Soziologie, 46. Jg., Heft 4, 2017, S. 429–452

Examensnoten

Langfristiger Verlauf und Vergleich unter Berücksichtigung des Faches Soziologie

Volker Müller-Benedict

Seit den 1970er Jahren werden die Examensnoten an deutschen Hochschulen in vielen Fächern immer besser. Dabei verläuft die Notenentwicklung in langfristigen Zyklen. Es gibt beträchtliche Unterschiede im Notenniveau desselben Fachs an verschiedenen Universitäten, für die gängige Erklärungsmuster wie die veränderte soziale Zusammensetzung der Studierenden, zum Beispiel in Bezug auf den Frauen-, Ausländer- oder Altersanteil zu kurz greifen. Zu diesen Ergebnissen kommt das von Gerd Grözinger und mir beantragte und durchgeführte DFG-Projekt, das die Notenentwicklung und mögliche Einflüsse darauf untersucht hat.[1] Der Wissenschaftsrat hatte diese Trends schon in drei Untersuchungen (2003; 2007; 2012) punktuell festgestellt. Auch in den USA läuft unter dem Stichwort *grade inflation* seit den 1990er Jahren eine kontroverse Diskussion, die unter anderem dazu führte, dass an einigen Universitäten wie zum Beispiel Harvard der Anteil der Bestnoten beschränkt wurde. Im Folgenden stelle ich die deutschen Ergebnisse zusammengefasst dar, wobei ich einen differenzierten Blick auf die Soziologie werfe.

Unterschiedliche Examensnoten sind natürlich kein Problem, wenn sie auf den Leistungen der Kandidaten beruhen. Deshalb liegt der Fokus da-

1 »Die Notengebung an Hochschulen in Deutschland von den 1960er Jahren bis heute. Trends, Unterschiede, Ursachen« (2012–2015). Für genauere Ausführungen vgl. Müller-Benedict, Grözinger (2017). Dieser Beitrag bezieht sich vor allem auf Gaens, Müller-Benedict (2017).

rauf, Unterschiede im durchschnittlichen Notenniveau zu untersuchen, die sich aus anderen Gründen ergeben. Erstere werden hier »leistungskonform« genannt, die nicht auf Leistung beruhenden Unterschiede »leistungsunabhängig«. Das Notenspektrum hängt entscheidend von den Prüfungsbedingungen und den angewendeten Testverfahren ab. Für diese Abhängigkeit gibt es Normen, die als Basis-Kriterium für die Beurteilung »abweichender« Notenentwicklung gelten können und die ich deshalb als Erstes kurz darstelle.

1. Wie sollten Examensnoten verteilt sein?

Mit Prüfungen und Tests befasst sich seit langem die Testpsychologie. Mit ihren Methoden kann man zum Beispiel nachprüfen, ob ein Test die geforderten Gütekriterien, insbesondere Reliabilität und Validität, erfüllt. Auch die Konstanz der Prüfungsbedingungen, allen voran die Konstanz der Prüfungspraxis der Prüfer und die der Vorbereitung der zu Prüfenden, ist notwendig für eine Vergleichbarkeit. Da es für HochschullehrerInnen keine spezielle Prüferausbildung gibt, besteht im Hinblick auf diese testtheoretischen Bedingungen im deutschen Hochschulalltag keine Stabilität, sondern eine große Varianz. Für die folgenden Ergebnisse, die auf Durchschnittsbildungen über sehr viele Prüfungen beruhen, wird wie in der sozialstatistischen Methodik üblich angenommen, dass sich dabei die voneinander unabhängigen individuellen Eigenschaften der Beteiligten und die lokalen Bedingungen ausgleichen. Unter dieser Annahme müssen Unterschiede zwischen den Aggregaten als überindividuelle, im weitesten Sinne soziale Einflüsse begriffen werden.

Notenniveaus sind im Folgenden Durchschnittswerte von Noten, die über verschiedene Aggregationsniveaus gebildet werden, zum Beispiel für Fächer über alle Hochschulen, für Hochschultypen, für Institute, für Prüfungsformen etc. Nur schwerlich würde man auf die Idee kommen, eine Universitätsgesamtnote zu bilden und damit Universitäten zu vergleichen, weil Universitäten zum Beispiel nicht unbedingt dieselben Fakultäten haben. Dasselbe gilt auch für Fakultäten: Sie sind nicht immer aus denselben Fächern zusammengesetzt. Diese Argumentation lässt sich fortsetzen für immer kleinere Einheiten bis hin zu der Frage, ob sich die Durchschnittsnote im Seminar gleichen Inhalts aus dem vorigen Semester mit der Note

des aktuellen Semesters vergleichen lässt. Für absolute Vergleichbarkeit müssten die Bedingungen für jede Prüfung dieselben sein. Da dies nicht der Fall ist, kann das Aggregationsniveau nicht durch die Frage bestimmt werden, was absolut vergleichbar ist. Vielmehr sind die Art und das Niveau der Aggregation seitens der Fragestellung normativ bestimmt: Wenn man fragt »Unterscheiden sich die Fächer in der Durchschnittsnote?« unterstellt man einen Vergleichsmaßstab für die Fächer. Mit der obigen Annahme des Ausgleichs über Durchschnittsbildung bezüglich anderer Merkmale lassen sich Unterschiede im Notenniveau dann mit Eigenschaften der Fächer – bzw. des jeweils in Frage stehenden Aggregats – in Verbindung bringen.

Ein Problem der Analyse leistungsunabhängiger Unterschiede ist die Abgrenzung zu Unterschieden, die tatsächlich auf Leistung beruhen. Von jeder Änderung oder jedem Unterschied in der Durchschnittsnote lässt sich behaupten, dass er durch die veränderten Leistungen der Examinierten zustande gekommen sei. Das lässt sich oft nur indirekt widerlegen, indem man zum Beispiel Sprünge im Niveau nach der Änderung einer Prüfungsordnung beobachtet oder längerfristige Notenzyklen feststellt. Solche Entwicklungen der Noten sind nicht mit dem vereinbar, was Examensnoten darstellen sollen.

Weil Noten eine Bewertung zum Ausdruck bringen, muss es einen Maßstab geben. Theoretisch wird zwischen drei Bezugsnormen für Noten unterschieden: der individuellen (Bewertung der individuellen Verbesserung), der sozialen (Bewertung im Vergleich zur Bezugsgruppe, zum Beispiel Klasse, Seminar) und der absoluten (Bewertung anhand eines geprüften Wissens- oder Kompetenzkanons). An den Hochschulen sollte für die Abschlussnoten die absolute Bezugsnorm im Vordergrund stehen, weil diese Noten den relativen Wissensstand des Absolventen in Bezug auf den aktuellen akademischen Wissensbestand signalisieren sollen. Da das akademische Wissen sich allerdings ständig weiterentwickelt, kann die absolute Bezugsnorm für den intertemporalen Vergleich nicht gelten – eine Einser-Leistung in Chemie 1930 würde heute vermutlich nicht einmal ein »ausreichend« erreichen. Die Notenskala gilt also je Zeitpunkt relativ zum aktuellen Wissen.

Eine Prüfung sollte sowohl schwierige als auch einfache und mittlere Aufgaben in einer gleichmäßigen Häufigkeit aufweisen, sonst gilt sie als »zu leicht« oder »zu schwer« in Bezug auf die jeweilige Bezugsnorm. In Schulen zum Beispiel gibt es Materialien für Tests, die die Fehlerpunkte für die Grenzen zwischen den Noten so festsetzen, dass es nicht zu viele »sehr gut« und »ausreichend« gibt, und die Mehrheit ein »gut« oder »befriedi-

gend« erhält. Andere Verteilungen gelten als didaktisch problematisch bzw. falsch konstruiert. Heute wird eine gleichmäßige Verteilung der zu testenden Leistungen in Bezug auf die absolute Bezugsnorm durch die genaue Definition aufsteigend höherer Kompetenzniveaus erreicht, wie sie zum Beispiel in den PISA-Unterlagen beschrieben werden. Aus diesen theoretischen Überlegungen heraus ist für Abschlussexamen an Hochschulen eine Streuung der Noten über die ganze Skala und eine etwa gleichbleibende Streubreite der Noten über die Zeit wünschenswert. Dies ist jedoch faktisch nicht der Fall, wie zu zeigen sein wird.

Neben der Funktion, die Leistung des Kandidaten zu repräsentieren, haben Noten bisweilen auch die Funktion, das Niveau des Abschlusses zu signalisieren. Jedem ist klar, dass eine 1 im Hauptschulexamen eine andere Leistung darstellt als eine 1 im Abitur; für die Schularten ist das bekannt. Die Promotion verdeutlicht dies schon seit jeher durch eigene Bezeichnungen von »summa cum laude« bis »rite«. Was aber ist mit Staatsexamen (zum Beispiel für Lehramt) und Diplom im selben Fach? Hier gibt es ein ungeschriebenes Einverständnis,[2] dass zum Beispiel die Mathematik-Diplomnoten besser ausfallen sollten als die Staatsexamensnoten. Und was ist mit Bachelor- und Masternoten? Es gibt Länder, in denen diesbezüglich festgelegt wird, dass im Master andere Bereiche der Notenskala gelten als für den Bachelor, zum Beispiel bedeutet in Mexiko mit einer Skala von 1 bis 100 eine Note unter 60 im Master »durchgefallen«, in der *Licenciatura* (dem Bachelor) dagegen nicht. An unseren Universitäten hört man oft, dass die Masternoten besser sein müssten als die Bachelornoten – das wird sich auch in der folgenden Analyse zeigen. In Fällen wie diesen wird die absolute Bezugsnorm vermischt mit einer nur vermuteten, in keiner Weise fixierten Abschlusshierarchienorm, was testtheoretisch völlig haltlos ist. Dadurch wird zum Beispiel unter der Hand festgelegt, dass die Verteilung der BA-Noten rechtssteiler zu sein habe als die der MA-Noten.

2. Der langfristige Verlauf der Examensnoten

Da die Noten vor 1997 aus einzelnen Universitätsarchiven erhoben werden mussten (Gaens 2013), wurden forschungspragmatisch sieben Hochschulen und zwölf Studiengänge ausgewählt, für die die Zeitreihen nach 1997

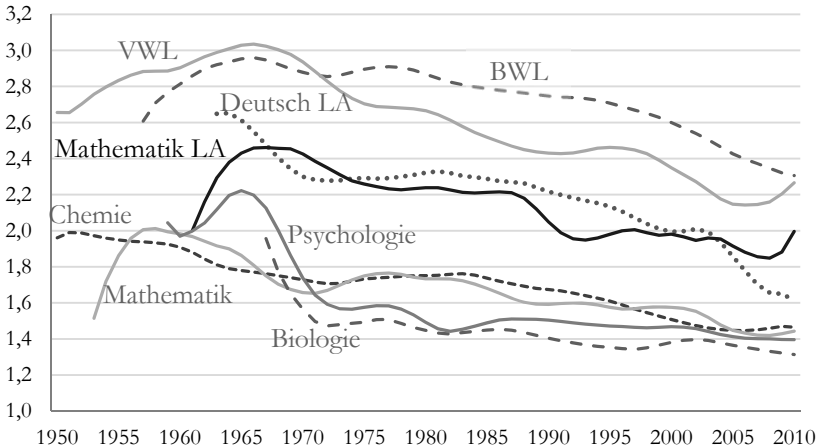
2 Siehe dazu auch die Ergebnisse weiter unten.

mit der Prüfungsdatenstatistik des Statistischen Bundesamts, ausgewertet im Forschungsdatenzentrum Kiel, fortgesetzt wurden.

2.1 Fächer mit *grade inflation*

Die Daten zeigen in acht der zwölf berücksichtigten Studiengänge eine nennenswerte Verbesserung der Notendurchschnitte im Zeitverlauf. Abbildung 1 präsentiert die Noten[3] für diese acht Fächer.

Abbildung 1: Verlauf der Abschlussnoten in Studiengängen mit langfristiger Notenverbesserung



Die Notenverbesserung setzt jeweils zu Beginn/Mitte der 1960er Jahre ein, was den Ergebnissen von Hitpass und Trosien (1987) entspricht. Die Verbesserungsprozesse vollziehen sich allerdings in unterschiedlichem Ausmaß, in Chemie zum Beispiel Verbesserung um ca. eine halbe Note seit 1960, in Deutsch Lehramt um mehr als eine ganze Note seit 1963.[4] Die durchschnittliche Abschlussnote in BWL ist 2010 trotz langfristiger Verbesserung immer noch signifikant schlechter als das Notenniveau in Chemie 1960 und in Biologie 1967. In Biologie und Psychologie kann man die Notenlage spätestens seit Beginn der 1970er Jahre als derart gut einstufen,

3 Diese und alle folgenden Abbildungen sind mit LOWESS (Bandbreite 0.2–0.4) geglättet.
4 Alle hier und im Folgenden benannten Differenzen sind statistisch signifikant.

dass die Leistungsdifferenzierung dort zwangsläufig durch eine Häufung der Noten im Bestbereich gefährdet ist. In Psychologie wurden 54,6% der 11.467 zwischen 1971 und 1997 bestandenen Prüfungen mit »sehr gut«, 95,5% mit »sehr gut« oder »gut« bewertet. In Biologie liegen diese Anteile im gleichen Zeitraum bei 62,8% bzw. 96,5% (n=11.611). Abbildung 2 verdeutlicht, dass die Streuung der Noten mit der Zeit dort ebenfalls sinkt. Auch in allen anderen Studiengängen mit sinkenden Notendurchschnitten nimmt die Streuung parallel ab, wie Abbildung 3 am Beispiel BWL zeigt. Die Verbesserungen gehen also unabhängig von der Begrenzung des Notenspektrums mit einer Verringerung der Streuung der Noten einher, die auf Englisch mit *grade compression* bezeichnet wird.

*Abbildung 2: Biologie – Abschlussnoten vs. Standardabweichungen*3*

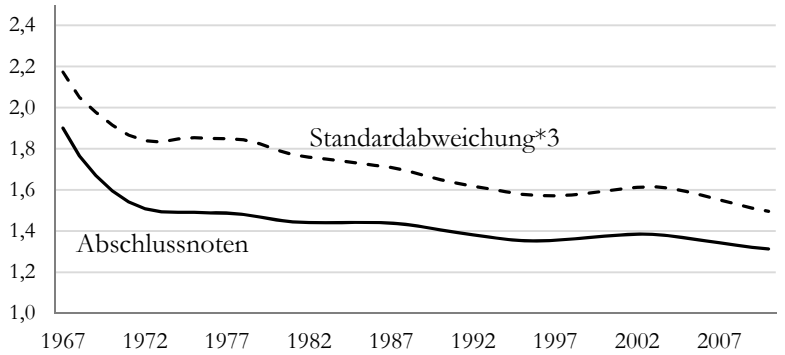
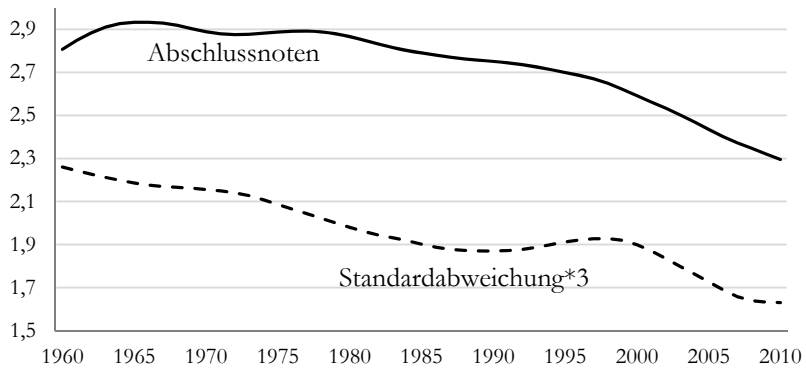


*Abbildung 3: BWL – Abschlussnoten vs. Standardabweichungen*3*



Gemeinsam ist allen Studiengängen, dass die Verbesserung nicht linear verläuft. Die Abwärtsbewegung wird von zyklischen Schwankungen begleitet, deren Aufwärtsbewegungen Plateauphasen erzeugen. Diese sorgen für zeitweilig stabile Notenniveaus. Die eigentliche Verbesserung im Zeitverlauf vollzieht sich damit in bestimmten Phasen von unterschiedlicher Länge.

2.2 Fächer ohne Notenverbesserung

Im Diplomstudiengang Maschinenbau, in den Magisterstudiengängen Soziologie und Germanistik sowie im ersten Staatsexamen der Rechtswissenschaften kann keine langfristige Verbesserung des Notenniveaus festgestellt werden (Abbildung 4). Während die Noten der ersten drei Studiengänge zyklisch verlaufen, bewegen sie sich in Jura über den gesamten Zeitverlauf im Rahmen einer maximalen Spannweite, so dass das Notenniveau als über den Zeitverlauf konstant eingestuft werden kann. Was könnte besonders sein an Jura? Faktische Voraussetzung für einen Eintritt in den Staatsdienst als Richter oder Staatsanwalt ist seit jeher eine über dem Durchschnitt von »voll befriedigend« (= 2,5) liegende gute Note, die nur ca. 15% aller Examenskandidaten erreichen. Die Noten werden jedes Jahr von den Justizministern der deutschen Länder veröffentlicht. Dadurch wird die Zahl potentieller Richter indirekt gesteuert. In Germanistik (-0,28) und Soziologie (+0,10) ist die Differenz zwischen Beginn und Ende der Zeitreihen nicht signifikant.[5] Die Streuung der Noten nimmt in den vier Studiengängen ohne Verbesserung im Zeitverlauf nicht ab (Abbildung 5). Dies belegt, dass eine sinkende Streuung tatsächlich in Verbindung mit der Verbesserung im Zeitverlauf zu sehen ist und keine generelle Tendenz der Notengebung darstellt.

5 Im Studiengang Maschinenbau liegen nur Noten von zwei Hochschulen vor. In Soziologie und Germanistik sind zu Beginn der Zeitreihen zwei bzw. sechs Datenpunkte mit geringen Fallzahlen ($n \leq 13$) entfernt worden.

Abbildung 4: Verlauf der Abschlussnoten in Studiengängen ohne langfristige Notenverbesserung

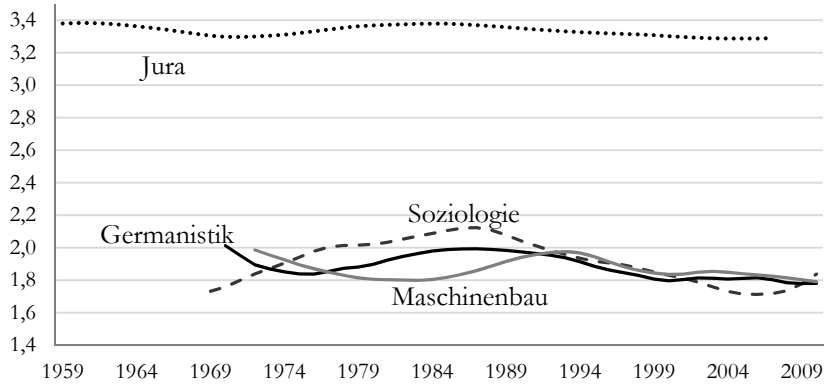
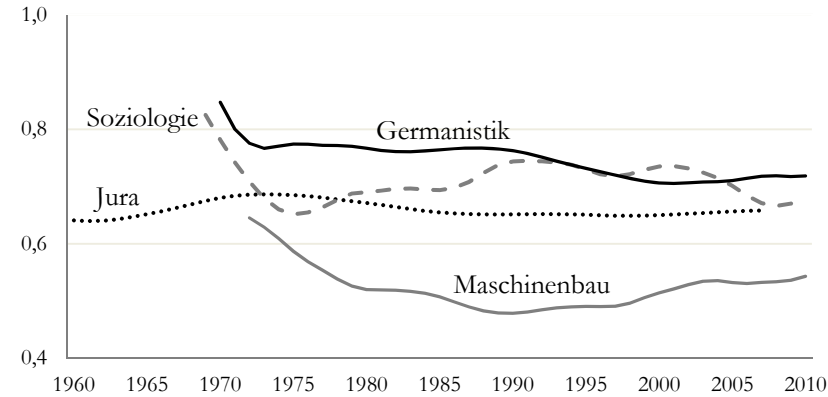


Abbildung 5: Standardabweichungen in den vier Studiengängen ohne Verbesserung



Damit gestaltet sich das Phänomen der *grade inflation* für die Studiengänge in verschiedener Form. Gemeinsam sind allen Fächern zyklische Notenschwankungen.

3. Langfristige Unterschiede zwischen Hochschulen in demselben Fach

Da die Noten auf Studiengangebene als Mittel aller Prüflinge berechnet wurden, ist der jeweilige Verlauf von den Noten an den Hochschulen mit den größeren Anteilen am gesamten Prüfungsvolumen abhängig. Betrachtet man die Noten desselben Studiengangs an verschiedenen Hochschulen, wird deutlich, dass die langfristige Notenentwicklung nicht nur studiengang-, sondern auch hochschulspezifisch verläuft. In Bezug auf das Fach Soziologie gibt es in der Datenbank des Statistischen Bundesamts die Abschlüsse »Diplom« und »Magister« und die Fachbezeichnungen »Soziologie« und »Sozialwissenschaften«. Bis zur Einführung von Bachelor und Master waren die häufigsten Studiengangs-Abschluss-Kombinationen »Diplom Sozialwissenschaften« und »Magister Soziologie«. Der Studiengang »Diplom Sozialwissenschaften« ist im Allgemeinen stärker mit Nebenfächern wie Jura und VWL verknüpft und auf betriebliche Praxis ausgerichtet, während der »Magister Soziologie« eher die akademische Soziologie widerspiegelt. An der Mehrzahl der Hochschulen war im Zeitraum 1997 bis 2010 der »Magister Soziologie« der weit häufigere oder der einzige Abschluss,[6] in anderen das Diplom in Sozialwissenschaften.[7] Nur im Studiengang Soziologie gibt es eine Reihe von Universitäten, an denen Diplom und Magister vergleichbar häufig studiert wurde.[8]

Abbildung 6 zeigt die Notenentwicklung der Studiengänge »Magister Soziologie« in Göttingen, Tübingen, Heidelberg und Münster, sowie »Diplom Soziologie« an der FU Berlin.[9] Hier sind nun drei Arten von Unterschieden in den Noten zu beobachten (3.1 –3.3):

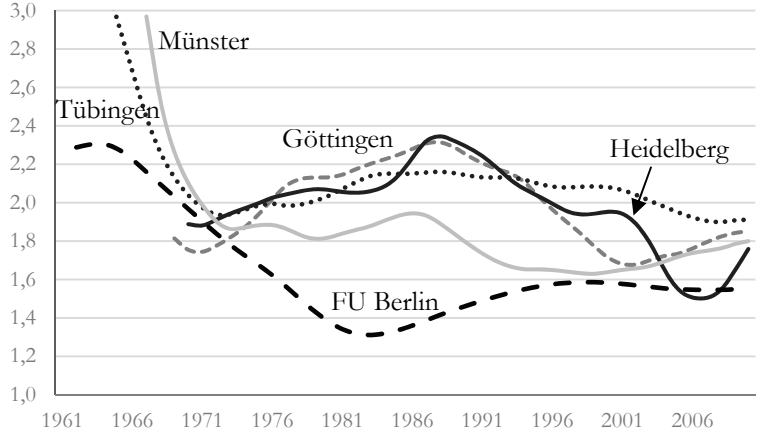
6 Universitäten Augsburg, Bamberg, Bielefeld, Bonn, Chemnitz, Frankfurt am Main, Freiburg, Jena, Konstanz, München, Münster sowie Technische Hochschule Aachen Freie Universität Berlin, Technische Universität Dresden, Universität Gesamthochschule Duisburg (und andere mit nur wenigen Prüfungen pro Semester).

7 Universitäten Bochum, Düsseldorf, Erlangen-Nürnberg, Göttingen, Hannover, Mannheim, Oldenburg, Osnabrück, Wuppertal sowie Humboldt Universität zu Berlin, Universität zu Köln (und andere mit nur wenigen Prüfungen pro Semester).

8 Universitäten Bremen, Frankfurt am Main, Hamburg, Heidelberg, Leipzig, Mainz, Marburg, Potsdam, Trier sowie Technische Universität Berlin.

9 Die Anzahl Prüfungen pro Semester bewegen sich im niedrigen Zehnerbereich für die Magister und um die Hundert für das Diplom an der FU Berlin. Maxima: Göttingen 30 (2009), FU Berlin 172 (1978), Tübingen 25 (1994), Heidelberg 43 (2007), Münster 65 (2003). Die Noten vor 1997 wurden für diese Studiengänge aus den jeweiligen Universitätsarchiven erhoben.

Abbildung 6: Notenentwicklung in »Magister Soziologie« bzw. »Diplom Soziologie«



3.1 Die über mittlere oder auch längere Fristen gemittelten Unterschiede im Durchschnittsniveau zwischen Universitäten

Über die Jahre 1969 bis 2010 gemittelt wird zum Beispiel an der FU Berlin durchschnittlich eine um 0,54 bessere Note erreicht als in Tübingen, eine um ca. 0,23 bessere als in Münster und eine um ca. 0,44 bessere als in den anderen drei Universitäten, die einen eher einheitlichen Verlauf aufwiesen.[10] Dass die Unterschiede im Notenniveau zwischen Hochschulen im gleichen Studiengang sogar bis zu einer ganzen Note reichen können, gilt auch für alle anderen Fächer. Das ist eklatant, denn damit könnte eine Studentin oder ein Student schon durch die Wahl der »richtigen« Universität ihre erwartete Abschlussnote entsprechend steigern. Dass es ein für die jeweilige Hochschule spezifisches Notenniveau gibt, zeigt sich schlüssig darin, dass in den Universitäten, in denen Soziologie sowohl mit Abschluss Diplom als auch mit Abschluss Magister studiert werden konnte (FN 8), der Notendurchschnitt in beiden Examensarten innerhalb derselben Universität immer gleich hoch ist,[11] während er zwischen den Universitäten unterschiedlich ist (siehe auch nächsten Punkt).

10 paired t-tests, n=41

11 Für den Zeitraum 1996 bis 2007; Ausnahme Universität Trier, dort ist das Diplomniveau ca. 0,4 schlechter.

3.2. Die Unterschiede im Durchschnittsniveau zwischen den verschiedenen Abschlussarten

Abbildung 6 deutet zusätzlich an, dass es im langfristigen Durchschnitt Unterschiede zwischen den verschiedenen Abschlussarten geben könnte, da der Berliner Studiengang sich auch durch den Abschluss »Diplom« unterscheidet. Dabei muss allerdings berücksichtig werden, dass das Diplom mehrheitlich im – praxisnäheren – Studiengang »Sozialwissenschaften« vorkommt; Berlin ist eine Ausnahme. Dass zwischen den Abschlüssen im Durchschnitt Unterschiede bestehen, zeigt die folgende Berechnung:

Tabelle 1: Notendurchschnitte 1996 – 2012, alle Hochschulen

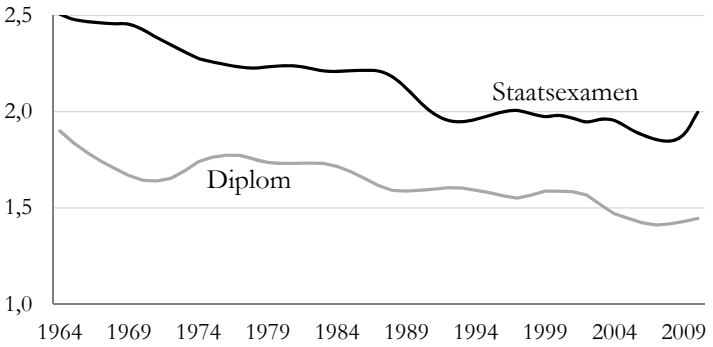
	Magister (bis 2006)	BA	MA	Diplom (bis 2008)
Sozialwissenschaften				
Notendurchschnitte	1,64	1,87	1,61	1,92
nachrichtlich: Studierende (gesamt 16.697)	*582*	*8.308*	*1.987*	*5.820*
Soziologie				
Notendurchschnitte	1,81	2,02	1,56	1,85
nachrichtlich: Studierende (gesamt 20.036)	*6.173*	*5.199*	*1.550*	*7.114*

Im Studiengang »Sozialwissenschaften« wurden im Diplom schlechtere Noten erzielt als im Magister. Interessant ist die Umstellung auf das BA/MA-System: Sie wurde erstens dazu genutzt, eine Hierarchie herzustellen: BA ist weniger wert als MA; zweitens dazu, im MA bessere Noten zu vergeben als im früheren Magister; und drittens scheinen sich nun die Unterschiede zwischen den beiden Studiengängen, die jetzt beide denselben Abschluss besitzen, anzugleichen.

Ein ähnlicher Unterschied besteht zwischen den Lehramtsstudiengängen und ihren entsprechenden akademischen Diplomstudiengängen. Abbildung 7 zeigt das deutlich für das Fach Mathematik. Im Projekt wurden Gruppendiskussionen mit Prüferinnen und Prüfern in den Fächern Germanistik und Mathematik durchgeführt, um verschiedene leistungsunabhängige Einflüsse auf die Noten aufzuspüren, die in der Prüfungspraxis ent-

stehen. Generell wurde die höhere Formalität und eingebrachte Beurteilungsroutine der externen Vorsitzenden als möglicherweise notenverschlechternd in Staatsexamensprüfungen beschrieben. Der Unterschied zwischen staatlichen und Hochschulprüfungen stellte sich dort aber auch als studiengangspezifisch heraus: In Mathematik wurden für die Diplom-Studierenden unter anderem deutlich höhere Leistungserwartungen und Betreuungshäufigkeiten als für die Lehramtsstudierenden berichtet, während in Germanistik diesbezüglich keine Unterschiede zwischen den Magister- und den Lehramtsstudierenden erwähnt wurden. Möglicherweise waren ähnliche Einflüsse[12] auch für die Unterschiede zwischen Magister- und Diplomstudierenden der Sozialwissenschaften verantwortlich.

Abbildung 7: Notenentwicklung im Fach Mathematik



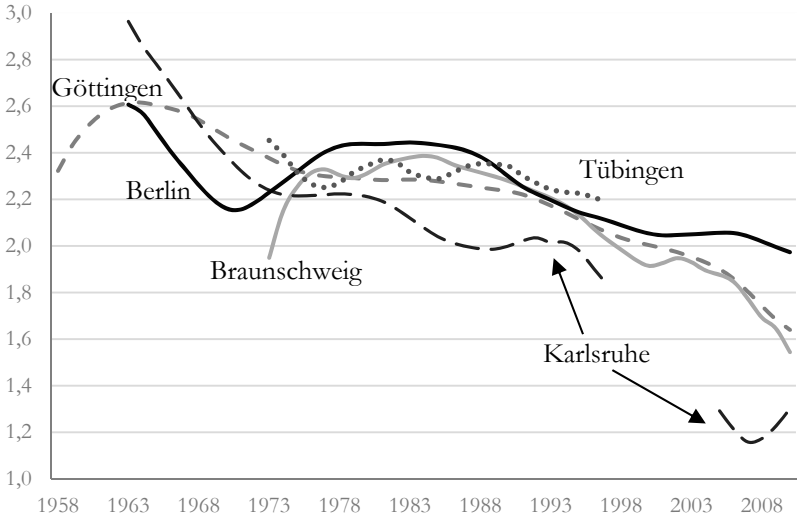
3.3. Die unterschiedlichen Dynamiken der Studiengänge

Abbildung 6 zeigt weiter, dass an der FU Berlin die Entwicklung mindestens bis zum Jahr 2000 zyklisch und im Vergleich zu den anderen vier Universitäten umgekehrt verläuft. Göttingen und Heidelberg folgen ebenfalls einem erkennbaren Zyklus. Die Noten zwischen 1980 und 1995 sind in Berlin deutlich besser, in den anderen Universitäten außer Münster klar schlechter als davor und danach. Abbildung 8 belegt dagegen für den Studiengang »Lehramt an Gymnasien im Fach Deutsch«, dass auch ein recht einheitlicher Verlauf möglich ist, der hier neben gemeinsamen Zyklen aus einer einheitlichen Notenverbesserung besteht.

12 Siehe insbesondere Tsarouha 2017.

Die Unterschiede in der langfristigen Dynamik bestehen offenbar zwischen verschiedenen Abschlussarten: Diplom, Magister oder Staatsexamen. Abbildung 6 deutet an, dass die Konjunktur der eher am Arbeitsmarkt orientierten Studiengänge mit Diplom-Abschluss möglicherweise anders verläuft als die der eher akademisch ausgerichteten (siehe unten Abschnitt 4.2). Abbildung 8 lässt vermuten, dass die staatlich geprüften Studiengänge noch einheitlichere Konjunkturen aufweisen.

Abbildung 8: Notenentwicklung im Fach Deutsch



4. Erklärungen

Wie jede(r) aus eigener Erfahrung als Geprüfte(r) oder PrüferIn weiß, gibt es sehr viele Gründe, die man abgesehen von der Leistung des Geprüften für eine bestimmte Note anführen kann. Hier interessieren vor allem Wirkungen, die langfristig bestehen, das heißt Jahrzehnte oder immer. Im Folgenden werde ich den häufig genannten Einfluss sozialer Strukturen (4.1) und den Einfluss von »Fächerkonjunkturen« (4.2) genauer untersuchen.[13]

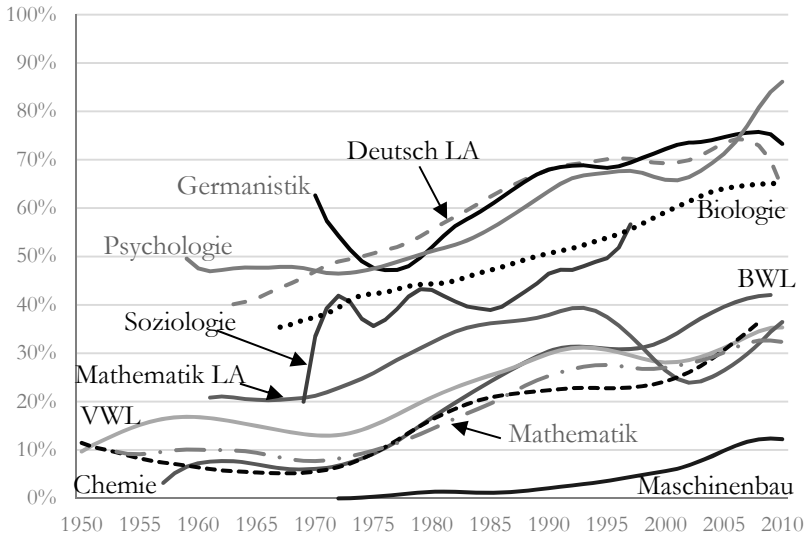
13 Weitere Einflüsse werden von Gaens, Müller-Benedict (2017) und Grözinger (2017) untersucht.

Letzterer hat sich als wichtige Erklärung für *grade inflation* herausgestellt. Für die Untersuchung solcher Einflüsse gilt generell, dass die erforderlichen Daten für alle zu untersuchenden Hochschulen vorhanden sein müssen.

4.1 Struktur der Studierenden, Professorenschaft und Prüfungsordnungen

Die Zusammensetzung der Studierenden nach Geschlecht und Alter, der Ausländeranteil, der Anteil der Stipendiaten etc. kann je Studiengang und Universität sehr unterschiedlich sein. Wenn diese Merkmale die Leistung der Studierenden beeinflussten, zum Beispiel Frauen durchschnittlich fleißiger wären, würde mit einer unterschiedlichen Verteilung von Männern und Frauen ein leistungskonformer Unterschied der Noten erzielt. Insbesondere der Frauenanteil hat sich in allen Studiengängen seit den Bildungsreformen der 1970er Jahre drastisch verändert, wie Abbildung 9 zeigt.

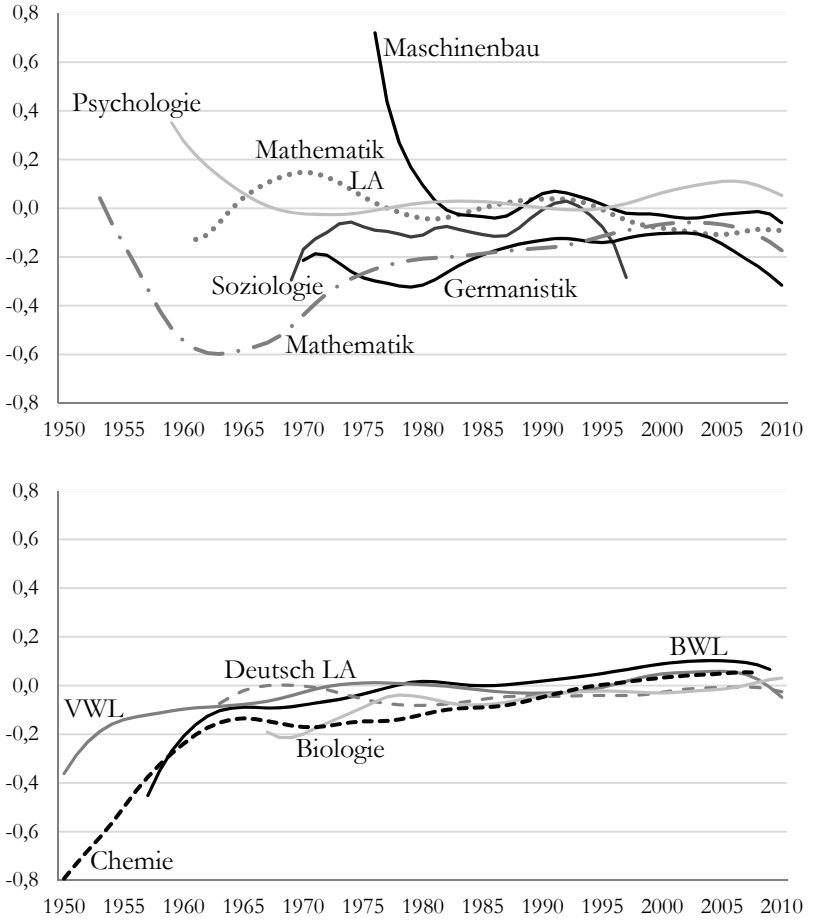
Abbildung 9: Anteil der weiblichen Studierenden nach Studiengang (in Prozent)



Vergleicht man Abbildung 9 mit den Abbildungen 6 und 8, könnte man vermuten, dass die *grade inflation* auf den steigenden Frauenanteil zurückzuführen sei. Das ist aber eine Scheinkorrelation. Bildet man pro Jahr die Differenz aus der durchschnittlichen Note der Männer und der der Frauen,

ist diese für alle Fächer sehr oft nicht signifikant von 0 verschieden (Abbildung 10):

Abbildung 10: Differenz aus der durchschnittlichen Abschlussnote der männlichen und der weiblichen Studierenden



Lesehilfe: Werte im positiven Bereich bedeuten im Durchschnitt bessere, Werte im negativen Bereich schlechtere Noten für weibliche gegenüber männlichen Studierenden. Zur besseren Erkennbarkeit wurden die Studiengänge aus Abbildung 9 auf zwei Grafiken verteilt.

Damit ist klar, dass die Zusammensetzung nach Geschlecht abgesehen von wenigen Zeitabschnitten und Fächern keinen Einfluss auf die Durchschnittsnote hatte. Insbesondere hängt der Verlauf der Noten nach Geschlecht nicht mit dem Verlauf des Anteils der Frauen zusammen.

Allerdings zeigen weitere Berechnungen, dass zum Beispiel der Ausländerstatus und ein höherer Anteil jüngerer und männlicher Professoren notenverschlechternde Wirkung haben. Das bedeutet, dass ein Vergleich der Durchschnittsnoten in einem Fach an verschiedenen Universitäten schon deshalb mit Fehlern behaftet sein wird, weil die Zusammensetzung der Studierenden und der Professorenschaft in der Regel unbekannt sind, aber diese Noten beeinflusst haben.

Darüber hinaus haben Analysen ergeben, dass die in den Prüfungsordnungen vorgeschriebene Zahl an Teilnoten und deren Gewichtungen die Notenhöhe *nicht leistungskonform* beeinflussen. Wird etwa die Abschlussarbeit an einer Hochschule bei derselben Zahl an Teilprüfungen weniger, an einer anderen höher gewichtet, kommen für dieselbe Leistung unterschiedliche Noten heraus.[14]

4.2 Konjunktur der Fächer

Eine wichtige Größe, die einen leistungsunabhängigen Einfluss auf die Noten hat, ist die Konjunktur eines Fachs, das heißt der Wechsel von »Überfüllung« und »Mangel«, sowohl an den Hochschulen als auch am Arbeitsmarkt. Nimmt man die Anzahl der abgelegten Prüfungen[15] als Indikator, sieht man in den meisten Fächern deutliche längerfristige Zyklen.

14 Das heißt, im Zeitverlauf können nicht einmal die Noten an derselben Universität verglichen werden, wenn sich die Prüfungsordnungen geändert haben. Die Vergleichbarkeit der Noten an verschiedenen Universitäten und im Zeitverlauf ist aus diesen Gründen prinzipiell nicht gegeben.

15 Genauer handelt es sich bei den hier abgebildeten Zeitreihen um die abgelegten und bestandenen Prüfungen. Die im ersten oder weiteren zugelassenen Versuchen nicht Bestehenden werden leider von keinem Archiv und keiner amtlichen Statistik erfasst, s. Gaens (2013). Nur die sehr geringe Zahl von »endgültig nicht Bestandenen« wird ab 1997 aufgeführt.

Abbildung 11: Zahl der abgelegten Prüfungen im Studiengang »Mathematik Lehramt«

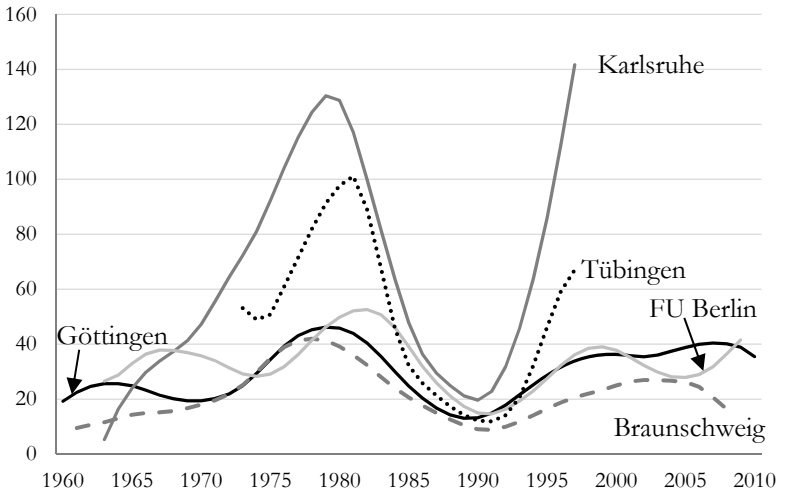
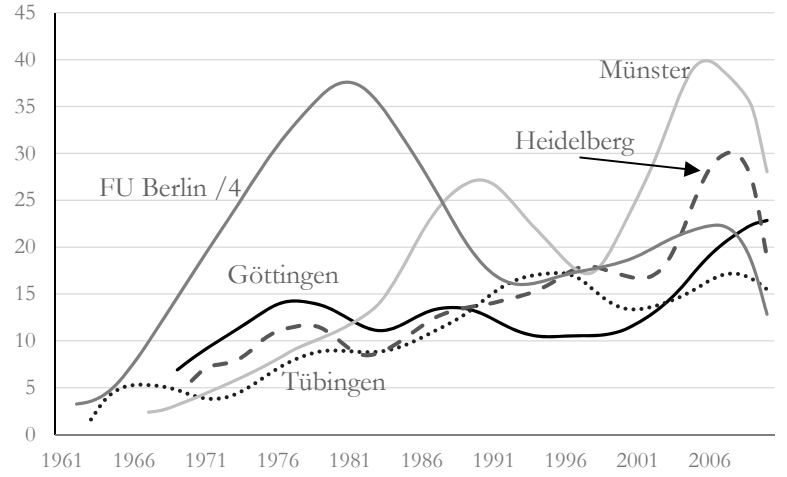


Abbildung 12: Zahl der abgelegten Prüfungen in »Magister« bzw. »Diplom Soziologie«



Diplom nur an der FU Berlin (für Berlin tatsächliche Anzahl geteilt durch 4, um die Zyklen auch der kleineren Universitäten besser sichtbar zu machen).

An den Zahlen abgelegter Prüfungen in den Abbildungen 11 und 12 wird deutlicher als an den Noten (Abbildungen 6 und 7), dass Unterschiede je nach Art des Abschlusses bestehen. In den staatlichen (Lehramts)-Karrieren mit dem staatlichen Nachfragemonopol gibt es die stärksten und für alle Hochschulen synchrone Zyklen (Müller-Benedict 2002). In den auf einen klaren fachspezifischen Arbeitsmarkt ausgerichteten (meist Diplom)-Studiengängen zeigen sich ebenfalls deutliche und synchrone Zyklen. In den Magisterstudiengängen dagegen sind die Zyklen an den Hochschulen am wenigsten synchron. Die Abgänger aus diesen Studiengängen haben oft heterogene Berufsperspektiven in verschiedenen Berufsfeldern. Die wechselnden Berufsaussichten auf dem Arbeitsmarkt beeinflussen dort die Erstsemesterzahlen kaum.

Deshalb stellt die Zahl der abgelegten Prüfungen einen mehrdeutigen Indikator für die Konjunktur der Fächer dar. Zum einen steht sie für die Arbeitsmarktaussichten der Absolventen. Wenn sich in einer Karriere, zum Beispiel im Lehramt, eine Überfüllung und damit schlechtere Berufsaussichten abzeichnen, beginnen die Erstsemesterzahlen immer weiter zu sinken. Sie nehmen erst wieder zu, wenn erneut Mangel auf dem Arbeitsmarkt in Sicht ist. Die Erstsemester reagieren dabei meist ein bis zwei Jahre eher als die öffentliche Wahrnehmung. Die Prüfungszahlen schwanken wie die Erstsemesterzahlen, nur eine Studiendauerlänge später. Deshalb zeigt ein Peak der Prüfungszahl an, dass etwa eine halbe Studiendauer vorher eine Überfüllungsphase begonnen hat. Zum anderen steht die Zahl der abgelegten Prüfungen für die Studienbedingungen einer »überfüllten« Kohorte, zum Beispiel viele TeilnehmerInnen in den Seminaren, wenig Betreuungszeit durch die DozentInnen etc. Im ersten Fall zeigt der Peak in den Prüfungen die Überfüllung auf dem Arbeitsmarkt etwa eine halbe Studiendauerlänge vorher an, im zweiten, dass die aktuelle Prüfungskohorte die »überfülltesten« Studienbedingungen hatte.

Um den Einfluss der Konjunktur der Prüfungszahlen auf die Noten zu prüfen, muss deshalb die Analyse unterschiedlich vorgehen. Für die stark vom Arbeitsmarkt abhängigen Studiengänge wird die zyklische Beziehung der Noten zu den um ca. eine halbe Studiendauerlänge zurück versetzten Zahlen abgelegter Prüfungen auf der Ebene von Studiengängen geprüft; für die Magisterstudiengänge die synchrone Beziehung zwischen beiden auf der Ebene einzelner Hochschulen.[16] Als besonders prägnante Beispiele

16 Dass diese Lag-Kombination der Zeitreihen auch die statistisch stärksten Effekte ergibt, war ein wichtiger Hinweis auf diese Hypothesen.

dienen der Studiengang »Lehramt an Gymnasien in Mathematik« insgesamt (Abbildung 13) und der Studiengang »Magister Germanistik« an der Universität Göttingen (Abbildung 14).

Abbildung 13: Noten und abgelegte Prüfungen im Studiengang »Mathematik Lehramt«

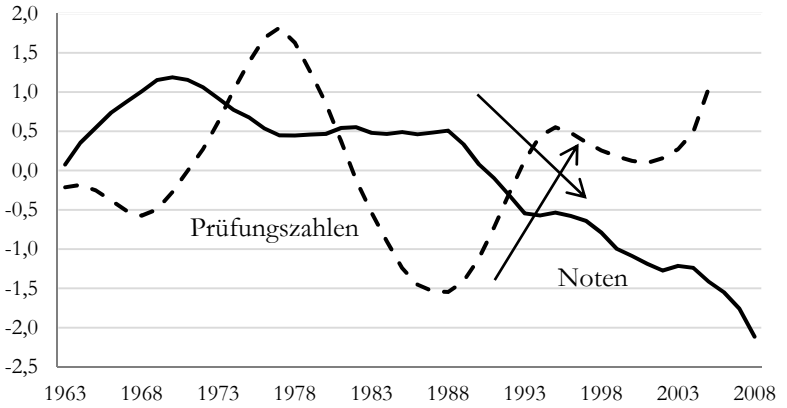
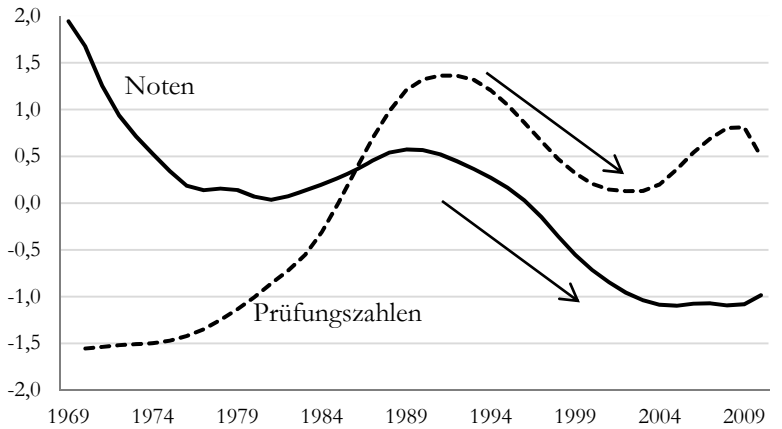


Abbildung 14: Noten und abgelegte Prüfungen im Studiengang »Magister Germanistik« an der Universität Göttingen



Prüfungszahlen 3 Jahre nach links verschoben; zum besseren Vergleich wurden alle Zeitreihen standardisiert (Skala y-Achse in Standardabweichungen vom zentrierten Mittelwert 0).

Wie man sieht, werden die Noten im Studiengang »Lehramt Mathematik« besser, wenn Mangel auf dem Arbeitsmarkt herrscht und umgekehrt ein wenig schlechter bei Arbeitsmarktüberfüllung. Eine mögliche Annahme wäre hier, dass die Prüfer – bewusst oder unbewusst – selektiver benoten, wenn die Aussichten der Geprüften auf einen Arbeitsplatz gering sind, weil sie dann Unterschiede zwischen den Studierenden deutlicher machen können. Im Mangelfall werden umgekehrt alle Geprüften benötigt, und die Prüfer benoten deshalb milder. Die Noten im Studiengang »Magister Germanistik« werden besser, wenn die Prüfungskohorten kleiner werden, und umgekehrt. Hier kann man annehmen, dass die Lehr- und Prüfungsbedingungen sich bei großen Kohorten verschlechtern und auf die Performance und die Benotung der Geprüften entsprechend einwirken. Diese zyklische Abhängigkeit der Notenhöhe von den Prüfungszahlen findet sich in der einen oder der anderen Form in jedem der untersuchten Studiengänge. Für die beiden Soziologie-Studiengänge in Tübingen und FU Berlin sieht man sie in den Abbildungen 15 und 16.

Die Gruppendiskussionen mit den mit Prüferinnen und Prüfern ergaben allerdings nur indirekte Hinweise auf Annahmen, wie die Prüfungszahl bei der Notenfindung eingehen könnte, etwa über Betreuungsintensität, die unterschiedliche Berücksichtigung der Lehrbefähigung bei den Lehrämtern oder die generell diffusen Arbeitsmarktaussichten für Magister.

Abbildung 15: Noten und abgelegte Prüfungen in Tübingen

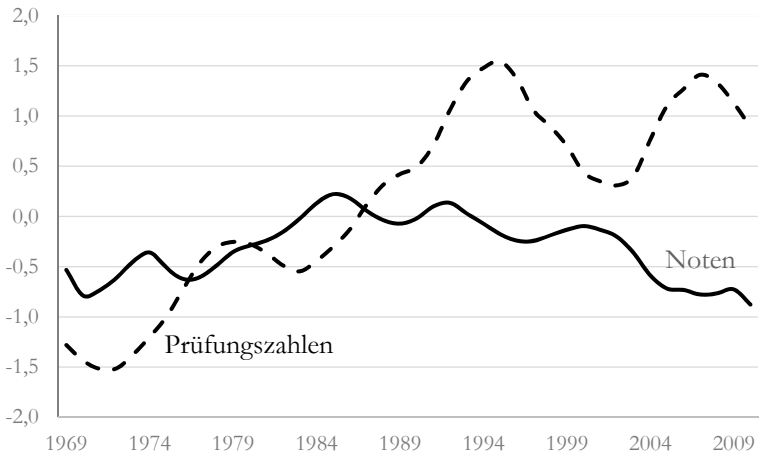
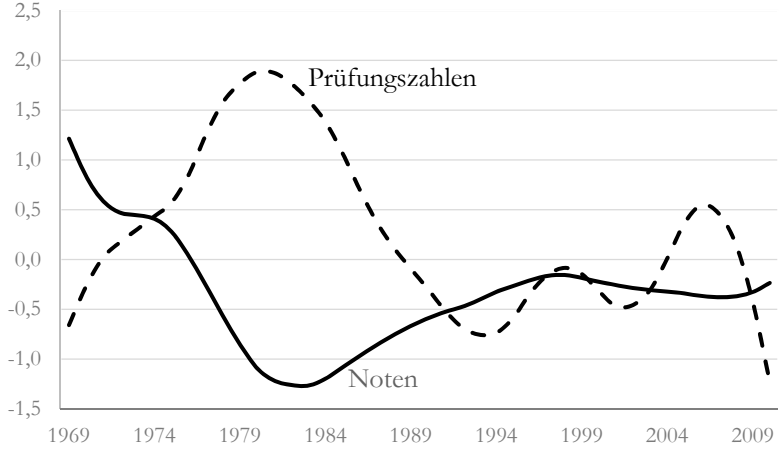


Abbildung 16: Noten und abgelegte Prüfungen an der FU Berlin[17]



4.3 Die Erklärung von *grade inflation*

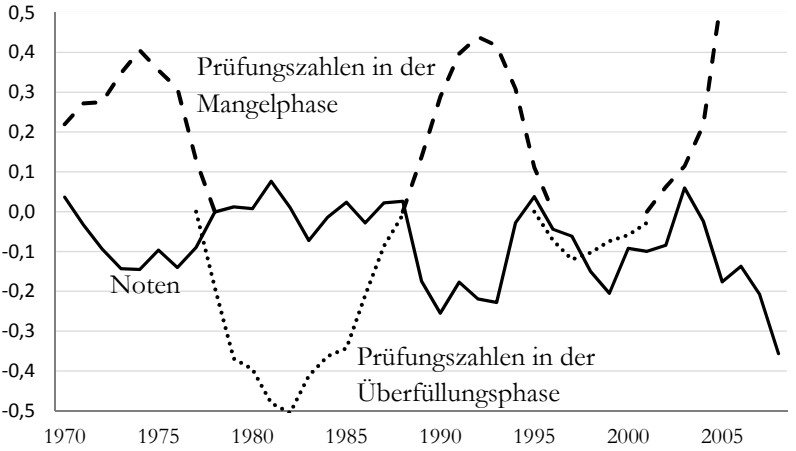
In Abbildung 13 werden die Noten bei Überfüllung nicht im gleichen Ausmaß schlechter wie sie bei Mangel besser werden. Das führt zu einer Hypothese über die generell unterschiedliche Elastizität von Noten: Nachdem sie sich – egal aus welchen Gründen – verbessert haben, werden sie sich in einer anschießenden Verschlechterungsphase weniger stark verschlechtern. Sie sind in Richtung »besser« elastischer als in Richtung »schlechter«. Eine solche Dynamik ist soziologisch gut bekannt, zum Beispiel als Phänomen sozialer Mobilität oder der intergenerationellen Bildungsmobilität: Aufstieg gern, aber auf keinen Fall Abstieg. An den Hochschulen sind bessere Noten bei allen Beteiligten erwünschter als schlechtere: bei den Studierenden, bei den DozentInnen, bei der Verwaltung. Deshalb ist eine »Drift« in Richtung *grade inflation* strukturell angelegt.

Durch Zerlegung des Verlaufs der Prüfungszahlen in Wachstums- und Schrumpfungsphasen konnte diese Hypothese statistisch für fast alle Fächer nachgewiesen werden. Abbildung 17 macht den unterschiedlich star-

17 Ein singulärer, allerdings nicht nachprüfbarer Einfluss zur Erklärung des speziellen Berliner Verlaufs könnte das große Symposium zur Notengebung sein, das 1981 an der FU Berlin stattgefunden hat (Klose, Lange 1981).

ken Einfluss am Beispiel des Lehramts für Mathematik sichtbar: stärkere Verbesserung der Noten in Wachstumsphasen als in Überfüllungsphasen.

Abbildung 17: Wachstum der Noten und der Prüfungszahlen im Studiengang »Mathematik Lehramt«



Wachstum = 1. Differenzen der Zeitreihen

Damit ergibt sich eine Erklärung für die *grade inflation*, die in den meisten Fächern zu beobachten ist: Durch die Zyklen der Zahlen abgelegter Prüfungen verändern sich auch die Noten zyklisch. Dabei verbessern sie sich aber in jedem kompletten Zyklus ein bisschen mehr, als sie sich verschlechtern. So verbessert sich ihr Niveau langfristig in absteigenden Zyklen, wie es in Abbildung 1 zu erahnen ist.

5. Empfehlungen

Um für die Praxis der Notenvergabe an den Hochschulen aus den Ergebnissen Empfehlungen abzuleiten, sollte man den wissenschaftstheoretischen Standpunkt verdeutlichen: Soll ein Hochschulexamen ähnlich wie die Schulausbildung vor allem als Signal für den Arbeitsmarkt funktionieren oder soll es weitgehend durch die Autonomie der wissenschaftlichen Aktivitäten der Institute und DozentInnen geprägt sein? Die systemfunk-

tionale Zwischenstellung des Hochschulsystems zwischen Bildungs- und Wissenschaftssystem schlägt sich hier nieder. Meines Erachtens sollte die Autonomie der Hochschulen als wissenschaftliche Einrichtungen gewahrt bleiben. Das heißt, Vergleichbarkeit durch Vorschriften – wie bei Jura – oder durch Berechnungen kann nicht hergestellt werden. Um die Verwertbarkeit der Noten trotzdem zu sichern, ist Transparenz das wichtigste Mittel. In norwegischen Examenszeugnissen steht hinter jeder Note ein kleines Balkendiagramm mit den Balken 1 bis 4, das die Häufigkeit dieser vier Noten anzeigt, die in den letzten fünf Jahren in diesem Seminar an dieser Hochschule vergeben wurden. Damit wird die Note im Vergleich zu anderen Fächern, Hochschulen und Zeiten einschätzbar. Wenn dann auf jedem Examenszeugnis in einigen Fächern oder an einigen Hochschulen nur noch die Balken für 1 und 2 erschienen, würden Konsequenzen für die Benotungspraxis möglicherweise schneller gezogen werden.

Eine abschließende Empfehlung[18] ergibt sich aus der generellen Unsicherheit, mit der jeder Notenvergleich auf Grund der inhärenten und nur teilweise aufklärbaren Differenzen zwischen Fächern, Examensarten, Hochschulen, Prüfungsordnungen etc. behaftet ist: Bei Aufnahmeentscheidungen, etwa beim Übergang vom Bachelor zum Master, sollte die in die zweite Kommastelle interpretierte Examensnote nicht für alle zu vergebenden Studienplätze ausschlaggebend sein, sondern es sollte ein Teil der Studienplätze, zum Beispiel 20 Prozent, für die Besetzung durch Verlosungen freigehalten werden.

Literatur

Gaens, T. 2013: Von einem, der auszog, einen Leistungsindikator zu erheben. Durchfallquoten und die Problematik ihrer Bildung. Das Hochschulwesen, 61. Jg., Heft 6, 200–206.

Gaens, T., Müller-Benedict, V. 2017: Die langfristige Entwicklung des Notenniveaus und ihre Erklärung. In V. Müller-Benedict, G. Grözinger (Hg.), Noten an Deutschlands Hochschulen. Wiesbaden: Springer VS, 17–78.

Grözinger, G. 2017: Einflüsse auf die Notengebung: eine Analyse ausgewählter Fächer auf Basis der Prüfungsstatistik. In V. Müller-Benedict, G. Grözinger (Hg.), Noten an Deutschlands Hochschulen. Wiesbaden: Springer VS, 79–116.

18 Weitere Empfehlungen finden sich in Müller-Benedict, Grözinger (2017: 183 ff.).

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

Klose, T., Lange, E.M. 1981: Diplomprüfungen im Widerstreit. Die Funktion von Hochschulabschlussprüfungen für das Studium und für den Beruf. Symposium am 29. und 30. April 1981. Dokumentationsreihe der Freien Universität Berlin.

Müller-Benedict, V. 2002: Ist Akademikermangel unvermeidbar? Eine Analyse einer Tiefenstruktur des Bildungssystems. Zeitschrift für Erziehungswissenschaft, 5. Jg., Heft 4, 672–691.

Müller-Benedict, V., Grözinger, G. (Hg.) 2017: Noten an Deutschlands Hochschulen. Wiesbaden: Springer VS.

Tsarouha 2017: Typologie der Einflussgrößen auf die Notengebung. In V. Müller-Benedict, G. Grözinger (Hg.), Noten an Deutschlands Hochschulen. Wiesbaden: Springer VS, 117–169.

Wissenschaftsrat 2003: Prüfungsnoten an Hochschulen 1996, 1998 und 2000 nach ausgewählten Studienbereichen und Studienfächern. Arbeitsbericht. Drucksache 5526-03.

Wissenschaftsrat 2007: Prüfungsnoten im Prüfungsjahr 2005 an Universitäten (einschließlich KH, PH, TH) sowie an Fachhochschulen (einschließlich Verwaltungsfachhochschulen) nach ausgewählten Studienbereichen und Studienfächern. Drucksache 7769-07.

Wissenschaftsrat 2012: Prüfungsnoten an Hochschulen im Prüfungsjahr 2010. Arbeitsbericht mit einem Wissenschaftspolitischen Kommentar des Wissenschaftsrates. Drucksache 2627-12.