ETH zürich

Leitfaden zur Notengebung bei schriftlichen Prüfungen

LET – Lehrentwicklung und -technologie

Eine mit Hyperlinks versehene PDF-Version des Leitfadens ist abrufbar unter
www.let.ethz.ch/docs/Leitfaden_NotengebungDE_2013_11.pdf

Eine Checkliste für die Praxis findet sich in Anhang 1 sowie unter
www.let.ethz.ch/docs/Checkliste_NotengebungDE_2013_11.pdf

Impressum
Herausgeber: ETH Zürich, Lehrentwicklung und -technologie
Redaktion: Tobias Halbherr, Claudia Schlienger
Druck: FO-Fotorotar AG
Auflage: 3000 Exemplare

1. Ausgabe, November 2013

ETH Zürich
Lehrentwicklung und -technologie
Haldenbachstrasse 44
8092 Zürich
www.let.ethz.ch

Inhaltsverzeichnis

Vorwort des Rektors

Die Qualität der Lehre an der ETH Zürich hochhalten und sie wo immer möglich steigern, das ist mir ein grosses Anliegen – nicht zuletzt mit Blick auf den internationalen Wettbewerb. Dabei kommt dem Prüfungssystem und hier wiederum der Notengebung eine besondere Stellung zu. In der Notengebung kommen wichtige Themen der Lehrqualität zum Ausdruck, wie zum Beispiel der Zweck von Prüfungen, aber auch die Art der Selektion. Die Benotung bildet ab, welche Leistungen genügend, welche sehr gut, aber auch welche Leistungen ungenügend sind.

Die Notengebung liegt ganz in Ihrer Verantwortung als Dozierende/r der ETH Zürich. Mit diesem Leitfaden wollen wir Sie dabei unterstützen. Er ist kein Ersatz für den Austausch mit Kolleginnen und Kollegen. Vielmehr soll er einerseits neue Kolleginnen und Kollegen anleiten, beim Aufsetzen von Prüfungen die Notengebung gebührend zu berücksichtigen. Andererseits werden hier zentrale Punkte angesprochen, die allen Dozierenden in ihrer Benotungspraxis nützliche Hinweise geben.

Der Leitfaden zur Notengebung ist knapp gehalten und liefert dennoch eine umfassende Darstellung des Themas. Fünf Grundsätze der Notengebung werden anhand von konkreten Beispielen illustriert und die Checkliste im Anhang fasst alles Wichtige für den Alltag zusammen. Ein theoretischer Grundlagenteil mit weiterführenden Literaturhinweisen bietet zudem allen Dozierenden die Möglichkeit, ihre eigene Praxis zu überprüfen und sich weiter in das Thema zu vertiefen.

Ich wünsche mir, dass Sie alle genau dies tun und den Leitfaden als Instrument nutzen, um Ihre Lehre weiter zu optimieren.

Prof. Dr. Lino Guzzella, Rektor der ETH Zürich

Notengebung: Ein wichtige Herausforderung mit messtheoretischen Fallstricken

Prof. Dr. Elsbeth Stern,
Professorin für Lehr- und Lernforschung an der ETH

Noten beeinflussen Bildungs- und Berufsentscheidungen, und ganze Lebensentwürfe hängen von ihnen ab. Aus wissenschaftlicher Sicht ist dies nicht unproblematisch, da die Notengebung eine mit Fehlern behaftete Form der Leistungsmessung ist. Ganz allgemein kann man von einer Messung immer dann sprechen, wenn den unterschiedlichen Ausprägungen einer Variablen unterschiedliche Zahlen zugeordnet werden. Das trifft auf Masse und Grösse von Objekten genauso zu wie auf die in einem Intelligenztest gelösten Aufgaben oder die Notengebung. Aber natürlich unterscheiden sich diese Messungen in ihrer Qualität, und für diese haben Messtheoretiker klare Kriterien entwickelt. Sie sind ebenso Gegenstand dieses Leitfadens wie die Frage, welche Schlüsse aus den Messwerten gezogen werden dürfen, und welche statistischen Berechnungen interpretierbar sind. Physikalische Grössen wie Masse und Grösse haben einen definierten Nullpunkt und deshalb dürfen Messwerte in das Verhältnis gesetzt werden. Noten hingegen erlauben lediglich Aussagen über Rangplätze. Die Note 5 ist besser als die Note 4, aber eine 4 ist nicht doppelt so gut wie eine 2. Nicht einmal die Abstände zwischen den Noten darf man interpretieren, da Noten im mittleren Bereich ein breiteres Spektrum abbilden als Noten im Extrembereich. Einen arithmetischen Mittelwert darf man aus messtheoretischer Sicht bei Noten eigentlich genauso wenig interpretieren wie die Varianz. Gleichwohl ist die Bildung von Durchschnittsnoten gängige Praxis, auch wenn man sich eigentlich auf den Median beschränken müsste. Mit sehr viel Aufwand kann man auch im nicht-physikalischen Bereich Messungen vornehmen, die über die Bestimmung von Rangplätzen hinausgehen. Die Messung von Intelligenz, Persönlichkeitsmerkmalen oder Kompetenzen sind Beispiele. Hier wird jede einzelne Aufgabe einer intensiven empirischen Prüfung unterzogen, um sicher zu stellen, dass sie ein einheitliches Konstrukt messen. Nur Aufgaben mit hoher Trennschärfe, die Personen mit einer hohen und niedrigen Ausprägung voneinander unterscheiden, werden beibehalten. Idealerweise folgt ein professioneller Test der Rasch-Skalierung: Personen die ein schwieriges Item gelöst haben, haben mit sehr hoher Wahrscheinlichkeit auch alle leichteren Aufgaben gelöst. Die Intervalle zwischen den Messwerten dürfen in diesem Fall interpretiert werden. Von solchen Qualitätsmerkmalen sind die selbstgestrickten Tests von Lehrern weit entfernt, und die Notengebung wird von Faktoren beeinflusst, die mit der Leistung nichts zu tun haben. Ein Beispiel ist die Bezugsnorm: Mit grosser Wahrscheinlichkeit wird die gleiche Arbeit in einer leistungsstärkeren Klasse strenger benotet als in einer leistungsschwächeren.

Auch aus lernpsychologischer Sicht spricht vieles gegen Notengebung. Diese wirkt sich negativ auf das Vertrauensverhältnis zwischen Lehrenden und Lernenden aus. Gerade in mathematisch-naturwissenschaftlichen Fächern, wo ein tiefgehendes Begriffsverständnis unabdingbar ist, zeichnet sich lernwirksamer Unterricht durch Fehlertoleranz auf. Lehrpersonen müssen sich mit den hinter den Fehlern liegenden Missverständnissen befassen und deshalb ein Klima schaffen, in dem die Lernenden keine Scheu haben, ihre Fehler zu offenbaren. In der Klausur ist es dann mit der Fehlertoleranz vorbei. Auch setzt Notengebung falsche Anreize und Akzente, die sich in der Unterscheidung von Lern- und Leistungsorientierung verdeutlichen lassen. Nicht das Verstehen des Stoffs steht im Mittelpunkt, sondern die Frage, wie man mit minimalem Aufwand die beste Note bekommt. Angesichts der vielen Probleme, die mit der Notengebung einhergehen, ist es nicht verwunderlich, dass Korrelationen zwischen professionell konstruierten Leistungstests und den Noten in dem entsprechenden Fach selten r=.40 übersteigen.

Trotz der Unzulänglichkeiten in der Notengebung ist es schwierig, das Unterrichten und die Leistungsmessung zu trennen und letztere an professionelle Testentwickler auszulagern. Dies kann gelingen, wenn es zu dem Themenbereich einen guten Test gibt, der laufend aktualisiert wird. Aber gerade im tertiären Bildungsbereich ist dies selten der Fall, und daran wird sich angesichts des grossen Stoffumfanges auf absehbare Zeit nichts ändern. An Universitäten werden wir weiterhin gleichzeitig als Lehrende und als Prüfer agieren müssen. Deshalb ist es umso wichtiger, die mit der Notengebung verbundenen Fallstricke zu kennen und so die Probleme abzumildern. Dazu kann der Leitfaden einen wichtigen Beitrag leisten.

Kapitel 1: Grundsätze zur Notengebung an der ETH

Folgendes gilt für alle Formen von Leistungskontrollen an der ETH Zürich

- Die einzelnen Dozierenden sind verantwortlich für die Prüfung. Sie verantworten insbesondere die inhaltliche Korrektheit und die methodische Angemessenheit von Prüfungsaufgaben, Prüfungsbewertung und Notengebung.
- Prüfungen sind Instrumente zur Messung der Lernzielerreichung oder zur Einschätzung des Potentials für die künftige Lernzielerreichung.
- Prüfungen genügen den Grundsätzen Aussagekraft, Fairness, Transparenz, Lerndienlichkeit und Verhältnismässigkeit, die im Folgenden beschrieben werden.
- Design, Durchführung, Auswertung und Notengebung einer Prüfung orientieren sich an praxistauglichen und wissenschaftlich fundierten Methoden und Standards.
- Formale Aspekte von Leistungskontrollen sind im Leitfaden für Dozierende[1] beschrieben. Reglementarische Grundlage ist die Verordnung über Lerneinheiten und Leistungskontrollen (VLK)[2] sowie deren Ausführungsbestimmungen[3].

Grundsätze für qualitativ hochwertige schriftliche Prüfungen[4]

1. Aussagekraft
Die Prüfung ist inhaltlich gültig (valide), genau und überprüft die Lernziele objektiv. Die Note ist eine sinnvolle normative Einschätzung der gesamten Prüfungsleistung.

a. Inhaltliche Gültigkeit (Validität): Prüfungsaufgaben ermöglichen die Überprüfung der in den Lernzielen festgehaltenen Kompetenzen auf inhaltlich gültige und methodisch zulässige Weise. Prüfungsaufgaben stehen in engem Bezug zu diesen Kompetenzen sowie zu den entsprechenden Lehr- und Lernaktivitäten. Die Aufgaben repräsentieren inhaltlich die Gesamtheit der Lernziele. Die Note ist ein sinnvoll gewichteter Kennwert der gesamten Lernzielerreichung. Der kognitive Prozess zur Lösung der Aufgaben, entspricht dem kognitiven Prozess, welcher dem entsprechenden Lernziel zugrunde liegt. Geprüft wird Wesentliches. Spitzfindigkeiten, komplizierte und missverständliche Formulierungen oder das Fokussieren auf nebensächliche Details werden vermieden.

b. Genauigkeit (Reliabilität): Die Prüfung ist ausreichend ausführlich und differenziert angemessen zwischen unterschiedlichen Leistungen. Sie unterscheidet Leistungen im kritischen Bereich (genügend/ungenügend) am genauesten. Repetitionsprüfungen bleiben vom Schwierigkeitsgrad vergleichbar.

[1] Siehe: http://www.ethz.ch/faculty
[2] Siehe: http://www.rechtssammlung.ethz.ch/pdf/322.021_leistungskontrollenverordnung_eth_zuerich.pdf
[3] Siehe: http://www.rektorat.ethz.ch/directives
[4] Die Grundsätze gelten auch für Prüfungen, die am Computer abgelegt werden.

c. **Objektivität:** Die Überprüfung der Lernzielerreichung ist unabhängig von den Durchführungs- und Auswertungsumständen und erfolgt unter einheitlichen Bedingungen. Unterschiede in der Bewertung von Prüfungsaufgaben sind Ausdruck tatsächlicher Leistungsunterschiede der Studierenden und nicht Ausdruck unterschiedlicher Bewertungskriterien verschiedener Examinatoren. Subjektive Einflüsse auf die Bewertung einer Prüfung werden möglichst ganz vermieden oder zumindest minimiert.

d. **Notenfestlegung:** Der Notenschlüssel wird so festgelegt, dass die Note eines Studierenden nicht von den Leistungen der anderen Studierenden abhängt. Nur genügende Noten bedeuten eine ausreichende Lernzielerreichung, die beste Note 6 muss erreichbar sein. Gleiche Notenunterschiede (z. B. 3 vs. 4 oder 5 vs. 6) spiegeln vergleichbare Unterschiede in der Lernzielerreichung wider.

2. Fairness

Die Studierenden sind in Bezug auf Inhalte, Durchführung und Auswertung der Prüfung keiner Willkür ausgesetzt. Alle Studierenden finden an der Prüfung die gleichen Bedingungen vor. Der gleiche Zugang zu Lerninfrastruktur und Lerninhalten wird gewährleistet. Die Prüfungen der Studierenden werden nach einheitlichen und objektiven Kriterien beurteilt.

Die Prüfungsumstände sind dem Abrufen persönlicher Spitzenleistungen zuträglich. Störungen, Ablenkungen oder anderweitige Beeinträchtigungen während der Prüfung werden vermieden. Nur die in den Lernzielen formulierten Kompetenzen sowie dafür inhärent notwendige Voraussetzungen haben einen Einfluss auf die Prüfungsleistung und -bewertung. Keinen Einfluss haben äusserliche und lernzielirrelevante Faktoren, wie soziodemographische Merkmale, Wertvorstellungen oder Gesinnungsfragen.

Die Prüfung sowie die Prüfungsresultate müssen in Vorbereitung, Durchführung und Nachbereitung vor Betrug geschützt werden. Prüfungen werden zuverlässig und fehlerfrei abgewickelt.

3. Transparenz

Die Studierenden kennen die inhaltlichen und formalen Anforderungen einer Prüfung. Diese Informationen sind leicht zugänglich, vollständig, verständlich und verbindlich. Grundlage ist der Eintrag im Vorlesungsverzeichnis. Prüfungen beziehen sich auf die kommunizierten Lernziele: Die Kompetenzen, welche von den Studierenden erwartet und geprüft werden, sind konkret, anschaulich und vollständig formuliert, insbesondere Stoffumfang sowie zugehörige kognitive Niveaus. Auch die erforderlichen Vorkenntnisse werden grob umschrieben und bekanntgegeben. Prüfungsform und -ablauf sind bekannt. Leistungskriterien und formale Antwortstruktur werden im Voraus festgelegt und kommuniziert. Den Studierenden wird bereits vor der Prüfung eine Rückmeldung zu ihrem Kompetenzniveau ermöglicht, zum Beispiel durch Übungen, Quizzes oder frühere Prüfungen.

4. Lerndienlichkeit

Prüfungen dienen der Ausbildung von Studierenden. Prüfungsaufgaben sollen in Form, Inhalt und Anspruch den anvisierten Kompetenzen entsprechen und in engem Zusammenhang mit Lehr- und Lernaktivitäten stehen. Damit schaffen sie Anreize, sich im Rahmen der Prüfungsvorbereitung die angestrebten Kompetenzen auf dem anvisierten kognitiven Niveau zu erarbeiten. Eine gute Prüfung ermöglicht Studierenden, herausragende Fertigkeiten zu zeigen und motiviert sie, über sich hinauszuwachsen. Prüfungen sind ein Anlass für Feedback und helfen Stärken oder Kompetenzlücken zu identifizieren. Sie dienen der Selektion und stellen sicher, dass die Studierenden der ETH Zürich den Leistungsanforderungen genügen. Prüfungen dienen dem Nachweis erreichter Lernziele.

5. Verhältnismässigkeit

Die Lernzielerreichung wird auf unmittelbar plausible und glaubhaft aussagekräftige Art und Weise überprüft. Aufwand und Umstände für Examinatoren und Studierende stehen in sinnvollem Verhältnis zum Nutzen bzw. zur Relevanz der Prüfung.

Kapitel 2:
Gute Praxis zur Notengebung

Die anschliessenden Erläuterungen zur Notengebung zielen darauf ab, jeweils einen oder mehrere der in Kapitel 1 formulierten Grundsätze zu gewährleisten.

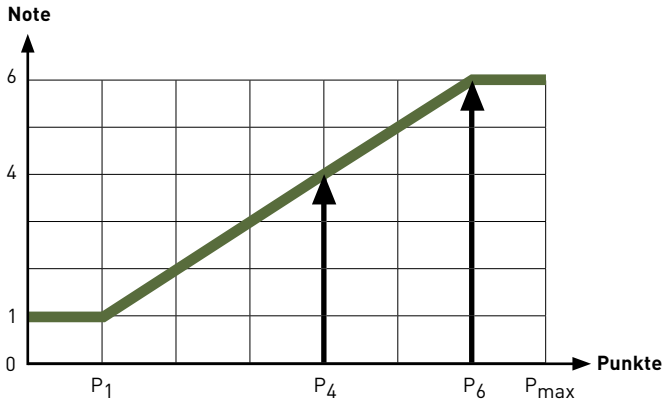
2.1 Notenmassstab und Punktevergabe

2.1.1 Festlegen des Notenmassstabs

Durch den Notenmassstab wird die Prüfungsleistung in Relation zu einer Bezugsnorm gesetzt. An der ETH Zürich soll kriteriumsorientiert benotet werden.

Die Abbildung der Prüfungspunkte auf die Notenskala wird von der 4 (und der 6) her verankert. Bei kriteriumsorientierten Prüfungen wird inhaltlich begründet, wie viele Punkte einer genügenden Leistung (Note 4), und wie viele einer herausragenden Leistung (Note 6) entsprechen. In der Regel können alle übrigen Noten aufgrund dieser zwei Punkte inhaltlich zufriedenstellend linear interpoliert werden. Alternativ werden die ungenügenden Noten separat interpoliert, indem man für die Note 1 ebenfalls eine erforderliche Punktzahl definiert. Die notwendige Anzahl Punkte für die Noten 4 und 6 wird vor der Prüfung festgelegt und den Studierenden mitgeteilt.

Einfache lineare Interpolation



Doppelte lineare Interpolation

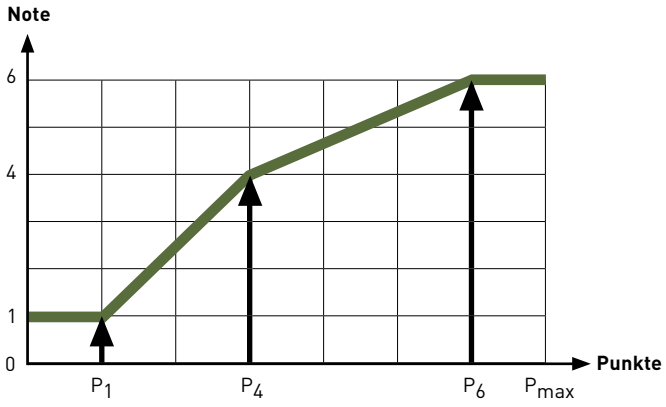


Abbildung 1: Berechnung der Noten

2.1.2 Festlegen der Note 4

Die Unterscheidung zwischen einer genügenden und einer ungenügenden Leistung ist essenziell. Die entsprechende Punktzahl muss vor der Prüfung feststehen und bereits bei der Entwicklung der Fragen beachtet werden. Die Note 4 bedeutet, dass die Lernziele gerade noch ausreichend erreicht wurden und darf sich nicht an der Leistung der anderen Studierenden orientieren. Wo genau die Grenze zu einer 4 festgelegt wird, bleibt eine Ermessensfrage. An folgenden inhaltlichen Fragen kann man sich beim Festlegen orientieren:

- Welche Kompetenzen bilden den Kern der Lernziele?
- Welche Leistung widerspiegelt eine minimal ausreichende Beherrschung dieser Kernkompetenzen respektive schliesst diese aus?
- Welche Leistung widerspiegelt eine Beherrschung der Kompetenzen, die zeitlich Bestand hat?
- Wie gut müssen die Kompetenzen beherrscht werden, damit in weiteren Lerneinheiten auf diesen aufgebaut werden kann?
- Welches Ausmass an Missverständnissen sowie fehlerhaft oder falsch Erlerntem verhindert eine ausreichende Lernzielerreichung?
- Welche Leistungen wurden bisher als genügend eingeschätzt?
- Welche Leistungen werden in vergleichbaren Prüfungen als genügend eingeschätzt?

Dabei sollte bereits bei der Aufgabenkonstruktion, der Definition des Bewertungsschemas und der Festlegung der Musterlösungen bedacht werden, was einer genügenden oder ungenügenden Leistung entspricht.

Die folgende **Methode** hilft, die Punktzahl für die 4 festzulegen:

- Die Prüfungsaufgaben werden einerseits mit den Lernzielen und andererseits mit Übungsaufgaben und früheren Prüfungsaufgaben verglichen. Die Lernziele geben vor, was von den Studierenden erwartet wird. Übungen und frühere Prüfungen helfen abzuschätzen, was von den Studierenden erwartet werden kann.
- Aufgrund dieser Vergleiche wird geschätzt, welche Punktzahl gerade noch als eindeutig genügend (Pg), respektive eindeutig ungenügend (Pu) zu betrachten ist. Die Punkte für die Note 4 liegen irgendwo dazwischen. Bei Prüfungen, die nicht primär der Selektion dienen («low stakes») werden die Punkte bei Pu + 1 gesetzt, im Zweifelsfalle soll die Prüfung bestanden werden.

Für präzisere Ergebnisse schätzen mehrere Personen die Punktzahl nach dieser Methode und die Punktzahl wird für die Teilbereiche der Prüfung einzeln festgelegt.

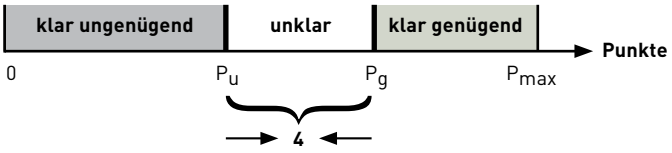


Abbildung 2: Ermittlung der Punktzahl für die Leistung «genügend»

2.1.3 Festlegen der Note 6 (und 1)

Anspruchsvolle aber eigenständig erreichbare Ziele sind starke Motivatoren. Das Erreichen einer 6 soll anspruchsvoll, aber machbar sein – und auch tatsächlich von Studierenden erreicht werden. Jedoch soll die Note 6 nicht einfach dem / der besten Studierenden gegeben werden, sondern soll eine ausserordentliche Leistung auszeichnen. Eine zu leichte Vergabe von guten Noten «entwertet» diese und schwächt ihre positiven Effekte ab.

Möchte man die ungenügenden Noten unabhängig von den genügenden interpolieren, muss zusätzlich eine ungenügende Note, zumeist die 1, verankert werden. Am einfachsten ist, wenn man für 0 Punkte die 1 setzt und linear bis zur 4 interpoliert. Je nach inhaltlichem Aufbau der Prüfung ist es auch möglich, die 1 erst für mehr als 0 Punkte zu geben. Enthält die Prüfung Multiple Choice Fragen, muss die Ratewahrscheinlichkeit berücksichtigt werden.

2.1.4 Punktevergabe: Quantifizierung von Prüfungsaufgaben

Prüfungsaufgaben werden in Punkten quantifiziert. Die Anzahl vergebener Punkte kann sich an folgenden Aspekten orientieren:

- die für einen Experten zur erfolgreichen Bearbeitung der Aufgabe erforderliche Zeit,
- der benötigte Zeitaufwand für den Erwerb der überprüften Kompetenz und/oder
- die Relevanz des überprüften Lernziels.

Im Idealfall entsprechen sich in der Prüfung alle drei Aspekte. Die für eine Aufgabe erforderliche Prüfungszeit entspricht also im Verhältnis in etwa der Relevanz des überprüften Lernziels sowie dem benötigten Zeitaufwand für den Erwerb desselben. Die subjektive Schwierigkeit einer Aufgabe, das heisst die Sicht einzelner Studierender, sollte keinen direkten Einfluss auf die Anzahl Punkte haben.

2.1.5 Punkte für Teilaufgaben

Die Gesamtanzahl Punkte für eine Aufgabe kann ausdifferenziert werden. Dafür gibt es drei Möglichkeiten:

1) Die Aufgabe wird in **Teilaufgaben** unterteilt.
2) Die (Teil-)Aufgabe wird nach mehreren **Kriterien** bewertet.
3) Die Leistung pro (Teil-)Aufgabe und/oder Kriterium wird **gestuft bewertet,** d. h. es werden je nach Leistung eine unterschiedliche Anzahl Teilpunkte vergeben.

Diese drei Möglichkeiten können kombiniert werden (siehe Abbildung 3).

Werden Teilaufgaben gebildet, gilt zu beachten, dass für jede Teilaufgabe immer die volle Teilpunktzahl erzielt werden kann, auch wenn andere Teilaufgaben (und damit Zwischenergebnisse) falsch sind.

Schliesslich können die einzelnen Teilaufgaben bzw. Kriterien gewichtet werden, indem ihre Teilpunkte zu unterschiedlichen Anteilen zur Gesamtpunktzahl der gesamten Aufgabe beitragen.

Eine zu feine Ausdifferenzierung der Punkte 1) bis 3) ist in der Regel nicht sinnvoll und kann die Objektivität der Prüfung beeinträchtigen. Insbesondere wird bei gestufter Bewertung eine höchstens fünfstufige Unterteilung (d. h. 0 bis max. 4 Teilpunkte) empfohlen.

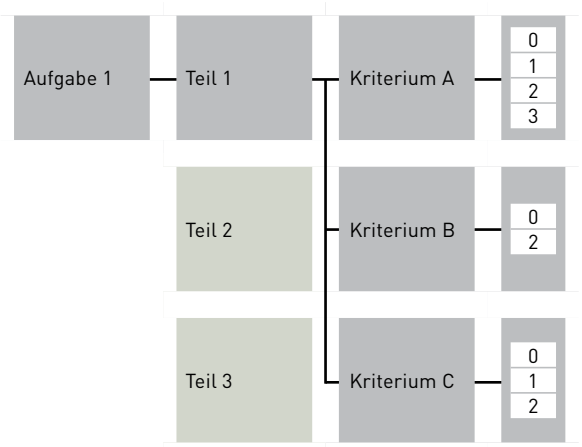


Abbildung 3: Ausdifferenzierung von Aufgaben und Punkten

2.1.6 Punktvergabe bei Multiple Choice Fragen

Bei Multiple Choice Fragen gibt es zwei geläufige Fragetypen:

- Bei **One-best-answer** Fragen ist eine von üblicherweise vier oder fünf Wahlantworten eindeutig die richtige oder beste Antwort. Die übrigen sind eindeutig falsch oder schlechter. Es werden nur Punkte für die Wahl der richtigen / besten Antwortalternative vergeben. Es werden keine Teilpunkte für die zweitbeste Antwort und keine Strafpunkte für falsche Antworten vergeben.

- **Wahr / Falsch Fragen** sind als Mehrfachwahlfragen im K-Prim (K') Format am geläufigsten. Dabei müssen vier Wahr / Falsch Fragen beantwortet werden, von welchen eine beliebige Anzahl eindeutig richtig oder falsch sein kann. Für das korrekte Beantworten aller vier Teilfragen wird die volle Punktzahl vergeben, für drei korrekte Teilfragen die halbe Punktzahl, ansonsten null Punkte. Alternativ können bei Wahr / Falsch Fragen auch ein Punkt bei richtigen Antworten und null Punkte bei falschen vergeben werden.

Alle anderen Multiple Choice Formate und Auswertungsschemata werden ausdrücklich nicht empfohlen. Bei der Festlegung des Notenmassstabs werden die durch reines Raten im Mittel erreichbaren Punkte mit berücksichtigt.

2.1.7 Bewertungsschema und Musterlösung

Ein **Bewertungsschema** definiert das formale Raster für die Punktvergabe und erleichtert dadurch die Korrektur. Es hält in Form einer Tabelle fest, für welche Teile der Prüfung und nach welchen Kriterien, wie viele Punkte vergeben werden. So können beim Korrigieren die verschiedenen Prüfungsteile einheitlich gewichtet und nach einheitlichen Kriterien bewertet werden. Als Kriterien für die Punktvergabe können Antworten sowohl konkret festgehalten («drei von vier Eigenschaften wurden erwähnt») als auch qualitativ umschrieben werden («Der Sachverhalt wurde umfassend begründet»). Um eine objektivere und reliablere Bewertung zu ermöglichen gilt: «So konkret wie möglich, so offen wie nötig.»

Das Bewertungsschema[5] wird vor der Prüfung festgelegt, mit den Korrektoren besprochen und den Studierenden in angemessener Form bekannt gemacht. Bei der Korrektur der Aufgaben wird nachvollziehbar notiert, weshalb wie viele Punkte vergeben wurden.

[5] Für ein Beispiel eines Bewertungsschemas siehe: www.let.ethz.ch/docs/Beispiel_BewertungsschemaDE_2013_11.pdf

Die **Musterlösung** hält exemplarisch fest, wie sich die Aufgabensteller mustergültig bearbeitete Prüfungsaufgaben vorstellen. Die Musterlösung erfüllt dabei zwei Funktionen: Erstens dient sie zur Orientierung für die Prüfungskorrektur. Zweitens erlaubt sie den Studierenden, nach der Prüfung nachzuvollziehen, wie die Prüfungsaufgaben hätten gelöst werden können. Dies ermöglicht einen zusätzlichen, gezielten und nachhaltigen Lerneffekt über die Prüfung hinaus.

2.2 Korrektur und Auswertung

2.2.1 Den Weg oder das Ergebnis bewerten?

Sowohl das Ergebnis einer Aufgabe als auch der Weg zum Ergebnis können bewertet werden. Das Ergebnis ist meist einfacher und effizienter zu bewerten. Die Studierenden wissen, was in der Aufgabe von ihnen erwartet wird. Ausserdem werden auch ungewöhnliche, aber zielführende Problemlösungen angemessen belohnt. Das Bewerten des Bearbeitungsweges erlaubt eine differenziertere Bewertung von Leistungen, gerade wenn der Kompetenzerwerb noch nicht abgeschlossen oder unvollständig ist. Das Formulieren klarer Bewertungskriterien ist allerdings anspruchsvoller und das Korrigieren der Aufgabe im Allgemeinen zeitaufwändiger. Diese Methode birgt ausserdem die Gefahr, dass dem/der Korrigierenden nicht vertraute oder unsympathische Lösungsansätze nicht angemessen belohnt werden.

2.2.2 Korrigieren von Prüfungen: Störfaktoren ausschalten

Die Lernziele beschreiben, was geprüft wird. Alle übrigen Aspekte sind Störfaktoren und ihr Einfluss auf die Prüfungsbewertung soll minimiert werden. Beispiele solcher Störfaktoren sind die Tagesform der Korrigierenden, Unterschiede von Person zu Person in der Art zu bewerten oder die Qualität der Sprache und Handschrift in den Antworten. In Anhang 5 finden Sie eine ausführliche Auflistung dieser Störfaktoren.

Die folgenden sechs Massnahmen helfen, die Auswirkungen solcher Störfaktoren in Grenzen zu halten.

1. Prüfungen werden **anonymisiert** ausgewertet. Papierprüfungen lässt man z. B. an vorbestimmten Stellen mit der Legi-Nummer und dem Namen kennzeichnen, welche vor der Auswertung überklebt, abgedeckt oder weggefaltet werden können. Bei Online-Prüfungen gestaltet sich das Anonymisieren besonders einfach.
2. Teilen sich mehrere Personen die Korrekturarbeit, werden **Prüfungsaufgaben** und nicht Studierende untereinander **aufgeteilt.** Die unabhängige Bewertung jeder Aufgabe durch zwei Personen macht die Bewertung objektiver. Bei Veranstaltungen mit mehreren Prüfenden werden die Aufgaben gemäss inhaltlicher Expertise aufgeteilt.
3. Die **Reihenfolge** der Bewertung von Studierenden wird mit jeder Aufgabe **variiert.**
4. **Macht man sich** die möglichen **Störfaktoren bewusst,** so schränkt man ihren Einfluss allein damit bereits beträchtlich ein (siehe Anhang: Übersicht der wichtigsten Störfaktoren).
5. **Bewertungsschema und Musterlösung** ermöglichen die konsistente und reproduzierbare Bewertung der Aufgaben nach einheitlichen Kriterien und schränken so den Einfluss von Störfaktoren ein. Bewertungsschema und Musterlösung werden vor der Prüfungskorrektur gemeinsam besprochen.
6. Anhand einiger **Referenzkorrekturen** «eichen» die Korrigierenden ihre Bewertung von Prüfungsaufgaben. Die Korrekturen werden stichprobenartig auf Konsistenz überprüft.

Störfaktoren dürfen keinesfalls nachträglich kompensiert werden. Dies wäre ein Akt der Willkür und würde zu einer weiteren Verschlechterung der Aussagekraft führen, da man der Messung lediglich einen neuen Störfaktor hinzufügt, welcher mit der eigentlichen Prüfungsleistung in keinem Zusammenhang steht.

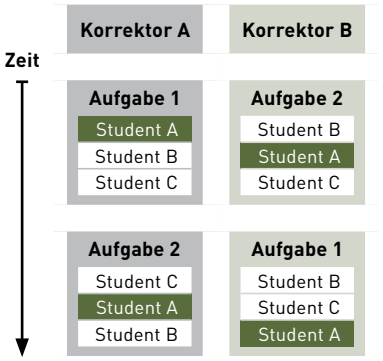


Abbildung 4: Mögliche Aufteilung von Aufgaben und Studierenden zu KorrektorInnen

2.3 Nachbereitung

2.3.1 Prüfungseinsicht

Die Prüfungskorrektur enthält differenzierte Informationen über die studentische Leistung. Die Prüfungseinsicht ermöglicht den Studierenden, diese Informationen für ihr weiteres Lernen zu nutzen. Ferner kann die Prüfungseinsicht, durch kritisches und konstruktives Feedback der Studierenden, einen Beitrag zur Verbesserung der Qualität zukünftiger Prüfungen leisten. Schliesslich erfüllt die Prüfungseinsicht auch eine rechtliche Funktion [6]. Nachkorrekturen aufgrund der Prüfungseinsicht sollen nur mit äusserster Zurückhaltung vorgenommen werden. Sie sind nur bei offensichtlich falschen Bewertungen oder schweren Ermessensfehlern in Härtefällen angezeigt.

2.3.2 Prüfungsreview

Folgende Schritte empfehlen sich bei einer Nachbereitung:

- Eine **Item-Analyse** hilft, zu anspruchsvolle oder zu einfache Aufgaben sowie Aufgaben mit fehlerhafter Bewertung für eine weitere Überprüfung zu identifizieren (siehe 2.3.3).
- Bei einem **inhaltlichen Review** wird die inhaltliche Korrektheit der Aufgabenstellung sowie der Bewertungen überprüft. Anschliessend wird geprüft, ob die unterschiedlichen Antworten der Studierenden tatsächlich Unterschieden in der Erreichung der Lernziele entsprechen und ob sich diese Unterschiede mit dem verwendeten Bewertungsschema zuverlässig erfassen lassen.
- Ein **Vergleich mit früheren Prüfungen** oder Prüfungen ähnlicher Veranstaltungen bietet zusätzliche Anhaltspunkte für allfällige Anpassungen.
- Eine **Prüfungsbesprechung im Plenum** mit Studierenden und/oder Assistierenden kann wertvolle qualitative Rückmeldungen bieten.
- Prüfungen werden im Rahmen der **Unterrichtsbeurteilung durch die Studierenden** evaluiert. Die Evaluation gibt wichtige Hinweise für die Gestaltung zukünftiger Prüfungen, im Speziellen zur Abstimmung der Prüfung mit den Lernzielen und dem Unterricht, der Formulierung der Prüfungsfragen sowie der Fairness.
- Allfällig notwendige **Änderungen an den Lernzielen** werden vorgenommen.
- Die wichtigsten Erkenntnisse aus Design, Durchführung, Auswertung und Nachbereitung der Prüfung werden fur die nächste Durchführung der Lerneinheit als **Prüfungsbericht** zusammengefasst.

[6] vgl. Anhang 3: Weisung zur Akteneinsicht und Aktenweitergabe im Rahmen von Leistungskontrollen

2.3.3 Problematische Fragen identifizieren: die Item-Analyse

In einer Item-Analyse werden für jede Aufgabe statistische Kennwerte berechnet. Diese erleichtern die Identifikation von Aufgaben, die näher überprüft werden sollen.

- Die **Aufgabenschwierigkeit** beschreibt, welcher Anteil der Studierenden eine Aufgabe erfolgreich bearbeitet hat. Zu viele zu schwierige und/oder zu einfache Aufgaben können die Reliabilität der Prüfung durch Boden- und Deckeneffekte beeinträchtigen.
- Die **Diskriminanz** beschreibt wie die Leistung in einer bestimmten Aufgabe mit der Leistung in der übrigen Prüfung korreliert. Eine Korrelation von Null bedeutet keinen Zusammenhang. Je tiefer die Diskriminanz einer Aufgabe, desto wahrscheinlicher sind Fehler in der Konstruktion und/oder der Bewertung der Aufgabe.

Gute Kennwerte in der Item-Analyse sind keine Garantie für gute Aufgaben, und schlechte Kennwerte bedeuten nicht zwingend, dass auch die Aufgabe schlecht ist. Die Kennwerte liefern Hinweise, die immer durch eine inhaltliche Überprüfung bestätigt werden müssen.

2.3.4 Umgang mit Fehlern in Prüfungen bei der Korrektur

Es kann vorkommen, dass man im Nachhinein bei einzelnen Aufgaben gravierende Designfehler entdeckt oder dass sich eine Prüfung als zu anspruchsvoll oder zu einfach herausstellt.

Prüfungsaufgaben mit einem **Designfehler** dürfen nachträglich nicht einfach aus der Prüfungsbewertung ausgeschlossen werden. Häufig kann eine Neubewertung aufgrund modifizierter Musterlösungen Abhilfe schaffen. Dabei ist darauf zu achten, dass diese im Einklang mit Bewertungsschema und -kriterien bleiben, welche den Studierenden kommuniziert wurden. Ist dies nicht möglich, kann man allen Studierenden die maximale Punktzahl für die Aufgabe geben, muss dabei aber sicherstellen, dass hieraus keinen Studierenden, welche die Aufgabe ganz oder in Teilen erfolgreich gelöst hatten, Nachteile entstehen.

Bei einer Neubewertung von Aufgaben gelten die gleichen Good-Practice-Regeln wie bei der ursprünglichen Bewertung.

Bestehen aufgrund der Notenverteilung oder aus anderen Gründen erhebliche **Zweifel an der Angemessenheit der Prüfungsschwierigkeit oder des Notenmassstabes,** wird

1. deren Angemessenheit und Korrektheit inhaltlich in einer Nachbereitung überprüft (siehe 2.3).
2. aufgrund der Nachbereitung das Bewertungsschema korrigiert, und die Aufgaben werden neu bewertet.
3. aufgrund der Nachbereitung wird der Notenmassstab inhaltlich neu festgelegt.

Bei kriteriumsorientierten Prüfungen wird der Notenmassstab vor der Prüfung festgelegt. Nachkorrekturen, wie oben beschrieben, sollten Ausnahmefälle sein. Werden Nachkorrekturen zur Regel, prüft man de facto normorientiert. Lässt sich dies nicht verhindern, ist es besser die Studierenden vor der Prüfung darüber zu informieren.

Kapitel 3: Grundlagen

3.1 Prüfungen sind auf Lehr- und Lernaktivitäten abgestimmt

Lernen ist ein aktiver Prozess, bei welchem neue Informationen mit bestehendem Wissen in Beziehung gesetzt wird. Neue Kompetenzen entwickeln sich durch neue Erfahrungen aus bestehenden heraus. Prüfung und Lehr-Lernaktivitäten sind so aufeinander abgestimmt, dass sie die Studierenden optimal in ihrem Lernen unterstützen.

3.1.1 Prüfen aus einem Guss (Alignment)

Die Unterrichtenden gestalten das Lernen durch die Vorgabe klarer Lernziele, die Gestaltung zielführender Aktivitäten sowie einer passenden Evaluation der Zielerreichung, indem Lernziele, Lehr-Lernaktivitäten sowie Prüfungsaufgaben in Form und Inhalt aufeinander abgestimmt werden. Diese Abstimmung bezeichnet man als **‹Alignment›.**

Durch Alignment wird sichergestellt, dass Studierende auf die Prüfung hin wesentliche, in den Lernzielen festgehaltene Kompetenzen erwerben. Prüfungsaufgaben mit gutem Alignment werden von den Studierenden als relevant wahrgenommen, setzen positive Anreize und sorgen so für eine bessere Lernmotivation. Dies wiederum bewirkt, dass Studierende intensiver und ausdauernder lernen.

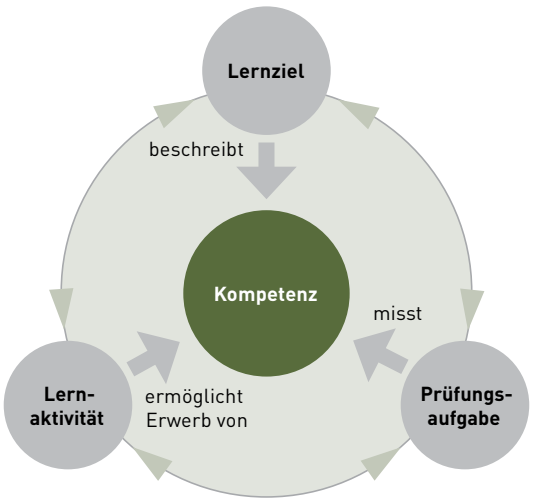


Abbildung 5: Gut abgestimmte Lernziele, Lernaktivitäten und Prüfungsaufgaben beziehen sich alle auf dieselben Kompetenzen. Diese Abstimmung nennt man Alignment.

3.1.2 Viel gewusst und nichts verstanden? (Lernzieltaxonomie)

«Studierende können Gleichgewichtskonzentrationen von Säuren und Basen in wässriger Lösung berechnen.» In diesem Beispiel besteht das **Lernziel** aus zwei Komponenten: Dem fachlichen Inhalt (dem «Lernstoff», hier Gleichgewichtskonzentrationen von Säuren und Basen), sowie der Handlung, welche mit dem Inhalt durchgeführt werden muss («Gleichgewichtskonzentrationen berechnen»). Die verlangte Leistung bedingt nun zweierlei: Die Speicherung von **Wissen** (Fakten, Konzepte, Prozeduren oder Wissen über Wissen, sogenannte Metakognition) sowie die Nutzung dieses Wissens **(kognitiver Prozess),** hier im Berechnungsvorgang. Der kognitive Prozess beschreibt die Art der Verarbeitung des Fachinhaltes: Erinnern, Verstehen, Anwenden, Analysieren, Evaluation und Synthese. Diese Reihenfolge kognitiver Prozesse kann man sich vereinfacht als eine Art «Verarbeitungstiefe» veranschaulichen.

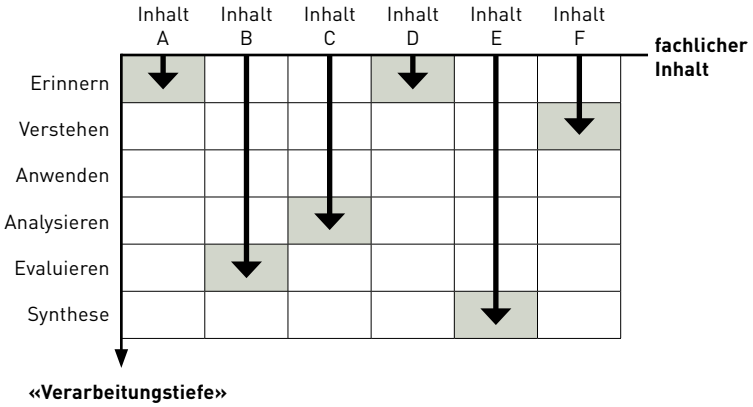


Abbildung 6: Lernziele können aufgrund der erwarteten «Verarbeitungstiefe» klassifiziert werden.

Gute Prüfungsaufgaben geben die Lernziele nicht nur in Bezug auf den fachlichen Inhalt korrekt wieder, sondern auch bezüglich des kognitiven Prozesses. Es macht wenig Sinn, die Lernziele einer Veranstaltung, in welcher grundlegende Konzepte verstanden und angewendet werden sollen, mit Aufgaben zu überprüfen, welche durch blosses Auswendiglernen erfolgreich beantwortet werden können. Umgekehrt, wenn man prüfen möchte, ob Studierende Gelerntes auf neue Kontexte übertragen können, muss man diese Erwartung auch in den Lernzielen transparent machen.

3.2 Prüfungen sind Messinstrumente

Prüfungen sind Instrumente zur Messung studentischer Kompetenzen. In einer Prüfung wird die Lernzielerreichung erhoben und diese zunächst auf eine Punkteskala und anschliessend auf die Notenskala von ‹1› bis ‹6› abgebildet. Um möglichst aussagekräftige Resultate zu erhalten, muss Folgendes so gut wie in der Unterrichtsrealität möglich umgesetzt werden.

3.2.1 Wann ist eine Prüfung gut? (Gütekriterien)

Eine Prüfung ist dann gut, wenn sie in der Lage ist, studentische Kompetenzen zu messen. Dies ist nie exakt möglich, da grundsätzlich jede Messung verschiedenen Messfehlern unterliegt. Durch das Einhalten der Gütekriterien Objektivität, Reliabilität und Validität können die Messfehler klein gehalten werden.

1. Die Messung muss inhaltlich gültig sein, d. h. sie muss auch tatsächlich das messen, was sie zu messen vorgibt **(Validität),**

2. sie muss verlässlich, präzise und reproduzierbar sein **(Reliabilität)** und
3. sie muss unabhängig von der messenden Person und den Messumständen sein **(Objektivität),**

Die Gütekriterien stehen in einer hierarchischen Beziehung zu einander. Objektivität ist eine notwendige aber nicht hinreichende Bedingung für Reliabilität und Reliabilität wiederum eine notwendige aber nicht hinreichende Bedingung für Validität. Das wichtigste Gütekriterium ist immer die Validität, da ohne Gültigkeit eine Messung ihre Berechtigung verliert.

3.2.2 Prüfungen messen Kompetenzen (Varianzerklärung)

Anders als bei einer einfachen physikalischen Grösse wie der Länge in Metern, handelt es sich bei der Lernzielerreichung, um eine nicht direkt beobachtbare, ‹latente› Eigenschaft. Die Lernzielerreichung kann nur indirekt gemessen werden, indem aus der Korrektur von bearbeiteten Prüfungsaufgaben auf sie rückgeschlossen wird. Dabei geht es letztendlich immer um Varianzaufklärung: Die durch die Prüfung erhobenen und durch Noten abgebildeten Unterschiede sollen den tatsächlichen Unterschieden in der Lernzielerreichung entsprechen.

Der Messvorgang lässt sich in fünf wesentliche Schritte unterteilen:

1) Die zu messende Zielkompetenz muss definiert und konkret festgehalten werden. Diese Funktion übernehmen die **Lernziele.**
2) In der **Operationalisierung** wird eine geeignete Methode zur (indirekten) Beobachtung der latenten Eigenschaft entwickelt. Dies entspricht dem Ausgestalten der Prüfungsaufgaben, inklusive Bewertungsschema und Musterlösung. Die Prüfungsaufgaben beziehen sich nicht auf die Zielkompetenz direkt, sondern auf die daraus abgeleiteten Lernziele,
3) Die Durchführung der Prüfung dient der eigentlichen Beobachtung, bzw. **Datenerhebung.**
4) Bei der **Quantifizierung** wird die Prüfung korrigiert und gemäss dem vorgängig definierten Bewertungsschema auf einer Punkteskala abgebildet. Bis hierhin stellt die Messung einen rein deskriptiven Vorgang dar.
5) Bei der **Normierung** werden die Punkte zu einem externen Massstab (Norm) in Beziehung gesetzt und auf eine neue Skala abgebildet. Man unterscheidet die individuelle, die soziale («normorientert» prüfen) und die sachliche («kriteriumsorientiert» prüfen) Bezugsnorm. Die Notenskala bildet Leistungen normativ als «sehr gut», «ungenügend», etc. ab.

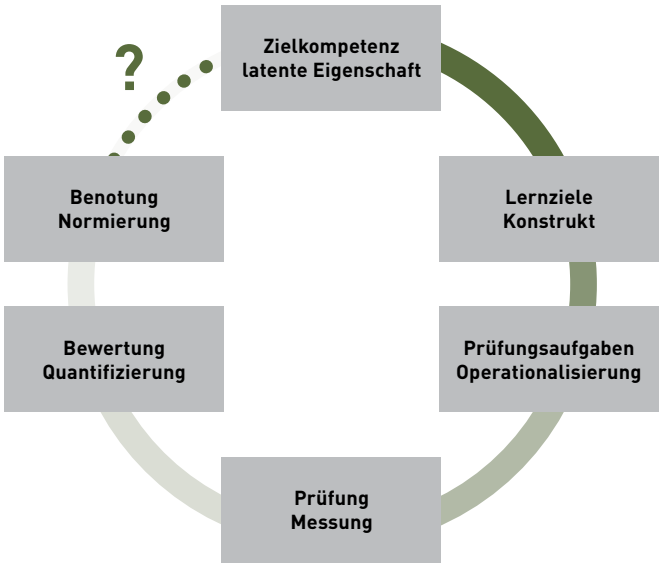


Abbildung 7: Die Messung von Kompetenzen erfolgt über mehrere, fehlerbehaftete Schritte.
Der grüne Pfeil repräsentiert die abnehmende Aussagekraft der Information über die Zielkompetenz.

Dabei ist zu bedenken, dass jeder dieser fünf Schritte fehlerbehaftet ist. Einerseits geht mit jedem Schritt ein Teil der ursprünglichen Varianz verloren, weil sie nicht mit erhoben wird: Zum Beispiel kann man bei der Gestaltung einer Prüfung kaum für alle Lerninhalte und auf allen vorgegebenen kognitiven Stufen Aufgaben gestalten. Andererseits wird der Messung durch Störfaktoren und Messungenauigkeiten Fehlervarianz hinzugefügt: Zum Beispiel werden bei der Durchführung einer Prüfung i. A. auch Faktoren, wie die Nervosität der Studierenden oder deren Sprachkompetenz, mit erhoben. Varianz, welche in früheren Schritten verloren ging, bleibt in späteren Schritten verloren und hinzugefügte Fehlervarianz wird man i. d. R. nicht mehr los.

Fehlervarianz kann auf zwei Arten minimiert werden.

- **Störfaktoren** können identifiziert und ihr Einfluss durch passende Gegenmassnahmen minimiert werden. So kann z. B. der Störfaktor Sprachkompetenz minimiert werden, indem Aufgabenstellungen immer in möglichst einfacher und klarer Sprache verfasst werden.
- **Unsystematische Fehlervarianz,** also die «Unschärfe» einer Messung, kann durch das Einfügen von Redundanz, bzw. Messwiederholungen, minimiert werden. Kernkonzepte oder fundamental wichtige Sachverhalte sollten deshalb, nach Möglichkeit, immer mit mehr als nur einer einzelnen Aufgabe überprüft werden.

3.2.3 Ist eine 6 doppelt so gut wie eine 3? (Skalenniveaus)

Messungen bilden Merkmale inhaltlich bedeutsam auf eine Skala ab. Das **Skalenniveau** hält dabei fest, wie diese Messwerte interpretiert werden können und welche statistischen Operationen für sie definiert sind. Eine **Ordinalskala** bildet Messwerte als hierarchisch geordnete Rangreihenfolge ab, welche Aussagen der Art «besser/schlechter», «grösser/kleiner», etc. erlauben. Beispiele ordinalskalierter Merkmale sind die durch eine Umfrage erhobene Zufriedenheit mit einer Lerneinheit oder der Dienstrang im Militär. Als zentrale Kennwerte sind für ordinalskalierte Daten Modus und Median definiert. Bei einer **Intervallskala** lässt sich ausserdem die Grösse von Abständen zwischen Messwerten vergleichen und diese ist auch inhaltlich bedeutungstragend. Beispiele intervallskalierter Merkmale sind Temperaturen in Grad Celsius, Kalenderdaten oder der Intelligenzquotient. Für intervallskalierte Daten sind zusätzlich als zentraler Kennwert der Durchschnitt sowie als Operationen die Addition und Subtraktion definiert.

Das Bestehen eines Prüfungsblocks und damit die Berechtigung weiter zu studieren, hängt an der ETH in aller Regel davon ab, ob in einem (gewichteten) Notendurchschnitt eine genügende Note erreicht wird. Die Berechnung von Durchschnitten setzt (1) Daten auf einer Intervallskala voraus und erfordert (2), dass Daten, welche zu einem Durchschnitt beitragen, sich auch inhaltlich auf eine gemeinsame Skala beziehen. Das Messen menschlicher Kompetenzen auf Intervallskalenniveau ist sowohl anspruchsvoll, als auch äusserst aufwändig und kann ohne die Unterstützung professioneller Testentwickler nicht gewährleistet werden. Prüfungsnoten sind deshalb grundsätzlich als ordinalskaliert zu betrachten. Ausserdem darf eine inhaltlich gemeinsame Skala bei unterschiedlichen Prüfungen nicht a priori angenommen werden. Notendurchschnitte sind aus messtheoretischer Sicht folglich weder zulässig noch interpretierbar. Daraus folgt, dass Entscheide aufgrund von Notendurchschnitten mit einem zusätzlichen, schwer schätzbaren Fehler behaftet sind, welcher über die Messfehler der einzelnen Prüfungen hinausgeht. Dennoch hat sich der Notendurchschnitt als Urteilsheuristik in der Hochschulpraxis bewährt, ist sowohl unter Studierenden, als auch Dozierenden akzeptiert und nur schwer aus der Schweizer Bildungslandschaft wegzudenken.

Je besser sich die einzelnen Prüfungsnoten eines Notendurchschnittes jeweils dem Intervallskalenniveau annähern und je stärker sich diese auf eine inhaltlich gemeinsame Skala beziehen, desto kleiner wird der durch die unzulässige Bildung des Durchschnitts eingeführte Fehler. Um sich dem Intervallskalenniveau annähern zu können, muss für die Punkte einer Prüfung eine gemeinsame Metrik gefunden werden, so dass ein Punkt immer etwas vergleichbares repräsentiert, egal in welcher Aufgabe er zustande kommt. Die Relevanz einer durch eine Prüfungsaufgabe erhobenen Kompetenz oder der relative Anteil betroffener Lernziele könnten zum Beispiel eine solche gemeinsame Metrik darstellen. Die Prüfungsnote würde dann in der Folge etwas über das Ausmass der Relevanz der in der Prüfung demonstrierten Kompetenzen sagen. Damit sich verschiedene Prüfungsnoten möglichst auf eine inhaltlich gemeinsame Skala beziehen können, ist es unerlässlich, dass sich die einzelnen Dozierenden absprechen und sich bezüglich ihrer Prüfungsbewertung austauschen.

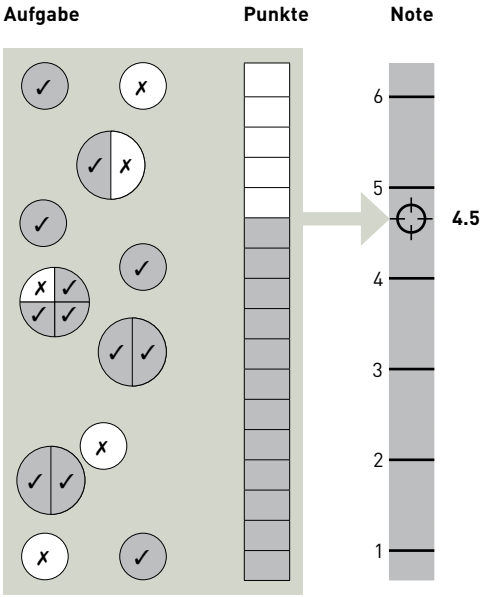


Abbildung 8: Bei der Quantifizierung werden diskrete Aufgaben in verrechenbare Punkte umgewandelt. Da dies nur näherungsweise gelingen kann, ist besondere Sorgfalt nötig.

Anhänge

Anhang 1: Checkliste Korrektur und Notengebung

Die Checkliste gibt einen schnellen Überblick der empfohlenen Vorgehensweisen zur Notengebung und hilft bei deren Umsetzung in der Praxis.

Aspekte	Leitfrage	
1. Vorbereitung		
→ **2.1.7**	Liegt das Bewertungsschema vor?	☐
→ **2.1.7**	Liegen die Musterlösungen vor?	☐
→ **2.1.7**	Wurden Musterlösung und Bewertungsschema auf Korrektheit überprüft?	☐
→ **2.1.2**	Liegen frühere oder ähnliche Prüfungen zum Vergleich vor?	☐
→ **2.1.1 – 2.1.3**	Wurde der Notenmassstab festgelegt?	☐
→ **Anhang 5**	Liegt das Übersichtsdokument zu den wichtigsten Störfaktoren vor?	☐
→ **2.2.2**	Wurden die Prüfungen nach Aufgaben und nicht nach Studierenden aufgeteilt?	☐
→ **2.2.2**	Wird die Reihenfolge der Bewertung von Studierenden für jede Aufgabe variiert?	☐
→ **2.2.2**	Wurden Bewertungsschema und Musterlösung gemeinsam vorbesprochen?	☐
→ **Anhang 5**	Werden die Aufgaben nach dem vier-Augen Prinzip ausgewertet?	☐
→ **2.2.2**	Liegen die Prüfungen in anonymisierter Form vor?	☐
→ **2.2.2**	Wurden ausreichend Pausen etc. eingeplant?	☐
2. Korrektur / Bewertung		
→ **2.1.7, 2.2.2**	Liegen Bewertungsschema und Musterlösungen zur Hand?	☐
→ **Anhang 5**	Bin ich ausreichend ausgeruht und fit?	☐
→ **2.2.2**	Habe ich mir die wichtigsten Störfaktoren in Erinnerung gerufen?	☐
→ **2.2.2**	Habe ich mich für die Korrektur geeicht, indem ich einige Referenzlösungen betrachtet habe?	☐

→ **2.2.2**	Habe ich meine Bewertungen so objektiv als möglich abgegeben?	☐
→ **2.2.2**	Habe ich meine Bewertungen abgegeben, ohne nachträglich Störfaktoren zu kompensieren versucht zu haben?	☐
→ **2.1.7**	Ist die Punktvergabe, inklusive Begründungen, in der Prüfung dokumentiert?	☐
→ **Anhang 5**	Ist die Aufgabenbewertung periodisch auf ihre Konsistenz überprüft worden?	☐
3. Prüfungseinsicht		
→ **2.3.1**	Wurde die Einsicht in die Ergebnisse und die Korrektur der Prüfung organisiert und durchgeführt?	☐
→ **2.3.1, 2.3.4**	Wurden allfällig notwendige Korrekturen vorgenommen?	☐
4. Prüfungsreview		
→ **2.3.3**	Wurde eine Item-Analyse durchgeführt?	☐
→ **2.3.2, 2.3.3**	Wurden Items mit schlechten Kennwerten inhaltlich überprüft?	☐
→ **2.3.2**	Wurden die Bewertungen überprüft?	☐
→ **2.3.2**	Wurde eine Prüfungsbesprechung durchgeführt?	☐
→ **2.3.2**	Wurde die Prüfung im Rahmen der Unterrichtsbeurteilung evaluiert?	☐
5. Massnahmen		
→ **2.3.4**	Lagen erhebliche Zweifel an der Angemessenheit der Prüfungsschwierigkeit und des Notenmassstabes vor?	☐
→ **2.3.4**	Wurde im Rahmen des Prüfungsreviews die Notwendigkeit von Nachkorrekturen abgeklärt?	☐
5.1 Neubewertung	**… falls Massnahmen notwendig sind:**	
→ **2.3.4**	Wurde das Bewertungsschema formal beibehalten und nur inhaltlich angepasst?	☐
→ **2.3.4**	Wurden die übrigen Good-Practice-Regeln bei der Neubewertung eingehalten?	☐
→ **2.3.4**	Sind die Neubewertungen konsistent mit den Bewertungskriterien, welche den Studierenden mitgeteilt wurden?	☐
5.2 Neuer Notenmassstab	**… falls die Neubewertung nicht möglich oder nicht erfolgreich war:**	
→ **2.3.4**	Wurde die neue Eichung des Notenmassstabes nach inhaltlichen Kriterien vorgenommen?	☐
6. Qualitätssicherung		
→ **2.3.2**	Wurden allfällig notwendige Änderungen an den Lernzielen vorgenommen?	☐
→ **2.3.2**	Wurden die wichtigsten Erkenntnisse aus dem Prüfungsreview und der gesamthaften Abwicklung der Prüfung in einem Prüfungsbericht für die nächste Durchführung zusammengefasst?	☐
→ **2.3.2, 2.3.4**	Wurden vorbeugende Massnahmen ergriffen, falls die Prüfung nachkorrigiert oder der Notenmassstab angepasst werden musste, um einer Wiederholung vorzubeugen?	☐

Anhang 2: Weiterführende Literatur

Anderson, L.W., Krathwohl, D.R., Airasian, P.W. et al. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom›s Taxonomy of Educational Objectives. New York: Longman.

Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32: 347 – 364.

Biggs, J., Tang, C. (2011): Teaching for Quality Learning at University. Maidenhead / U.K.: Open University Press.

Hattie, J.A.C. (2002). What are the attributes of excellent teachers? In Teachers make a difference: What is the research evidence? (pp. 3-26). Wellington: New Zealand Council for Educational Research.

Eidgenössische Technische Hochschule Zürich (2013). Qualitätskriterien für die Lehre: Teil «Studiengänge und Lehrveranstaltungen». Retrieved from: http://www.let.ethz.ch/docs/Qualitaetskriteren_LehreETH.pdf

Krathwohl, D.R. (2002). A Revision of Bloom›s Taxonomy: An Overview. Theory Into Practice, 41 (4), 212 – 218.

Krebs, R. (2004). Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. Bern: Institut für Medizinische Lehre IML, Abteilung für Ausbildungs- und Examensforschung. Retrieved from: http://www.iml.unibe.ch/dienstleistung/assessment_pruefungen/pruefungsmethoden/wahlantwortfragen_mc/

Metzger, Ch., Nüesch, Ch. (2004): Fair prüfen - Ein Qualitätsleitfaden für Prüfende an Hochschulen. St. Gallen: Institut für Wirtschaftspädagogik, Universität St. Gallen.

Race, P., Brown, S., Smith, B. (2005). 500 Tips on Assessment, 2nd edition. RoutledgeFalmer: New York, pp. 2 – 11, 21 – 22.

Schneider, M., Stern E. (2010). The cognitive perspective of learning: ten cornerstone findings. In H. Dumont, D. Istance & F. Benavides (eds.): The Nature of Learning: Using Research to Inspire Practice (S. 69 – 90). Paris: OECD. doi: http://dx.doi.org/10.1787/9789264086487-5-en

Anhang 3: Weitere Angebote

didactica: Hochschuldidaktik-Weiterbildung an der ETH und Universität Zürich:
http://www.didactica.ethz.ch

LET Beratungsangebot: Das Beratungsteam des LET ist für Sie Ansprechpartner für alle Fragen rund um die Lehre:
beratung@let.ethz.ch

Leitfaden für Dozierende der ETH Zürich unter:
http://www.ethz.ch/faculty

Weisung zur Akteneinsicht und Aktenweitergabe im Rahmen von Leistungskontrollen:
http://www.rektorat.ethz.ch/directives

Leistungskontrollenverordnung ETH Zürich:
http://www.rechtssammlung.ethz.ch/pdf/322.021_leistungskontrollenverordnung_eth_zuerich.pdf

Ausführungsbestimmungen des Rektors zur Leistungskontrollenverordnung ETH Zürich:
http://www.rektorat.ethz.ch/directives

Anhang 4: Mathematische Formeln

Aufgabenschwierigkeit:
Durchschnittliche Punktzahl einer Aufgabe geteilt durch die maximal erreichbare Punktzahl der Aufgabe

Formel: $p = \frac{\bar{x}}{x_{max}}$

$\overline{x}$ = arithmetisches Mittel (Durchschnitt) der erreichten Punktzahl der Aufgabe

x_{max} = maximal erreichbare Punktzahl der Aufgabe

Trennschärfe:
Die Trennschärfe ist die Korrelation der Leistung in einem bestimmten Item mit der Leistung im Gesamttest ohne dieses Item.

Formel: $r_{i(t-i)} = \frac{\sigma(x_i, x_{t-i})}{\sigma(x_i)\sigma(x_{t-i})}$

x_i : Werte für das Item i

x_{t-i} : Werte für den Gesamttest ohne das Item i

Umsetzung in Excel:
Die Beispiele beruhen auf einer Tabelle, in welcher die Aufgaben den Spalten A – O zugeordnet sind und die Zeilen 1 – 10 die Studierenden repräsentieren.

A1 – A10: Zeilen mit den Punkten der Studierenden
A11: Feld mit der maximal erreichbaren Punktzahl der Aufgabe
Q1 – Q10: Summe der Punkte aus allen Aufgaben ohne Punkte der Aufgabe der Spalte A

Schwierigkeit der Aufgabe in Spalte A:
MITTELWERT(A1:A10)/A11

Trennschärfe der Aufgabe in Spalte A:
KORREL(A1:A10;Q1:Q10)

Anhang 5: Übersicht Störfaktoren

Störfaktor	Massnahme
«Äussere» Eigenschaften der beurteilten Person, wie das Erscheinungsbild, Auftreten, Beteiligung und Interesse am Unterricht, Fleiss, Geschlecht, etc.	Der Einfluss dieser Faktoren wird durch das Anonymisieren der Prüfungen ausgeschlossen.
Aspekte der zu beurteilenden Arbeit, ohne direkten Zusammenhang mit den eigentlichen Lernzielen, wie Länge, Bestimmtheit und Layout der Antworten, sprachliche Gewandtheit, etc.	Konkrete, objektive Bewertungskriterien im Bewertungsschema schränken solche Einflüsse ein.
Die **Qualität der Handschrift** hat einen starken Einfluss auf die Bewertung von Aufgaben.	Dieser Störfaktor kann minimiert werden, indem man Fragen stellt welche nur relativ kurze Antworten erfordern und indem handschriftliche Prüfungen ohne Zeitdruck stattfinden. Prüfungen am Computer eliminieren den Störfaktor Handschrift gänzlich.
Sich **ändernde innere Zustände** und Verfassung der korrigierenden Person führen zu inkonsistenten Bewertungen. Man bewertet strenger am frühen Morgen, bei schlechter Laune und vor dem Mittag, wenn man Hunger hat. Man bewertet die ersten Aufgaben nach besonders guten Arbeiten strenger. Umgekehrt bewertet man besonders nachsichtig vor Feierabend, bei guter Laune, nach dem Mittagessen und bei den letzten Aufgaben, wenn man müde ist. Man bewertet weniger streng, wenn man zuvor eine eher schlechte Arbeit korrigiert hat.	Durch regelmässige Pausen und «Eichen» anhand einiger Referenzbewertungen von Aufgaben können diese Störfaktoren verringert werden. Indem man eine Aufgabe nach der anderen, jeweils für alle Studierenden, bewertet und dabei bei jeder Aufgabe die Reihenfolge der Studierenden, zufällig oder systematisch, variiert, kann man die Störfaktoren gleichmässig auf alle Studierenden verteilen.
Unterschiedliche Personen beurteilen dieselbe Arbeit unterschiedlich. Teilen sich mehrere Personen die Korrekturarbeit werden Prüfungsaufgaben und nicht Studierende untereinander aufgeteilt.	Bewertungsschema und Musterlösung werden vor der Prüfungskorrektur gemeinsam besprochen, um eine einheitliche Umsetzung in der Korrektur zu gewährleisten. Die unabhängige Bewertung jeder Aufgabe durch zwei Personen und die wenigstens stichprobenartige Überprüfung der Korrekturen auf Konsistenz helfen die Bewertungen einheitlicher und objektiver zu machen. Bei Veranstaltungen mit mehreren Prüfenden werden die Aufgaben gemäss inhaltlicher Zuständigkeit, beziehungsweise gemäss Expertise aufgeteilt.