schweizerischen Kaufmännischen Verbandes.
- Dudenredaktion. o. J. „Feedback“ auf Duden online.“ <https://www.duden.de/node/45632/revision/1260455>. Zugriff 12.6.2023.
- Eagleman, David. 2020. *Livewired: The Inside Story of the Ever-Changing Brain*. Edinburgh: Canongate Books.
- Eberle, Franz und Christoph Metzger. 1988. *Gültigkeit und Zuverlässigkeit von Aufsatzkorrekturen am Beispiel der Zulassungsprüfung für ausländische Studienbewerber der Hochschule St. Gallen*. Unter Mitarbeit von Peter Sieber. Studien und Berichte des IWP 7. St. Gallen: Institut für Wirtschaftspädagogik.
- Eberle, Thomas S. 1984. *Sinnkonstitution in Alltag und Wissenschaft: Der Beitrag der Phänomenologie an die Methodologie der Sozialwissenschaften*. Veröffentlichungen der Hochschule St. Gallen für Wirtschafts- und Sozialwissenschaften. Schriftenreihe Kulturwissenschaft 5. Bern: Haupt.
- Eberly Center. 2008/2015. „Why Should Assessments, Learning Objectives, and Instructional Strategies Be Aligned?“ <https://www.cmu.edu/teaching/assessment/basics/alignment.html>. Zugriff 24.5.2023.
- Edelmann, Walter. 2005. *Lernpsychologie: Lehrbuch*. 6., vollständig überarbeitete Auflage. Weinheim: Beltz PVU.
- Effinger, Herbert. 2018. „Ganzheitlich arbeiten, fragmentiert studieren: Das Studiengangskonzept des Masterstudiengangs Soziale Arbeit an der ehs Dresden.“ In *Lernortverknüpfung. Didaktische Ansätze und Perspektiven berufsintegrierenden Studierens. Tagungsband*, hrsg. v. Claudia Albrecht und Johanna Schneider, 46–55. Dresden.
- Elbow, Peter und Mary D. Sorcinelli. 2006. „How to Enhance Learning by Using High-Stakes and Low-Stakes Writing.“ In *McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers*, hrsg. v. Wilbert J. McKeachie und Marilla D. Svinicki. 12. Auflage, 192–212. College Teaching Series. Boston: Houghton Mifflin.
- Embretson, Susan E. und Steven P. Reise. 2015. *Item Response Theory*. 2. Auflage. Multivariate Applications Books Series 4. Mahwah, NJ: Routledge Academic.
- Engelmann, Katharina, Maria Bannert und Nadine Melzner. 2021. „Do Self-Created Metacognitive Prompts Promote Short- and Long-Term Effects in Computer-Based Learning Environments?“ *Research and Practice in Technology Enhanced Learning* 16 (1). doi:10.1186/s41039-021-00148-w.
- Engfer, Dagmar. 2018. „Blended Coaching: Ein Konzept zur Verbindung von Weiterbildung und Beratung.“ In *Weiterbildung an Hochschulen: Über Kurse und Lehrgänge hinaus*, hrsg. v. Tobias Zimmermann, Geri Thomann und Denise Da Rin, 141–63. Forum Hochschuldidaktik und Erwachsenenbildung 7. Bern: hep verlag.
- Eren, Ozkan und Naci Mocan. 2018. „Emotional Judges and Unlucky Juveniles.“ *American Economic Journal: Applied Economics* 10 (3): 171–205. doi:10.1257/app.20160390.

- Ericsson, Karl A., Ralf T. Krampe und Clemens Tesch-Römer. 1993. „The Role of Deliberate Practice in the Acquisition of Expert Performance.“ *Psychological Review* 100 (3): 363–406. doi:10.1037/0033-295X.100.3.363.
- Ericsson, Karl A. und Robert Pool. 2016. *Top. Die neue Wissenschaft vom bewussten Lernen [Original: Peak. Secrets from the new Science of Expertise]*. München: Pattloch. Aus dem Amerikanischen von Gabriele Gockel, Barbara Steckhan und Claus Varrelmann.
- Eugster, Balthasar und Peter Tremp. 2018. „Lehre als Zugang zum Fach: Plädoyer für eine didaktische Wissenschaftssozialisation.“ In *Zukunftslabor Lehrentwicklung: Perspektiven auf Hochschuldidaktik und darüber hinaus*, hrsg. v. Markus Weil, 75–93. Münster, New York: Waxmann.
- Fastré, Greet M. J., Marcel R. van der Klink und Jeroen J. G. van Merriënboer. 2010. „The Effects of Performance-Based Assessment Criteria on Student Performance and Self-Assessment Skills.“ *Advances in Health Sciences Education: Theory and Practice* 15 (4): 517–32. doi:10.1007/s10459-009-9215-x.
- Faulstich, Peter. 2002. „Verteidigung von ‚Bildung‘ gegen die Gebildeten unter ihren Verächtern.“ In *Thema: Kompetenzentwicklung statt Bildungsziele?* hrsg. v. Ekehard Nuissl, Christiane Schiersmann und Horst Siebert, 15–25. Literatur- und Forschungsreport Weiterbildung 49. Bielefeld: Bertelsmann.
- Fehm, Lydia, Thomas Fydrich und Klara Sommer. 2022. *Prüfungsangst. 2., überarbeitete Auflage*. Fortschritte der Psychotherapie 44. Göttingen: Hogrefe.
- Feixas, Mònica und Tobias Zimmermann. 2018. „Competence-Oriented Approaches to Assessment.“ In *Competence-Oriented Teaching and Learning in Higher Education – Essentials*, hrsg. v. Heinz Bachmann, 74–108. Bern: hep.
- Feixas, Mònica und Tobias Zimmermann. 2023. „Authenticity and Inclusiveness in University Assessment: An Examination of Current Visions of Authentic and Inclusive Assessment and of the Systemic Barriers in Institutional Policies, Processes and Practices that Limit its Implementation.“ Präsentation gehalten am *Jahreskongress 2023 der Schweizerischen Gesellschaft für Bildungsforschung (SGBF) und der Schweizerischen Gesellschaft für Lehrerinnen- und Lehrerbildung (SGL): Zeiten des Umbruchs? Bildung zwischen Persistenz und Wandel*, Zürich, 28.-30. Juni. https://express.converia.de/frontend/index.php?page_id=24265&v=List&do=15&day=all&ses=26193#. Zugriff 15.2.2024.
- Feynman, Richard P. 1988. *“What Do You Care What Other People Think?”: Further Adventures of a Curious Character*. Richard P. Feynman as told to Ralph Leighton. New York: Norton.
- Finke, Jasper. 2019. „Praxisbeispiel Jura: Moot Courts in der juristischen Ausbildung.“ In *Kompetenzorientiert prüfen und bewerten an Universitäten: Didaktische Grundannahmen, rechtliche Rahmenbedingungen und praktische Handlungsempfehlungen*, hrsg. v. Susanne Frölich-Steffen, Hendrik den Ouden und Ursula Gießmann, 171–80. Opladen, Berlin, Toronto: Verlag Barbara Budrich.
- Fischer, Walther L. 1991. „Mathematische Kritik der Ziffernnoten und ihrer Interpretation.“ In *Schülerbeurteilungen und Schulzeugnisse: Historische und systematische Aspekte*, hrsg. v. Johann G. von Hohenzollern und Max Liedtke, 225–49. Bad Heilbrunn: Klinkhardt.

- Fleischmann, Andreas. 2023. „ChatGPT in der Hochschullehre: Wie künstliche Intelligenz uns unterstützen und herausfordern wird.“ In *Neues Handbuch Hochschullehre: Praxisnah – erfahrungsbasiert – wissenschaftsfundiert*. A 1.30 (110. EL), hrsg. v. Brigitte Berendt, Andreas Fleischmann, Gudrun Salmhofer, Niclas Schaper, Birgit Szczyrba, Matthias Wiemer und Johannes Wildt. Berlin: DUZ Verlags- und Medienhaus.
- Floridi, Luciano. 2019. *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford: Oxford University Press.
- Foucault, Michel. 2004. *Die Geburt der Biopolitik. Geschichte der Gouvernementalität II. Vorlesung am College de France 1978–1979*, hrsg. v. Michel Sennelart. Aus dem Französischen von Jürgen Schröder. Frankfurt a. M.: Suhrkamp.
- Frederick, Shane. 2005. „Cognitive Reflection and Decision Making.“ *Journal of Economic Perspectives* 19 (4): 25–42. doi:10.1257/089533005775196732.
- Frölich-Steffen, Susanne und Hendrik den Ouden. 2019. „Hochschulprüfungen im Fokus. Vom autodidaktischen Abschauren zu didaktisch-gereifter Prüfungskompetenz.“ In *Kompetenzorientiert prüfen und bewerten an Universitäten: Didaktische Grundannahmen, rechtliche Rahmenbedingungen und praktische Handlungsempfehlungen*, hrsg. v. Susanne Frölich-Steffen, Hendrik den Ouden und Ursula Gießmann, 11–27. Opladen, Berlin, Toronto: Verlag Barbara Budrich.
- Furck, Carl-Ludwig. 1972. *Das pädagogische Problem der Leistung in der Schule*. 4. Auflage. Weinheim: Beltz.
- Galton, Francis. 1889. *Natural Inheritance*. London: MacMillan.
- Gardner, Howard. 1993. *Frames of Mind: The Theory of Multiple Intelligences. With a New Introduction by the Author*. Ausgabe zum zehnjährigen Jubiläum. New York, NY: BasicBooks.
- Gast, Lilli. 2009. „„Kein Ort. Nirgends?“ Das Subjekt der Erkenntnis und die Idee der Universität. Einige Gedanken aus psychoanalytischer Perspektive.“ *Psychologie und Gesellschaftskritik* 33/34 (132/133): 153–71.
- Gerholz, Karl-Heinz, Ilona Maidanuk und Philipp Schlottmann. 2022. „Virtual Reality in der (beruflichen) Lehrerinnen- und Lehrerbildung – Hochschuldidaktische Einordnung und empirische Befunde auf Basis eines systematischen Literaturreviews.“ In *Digital Literacy in der beruflichen Lehrer:innenbildung: Didaktik, Empirie und Innovation*, hrsg. v. Karl-Heinz Gerholz, Philipp Schlottmann, Peter Slepcevic-Zach und Michaela Stock, 185–98. Berufsbildung, Arbeit und Innovation 68. Bielefeld: wbv Media GmbH & Co. KG.
- Gerick, Julia, Angela Sommer und Gernot Zimmermann, Hrsg. 2018. *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Ghisla, Gianni. 2005. „Editorial: Modularisierung der Bildung: Flexibilität, aber zu welchem Preis?“ *Schweizerische Zeitschrift für Bildungswissenschaften* 27 (2): 157–74.
- Gibbs, Graham und Claire Simpson. 2004. „Conditions under which Assessment Supports Students' Learning.“ *Learning and Teaching in Higher Education* 1 (1): 3–31.

- Gießmann, Ursula. 2019. „Grundlagen zum Prüfen bei wissenschaftlichen Arbeiten.“ In *Kompetenzorientiert prüfen und bewerten an Universitäten: Didaktische Grundannahmen, rechtliche Rahmenbedingungen und praktische Handlungsempfehlungen*, hrsg. v. Susanne Frölich-Steffen, Hendrik den Ouden und Ursula Gießmann, 121–23. Opladen, Berlin, Toronto: Verlag Barbara Budrich.
- Gigerenzer, Gerd und Wolfgang Gaissmaier. 2006. „5. Kapitel: Denken und Urteilen unter Unsicherheit: Kognitive Heuristiken.“ In *Denken und Problemlösen*, hrsg. v. Joachim Funke, 329–74. Enzyklopädie der Psychologie 8. Göttingen: Hogrefe Verlag für Psychologie.
- Gill, Phupinder. 2017. „Global Developments in Equity Trading.“ In *Equity Markets in Transition*, hrsg. v. Reto Francioni und Robert A. Schwartz, 391–409. Cham: Springer International Publishing.
- Girgensohn, Katrin und Nadja Sennwald. 2012. *Schreiben lehren, Schreiben lernen: Eine Einführung*. Darmstadt: WBG.
- Gladwell, Malcolm. 2008. *Outliers: The Story of Success*. 1. Auflage. New York: Little Brown and Co.
- Gottselig, J. M., G. Hofer-Tinguely, A. A. Borbély, S. J. Regel, H-P Landolt, J. V. Rétey und P. Achermann. 2004. „Sleep and Rest Facilitate Auditory Learning.“ *Neuroscience* 127 (3): 557–61. doi:10.1016/j.neuroscience.2004.05.053.
- Graf, Dittmar. 2001. „Welche Aufgabentypen gibt es?“ *Der mathematische und naturwissenschaftliche Unterricht* 54 (7): 422–25.
- Gräsel, Cornelia, Johannes Bruhn, Heinz Mandl und Frank Fischer. 1997. „Lernen mit Computernetzen aus konstruktivistischer Perspektive.“ *Unterrichtswissenschaft* 25 (1), 4–18.
- Gregory, Richard. 1998. „Brainy Mind.“ *British Medical Journal* 317 (7174): 1693–95. doi:10.1136/bmj.317.7174.1693.
- Greve, Werner und Dirk Wentura. 1997. *Wissenschaftliche Beobachtung: Eine Einführung*. 2. Auflage. Weinheim: Beltz.
- Guba, Egon G. und Yvonna S. Lincoln. 1989. *Fourth Generation Evaluation*. Newbury Park, CA: SAGE Publications.
- Gugler, Anne. 2023. „KI darf nicht das Denken von Studierenden ersetzen.“ *Freiburger Nachrichten*, 17.4.2023, 2.
- Gunstone, Richard F. und Richard T. White. 1981. „Understanding of Gravity.“ *Science Education* 65 (3): 291–99. doi:10.1002/sci.3730650308.
- Gutsmiedl-Schumann, Doris. 2018. „Beitrag in einer formalisierten Debatte bzw. im Debattierclub.“ In *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*, hrsg. v. Julia Gerick, Angela Sommer und Gernot Zimmermann, 25–29. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Hadjar, Andreas und Rolf Becker. 2009. „Erwartete und unerwartete Folgen der Bildungsexpansion in Deutschland.“ In *Lehrbuch der Bildungssoziologie*, hrsg. v. Rolf Becker, 195–213. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Hagemann, Dirk, Frank M. Spinath und Erik M. Mueller. 2022. *Differentielle Psychologie und Persönlichkeitsforschung*. Unter Mitarbeit von M. Hasselhorn, W. Kunde und S. Schneider. 9. überarbeitete Auflage. Stuttgart: Kohlhammer Verlag.

- Haladyna, Thomas M. 2004. *Developing and Validating Multiple-Choice Test Items*. 3. Auflage. London, New York: Routledge Taylor & Francis Group.
- Hancock, Jessica C. 2022. „Haikus for Learning.“ In *100 Ideas for Active Learning*, hrsg. v. Tab Betts und Paolo Oprandi, 411–16: The Active Learning Network. <https://openpress.sussex.ac.uk/ideasforactivelearning/open/download?type=pdf>. Zugriff 15.2.2024.
- Hanft, Anke, Olaf Zawacki-Richter und Willi B. Gierke, Hrsg. 2015. *Herausforderung Heterogenität beim Übergang in die Hochschule*. Münster, New York: Waxmann.
- Hassler, Dominic. 2022. „Pädagogische Hochschule Zürich: Multimediale Formen von Feedback.“ In *E-Learning: Digitale Lehr- und Lernangebote in Zeiten von Smart Devices und Online-Lehre*, hrsg. v. Ullrich Dittler. 5. aktualisierte Auflage, 143–58. Berlin: de Gruyter.
- Hattie, John A. C. 2009. *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. London: Routledge.
- Hattie, John A. C. und Shirley Clarke. 2019. *Visible Learning: Feedback*. London: Routledge.
- Hattie, John A. C. und Gregory M. Donoghue. 2016. „Learning Strategies: a Synthesis and Conceptual Model.“ *NPJ Science of Learning* 1.
- Hattie, John A. C. und Helen Timperley. 2007. „The Power of Feedback.“ *Review of Educational Research* 77 (1): 81–112. doi:10.3102/003465430298487.
- Hattie, John A. C. und Klaus Zierer. 2017. *10 Mindframes for Visible Learning: Teaching for Success*. Milton: Taylor and Francis.
- Hazel, Elizabeth, Michael Prosser und Keith Trigwell. 2002. „Variation in Learning Orchestration in University Biology Courses.“ *International Journal of Science Education* 24 (7): 737–51. doi:10.1080/09500690110098886.
- Hebb, Donald O. 1949. *The Organization of Behavior: A Neuropsychological Theory*. New York: John Wiley.
- Heersmink, Richard. 2022. „Human Uniqueness in Using Tools and Artifacts: Flexibility, Variety, Complexity.“ *Synthese* 200 (6). doi:10.1007/s11229-022-03892-8.
- Heid, Helmut. 1989. „Pädagogische Rückfragen an herkömmliche Bestimmungen von ‚Verhaltensauffälligkeit‘ und Schulleistungsversagen.“ In *Beratungsaufgaben in der Schule: Psychologisch-pädagogische Hilfen aus Theorie und Praxis für erzieherische und unterrichtliche Beratungsanlässe*, hrsg. v. Helmut Lukesch, Wolfgang Nöldner und Helmut Peez, 41–52. München: Reinhardt.
- Heintz, Bettina. 2016. „„Wir leben im Zeitalter der Vergleichung.“ Perspektiven einer Soziologie des Vergleichs.“ *Zeitschrift für Soziologie* 45 (5): 305–23. doi: 10.1515/zfsoz-2015-1018.
- Helsper, Werner. 1996. „Antinomien des Lehrerhandelns in modernisierten pädagogischen Kulturen: Paradoxe Verwendungsweisen von Autonomie und Selbstverantwortlichkeit.“ In *Pädagogische Professionalität: Untersuchungen zum Typus pädagogischen Handelns*, hrsg. v. Arno Combe und Werner Helsper, 521–69. Frankfurt a. M.: Suhrkamp.

- Helsper, Werner. 2016. „Lehrerprofessionalität – der strukturtheoretische Ansatz.“ In *Beruf Lehrer/Lehrerin: Ein Studienbuch*, hrsg. v. Martin Rothland, 103–25. Münster, New York: Waxmann.
- Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan C. Cardenas et al. 2010. „Markets, Religion, Community Size, and the Evolution of Fairness and Punishment.“ *Science* 327 (5972): 1480–84. doi:10.1126/science.1182238.
- Hessler, Gudrun und Mechthild Oechsle. 2012. „Studium und Beruf – Praxiskonzepte von Studierenden der Soziologie und Sozialwissenschaften.“ In *Studium nach Bologna: Praxisbezüge stärken?! Praktika als Brücke zwischen Hochschule und Arbeitsmarkt*, hrsg. v. Wilfried Schubarth, Karsten Speck, Andreas Seidel, Corinna Gottmann, Caroline Kamm und Maud Krohn, 113–25. Wiesbaden: Springer Fachmedien.
- Heyes, Anthony und Soodeh Saberian. 2019. „Temperature and Decisions: Evidence from 207,000 Court Cases.“ *American Economic Journal: Applied Economics* 11 (2): 238–65. doi:10.1257/app.20170223.
- Hiller, Tammy B. und Amy B. Hietapelto. 2001. „Contract Grading: Encouraging Commitment to the Learning Process through Voice in the Evaluation Process.“ *Journal of Management Education* 25 (6): 660–84. doi:10.1177/105256290102500605.
- Hinds, Pamela J. 1999. „The Curse of Expertise: The Effects of Expertise and Debiasing Methods on Predictions of Novice Performance.“ *Journal of Experimental Psychology: Applied* 5 (2): 205–21.
- Hirdina, Ralph. 2019. „Ausgewählte Rechtsaspekte zum Prüfen an Hochschulen.“ In *Kompetenzorientiert prüfen und bewerten an Universitäten: Didaktische Grundannahmen, rechtliche Rahmenbedingungen und praktische Handlungsempfehlungen*, hrsg. v. Susanne Frölich-Steffen, Hendrik den Ouden und Ursula Gießmann, 39–60. Opladen, Berlin, Toronto: Verlag Barbara Budrich.
- Huber, Ludwig. 1995. „Hochschuldidaktik als Theorie der Bildung und Ausbildung.“ In *Ausbildung und Sozialisation in der Hochschule*, hrsg. v. Ludwig Huber, 114–38. Enzyklopädie Erziehungswissenschaft 10, hrsg. von Dieter Lenzen. Stuttgart, Dresden: Klett.
- Hucho, Carsten. 2010. „Jung, erfolgreich, exzellent – aber nur auf den ersten Blick: Das Quantifizierungsritual als falsche Fährte.“ *Gegenworte* 13 (2): 72–73.
- Huntley, Christopher D., Bridget Young, James Temple, Melissa Longworth, Catrin T. Smith, Vikram Jha und Peter L. Fisher. 2019. „The Efficacy of Interventions for Test-Anxious University Students: A Meta-Analysis of Randomized Controlled Trials.“ *Journal of anxiety disorders* 63: 36–50. doi: 10.1016/j.janxdis.2019.01.007.
- Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.
- Iller, Carola und Alexander Wick. 2009. „Prüfungen als Evaluation der Kompetenzentwicklung im Studium.“ *Das Hochschulwesen* 57 (6): 195–201.
- Ingenkamp, Karlheinz. 1971. „Sind Zensuren aus verschiedenen Klassen vergleichbar?“ In *Die Fragwürdigkeit der Zensurenggebung: Texte und Untersuchungsberichte*, hrsg. v. Karlheinz Ingenkamp, 156–63. Weinheim, Basel: Beltz.

- Ingenkamp, Karlheinz und Urban Lissmann. 2008. *Lehrbuch der Pädagogischen Diagnostik*. 6., neu ausgestattete Auflage Beltz Pädagogik. Weinheim, Basel: Beltz.
- Institute for Teaching and Learning Innovation, The University of Queensland. 2023. „Teaching Frameworks and Models.“ <https://itali.uq.edu.au/teaching-guidance/principles-learning/guiding-theories-and-frameworks/teaching-frameworks-and-models>. Zugriff 19.4.2023.
- Ireland, Jo und Melissa Mouthaan. 2020. „Perspectives on Curriculum Design: Comparing the Spiral and the Network Models.“ *Research Matters: A Cambridge Assessment Publication* 30 (2): 7–12.
- Jakobs, Eva-Maria. 1993. „„Das kommt mir so bekannt vor...“. Plagiate als verdeckte Intertextualität.“ *Zeitschrift für Germanistik* 3 [14] (2): 377–90.
- Jäncke, Lutz. 2018. „Die Neurobiologie des menschlichen Lernens.“ In *Kompetenzorientierte Hochschullehre: Die Notwendigkeit von Kohärenz zwischen Lernzielen, Prüfungsformen und Lehr-Lern-Methoden*, hrsg. v. Heinz Bachmann. 3., überarbeitete Auflage, 122–44. Forum Hochschuldidaktik und Erwachsenenbildung 1. Bern: hep.
- Jenert, Tobias, Miriam Barnat, Peter Salden und Bernadette Dilger. 2018. „Struktur, Prozess oder Didaktik als Ausgangspunkt? – Ein integratives Modell der Curriculumentwicklung an Hochschulen.“ In *Gelingende Lehre: erkennen, entwickeln, etablieren: Beiträge der Jahrestagung der Deutschen Gesellschaft für Hochschuldidaktik (dghd) 2016*, hrsg. v. Martina Schmohr, Kristina Müller und Julia Philipp, 149–63. Blickpunkt Hochschuldidaktik 132. Bielefeld: wbv Publikation.
- Jennings, Jennifer L. und Douglas L. Lauen. 2016. „Accountability, Inequality, and Achievement: The Effects of the No Child Left Behind Act on Multiple Measures of Student Learning.“ *The Russell Sage Foundation Journal of the Social Sciences* 2 (5): 220–41. doi:10.7758/rsf.2016.2.5.11.
- Jonsson, Anders und Ernesto Panadero. 2018. „Facilitating Students’ Active Engagement with Feedback.“ In *The Cambridge handbook of Instructional Feedback*, hrsg. v. Anastasiya A. Lipnevich und Jeffrey K. Smith, 531–53. Cambridge, New York, NY: Cambridge University Press.
- Jornitz, Sieglinde und Felicitas Macgilchrist. 2021. „Datafizierte Sichtbarkeiten.“ *MedienPädagogik* 45: 98–122. doi.org/10.21240/mpaed/45/2021.12.21.X. Zugriff 15.2.2024.
- Jötten, Bernd. 2007. „Von der Schuldzuweisung zur verlässlichen Kooperation – Vereinbarungskultur zwischen schulischen Bildungspartnern.“ In *Bildungs- und Entwicklungskontrakte als Instrumente von Schulentwicklung: Expertise*, hrsg. v. Marianne Horstkemper, Joachim Ludwig und Wilfried Schubarth, 273–83. Bildungsforschung 21. Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Kahneman, Daniel. 2012. *Thinking, Fast and Slow*. London: Penguin Books.
- Kahneman, Daniel, Olivier Sibony und Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgement*. London: William Collins.
- Kahrs, Miriam und Sandra Kohl. 2018. „Blogbeitrag.“ In *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*, hrsg. v. Julia Gerick, Angela

- Sommer und Germa Zimmermann, 38–41. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Kant, Immanuel. 1803. *Über Pädagogik*. Hrsg. v. D. Friedrich Theodor Rink. Königsberg: Friedrich Nicolovius.
- Kapur, Manu. 2012. „Productive Failure in Learning the Concept of Variance.“ *Instructional Science* 40 (4): 651–72. doi:10.1007/s11251-012-9209-6.
- Karcher, Florian. 2018. „Fachlich reflektierte Hospitation mit Bericht.“ In *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*, hrsg. v. Julia Gerick, Angela Sommer und Germa Zimmermann, 76–79. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Katopodis, Christina und Cathy N. Davidson. 2020. „Contract Grading and Peer Review.“ In *Ungrading: Why Rating Students Undermines Learning (and what to Do instead)*, hrsg. v. Susan D. Blum, 105–22. Morgantown: West Virginia University Press.
- Kautz, Christian und Katrin Billerbeck. 2018. „Kombination von Einzel- und Gruppenprüfung im Zwei-Stufen-Verfahren.“ In *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*, hrsg. v. Julia Gerick, Angela Sommer und Germa Zimmermann, 105–8. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Kearney, Sean. 2013. „Improving Engagement: The Use of ‚Authentic Self-and Peer-Assessment for Learning‘ to Enhance the Student Learning Experience.“ *Assessment & Evaluation in Higher Education* 38 (7): 875–91. doi:10.1080/02602938.2012.751963.
- Keck, Rudolf W. 1991. „Zensieren und Zertieren: Zur Kontroll- und Gratifikationsleistung der katholischen Pädagogik im jesuitischen Einflussbereich.“ In *Schülerbeurteilungen und Schulzeugnisse: Historische und systematische Aspekte*, hrsg. v. Johann G. von Hohenzollern und Max Liedtke, 69–88. Bad Heilbrunn: Klinkhardt.
- Keller, Gottfried. 1975 [1874]. *Kleider machen Leute: Novelle*. [Nachdr.] 7470. Stuttgart: Reclam.
- Kennedy, Declan. 2006. *Writing and Using Learning Outcomes: A Practical Guide*. Cork: University College Cork.
- Khan, Kamran Z., Sankaranarayanan Ramachandran, Kathryn Gaunt und Piyush Pushkar. 2013. „The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An Historical and Theoretical Perspective.“ *Medical teacher* 35 (9): e1437–46. doi:10.3109/0142159X.2013.818634.
- Kiener, Urs. 2012. „Wer sind die Studierenden an Fachhochschulen? Hinweise auf eine zunehmende Vielfalt.“ In *Lernendenorientierung: Studierende im Fokus*, hrsg. v. Tobias Zimmermann und Franziska Zellweger, 14–28. Forum Hochschuldidaktik und Erwachsenenbildung 3. Bern: hep.
- Kim, Kyung H. und Robert A. Pierce. 2013. „Convergent versus Divergent Thinking.“ In *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship*, hrsg. v. Elias G. Carayannis, 245–50. New York, NY: Springer New York.
- Kim, Rebecca H., Susan C. Day, Dylan S. Small, Christopher K. Snider, Charles A. L. Rareshide und Mitesh S. Patel. 2018. „Variations in Influenza Vaccination by

- Clinic Appointment Time and an Active Choice Intervention in the Electronic Health Record to Increase Influenza Vaccination.“ *JAMA network open* 1 (5): e181770. doi:10.1001/jamanetworkopen.2018.1770.
- Kinzie, Jillian. 2013. „Taking Stock of Capstones and Integrative Learning.“ *Peer Review* 15 (4): 27–30.
- Kluger, Avraham N. und Angelo DeNisi. 1996. „The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory.“ *Psychological Bulletin* 119 (2): 254–84. doi:10.1037/0033-2909.119.2.254.
- Koller, Marcel. 2002. *Was heißt ‚Faire Prüfung‘? Die wesentlichen rechtlichen Aspekte bei Prüfungen an schweizerischen Mittel- und Hochschulen*. St. Gallen: Salinas Press. Dissertation.
- Konersmann, Ralf. 2021. *Welt ohne Maß*. Originalausgabe. Frankfurt a. M.: S. Fischer.
- Konietzka, Dirk. 2009. „Berufsbildung im sozialen Wandel.“ In *Lehrbuch der Bildungssoziologie*, hrsg. v. Rolf Becker, 257–80. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Koskinen, Heli. 2007. „Evaluation of the Level of Difficulty of Patient Cases for Veterinary Problem-Solving Examination: A Preliminary Comparison of Three Taxonomies of Learning.“ *Journal of Veterinary Medical Education* 34 (2): 106–11. doi:10.3138/jvme.34.2.106.
- Kostorz, Peter. 2016. „Bewertungsmaßstäbe und Bezugsnormen bei der Notenvergabe unter der Lupe des Schulrechts – Was ist pädagogisch sinnvoll, was juristisch möglich?“ *Recht der Jugend und des Bildungswesens* 64 (2): 270–89. doi:10.5771/0034-1312-2016-2-270.
- Krathwohl, David R., Benjamin S. Bloom und Bertram B. Masia. 1975. *Taxonomie von Lernzielen im affektiven Bereich* [Original: *Taxonomy of Educational Objectives*]. Übersetzt von Helmut Dreesmann. Weinheim: Beltz.
- Krebs, René. 2004. „Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung.“ https://www.iml.unibe.ch/attachment/7/download/mc_anleitung.pdf. Zugriff 15.2.2024.
- Krebs, René. 2019. *Prüfen mit Multiple Choice: Kompetent planen, entwickeln, durchführen und auswerten*. Bern: Hogrefe.
- Kreis, Annelies. 2015. „Kollegiale Hospitation – Chancen und Realisierungsmöglichkeiten.“ In *Personalentwicklung in der Schule zwischen Fremdsteuerung und Selbstbestimmung*, hrsg. v. Katja Kansteiner und Christoph Stamann, 185–99. Bad Heilbrunn: Klinkhardt.
- Krist, Horst. 1999. „Die Integration intuitiven Wissens beim schulischen Lernen.“ *Zeitschrift für Pädagogische Psychologie* 13 (4): 191–206. doi:10.1024/1010-0652.13.4.191.
- Kronig, Winfried. 2007. *Die systematische Zufälligkeit des Bildungserfolgs: Theoretische Erklärungen und empirische Untersuchungen zur Lernentwicklung und zur Leistungsbewertung in unterschiedlichen Schulklassen*. Bern: Haupt Verlag.

- Krüger, Hans-Peter, Hrsg. 2007. *Hirn als Subjekt? Philosophische Grenzfragen der Neurobiologie*. Deutsche Zeitschrift für Philosophie Sonderband 15. Berlin: Akademie-Verlag.
- Kuppens, Toon, Russell Spears, Antony S. Manstead, Bram Spruyt und Matthew J. Easterbrook. 2018. „Educationism and the Irony of Meritocracy: Negative Attitudes of Higher Educated People towards the Less Educated.“ *Journal of Experimental Social Psychology* 76: 429–47. doi: 10.1016/j.jesp.2017.11.001.
- Lacey, Simon, Randall Stilla und K. Sathian. 2012. „Metaphorically Feeling: Comprehending Textural Metaphors Activates Somatosensory Cortex.“ *Brain and Language* 120 (3): 416–21. doi:10.1016/j.bandl.2011.12.016.
- Lamborn, S. D., N. S. Mounts, L. Steinberg und S. M. Dornbusch. 1991. „Patterns of Competence and Adjustment among Adolescents from Authoritative, Authoritarian, Indulgent, and Neglectful Families.“ *Child Development* 62 (5): 1049–65. doi:10.1111/j.1467-8624.1991.tb01588.x.
- Larsen, Malene R., Hanna B. Sommersel und Sogaard L. Larsen. 2013. „Evidence on Dropout Phenomena at Universities.“ <https://edudoc.ch/record/115243?ln=de>. Zugriff 15.2.2024.
- Laughlin, Patrick R. und Alan L. Ellis. 1986. „Demonstrability and Social Combination Processes on Mathematical Intellectual Tasks.“ *Journal of Experimental Social Psychology* 22 (3): 177–89. doi:10.1016/0022-1031(86)90022-3.
- Lepsinger, Richard und Anntoinette D. Lucia. 2009. *The Art and Science of 360-Degree Feedback*. 2. Auflage. Chichester, San Francisco CA: Wiley; Jossey-Bass.
- Lewis, Robert, Paul Berghoff und Pierette Pheeneey. 1999. „Focusing Students: Three Approaches for Learning Through Evaluation.“ *Innovative Higher Education* 23 (3): 181–96. doi:10.1023/A:1022994417651.
- Lienert, Gustav A. und Ulrich Raatz. 1998. *Testaufbau und Testanalyse*. 6. Auflage. Weinheim: Beltz Psychologie Verlags Union.
- Liessmann, Konrad P. 2006. *Die Theorie der Unbildung: Die Irrtümer der Wissensgesellschaft*. Wien: Zsolnay.
- Linder, Jeffrey A., Jason N. Doctor, Mark W. Friedberg, Harry Reyes Nieva, Caroline Birks, Daniella Meeker und Craig R. Fox. 2014. „Time of Day and the Decision to Prescribe Antibiotics.“ *JAMA Internal Medicine* 174 (12): 2029–31. doi: 10.1001/jamainternmed.2014.5225.
- Lipnevich, Anastasiya A. und Jeffrey K. Smith. 2009. „Effects of Differential Feedback on Students’ Examination Performance.“ *Journal of Experimental Psychology. Applied* 15 (4): 319–33. doi:10.1037/a0017841.
- Loughlin, Colin, Simon Lygo-Baker und Åsa Lindberg-Sand. 2021. „Reclaiming Constructive Alignment.“ *European Journal of Higher Education* 11 (2): 119–36. doi:10.1080/21568235.2020.1816197.
- Maasen, Sabine und Peter Weingart. 2008. „Unternehmerische Universität und neue Wissenschaftskultur.“ In *Wissenschaft unter Beobachtung: Effekte und Defekte von Evaluationen*, hrsg. v. Hildegard Matthies und Dagmar Simon, 141–60. Leviathan Sonderheft 24. Wiesbaden: VS Verlag für Sozialwissenschaften.

- Macnamara, Brooke N., David Z. Hambrick und Frederick L. Oswald. 2014. „Deliberate Practice and Performance in Music, Games, Sports, Education, and Professions: A Meta-Analysis.“ *Psychological Science* 25 (8): 1608–18. doi:10.1177/0956797614535810.
- Mahoney, Paige, Susie Macfarlane und Rola Ajjawi. 2019. „A Qualitative Synthesis of Video Feedback in Higher Education.“ *Teaching in Higher Education* 24 (2): 157–79. doi:10.1080/13562517.2018.1471457.
- Maier, Uwe, Marc Kleinknecht, Kerstin Metz und Thorsten Bohl. 2010. „Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben.“ *Beiträge zur Lehrerbildung* 28 (1): 84–96. doi:10.25656/01:13734.
- Mallinckrodt, Rebekka von und Angela Schattner. 2016. „Introduction.“ In *Sports and Physical Exercise in Early Modern Culture: New Perspectives on the History of Sports and Motion*, hrsg. v. Rebekka von Mallinckrodt und Angela Schattner, 1–17. London, New York: Routledge Taylor & Francis Group.
- Marriott, Chris und Menaka Abraham. 2023. „Labor-based Grading in Computer Science: A Student-Centered Practice.“ In *2023 ASEE Annual Conference & Exposition*, hrsg. v. American Society for Engineering Education.
- Marx, Alexandra. 2020. „Notenunabhängige Leistungserfassung und -beurteilung: Ansätze und Effekte.“ In *Lernen ohne Noten: Alternative Konzepte der Leistungsbeurteilung*, hrsg. v. Silvia-Iris Beutel und Hans A. Pant. 1. Auflage, 59–89. Stuttgart: Verlag W. Kohlhammer.
- McKeachie, Wilbert J. und Marilla D. Svinicki, Hrsg. 2006. *McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers*. 12. Auflage. College Teaching Series. Boston: Houghton Mifflin.
- Mercier, Hugo und Dan Sperber. 2011. „Why Do Humans Reason? Arguments for an Argumentative Theory.“ *The Behavioral and Brain Sciences* 34 (2): 57–74; discussion 74–111. doi:10.1017/S0140525X10000968.
- Mercier, Hugo und Dan Sperber. 2017. *The Enigma of Reason: A New Theory of Human Understanding*. Cambridge, Massachusetts: Harvard University Press.
- Metzger, Christoph und Charlotte Nüesch. 2004. *Fair prüfen: Ein Qualitätsleitfaden für Prüfende an Hochschulen*. Hochschuldidaktische Schriften 6. St. Gallen: IWP-HSP Universitätsinstitut für Wirtschaftspädagogik.
- Micceri, Theodore. 1989. „The Unicorn, the Normal Curve, and other Improbable Creatures.“ *Psychological Bulletin* 105 (1): 156–66. doi:10.1037/0033-2909.105.1.156.
- Mohammed, Susan und Brad C. Dumville. 2001. „Team Mental Models in a Team Knowledge Framework: Expanding Theory and Measurement across Disciplinary Boundaries.“ *Journal of Organizational Behavior* 22 (2): 89–106. doi:10.1002/job.86.
- Morgan, Whitney R. und Erik S. Wright. 2021. „Ten Simple Rules for Hitting a Home Run with Your Elevator Pitch.“ *PLoS Computational biology* 17 (3): e1008756. doi:10.1371/journal.pcbi.1008756.
- Moshman, David und Molly Geil. 1998. „Collaborative Reasoning: Evidence for Collective Rationality.“ *Thinking & Reasoning* 4 (3): 231–48. doi:10.1080/135467898394148.

- Moss, Pamela A. 1994. „Can there Be Validity without Reliability?“ *Educational Researcher* 23 (2): 5–12. doi: 10.3102/0013189X023002005.
- Moulton, Carol-Anne E., Adam Dubrowski, Helen Macrae, Brent Graham, Ethan Grober und Richard Reznick. 2006. „Teaching Surgical Skills: What Kind of Practice Makes Perfect? A Randomized, Controlled Trial.“ *Annals of Surgery* 244 (3): 400–409. doi:10.1097/01.sla.0000234808.85789.6a.
- Mueller, Claudia M. und Carol S. Dweck. 1998. „Praise for Intelligence Can Undermine Children’s Motivation and Performance.“ *Journal of Personality and Social Psychology* 75 (1): 33–52. doi:10.1037/0022-3514.75.1.33.
- Muller, Jerry Z. 2018. *The Tyranny of Metrics*. Princeton, Oxford: Princeton University Press.
- Müller, Andreas und Hartmut Ditton. 2014. „Feedback: Begriff, Formen und Funktionen.“ In *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder*, hrsg. v. Hartmut Ditton und Andreas Müller, 11–28. Münster, New York: Waxmann.
- Narciss, Susanne. 2006. *Informatives tutorielles Feedback: Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse*. Münster: Waxmann.
- Narciss, Susanne. 2020. „Feedbackstrategien für interaktive Lernaufgaben.“ In *Handbuch Bildungstechnologie: Konzeption und Einsatz digitaler Lernumgebungen*, hrsg. v. Helmut M. Niegemann und Armin Weinberger, 1–24. Berlin: Springer.
- Nathan, Mitchell J. und Anthony Petrosino. 2003. „Expert Blind Spot among Preservice Teachers.“ *American Educational Research Journal* 40 (4): 905–28. doi: 10.3102/00028312040004905.
- Neumann, Martin. 2019. „Praxisbeispiel Rechtswissenschaft: Mündliche Prüfungen in den Rechtswissenschaften – Baustellen, Herausforderungen, Potenziale und Handlungsempfehlungen aus rechtsdidaktischer Sicht.“ In *Kompetenzorientiert prüfen und bewerten an Universitäten: Didaktische Grundannahmen, rechtliche Rahmenbedingungen und praktische Handlungsempfehlungen*, hrsg. v. Susanne Frölich-Steffen, Hendrik den Ouden und Ursula Gießmann, 72–84. Opladen, Berlin, Toronto: Verlag Barbara Budrich.
- Nicht, Jörg und Thomas Müller. 2017. „Zwischen individueller Zurechnung und sozialer Hervorbringung: Vorüberlegungen zu einem sozialtheoretischen Verständnis von Leistung.“ In *Leistung – Anspruch und Scheitern*, hrsg. v. Carsten Büniger, Ralf Mayer, Sabrina Schröder und Britta Hoffarth, 61–83. Wittenberger Gespräche IV. Halle.
- Nicol, David. 2010. „From Monologue to Dialogue: Improving Written Feedback Processes in Mass Higher Education.“ *Assessment & Evaluation in Higher Education* 35 (5): 501–17. doi:10.1080/02602931003786559.
- Nietzsche, Friedrich. 1999 [1878]. *Kritische Studienausgabe, Bd. 2: Menschliches, Allzumenschliches*, hrsg. v. Giorgio Colli u. Mazzino Montinari. München: dtv.
- Niggli, Martina. 2023. „Focus Groups – a New Type of Exercise Classes Fostering Equality of Opportunities.“ Präsentation gehalten am *Jahreskongress 2023 der Schweizerischen Gesellschaft für Bildungsforschung (SGBF) und der Schweizerischen Gesellschaft für Lehrerinnen- und Lehrerbildung (SGL): Zeiten des Um-*

- bruchs? *Bildung zwischen Persistenz und Wandel*, Zürich, 28.-30. Juni. https://express.converia.de/frontend/index.php?page_id=24265&v=List&do=15&day=4470&ses=26243#anker_session_26243. Zugriff 17.2.2024.
- Niv, Yael und Geoffrey Schoenbaum. 2008. „Dialogues on Prediction Errors.“ *Trends in Cognitive Sciences* 12 (7): 265–72. doi:10.1016/j.tics.2008.03.006.
- Nölte, Björn und Philippe Wampfler. 2021. *Eine Schule ohne Noten: Neue Wege zum Umgang mit Lernen und Leistung*. Bern: hep.
- Offe, Claus. 1970. *Leistungsprinzip und industrielle Arbeit: Mechanismen der Statusverteilung in Arbeitsorganisationen der industriellen Leistungsgesellschaft*. Frankfurt a. M.: Europäische Verlagsanstalt.
- Orsmond, Paul und Stephen Merry. 2013. „The Importance of Self-Assessment in Students’ Use of Tutors’ Feedback: A Qualitative Study of High and Non-High Achieving Biology Undergraduates.“ *Assessment & Evaluation in Higher Education* 38 (6): 737–53. doi:10.1080/02602938.2012.697868.
- Ortner, Hanspeter. 2000. *Schreiben und Denken*. Germanistische Linguistik 214. Tübingen: Max Niemeyer Verlag.
- Ouden, Hanneke E. M. den, Peter Kok und Floris P. de Lange. 2012. „How Prediction Errors Shape Perception, Attention, and Motivation.“ *Frontiers in Psychology* 3. doi: 10.3389/fpsyg.2012.00548.
- Paivio, Allan. 1971. *Imagery and Verbal Processes*. New York: Holt Rinehart and Winston.
- Panadero, Ernesto und Anders Jonsson. 2013. „The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review.“ *Educational Research Review* 9: 129–44. doi 10.1016/j.edurev.2013.01.002.
- Parks, Melissa. 2023. „Contract Grading.“ *The National Teaching & Learning Forum* 32 (4): 1–4. doi:10.1002/ntlf.30367.
- Paul, Annie M. 2021. *The Extended Mind: The Power of Thinking Outside the Brain*. Boston, New York: Houghton Mifflin Harcourt.
- Pautsch, Arne und Anja Dillenburger. 2011. *Kompandium zum Hochschul- und Wissenschaftsrecht*. De Gruyter-Handbuch. Berlin u. a. de Gruyter.
- Perrez, Meinrad, Günter L. Huber und Karlheinz A. Geissler. 2001. „Psychologie der pädagogischen Interaktion.“ In *Pädagogische Psychologie: Ein Lehrbuch*, hrsg. v. Andreas Krapp und Bernd Weidenmann. 4., vollständig überarbeitete Auflage, 357–413. Weinheim: Beltz Psychologie Verlags Union.
- Philipp, Maik. 2023a. „Epistemisches Schreiben – was es braucht, was es leistet, was sich gerade ändert.“ Vortrag zum 20-jährigen Jubiläum des Schreibzentrums der PH Zürich. https://www.researchgate.net/publication/374001211_Epistemisches_Schreiben_-_was_es_braucht_was_es_leistet_was_sich_gerade_ändert_Epistemic_Writing_-_Its_Requirements_Its_Outcomes_Its_Current_Challenges. Zugriff 30.12.2023.
- Philipp, Maik. 2023b. „Schreiben an Hochschulen in Zeiten von KI – das tiefe Nachdenken wird zukünftig noch nötiger.“ <https://blog.phzh.ch/zhe/schreiben-an-hochschulen-in-zeiten-von-ki-das-tiefe-nachdenken-wird-zukuenftig-noch-noetiger/>. Zugriff 27.12.2023.

- Philipps, Lothar. 1978. *Testaufgaben in der Rechtswissenschaft: Konstruktionsprinzipien und Auswertung durch den Computer*. Heidelberg, Karlsruhe: Mueller Juristischer Verlag.
- Phillips, Paul E. M., Garret D. Stuber, Michael L. A. V. Heien, R. M. Wightman und Regina M. Carelli. 2003. „Subsecond Dopamine Release Promotes Cocaine Seeking.“ *Nature* 422 (6932): 614–18. doi:10.1038/nature01476.
- Philpot, Lindsey M., Bushra A. Khokhar, Daniel L. Roellinger, Priya Ramar und Jon O. Ebbert. 2018. „Time of Day Is Associated with Opioid Prescribing for Low Back Pain in Primary Care.“ *Journal of General Internal Medicine* 33 (11): 1828–30. doi:10.1007/s11606-018-4521-8.
- Pine, Alex, Noa Sadeh, Aya Ben-Yakov, Yadin Dudai und Avi Mendelsohn. 2018. „Knowledge Acquisition Is Governed by Striatal Prediction Errors.“ *Nature Communications* 9 (1): 1673. doi:10.1038/s41467-018-03992-5.
- Platon. ca. 400 v. Chr. „Phaidros: In der Übersetzung von Friedrich Daniel Ernst Schleiermacher.“ <https://www.projekt-gutenberg.org/platon/platowr1/phaidro1.html>. Zugriff 26.12.2023.
- Pohl, Thorsten. 2007. *Studien zur Ontogenese wissenschaftlichen Schreibens*. Germanistische Linguistik 271. Tübingen: Niemeyer. [Dissertation, Universität Siegen 2005].
- Pospeschill, Markus. 2022. *Testtheorie, Testkonstruktion, Testevaluation: Mit 71 Abbildungen und 77 Fragen zur Wiederholung*. 2., aktualisierte Auflage. utb Psychologie. München: Ernst Reinhardt Verlag.
- Power, Michael. 1997. *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.
- Prahl, Hans-Werner. 1995. „Prüfungen.“ In *Ausbildung und Sozialisation in der Hochschule*, hrsg. v. Ludwig Huber, 438–50. Enzyklopädie Erziehungswissenschaft 10, hrsg. von Dieter Lenzen. Stuttgart, Dresden: Klett.
- Pulvermüller, Friedemann, Olaf Hauk, Vadim V. Nikulin und Risto J. Ilmoniemi. 2005. „Functional Links between Motor and Language Systems.“ *The European Journal of Neuroscience* 21 (3): 793–97. doi:10.1111/j.1460-9568.2005.03900.x.
- Race, Phil. 2001. „A Briefing on Self, Peer and Group Assessment.“ https://phil-race.co.uk/wp-content/uploads/Self_peer_and_group_assessment.pdf. Zugriff 2.6.2023.
- Race, Phil. 2014. *Making Learning Happen: A Guide for Post-Compulsory Education*. 3. Auflage. Los Angeles u.a.: SAGE.
- Rattan, Aneeta, Catherine Good und Carol S. Dweck. 2012. „‘It’s Ok – not Everyone Can Be Good at Math’: Instructors with an Entity Theory Comfort (and Demotivate) Students.“ *Journal of Experimental Social Psychology* 48 (3): 731–37. doi:10.1016/j.jesp.2011.12.012.
- Ray, Partha P. 2023. „ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope.“ *Internet of Things and Cyber-Physical Systems* 3: 121–54. doi: 10.1016/j.iotcps.2023.04.003.

- Reddy, Y. M. und Heidi Andrade. 2010. „A Review of Rubric Use in Higher Education.“ *Assessment & Evaluation in Higher Education* 35 (4): 435–48. doi:10.1080/02602930902862859.
- Redelmeier, Donald A. und Simon D. Baxter. 2009. „Holiday Review. Rainy Weather and Medical School Admission Interviews.“ *CMAJ: Canadian Medical Association Journal = Journal de l'association medicale Canadienne* 181 (12): 933. doi:10.1503/cmaj.091546.
- Reh, Sabine, Kathrin Berdelmann und Joachim Scholz. 2015. „Der Ehrtrieb und unterrichtliche Honorierungspraktiken im Schulwesen um 1800: Die Entstehung des Leistungs-Dispositivs.“ In *Leistung*, hrsg. v. Alfred Schäfer und Christiane Thompson, 37–60. Paderborn: Ferdinand Schöningh.
- Reinmann, Gabi. 2015. „Lehrkompetenzen von Hochschullehrern: Kritik des Kompetenzbegriffs in fünf Thesen.“ In *Lehrkompetenzen in der wissenschaftlichen Weiterbildung*, hrsg. v. Olaf Hartung und Marguerite Rumpf, 17–36. Wiesbaden: Springer Fachmedien.
- Reinmann, Gabi. 2018. „24-Stunden-Hausarbeit.“ In *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*, hrsg. v. Julia Gerick, Angela Sommer und Gernot Zimmermann, 20–24. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Renkl, Alexander. 1994. *Träges Wissen: Die ‚unerklärliche‘ Kluft zwischen Wissen und Handeln. (Forschungsbericht Nr. 41)*. München: LMU München: Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie.
- Richter, Ingo. 2007. „Rechtliche Grundlagen von Bildungskontrakten.“ In *Bildungs- und Erziehungskontrakte als Instrumente von Schulentwicklung: Expertise*, hrsg. v. Marianne Horstkemper, Joachim Ludwig und Wilfried Schubarth, 313–23. Bildungsforschung 21. Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Ricken, Norbert. 2018. „Konstruktionen der ‚Leistung‘. Zur (Subjektivierungs-)Logik eines Konzepts.“ In *Leistung als Paradigma: Zur Entstehung und Transformation eines pädagogischen Konzepts*, hrsg. v. Sabine Reh und Norbert Ricken, 43–60. Wiesbaden: Springer Fachmedien.
- Rietveld, Cornelius A., Sarah E. Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W. Martin, Harm-Jan Westra et al. 2013. „GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment.“ *Science* 340 (6139): 1467–71. doi:10.1126/science.1235488.
- Robins, Richard W. und Jennifer L. Pals. 2002. „Implicit Self-Theories in the Academic Domain: Implications for Goal Orientation, Attributions, Affect, and Self-Esteem Change.“ *Self and Identity* 1 (4): 313–36. doi:10.1080/15298860290106805.
- Roediger, Henry L. und Jeffrey D. Karpicke. 2006. „Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention.“ *Psychological Science* 17 (3): 249–55. doi:10.1111/j.1467-9280.2006.01693.x.
- Roediger, Henry L. und Kathleen B. McDermott. 1995. „Creating False Memories: Remembering Words not Presented in Lists.“ *Journal of Experimental Psycholo-*

- gy: *Learning, Memory, and Cognition* 21 (4): 803–14. doi: <https://doi.org/10.1037/0278-7393.21.4.803>.
- Röhl, Tobias. 2023. „Das Ende der Hausaufgaben?“ *NZZ*, 12. April.
- Rosenbusch, Christa. 2004. „Nonverbale Kommunikation in schulischen Beratungssituationen.“ In *Körpersprache und Pädagogik: Das Handbuch*, hrsg. v. Heinz S. Rosenbusch und Otto Schober. 4., überarbeitete Auflage, 185–209. Baltmannsweiler: Schneider Verlag Hohengehren.
- Rosenthal, Robert und Lenore Jacobson. 1968. „Pygmalion in the Classroom.“ *The Urban Review* 3 (1): 16–20. doi:10.1007/BF02322211.
- Rost, Detlef H., Jörn R. Sparfeldt und Susanne Buch, Hrsg. 2018. *Handwörterbuch pädagogische Psychologie*. 5., überarbeitete und erweiterte Auflage. Weinheim, Basel: Beltz.
- Roth, Philip, Craig BeVier, Fred Switzer und Jeffery Schippmann. 1996. „Meta-Analyzing the Relationship Between Grades and Job Performance.“ *Journal of Applied Psychology* 81 (5): 548–56. doi: 10.1037/0021-9010.81.5.548.
- Rubie-Davies, Christine. 2016. „High and Low Expectation Teachers. The Importance of the Teacher Factor.“ In *Interpersonal and Intrapersonal Expectancies*, hrsg. v. Sławomir Trusz und Przemysław Bąbel, 145–56. London: Taylor and Francis.
- Ruf, Urs und Peter Gallin. 2003. *Austausch unter Ungleichen: Grundzüge einer interaktiven und fächerübergreifenden Didaktik*. 2. Auflage. Dialogisches Lernen in Sprache und Mathematik 1. Seelze-Velber: Kallmeyer.
- Ryan, Allison M. und Huiyoung Shin. 2011. „Help-Seeking Tendencies During Early Adolescence: An Examination of Motivational Correlates and Consequences for Achievement.“ *Learning and Instruction* 21 (2): 247–56. doi:10.1016/j.learninstruc.2010.07.003.
- Ryan, Mary. 2012. „Conceptualising and Teaching Discursive and Performative Reflection in Higher Education.“ *Studies in Continuing Education* 34 (2): 207–23. doi:10.1080/0158037X.2011.611799.
- Ryan, Richard M. und Edward L. Deci. 2000. „Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being.“ *The American Psychologist* 55 (1): 68–78. doi:10.1037/0003-066X.55.1.68.
- Ryan, Richard M. und Edward L. Deci. 2017. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York, London: The Guilford Press.
- Sacher, Werner. 2014. *Leistungen entwickeln, überprüfen und beurteilen: Bewährte und neue Wege für die Primar- und Sekundarstufe*. 6., überarbeite und erweiterte Auflage Bad Heilbrunn: Klinkhardt.
- Sadler, D. R. 1989. „Formative Assessment and the Design of Instructional Systems.“ *Instructional Science* 18 (2): 119–44.
- Sanchez, Albert, Ernest Redondo, David Fonseca und Isidro Navarro. 2014. „Academic Performance Assessment using Augmented Reality in Engineering Degree Course.“ Paper presented at 2014 IEEE Frontiers in Education Conference (FIE). Madrid, 22.-25.10.2014. Madrid: IEEE Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/FIE.2014.7044238>. Zugriff 17.2.2024.

- Sandel, Michael J. 2021. *The Tyranny of Merit: What's Become of the Common Good?* London: Penguin Books.
- Sapolsky, Robert M. 2004. *Why Zebras Don't Get Ulcers*. 3. Auflage. New York: Holt.
- Sapolsky, Robert M. 2018. *Behave: The Biology of Humans at Our Best and Worst*. London: Vintage.
- Sauder, Michael und Wendy N. Espeland. 2009. „The Discipline of Rankings: Tight Coupling and Organizational Change.“ *American Sociological Review* 74 (1): 63–82.
- Saunders, Samuel. 2022. „The ‚Reflective Elevator Pitch‘.“ In *100 Ideas for Active Learning*, hrsg. v. Tab Betts und Paolo Oprandi, 481–86: The Active Learning Network. <https://openpress.sussex.ac.uk/ideasforactivelearning/open/download?type=pdf>. Zugriff 15.2.2024.
- Schacter, Daniel L. 2021. *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Aktualisierte Auflage. Boston: Mariner Books.
- Schacter, Daniel L. und Donna R. Addis. 2007. „The Cognitive Neuroscience of Constructive Memory: Remembering the Past and Imagining the Future.“ *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362 (1481): 773–86. doi:10.1098/rstb.2007.2087.
- Schäfer, Alfred. 2015. *Schulische Leistungsdiskurse: Zwischen Gerechtigkeitsversprechen und pharmazeutischem Hirndoping*. Paderborn: Verlag Ferdinand Schöningh.
- Schäfer, Thomas. 2016. *Methodenlehre und Statistik*. Wiesbaden: Springer Fachmedien.
- Schiefele, Hans. 2000. „Befunde – Fortschritte – neue Fragen.“ In *Interesse und Lernmotivation: Untersuchungen zu Entwicklung, Förderung und Wirkung*, hrsg. v. Ulrich Schiefele und Klaus-Peter Wild, 227–41. Münster, München, Berlin: Waxmann.
- Schiefele, Ulrich und Inge Schreyer. 1994. „Intrinsische Lernmotivation und Lernen. Ein Überblick zu Ergebnissen der Forschung.“ *Zeitschrift für Pädagogische Psychologie* 8 (1): 1–13.
- Schiefner, Mandy und Markus Weil. 2010. „Global übertragbar oder doch kontextgebunden.“ *Education Permanente* 44 (2): 38–39.
- Schimank, Uwe. 2010. *Handeln und Strukturen: Einführung in die akteurtheoretische Soziologie*. 4., völlig überarbeitete Auflage. Grundlagentexte Soziologie. Weinheim, München: Juventa.
- Schlie, Frank. 1988. „Die Vielfalt der Leistungsbegriffe.“ In *Krise der Leistungsgesellschaft? Empirische Analysen zum Engagement in Arbeit, Familie und Politik*, hrsg. v. Karl O. Hondrich, Jürgen Schumacher, Klaus Arzberger, Frank Schlie und Christian Stegbauer, 50–67. Opladen: Westdeutscher Verlag.
- Schneider, Michael und Franzis Preckel. 2017. „Variables Associated with Achievement in Higher Education: A Systematic Review of Meta-Analyses.“ *Psychological Bulletin* 143 (6): 565–600. doi:10.1037/bul0000098.
- Schön, Donald A. 1991. *The Reflective Practitioner: How Professionals Think in Action*. Farnham, Surrey: Ashgate.

- Schultz, W., L. Tremblay und J. R. Hollerman. 2000. „Reward Processing in Primate Orbitofrontal Cortex and Basal Ganglia.“ *Cerebral Cortex* 10 (3): 272–84. doi: 10.1093/cercor/10.3.272.
- Schulz von Thun, Friedemann. 1981. *Störungen und Klärungen: Allgemeine Psychologie der Kommunikation: Miteinander Reden 1*. Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.
- Schunk, Dale H. 2020. *Learning Theories: An Educational Perspective*. 8. Auflage. Hoboken, NJ: Pearson.
- Schütz, Gabriela, Heinrich W. Ursprung und Ludger Wössmann. 2008. „Education Policy and Equality of Opportunity.“ *Kyklos* 61 (2): 279–308. doi:10.1111/j.1467-6435.2008.00402.x.
- Schwandt, Thomas. 2007. *The SAGE Dictionary of Qualitative Inquiry*. Thousand Oaks CA: SAGE Publications.
- Scouller, Karen. 1998. „The Influence of Assessment Method on Students’ Learning Approaches: Multiple Choice Question Examination vs. Essay Assignment.“ *Higher Education* 35 (4): 453–72. doi: 10.1023/A:1003196224280.
- Shaffer, David W. und Mitchel Resnick. 1999. „Thick Authenticity: New media and Authentic Learning.“ *Journal of Interactive Learning Research* 10 (2): 195–215.
- Shapiro, Lawrence und Shannon Spaulding. 2021. „Embodied Cognition: The Stanford Encyclopedia of Philosophy.“ <https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/>. Zugriff 30.3.2023.
- Shore, Cris und Susan Wright. 2015. „Audit Culture Revisited.“ *Current Anthropology* 56 (3): 421–44. doi:10.1086/681534.
- Shulman, Lee S. 1970. „Cognitive Learning and the Educational Process.“ *Journal of Medical Education* 45 (11): 90–100.
- Simonson, Itamar und Peter Nye. 1992. „The Effect of Accountability on Susceptibility to Decision Errors.“ *Organizational Behavior and Human Decision Processes* 51 (3): 416–46. doi:10.1016/0749-5978(92)90020-8.
- Sorensen-Unruh, Clarissa. 2020. „A STEM Ungrading Case Study: A Reflection on First-Time Implementation in Organic Chemistry II.“ In *Ungrading: Why Rating Students Undermines Learning (and what to Do instead)*, hrsg. v. Susan D. Blum, 140–60. Morgantown: West Virginia University Press.
- Spannagel, Christian. 2023a. „ChatGPT und die Zukunft des Lernens: Evolution statt Revolution.“ <https://hochschulforumdigitalisierung.de/blog/chatgpt-und-die-zukunft-des-lernens-evolution-statt-revolution/>. Zugriff 3.1.2024.
- Spannagel, Christian. 2023b. „Rules for Tools: Version 2.2.“ <https://csp.uber.space/phhd/rulesfortools.pdf>. Zugriff 3.1.2024.
- Spruit, Edward N., Guido P. H. Band und Jaap F. Hamming. 2015. „Increasing Efficiency of Surgical Training: Effects of Spacing Practice on Skill Acquisition and Retention in Laparoscopy Training.“ *Surgical Endoscopy* 29 (8): 2235–43. doi:10.1007/s00464-014-3931-x.
- Stangl, Werner. 2023. „Die konstitutiven Antinomien des Lehrerhandelns.“ <https://arbeitsblaetter.stangl-taller.at/LEHREN/Antinomien-Lehrerhandeln.shtml>. Zugriff 16.3.2023.

- Stavrova, Olga, Tila Pronk und Michail D. Kokkoris. 2019. „Choosing Goals that Express the True Self: A Novel Mechanism of the Effect of Self-Control on Goal Attainment.“ *European Journal of Social Psychology* 49 (6): 1329–36. doi:10.1002/ejsp.2559.
- Steinhoff, Torsten. 2007. *Wissenschaftliche Textkompetenz: Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten*. Germanistische Linguistik 280. Tübingen: Niemeyer.
- Stevens, Dannelle D. und Antonia Levi. 2013. *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. 2. Auflage [Kindle-Ausgabe]. Sterling VA: Stylus.
- Strathern, Marilyn. 2000. „The Tyranny of Transparency.“ *British Educational Research Journal* 26 (3): 309–21. doi:10.1080/713651562.
- Studiengangsteam Data Science. 2022. „Ausbildungskonzept Studiengang Data Science FHNW.“ <https://www.fhnw.ch/plattformen/sgds/wp-content/uploads/sites/339/Ausbildungskonzept-SG-DS-HS22-Sign.pdf>. Zugriff 9.8.2023.
- Stumm, Sophie von und Robert Plomin. 2015. „Socioeconomic Status and the Growth of Intelligence from Infancy Through Adolescence.“ *Intelligence* 48: 30–36. doi:10.1016/j.intell.2014.10.002.
- Sun, Bo, Uzoma Chikwem und Donald Nyingifa. 2019. „VRLearner: A Virtual Reality Based Assessment Tool in Higher Education.“ Paper presented at *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. Osaka, 23.-27. März 2019. Osaka: IEEE Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/VR.2019.8798129>. Zugriff 17.2.2024.
- Sweller, John. 1988. „Cognitive Load During Problem Solving: Effects on Learning.“ *Cognitive Science* 12 (2): 257–85. doi:10.1207/s15516709cog1202_4.
- Tai, Joanna, Rola Ajjawi, David Boud, Phillip Dawson und Ernesto Panadero. 2018. „Developing Evaluative Judgement: Enabling Students to Make Decisions about the Quality of Work.“ *Higher Education* 76 (3): 467–81. doi:10.1007/s10734-017-0220-3.
- Taylor, Catherine. 1994. „Assessment for Measurement or Standards: The Peril and Promise of Large-Scale Assessment Reform.“ *American Educational Research Journal* 31 (2): 231–62. doi:10.2307/1163308.
- Terhart, Ewald. 2019. *Didaktik: Eine Einführung*. Ditzingen: Reclam.
- Thaler, Richard. 1980. „Toward a Positive Theory of Consumer Choice.“ *Journal of Economic Behavior & Organization* 1 (1): 39–60. doi:10.1016/0167-2681(80)90051-7.
- Theobald, Maria. 2021. „Self-Regulated Learning Training Programs Enhance University Students' Academic Performance, Self-regulated Learning Strategies, and Motivation: A Meta-Analysis.“ *Contemporary Educational Psychology* 66. doi:10.1016/j.cedpsych.2021.101976.
- Thomann, Geri. 2017a. „Grundlagen der Beratung für die Hochschullehre.“ In *Zwischen Beraten und Dozieren: Praxis, Reflexion und Anregungen für die Hochschullehre*, hrsg. v. Geri Thomann, Monique Honegger und Peter Suter, 2., überarbeitete und erweiterte Auflage, 12–32. Forum Hochschuldidaktik und Erwachsenenbildung 2. Bern: hep.

- Thomann, Geri. 2017b. „Instrument 3: Der Beratungskontrakt.“ In *Zwischen Beraten und Dozieren: Praxis, Reflexion und Anregungen für die Hochschullehre*, hrsg. v. Geri Thomann, Monique Honegger und Peter Suter, 2., überarbeitete und erweiterte Auflage, 62–65. Forum Hochschuldidaktik und Erwachsenenbildung 2. Bern: hep.
- Thomann, Geri. 2017c. „Kompetenzorientierung und Bildung auf Tertiärstufe: drei unterschiedliche Sichtweisen.“ *Case Management* 12 (4): 148–51.
- Thomann, Geri. 2019. *Ausbildung der Auszubildenden: Professionelles Handeln in der Erwachsenenbildung und Weiterbildung*. Unter Mitarbeit von E. Haberzeth, M. Honegger, F. Zellweger und T. Zimmermann. 5., vollständig überarbeitete und erweiterte Auflage. Bern: hep.
- Thomann, Geri, Monique Honegger und Peter Suter, Hrsg. 2017. *Zwischen Beraten und Dozieren: Praxis, Reflexion und Anregungen für die Hochschullehre*. 2., überarbeitete und erweiterte Auflage. Forum Hochschuldidaktik und Erwachsenenbildung 2. Bern: hep.
- Thomann, Geri und Anja Pawelleck. 2013. *Studierende beraten*. UTB Schlüsselkompetenzen 4012. Opladen, Toronto: Verlag Barbara Budrich.
- Tomasello, Michael, Ann C. Kruger und Hilary H. Ratner. 1993. „Cultural Learning.“ *The Behavioral and Brain Sciences* 16 (3): 495–511. doi:10.1017/S0140525X0003123X.
- Trapmann, Sabrina, Benedikt Hell, Sonja Weigand und Heinz Schuler. 2007. „Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse.“ *Zeitschrift für Pädagogische Psychologie* 21 (1): 11–27. doi:10.1024/1010-0652.21.1.11.
- Trigwell, Keith und Michael Prosser. 2020. *Exploring University Teaching and Learning: Experience and Context*. 1. Auflage. Cham: Springer International Publishing; Imprint Palgrave Pivot.
- Tyler, Ralph W. 1949. *Basic Principles of Curriculum and Instruction*. Chicago: The University of Chicago Press.
- Ulanoff, Sharon H., Joan C. Fingon und Dolores Beltrán. 2009. „Using Case Studies to Assess Candidates' Knowledge and Skills in a Graduate Reading Program.“ *Teacher Education Quarterly* 36 (2): 125–42.
- van der Linden, Wim J. und Ronald K. Hambleton. 1997. „Item Response Theory: Brief History, Common Models, and Extensions.“ In *Handbook of Modern Item Response Theory*, hrsg. v. Wim J. van der Linden und Ronald K. Hambleton, 1–28. New York, Berlin, Heidelberg: Springer.
- VanLehn, Kurt, Stephanie Siler, Charles Murray, Takashi Yamauchi und William B. Baggett. 2003. „Why Do only Some Events Cause Learning During Human Tutoring?“ *Cognition and Instruction* 21 (3): 209–49. doi:10.1207/S1532690XC12103_01.
- Verheyen, Nina. 2014. „Die soziale Konstruktion individueller Leistung: Forschungsperspektiven zwischen Geschichts- und Sozialwissenschaften.“ *Neue Politische Literatur* 59 (1): 63–87. doi: 10.3726/91500_63.
- Verheyen, Nina. 2018. *Die Erfindung der Leistung*. München: Hanser Berlin.

- Vermunt, Jan D. und Vincent Donche. 2017. „A Learning Patterns Perspective on Student Learning in Higher Education: State of the Art and Moving Forward.“ *Educational Psychology Review* 29 (2): 269–99. doi:10.1007/s10648-017-9414-6.
- Vierlinger, Rupert. 1999. *Leistung spricht für sich selbst: direkte Leistungsvorlage' (Portfolios) statt Ziffernzensuren und Notenfetischismus*. Heinsberg: Dieck.
- Villarroel, Verónica, Susan Bloxham, Daniela Bruna, Carola Bruna und Constanza Herrera-Seda. 2018. „Authentic Assessment: Creating a Blueprint for Course Design.“ *Assessment & Evaluation in Higher Education* 43 (5): 840–54. doi:10.1080/02602938.2017.1412396.
- Vlach, Haley A. und Catherine M. Sandhofer. 2012. „Distributing Learning over Time: The Spacing Effect in Children's Acquisition and Generalization of Science Concepts.“ *Child Development* 83 (4): 1137–44. doi:10.1111/j.1467-8624.2012.01781.x.
- von der Embse, Nathaniel, Dane Jester, Devlina Roy und James Post. 2018. „Test Anxiety Effects, Predictors, and Correlates: A 30-Year Meta-Analytic Review.“ *Journal of Affective Disorders* 227: 483–93. doi: 10.1016/j.jad.2017.11.048.
- Wagner, Anthony D., Anat Maril und Daniel L. Schacter. 2000. „Interactions Between Forms of Memory: When Priming Hinders New Episodic Learning.“ *Journal of Cognitive Neuroscience*, 12 Suppl 2: 52–60. doi: 10.1162/089892900564064.
- Wahl, Diethelm. 2001. „Nachhaltige Wege vom Wissen zum Handeln.“ *Beiträge zur Lehrerbildung* 19 (2): 157–74. doi:10.25656/01:13453.
- Wahl, Diethelm. 2011. „Der Advance Organizer: Einstieg in eine Lernumgebung.“ In *Lehren und Lernen im Unterricht*, hrsg. v. Sandra Brandt, 185–202. Professionswissen für Lehrerinnen und Lehrer 2. Baltmannsweiler: Schneider Verlag Hohengehren.
- Walden, Victoria G. 2022. „Using the ‚Unessay‘ as Active Co-Creation of Marking Criteria.“ In *100 Ideas for Active Learning*, hrsg. v. Tab Betts und Paolo Oprandi, 432–38: The Active Learning Network. <https://openpress.sussex.ac.uk/ideas/foractivelearning/open/download?type=pdf>. Zugriff 15.2.2024.
- Walvoord, Barbara E. und Virginia J. Anderson. 1998. *Effective Grading: A Tool for Learning and Assessment*. San Francisco: Jossey-Bass.
- Walzik, Sebastian. 2012. *Kompetenzorientiert prüfen: Leistungsbewertung an der Hochschule in Theorie und Praxis*. UTB Schlüsselkompetenzen 3777. Opladen: Budrich.
- Wampfler, Philippe. 2021. „Können Programme gute Texte schreiben? Streifzug durch die Welt der schreibenden Maschinen.“ *Sprachspiegel* 77 (5): 132–42.
- Wampfler, Philippe. 2022. „Grundlagenartikel: Umgang mit KI-Programmen im Schreibunterricht.“ <https://schulesocialmedia.com/2022/10/15/grundlagenartikel-umgang-mit-ki-programmen-im-schreibunterricht/>. Zugriff 29.12.2023.
- Wason, Peter C. 1966. „Reasoning.“ In *New Horizons in Psychology*, hrsg. v. Brian M. Foss, 135–51. Harmondsworth: Penguin.
- Wegner, Elisabeth und Matthias Nückles. 2012. „Mit Widersprüchen umgehen lernen: Reflektiertes Entscheiden als hochschuldidaktische Kompetenz.“ In *Lernwelt*

- Universität: Entwicklung von Lehrkompetenz in der Hochschullehre*, hrsg. v. Rudolf Egger und Marianne Merkt, 63–81. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Weil, Markus. 2020. „Rollengestaltung in der Hochschullehre.“ In *Grundlagen der Hochschullehre: Teaching in Higher Education*, hrsg. v. Sandra Hummel, 83–108. Doing Higher Education. Wiesbaden, Heidelberg: Springer VS.
- Weinert, Franz E. 1994. „Lernen lernen und das eigene Lernen verstehen.“ In *Verstehen: Psychologischer Prozess und didaktische Aufgabe*, hrsg. v. Kurt Reusser und Marianne Reusser-Weyeneth, 183–205. Bern: Huber.
- Weinert, Franz E. 2001. „Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit.“ In *Leistungsmessungen in Schulen*, hrsg. v. Franz E. Weinert, 17–31. Weinheim, Basel: Beltz.
- Weitzel, Julia, Anke Timmann, Melanie Franz-Özdemir, Cornelia Grunert, Jürgen Reimann, Anna-Linda Sachse, Sara Salzmann, Petra Weiss und Karin Wessel. 2022. „Dialogische Feedback- und Evaluationsverfahren für die Hochschulentwicklung.“ In *Hochschullehre als Gemeinschaftsaufgabe: Akteur:innen und Fachkulturen in der lernenden Organisation*, hrsg. v. Nora Leben, Katja Reincke und Ulrike Sonntag, 107–19. Blickpunkt Hochschuldidaktik 139. Bielefeld: wbv.
- Werner, Julia, Christian Ebel, Christian Spannagel und Stephan Bayer. 2018. „Flipped Classroom – Zeit für deinen Unterricht.“ In *Flipped Classroom – Zeit für deinen Unterricht: Praxisbeispiele, Erfahrungen und Handlungsempfehlungen*, hrsg. v. Julia Werner, Christian Ebel, Christian Spannagel und Stephan Bayer, 13–16. Gütersloh: Verlag Bertelsmann Stiftung.
- Westermann, Katharina und Nikol Rummel. 2012. „Delaying Instruction: Evidence from a Study in a University Relearning Setting.“ *Instructional Science* 40 (4): 673–89. doi:10.1007/s11251-012-9207-8.
- Whalley, Jacqueline L., Raymond Lister, Errol Thompson, Tony Clear, Phil Robbins, P. K. A. Kumar und Christine Prasad. 2006. „An Australasian Study of Reading and Comprehension Skills in Novice Programmers, Using the Bloom and SOLO Taxonomies.“ In *Computing Education 2006: Proceedings of the Eighth Australasian Computing Education Conference (ACE2006)*, Hobart, Tasmania, Australia 16–19 January 2006, hrsg. v. Denise Tolhurst, unpaginiert.
- Whitehead, Alfred N. 1929. *The Aims of Education and Other Essays*. New York: Free Press.
- Wieman, Carl E., Georg W. Rieger und Cynthia E. Heiner. 2014. „Physics Exams that Promote Collaborative Learning.“ *The Physics Teacher* 52 (1): 51–53. doi:10.1119/1.4849159.
- Wiggins, Grant P. 1993. *Assessing Student Performance: Exploring the Purpose and Limits of Testing*. San Francisco: Jossey-Bass.
- Wiggins, Grant P. und Jay McTighe. 2005. *Understanding by Design*. 2., erweiterte Auflage. Alexandria VA: Association for Supervision and Curriculum Development.
- Wild, Elke, Manfred Hofer und Reinhard Pekrun. 2001. „Psychologie des Lernens.“ In *Pädagogische Psychologie: Ein Lehrbuch*, hrsg. v. Andreas Krapp und Bernd

- Weidenmann. 4., vollständig überarbeitete Auflage, 207–70. Weinheim: Beltz Psychologie Verlags Union.
- Winstone, Naomi E., Robert A. Nash, James Rowntree und Michael Parker. 2017. „It'd Be Useful, but I Wouldn't Use it': Barriers to University Students' Feedback Seeking and Recipience.“ *Studies in Higher Education* 42 (11): 206–41. doi:10.1080/03075079.2015.1130032.
- Winter, Felix. 2015. *Lerndialog statt Noten: Neue Formen der Leistungsbeurteilung*. Weinheim, Basel: Beltz.
- Wirtz, Markus A. und Franz Caspar. 2002. *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.
- Wolf, Maryanne. 2007. *Proust and the Squid: The Story and Science of the Reading Brain*. New York NY: Harper.
- Wolter, Stefan C., Maria A. Cattaneo, Stefan Denzler, Andrea Diem, Stefanie Hof, Ramona Meier und Chantal Oggenfuss. 2018. *Bildungsbericht Schweiz 2018*. Aarau: SKBF/CSRE.
- Yorke, Mantz. 2003. „Formative Assessment in Higher Education: Moves towards Theory and the Enhancement of Pedagogic Practice.“ *Higher Education* 45 (4): 477–501. doi:10.1023/A:1023967026413.
- Zachrau, Sebastian. 2023. „ChatGPT zeigt uns, wie dumm unsere Prüfungen sind.“ *Zeit Campus*, 17. April 2023. <https://www.zeit.de/campus/2023-04/chatgpt-ki-uni-versitaet-pruefungen-schummeln>. Zugriff 3.1.2024.
- Zhang, Jiawei. 2019. „Basic Neural Units of the Brain: Neurons, Synapses and Action Potential.“ <https://arxiv.org/abs/1906.01703>. Zugriff 24.3.2023.
- Ziegenspeck, Jörg. 1999. *Handbuch Zensur und Zeugnis in der Schule: Historischer Rückblick, allgemeine Problematik, empirische Befunde und bildungspolitische Implikationen; ein Studien- und Arbeitsbuch*. Unter Mitarbeit von Jens Lehmann. Bad Heilbrunn: Klinkhardt.
- Zimmermann, Germa. 2018. „Pecha Kucha.“ In *Kompetent Prüfungen gestalten: 53 Prüfungsformate für die Hochschullehre*, hrsg. v. Julia Gerick, Angela Sommer und Germa Zimmermann, 168–71. utb Hochschuldidaktik 4840. Münster, New York: Waxmann.
- Zimmermann, Tobias. 2014. *Transaktivität in Rückmeldungen: Linguistische und soziokognitive Analysen zur Bezugnahme auf andere Texte in Online-Foren*. Zürich: Universität Zürich. Dissertation. doi: 10.5167/uzh-164372.
- Zimmermann, Tobias. 2018. „Durchführen von lernzielorientierten Leistungsnachweisen.“ In *Kompetenzorientierte Hochschullehre: Die Notwendigkeit von Kohärenz zwischen Lernzielen, Prüfungsformen und Lehr-Lern-Methoden*, hrsg. v. Heinz Bachmann. 3., überarbeitete Auflage, 50–85. Forum Hochschuldidaktik und Erwachsenenbildung 1. Bern: hep.
- Zimmermann, Tobias, Karen-Lynn Bucher und Daniel Hurtado. 2010. „Hybrid Dialog: Dialogic Learning in Large Lecture Classes.“ In *Learning Management System Technologies and Software Solutions for Online Teaching: Tools and Applications*, hrsg. v. Yefim Kats, 314–31. Hershey PA: IGI Global.

- Zimmermann, Tobias und Alex Rickert. 2015. „Austausch in Onlineforen: Wie Kategorien die Lernwirksamkeit von Diskussionen steigern.“ In *Schreiben und Reflektieren: Denkspuren zwischen Lernweg und Leerlauf*, hrsg. v. Monique Honnegger, Daniel Ammann und Thomas Hermann. 1. Auflage, 83–96. Forum Hochschuldidaktik und Erwachsenenbildung 5. Bern: hep.
- Zlatkin-Troitschanskaia, Olga und Jana Seidel. 2011. „Kompetenz und ihre Erfassung – das neue ‚Theorie-Empirie-Problem‘ der empirischen Bildungsforschung?“ In *Stationen empirischer Bildungsforschung: Traditionslinien und Perspektiven*, hrsg. v. Olga Zlatkin-Troitschanskaia, 218–33. Wiesbaden: VS Verlag für Sozialwissenschaften.



Auferkorte-Michaelis | Linde
Bonnes | Haschke | Hintze

Feedback für den Lehralltag

Lehren und Lernen im Dialog

Kompetent lehren, Band 15

utb S • 2023 • 122 Seiten • Kart. • 12,90 € (D) • 13,30 € (A)

ISBN 978-3-8252-6131-3 • eISBN 978-3-8385-6131-8

Als idealer Begleiter für Lehrende in der Hochschule widmet sich das Buch sowohl dem Austausch zwischen Lehrenden und Studierenden als auch dem Peer- und dem kollegialen Feedback. Nach einer theoretischen Einführung laden Methodensteckbriefe zum Ausprobieren ein. Neben bewährten Klassikern wie dem „Blitzlicht“, warten auch neue, praxiserprobte Methoden auf die Lesenden.

www.utb-shop.de



Renate Lichtenstern

Kernelemente herausragender Lehre

Zur hochschuldidaktischen
Personalentwicklung für
angewandte Wissenschaften

2024 • 300 Seiten • kart. • 39,90 € (D) • 41,10 € (A)
ISBN 978-3-96665-082-3 • eISBN 978-3-96665-916-1

Welche Kompetenzen zeichnen herausragende Lehrende aus? Wie haben sich die Anforderungen an Lehrende verändert? Wo und wie erwerben Lehrende Lehrkompetenz und welche Implikationen können hieraus für die hochschuldidaktische Personalentwicklung abgeleitet werden? Die Autorin zeigt mittels einer Replikationsstudie, dass herausragende Lehre auf fünf Kernelementen beruht, die die Bedeutung von Werten in der Lehrkompetenz betonen und sich auf die Weiterbildung auswirken.

www.shop.budrich.de



Rieckmann | Giesenbauer | Nölting |
Potthast | Schmitt (Hrsg.)

Nachhaltige Entwicklung von Hochschulen

Erkenntnisse und Perspektiven
zur gesamtinstitutionellen
Transformation

*Schriftenreihe **Ökologie und Erziehungswissenschaft** der Kommission
Bildung für nachhaltige Entwicklung der DGfE*

2024 • 233 Seiten • kart. • 39,90 € (D) • 41,10 € (A)

ISBN 978-3-8474-2551-9 • eISBN 978-3-8474-1698-2 (Open Access)

Welchen Beitrag zur Nachhaltigen Entwicklung können Hochschulen in den Bereichen Lehre, Forschung, Transfer und Betrieb leisten? Wie gelingt die gesamtinstitutionelle Nachhaltige Entwicklung von Hochschulen? Diese Fragen standen im Fokus des Verbundprojekts „Nachhaltigkeit an Hochschulen“ mit elf deutschen Hochschulen. Der Sammelband präsentiert und diskutiert die Ergebnisse.

www.shop.budrich.de

Leistungsbeurteilungen an Hochschulen lernförderlich gestalten

Wie können Hochschullehrende Leistungsnachweise sowie die Beurteilung und Rückmeldung von Lernleistungen möglichst lernförderlich gestalten? Das Buch bietet anwendungsorientierte Hinweise zur Gestaltung, Beurteilung und Rückmeldung von Prüfungen und vielen weiteren Formaten von Leistungsnachweisen. Zudem thematisiert es lernpsychologische Grundlagen, deren Kenntnis für eine wirkungsvolle und valide Beurteilungspraxis wertvoll ist.

Der Autor:

Dr. Tobias Zimmermann, Zentrum für Hochschuldidaktik und -entwicklung, Abteilung Hochschuldidaktik und Erwachsenenbildung, Pädagogische Hochschule Zürich, Schweiz

ISBN 978-3-8474-3045-2



www.budrich.de

Titelbildnachweis: KI-generiert von DALL-E /OpenAI