Universität Stuttgart
Zentrum für Lehre und
Weiterbildung | zlw

Die Klausur als Orakel?

Arbeitsergebnisse einer Klausurentwicklung in der Technischen Thermodynamik

zlw working paper
1 / 2018

working paper 1/2018

Impressum
ISSN 2363–8834
Herausgeber: Dr. Edith Kröber
Redaktion: Dipl.-Soz. Thorsten Braun
Titelfoto: Avni Qekaj, M.A.

Kontakt
Zentrum für Lehre und Weiterbildung | zlw
Universität Stuttgart
Azenbergstraße 16
70174 Stuttgart

Phone: +49 (0) 711 685 820 21
Mail: sekretariat@zlw.uni-stuttgart.de
Web: www.uni-stuttgart.de/zlw

Inhaltsverzeichnis

THORSTEN BRAUN

Die Klausur als Orakel?

Die Gestaltung von Klausuren ist nicht trivial. Wenn man den Anspruch an aussagekräftige Klausurergebnisse stellt, geraten Aufgabengestaltung, Punkteverteilung und Korrekturverfahren schnell in die Kritik. Wünschenswert wäre eine empirisch fundierte Kompetenzmodellierung für eine Klausur. Verfahren hierzu haben sich noch nicht an der Hochschule etabliert. Der folgende Aufsatz stellt Zwischenergebnisse einer hochschuldidaktischen Kooperation mit dem Institut für Technische Thermodynamik und thermische Verfahrenstechnik (ITT) an der Universität Stuttgart vor. Es werden typische Schwachpunkte einer Klausur aus Perspektive einer Kompetenzmodellierung aufgezeigt und erläutert. Ziel ist es dabei, für Gütekriterien bei der Klausurgestaltung zu sensibilisieren und Ansatzpunkte aufzuzeigen, wie diese Kriterien umgesetzt werden können. Auch wenn die Ausführungen implizit auf eine Prüfungsmodellierung mittels Item Response Theory hinarbeiten, sind sie ohne Vorkenntnisse in der Testtheorie verständlich und umsetzbar.

1 Einführung

Die Klausur als schriftliche Prüfungsform mit zeitlich strikter Begrenzung ist ein Klassiker des Hochschulstudiums. Ihre Ausgestaltung spiegelt die Vielfalt der Fachdisziplinen, Studiengänge, Fachkulturen und organisatorischen Rahmenbedingungen wieder, denen sich Lehrende und Lernende ausgesetzt sehen. Entsprechend zahlreich sind also die Spielarten der Klausur. Von juristischen Fallbearbeitungen über philosophische oder politikwissenschaftliche Essays bis hin zu Konstruktionszeichnungen, mathematischen Problemlösungen oder Multiple-Choice-Klausuren reicht das Spektrum. Sie alle teilen zumeist zwei Gemeinsamkeiten. Einerseits sollen erworbene Kompetenzen der Studierenden sichtbar gemacht werden. Dies dient Lehrenden wie Lernenden zur Reflexion über den Lernstand sowie Erfolg der Lehrbemühungen. Andererseits ist die Prüfung ein formaler Verwaltungs-

akt, der entsprechende Konsequenzen für die Mitglieder der Organisation Hochschule mit sich bringt.

Die Frage nach einer effizienten, effektiven und gerechten Prüfungsgestaltung ist ein immer wieder kehrendes Thema in hochschuldidaktischen Kontexten. Am Zentrum für Lehre und Weiterbildung (zlw) der Universität Stuttgart zeigt sich dies insbesondere bei der Bearbeitung konkreter Praxisanliegen mit Professorinnen und Professoren. Das Thema Prüfung ist konstant präsent und treibt um. In einem früheren working paper haben wir einige Handreichungen zusammengestellt, die eine Entscheidung für oder gegen eine bestimmte Prüfungsform unterstützen können.[1] Im vorliegenden working paper steht die Klausur im Mittelpunkt.

Ziel der Handreichung ist es, Ihnen einige praktische Hinweise für den Entwurf und die Gestaltung von schriftlichen Klausuren zu geben. Dies erfolgt dies aus Perspektive einer empirischen Kompetenzmodellierung, also Bildungsforschung. Was hat es damit auf sich? Eine auf empirische Kompetenzmodellierung hin ausgerichtete Klausurgestaltung möchte eine wissenschaftlich fundierte Feststellung der tatsächlich von Studierenden erworbenen Fähigkeiten erreichen. Neben der formalen Leistungs- und Eignungsfeststellung ist damit die Absicht verbunden, ein differenzierteres Bild von Studienleistungen, Lernschwächen, Effektivität der Lehrmethoden, Lernhürden usw. zu erhalten. Es spielt also die *Gütebeurteilung* von Klausuren eine wichtige Rolle.

Eine kompetenzorientierte Prüfungsmodellierung mit wissenschaftlichem Anspruch ist faktisch Bildungsforschung. Sie lässt sich leider nicht gebrauchsfertig auf wenigen Seiten zusammen fassen. Es kann im Folgenden also nicht um eine Einführung in die empirische Bildungsforschung oder probabilistische Testtheorie gehen. Hierzu muss auf entsprechende Einführungen verwiesen werden.[2] Ebenso ist zu ergänzen, dass zahlreiche Hochschulen spezielle Einrichtungen zur Entwicklung von Lehr-/Lernformen besitzen, an denen man fachkundige Unterstützung erhalten kann. Die folgenden praktischen Ratschläge sind entsprechend allgemein formuliert und sollen dabei helfen, einige zeitgemäße Prinzipien der kompetenzorientierten

[1] vgl. Rapp 2014.
[2] vgl. Bühner 2011, S. 477–602.

Prüfungsmodellierung auch ohne vertieftes Fachwissen zu berücksichtigen. Wir sind davon überzeugt, dass sich diese empirisch reflektierte Form der Prüfungsgestaltung an Hochschulen immer mehr durchsetzen wird. Im Bereich nicht-akademischer, beruflicher Ausbildung[3] oder bei großen Bildungsstudien[4] gehören sie schon zum etablierten Standard der Leistungsmessung. Kompetenzmodellierungen von Prüfungen im Bereich der hochschulischen Grundlagenausbildungen wurden bereits versucht. [5] Das Thema wird also in naher Zukunft eher vermehrt in Fakultätsratssitzungen, bei Projektanträgen oder Studiengangentwicklungen Gegenstand von Diskussionen sein.

2 Die Klausur als Orakel – Problemanalyse

Im Folgenden werden typische Probleme von Klausuren anhand einer etwas zugespitzten und spielerischen Darstellung angesprochen und in grellen Farben überzeichnet. Ähnlich einem Orakel, hat die Klausur ja durchaus etwas Archaisches an sich. Sie hebt sich aus dem Dunkel einer scheinbar vorgeschichtlichen (besser vor-empirischen) Zeit herauf. Vor der Klausur war Ungewissheit. Der Wissensstand der Studierenden war unklar und das Semester als Lehrprogramm bleibt bis zum Ende in seinem Ausgang unsicher. Die Klausur tritt sodann als heilsames Mittel gegen dieses Dunkel an, ist selber aber nicht minder bedrohlich. Einem Orakel gleich, erscheint den Studierenden die Klausur als verdeckter Griff der Hand des Schicksals (Lehrende?) in einen Beutel mit Runensteinen (Themenplan?). Den bangen Lernenden offenbart sich das eigene Schicksal oftmals erst mit ihrem starren Blick auf die kryptischen Rätsel, die ihnen die Klausur auferlegt.

Eine Klausur kann mit Fug und Recht als Orakel bezeichnet werden, wenn weder Lehrende noch Lernende genau wissen was geprüft wird, auf welchem Niveau dies geschehen soll und warum. An *Transparenz* fehlt es

[3] vgl. Nickolaus und Seeber 2013.

[4] vgl. Organisation for Economic Co-operation and Development 1999.

[5] So etwa im Projekt »Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen (KoKoHs)«; siehe http://www.kompetenzen-im-hochschulsektor.de (letzte Überprüfung 02. März 2017).

dann völlig, Ziele und Inhalte gehorchen höheren Mächten.

Eine Klausur unterstellt üblicherweise, dass verschiedene *Wissensgebiete und Handlungsweisen* (z. B. Methoden) legitime Bestandteile einer Leistungsüberprüfung sind. Legitim sind sie dann, wenn sie das Thema des Fachs, Moduls oder einer konkreten Veranstaltung betreffen. *Leistung* wird dann im technischen Sinne verstanden, als Arbeit in begrenzter Zeit. Da aber eine umfängliche Prüfung aller legitimen Inhalte nicht möglich ist, stellt jede Prüfung eine *Selektion* dar. Aus dem Universum gültiger Fragen und Herausforderungen des Studiums werden also Inhalte und Aufgabentypen mehr oder weniger begründet »gezogen« (etwa wie aus einer Urne in der Lotterie oder Runensteine aus dem Beutel). Dabei unterstellen die Prüferinnen und Prüfer normalerweise, dass manche Aufgaben schwerer sind als andere – und dass dies aus gutem Grund so ist. Jede mögliche Aufgabe hat also einen *Schwierigkeitsparameter*, so wie es auch farbige Kugeln in der Urne gibt. Man trägt diesem Umstand Rechnung, indem verschiedene Klausuraufgaben verschiedene Punkte wert sind. Wer also die schwere Rune des Scheiterns zieht, aber dennoch die Aufgabe meistert, der hat *mehr Punkte* verdient. Dabei spielt auch oft die Zeit eine Rolle. Denn wenn eine Aufgabenbearbeitung länger dauert, muss sie fast zwangsläufig mehr Punkte einbringen, da sonst im Verhältnis zu anderen Aufgaben ein Ungleichgewicht entstünde.

Nun muss sich natürlich auch die Spreu vom Weizen trennen, oder – um bei dem Bild des Orakels zu bleiben – manchen armen Seelen ist eine güldene Zeit beschert, andere erwartet ein schweres Schicksal. Die *magische Bestehensgrenze* ist die diagnostische Seelenwaage. Wer die wissenschaftlichen Inhalte pflichtbewusster angeeignet hat als andere, so kann man schließen, wird die Aufgaben auch schneller lösen. Daher macht es ja vermeintlich Sinn, die Klausur mit ein oder zwei besonders anspruchsvollen Aufgaben zu versehen, welche nur die oberen 15% der Studierenden erfolgreich lösen können. Die *Leistungsspannweite* ist begrüßenswert! Schade nur, dass die obere Leistung *in der Zeit* dann doch nur von 5% erbracht werden kann, weil die anderen 10% dem Prüfungsstress, Flüchtigkeitsfehlern oder Sprachbarrieren zum Opfer fallen. Da alle Studierenden aber dieselbe Prüfung schreiben – so hört man es sagen – sei auch der Gleichberechtigung genüge getan, denn die Prüfung ist ja für alle dieselbe. Wer eine

Aufgabe löst, der hat erfolgreich gelernt. Dass manche Studierende vielleicht verschiedene *Lösungsstrategien* verfolgen oder *andere Kompetenzen* einsetzen, bleibt unbeachtet.

Wie dem auch sei, auf jeden Fall braucht man am Ende eine *Note*. Die erhält man durch die *Aufsummierung der Teilleistungen* zu einer Gesamtsumme. Das ist ein bequemer Weg, denn am Ende hat jede Person eine Zahl, welche über besser und schlechter entscheidet. Direkt nach dem Würfelwurf, der Vogelschau und den Runensteinen ist dies wohl die unschärfste Bestimmung des studentischen *Leistungsniveaus*. Sie entspricht der magischen Hand höherer Mächte, die sich aus dem Himmel heraus über die Studierenden waltend niederlegt. Diese Hand folgt vielleicht manchmal der mildtätigen Erwägung, dann doch nicht zu viele Studierende durchfallen zu lassen und deshalb lieber ein wenig *nach oben zu korrigieren*. Oder es wird streng dem Hilfs-Gott der Statistik vertraut: Es wird am Mittelwert geteilt, oder am unteren Quantil. Vielleicht trifft sich auch das Pantheon der Korrektoren und beratschlagt nach antiker Sitte, wo die beste Bestehensgrenze zu verlaufen habe.

Überhaupt sind die *Korrektorinnen und Korrektoren* wichtige Personen, denn sie stellen die Leistung der Studierenden faktisch fest und erwägen Punktabzug, Teilerfolge, hinreichende Leistungen und Flüchtigkeitsfehler. Im besten Fall ist dieser Prozess des Abwägens dann einheitlich zwischen allen Korrektorinnen und Korrektoren abgestimmt. Im schlimmsten Fall predigt der Eine Wasser, der Andere Wein und der *Korrekturleitfaden* entspricht dem jeweiligen Bauchgefühl.

Diese ironisch überspitzten Zustände treffen in dieser Reinform hoffentlich selten zu. Sie sind für sich genommen aber auch keine bloße Fiktion. Tradition, Gewohnheit und organisatorische Verpflichtungen belasten Klausuren bisweilen mit starken Gütebedrohungen. Im Folgenden werden Aspekte dieser negativen Geschichte aufgegriffen und mit Handlungsempfehlungen kommentiert.

3 Kompetenzmodell des Moduls / der Lehrveranstaltung

Eine Lehrveranstaltungsplanung beginnt üblicherweise mit der Formulierung von Zielen, die Studierende bestenfalls nach Abschluss des Moduls erreicht haben. Unmittelbar nach der Zielformulierung springt man für gewöhnlich zu der Frage, wie man diese Zielerreichung am Ende sichtbar machen, also auch prüfen, kann. Insofern beginnt jede Klausur mit den Zielen der Veranstaltung. Es haben sich in der Praxis der Studiengangentwicklung einige Dinge als sehr hilfreich herausgestellt, die man einer Prüfungsgestaltung vorausschicken sollte.

Dies ist zum einen die Formulierung eines möglichst anschaulichen Kompetenzmodells für das Studienmodul, in dem geprüft werden soll. Dieses Kompetenzmodell ist zuerst einmal inhaltlich zu verstehen, also als *fachliche Themensammlung*. Es sollte aber auch Methoden und Prozeduren umfassen, also die erwarteten *studentische Arbeitsweisen*. Hierfür bieten sich verschiedene Methoden an. Von einer schlichten Gliederung bis zur kreativen, dynamischen Mind-Map ist vieles denkbar.[6] Solch ein Kompetenzmodell entwickelt sich am einfachsten (und auch am besten) im Austausch mit Kolleginnen und Kollegen, zum Beispiel im Rahmen einer Klausurtagung oder eines Workshops. Das Ergebnis dient unmittelbar der Eingrenzung und näheren Bestimmung dessen, was relevant oder irrelevant für die Prüfung ist.

Ausgehend von solch einer qualitativ gehaltvollen Modulbeschreibung können dann Lehr-/Lernziele formuliert werden. Lernziele sind das angestrebte und zu prüfende Ergebnis des erfolgreichen Lernprozesses. Lernziele sollen daher als *beobachtbare Handlungen der Studierenden* formuliert sein. Diese Handlungen werden *prototypisch* in einer Prüfung verlangt. Zu solchen Lehr-/Lernzielen gibt es zahlreiche Literatur, auf die hier nur ansatzweise verwiesen wird.[7] Im Kern bedarf es einer Lernzieltaxonomie, anhand derer man die von den Studierenden erwarteten Leistungen nach Leistungsniveau differenziert. Etwas betagt aber immer noch aktuell ist die sehr

[6] vgl. Schaper o.D., S. 6 f.

[7] vgl. Arbeitsstelle für Hochschuldidaktik der Universität Zürich 2008.

verbreitete Taxonomie von Bloom.[8] Es sind aber auch andere Fassungen verbreitet.[9]

Bevor es also überhaupt an die Gestaltung einer Prüfung geht, bedarf es eines konkreten Bildes und konkreter Grenzen, die das, was legitim zu prüfen ist, klar bestimmen: Was sollen Studierende können, auf welchem Niveau und wie kann ich es prüfen? Streng genommen sollte zu diesem Zeitpunkt die Entscheidung für oder gegen eine Klausur noch gar nicht erfolgt sein. Andere Prüfungsformen sind vielleicht angemessener.[10] Wir gehen im Folgenden davon aus, dass eine Klausur als sinnvoll erachtet wird.

4 Transparenz und Constructive Alignment

Eine Klausursituation ist vergleichsweise künstlich und wenig realistisch, wenn es darum geht wissenschaftliche Kompetenz sichtbar zu machen. Ob der mit einer Klausur verbundene Leistungsdruck, die Nervosität, Begrenztheit der Hilfsmittel, Einsamkeit und künstliche Zeitknappheit der Idee authentischer, wissenschaftlicher Arbeit entspricht, kann man sicherlich sehr kontrovers diskutieren. Sofern die Fachinhalte nicht explizit Stressbewältigung, Arbeiten unter Zeitdruck oder Umgang mit dem Ungewissen beinhalten, besteht eigentlich kein Grund, Studierende über die geforderten Inhalte, Handlungen oder das erwartete Niveau der Ergebnisse einer Klausur im Unklaren zu lassen. Die Anforderungsspezifikationen, welche für eine Klausur erarbeitet werden, sollten also auch den Studierenden vollumfänglich transparent und zugänglich sein. Entwickelte Kompetenzmodelle, Epitome oder Lernziele für das eigene Modul sind gleichzeitig sehr effektive Lernhilfen für Studierende. Insofern produziert eine solide Klausurentwicklung auch didaktisches Material, das der Steuerung des Lernprozesses zugute kommen kann.

Im Semesterverlauf sollten Themen und Methoden behandelt werden, die auch legitim zu erwartende Gegenstände der Prüfung sind. Umgekehrt

[8] vgl. Bloom und Krathwohl 1969.
[9] vgl. Biggs 2003.
[10] vgl. Rapp 2014.

sollte die Prüfung nur Dinge enthalten, die auch in irgendeiner Weise fair vorbereitet werden konnten. Man spricht hier vom *constructive alignment* und meint damit eine konsistente und faire Abstimmung zwischen Lernzielen, Prüfung und dem didaktisch angeregten Lernhandeln während des Semesters.[11] Angestrebt sind solche Prüfungsaufgaben, die valide gewünschte Kompetenzen abbilden und eine reliable Diagnostik (also Auswertungsmethode) zulassen.[12] Das bedeutet auch, dass die Korrekturrichtlinien für die Studierenden ebenfalls transparent sein sollten.

Während sich Transparenz vergleichsweise einfach herstellen lässt, ist eine konsistente Abstimmung von Zielen, Lernhandeln und Prüfungsaufgaben nicht so trivial. Grundsätzlich sollte eine Prüfung von den Studierenden eine Leistung und Handlungsweise verlangen, die dem Lernprozess des Semesters entspricht. Überraschungen sollten in Klausuren nicht eingebaut werden, insbesondere nicht in Bezug auf die erwartete Arbeitsweise. Schlecht wäre es also, nach einem Semester kleinschrittiger Mathematikaufgaben plötzlich große Aufgaben zu stellen, in denen verschiedene Aufgabengebiete integriert geprüft werden.

5 Klausurspezifizierung

Wie erstellt man effektive Klausuraufgaben? Ausgehend von dem bisher gesagten, orientieren sich Aufgaben inhaltlich und arbeitstechnisch an dem, was im Semester auch üblicherweise gelehrt und gelernt wurde. Es besteht bei erfolgter Vorbereitung mittels Kompetenzmodell und Lehr-/Lernzielen eine recht klare Vorstellung davon, welche Kompetenzen am Ende geprüft werden sollen. Das sind zuerst einmal thematische Felder und methodische Vorgehensweisen (z. B. in der Thermodynamik: Feuchte Luft, Kreisprozesse, Mischungsprozesse, Hauptsätze, Interpolationen, Arbeit mit Grafen und Tabellen). Dies bildet die mögliche *Kompetenzbreite*, aus der eine Prüfung schöpfen kann. Es empfiehlt sich nun zum Beispiel im Team mit Kolleginnen und Kollegen einen größeren *Test-Pool von Klausuraufgaben* zu entwerfen, der die gesamte Kompetenzbreite abdeckt und zusätzlich dabei auch

[11] J. Wildt und B. Wildt 2012, vgl.

[12] Der Begriff der Diagnostik ist in diesem Zusammenhang von Reis 2015, übernommen.

das Spektrum der vermuteten *Schwierigkeitsgrade* umspannt. Für jedes typische Themenfeld oder Arbeitsweise können also leichte bis schwere Prüfungsaufgaben entworfen werden. Dabei ist es zuerst einmal auch egal, ob viel mehr Aufgaben entworfen werden als später in der Prüfung Verwendung finden. Es geht vielmehr darum, das Universum der möglichen Themen und Schwierigkeitsgrade abzustecken sowie mögliche Aufgabenformate zu erproben.

Was macht eine Aufgabe nun schwerer als andere? Eine Antwort kann darauf oftmals a priori nur sehr schwer gegeben werden. Zum einen ist dies abhängig vom Inhalt, der Zahl von Arbeitsschritten und dem Grad der Abstraktion. Zum anderen ist die vorherige Einschätzung der Schwierigkeit häufig nicht mit der tatsächlichen identisch. Erst nach einer Prüfung zeigt sich, ob die Perspektiven der Lehrenden und Lernenden eine gleiche Einschätzung der Aufgabenschwierigkeit liefern. Dennoch kann allgemein gesagt werden, dass eine Prüfungsaufgabe umso schwieriger wird, je weniger Vorstrukturierung die Aufgabenstellung bietet. Es wird dann ein hohes Maß an *selbstgeleitetem Arbeiten* verlangt. Sind die Aufgaben kleinschrittiger untergliedert und so beschaffen, dass sie die Studierenden durch das Problem führen, ist die Aufgabe in der Regel leichter. Diese Unterscheidung steht oft (aber nicht zwangsläufig) im Zusammenhang mit der Taxonomiestufe (und damit den Lernzielen) einer Aufgabe. Das Aufzählen, Wiedergeben oder Anwenden einzelner Wissensbestände oder Methoden (Taxonomiestufen Erinnern, Anwenden) ist einfacher als globale Fragen nach Begründungszusammenhängen, kritischer Bewertung oder dem Ableiten von Konsequenzen aus gegebenen, nicht vorstrukturierten Sachverhalten (Taxonomiestufen Evaluation, Innovation). Eine verbreitete Taxonomie ist in Tabelle 1 verkürzt zusammengefasst.

Ein solcherart gestalteter Aufgaben-Pool sollte also für jedes Thema und für jeden Anspruchsgrad eine oder mehrere Beispielaufgaben beinhalten. Lassen Sie diese Aufgaben dann nach Möglichkeit von Studierenden einmal durchrechnen! Zum Beispiel in Gruppenübungen oder auch in einer Vorlesung. Sammeln Sie Erfahrung mit den Aufgabentypen. Was fällt Studierenden leicht, was schwer? Wo liegen Fehler oder Missverständnisse? Gehen Sie mit Studierenden über die Aufgabentypen in den Dialog. Es sollte Aufgaben für jeden Schwierigkeitsgrad geben.

Tabelle 1: Taxonomiestufen nach Benjamin S. Bloom

Stufe	Beispieltätigkeiten
Wissen und Verstehen	beschreiben, ordnen, wiedergeben
Anwenden	ausführen, berechnen, zeichnen, erstellen
Analysieren	ableiten, auswerten, prüfen, vergleichen
Synthetisieren	einordnen, planen, Thesen bilden
Bewerten	begründen, einschätzen, urteilen, werten
Innovieren	entdecken, erzeugen, konzipieren, erfinden

An dieser Stelle wird ersichtlich, dass die Entwicklung einer Klausur perspektivisch über ein Semester hinausreicht. Es ist völlig legitim und auch notwendig, dass sich Prüfungen über die Zeit ändern. Insofern fließen auch Erfahrungen vergangener Klausuren mit in die Beurteilung ein. Es werden sich wahrscheinlich Erkenntnisse ansammeln, die zu einem gewissen Zeitpunkt einen kleinen oder größeren Bruch in der Klausurgestaltung nötig machen. In diesem Fall sollte die Anpassung der Prüfung zu Beginn des Semesters erfolgen, damit auch didaktische Konsequenzen in die Lehre einfließen können. Sonst könnte es unter Umständen zu Missverständnissen kommen, da Studierende eine den Vorsemestern vergleichbare Klausur erwarten. Auch hier empfiehlt sich also ein hohes Maß an Transparenz bezüglich Klausurreformen.

Entwerfen Sie dann eine erste »echte« oder »revidierte« Prüfung, indem Sie die gemachten Erkenntnisse verwerten. Eine Klausurkonstruktion erfolgt dabei anhand der beabsichtigten Klausurziele. Je nachdem, ob Sie zwischen bestimmten Kompetenzstufen diskriminieren, die Bandbreite des Leistungsvermögens erfassen oder einen bestimmten Kompetenzbereich besonders untersuchen möchten, sind entsprechende Klausuraufgaben auszuwählen. Im ersten Fall nehmen Sie Aufgaben aus dem gesamten Schwierigkeitsspektrum, wobei insbesondere mehr Fokus auf besonders leichte und schwere gelegt werden sollte. Im zweiten Fall nehmen Sie mehr Aufgaben eines ähnlichen Schwierigkeitsgrades in die Prüfung auf, um hier mehr Information zu erhalten. Im dritten Fall fokussieren Sie zum Beispiel

einen bestimmten Themenbereich oder eine bestimmte Arbeitsweise. Es macht generell Sinn, Aufgaben nach *steigendem Schwierigkeitsgrad* zu ordnen, also mit den leichten zu beginnen. Dies gilt auch innerhalb von Aufgabenbündeln, die sich auf einen gemeinsamen Textimpuls beziehen und daher eine Einheit bilden. Auch innerhalb solcher Bündel ist es von Vorteil mit leichten Aufgaben zu beginnen.

Zu berücksichtigen ist, dass im Sinne der Testfairness und des constructive alignment den Studierenden früh bekannt sein sollte, was die Prüfung erreichen will und mit welchem Typus von Aufgaben prinzipiell zu rechnen ist. *Außerdem – und dies kann hier nicht stark genug betont werden – ist es für eine faire und aussagekräftige Prüfungsgestaltung sehr schädlich, wenn eine Klausur zu umfangreich für die verfügbare Zeit ist*. Als Faustregel gilt, dass eine Person mit mittlerem Leistungsniveau in der Lage sein sollte die gesamte Klausur zu lösen ohne in Zeitnot zu geraten. Man könnte nun argumentieren, dass man im Grenzfall also unbegrenzt Zeit geben müsste, eine Klausur zu bearbeiten. Der anzustrebende Grenzfall sollte aber eher sein, eine aussagekräftige Bearbeitung aller Aufgaben innerhalb der verfügbaren Zeit zu ermöglichen. Oder anders gewendet: wer die Kompetenz zur Lösung im Prüfungskontext besitzt, sollte genügend Zeit haben diese Kompetenz erfolgreich zu demonstrieren.

6 Punktesummen und erschöpfende Statistiken

Ein sehr verbreitetes Verfahren zur Bewertung von Klausuren ist die Vergabe von Teilpunkten, die zu einer Punktesumme aufgerechnet werden. Auf Grundlage dieser Summe wird dann der prozentuale Anteil von der maximal zu erbringenden Leistung berechnet. Dabei sind die Schwierigkeit oder der Bearbeitungsaufwand einer Teilaufgabe oftmals durch eine gewichtete Punktzahl berücksichtigt. So erbringt schwere Aufgabe A den Studierenden 15 Punkte, eine einfachere Aufgabe nur 5 Punkte. Diese Logik ist vertretbar, um pragmatisch zu einer Notenfindung zu gelangen. Vielleicht drückt sie aber auch eine gewisse, intuitiv spürbare, Hilflosigkeit der Korrektoren aus, jenseits der Punktzahl kein Benotungskriterium zur Hand zu haben. Die Benotung nach solchen gewichteten Punktesummen birgt jedenfalls

erhebliche Nachteile. Hierzu ein Beispiel.

Besteht eine Klausur aus vier Teilaufgaben verschiedener Schwere, dann erhält die erste Aufgabe 5, die zweite 10, die dritte 10 und die vierte 20 Punkte. Wir würden nun sagen, dass, wer die letzte Aufgabe erfolgreich löst, einen fachlichen Transfer geleistet hat, der auf jeden Fall dem höchsten Anspruch der Lernziele im Modul genügt. Wenn nun eine Studierende diese Aufgabe löst, sonst aber keine andere, erhielte sie 20 von 45 möglichen Punkten, was unter der Hälfte läge. Sagen wir für dieses Beispiel, sie wäre durchgefallen. Würde man ihr ein Nichtbestehen mit gutem Gewissen bescheinigen wollen?

Ebenso denkbar ist es, dass ein Studierender in jeder der vier Aufgaben ungefähr die Hälfte der Punkte erringt. Er hätte dann vielleicht 23 Punkte erreicht und damit knapp bestanden. Tatsächlich brachte er aber keine der Aufgaben erfolgreich zum Abschluss, sondern lediglich einige Ansätze zu Papier. Könnte man hier sagen, er hat das Modulziel erreicht?

Der Kern des Problems besteht hier aus Sicht der Kompetenzmodellierung, dass die Punktesumme mit einer sehr großen Unschärfe belastet ist. Sie kaschiert und verdeckt tatsächliche Leistungsniveaus und macht eine qualitative Rückmeldung anhand dieser Punktesumme unmöglich. Man kann auch sagen, dass ohne Kenntnis von *Antwortmustern* eine einfache *Summenskalierung* in der Klausur nicht eindeutig als *Ausprägung der Personenfähigkeit* interpretierbar ist. Die Punktesummen sind keine *erschöpfende Statistik*. »Eine erschöpfende Statistik ist ein Kennwert, das sämtliche Informationen enthält, die man zu seiner Interpretation benötigt.«[13] Dies ist im konstruierten Beispiel nicht der Fall. Und es sind viele Variationen dieses Beispiels denkbar. Aspekte wie Strafpunkteabzug, Formfehler oder unklare Behandlung von Rechenfehlern verschlimmern die Lage noch.

Das Problem betrifft die Beurteilung der Studierenden, spiegelt sich aber ebenso in der Beurteilung von Prüfungsaufgaben wieder. Auch die *Aufgabenschwierigkeit* wird nicht einfach durch die Summe der korrekten Bearbeitungen erschöpfend abgebildet, da die Antwortmuster über Personen unberücksichtigt bleiben. Vielleicht haben nur zehn Prozent eine Aufgabe gelöst. Waren diese zehn Prozent nun besonders leistungsschwach oder

[13] Bühner 2011, S. 483.

leistungsstark? Welche Aufgaben haben sie noch gelöst? Diese Information liegt nicht vor, wenn man lediglich die Zahl erfolgreicher Bearbeitungen zur Interpretation von Aufgaben heranzieht.

Man kann also zusammenfassen: Eine Punktesumme ist keine erschöpfende Statistik, da die Schwierigkeit der Aufgaben und Fähigkeitsgrade der Personen nicht aufeinander bezogen werden; Antwortmuster bleiben unberücksichtigt. Ein Großteil der probabilistischen Testtheorie und kompetenzorientierten Prüfungsmodellierung dreht sich um diese Problematik. Klausurergebnisse lassen sich aber auch ohne vertiefte Kenntnisse der Statistik relativ leicht auf den Grad von Verzerrungen dieser Art prüfen.

Liegen Klausurergebnisse in einer Tabelle vor, in der jede Spalte eine Teilaufgabe und jede Zeile eine Person ist, dann beinhaltet jede Zelle die erreichte Teilpunktzahl. Wenn man nun für jede Spalte und Zeile die Summe bildet (und zum Beispiel am Rand jeweils eine Spalte und Zeile hierfür hinzufügt), dann erhält man eine Rangordnung der Personen und Teilaufgaben. Eine Person mit hoher Summe ist leistungsstärker als andere. Eine Aufgabe mit hoher Summe ist tendenziell leichter als andere (da sie mehr Personen erfolgreich bearbeiten konnten). Sortiert man nun diese Tabelle nach den Randsummen in der Art, dass die Spalten von leicht nach schwer und die Personen von schwach nach stark streben, dann sollte sich im idealen Fall eine deutliche Tendenz zeigen: je leistungsstärker eine Person ist (je höher also ihr Summenwert), desto mehr schwere Aufgaben sollte sie erfolgreich bearbeitet haben. [14]

Muster, nach denen viele Personen schwerere Aufgaben schaffen, manche leichteren aber nicht, sind Hinweise auf unerwartete Prüfungseigenschaften. Oder dass sich überhaupt ein ziemliches Chaos abbildet und es keine klaren Tendenzen gibt. Manche Teilaufgaben werden vielleicht von niemandem oder aber von jeder Person erfolgreich bearbeitet. Zu erwarten und wünschenswert wäre es hingegen, dass eine Person mit einer beliebigen Fähigkeit alle Aufgaben bis zu einer gewissen Schwierigkeit löst, aber

[14] Wenn die Klausur stark unterschiedlich gewichtet ist, funktioniert diese Logik ggf. nicht. Eine leichtere Aufgabe erhält dann vielleicht doch mehr Punkte als eine schwere, weil sie viel Zeit zur Bearbeitung benötigt. In diesem Fall empfiehlt es sich die erreichten Punktzahlen zu recodieren und beispielsweise einen Punkt für teilweise und zwei Punkte für vollständige Bearbeitung zu vergeben.

keine darüber hinaus.

Mit relativ geringem Aufwand verschafft einem eine solche Auswertung ein Gefühl für die Struktur einer Klausur und eventuelle Inkonsistenzen. Diese Vertrautheit kann insbesondere dann helfen, wenn die Klausur von sehr vielen Studierenden geschrieben oder vielen Korrektoren korrigiert wird und eine intuitive Einschätzung der Gesamtergebnisse nur schwer möglich ist. Die Alternative zur Punktesummierung besteht in einer angemessenen Leistungsskalierung, zum Beispiel mit Hilfe der Item Response Theorie.[15] Dies macht meistens externe Unterstützung notwendig. Aber auch ohne diesen Schritt kann es sehr hilfreich für die weitere Prüfungsgestaltung sein, wenn man notorisch problematische oder irritierende Teilaufgaben zukünftig ersetzt oder auslässt. Auch hier können Studierende in den Analyseprozess aktiv eingebunden werden, indem man die Befunde zum Beispiel bespricht und so irritierenden Bearbeitungsmustern auf die Spur kommt.

7 Unabhängigkeit der Aufgaben

Viele Klausuren bieten den Studierenden mehr als nur eine Aufgabe zur Bearbeitung an. [16] Die Zahl der Aufgaben variiert mit dem erwarteten Bearbeitungsumfang jeder Teilaufgabe. Dabei spielt die Art des zu prüfenden Wissens eine wichtige Rolle bei der »Korngröße« beziehungsweise Kleinteiligkeit von Teilaufgaben. Jedenfalls wird man erwarten, dass die Teilaufgaben sich nach Schwere und möglicherweise auch Themenfeld unterscheiden, es gibt vielleicht auch Redundanzen, also Aufgaben, die sich prinzipiell ähneln.

Aus Perspektive der Kompetenzmodellierung ist es ein wichtiges Anliegen, dass sich Teilaufgaben nicht in solcher Abhängigkeit untereinander befinden, dass ein Studierender zum Beispiel keine Chance hat Aufgabe B zu lösen, wenn zuvor nicht A gelöst worden ist. Ebenso wenig wäre es zu begrüßen, dass Aufgabe A die Lösung von Aufgabe B in irgendeiner nicht intendierten Form beeinflusst, etwa durch Hinweise im Aufgabentext,

[15] vgl. Bühner 2011, S. 477–602.

[16] Bei einigen Formaten ist das vordergründig nicht der Fall, siehe hierzu weiter unten Abschnitt »Quantifizierung offener Aufgaben«, S. 24.

die für andere Aufgaben unmittelbar lösungsrelevant sein könnten. Man spricht von *lokaler Unabhängigkeit* von Teilaufgaben und meint damit im statistischen Sinne, dass bei einer gleichbleibenden Personenfähigkeit keine Korrelation mehr zwischen den Antworten auf die Teilaufgaben besteht. Es gibt dann keine anderen Variablen neben der Fähigkeit der Studierenden, die für einen Zusammenhang im Antwortverhalten verantwortlich sind. Die Ausprägung der Personenkompetenz ist dann im Idealfall die einzige Variable, die für eine Korrelation zwischen den Aufgabenergebnissen sorgt.[17] Diese Eigenschaft wird von Prüfungen explizit verlangt, wenn sie nach der Item Response Theorie probabilistisch modelliert werden soll. Die Forderung wird als *lokale stochastische Unabhängigkeit* bisweilen noch verschärft. Damit ist gefordert, dass bei konstanter Fähigkeit einer Person die Beantwortung von Teilaufgaben *unabhängig* voneinander erfolgt. Dies wird mathematisch dadurch ausgedrückt, dass das Produkt aller einzelnen Lösungswahrscheinlichkeiten gleich der Gesamtlösewahrscheinlichkeit der Prüfung ist. »[B]ei lokaler stochastischer Unabhängigkeit [ist] gegeben, dass die Korrelationen zwischen den Items alleine auf die spezifizierte latente Variable zurückgeführt werden können und nicht auf andere Einflussgrößen, z. B. auf andere Eigenschaften und Fähigkeiten.«[18]

Der Grund dafür, darauf bei der Prüfungserstellung zu achten, liegt darin, dass eine Prüfung möglichst frei von Verzerrungen sein sollte. Wenn alle Aufgaben einer Prüfung in der Summe eine Gesamtleistung abbilden, dann bedeutet eine Abhängigkeit der Aufgaben untereinander eine drohende Verzerrung. Hängt die Hälfte der Teilaufgaben so eng zusammen, dass Studierende diese Aufgaben tendenziell ganz oder gar nicht lösen, dann bringt dies die Beurteilungsgrundlage in eine Schieflage. Es kann aber auch andere Beispiele dafür geben. Wenn gewisse Aufgaben durch eine besondere Lösungsstrategie abgekürzt werden können, dann wird nicht mehr die gleiche Fähigkeit für alle Aufgaben abgeprüft. Die fraglichen Aufgaben korrelieren dann über die alternative Lösungsstrategie, die möglicherweise so nicht im Prüfungsdesign vorgesehen war. Weiterhin könnten sich Aufgabentypen so stark ähneln, dass sich ein Gewöhnungseffekt einstellt etc.

[17] vgl. Bühner 2011, S. 32 f.
[18] Ebd., S. 485.

Um Verletzungen der lokalen Unabhängigkeit zu vermeiden, können einige Praxisregeln beachtet werden:[19]

- Es sollte *keine generellen und spezifischen Fragen zur gleichen Sache* geben, so dass die generelle Frage nahezu perfekt durch die speziellen vorausgesagt werden kann oder umgekehrt.
- Fragen in einem *nicht vertrauten Antwort- oder Arbeitsformat* führen zu Eingewöhnungseffekten, so dass sich das Antwortverhalten mit der Zeit verändert. Klausurfragen sollten vielmehr in vertrautem Format formuliert und bearbeitet werden.
- Aufgaben sollten *nicht aufeinander aufbauen*, so dass ein Fehler oder eine fehlende Antwort in der einen Frage zwangsläufig zu Folgefehlern oder Nichtbearbeitung führt.
- Es sollte vermieden werden, dass *Klausuraufgaben mit einem einheitlichen Satzstamm* beginnen (z. B. »Welcher der folgenden Aussage stimmen Sie am meisten zu. . . «). Besser ist es, Formulierungen zu variieren.
- Die *Vertrautheit mit gewissen Aufgabentypen* kann zu Verzerrungen führen. Eine zu homogene Aufgabenstruktur sollte vermieden werden.

Es muss abschließend bezüglich der lokalen Unabhängigkeit betont werden, dass es natürlich nicht darum geht Aufgaben zu formulieren, die nichts miteinander zu tun haben. Alle Aufgaben sollen die fragliche Fähigkeit prüfen. Nicht zusammenhangslose Aufgaben sind das Ziel, sondern solche, deren Beantwortung ausschließlich über die angestrebte Fachkompetenz korrelieren.

8 Aufgabenbündel und Textimpulse

Die im letzten Abschnitt formulierte Forderung nach einer Unabhängigkeit aller Teilaufgaben erscheint einerseits intuitiv, andererseits entspricht

[19] vgl. Bühner 2011, S. 486.

sie oft nicht den Erwartungen an eine angemessene Aufgabenstellung im Fach. Oftmals sind es nämlich Aufgabenbündel, die sich auf einen gemeinsamen Textimpuls beziehen und so beispielsweise eine exemplarische Anwendung verschiedener Verfahren und Methoden an einem Fallbeispiel verlangen. Ein klassisches Beispiel ist eine komplexe Situationsbeschreibung zu einem technischen Vorgang oder einer anderen Problemstellung, die als Grundlage für eine Reihe von Prüfungsaufgaben dient. Man spricht in solchen Fällen von Aufgabenbündeln oder auch Testlets. Es gibt gute Gründe dies so zu tun. Aus ökonomischen Erwägungen möchte man eine umfangreiche Aufgabenschilderung auch möglichst zeiteffizient ausnutzen und so viel Beurteilungsmaterial wie möglich von Studierenden gewinnen. Auf einen halbseitigen Textimpuls nur eine Frage zu stellen, verschenkt Potential. Auch erscheint es aus fachlicher Perspektive oft angemessen, Problemschilderungen als Ausgangspunkt für verschiedene, gestufte Arbeitsschritte festzulegen, die dann sukzessive (und meist mit steigender Schwierigkeit) bearbeitet werden sollen. Dies steht allerdings im direkten Widerspruch zur Forderung nach lokaler Unabhängigkeit, da fast per Definition die Unabhängigkeit von Teilaufgaben verletzt ist, wenn sich diese auf einen gemeinsamen Frageimpuls oder eine Situationsbeschreibung beziehen. Versteht eine Studierende zum Beispiel die Ausgangsschilderung nicht (vielleicht aufgrund sprachlicher Missverständnisse), scheitert sie zwangsläufig an allen weiteren Teilaufgaben des Testlets. Die methodischen Fähigkeiten hätte sie aber vielleicht mitgebracht.

In solchen Situationen ist aus kompetenzanalytischer Perspektive eine Abwägung vorzunehmen. Die Prüfung und die Interpretation von Prüfungsergebnissen gewinnen auf jeden Fall durch Aufgabenbündel an Komplexität. Es existieren zwar statistische Verfahren, um solch einen Einfluss von Aufgabenbündel bei der Bestimmung studentischer Kompetenzen rechnerisch zu berücksichtigen, diese sind jedoch voraussetzungsvoll und können nicht ohne weiteres improvisiert werden. Da sich eine Prüfung vorrangig an den Erfordernissen des Fachs und der Lernziele orientieren sollte, kann die statistische Konsequenz nicht ausschlaggebend für die Gestaltung sein. In gewissen Fällen hat man sich also mit solchen Testlet-Effekten durch Aufgabenbündel zu arrangieren.

Zwei wichtige, grundsätzliche Erwägungen sollten bei Prüfungen mit

Aufgabenbündeln bedacht werden. Erstens ist die Verletzung der lokalen Unabhängigkeit innerhalb solcher Bündel zwar bis zu einem gewissen Grad unvermeidbar, keinesfalls sollte sich eine solche Abhängigkeit aber auf andere Aufgabenbündel erstrecken. Die *Aufgabenimpulse von Testlets sollten sich untereinander also deutlich unterscheiden*.[20] Zweitens kann der Verzerrungseffekt innerhalb von Aufgabenbündel dadurch reduziert werden, dass man den Studierenden bei aufeinander aufbauenden Arbeitsschritten *alternative Zwischenergebnisse* an die Hand gibt, anhand derer sie im Zweifelsfall weiter arbeiten können. So ist ausgeschlossen, dass Studierende gewisse Fähigkeiten nicht demonstrieren können, nur weil sie an anderer Stelle nicht zum Erfolg kamen.

Weiter sind auch die folgenden Hinweise bei Aufgabenbündel sinnvoll:[21]

- *Kreuzinformationen zwischen Aufgabenbündel* sind möglichst zu vermeiden. Das bedeutet, dass Aufgabenstellungen oder Teilergebnisse eines Testlets nicht die Lösungswahrscheinlichkeit eines anderen Testlets beeinflussen können sollten.

- *Unausgeglichener Inhalt ist zu vermeiden*. Wenn die zu prüfende Fähigkeit (z. B. »Thermodynamische Kompetenz«) verschiedene Inhaltsbereiche und Arbeitsformen abdeckt, sollten sich diese ausgewogen wiederfinden.

- Des Weiteren sollen im Falle *lebensweltbezogener Problemschilderungen* keine Studierendengruppen systematisch benachteiligt werden, indem etwa Aufgabenimpulse zu stark auf ein gewisses Geschlecht, Alter oder eine gewisse soziale oder geographische Herkunft aufbauen.

Bisweilen stellt sich bei der Konstruktion von Aufgabenbündeln die Frage, welcher Umfang und Komplexitätsgrad angemessen sind. Aus Perspektive der Kompetenzmodellierung kann es hierzu vorab keine Aussage geben. Maßgeblich dafür ist vielmehr das *latente Konstrukt*, also die zu prüfende

[20] vgl. Wainer, Bradlow und Wang 2007, S. 64.
[21] vgl. ebd., S. 55 f.

Personenfähigkeit. Aus den Inhalten, typischen Arbeitsweisen und den geforderten Taxonomiestufen ergibt sich die Ausgestaltung von Aufgabenbündeln.[22] Wie bereits erwähnt ist der Grad der verlangten Selbststrukturierung hierbei eine wichtige Stellschraube.[23] Je höher das von den Studierenden erwartete Kompetenzniveau ist, desto größer wird dann die Zahl der Bearbeitungsschritte werden und desto geringer wird die Vorstrukturierung des Lösungsprozesses durch den Aufgabentext.

9 Eindimensionalität der Prüfung

Die Eindimensionalität einer Prüfung bedeutet, dass eine Klausur nur die gewünschte Fähigkeit abprüft, also zum Beispiel »Thermodynamische Kompetenz«. Damit ist nicht gesagt, dass eine Prüfung nur ein Thema oder eine Methode zum Gegenstand hat. Vielmehr ist es die Regel, dass sich eine Kompetenz, beschrieben durch eine Reihe von Lernzielen, als eine Anzahl von methodischen Handlungsweisen darstellt, die sich in ausgewählten Themenbereichen des Fachs zur Anwendung bringen lassen. Insofern definieren die Lernziele und das Prüfungsdesign, was als »eindimensional« zu verstehen ist. Darin können sehr unterschiedliche Themen und Methoden enthalten sein.

Eine Verletzung der Eindimensionalität liegt dann vor, wenn die Klausuraufgaben auch andere Kompetenzen abprüfen oder andere Fähigkeiten eine solche Rolle spielen, dass sie eine Interpretation der Prüfungsergebnisse gefährden.[24] Hier ist ein weit verbreitetes Problem der hohe Zeitdruck von Prüfungen sowie deren spezifische Situation *als Prüfung*. Unter schlechten Umständen können Zeitknappheit und Prüfungsstress die Prüfungsgüte erheblich mindern. Im Falle von Studierenden, deren Muttersprache nicht der Klausursprache entspricht, kann eine Prüfung ungewollt die Sprachkompetenz erfassen, so dass die eigentliche Fachkompetenz nicht angemessen durch die Prüfung festgestellt werden kann. Als drittes Beispiel kann der Einsatz von unterschiedlichen Fähigkeiten durch die Studierenden genannt

[22] vgl. Wainer, Bradlow und Wang 2007, S. 57.
[23] vgl. Reis 2015.
[24] vgl. Strobl 2015, S. 23 f.

werden. Vielleicht gibt es zwei Gruppen in der Lehrveranstaltung. Die eine löst Aufgaben anhand der erwarteten Problemlösekompetenz, die andere gelangt durch vorher nicht bedachte Hilfsstrategien auf den richtigen Weg.

Durch eine Kompetenzmodellierung lässt sich solch eine Mehrdimensionalität gegebenenfalls feststellen. Im Alltag der Prüfungsgestaltung helfen einige Faustregeln, die aber sehr effektiv sein können, um die Klausurgüte zu steigern:

- Klausuren sollten so konstruiert sein, dass *nur die angestrebten Lernziele mit vertrauten Arbeitsweisen* geprüft werden. Das setzt eine valide Abstimmung zwischen Klausurinhalte und Fachkompetenz voraus, gefolgt von einer Transparenz gegenüber den Studierenden.
- Klausuraufgaben sollten so bemessen sein, dass aufgrund von *Zeitdruck oder Sprachbarrieren* die kompetente Beantwortung der Aufgabe nicht verhindert wird.
- Es sind *universell vertraute Situationsschilderungen* im Aufgabentext zu verwenden. Es sollten keine Textimpulse gewählt werden, die Teile der Studierenden absehbar kognitive, soziale oder emotionale Verständnisprobleme bereiten.[25]
- Studierende können gegebenenfalls auf ihre tatsächlichen Lösestrategien hin befragt werden. So wird festgestellt, ob Prüfungsaufgaben tatsächlich in der erwarteten Weise angegangen und gelöst werden, oder vielleicht unerwartetes Wissen oder unerwartete Methoden zum Tragen kommen.

10 Korrekturrichtlinien und differenzierte Datenerhebung

Insbesondere beim Einsatz mehrerer Korrektorinnen und Korrektoren für die Bewertung von Prüfungen ist eine Abstimmung über die Kriterien ganz

[25] vgl. Wainer, Bradlow und Wang 2007, S. 48 f.

zentral. Es empfiehlt sich nachdrücklich, eine gemeinsame Verständigung darüber zu treffen, was eine hinreichende Leistung für eine Teilaufgabe ist, welche Lösungen mustergültig sind und wann ein Punktabzug erfolgt. Die Objektivität der Korrektur kann anhand von Beispielfällen erprobt werden, in denen verschiedene Korrektoren die gleiche Klausur korrigieren und die Bewertungen untereinander im Detail vergleichen.

Sowohl bei mehreren Personen, aber auch bei einer Korrektur in Eigenverantwortung, ist das Ausformulieren einer Korrekturrichtlinie hilfreich. Sie hilft der Reflexion auf eigene Standards und Erwartungen, erlaubt eine spätere Referenz bei Folgeklausuren und kann auch als didaktisches Mittel für die Studierenden aufbereitet werden. In vielen Fällen (insbesondere im technisch-naturwissenschaftlichen Bereich) wird die Korrekturrichtlinie die Form einer Musterlösung annehmen. [26]

Im Zuge einer Abstimmung über Bewertungsrichtlinien findet auch eine Punkteverteilung statt. Es wurde oben bereits auf die Problematik der gewichteten Summenscores hingewiesen. In den Richtlinien sollte klar geregelt sein, welche Teilleistung wie viel Punkte wert ist und was zu Punktabzug führt. Hier ist der Hinweis zu ergänzen, dass probabilistische Verfahren der Testmodellierung in der Regel auf eine einfache Codierung von Ergebnissen ausweichen, die keine Gewichtung mehr umfasst. Eine bestandene Teilaufgabe erhält dann zum Beispiel eine 1, unabhängig davon, wie viel Punkte die Aufgabe den Klausurgestalterinnen »wert« war. Dies lässt sich dadurch rechtfertigen, dass in probabilistischen Modellierungen die Schwere einer Aufgabe statistisch angemessen berücksichtigt wird. Beim einfachen Summieren aller Einser wäre das irreführend. Dennoch kann es sehr erhellend sein, die Prüfungsergebnisse nach einem simpleren Schema zu recodieren. [27] Dabei wird die gewichtete Punktesumme einer Teilaufgabe

[26] Auch bei der Anwendung von Musterlösungen zur Klausurkorrektur muss im konkreten Einzelfall eine Korrektorentscheidung getroffen werden. So ergeben sich schnell Spielräume oder unklare Graubereiche, in denen ein Punktabzug oder eine Punktvergabe befürwortet oder abgelehnt werden kann. Eine Musterlösung alleine ist also unter Umständen nicht ausreichend für ein einheitliches Vorgehen einer Gruppe von Korrektorinnen und Korrektoren.

[27] Als Recodierung bezeichnet man in der empirischen Sozialforschung das Ersetzen von Datenwerten mit normierten Größen. Ziel ist es dabei, die Auswertung der Daten in mathematischen Verfahren oder grafischen Darstellungen zu erleichtern.

mit einem neuen Zahlenwert versehen, zum Beispiel mit 0 für »nicht gelöst«, 1 für »teilweise gelöst« und 2 für »vollständig gelöst«. Auch die Codierung nach 0 und 1 kann schon sehr aussagekräftig sein. Berechnet man für eine solche Recodierung dann die Randsummen, wie es oben erläutert wurde (S. 13), erhält man einen guten Eindruck davon, welche Aufgaben die schwierigeren waren und welche Personen die meisten vollständigen Bearbeitungen aufweisen. Aufgaben, die *zu einfach oder zu schwer* sind, also immer oder extrem selten komplett gelöst werden, sind dann in Zukunft zu vermeiden. Weiterhin können über solche Randstatistiken (also die Summen erfolgreicher Bearbeitungen nach Teilaufgaben und Personen) aussagekräftige Verteilungen schnell und einfach erstellt werden (Histogramme, Balkendiagramme). Auch ohne viel Vorwissen oder Aufwand hilft dies der Interpretation von Prüfungsergebnissen sehr.

Im Falle von Aufgabenbündeln (Testlets), die gemeinsam als eine umfassende Teilaufgabe gewertet werden, sollte zumindest für die Korrektoren eine *klare Ausweisung von Teilschritten* erfolgen. Wenn also eine umfangreiche Aufgabe in der Prüfung nur aus einer einzigen Frage besteht (»Welches der beiden geschilderten Verfahren ist effizienter?«), besteht sie nach ihrer inneren Logik vielleicht doch aus sehr klar abgrenzbaren Teilschritten. Diese Teilschritte stehen zwar nicht auf dem Aufgabenblatt, sind aber für die Lösung nötig. In diesem Fall kann es für mehr Transparenz bei der Bewertung sorgen, wenn der Korrekturleitfaden eine differenzierte Punkteverteilung innerhalb dieser einen, großen Aufgabe vorsieht und die digitale Erfassung auch entsprechend erfolgt.

Das geschilderte Verfahren macht auf jeden Fall eine *differenzierte Datenerhebung* notwendig. Vom Rotstift auf dem Klausurbogen bis zur Notentabelle in Excel sollte möglichst wenig Information verloren gehen. Es rentiert sich hier bei der Digitalisierung von Prüfungsergebnissen die erreichten Teilpunkte für jede Teilaufgabe mit zu erfassen, nicht nur die Gesamtsumme. So erhält man sich eine große Flexibilität bei der Ergebnisbeurteilung.

11 Niveaumodellierung und begründete Bestehensgrenzen

Wünschenswert wäre es aus Sicht der Kompetenzmodellierung, dass die Grenze zwischen Bestehen und Nichtbestehen anhand tatsächlicher, qualitativer Leistungsunterschiede festgelegt werden kann. Es sollte deutlich geworden sein, dass dies durch die verbreitete Bewertung auf Grundlage von Punktesummen nicht immer möglich ist. Eine qualitativ begründete Bestehensgrenze ist gegeben, wenn Studierende nur dann eine bessere Note erhalten, wenn sie eine (schwierigere) Kompetenz erfolgreich zur Anwendung brachten – und nicht wenn sie an der ein oder anderen Stelle ein paar Punkte mehr sammeln konnten. Wenn dies gelänge, wäre nicht nur die Notenvergabe gehaltvoller, sie würde den Studierenden zugleich auch eine qualitative Rückmeldung an die Hand geben, aus der sie ihre tatsächlichen Schwächen ablesen können. Erforderlich wäre es hierzu, eine umfassende Kompetenzmodellierung erfolgreich vorzunehmen und die Studierenden auch mit einem statistisch robusten Punktschätzer (also eine konkrete, persönliche Leistungsziffer) zu beurteilen.[28]

In der Praxis kann dies ohne Unterstützung aus der empirischen Bildungsforschung nur selten erfolgen. Daher muss mit Einschränkungen gelebt werden. Die hier dargestellten Anregungen können wichtige Schritte sein, dennoch die Güte von Klausuren zu verbessern. Die Begründung einer Notengrenze wird aber oft willkürlich bleiben. Der Reflex, eine Mindestleistung sowie eine Kür festzulegen, ist verständlich und vertretbar. Natürlich ist es wichtig anzugeben, ab welcher Leistung die Prüfung grade so bestanden beziehungsweise voll erreicht wurde. Problematisch ist, dass diese Grenze nach der Logik einer summenbasierten Bewertung irgendwie *quantifiziert* werden muss. Und diese Quantifizierung verwischt oftmals die *qualitative* Grenze zwischen Kompetenzniveaus erheblich. Das wurde in diesem Leitfaden wiederholt thematisiert. Praktisch wird man aber vorerst so verfahren müssen, bis sich fortgeschrittene (d. h. robustere) Kompetenzmodellierungen im Prüfungskontext flächendeckender umsetzen lassen.

[28] Beispielhaft für die Ingenieurwissenschaften vgl. Behrendt u. a. 2015, S. 6 ff.

12 Zur Quantifizierung von offenen Klausurformaten

Eingangs wurde gesagt, dass Klausuren ein Klassiker der Prüfungsformate sind und in sehr vielen Formen eingesetzt werden. Auch wenn die angeführten Hinweise und Praxisratschläge auf viele Prüfungen anwendbar sind, stand stillschweigend doch ein gewisser Typus im Fokus. Die auf quantifizierenden Verfahren beruhende Kompetenzmodellierung setzt eine Prüfung voraus, die aus einer mittleren bis größeren Zahl von klar abgegrenzten Teilaufgaben besteht, die man nach schematischen Kriterien bewerten kann. Eine Klausur im Bereich der Konstruktionslehre oder Sprachwissenschaften entspricht oftmals diesem Bild. Es gibt aber natürlich auch Klausuren, die eine viel offenere und weniger formalisierte Struktur der Antworten verlangen.

Klausuren mit sehr offenen Antworträumen ähneln Essays oder schriftlichen Erörterungen und sind wesentlich textorientierter als die bisher im Fokus genommenen Prüfungen. Eine Prüfungsfrage der Art »Welche Entwicklung hat die ethnologische Forschung des frühen 20. Jahrhunderts auf die Methoden der Rechtswissenschaften gehabt?«, lässt sich viel schwerer formalisieren als eine Multiple-Choice-Prüfung. Lassen sich die hier genannten Praxishinweise auch auf solche Prüfungen anwenden?

Die Antwort fällt zwiegespalten aus. Einerseits werden solche Prüfungen, die aus offenen Aufsätzen und sehr freien Antwortmöglichkeiten bestehen, selten für solch große Studierendengruppen eingesetzt, dass damit die quantitative Kompetenzmodellierung angemessen möglich ist. Zudem ist eine Codierung von korrekten Antwortaspekten in einem Aufsatz oder Essay viel schwerer anhand einer Musterlösung vorzunehmen, so dass Aspekte wie Gesamteindruck, Schlüssigkeit der Argumentation oder Grad der Reflexion kaum angemessen mit Punkten zu bewerten (und damit auch zu vergleichen) sind.

Andererseits ist eine methodisch verfahrende und reflektierte Prüfungsmodellierung gerade im unscharfen Feld freier, sprachlicher Ausdrucks- und Argumentationsweise enorm wichtig. Auch wenn Themen wie lokale Unabhängigkeit oder Summenbildung nur untergeordnete Rollen spielen,

findet an *irgendeinem Punkt der Bewertung eine Quantifizierung in Form einer Note* statt. Dem quantitativen Paradigma soll damit nicht das Wort geredet werden. Die Einsicht in eine gut kommentierte Klausur ist sicherlich gehaltvoller als eine (mehr oder weniger robuste) Leistungsziffer. Dennoch gibt es diesen magischen Punkt der Notenbildung zwangsläufig. Und genau hier können auch Gütekriterien der Klausurbewertung ansetzen, wie sie prinzipiell auch für eher quantifizierende Klausurtypen gültig sind:

- Korrekturrichtlinien sollten im Team erarbeitet und vielleicht auch mit den Studierenden erarbeitet werden. Kriterien der Prüfung sind transparent zu machen.
- Solche Korrekturrichtlinien gehen von formulierten Lernzielen aus und beginnen damit, dass Beispiele für Mindestleistungen und Höchstleistungen exemplarisch formuliert oder über Semester gesammelt werden. Diese Beispiele dienen als Referenz.
- Subjektive Kriterien (Bauchgefühl, Stimmigkeit des Textes, Überzeugungskraft) können durch Korrektorinnen und Korrektoren verschriftlicht werden. Man kann sie diskutieren und über die Zeit professionell in objektive Kriterien überführen.
- Gewisse Aspekte textueller, offener Klausurantworten lassen sich auch quantifizieren. Dies ist vor allem bei formalen Kriterien der Fall, aber auch bei erforderlichen Argumenten, der vollständigen Aufzählung oder Berücksichtigung von inhaltlichen Aspekten usw.

Insgesamt sollte also insbesondere bei sehr offenen und weichen Prüfungskriterien eine reflektierte und möglichst im Team validierte Arbeitsweise angestrebt werden. Auch hier gilt wieder, dass sich daraus auch didaktische Mittel ableiten lassen. Indem Studierende nachvollziehen und verstehen, was zu einer erfolgreichen, schriftlichen Prüfung im Fach führt, kann zielgerichteter gelernt, bewertet und verbessert werden.

13 Zusammenfassung

Die bisher behandelten Aspekte einer Klausurgestaltung aus Perspektive bildungswissenschaftlicher Kompetenzmodellierung orientierten sich am intuitiven Gestaltungsprozess für eine Klausur. Eingangs wurden das Kompetenzmodell und die Lernziele als Ausgangspunkt behandelt. Davon ausgehend wurde der Begriff des constructive alignment eingeführt, um eine Abstimmung zwischen Zielen, Lernhandeln und Prüfung zu betonen. Die Rolle der Transparenz gegenüber den Studierenden und deren aktive Beteiligung war häufiges Thema. Die eigentliche Klausurgestaltung (Klausurspezifizierung) wurde dann in den Kontext eines prinzipiell fortgesetzten Prozesses gesetzt. Aufgaben für Klausuren werden entwickelt und stets auch weiterentwickelt. Die Annahmen über Schwere und Kompetenzanforderung von Prüfungsaufgaben sind mit der empirischen Wirklichkeit immer wieder abzugleichen. Diesem Prozess sollte eine gewisse Systematik anhand der Lernziele und Taxonomiestufen inne wohnen. Dann wurde die Be- und Auswertung in den Blick genommen. Das verbreitete Verfahren einer Summenbildung zur Notengrundlage wurde kritisiert. Hier konnten einige Anregungen gegeben werden, die Klausurergebnisse mit einfachen Statistiken etwas anders als üblich in den Blick zu nehmen. Ziel ist es, eine Vertrautheit mit den tatsächlichen Effekten der Prüfungsaufgaben zu entwickeln. Gerade bei großen Studierendenzahlen ist das intuitiv nicht fassbar. Mit den Themen »lokale Unabhängigkeit der Teilaufgaben« und »Aufgabenbündel« wurde ein zentrales Thema der Kompetenzmodellierung angesprochen. Ziel ist es letztlich immer, Prüfungsaufgaben zu gestalten, die möglichst trennscharf und frei von störenden Verzerrungen sind. Das setzte sich mit dem Thema der »Eindimensionalität« fort, wonach Prüfungen auch durch oftmals nicht erwartete Effekte wie Stress, Zeitdruck, Sprachbarrieren oder alternative Lösungskompetenzen an Güte einbüßen. Alsdann wurde die Arbeit der Korrektorinnen und Korrektoren angesprochen, denen eine besondere Rolle in jeder Prüfungsform zukommt. Korrekturrichtlinien können helfen die Qualität zu verbessern. Dazu gehören auch die Codierungsstrategie sowie Digitalisierung der Prüfungsdaten, die in einer möglichst umfassenden Weise erfolgen sollte. So erhält sich für die Prüfungsbeurteilung eine große Analyse- und Interpretationsgrundlage. Zuletzt wurde ein

Bezug zu Klausuren mit stark offenen und textorientierten Antwortformaten hergestellt.

Im Folgenden sind die wichtigsten der besprochenen Handlungsempfehlungen zur Klausurgestaltung noch einmal zusammengefasst:

- Prüfungsaufgaben sollten nur die *angestrebten Kompetenzen und Lernziele* adressieren. Diese Prüfungsziele sind sachlich zu begründen und transparent zu machen.
- *Kreuzinformationen* zwischen Aufgaben und Aufgabenbündeln sind zu vermeiden. Eine Aufgabe(-nstellung) darf keinen inhaltlichen Vorteil bei einer anderen Aufgabe erlauben.
- Aufgabenbündel (Testlets) können sich auf einen *gemeinsamen Aufgabenimpuls* beziehen. Unterschiedliche Testlets sollten sich allerdings deutlich unterscheiden.
- Die Entscheidung, ob eine Aufgabe kleinschrittig vorgibt welche Arbeitsschritte zu gehen sind oder ohne Hinweise auf Teilschritte gestellt wird, ist alleine von der *Taxonomiestufe* (Komplexität der geforderten Leistung) abhängig.
- Prüfungsaufgaben sollten sich in einem vertrauten und im Lernprozess *erschlossenen Erwartungsraum* bewegen.
- Aufgaben und Teilaufgaben sollten *unabhängig voneinander lösbar* sein, so dass ein Fehler oder eine fehlende Antwort nicht zu Folgefehlern oder Nichtbearbeitung führt.
- Zeitdruck, Sprachbarrieren und ähnliche Faktoren verzerren die Einschätzung der Kompetenz und verletzen die *Eindimensionalität* der Messung.
- In Prüfungstexten sind *universell vertraute Situationsschilderungen* zu verwenden, die keine Studierenden aufgrund kognitiver, sozialer oder emotionaler Barrieren ausschließen.

- Als *Alternative zu gewichteten Summenwerten* können Teilaufgaben nach den Kriterien 0 (»unzureichende Leistung«) oder 1 (»ausreichende Leistung«) codiert werden. Die Betrachtung von Randverteilungen liefert dann wichtige Aufschlüsse über Eigenschaften der Prüfung.
- Jede Prüfungsaufgabe muss *klare Regeln zur Codierung* besitzen. Auch unstrukturierte Aufgaben ohne vorgegebene Teilschritte sind im Korrekturleitfaden klar zu differenzieren.
- Alle Teilleistungen sollten bei der Korrektur *digital erfasst* werden.
- Es ist ein *Aufgabenpool* anzulegen, der alle erwarteten Themenfelder und methodischen Fähigkeiten abdeckt und Aufgaben verschiedenen Schwierigkeitsgrades enthält. Der Pool kann als Grundlage für eine langfristige Prüfungsentwicklung dienen.

14 Ausblick und Anschluss

Es wurden einige Handlungsvorschläge für die Gestaltung von Klausuren vorgestellt, die sich aus der Perspektive einer empirischen Bildungswissenschaft ergeben. Kompetenzmodellierung im umfänglichen Sinne ist im Lehralltag ohne Unterstützung entsprechender Beratungsstellen oder Kooperationspartner oft nicht möglich. Dennoch lassen sich pragmatische Erkenntnisse umsetzen. Sinnvoll scheint dies, da die Kompetenzmodellierung von Prüfungen im Hochschulkontext an Bedeutung zu gewinnen scheint. Eine Hinführung zu diesem etwas anderen Blick auf die Klausurgestaltung war das Ziel dieses working papers.

Eine zentrale Botschaft ist es dabei, dass Prüfungen aller Art prinzipiell im Dialog mit den Studierenden entwickelt werden können. Damit sind einige Vorteile verbunden. Transparenz und Offenheit sorgen nicht nur für eine Prüfungsfairness sondern haben auch einen didaktischen Mehrwert. Die Bedeutung von Lerninhalten wird durch die Studierenden besser einzuordnen sein. Des Weiteren sorgt eine Beteiligungsoffenheit auch für eine bessere Einsicht in das Verhältnis von Lehre, Lernen und Prüfung. Wie *Prüfung und Lernen* zusammenhängen, lässt sich nicht einfach aus der

bloßen Korrekturleistung ablesen. Ein Dialog ist hierzu wichtig und kann dunkle Felder des Lehr-/Lerngeschehens erhellen.

Zuletzt gilt auch, dass die besten Erkenntnisse zum Lehr-/Lernverhalten und den statistischen Eigenschaften einer Prüfung ihren Mehrwert erst in der Anwendung auf die Gestaltung von Folgeprüfungen und der Anpassung didaktischer Maßnahmen entwickeln.[29] Prüfungsentwicklung ist also – wie die Lehre überhaupt – ein dynamischer Prozess. Die perfekte Klausur ist demnach eine im kontinuierlichen Werden begriffene.

[29] vgl. Thissen und Steinberg 1988, S. 385.

15 Literatur

Arbeitsstelle für Hochschuldidaktik der Universität Zürich (2008). *Lernziele formulieren in Bachelor- und Masterstudiengängen*. af. Bd. 1/2008. Dossier Unididaktik. ISSN 1662-579X. Zürich.

Behrendt, Stefan u. a. (2015). »Physical-technical prior competencies of engineering students«. In: *Empirical Research in Vocational Education and Training* (7:2 15.02.2015), S. 1–19.

Biggs, John B. (2003). *Teaching for quality learning at university. What the student does*. en. 2. Aufl. Zu Lehrzielen: "Formulating and clarifying curriculum objectives", S. 34-55. Buckingham und Philadelphia, PA: Society for Research into Higher Education und Open University Press. xii, 309. ISBN: 9780335211685.

Bloom, Benjamin Samuel und David R. Krathwohl (1969). *Taxonomy of educational objectives. The classification of educational goals : Handbook I, Cognitive domain*. en. Band 1 von 2. New York: McKay. 219 S.

Bühner, Markus (2011). *Einführung in die Test- und Fragebogenkonstruktion*. 3. Auflage. Pearson Studium. ISBN: 9783868940336.

Nickolaus, Reinhold und Susan Seeber (2013). »Berufliche Kompetenzen. Modellierungen und diagnostische Verfahren«. In: *Handbuch Berufspädagogische Diagnostik*. Hrsg. von Andreas Frey, Urban Lissmann und Bernd Schwarz. 1. Aufl. Pädagogik 2014. s.l.: Beltz, S. 166–195. ISBN: 978-3407831736.

Organisation for Economic Co-operation and Development (1999). *Measuring Student Knowledge and Skills. A New Framework for Assessment*. Paris: OECD Publications. ISBN: 92-64-17053-7.

Rapp, Sonja (2014). »Entscheidungshilfen zur Wahl der Prüfungsform. Eine Handreichung zur Prüfungsgestaltung«. In: *zlw working paper* (1). ISSN: 2363-8834.

Reis, Oliver (2015). *Prüfungsformate im Forschenden Lernen. Alles eine Frage des Learning-Outcomes*. Tagungsbeitrag zur nexus Tagung am 27.11.2015. Hohenheim.

Schaper, Niclas (o.D.). »Lernprozesse mit Instruktionsmethoden wirkungsvoll gestalten«. de. In: *Neues Handbuch Hochschullehre* (B 1.5). ISSN 2198-5693; http://d-nb.info/1049897242. ISSN: 2198-5693.

Strobl, Carolin (2015). *Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis*. ger. 3. erweiterte Auflage. Bd. 2. Sozialwissenschaftliche Forschungsmethoden. Matiaske, Wenzel (BeteiligteR) Fantapié Altobelli, Claudia (BeteiligteR). München und Mering: Reiner Hampp Verlag. 116 S. ISBN: 978-3957100504.

Thissen, David und Lynne Steinberg (1988). »Data Analysis Using Item Response Theory«. In: *Psychological Bulletin* 104 (3). Quantitative Methods in Psychology, S. 385–395.

Wainer, Howard, Eric T. Bradlow und Xiaohui Wang (2007). *Testlet response theory and its applications*. eng. Cambridge: Cambridge University Press. 267 S. ISBN: 978-0-521-86272-1.

Wildt, Johannes und Beatrix Wildt (2012). »Lernprozessorientiertes Prüfen im "Constructive Alignment. Ein Beitrag zur Förderung der Qualität von Hochschulbildung durch eine Weiterentwicklung des Prüfungssystems«. In: *Neues Handbuch Hochschullehre* (H 6.1). ISSN: 2198-5693.

16 Über den Autor

Thorsten Braun arbeitet am Zentrum für Lehre und Weiterbildung | zlw der Universität Stuttgart. Als Soziologe und Hochschuldidaktiker unterrichtet er hochschuldidaktische Grundlagen, Gruppendynamik und Interdisziplinarität für das Weiterbildungsangebot des zlw. Im Rahmen von Lehrcoachings und Kooperationen mit Fakultäten und Instituten betreut er Professorinnen und Professoren verschiedener Disziplinen. Besondere Arbeitsschwerpunkte sind hochschuldidaktische Begleitforschung, Wissenschaftstheorie, Forschendes Lernen und die Implikationen des bildungspolitischen Programms Lebenslangen Lernens (lifelong learning).