Prof. Dr. Torsten Kirstges
Fachhochschule in Wilhelmshaven
Studiengang Tourismuswirtschaft

Gerechte Noten:
zur Gestaltung von Notensystemen
für die Beurteilung von Leistungen in Klausuren

Wilhelmshaven 2007
ISBN: 978-3-935923-08-8

Ob in der Schule, an der Hochschule oder bei sonstigen Prüfungen: **Prüfende vergeben Noten an die Geprüften**, um damit deren **Leistung zu beurteilen** und möglichst deren Leistungsbereitschaft zu steigern (wobei nicht auszuschließen ist, dass – schlechte – Noten auch eher demotivieren).

Doch **wie kommt ein Prüfender zu einer der Leistung des Geprüften entsprechenden Note**? Wann ist eine Note bzw. ein Notensystem **(leistungs-)gerecht**? Worin liegen die **Probleme der Benotung**? Diesen Fragen soll in den folgenden Ausführungen nachgegangen werden.

1. Wie wird die Leistung definiert?

Bewertet wird die vom Prüfling **in der Prüfung erbrachte Leistung**, nicht aber dessen ansonsten - also z.B. zu anderen Zeitpunkten, im Unterricht, bei Präsentationen etc. - gezeigte Leistung, **nicht** dessen **Leistungsvermögen** oder dessen **Leistungsbereitschaft**/-wille. Leistung, so lehrt uns die Physik, ist „**Arbeit dividiert durch Zeit**“: Wer eine Arbeit in einer beliebig langen Zeit erbringt, leistet eher wenig. Wer nur wenig Arbeit in einer bestimmten Zeitspanne erledigt, leistet ebenfalls wenig.

Prüfende lehren i.d.R. einen bestimmten **prüfungsrelevanten Lehrstoff**, den sie dann – meist in Auszügen - in der Prüfung, die eine bestimmte Zeit dauert, abfragen. Geprüft wird dabei üblicherweise nur die **kognitive Leistung** (nicht z.B. die körperliche oder soziale Leistung) des Prüflings. Folgende, für die spätere Benotung relevante Variablen ergeben sich daraus:

- **Quantität / Umfang des Lehrstoffes**: Wird in Relation zu der zur Verfügung stehenden Unterrichtszeit (dies auch im Vergleich zu anderen Lehrenden) viel oder wenig Stoff vermittelt?

- **Qualität / Anspruchniveau** des Lehrstoffes: Wird in Relation zu den von den Prüflingen mitgebrachten Voraussetzungen sowie zu anderen Lehrenden ein niedriges oder ein hohes Anspruchsniveau an den Tag gelegt?
- **Deckungsgleichheit von gelehrtem Stoff und prüfungsrelevantem Stoff**: Wird nur das abgefragt, was auch (mehr oder weniger wörtlich) gelehrt wurde, oder verlangt der Prüfende, dass die Prüflinge sich auch darüber hinaus gehend fachrelevantes Wissen aneignen (z.B. durch Lektüre/Selbststudium, Bearbeitung von Fallstudien etc.)? Ist die **Prüfungszeit** dem geprüften Lehrstoff angemessen?

Es ist offensichtlich, dass bereits hier, bei der Frage der Leistungsdefinition, **große Unterschiede zwischen verschiedenen Lehrenden** auftreten können: Wer viel und – aus Sicht der Lernenden – komplizierten Stoff vermittelt, viel an Vorkenntnissen voraussetzt und darüber hinaus erwartet, das die Prüflinge den Stoff selbständig vertiefen und ergänzen, gilt als „schwer“, das von ihm gelehrte Fach als „schwierig“.

Um zu ermitteln, ob der ein oder andere Dozent zumindest in der Wahrnehmung der Prüflinge „schwerer“ ist als andere, könnte man die Prüflinge – die idealerweise mehrere Dozenten genossen haben - z.B. wie folgt befragen:

Wie beurteilen Sie den **Umfang** des vom Dozenten vermittelten Lehrstoffes angesichts der für die Lehrveranstaltung zur Verfügung stehenden Unterrichtszeit?
[] es wird zu wenig Lehrstoff vermittelt; ich möchte eher mehr lernen.
[] die Lehrstoff ist der Unterrichtszeit angemessen.
[] es wird zu viel Lehrstoff vermittelt; ich kann nicht alles aufnehmen.

Der Dozent setzt **Vorkenntnisse** voraus. Wie beurteilen Sie diese Anforderung an Vorkenntnisse?
[] Der Dozent setzt nur geringe Vorkenntnisse voraus, damit man seinem Lehrstoff folgen kann. Diese erfülle ich problemlos.
[] Der Dozent setzt angemessene Vorkenntnisse voraus. Diese erfülle ich in der Regel.
[] Der Dozent setzt zu hohe Vorkenntnisse voraus. Diese bringe ich i.d.R. nicht mit.

Wie beurteilen Sie das **Anspruchniveau** des Dozenten bzw. der Lehrveranstaltung?
[] Der Lehrstoff ist relativ leicht; ich möchte eher mehr gefordert werden.
[] Das Anspruchniveau ist angemessen.
[] Der Lehrstoff ist relativ schwierig; ich fühle mich oft überfordert.

Wie beurteilen Sie das **Deckungsverhältnis** von dem vom Dozenten **präsentierten Lehrstoff** einerseits und den Anforderungen in der **Klausur** andererseits?

[] Der Dozent fragt in der Klausur i.d.R. weniger ab, als er unterrichtet hat (z.B. indem der Klausurstoff eingegrenzt bzw. Bereiche des Faches vor der Klausur ausgeschlossen werden).
[] Der Dozent erwartet all das, was er auch gelehrt hat.
[] Der Dozent erwartet über das von ihm präsentierte hinaus weitergehende fachliche Kenntnisse, die wir uns selbständig aneignen sollen.

Wie beurteilen Sie die in der Klausur zur Verfügung stehende **Bearbeitungszeit** für die einzelnen Aufgaben?

[] Die Bearbeitungszeit ist großzügig bemessen. Ich habe ausreichend Zeit zur Beantwortung der Fragen.
[] Die Bearbeitungszeit ist angemessen. Die Zeit reicht mir gerade so aus.
[] Die Bearbeitungszeit ist zu knapp bemessen. Ich habe keine ausreichende Zeit zur Beantwortung aller Fragen.

Eine solche Befragung birgt natürlich eine Reihe **methodischer Probleme**, auf die hier jedoch nicht eingegangen werden soll. Es ließen sich auf diese Weise jedoch Anhaltspunkte darüber gewinnen, ob ein Dozent zu viel, zu wenig oder in angemessenem Umfang Leistung fordert.

2. Wie wird das Leistungsniveau gemessen?

Nun kommt die **Erwartung des Prüfenden an eine Idealleistung** zum Tragen: Welchen Erwartungshorizont hat der Prüfende gegenüber seinen Prüflingen? Wie sieht in seinen Augen eine zu hundert Prozent richtige Antwort („Musterlösung“) zu einer Prüfungsfrage aus? Diese stellt somit die **Bezugsnorm** für die möglichst objektive Beurteilung der erbrachten Leistung dar. Ihr übergeordnet ist i.d.R. ein vorgeschriebenes Curriculum (**Curricularnorm**, in Hochschulen i.d.R. durch die Prüfungsordnung definiert), das jedoch – u.a. aufgrund der Freiheit der Lehre an Hochschulen – allenfalls Orientierungscharakter hat.

Der **Lehrende vergleicht die Ausführungen des Prüflings mit seinen Erwartungen**, denen bestimmte – i.d.R. von ihm selbst definierte – **Beurteilungskriterien** zugrunde liegen. Dabei kann er – gedanklich – auf ein Raster **von 0% bis 100%** zurückgreifen: Zu wie viel Prozent, d.h. zu welchem Grad, hat der Geprüfte mit seinen Leistungen die Erwartungen erfüllt? Es

macht Sinn, diese Beurteilung **für jede einzelne Teilleistung**, d.h. i.d.R. für jede einzelne Frage/Aufgabe einer Klausur, anzuwenden.

Bei **mathematisch** orientierten Aufgaben („Kalkulieren Sie den Reisepreis auf Basis folgender Kosten ….“) und reinen „**Faktenfragen**“ („Wie hoch ist die Reiseintensität …“) lässt sich das Leistungsniveau i.d.R. einfacher messen als bei **verbalen „Aufsatz-Aufgaben“**, doch stellt sich auch hier die Frage, wie korrekte **Zwischenlösungen** (bei einem letztlich falschen Ergebnis) bewertet werden. Werden sog. **Folgefehler** (ein im Prinzip richtiges Verfahren wird auf ein falsches Zwischenergebnis angewandt, so dass die weiteren Schritte zwar korrekt, das Endergebnis aber dennoch falsch ist) negativ bewertet? Diese Frage stellt sich übrigens nicht nur bei mathematischen Aufgaben, sondern z.B. auch bei Textaufgaben im juristischen Kontext (z.B. falsche Anspruchsgrundlage im Reiserecht geprüft, darauf aufbauende nachfolgende Überlegungen aber in sich korrekt).

Bei verbal zu lösenden Aufgaben macht es Sinn, wenn der Prüfende die wesentlichen bzw. möglichen **Gedankengänge und Argumente** im Sinne einer **Musterlösung als Beurteilungskriterium** definiert: Der Prüfling hat vier von fünf Argumenten erkannt und diskutiert, also hat er 80% Leistung auf diese Aufgabe bezogen erbracht. Während in der Grundschule neben Fehlerquote oder Qualität eines Aufsatzes noch „Form & Schrift“ als Beurteilungskriterium herangezogen werden, beschränkt sich Beurteilung in späteren Jahren i.d.R. auf die inhaltlichen Aspekte einer vom Prüfling niedergeschriebenen Lösung.

Auch hier ist offensichtlich, dass es **große Unterschiede hinsichtlich der Beurteilung** des Leistungsniveaus geben kann. Als „streng“ oder vielleicht schon „ungerecht“ dürfte ein Prüfender gelten, der nur solche Antworten als korrekt bewertet, die voll und ganz seinen Erwartungen (also seiner Musterlösung) entsprechen. Schwierig wird es auch, wenn **Werturteile**, also nicht wahrheitsfähige Aussagen, gefragt sind („Beurteilen Sie …“): Wie beurteilt der Prüfende Werturteile des Geprüften, die nicht seinem eigenen Urteil (oder einer herrschenden Meinung) entsprechen, aber durchaus möglich (und argumentativ untermauert) sind?

Seinen **Erwartungshorizont** sollte der Prüfende den Prüflingen bei der konkreten Aufgabenstellung **erkennbar machen**. Gerade in den Wirtschaftswissenschaften kann man zu jeder Frage und Problemstellung sowohl nur einige Worte als Lösung schreiben („Lexikon-Stichwort“) als auch eine seitenlange Abhandlung, geleitet über Assoziationsketten, verfassen. Der **Prüfling muss aber wissen, in welchem Umfang der Prüfer die Beantwortung einer Frage erwartet**. Dies kann der Prüfende z.B. durch die **Angabe von Bearbeitungszeiten** je Aufgabe gewährleisten, wodurch die gesamte zur Verfügung stehende Klausurbearbeitungszeit den einzelnen Prüfungsfragen entsprechend aufgeteilt wird.

Darüber hinaus stellt sich hier das Problem, dass der **Schwierigkeitsgrad der einzelnen Aufgaben unterschiedlich hoch** sein kann. Wird z.B. eine Frage mit hohem Schwierigkeitsgrad richtig beantwortet, so macht es Sinn, wenn der Geprüfte hierfür einen höheren Notenanteil gut geschrieben erhält als bei der erfolgreichen Lösung einer einfacheren Frage. Hierbei stellt sich jedoch das Problem, **wer den relativen Schwierigkeitsgrad definiert** und wie er **bei der Ermittlung des gesamten Leistungsniveaus Berücksichtigung** findet.

Nun, i.d.R. wird es der Prüfende sein, der – quasi objektiv - festlegt, ob eine Frage schwierig (und damit hoch gewichtet) oder leicht (und damit gering gewichtet) ist. Doch kann es sein, dass seine Einschätzung von der der Prüflinge abweicht: Diese empfinden eine andere Aufgabe als viel schwieriger. Somit könnte der **subjektiv definierte Schwierigkeitsgrad** erst bei der Auswertung der Klausurergebnisse evaluiert werden, indem statistisch ausgewertet wird, welche Fragen von welcher Anzahl der Kandidaten richtig bzw. falsch beantwortet wurden. Dies ist jedoch nicht nur sehr aufwändig, sondern birgt auch das methodische Problem, dass damit das Anspruchniveau letztlich erst nachträglich (und damit nicht mehr unabhängig von der jeweiligen Prüflingsgruppe) definiert wird.

Hier deutet sich ein weiteres Problem der Messung des Leistungsniveaus an: Es mag **Klausuren** geben, die – vielleicht sogar vom Prüfenden unbeabsichtigt – aus Sicht der Prüflinge **schwieriger sind als andere**. Und es mag Gruppen geben, die ein generell niedrigeres Leistungsniveau aufweisen als andere. Letzteres kann z.B. daher kommen, dass zwei Studentengruppen aus unterschiedlichen Studiengängen, die einen unterschiedlich hohen Numerus Clausus aufweisen, dieselbe Klausur schreiben: Die Studentengruppe, die den höheren NC erfüllt, schneidet auch deutlich besser bei der Klausur ab. Soll der Prüfende derartigen **Unterschieden inter-temporaler** (die Kohorte (selbe Gruppe) war bei früheren Prüfungen signifikant besser/schlechter als jetzt) **oder inter-personeller Art** (andere Prüflinge sind/waren bei derselben Prüfung wesentlich besser/schlechter als die jetzt bewertete Gruppe) nun Rechnung tragen? Und falls ja: Wie bzw. in welchem Umfang?

Auf relativ einfache Weise könnte man dies durch eine **Variation der Grenze**, ab der der Prüfling **durchgefallen** ist, berücksichtigen (vgl. unten zur Frage der Notensysteme): Würde die Klausur ungewöhnlich schlecht ausfallen, könnte man die Hürde zum Bestehen absenken: statt bei 50% Mindestleistung dann schon bei z.B. 40% bestanden. Andere Systeme legen keinen absoluten, d.h. über verschiedene Gruppen von Prüflingen geltenden Maßstab an, sondern nutzen einen nur jeweils auf die geprüfte Gruppe angewandten Leistungsindex.

Bereits aus den bisherigen Überlegungen erkennt man unschwer, dass es **keine objektive, absolute (d.h. immer und überall gleich gültige) und damit (leistungs-)„gerechte“ Defini-**

tion und Messung von Leistung durch Prüfende geben kann. Für manche Prüflinge mag dies schon höchst „ungerecht“ klingen, sehen sie sich doch dem Goodwill des Prüfenden ausgeliefert. Und in der Tat ist es **dem Anspruchniveau und der Urteilskraft des Prüfenden anheim gegeben, wie er die Leistung des Prüflings misst und beurteilt**. Jeder Professor stellt etwas andere Anforderungen, hat andere Schwerpunktsetzungen, spricht andere Fähigkeiten der Studierenden an und stellt letztlich auf seine Weise Leistung fest. Dabei gibt der Prüfende Signale an die Prüflinge und ggf. an ihre späteren Verwender (z.B. Arbeitgeber), indem die **Prüflinge letztlich in Qualitäts-Cluster eingeteilt** werden. Diese Signale können (sollen) dazu dienen, dass der Prüfling seine Leistung ggf. noch steigert.

Unter Gerechtigkeitsaspekten ist dabei auf zweierlei zu achten:
a) **Horizontale Bewertungsgerechtigkeit**: Gleiche Leistungen müssen zu gleicher Beurteilung führen! Hier besteht, wie oben ausgeführt, das **Problem** jedoch darin, **gleiche Leistungen inter-temporal und inter-personell zu identifizieren**.

b) **Vertikale Bewertungsgerechtigkeit**: Deutlich bessere Leistungen müssen zu besserer Beurteilung führen, und die Rangreihe muss **transitiv** sein (wenn A besser als B bewertet ist und B besser als C, dann muss A auch besser als C bewertet sein). Hier besteht das Problem darin, **Grenzen bzw. Abweichungsniveaus festzulegen, ab denen eine Leistung als so viel besser zu beurteilen ist**, dass sie zu einer besseren Beurteilung führt und den Prüfling somit in ein anderes Leistungs-Cluster einteilt. Wir vertiefen diesen Aspekt unter 3. bei der Frage nach der Gestaltung der Notensysteme.

Aus meiner Erfahrung hat sich folgendes **System zur Festlegung des Anspruchsniveaus** je Aufgabe und **zur Bemessung der Leistung** als fruchtbar erwiesen:

Jede **Aufgabe** einer Klausur erhält eine **Punktzahl**. Diese Punktzahl **entspricht den Minuten**, die der Studierende für diese Aufgabe (inkl. der Lektüre des Aufgabentextes) verwenden soll. Durch diese Punkte- = Minutenzahl wird dem **Schwierigkeitsgrad** einer Aufgabe Rechnung getragen. Tendenziell wird bei einer höheren Punktzahl auch eine umfangreichere Ausarbeitung erwartet. Die **Summe aller Punkte** entspricht der **Gesamtbearbeitungszeit** der Klausur. Somit hat der Prüfling unmittelbar einen **Orientierungsmaßstab** (nämlich die Zeit) hinsichtlich der erwarteten Bearbeitungsintensität einer Aufgabe. Die Punktzahl (bzw. die Leistung) kann dann auf 100 Punkte = 100% umgerechnet werden, um im nächsten Schritt schließlich Noten festzulegen.

Auch diese **Frage, wie das Leistungsniveau gemessen wird**, könnte man für die verschiedenen Dozenten durch eine **Befragung** zu ermitteln versuchen. Anders als bei 1. müsste hier

jedoch der Dozent (sich) selbst beurteilen, da der Prüfling nur in den seltensten Fällen Einblick haben kann:

- Wird eine „**Musterlösung**“ erstellt, der **klare und nachvollziehbare Beurteilungskriterien** zugrunde liegen?
- Wie werden die **Punkte** (Prozente der erreichten Leistung) auf die einzelnen Aspekte der (Muster-)Lösung **verteilt**? Inwiefern wird dabei der **Schwierigkeitsgrad** einer Aufgabe **berücksichtigt**?
- Wie werden **Zwischenlösungen** und **Folgefehler** bewertet?
- Wie werden vom Prüfling geforderte **Werturteile** bewertet?
- (Wie) Ist erkennbar, in welchem **Umfang** der Prüfer die **Beantwortung einer Frage** erwartet?
- (Wie/Inwiefern) Werden **Leistungsunterschiede** inter-temporaler oder inter-personeller Art zwischen verschiedenen Prüflingsgruppen **ausgeglichen**?

Um hier die **Transparenz** für die Prüflinge zu erhöhen, macht es Sinn, die **Musterlösungen** und auch die **Punkteverteilung** auf die einzelnen Teilaspekte der Musterlösung im Anschluss an die Klausur **bekannt zu machen**. Nur so können die Prüflinge – nachträglich, aber ggf. auch im Hinblick auf künftige Prüfungen – erfahren, auf welche Aspekte der Prüfer Wert legt, wie er bewertet etc.

3. Wie wird das gemessene Leistungsniveau in einem Notensystem codiert?

3.1. Verschiedene Ansätze zur Definition der Notensysteme

Um Leistungen, die Prüflinge bei verschiedenen Prüfern oder auch in verschiedenen Bildungseinrichtungen erzielen, vergleichbar zu machen, wird das gemessene **Leistungsniveau in ein Notensystem übertragen**, d.h. umcodiert. Verschiedene Länder, Kulturen oder auch nur Schulformen verwenden hierzu durchaus **unterschiedliche Systeme**, was letztlich die Vergleichbarkeit erschwert. Als „**Note**“ können dabei Zahlenwerte, Buchstaben oder Worte zur Anwendung kommen.

Die **Notensysteme**, die auf **Zahlenwerten** basieren, müssen folgendes festlegen:

- **Zahlenbereich**: von welcher Zahl bis zu welcher Zahl reicht die Notenskala?
- **Schrittweite/Abstufung**: Werden nur ganzzahlige Noten vergeben oder auch Nachkommastellen berechnet und angegeben?
- **Richtung**: Entspricht die größere Zahl der besseren oder der schlechteren Leistung?

- Ggf. **verbale Umschreibung** der Note.
- Was ist die **minimale Leistung** bzw. die Mindestnote, ab der eine Leistung noch akzeptabel und die Prüfung somit **bestanden** ist?
- **Umsetzungsmaßstab**/Code-Plan: Wie wird der gemessene **Leistungsgrad in Noten umgesetzt/umgerechnet**?

Im deutschen (Hoch-)Schulsystem herrscht weitgehend Einigkeit über folgende Festlegungen:

- Es gibt die (ganzen) **Noten 1, 2, 3, 4, 5, 6** (wobei in Hochschulen oft 5 und 6 als 5 zusammengefasst, somit nicht mehr unterschieden werden).
- Es gibt **Komma-Noten**, wobei manchmal jede Zehntel-Note möglich ist, manchmal nur die „…,3“ bzw. „…,7“ im Sinne von „minus“ bzw. „plus“ unterschieden wird.
- Je kleiner die Zahl, desto besser die Leistung (1 ist besser als 2 ist besser als 3 ….).
- Verbale Umschreibung: 1 = „sehr gut“, 2 = „gut“, 3 = „befriedigend“, 4 = „ausreichend“, 5 = „mangelhaft“, 6 = „ungenügend“.
- **Bestanden** gilt eine Prüfung ab der **Note 4**.

Zum Vergleich das Notensystem, wie es in **Frankreich** Anwendung findet:

Abbildung: Notensystem in Frankreich

"Mention" [Zeugnisnote]	entspricht Punktwerten von:	Bemerkung
très bien	ab 16 - 20 Punkten	Bestehensbereich 10 - 20 Punkte
bien	ab 14 Punkten	
assez bien	ab 12 Punkten	
passable	ab 10 Punkten	
médiocre	ab 8 Punkten	nicht bestanden
faible	ab 6 Punkten	
rès faible	ab 3 Punkten	
nul	0 - 2,9 Punkten	

Da es in Europa sehr verschiedene Benotungssysteme gibt, wurde das **ECTS-Notenschema** (ECTS-grading scale; ECTS = European Community Credit Transfer System) entwickelt, das den Hochschulen helfen soll, die von den Gasthochschulen vergebenen Noten dem heimischen System entsprechend zu transponieren und vice versa. Die ECTS-grades ersetzen hierbei nicht die heimische Art der Bewertung. Auf die **Problematik der Umrechnung** von Leistungsbewertungen („Noten“) zwischen den verschiedenen Benotungssystemen soll in diesem Beitrag nicht eingegangen werden; in der nachfolgenden Übersicht sind die allgemein als entsprechend geltenden deutschen Noten zur Information aufgeführt.

Abbildung: ECTS-Notensystem

ECTS-Noten	ECTS-Definition	entsprechende deutsche Noten
A	Excellent - outstanding performance and only a few minor mistakes	1,0 bis inkl. 1,5
B	Very good - above average performance, but some mistakes	schlechter als 1,5 bis inkl. 2,0
C	Good - overall good, solid work, but containing a few basic errors	schlechter als 2,0 bis inkl. 3,0
D	Satisfactory - undistinguished performance but no serious shortcomings	schlechter als 3,0 bis inkl. 3,5
E	Sufficient - meets the minimum requirements	schlechter als 3,5 bis inkl. 4,0
FX/F	Fail - improvement is essential before the work can be counted	schlechter als 4,0
F	Fail - major improvement required	

An der Fachhochschule in Wilhelmshaven z.B. gilt folgendes Notenschema:

Abbildung: Notensystem an der Fachschule in Wilhelmshaven

ganze Note	abgestuft in folgende Komma-Noten
1 = sehr gut	1,0 1,3
2 = gut	1,7 2,0 2,3
3 = befriedigend	2,7 3,0 3,3
4 = ausreichend	3,7 4,0
5 = nicht bestanden	5,0

Des Weiteren herrscht bei den meisten Professoren unserer Hochschule Einigkeit darüber, dass **mindestens 50%** der erwarteten Leistung vom Studierenden erbracht sein müssen, damit die **Note 4,0** gewährt wird. Wie oben angedeutet wird von manchen Prüfern jedoch hiervon **abgewichen**: War die Klausur im Vergleich zu früheren Klausuren überdurchschnittlich schwer und/oder waren die Leistungen der Prüflinge ungewöhnlich schlecht (beides stellt der Prüfende i.d.R. erst bei der Korrektur fest), so wird die Grenze auf z.B. 40% abgesenkt, um auch inter-temporal bzw. inter-personell einer Beurteilungsgerechtigkeit näher zu kommen. Denselben Effekt ein Notensystem, das jeweils nur das Leistungsniveau der geprüften Gruppe

betrachtet („**Sozialnorm**“). So kann das ECTS-Notenschema auch wie folgt angewandt werden:

Abbildung: ECTS-Notensystem mit Leistungs-Clustern

ECTS-Noten	Definition mittels Leistungs-Cluster
A	die 10% Besten
B	die darauf folgenden 25% Guten
C	die darauf folgenden 30%
D	die darauf folgenden 25%
E	die schlechtesten 10%, die noch bestanden haben
FX/F	durchgefallen

Gerade um eine inter-temporale Bewertungsgerechtigkeit anstreben zu können, sollte diese Art der Bewertung jedoch nur dann angewandt werden, wenn **über mehrere Gruppen von Prüflingen Erfahrungswerte** vorliegen, da eine isolierte Betrachtung nur einer Gruppe u.U. zu sehr starken (nicht gerechtfertigten) Verschiebungen im Notensystem führen könnte.

3.2. Noten als ordinale Messsysteme

Einigkeit herrscht auch darüber, dass zahlenmäßige Notensysteme **ordinal skaliert** sind, somit eine **Rangskala** darstellen, also **nicht metrisch** zu interpretieren sind. Daraus folgt z.B.:

- Eine „1“ ist besser als eine „2“, aber keinesfalls doppelt so gut oder gar nur halb so gut (die Hälfte von 2 ist 1; da die Beurteilungsrichtung aber umgekehrt läuft könnte man auf die Idee kommen, „1“ wäre doppelt so gut wie „2“ – dies ist i.d.R. nicht der Fall).
- Eine „4“ ist schlechter, aber eben nicht doppelt so „schlecht“ wie eine „2“ (dieses Adjektiv würde bei einer „2“ ohnehin kaum passen).
- Um von einer „4“ auf eine „3“ zukommen bedarf es **nicht unbedingt derselben Mehrleistung** wie man sie benötigt, um von einer „2“ auf eine „1“ zu kommen. So könnte der Prüfer die Meinung vertreten, dass der Sprung von „4“ auf „3“ relativ leicht zu schaffen sein soll, während man für den Sprung von der „2“ auf die „1“ schon hervorragend sein soll, so dass der prozentuale Leistungsbereich der „2“ viel weiter gefasst wird als derjenige der „4“. Er könnte aber auch die erforderlichen Mehrleistungen genau umgekehrt ansetzen …

Daraus folgt natürlich, dass eine „sture“ rechnerische Zuordnung von Leistungsprozentsätzen („Punkten“) zu Noten im Allgemeinen nicht sinnvoll ist. Auch macht es eigentlich keinen

Sinn, **Durchschnittsnoten** (arithmetisches Mittel) zu ermitteln, da man hierzu eigentlich **metrische Skalen** (**Verhältnisskalen** (Gleichheit von Verhältnissen) oder zumindest **Intervallskalen** (Gleichheit von Differenzen; Äquidistanz zwischen den Notenwerten)) vorliegen haben müsste. Auch ist der „**Nullpunkt**“ der Notenskala nicht eindeutig definiert: Ist „Null-Leistung“ eine „5,0“ oder z.B. „0%“ oder „< 50%“ der erreichbaren Leistungspunkte? Falls alles unter 50% erreichte Leistung als „Null-Leistung“ definiert wird, wäre jemand, der 80% richtige Antworten hat, nicht um ¼ besser als jemand, der 64% erreicht hat (64 + 40% von 64 = 80), sondern sogar 214% besser ((64-50) + 214% = (80-50)) – doch ist auch diese Interpretation unzulässig bzw. ohne wahren Sinn, dies nicht nur, weil der „wahre“ Nullpunkt kaum sinnvoll zu ermitteln ist, sondern auch, weil man bei Noten eben von einer Ordinalskala ausgeht.

3.3. Zuordnung von Leistungsniveau und Notensystem

Mehr oder weniger große **Uneinigkeit** herrscht nun aber darüber, **wie das mit Prozentwerten festgestellte Leistungsniveau auf die Notenskala übertragen werden soll**. Hier gibt es durchaus **unterschiedliche Philosophien** bei verschiedenen Prüfern, wie eine – sicherlich nicht repräsentative, aber doch sehr informative – Umfrage unter Professorenkollegen unserer Hochschule gezeigt hat – wobei vielen Prüfern die hinter ihrer Notengebung stehende „Philosophie“ gar nicht gewusst sein dürfte.

Ausgangspunkt für diese Untersuchung – und damit auch für den gesamten hiermit vorliegenden Aufsatz – war die Feststellung, dass viele Studierende der Meinung sind, bei mir schlechtere Noten zu bekommen als bei Kollegen. Dies könnte, so meine Überlegung, ja daran liegen, dass mein Notenschema „strenger“ ist als das der Kollegen. Nun, dem ist nicht so, wie nachfolgende Analyse zeigt. Somit liegen, falls denn die Vermutung der Studierenden zutreffend ist, die Gründe in den unter 1. und 2. genannten Aspekten der Leistungsbeurteilung.

Dem von mir bislang verwendeten **Notenschlüssel** liegt folgende **Philosophie** zugrunde:

- **50%** richtige Antworten müssen mindestens erreicht werden, um eine Klausur zu bestehen.
- Jede **ganze Note** (1, 2, 3, 4) umfasst eine **identische Spanne von Punkten** (bzw. Prozenten richtiger Antworten).
- Da es keine 0,7 und keine 4,3 gibt, werden die Punkte für diese beiden Noten der 1,0 bzw. 4,0 zugeschrieben.

- Es ist relativ leicht, in den **4er-Bereich** zu gelangen (z.B. durch Ausgleich von schlechten/schwachen Leistungen in vielen Aufgaben durch gute Leistungen in einigen anderen Aufgaben). Daher wird dieser Bereich (obwohl nur mit zwei Notenstufen abgedeckt, nämlich 4,0 und 3,7) ebenso **groß gefasst** wie die nachfolgenden Notenbereiche (Spanne von 12,5 Prozentpunkten).
- Es ist sehr schwierig, in den **obersten Notenbereich** zu gelangen (1,3 bzw. 1,0), da der Student hierfür auf *allen* Gebieten der Klausur hervorragend sein muss (kein Ausgleich möglich). Daher wird auch dieser Bereich (obwohl nur mit zwei Notenstufen abgedeckt) ausreichend **groß gefasst** (Spanne von 12,5 Prozentpunkten).

Abbildung: Notenschlüssel Prof. Dr. Kirstges

Punkte			Note	Spanne	Spanne
ab inkl.		bis unter			ganze Noten
0,0	-	50,0	5,0	50,0	50,0
50,0	-	58,5	4,0	8,5	12,5
58,5	-	62,5	3,7	4,0	
62,5	-	66,5	3,3	4,0	
66,5	-	71,0	3,0	4,5	12,5
71,0	-	75,0	2,7	4,0	
75,0	-	79,0	2,3	4,0	
79,0	-	83,5	2,0	4,5	12,5
83,5	-	87,5	1,7	4,0	
87,5	-	91,5	1,3	4,0	12,5
91,5	bis inkl.	**100,0**	1,0	8,5	

Die **Industrie- und Handelskammern** verwenden einen Notenschlüssel, an den sich manche Kollegen anlehnen. Wenngleich nicht explizit so formuliert, kann man daraus folgende Philosophie erkennen:

- Nur die **ganzen Noten** 1, 2, 3, 4, 5 und 6 werden von der IHK eingeteilt (hier fett eingetragen). Dabei wird die Leistungs-/Prozentspanne zu den guten Noten hin enger, d.h. man kann **mit immer weniger Punkten schneller "aufsteigen"**: Bis zur 4 braucht man 50%, bis zur 3 dann nur noch 17%, bis zur 2 noch 14%, bis zur 1 noch 11% (Leistungsspannen verringern sich immer um 3 Prozentpunkte).
- Die hier abgebildete Unterteilung in die Zwischennoten wurde gemäß diesem System ergänzt.
- Zu den **guten Noten hin "strenger" Ansatz**: Wer eine gute 2 oder gar 1 bekommen möchte, muss nahezu alle Aufgaben perfekt gelöst haben (geringe Spanne in den oberen Noten).

Abbildung: IHK-Notenschlüssel

Punkte			Note	Spanne	Spanne
ab inkl.		bis unter			ganze Noten
0,0	-	50,0	5,0	50,0	50,0
50,0	-	60,0	4,0	10,0	17,0
60,0	-	67,0	3,7	7,0	
67,0	-	72,0	3,3	5,0	
72,0	-	77,0	3,0	5,0	14,0
77,0	-	81,0	2,7	4,0	
81,0	-	85,0	2,3	4,0	
85,0	-	89,0	2,0	4,0	11,0
89,0	-	92,0	1,7	3,0	
92,0	-	97,0	1,3	5,0	8,0
97,0	bis inkl.	**100,0**	1,0	3,0	

Ein anderer Kollege verwendet folgenden Ansatz:

- 50% richtige Antworten müssen mindestens erreicht werden, um eine Klausur zu bestehen.
- Zu den guten Noten hin "strenger" Ansatz: Wer eine gute 2 oder gar 1 bekommen möchte, muss nahezu alle Aufgaben perfekt gelöst haben (geringe Spanne in den oberen Noten).

Abbildung: Notenschlüssel Kollege A

Punkte			Note	Spanne	Spanne
ab inkl.		bis unter			ganze Noten
0,0	-	50,0	5,0	50,0	50,0
50,0	-	60,0	4,0	10,0	17,0
60,0	-	67,0	3,7	7,0	
67,0	-	73,0	3,3	6,0	
73,0	-	78,0	3,0	5,0	16,0
78,0	-	83,0	2,7	5,0	
83,0	-	87,0	2,3	4,0	
87,0	-	91,0	2,0	4,0	12,0
91,0	-	95,0	1,7	4,0	
95,0	-	98,0	1,3	3,0	5,0
98,0	bis inkl.	**100,0**	1,0	2,0	

Wieder ein anderer Kollege verwendet folgendes Schema:

- 50% richtige Antworten müssen mindestens erreicht werden, um eine Klausur zu bestehen.
- **Gleichmäßige Aufteilung in 5%-Schritten.**

Abbildung: Notenschlüssel Kollege B

Punkte			Note	Spanne	Spanne
ab inkl.		bis unter			ganze Noten
0,0	-	50,0	5,0	50,0	50,0
50,0	-	55,0	4,0	5,0	10,0
55,0	-	60,0	3,7	5,0	
60,0	-	65,0	3,3	5,0	
65,0	-	70,0	3,0	5,0	15,0
70,0	-	75,0	2,7	5,0	
75,0	-	80,0	2,3	5,0	
80,0	-	85,0	2,0	5,0	15,0
85,0	-	90,0	1,7	5,0	
90,0	-	95,0	1,3	5,0	10,0
95,0	bis inkl.	**100,0**	1,0	5,0	

Würde man, um einen letzten Ansatz vorzustellen, die Punkte/Prozente zwischen 50 und 100 **proportional zu den Zehntel-Notenschritten** aufteilen, so ergäbe dies folgenden Notenschlüssel:

Von 1,0 bis 5,0 gibt es 41 Zehntel-Notenschritte:

1,0	2,0	3,0	4,0	5,0
1,1	2,1	3,1	4,1	
1,2	2,2	3,2	4,2	
1,3	2,3	3,3	4,3	
1,4	2,4	3,4	4,4	
1,5	2,5	3,5	4,5	
1,6	2,6	3,6	4,6	
1,7	2,7	3,7	4,7	
1,8	2,8	3,8	4,8	
1,9	2,9	3,9	4,9	

Diesen 41 Notenschritten entsprechen 51 Prozentpunkte (von 50 bis 100 Punkte):

51 Punkte = 41 Zehntel-Notenschritte, d.h.
0,3 Notenschritt = 3 Zehntel = 3,73
0,4 Notenschritt = 4 Zehntel = 4,98
1,0 Notenschritt = 10 Zehntel = 12,44

Daraus resultiert - ausgehend von Note 1,0 = 100 Punkte - folgendes Notenschema:

Abbildung: Notenschlüssel Zehntel-Gleichverteilung

Punkte			Note	Spanne	Spanne
ab inkl.		bis unter			ganze Noten
0,0	-	50,2	5,0	50,2	50,2
50,2	-	62,7	4,0	12,4	16,2
62,7	-	66,4	3,7	3,7	
66,4	-	71,4	3,3	5,0	
71,4	-	75,1	3,0	3,7	12,4
75,1	-	78,9	2,7	3,7	
78,9	-	83,8	2,3	5,0	
83,8	-	87,6	2,0	3,7	12,4
87,6	-	91,3	1,7	3,7	
91,3	-	96,3	1,3	5,0	8,7
96,3	bis inkl.	**100,0**	1,0	3,7	

Es sei nochmals betont, dass Notensysteme ein **ordinales Skalenniveau** aufweisen. Es ist also keinesfalls zwingend erforderlich, dass das Leistungsniveau eines Prüflings zwischen allen Notenstufen jeweils um denselben Wert ansteigt bzw. abnimmt. Denkbar wären also auch nachfolgend dargestellte Zuordnungen von Leistungsniveaus zu Noten:

Abbildung: Stark verzerrte ordinale Notensysteme

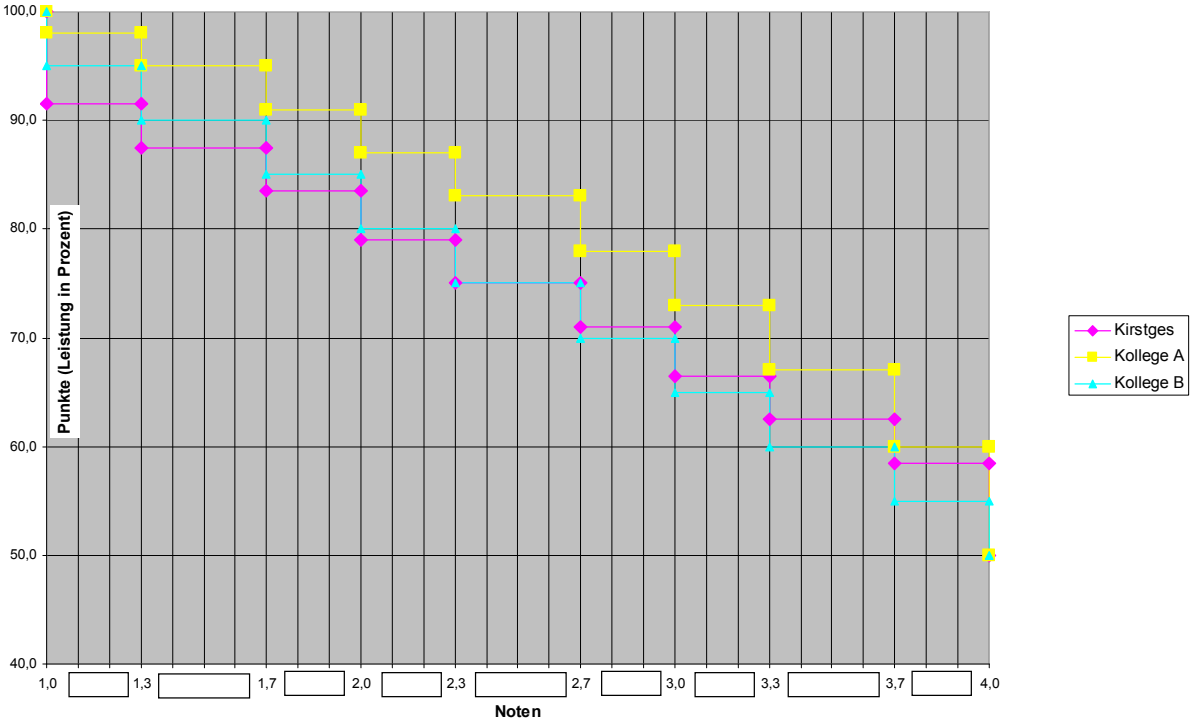
Punkte			Note	Spanne	Spanne
ab inkl.		bis unter			ganze Noten
0,0	-	50,0	5,0	50,0	50,0
50,0	-	61,0	4,0	11,0	21,0
61,0	-	71,0	3,7	10,0	
71,0	-	77,0	3,3	6,0	
77,0	-	83,0	3,0	6,0	18,0
83,0	-	89,0	2,7	6,0	
89,0	-	92,0	2,3	3,0	
92,0	-	95,0	2,0	3,0	9,0
95,0	-	98,0	1,7	3,0	
98,0	-	99,0	1,3	1,0	2,0
99,0	bis inkl.	**100,0**	1,0	1,0	

Punkte			Note	Spanne	Spanne
ab inkl.		bis unter			ganze Noten
0,0	-	50,0	5,0	50,0	50,0
50,0	-	51,0	4,0	1,0	2,0
51,0	-	52,0	3,7	1,0	
52,0	-	55,0	3,3	3,0	
55,0	-	58,0	3,0	3,0	9,0
58,0	-	61,0	2,7	3,0	
61,0	-	67,0	2,3	6,0	
67,0	-	73,0	2,0	6,0	18,0
73,0	-	79,0	1,7	6,0	
79,0	-	89,0	1,3	10,0	21,0
89,0	bis inkl.	**100,0**	1,0	11,0	

Gleichwohl **führt eine zu starke Verzerrung automatisch zu einem gewissen Störungsgefühl**, zu einem Eindruck von „Ungerechtigkeit“, der vermieden werden sollte.

Die **Auswirkungen der unterschiedlichen Notensysteme auf die Beurteilung** eines einzelnen Prüflings kann man am besten grafisch erkennen.

Abbildung: Vergleich der von verschiedenen Prüfern verwendeten Notensysteme



Man erkennt, dass das **Kirstges’che Notenspektrum** über alle Notenstufen (mit Ausnahme von 4,0 und 1,0) „**studentenfreundlicher**“ ist **als das des Kollegen A**: So gibt es bei z.B. 70% Leistungsniveau bei Kirstges eine „3,0“, bei Kollege A nur eine „3,3“. Für 90% gibt es

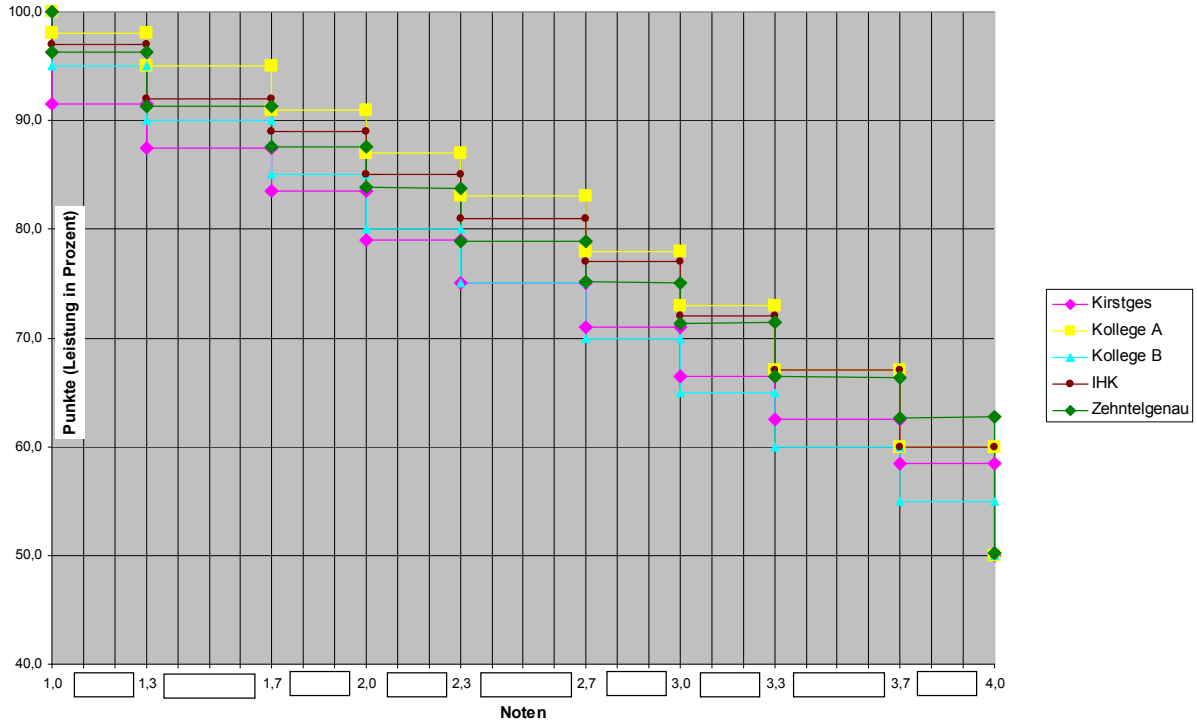
bei Kirstges eine „1,3“, bei A nur eine „1,7“. Oder anders herum gelesen: Eine „2,3“ gibt es bei Kirstges schon für 75% der maximal erzielbaren Leistung, bei A erst ab 83%.

Gegenüber **Kollege B** ist das Kirstges’che Notensystem im Bereich der besseren Noten (ab „2,7“) studentenfreundlicher, da man in diesem Bereich mit weniger Leistungszuwächsen, also sozusagen leichter zu besseren Noten kommt. Im Bereich der schlechteren Noten ist Kollege B jedoch studentenfreundlicher: Der Prüfling kommt hier schon bei geringeren Leistungsverbesserungen zu einer weniger schlechten Note.

Nun, die These, die den Impuls für diese Untersuchung gab, dass nämlich Studierende bei Kirstges eher schlechtere Noten deshalb bekommen, weil der Notenspiegel strenger ist, kann also nicht aufrechterhalten werden. Falls der Notendurchschnitt bei den Kirstges’chen Noten tatsächlich schlechter ist, muss dies an den unter 1. und 2. beschriebenen Effekten liegen.

Die nachfolgende Abbildung zeigt abschließend den Vergleich aller fünf dargestellten Notensysteme.

Abbildung: Vergleich von fünf verschiedenen Notensystemen



Derart unterschiedliche Systeme der Zuordnung von Noten zu Leistungsniveaus können von Studierenden als „ungerecht“ wahrgenommen werden. Für den einzelnen Studierenden ist es schon ein deutlicher Unterschied, ob er bei z.B. 80% erzieltem Leistungsniveau eine „2,0“ (Kirstges, Kollege B), eine „2,3“ (zehntelgenaues System) oder eine „2,7“ (Kollege A und IHK-System) bekommt.

Andererseits würde auch eine Einheitlichkeit des Maßstabes nicht weiterhelfen, da – wie oben ausgeführt – jeder Prüfende mit je seiner Methode spezifische Fähigkeitskombinationen feststellt und benotet. Ein (lediglich) einheitlicher Maßstab würde eine „**Scheinobjektivierung**“ bedeuten. Lediglich über alle Noten aller Dozenten kann sich somit ein „richtiges“ Gesamtbild der Leistungskraft eines Studierenden ergeben.

Es kann also **kein wirklich „gerechtes“**, vermutlich nicht einmal ein wirklich leistungsgerechtes **Notensystem geben**. Es würde zu weit führen, hier auf diese grundsätzliche Frage näher einzugehen. Kurz sei z.B. auf Lyotard hingewiesen, gemäß dem es **keine absolute Gerechtigkeit innerhalb einer Gesellschaft geben kann**, da Gerechtigkeit naturgesetzmäßig Ungerechtigkeit beinhaltet. Diese unvermeidbare Ungerechtigkeit besteht darin, dass man sich immer für eine Möglichkeit (hier: ein Notensystem) entscheiden muss und dabei andere unrealisierbar bleiben, obwohl sie völlig gleichberechtigt sind, weil wegen des Fehlens von Metaregeln eine Legitimation des einen oder des anderen Tuns unmöglich ist. (Lyotard, J.F. : Der Widerstreit. München 1987). Dennoch wird die quasi institutionell ausgeübte „Notengerechtigkeit“ im Bildungssystem in der liberalen Demokratie am effizientesten verwirklicht, wenngleich sie ohne Zweifel auch im Detail noch verbesserungsbedürftig ist. Da eine allgemeine subjektive **Definition von Gerechtigkeit** in das Charaktermerkmal von Menschen einwirkt und in der Tradition – neben Klugheit, Besonnenheit und Tapferkeit – zu den vier Kardinalstugenden zählt, sind wir **Professoren doch geradezu prädestiniert, für (Noten-) Gerechtigkeit unter unseren Prüflingen zu sorgen** … ;-)

Daher macht es m.E. Sinn, wenn zumindest in diesem Bereich der Notenfestlegung **Transparenz** – nicht unbedingt **Einheitlichkeit -** herrscht. Darauf hinzuwirken ist Sinn und Zweck dieses Beitrages.

Ich danke meinen Kollegen Prof. Dr. Christian-Uwe Behrens, Prof. Dr. Uwe Weithöner und Prof. Dr. Knut Scherhag für ihre Informationen und Anregungen!

Prof. Dr. Torsten Kirstges