Sind Examensnoten vergleichbar? Und was, wenn Noten immer besser werden?

Der Versuch eines Tabubruchs

Volker Müller-Benedict
Thomas Gaens
Flensburg

Die Note im Abschlussexamen stellt einen Indikator sowohl für den Ertrag der eigenen im Studium erbrachten Anstrengung als auch für die Chancen des beruflichen Werdegangs dar. Darum ist Gerechtigkeit, das heißt gleiche Noten für gleiche Leistung, bei der Notenvergabe Pflicht. Für einen einzelnen Prüfling ist es schwer nachzuweisen, dass er trotz gleicher Leistung nicht dieselbe Note erhalten hat wie ein anderer. Da auch kein Prüfer ohne Folgen öffentlich eingestehen kann, dass er nicht gerecht prüft, werden die Höhe der Noten und deren Zustandekommen im Allgemeinen nicht verglichen – das ist ein Tabu an den Hochschulen. Dieses Schweigen betrifft nicht nur den Vergleich einzelner Prüfungen, sondern auch den Vergleich von Noten zwischen Hochschulen, Fächern, und Studierenden mit unterschiedlichen Merkmalen in Geschlecht, Nationalität, sozialer Herkunft etc. Neuerdings wird inbesondere der zeitliche Vergleich der Noten in Frage gestellt, indem *grade inflation*[1], eine ständige Verbesserung des Notendurchschnitts über viele Jahre hinweg, konstatiert wird.

Erste Vergleiche von Noten im Verlauf der letzten 15 Jahre zwischen vielen Studiengängen und Universitäten zeigen allerdings, dass es unbezweifelbare Niveauunterschiede und Notentrends gibt (Wissenschaftsrat 2003, 2007, 2012; Müller-Benedict/Tsarouha 2011). Diese bedeuten allerdings nicht sofort, dass die Vergleichbarkeit der Noten in Frage gestellt ist. Das gilt nur, wenn die Unterschiede nicht auf unterschiedlichen Leistungen beruhen. Dann sollte man das Tabu brechen und auf systematische, nicht auf unterschiedlichen Prüfungsleistungen beruhende Unterschiede im Notenniveau hinweisen. Dieser Beitrag soll neben der Beschreibung der langfristigen Notenentwicklung von 1960 bis heute insbe-

[1] Wenn die Noten auf Grund besserer Leistungen besser werden, ist dies nach der gebräuchlichen Definition keine *grade inflation*. Wir verwenden den Begriff hier dennoch allgemein für den langfristigen Trend zu besseren Noten.

sondere zeigen, wie aus leistungsfremden Einflüssen *grade inflation* entstehen kann. Aus dieser Analyse ergibt sich, über welche Auffälligkeiten dieser Unterschiede zu sprechen sich lohnen könnte.

Eine andere institutionalisierte Art der Leistungsmessung, die von möglichen systematischen Einflussfaktoren unabhängiger wäre als Noten, existiert an Hochschulen nicht. Im Schulsystem wird eine solche Alternative durch einheitliche hochstandardisierte Abiturprüfungen in einigen Bundesländern angestrebt. Selbst dabei lassen sich Notenunterschiede zwischen Bundesländern feststellen, die nicht auf Kompetenzunterschieden beruhen (Neumann et al. 2009). Im Folgenden wird deshalb zunächst theoretisch entwickelt, wie die Notenniveaus sein müssten, wenn sie nur die Bandbreite der Leistungen wiedergeben würden (Abschnitt 1). Diese Annahmen werden dann mit den empirischen Noten verglichen. Dazu werden die Notenentwicklung einiger Fächer sowie ihre Folgen beschrieben (Abschnitt 2). Es zeigt sich, dass die Dynamik fachspezifisch ist und meist zu *grade inflation* führt. Zur Erklärung dieser Phänomene werden erst Ursachen für zyklische Verläufe (Abschnitt 3) und dann für *grade inflation* (Abschnitt 4) erläutert.

1. Theoretische Überlegungen zum Notenniveau

Theoretisch wird zwischen drei Bezugsnormen für Noten unterschieden: der individuellen (Bewertung der individuellen Verbesserung), der sozialen (Bewertung im Vergleich zur Bezugsgruppe, z.B. Klasse, Seminar) und der absoluten (Bewertung anhand eines geprüften Wissens-/Kompetenzkanons) (Rheinberg 2002). An den Hochschulen sollte in den Abschlussnoten die absolute Bezugsnorm im Vordergrund stehen, da sie den relativen Wissensstand des Absolventen in Bezug auf den akademischen Wissensbestand signalisieren sollen. Da das akademische Wissen sich allerdings ständig weiterentwickelt, kann die absolute Bezugsnorm für den intertemporalen Vergleich nicht gelten – eine 1-er Leistung in Chemie 1930 würde heute vermutlich nicht einmal eine 4 erreichen. Die Notenskala gilt also je Zeitpunkt relativ zum aktuellen Wissen.

Das Anliegen, unterschiedliche Leistungen differenzieren zu können, beginnt bei der Testkonstruktion. Ein Test sollte sowohl schwierige als auch leichte und mittlere Aufgaben in einer gleichmäßigen Häufigkeit aufweisen, sonst gilt er als „zu leicht“ oder „zu schwer“ in Bezug auf das absolute Bezugsniveau. Schulmaterialien für Tests setzen die Fehlerpunkte für die Grenzen zwischen den Noten so fest, dass es nicht zu viele „sehr gute“ und „ausreichende“ gibt, und die Mehrheit ein „gut“ oder „befriedigend“ erhält (Lehnen/Loch 1978). Andere Verteilungen gelten

als didaktisch problematisch bzw. falsch konstruiert. Wenn etwa alle eine ähnliche Note erreichen, kann das nur schwer mit einer zufällig gleichen Leistung aller Kandidaten erklärt werden. In der Schule vermischt sich die absolute Testleistung mit dem relativen Niveau der Klasse und dem individuellen Lernniveau der Schüler zu einer Note. Dadurch können sich die Noten von der Testleistung unterscheiden. An Hochschulen, vor allem in abschließenden Examensnoten, die ja nicht mehr für weitere Lernprozesse, z.B. als Motivation, verwendet werden können, sollten solche Einflüsse nicht vorhanden sein. Aus diesen Überlegungen heraus ist für Abschlussexamen an Hochschulen eine etwa gleich bleibende Streuung der Noten über die Zeit erwartbar.

Entspricht die tatsächliche Entwicklung diesen Annahmen? Im Folgenden analysieren wir für ausgewählte Studiengänge und Universitäten Zeitreihen von Noten über die letzten 50 Jahre. Die Daten stammen aus dem DFG-Forschungsprojekt „Die Notengebung an Hochschulen in Deutschland von den 1960er Jahren bis heute. Trends, Unterschiede, Ursachen." Im Projekt wurden an sieben Hochschulen in bis zu 12 Studiengängen die Examensnoten der letzten Jahrzehnte erhoben und mit aggregierten Daten der amtlichen Hochschulprüfungsstatistik verknüpft (Stat. Landesamt Schleswig-Holstein 2012). Aus diesen Daten werden aus Platzgründen nur einige Beispiele herausgegriffen, die Ergebnisse sind jedoch auch für weitere Daten bestätigt worden.

2. Veränderungen des Notenniveaus und *grade inflation*

Wenn das Durchschnittsniveau der Noten sich immer mehr dem unteren Rand (der besten Noten) nähert, muss die Streuung notwendigerweise wegen der Begrenztheit der Skala kleiner werden. Dann können aber wegen der feststehenden Unterteilungen[2] auch nur weniger unterschiedliche Noten vergeben werden, also werden viele dieselbe Note erhalten. Deshalb führt das Auftreten von *grade inflation* direkt auf Gerechtigkeitsprobleme, weil immer mehr unterschiedliche Leistungen gleich bewertet werden müssen. Die betroffenen Absolventen werden das eher nicht als problematisch ansehen, da sie ja dann meist sehr gute Noten erhalten. Vergleichen sich Absolventen anderer Fächer, Universitäten oder Zeit-

[2] Da die möglichen auf dem Abschlusszeugnis erscheinenden Noten an jeder Universität individuell geregelt werden, gibt es hier eine überraschende Spannbreite von nur den vier möglichen ganzen Zahlen 1, 2, 3 oder 4 bis zu beliebig unterteilten Rationalzahlen, z.B. 1,41.

räume jedoch mit ihnen, wird die unterschiedliche Bewertung zum Problem.

Grade inflation stellt deshalb per se eine Bedrohung gerechter Beurteilung dar, wenn man von der extremen Annahme absieht, dass alle Prüflinge immer einheitlichere, bessere Leistungen erbringen. Die Verbesserungen im erfassten Zeitraum zeigt Tabelle 1. Das Notenniveau hat sich bis 2010 in 9 der 12 Studiengänge um eine halbe bis eine Note verbessert. Hier sieht man aber schon die Notwendigkeit von Differenzierungen: Soziologie Magister, Maschinenbau und Jura sind nicht betroffen.

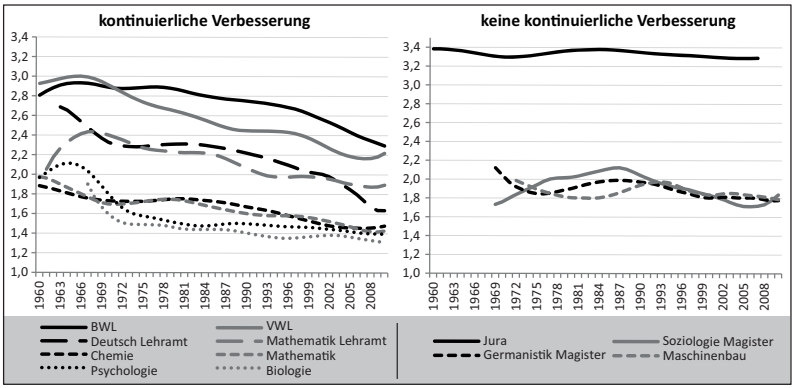
Tabelle 1: Durchschnittliche Notenverbesserungen (Extreme) zwischen den 1960er Jahren und 2010

Studiengang	Schlechteste Durchschnittsnote im ersten Jahrzehnt (Jahr)	Beste Durchschnittsnote im letzten Jahrzehnt (Jahr)	Differenz	Verbesserung in % ggü. dem schlechtesten Niveau
Deutsch Lehramt	2,71 (1965)	1,58 (2006)	1,13	41,7 %
VWL	3,09 (1964)	2,00 (2006)	1,09	35,3 %
Psychologie	2,40 (1965)	1,39 (2004)	1,01	42,1 %
Mathematik Lehramt	2,56 (1965)	1,68 (2009)	0,88	34,4 %
Mathematik	2,17 (1963)	1,38 (2002)	0,79	36,4 %
BWL	3,05 (1965)	2,30 (2009)	0,75	24,6 %
Biologie	1,92 (1967)	1,31 (2010)	0,61	31,8 %
Chemie	1,98 (1960)	1,40 (2006)	0,58	29,3 %
Germanistik (Magister)	2,15 (1969)	1,60 (2002)	0,55	25,6 %
Maschinenbau	2,00 (1972)	1,79 (2009)	0,21	10,5 %
Soziologie (Magister)	1,71 (1969)	1,57 (2005)	0,14	8,2 %
Jura (1. Staatsexamen)	3,41 (1964)	3,27 (2005)	0,14	4,1 %

Allerdings gibt die Differenz nur das Verhältnis zweier Zeitpunkte wieder. Die Extremwerte könnten nur Minimum bzw. Maximum einer zyklischen Entwicklung sein. So zeigt sich, dass in den ersten 8 Studiengängen in Tabelle 1 tatsächlich *grade inflation* vorliegt, die letzten vier jedoch ein weitgehend konstantes Niveau zeigen (Abb. 1)[3].

[3] Die Zeitreihen sind grundsätzlich mit der lowess-Technik geglättet, um die kurzfristigen Schwankungen zu unterdrücken.

Abbildung 1: Studiengänge mit und ohne kontinuierliche Verbesserung



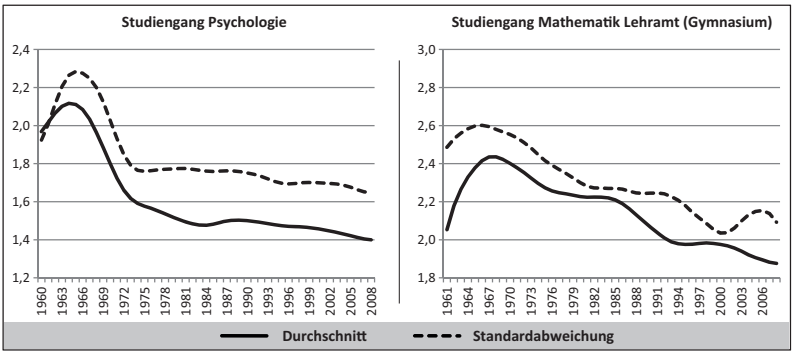
Ersichtlich existiert eine Notenhierarchie, die über lange Phasen im Zeitraum konstant bleibt: Biologie und Psychologie vergeben die besten Noten, Jura und BWL die schlechtesten. In das Notenspektrum eines Faches wird man im Laufe des Studiums hineinsozialisiert: nur wenige Erstsemester in Biologie und wenige zum ersten Mal bewertende Dozierende werden wissen, dass Sie mindestens eine 1,5 bekommen bzw. im Durchschnitt geben werden. Es ist unhinterfragtes Professionswissen, Teil einer Fachkultur. Aber dieser Teil ist Tabu für nicht im jeweiligen Fach Tätige.

Die Tatsache, dass es in Biologie und Psychologie seit ca. 30 Jahren fast nur Examensnoten zwischen 1 und 2 gibt, ist nicht nur für die Öffentlichkeit eine Überraschung (z.B. der SPIEGEL vom 28.2.2011: Hochschulen – Alles Spitze). Leider werden in der Regel in den Universitätsarchiven und in der Hochschulstatistik Nichtbestehende gar nicht erfasst (Gaens 2013). Wenn sie erfasst wären, wie sollte die „Lücke" zwischen „nicht bestanden" und „gut" erklärt werden? Sehr wirksam und sogar gerichtswirksam (Verwaltungsgericht Münster 2010) wird diese Tatsache allerdings, wenn Notenniveaus als Eingangsvoraussetzungen, etwa für den Masterstudiengang oder für das Lehramtsreferendariat, festgeschrieben werden. Durch solche Regelungen werden die Studiengänge und Universitäten mit hoher *grade inflation* systematisch bevorzugt.

Weiter sind überall zyklische Schwankungen von längerer Dauer (im Mittel 20 Jahre) beobachtbar, bei den Studiengängen mit *grade inflation* überlagert vom generellen Trend zu besseren Noten. Das Vorhandensein von Zyklen ist nur durch nicht leistungsbedingte, systematische Einflüsse auf die Notengebung erklärbar, weil es keine Theorien gibt, die erklären könnten, dass die Leistung von Studierenden in langjährigen Zyklen schwankt.

Diese deskriptiven Ergebnisse bedeuten, dass jeder Studiengang sein eigenes Notenniveau hat, das man kennen muss, wenn man die dortigen Examensnoten beurteilen möchte. Obwohl dies dem Konzept einer Notenskala widerspricht, könnte man das als Teil der Fachkultur für vertretbar halten. Nicht nur im Rahmen zunehmender Interdisziplinarität sollte darüber jedoch gesprochen werden. Zudem sind diese fachkulturellen Differenzen nicht stabil, die Niveaus und ihre Beziehung zueinander haben sich in kürzeren oder längeren Zeiträumen verändert, einige Studiengänge haben sich sogar bis zu einer Note verbessert. Dass diese Entwicklung zu besseren Noten direkt mit einer Einschränkung der Verteilung der Noten auf einen kleineren Bereich einhergeht, zeigt die Abbildung 2 für Psychologie und Mathematik (höheres) Lehramt. Die Streuungen der Noten (in der Grafik mit 3 multipliziert) werden desto kleiner, je besser die Noten werden.

Abbildung 2: Durchschnitt und Standardabweichung der Noten in den Studiengängen Psychologie und Mathematik Lehramt (Gymnasium)



Alle Entwicklungen führen dazu, dass die Noten ihre Funktion des Leistungsvergleichs verlieren. Darüber sollte auf jeden Fall gesprochen werden. Im folgenden Kapitel soll eine mögliche Erklärung für die Zyklen und im letzten Kapitel für *grade inflation* gegeben werden. Damit werden Aspekte aufgezeigt, unter denen diese Tabus der Notendifferenzen angesprochen werden können.

Die vorstellbaren Einflüsse auf die Notengebung sind zahllos (Hu 2005). Genannt wird z.B. die Reputation eines Instituts. Sie kann sowohl zu einer unterdurchschnittlichen („wir prüfen besonders hart“) wie zu einer überdurchschnittlichen („wir sind Exzellenzuni, unsere Noten sind deshalb besser“) Notengebung führen. Insbesondere in der amerikanischen Diskussion werden Lehrevaluationen und andere Anreizsysteme

diskutiert, die zum Tauschhandel „bessere Noten gegen bessere Bedingungen/Bewertungen für die Lehrperson“ führen (Kuh 2003).

Für viele dieser theoretisch denkbaren Einflüsse gibt es keine empirischen Daten, für so gut wie gar keine Längsschnittdaten, einige – wie die Lehrevaluation – gibt es erst seit wenigen Jahren. Deshalb müssen wir uns im Folgenden auf wenige Einflüsse beschränken, die aber gerade in der Lage sind, langfristige Dynamiken und insbesondere Zyklen zu erklären, was für viele in der Literatur genannte Gründe nicht zutrifft.

3. Zyklen, der fachspezifische Arbeitsmarkt und Prüfungsbelastung

Womit jede Universität zu kämpfen hat, sind die Schwankungen der Studierendenzahlen, die sich bei den Examenssemestern als zyklische Bewegung der Prüfungszahlen zeigen. Abbildung 3 zeigt die Konjunktur der Prüfungszahlen für fünf verschiedene Universitäten in Mathematik Lehramt. Um die Schwankungen der Zeitreihen unabhängig vom Niveau besser vergleichen zu können, werden die Reihen in allen folgenden Auswertungen standardisiert. Man sieht, dass die Prüfungszahlenkonjunktur alle Universitäten dieses Fachs gleichermaßen trifft, und weiter, dass diese Zyklen Dauern um die 20 Jahre haben.

Abbildung 3: Prüfungszahlen in Mathematik Lehramt an fünf verschiedenen Universitäten

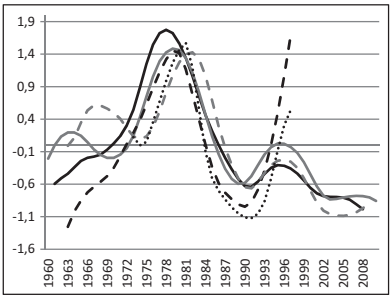


Abbildung 4: Prüfungszahlen in Germanistik Magister an sechs verschiedenen Universitäten

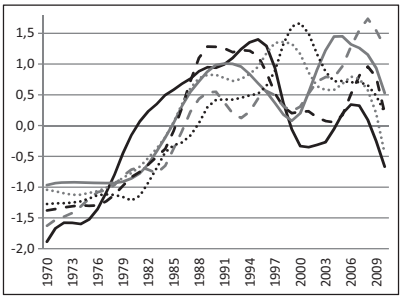


Abbildung 4 zeigt das Fach Germanistik Magister. Hier sind die Konjunkturen weniger stark und universitätsspezifisch. Die Länge der Zyklen der Prüfungszahlen findet sich jedoch in beiden Fällen in den zyklischen Bewegungen des Notenniveaus wieder. Welche Verbindungen könnten zwischen diesen beiden Zeitreihen bestehen?

Die Prüfungszahlen wirken erstens als Indikator für die Lehrbedingungen. Dabei repräsentieren sie zwei Einflüsse: größere Veranstaltungen, damit schlechtere Betreuung im Studium und daraus folgend schlechtere Prüfungsleistungen, und mehr Belastung der PrüferInnen – die KandidatInnen z.B. sind weniger bekannt, die Notenskala wird bei einer größeren Zahl mehr ausgereizt (Birkel 1978). Beide Einflüsse wirken gleich: die Noten sollten genau bei der Maximalzahl von Prüfungen am schlechtesten sein und umgekehrt. Zu diesem Zeitpunkt werden die geprüft, die die schlechtesten Lehrbedingungen (die meisten Mitstudierenden) hatten, und dann ist auch die Prüferbelastung am höchsten. Daraus müsste sich ein paralleler Verlauf von Prüfungszahlen und Noten ergeben.

Der zweite Einfluss der Prüfungszahlen ergibt sich aus ihrer Zyklizität. Prüfungszahlen heute sind (leicht reduzierte) Erstsemesterzahlen eine Studiendauer zuvor. Eine Entscheidung für ein Studienfach wird neben einer intrinsischen Motivation auch von der Arbeitsmarktlage beeinflusst. Von ihr ist bekannt, dass sie bei vielen akademischen Berufen in langen Zyklen zwischen Überfüllung und Mangel schwankt (Titze 1990; Müller-Benedict 2005). Die Erstsemesterzahlen sind also Indikator für die Arbeitsmarktsituation: Solange sie steigen, herrscht Mangel, wenn sie wieder sinken, Überfüllung. Es gibt allerdings Fächer, die keine ausgeprägten Konjunkturen aufweisen. So ist der Ingenieurs- oder Lehrermangel vieldiskutiertes öffentliches Thema, aber wann hat man jemals in der Zeitung gelesen: „Eklatanter Soziologenmangel befürchtet“?

Aber wie könnten die Berufsaussichten auf die Notengebung Einfluss nehmen? Hier gibt es zwei gegensätzliche Hypothesen. H1 lautet: Wenn in einem Fach Arbeitsmarktüberfüllung herrscht, wird „milder“ benotet. Daraus ergibt sich ein paralleler Verlauf der Erstsemester- und Notenzyklen: Solange jene sinken (Überfüllungsphase), sinken diese ebenfalls (werden besser), und umgekehrt. H2 lautet: Wenn in einem Fach Arbeitsmarktüberfüllung herrscht, wird strenger selektiert (schlechter benotet). Damit erhält man genau die gegensätzliche Bewegung: solange die Erstsemesterzahlen sinken, steigen die Noten (werden schlechter).

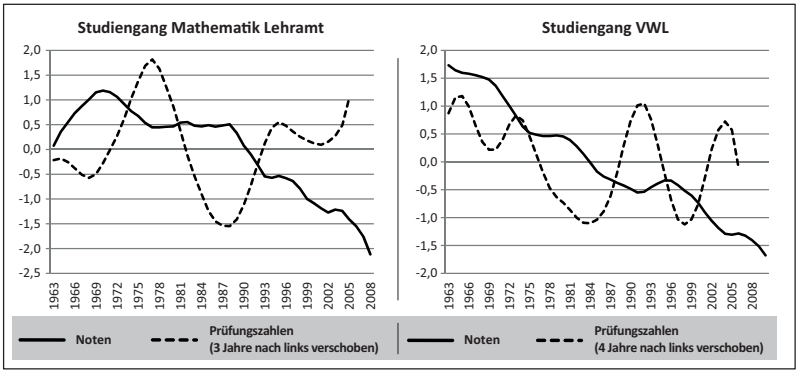
Für H1 gilt als Begründung, dass die Prüfer und Prüferinnen Milde walten lassen wollen, um für die schlechten Aussichten und die schlechten Lehrbedingungen zu trösten. (Hitpass/Trosien 1987:XI). Für H2 wird angenommen, dass wegen der Überfüllung ein strengeres Selektionsklima herrscht (Nath et al. 2004). Die Prüfungszahlenkonjunktur folgt nun der Erstsemesterkonjunktur im Abstand einer Studiendauer. Die Zeitreihe der Prüfungszahlen, wenn sie um eine Studiendauerlänge in die Vergangenheit verschoben (gelagt) wird, zeigt so gerade die Arbeitsmarktkonjunktu-

ren an: sinkt die verschobene Zeitreihe, herrscht Überfüllung, wächst sie, herrscht Mangel.

Welche der beiden Einflüsse und welche der beiden Hypothesen sind eher mit den Daten vereinbar? Die Antwort hängt von der Lag-Struktur und dem Vorzeichen der Beziehung zwischen Prüfungszahlen und Noten ab. Wenn die Selektionshypothese H2 gilt, müssten die gelagten Prüfungszahlen sich gegenläufig zu den Noten verändern (negatives Vorzeichen). Wenn die Mildehypothese H1 gilt, müsste es genau umgekehrt sein: die gelagten Prüfungszahlen bewegen sich parallel zu den Noten (positives Vorzeichen). Wenn nicht der Arbeitsmarkt, sondern die Lehrbedingungen die Noten verändern, müssten die Noten sich genau synchron zu den Prüfungszahlen verhalten (kein Lag, positives Vorzeichen).

Um die Verhältnisse zu überprüfen, wird eine Regression durchgeführt. Da es sich bei den Daten um Zeitreihen mit autokorrelierten Residuen handelt, wird hier eine Prais-Winsten-Regression durchgeführt und zusätzlich, wenn nötig, die Zeit als Variable berücksichtigt, um *grade inflation* zu neutralisieren. Diese Regression wird mit Verschiebungen der Prüfungszahl in die Vergangenheit um 0 bis 6 Jahre durchgeführt, und das Lag mit der höchsten Anpassung (Signifikanz, R-Quadrat) gewählt. Die Ergebnisse zeigen, dass eine Entscheidung zwischen den Einflüssen und den Hypothesen je nach Studiengang unterschiedlich ist. Beispielhaft werden hier zunächst Mathematik Lehramt und VWL gezeigt (Abb. 5; Tab. 2).

Abbildung 5: Noten und Prüfungszahlen im Studiengang Mathematik Lehramt und VWL



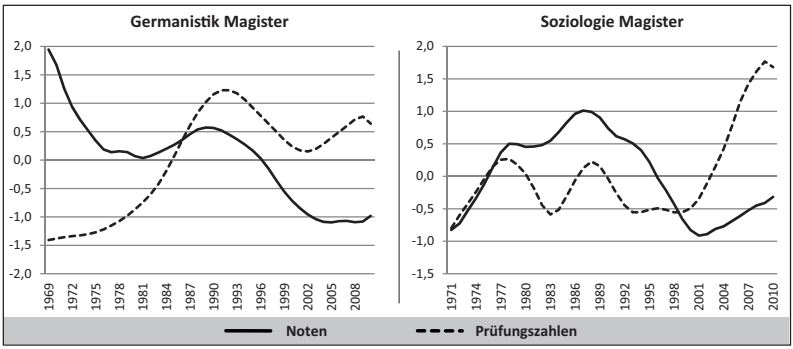
Hier gilt die Hypothese 2 (negatives Vorzeichen bei 3 Jahren (Mathematik) bzw. 4 Jahren (VWL) Verschiebung): Vom Minimum der verschobe-

Tabelle 2: P-W Regression Mathematik Lehramt und VWL

	Coef.	SE	P>|t
Note_MatheLA			
nMatheLA (Lead 3)	–0.208	0.062	0.002
time	–0.030	0.015	0.061
constant	+58.27	30.53	0.063
Note_VWL			
nVWL (Lead 4)	–0.135	0.030	0.000
time	–0.073	0.007	0.000
constant	+145.3	13.09	0.000

nen Prüfungszahlen beginnend (Mangelphase), verbessern sich (sinken) auch die Noten, und umgekehrt. Diese Situation kann für alle Studiengänge mit deutlichen Fachkonjunkturen bestätigt werden. Für Soziologie Magister und Germanistik Magister muss die Analyse wegen des Fehlens einer Fachkonjunktur für einzelne Universitäten durchgeführt werden (Abb. 6).

Abbildung 6: Noten und Prüfungszahlen im Studiengang Germanistik Magister und im Studiengang Soziologie Magister an der Universität Göttingen



Hier zeigt sich der Einfluss der Lehrbedingungen (positives Vorzeichen bei Germanistik und Soziologie): Die Noten schwingen parallel (bei Germanistik um 1 Jahr verschoben) zu den Prüfungszahlen. Für die Mildehypothese wurde dagegen gar keine Bestätigung gefunden.

Besonderheiten der Arbeitsmärkte und der Professionen, die für die starken Unterschiede zwischen Fächern bzw. Studiengängen verantwortlich sein könnten, ließen sich zahlreiche und widersprüchliche aufzählen (Whitley 1984; Becher 1989). Interessant ist die oben erwähnte Zweitei-

Tabelle 3: P-W Regression Germanistik Magister und Soziologie Magister

	Coef.	SE	P>|t
Note_GerMA			
nGerMA (Lead 1)	+0.580	0.125	0.000
time	–0.105	0.015	0.000
constant	+208.8	29.78	0.000
Note_SozMA			
nSozMA	+0.363	0.112	0.003
time	–0.014	0.019	0.481
constant	+26.43	37.93	0.490

lung in einerseits ihrem spezifischen Arbeitsmarkt „hart" unterworfene und andererseits eher lose mit einem diffusen Arbeitsmarkt gekoppelte Berufe. Dass Studierende der Fächer dieser beiden Kategorien auf die Arbeitsmarktlagen unterschiedlich reagieren, haben Reisz/Stock (2013) nachgewiesen. Lehramtstudiengänge gehören eher zur ersten, Germanistik und Soziologie als Magisterstudiengänge eher zur zweiten Gruppe.

Wenn auch noch unklar ist, wie die Arbeitsmarktlage oder ein überfüllter Studiengang auf die konkrete Benotungspraxis der Prüfenden einwirken, wurde der Zusammenhang zwischen Prüfungszahlen und Prüfungserfolg auch schon früher sowohl auf der Grundlage historischer Daten (Titze 1990; Müller-Benedict 2005) als auch bei der ersten langfristigen Analyse der bundesdeutschen Hochschulnoten durch Hitpass/Trosien (1987) festgestellt. Mit diesen Analysen wird der Zusammenhang wieder bestätigt, allerdings zeigt er sich fachspezifisch. Im Falle der Arbeitsmarktkonjunktur ist der Einfluss außerhalb der Hochschulen selbst angesiedelt und wirkt trotzdem auf die Prüfungen. Weitere externe systematische Einflüsse sind denkbar, z.B. die regionale Arbeitsmarktlage (s. Grözinger in diesem Heft).

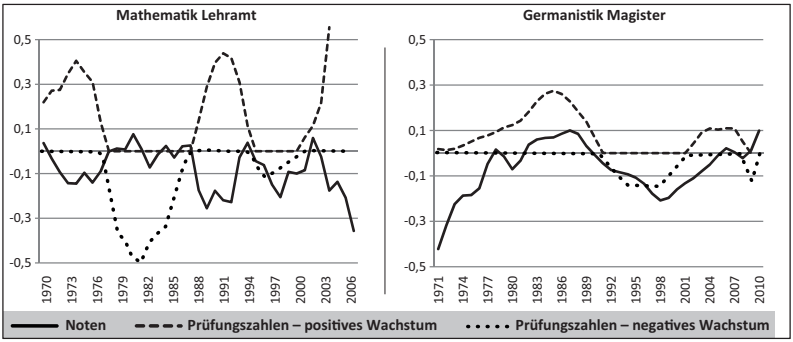
Zyklische Veränderungen der Noten sind nicht durch Leistungsänderungen erklärbar. Wenn Zyklen bei der Notenentwicklung vermieden werden sollen, muss auch über solche möglicherweise bisher unbewussten Einwirkungen bei der Benotungspraxis gesprochen werden. Bis jetzt ist es eher ein Tabu, davon auszugehen, dass Berufsaussichten Noten beeinflussen könnten. Aber vielleicht hat der eine Prüfer oder die andere Prüferin schon einmal bei sich gedacht: „bei derart schlechten Berufaussichten sollte man die Besten doch deutlich von den Schlechten separieren, weil sie eher eine Chance verdient haben"? Darüber sollte mehr und systematisch gesprochen werden.

4. Eine mögliche Ursache für *grade inflation*

Aus dem Vorhergehenden folgt noch nicht, dass die Noten immer besser werden – mit zyklischen Bewegungen in den Prüfungszahlen sollten sich eigentlich auch die Noten zyklisch bewegen. Wie kann es dennoch zu *grade inflation* kommen?

Um diese Frage zu beantworten wurde eine besondere Berechnung durchgeführt, die die Reaktion auf die Veränderung der Prüfungszahlen in zwei Komponenten zerlegt: den Einfluss von Erhöhung und den Einfluss von Verringerung. Weil es um eine Beziehung zwischen zwei Veränderungen geht, wird mit den ersten Differenzen der Zeitreihen gerechnet. Die Veränderung der nach den Ergebnissen aus dem vorigen Abschnitt gelagten Prüfungszahl wurde in zwei Variablen, positives (steigende Zahlen) bzw. negatives (fallende Zahlen) Wachstum, aufgespalten[4] und eine Regression der Veränderung der Noten auf beide berechnet. Das Ergebnis für Mathematik Lehramt und Germanistik Magister mit den dazu gehörigen Zeitreihen zeigt die Abbildung 7 sowie die Tabelle 4.

Abbildung 7: Wachstum der Noten und Prüfungszahlen im Studiengang Mathematik Lehramt und in Germanistik Magister an der Universität Göttingen



In Mathematik Lehramt werden in der Mangelphase (wenn die – verschobenen – Prüfungszahlen wachsen) die Noten um den Betrag 0,377 pro Standardabweichung besser (geringer), in der Überfüllungsphase dagegen bleiben sie gleich (0,023 ist nicht signifikant von 0 verschieden). D.h. die Noten sind unterschiedlich elastisch in Bezug auf die Phasen: Auf dieselbe Veränderung der Prüfungsanzahl verändern sich die Noten nur in Rich-

[4] Für die Phasen des jeweils anderen Wachstums wurden die Variablen auf 0 gesetzt. Simulationsrechnungen zeigen, dass die mit diesem Verfahren gewonnenen zwei Koeffizienten gemittelt genau den Wert des Koeffizienten der nicht aufgespaltenen Variable ergeben.

Tabelle 4: OLS Regression Mathematik Lehramt und Germanistik Magister

	Coef.	SE	P>|t
D.Note_MatheLA			
posWachstum	–0.377	0.057	0.000
negWachstum	+0.023	0.067	0.731
D.Note_GerMA			
posWachstum	+0.0977	0.138	0.482
negWachstum	+1.006	0.261	0.000

tung „besser". Dadurch wird das Niveau nach jedem Zyklus ein wenig besser und über die ganze Zeitspanne hinweg ergibt sich *grade inflation*. Für Germanistik Magister (Uni Göttingen) ergibt sich dieselbe unterschiedliche Elastizität, obwohl hier die Noten ja anders mit der Prüfungsanzahl gekoppelt sind. Wenn die Prüfungszahlen steigen, stagnieren die Noten (Koeffizient 0,0977 ebenfalls nicht signifikant), in der Überfüllung jedoch verbessen sich (sinken) die Noten um 1,006. Also ist auch bei einer anderen Art des Zusammenhangs der Prüfungszahlen mit den Noten die Reaktion in der Phase der Verbesserung der Noten stärker als in der Phase der Verschlechterung. Demnach besteht auch hier die ständige Tendenz zur Verbesserung.

Dieses Ergebnis könnte eine mögliche Ursache für den langfristigen Trend zur *grade inflation* sein. In den Verbesserungsphasen verbessern sich die Noten stärker, als sie sich in den Verschlechterungsphasen verschlechtert haben. Dies ist möglicherweise auch die Ursache dafür, dass das einmal nach einer Mangelphase um ca. 1980 erreichte sehr gute Notenniveau in Biologie und Psychologie sich seitdem konstant hält und nicht wieder hebt. Und es ist auch eine Anregung, über ein weiteres Tabu zu sprechen: versucht nicht jede Dozentin und jeder Dozent, eine Verschlechterung seiner Klausur- oder Seminarergebnisse eher zu vermeiden als eine Verbesserung?

5. Fazit

Auch wenn der Wissenschaftsrat bereits 2003 Ungleichheiten bei der Notengebung und auffallend gute Noten einiger Fächer benannte, blieb das Thema in der Bildungspolitik und an den Hochschulen bis heute ein Tabu. Die Analyse der langfristigen Notenentwicklung zeigt demgegenüber folgende Entwicklungen auf, die im Hinblick auf die Aussagekraft von Noten bedenklich sind:

1. Es gibt langfristige stable Notenunterschiede zwischen den Studiengängen und in den meisten von ihnen *grade inflation*.
2. Es besteht ein Zusammenhang zwischen der Entwicklung der Prüfungszahlen und der Notenentwicklung. Dieser äußert sich fachspezifisch: Existiert eine durch die Arbeitsmarktenwicklung determinierte einheitliche Fachkonjunktur in den Studierendenzahlen, lässt sich hochschulübergreifend zeigen, dass sich die Noten in Mangelphasen verbessern, in Überfüllungsphasen verschlechtern. Dieser Einfluss lässt sich in Magisterstudiengängen, die nicht an ein klares Berufsfeld gebunden sind, nicht nachweisen. Allerdings gibt es hier einen hochschulspezifischen Effekt der Prüfungszahlen: steigen diese an, werden die Noten schlechter, sinken sie, werden die Noten besser, was auf entsprechende Schwankungen der Qualität der Lehrbedingungen zurückzuführen sein könnte.
3. Unabhängig davon, welcher Zusammenhang gilt, lässt sich weiter aufzeigen: In Phasen der Notenverbesserung verbessern sich die Noten stärker, als sie sich in Zeiten der Notenverschlechterung verschlechtern. Die dadurch erzeugte *grade inflation* hat in einigen Fächern dazu geführt, dass mehrheitlich nur noch sehr gute Noten vergeben werden. Wie auch immer diese Phänomene zustande kommen: es sind Phänomene, über die geredet werden muss. Denn sollen Noten vergleichbar sein und ihrer zentralen Aufgabe, der Abbildung von Leistung, gerecht werden, dürfen sie nicht in Zyklen schwanken und nicht bis zum unteren Ende der Notenskala sinken. Auch wenn auf diese Weise vielleicht einige unangehme Wahrheiten über die Notengebung ans Licht kommen: Nur wenn die Diskussion über die Praxis der Notengebung enttabuisiert wird, kann die Notengebung von leistungsunabhängigen Einflüssen entkoppelt werden.

Literatur

Becher, Tony (1989): Academic Tribes and Territories. Intellectual Enquiry and the Cultures of Disciplines. Milton Keynes: Open University Press.

Birkel, Peter (1978): Mündliche Prüfungen. Zur Objektivität und Validität der Leistungsbeurteilung. Bochum: Kamp.

Gaens, Thomas (2013): Von einem, der auszog, einen Leistungsindikator zu erheben. Durchfallquoten und die Problematik ihrer Bildung. *Das Hochschulwesen* Vol. 6, S. 200–206.

Hitpass, Josef/Jürgen Trosien (1987): Leistungsbeurteilung in Hochschulabschlussprüfungen innerhalb von drei Jahrzehnten – Wandel von Prüfungsergebnis und Prüfungserlebnis an deutschen Universitäten. Bad Honnef: Bock.

Hu, Shouping (2005): Beyond Grade Inflation. Grading Problems in Higher Education. San Francisco: Jossey-Bas.

Lehnen, Alfred/Werner Loch (1978): Objektivierte Leistungsmessung durch Test-Diktate. Limburg: Frankonius.
Kuh, George D. (2003): What We're Learning about Student Engagement From NSSE. *Change* Vol. 35 (2), S. 24–32.
Müller-Benedict, Volker (2005): Sind Universitätsprüfungen objektiv? Der langfristige historische Zusammenhang zwischen Erfolg in akademischen Prüfungen und Überfüllung der akademischen Berufe. *Soziologie* Vol. 34, S. 191–208.
Müller-Benedict, Volker/Elena Tsarouha (2011): Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen. *Zeitschrift für Soziologie* Vol. 40, S. 288–309.
Nath, Axel/Corinna Dartenne/Carina Oelerich (2004): Der historische Pygmalioneffekt der Lehrergenerationen im Bildungswachstum von 1884 bis 1993. *Zeitschrift für Pädagogik* Vol. 50, S. 539–564.
Neumann, Marko/Gabriel Nagy/Ulrich Trautwein/Oliver Lüdtke (2009): Vergleichbarkeit von Abiturleistungen. Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen. *Zeitschrift für Erziehungswissenschaft* Vol. 12, S. 691–714.
Reisz, Robert, D./Manfred Stock (2013): Hochschulexpansion, Wandel der Fächerproportionen und Akademikerarbeitslosigkeit in Deutschland. *Zeitschrift für Erziehungswissenschaft* Vol. 16 (1), S. 137–156.
Rheinberg, Falko (2002): Bezugsnormen und schulische Leistungsbewertung, in: Franz E. Weinert (Hg.), Leistungsmessungen in Schulen. Weinheim: Beltz, S. 59-71.
Statistisches Landesamt Schleswig-Holstein, Forschungsdatenzentrum (2012): Hochschulprüfungsstatistik 1995–2010.
Titze, Hartmut (1990): Der Akademikerzyklus: Historische Untersuchungen über die Wiederkehr von Überfüllung und Mangel in akademischen Karrieren. Göttingen: Vandenhoeck & Ruprecht.
Verwaltungsgericht Münster, Beschluss vom 15. November 2010 – 9 L 529/10.
Whitley, Richard (1984): The intellectual and social organization of the sciences. Oxford: Clarendon.
Wissenschaftsrat (2003, 2007, 2012): Prüfungsnoten an Hochschulen 1996, 1998 und 2000 (bzw. 2005, 2010) – Arbeitsbericht. Hrsg. v. d. Geschäftsstelle des Wissenschaftsrats, Drucksachen 5536–03 (2003), 7769-07 (2007), 2627-12 (2012).