Wissenschaftliche Beiträge

Zur Benotung in der Examensvorbereitung und im ersten Examen

Eine empirische Analyse

Emanuel Towfigh/Christian Traxler/Andreas Glöckner[*]

Bislang gibt es zu den Erfolgsfaktoren des staatlichen Teils der Ersten Juristischen Staatsprüfung wenig gesicherte, empirisch belastbare Erkenntnisse. Lediglich die von den Landesjustizprüfungsämtern jährlich veröffentlichten Überblicksdaten über das allgemeine Abschneiden der Kandidatinnen und Kandidaten sind bekannt. Das verwundert vor allem mit Blick darauf, dass die Staatsexamina seit Jahrzehnten und in allen Bundesländern in ähnlicher Form geprüft werden, dass sie für die spätere berufliche Entwicklung von Heerscharen von Juristen große Bedeutung haben, dass sich Jahr für Jahr Tausende Kandidatinnen und Kandidaten dieser Prüfung unterziehen und dass es mit den privaten Repetitorien eine ganze Industrie kommerzieller Examensvorbereiter gibt. Auch aus didaktischer Sicht ist der Mangel an Evidenz unbefriedigend. Gibt es identifizierbare Faktoren, die für den Erfolg in der Examensprüfung eine Rolle spielen? Sind etwa „kluge Köpfe" mit gutem Abitur auch die in der Staatsprüfung erfolgreicheren Juristen (B. II.)? Lohnt sich das Schreiben von Probeklausuren (B. I. 1.), und erwirbt man dabei fachspezifische Fähigkeiten, oder wirkt der Lernfortschritt fächerübergreifend; wie sieht die Lernkurve aus? Gibt es Unterschiede zwischen den Fakultäten (B. III. 2.)? Gibt es Unterschiede zwischen Männern und Frauen (B. I. 2. und II. 3.), Deutschen und Ausländern (B. II. 4.)?

Nachdem wir in einem lernpsychologisch ausgerichteten Beitrag[1] mit Hilfe eines Datensatzes aus dem Examensklausurenkurs der Universität Münster vor allem die Form der Lernfunktion fortgeschrittener Juristen anhand ihrer Ergebnisse im Klausurenkurs untersucht haben, wurden wir eingeladen, hier einen Beitrag zu veröffentlichen, der unsere Ergebnisse der juristischen Fachwelt zugänglich machen soll. In Ergänzung zu dem im vorgenannten Beitrag analysierten Datensatz untersuchen wir erstmalig auch Examensergebnisse von beim Oberlandesgericht Hamm abgelegten

* Dr. iur. *Towfigh* ist Senior Research Fellow am Max-Planck-Institut zur Erforschung von Gemeinschaftsgütern in Bonn. Prof. Dr. oec. pub. *Traxler* ist Professor für Ökonomie an der Hertie School of Governance, Berlin. Prof. Dr. phil. *Glöckner* ist Professor für Psychologie an der Universität Göttingen.

1 *Glöckner/Towfigh/Traxler*, in: Instructional Science 41 (2013), S. 989 ff.

Prüfungen und versuchen, einige der aufgeworfenen Fragen empirisch belastbar zu beantworten.[2]

A. Daten

Für die Analysen haben wir drei Datensätze verwendet. Der erste Datensatz wurde uns vom Oberlandesgericht Hamm zur Verfügung gestellt. Er enthält Daten sämtlicher Examenskandidaten, die dort im Zeitraum von September 2007 bis Dezember 2010 geprüft wurden. Neben dem Geschlecht und dem Geburtsdatum sind die Abiturnote, der Hochschulort und ein „Abschichter"-Merkmal sowie die Noten der schriftlichen Klausurleistungen (in Nordrhein-Westfalen drei zivilrechtliche, zwei öffentlich-rechtliche, eine strafrechtliche; jede Klausur trägt 10 % zur Gesamtnote bei), die Note für das Prüfungsgespräch (über alle Fächer, 30 % der Gesamtnote) sowie die Note für den Kurzvortrag (10 %) vorhanden. Ferner sind die Daten der Klausuren und die Herkunft der Klausur (in NRW: Prüfungsämter, andere Bundesländer mit Tauschvereinbarung) vermerkt. Für die im Datensatz enthaltenen Kandidaten von der Universität Münster hat uns das Oberlandesgericht außerdem unter strengen datenschutzrechtlichen Auflagen Klarnamen geliefert, mit dem Ziel, sie mit Daten aus dem Examensklausurenkurs der Universität zu „akademischen Biographien" zusammenzuführen; die übrigen Daten (d.h. der Kandidaten von den Universitäten Bielefeld und Bochum) haben wir anonymisiert erhalten. Die Examens- und Klausurenkursdaten der Münsteraner Studierenden wurden nach der Zusammenführung anonymisiert.

Die Übermittlung der Namen hat es uns ermöglicht, auch eine Namenskodierung nach Herkunftsregion vorzunehmen. Das geschah zum einen durch eine sog. Onomastikbearbeitung;[3] dabei wurden die Namen Herkunftsregionen zugeordnet, wobei

2 Mit Blick auf den hier verfügbaren Platz berichten wir nur die zentralen Ergebnisse unserer Untersuchung. Weitere Auswertungen und Details zu den statistischen Ergebnissen können, soweit sie die Daten des Münsteraner Universitätsklausurenkurses betreffen, in unserem Beitrag in *Instructional Science* (Fn. 1) und soweit sie die Auswertung der Examensdaten des Oberlandesgerichts Hamm betreffen, in einem Online-Addendum nachgelesen werden. Beides findet sich im Internet unter der Adresse http://www.towfigh.net/zdrw-2014/. Für die Bereitstellung der Daten sind wir beim Oberlandesgericht Hamm dem Vorsitzenden des Justizprüfungsamtes Herrn Vorsitzenden Richter am Oberlandesgericht *Josef Schulte* sowie Herrn Justizamtsrat *Detlef Coböken* außerordentlich dankbar; bei der Universität Münster gilt unser Dank Herrn Dekan Professor Dr. *Thomas Hoeren* sowie Herrn Akademischen Direktor Dr. *Ulrich Weber-Steinhaus*. Herrn Dr. *Manfred Antoni* und Frau *Johanna Eberle* vom German RLC danken wir für die außerordentlich zügige Zusammenführung und Anonymisierung der Datensätze. Für die automatische onomastische Bearbeitung danken wir Herrn Dr. *Andreas Humpert*; die Herren *David Faßbender* und *Patrick Schwentker* haben dankenswerter Weise die manuelle Namenskodierung vorgenommen. Herrn Prof. Dr. *Janbernd Oebbecke*, Frau Dr. *Katharina Towfigh* und Frau *Stefanie Egidy* danken wir für Hinweise zum Manuskript; die verbleibenden Fehler sind freilich unsere.

3 Diese wurde automatisiert vorgenommen durch Humpert & Schneiderheinze Sozial- und Umfrageforschung. Dabei wurden durch einen Abgleich mit einer zentralen Namendatei mit 26.021.479 Namen, die übermittelten Namen 126 Regionen zugeordnet. Die Plausibilität der Zuordnung wird in drei Zuordnungsstufen angegeben (hoch = nur eine Zuordnung des Namens zu einer Region plausibel; mittel = zwei Zuordnungen plausibel, wahrscheinlichste codiert; niedrig = mehr als zwei Zuordnungen plausibel, wahrscheinlichste wird codiert). Zu methodischen Fragen siehe: *Humpert/Schneiderheinze*, in: Gabler/Häder (Hrsg.), S. 187 ff.; *Humpert/Schneiderheinze*, in: ZUMA-Nachrichten 47 (2000), S. 36 ff.; *Humpert*, in: ZUMA-Nachrichten 54 (2004), S. 141 ff.

die Zuordnung 150 Kandidaten mit Migrationshintergrund aus 40 verschiedenen Regionen ergab. Ferner kodierten zwei unabhängig voneinander arbeitende studentische Hilfskräfte die (subjektive) Wahrscheinlichkeit, dass bei einem Namen ein Migrationshintergrund vorliegt, mit einem Wert zwischen 0 (unwahrscheinlich) und 1 (sicher). Die manuellen studentischen Kodierungen waren untereinander hoch korreliert (*Korrelation* = 0,76), und die Korrelation der auf der onomastischen Kodierung basierenden binären Einteilung in deutsche bzw. nicht-deutsche Namensherkunft mit der manuellen war ebenfalls hoch (*Korrelation* = 0,65).

Ferner konnten wir mit zwei Datensätzen der Universität Münster arbeiten. Der erste dieser Datensätze lag bereits unserer ersten Untersuchung zugrunde. Er enthält insgesamt 71.405 anonymisierte Klausurergebnisse von 2.979 Studierenden, die zwischen Oktober 1999 und Januar 2008 mindestens zehn Klausuren im Examensklausurenkurs bzw. im Ferien-Examensklausurenkurs der Juristischen Fakultät geschrieben haben. Ferner beinhaltet der Datensatz unter anderem Geschlecht, Fachsemester und die laufende Nummer der Klausur. Für weitere Einzelheiten sei hier auf die andere Studie[4] verwiesen. Schließlich hat uns die Fakultät für die aktuelle Untersuchung einen zweiten Datensatz mit 2.119 Klausurergebnissen von insgesamt 150 Teilnehmern bereitgestellt, die zwischen dem WS 2005/2006 und dem WS 2008/09 eine Klausur im Examensklausurenkurs bzw. im Ferien-Examensklausurenkurs geschrieben haben. Die übermittelten Daten entsprechen im Wesentlichen denen des vorangegangenen Projekts.

Die Verknüpfung der Daten wurde durch die Mitarbeiter des *German Record Linkage Center* (German RLC) durchgeführt.[5] Von den übermittelten Klausurenkursteilnehmer/innen stimmten 96 namentlich mit Examenskandidaten überein; allerdings haben nur 61 Personen mehr als die für eine sinnvolle statistische Auswertung erforderliche Zahl von zehn oder mehr Klausurenkurs-Klausuren geschrieben. Da eine solch geringe Zahl nur begrenzt belastbare statistische Analysen erlaubt und auch der Auswahlmechanismus für die 150 Kandidaten für uns nicht nachvollziehbar ist (und so mögliche Selektionseffekten nicht beurteilt werden können), können kaum empirisch belastbare Aussagen über „akademische Biographien" gemacht werden.

4 *Glöckner/Towfigh/Traxler*, in: Instructional Science 41 (2013), S. 989 ff.

5 Eine Beschreibung des Vorgehens bei der Verknüpfung findet sich bei *Antoni et al.*, Record linkage of data on state exams in law from the Hamm Court of Appeals and the Faculty of Law of the University of Münster, German Record Linkage Center Working Paper No. wp-grlc-2014-01 (2014), www.record-linkage.de (18.12.2013). Bei Fragen zur Verknüpfung wenden Sie sich bitte direkt an das German RLC. Das German RLC wird durch die Deutsche Forschungsgemeinschaft (DFG) gefördert.

B. Analysen

I. Lernen im Klausurenkurs

Die Studierenden in unserem ersten Klausurenkurs-Datensatz haben durchschnittlich 24 Klausuren geschrieben; zwischen der ersten und letzten lagen durchschnittlich 43 Wochen. Das linke Bild der Abbildung 1 zeigt die Verteilung der durchschnittlichen Noten der Studierenden (M = 5,81 Punkte); auf der rechten Seite ist die Verteilung der einzelnen Klausurergebnisse zu sehen (M = 5,96, MD = 6 Punkte).[6]

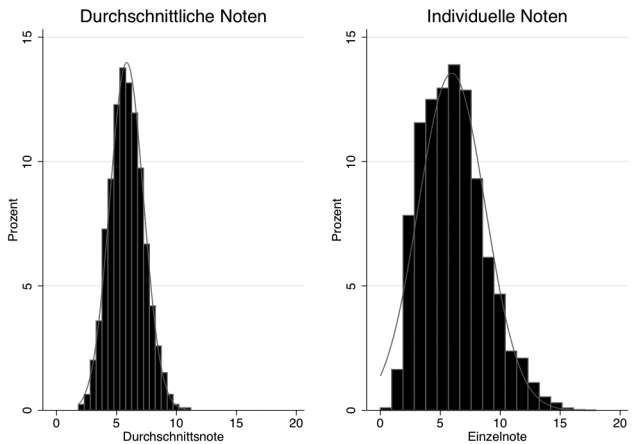

Abb. 1: Durchschnittliche und individuelle Noten

6 Im Folgenden verwendet werden die konventionellen statistischen Abkürzungen: M = Mittelwert; MD = Median; p = Signifikanzwert (hochsignifikant: $p \leq 0{,}01$, signifikant: $p \leq 0{,}05$, marginal signifikant: $p \leq 0{,}1$).

1. Einfluss der Anzahl der geschriebenen Probeklausuren

Zunächst untersuchen wir (anhand des „großen" Klausurenkurs-Datensatzes)[7] die Leistungsentwicklung, gemessen durch die Note, über die Zeit, gemessen anhand der Anzahl der geschriebenen Klausuren. Die Daten (vgl. Abbildung 2) zeigen, dass die Noten mit zunehmender Klausurerfahrung beinahe monoton steigen. Bei der ersten Klausur beginnen Studierende mit einer durchschnittlich signifikant unter 5,5 Punkten liegenden Note; nach 20 Klausuren erreichen sie Noten signifikant über 6,0 Punkten und nach 30 Klausuren erreichen sie die Durchschnittsnote 6,5 Punkte. Dabei sind kleinere Ausschläge zu beobachten, die allerdings nicht in dem Sinne statistisch signifikant sind, dass die Note am Ausschlag signifikant von den Noten der vorigen oder nächsten Klausur abweicht. Eine Regressionsanalyse, die individuelle, saisonale (Sommer/Winter), fachspezifische (Zivilrecht, Öffentliches Recht, Strafrecht) und Reihenfolge-Effekte kontrolliert, bestätigt diesen Befund. Sie ergibt, dass die Studierenden eine leicht konkave Lernkurve durchlaufen.

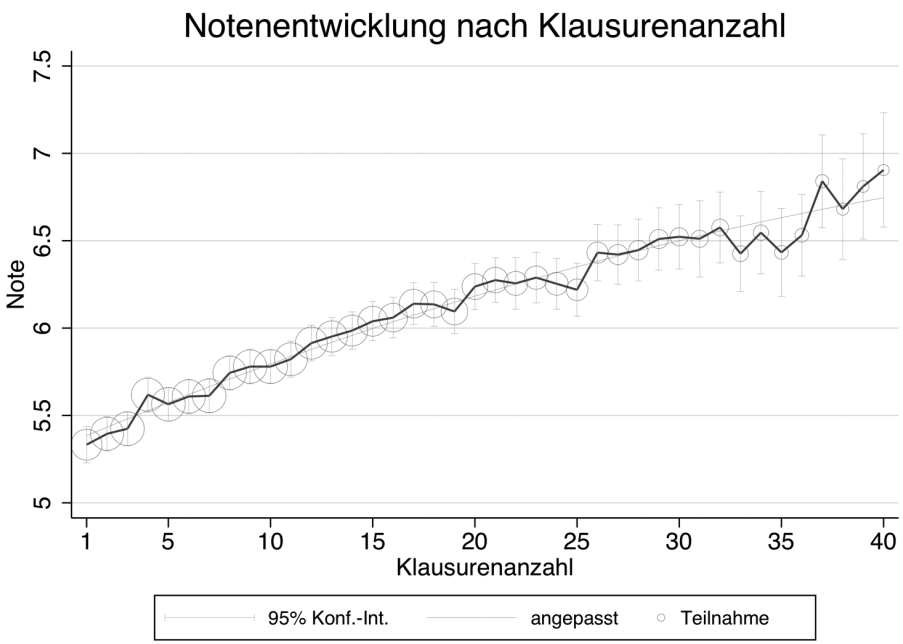

Abb. 2: Notenentwicklung nach Klausurenanzahl. Fehlerbalken bezeichnen das 95 %-Konfidenz-Intervall (basierend auf gepoolten Standardfehlern). Teilnahme stellt die Anzahl der Studierenden dar, die an der jeweiligen Anzahl von Klausuren noch teilgenommen haben (d.h., der Durchmesser des Kreises ist proportional zur Teilnahme)

7 Wir referieren hier im Wesentlichen Ergebnisse der früheren Studie, sofern sie für das juristische Publikum interessant erscheinen. Für eine genauere statistische Auswertung s. *Glöckner/Towfigh/Traxler*, in: Instructional Science 41 (2013), S. 989 ff.

Die Daten weisen darüber hinaus auf Lernfortschritte auf verschiedenen Ebenen hin. Einerseits gibt es „allgemeine Lerneffekte", also unabhängig vom Fach, in dem eine Klausur geschrieben wird. So würde beispielsweise eine Studentin, die zehn zivilrechtliche Klausuren geschrieben hat, in der elften Klausur ein besseres Ergebnis erzielen, auch wenn sie diese im Strafrecht schreibt. Andererseits ist auch ein zusätzlicher „fachspezifischer Lerneffekt" zu beobachten. Je mehr Klausuren ein/e Kandidat/in in einem Fach geschrieben hat, umso besser wird er/sie bei künftigen Klausuren in *diesem* Fach abschneiden; dieser Lerneffekt kommt zum allgemeinen Lerneffekt hinzu. Die Effekte haben jeweils eine durchschnittliche Größenordnung von rund 0,5 % Notensteigerung pro geschriebener Klausur (das heißt, ein „durchschnittlicher" Kandidat mit 6 Punkten verbessert sich pro Klausur um rund 0,03 Punkte); die Steigerungsraten sind also verhältnismäßig klein. Allerdings ist darauf hinzuweisen, dass diese spezifischen Verbesserungen bei Studierenden zu erwarten sind, die das Schreiben von Probeklausuren mit einer davon unabhängigen „normalen" Examensvorbereitung verbinden, die sich also nicht auf das Bearbeiten von Probeklausuren beschränken. Denn in unseren Daten schlägt sich nicht allein der Effekt durch das Schreiben von Klausuren nieder, sondern vielmehr auch der – unbeobachtete und daher nicht isolierbare – Effekt der allgemeinen Examensvorbereitung.

Dabei profitieren die entsprechend der Leistung in den ersten fünf Klausuren besten ($M > 6{,}2$ Punkte), mittleren und schlechtesten ($M < 4{,}6$ Punkte) Studierenden in etwa gleich stark von der Übung durch die Probeklausuren, wenn man die Entwicklung von der 5. bis zur 25. Klausur berücksichtigt (vgl. Abbildung 3). Danach profitieren die schwächeren Studierenden deutlich weniger als die anderen beiden Gruppen von der Teilnahme an zusätzlichen Klausuren.

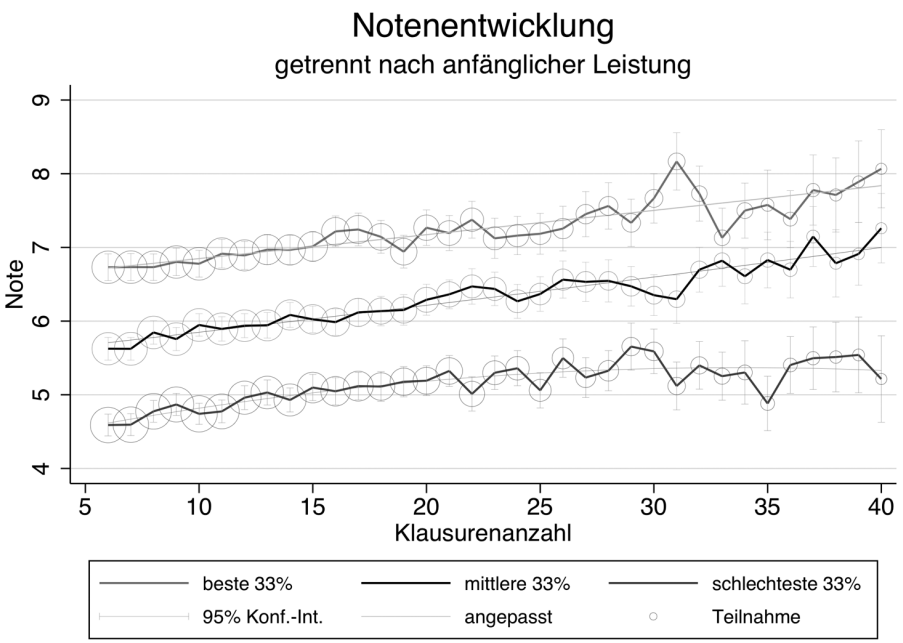

Abb. 3: Notenentwicklung nach Klausurenanzahl, getrennt nach guten, mittleren und schlechten Studierenden. Fehlerbalken bezeichnen das 95 %-Konfidenz-Intervall (basierend auf gepoolten Standardfehlern). Teilnahme stellt die Anzahl der Studierenden dar, die an der jeweiligen Anzahl von Klausuren noch teilgenommen haben (d.h., der Durchmesser des Kreises ist proportional zur Teilnahme)

In einer Regressionsanalyse zeigt sich außerdem, dass bei den „besseren" Studierenden der fachspezifische Lerneffekt stärker ist (das heißt, sie lernen von einer Zivilrechts-Klausur mehr für die nächste Zivilrechtsklausur), während die „schlechteren" Studierenden eher allgemein vom Klausuren Schreiben profitieren; der fachspezifische Effekt dagegen ist nicht so stark. Daraus kann man mit Blick auf die Möglichkeit des Abschichtens von Klausuren vorsichtig schließen, dass dies vor allem für bessere Studierende lohnend sein dürfte. Im Übrigen zeigt sich für die beiden oberen Gruppen ein fast linearer Lernfortschritt. Die konkave Form der Lernfunktion scheint also in erster Linie durch die schwächeren Kandidaten getrieben zu sein. Die Noten im Strafrecht fallen durchweg um ca. 0,5 Punkte schlechter aus als die Noten in den beiden anderen Rechtsgebieten. Der Anstieg der Lernkurve unterscheidet sich anfänglich noch etwas, gleicht sich aber später zunehmend an.

Schließlich zeigen unsere Daten einen interessanten „8-Wochen-Sturz": Nach acht Wochen sacken die Noten substanziell und hochsignifikant ab. Bei genauerer Betrachtung scheint die Deutung plausibel, dass es sich dabei im Wesentlichen um einen motivationalen Effekt handelt – nach etwa acht Wochen intensiver Examensvorbe-

reitung scheint „die Luft ‘raus“ zu sein. Wir beobachten ähnliche, aber schwächere Leistungseinbrüche in den Wochen 15 und 25, allerdings sind diese nicht statistisch signifikant. Das könnte damit zu erklären sein, dass sich das statistische „Rauschen“, also die Messungenauigkeiten, bei der Aggregierung heterogener Individuen mit sich überlappenden, nicht vollständig synchronen „Motivationszyklen“ (z.B. 7 vs. 7,5 vs. 8 Wochen) mit jedem Intervall vergrößert.

2. Geschlechtseffekte

Interessanter Weise haben wir in unseren Daten einen allgemeinen Geschlechtseffekt gefunden, demzufolge Frauen (*M* = 5,83) bei den Probeklausuren schlechter abgeschnitten haben als Männer (*M* = 6,10); der Effekt ist statistisch hochsignifikant. Auch die beiden Lerneffekte – allgemein und fachspezifisch – sind bei Männern ausgeprägter als bei Frauen, das heißt, in unserem Datensatz verbessern sich Männer beim Klausurenschreiben stärker. Die Effekte sind stabil; sie zeigen sich auch im zweiten, kleineren Münsteraner Datensatz.

II. Examensergebnisse

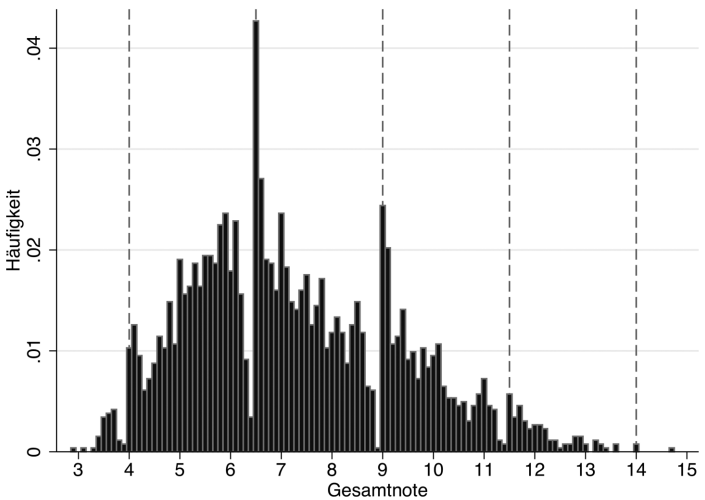

Abb. 4: Notenverteilung der Gesamtnoten im staatlichen Teil der ersten juristischen Prüfung

Wie sieht nun die Benotung im staatlichen Teil der Examensprüfung aus? Abbildung 4 zeigt die Notenverteilung der Gesamtnoten im Examen. In den uns vorliegenden Examensergebnissen haben die Studierenden insgesamt im Durchschnitt 7,27 Punkte erzielt (*MD* = 6,95). Die schriftlichen Noten in den einzelnen Fächer lagen allesamt eng beieinander (Zivilrecht – *M* = 5,39, *MD* = 5; Öffentliches Recht – *M* = 5,47, *MD* = 5; Strafrecht – *M* = 5,44, *MD* = 5). Die Noten des Kurzvortrages (*M* = 7,51, *MD* = 7) und des Prüfungsgespräches (*M* = 8,74, *MD* = 9) und das Gesamtergebnis

sind signifikant besser, wobei zu berücksichtigen ist, dass ca. 1.030 Kandidaten (ca. 28 %) in unserem Datensatz nicht zur mündlichen Prüfung zugelassen wurden oder aus anderen Gründen nicht angetreten sind; der Unterschied bleibt allerdings auch signifikant, wenn man lediglich die Menge der Teilnehmer betrachtet, die beide Prüfungsteile abgelegt haben (und deren durchschnittliche Klausurnoten daher besser sind).

Die Examenskandidaten hatten eine durchschnittliche Abiturnote von 2,13 (*MD* = 2,1). Setzt man Abitur- und Examensnote in Beziehung, so zeigt sich eine signifikante Korrelation (-0,45): Ein um einen ganzen Notengrad besseres Abitur (Schulnotenskala 0,7-6) korreliert in unserem Datensatz mit einem um 1,56 Punkte besseren Examensergebnis (Notenskala 0-18); die Variation in der Abiturnote erklärt dabei etwa 20 % der Varianz in den Examensnoten. Man erhält im Wesentlichen dieselben Ergebnisse, wenn man das Abitur und die Noten der einzelnen Fächer in Beziehung setzt. Wichtig ist, dass es sich bei diesen Auswertungen um Korrelationsanalysen handelt, dass mithin nichts über die *Kausalität* des Abiturs für das Examensergebnis gesagt werden kann.

Analysiert man die Examensergebnisse und die Klausurenkurs-Klausuren jener Personen, die sowohl am Münsteraner Klausurenkurs teilgenommen, als auch in Hamm ihr Examen abgelegt haben, so zeigt sich auch hier eine positive Korrelation zwischen den Durchschnittsnoten der Probeklausuren und im Examen (Korrelation = 0,57). Die Examensnote (8,30) bzw. die durchschnittliche schriftliche Note (7,21) liegen allerdings im Durchschnitt um 2,2 bzw. 1,1 Punkte höher als der Durchschnitt der Klausurenkursnoten (6,10). Wie bereits erwähnt, sind diese Ergebnisse jedoch mit Vorsicht zu interpretieren, weil es sich um eine sehr kleine Stichprobe handelt, bei der zudem der Selektionsmechanismus unklar ist.

3. Muster der Leistungsbewertung im Examen

Betrachtet man Abbildung 4, so zeigen sich erhebliche „Lücken" in der Notenverteilung bei den zwei bis drei Zehntelpunkten unmittelbar vor den Notensprüngen zum „ausreichend", „befriedigend", „vollbefriedigend" und „gut" sowie eine ungewöhnlich hohe Häufigkeit genau auf der Notenschwelle oder in den Zehntel-Punkten kurz danach. In der Verteilung der *durchschnittlichen* schriftlichen Noten (vgl. Abb. 5, links) sieht man zwar auch ein an den Notenschwellen orientiertes Muster, allerdings in sehr viel schwächerem Ausmaß; schaut man sich die tatsächlich vergebenen Einzelnoten an (vgl. Abb. 6, hier sortiert nach Fächern), so zeigen sich in der Verteilung jedoch keine Besonderheiten. Das bedeutet, dass das auffällige Muster in der Verteilung der Gesamtnoten durch die Ergebnisse der mündlichen Prüfung getrieben werden muss. Ein Blick auf die Verteilung der durchschnittlichen Noten in der mündlichen Prüfung (vgl. Abb. 5, rechts) unterstützt den Verdacht, dass die Prüfer in der mündlichen Prüfung die Noten im Bewusstsein und in Abhängigkeit der Ergebnisse der schriftlichen Prüfungen so vergeben, dass bestimmte Notenstufen (nicht) erreicht werden.

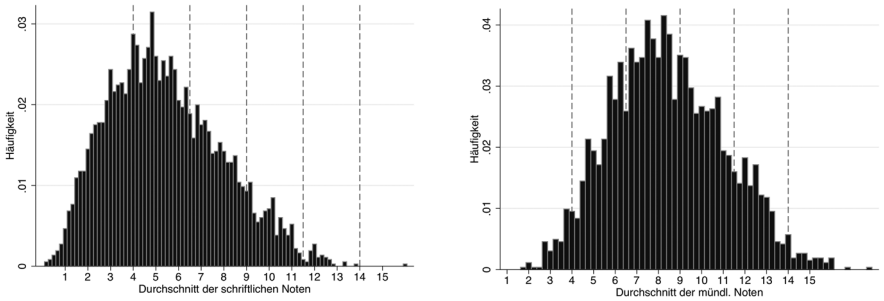

Abb. 5: links: Durchschnitt der schriftlichen Noten, rechts: Durchschnitt mündliche Noten

Das zeigt sich insbesondere in Abbildung 7, die eine besondere Form der Notenverteilung um die Schwellenwerte darstellt. Dazu wurden zuerst alle schriftlichen Durchschnittsnoten („Vorpunkte"), die ±1 Notenpunkt rund um einen Schwellenwert – also die nächste Notenstufe – liegen (4: 3–5; 6.5: 5,5–7,5; 9: 8–10; 11,5: 10,5–12,5 Punkte) herangezogen. Für jeden dieser „knappen" Fälle, wurde dann die Note berechnet, die in den mündlichen Prüfungsteilen benötigt wird, um genau den Schwellenwert zu erreichen; diese im mündlichen Teil insgesamt zu erreichende Note nennen wir Ziel-Note.

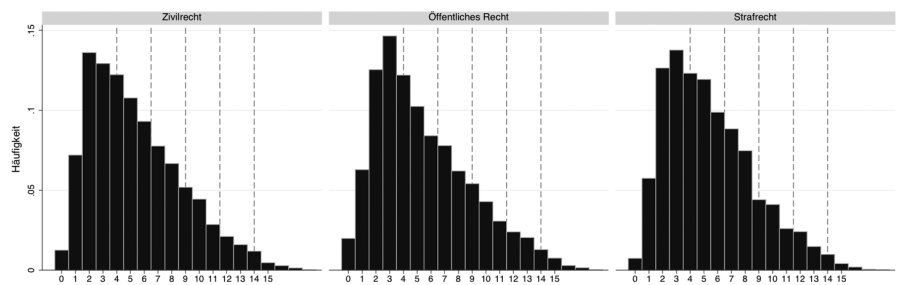

Abb. 6: vergebene Einzelnoten nach Fächern

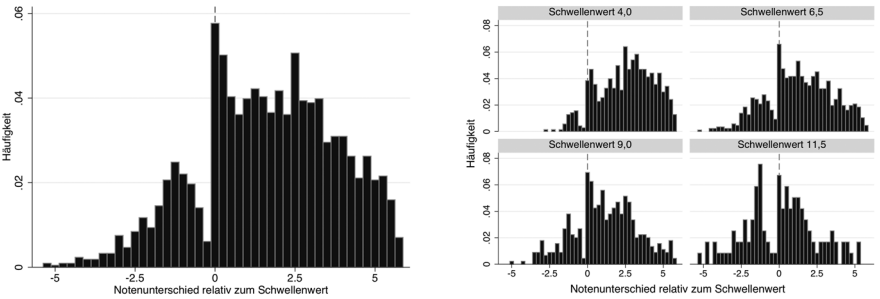

Abb. 7: Verteilung der mündlichen Noten um Schwellenwerte

Die Kandidaten, die in der Grafik (Abb. 7) auf 0 oder darüber liegen, erreichen also die bessere Note; wer darunter liegt, verpasst sie. Wir zeigen diesen Effekt für die unterschiedlichen Schwellenwerte. Die Grafik macht sichtbar, dass systematisch mehr Studierende genau den Schwellenwert erreichen (bzw. knapp darüber liegen) als umgekehrt Studierende diesen knapp verpassen; statistisch gesprochen gibt es „zu wenige" Fälle, die knapp unter dem Schwellenwert liegen. Jene, die den Schwellenwert verpassen, verpassen ihn nicht knapp, sondern in der Regel mit mehr als einem halben Notenpunkt. Die Wahrscheinlichkeit, dass eine solche Notenverteilung zufällig zustande kommt, ist äußerst gering; die Daten legen vielmehr nahe, dass die Prüfer in der mündlichen Prüfung ihre Noten „strategisch", das heißt unter Berücksichtigung der schriftlichen Prüfungsleistungen mit Blick auf das (mögliche) Gesamtergebnis des Examens vergeben.

4. Universitäre Unterschiede

Beim Oberlandesgericht Hamm werden regelmäßig Kandidaten von drei Universitäten geprüft: Bielefeld, Bochum und Münster. Hier soll es nicht spezifisch um die drei rechtswissenschaftlichen Fakultäten gehen; vielmehr ist die Frage, ob man mit der durch die drei Hochschulorte der Studierenden eingeführte Varianz der Daten mehr über die Erfolgsfaktoren des Examens erfahren kann. Betrachtet man zunächst die deskriptiven Examensergebnisse (Gesamtnote) der drei Universitäten, so schneidet Münster (M = 7,45) vor Bielefeld (M = 6,93) und Bochum (M = 6,71) ab; die Unterschiede sind statistisch hoch signifikant. Allerdings fällt auch auf, dass die Studierendenpopulation an den drei Universitäten deutliche Unterschiede aufweist. So liegt die Abiturnote in Münster bei 2,17, in Bochum bei 2,60 und in Bielefeld bei 2,51; in Bochum sind 58 % der Studierenden weiblich, in Münster 53 %, in Bielefeld nur 49 %.

Um die Heterogenität zwischen den Studierenden im Vergleich der Standorte zu berücksichtigen, haben wir eine Regressionsanalyse durchgeführt, bei der wir für die Variablen Abiturnote, Geschlecht, Alter, Prüfungsjahr und -monat sowie für die

„Abschichter" kontrolliert haben.[8] Bildlich gesprochen haben wir damit hinsichtlich *der angegebenen Variablen* „statistische Zwillinge" erzeugt, die sich nur in ihrem Studienort (und natürlich in vielen unbeobachteten Dimensionen!) unterscheiden. Das Ergebnis dieser Regressionen ist, dass der Unterschied zwischen Münster und Bielefeld insignifikant wird, nicht aber der zwischen Münster und Bochum; relativ zur Gesamtnote in Münster sind Bochumer Studierende (mit gleicher Abiturnote, gleichem Alter, gleichen Geschlecht, etc.) *ceteris paribus* noch immer 0,35 Notenpunkte schlechter.

Ein ähnliches Bild zeigt sich, wenn man die Klausurnoten in den einzelnen Fächern betrachtet: Sowohl im Zivilrecht wie auch im Öffentlichen Recht erzielen Studierende aus Bielefeld und Bochum signifikant schlechtere Ergebnisse als jene aus Münster; kontrolliert man für oben genannte Variablen, verliert der Unterschied zwischen Münster und Bielefeld jedoch an Signifikanz. Vergleicht man Münster und Bochum, bleiben die Ergebnisse in den beiden Fächern signifikant schlechter (im Zivilrecht um durchschnittlich 0,69 Punkte, im Öffentlichen Recht um durchschnittlich 0,44 Punkte). In der schriftlichen Strafrechtsprüfung verschwinden durch die Berücksichtigung der zusätzlichen Variablen signifikante Unterschiede zwischen den drei Universitäten ganz.

Ein interessantes Ergebnis ergibt sich bei der Betrachtung der Ergebnisse der mündlichen Prüfungen. Hier schneiden vor Berücksichtigung der Kontrollvariablen Bielefeld und Bochum wiederum hochsignifikant schlechter ab als Münster (Münster: $M = 8{,}77$; Bochum relativ -0,63; Bielefeld -0,47). Nach Einführung der Kontrollvariablen wird der Unterschied zwischen Bochum und Münster insignifikant; die Bielefelder Studierenden schneiden dagegen statistisch signifikant im Mündlichen um 0,35 Punkte besser ab als die Münsteraner Studierenden.

Die Daten dokumentieren, dass – selbst wenn man Abiturnoten, Geschlecht, Alter, Prüfungsjahr und -monat sowie das Merkmal „Abschichter" konstant hält – die Bochumer Studierenden im Zivil- und im Öffentlichen Recht um 6–10 % schlechtere Ergebnisse erzielen. Wie sind diese Unterschiede zu interpretieren, woran mag das liegen? Zum einen könnten die Unterschiede in der Qualität der Ausbildung liegen; dafür spricht, dass in Bochum die Unterschiede zwischen den Fächern signifikant bleiben. Unsere Daten lassen eine solche *kausale* Interpretation nicht notwendigerweise zu, zumal die Regressionsanalyse nur *Korrelationen* dokumentiert. So können die Differenzen auch an systematischen und in unserem Datensatz nicht beobachteten Eigenschaften der in Bochum studierenden Personen liegen, etwa an ihrer Motivation oder ihrem Ehrgeiz. Wir können also, um im Bild unserer Zwillinge zu bleiben, nicht ausschließen, dass zwei tatsächlich vollkommen identische Zwillingsgeschwister nicht doch völlig identische Examensergebnisse erreichen würden, obwohl eines in Münster und eines in Bochum studiert.

8 Zur den Korrelationen bezüglich Geschlecht, Prüfungszeitpunkt und „Abschichtern" siehe sogleich.

Interessanter Weise ist die Bedeutung der Abiturnote an den verschiedenen Studienorten unterschiedlich. Während für Studierende in Münster die Abiturnote 22 % der Varianz der Gesamtnoten des Examens (R^2) erklärt, ist das in Bochum mit nur 11 % gerade die Hälfte; Bielefeld kommt auf 14 %. Betrachtet man die Korrelation von Abiturnote und Gesamtexamensnote, so ist der Zusammenhang in Münster signifikant stärker als in Bochum. Die Interpretation dieses Ergebnisses fällt nicht leicht. Eventuell sammeln sich in Bochum Studierende, deren Abiturnote im Vergleich zu Münster nicht die „tatsächliche Qualität" der Studierenden wiedergibt. Eine andere mögliche Erklärung sind Unterschiede im Auswahlverfahren der Universitäten. Man wird jedenfalls sagen können, dass in Bielefeld und Bochum andere Faktoren als das Abitur wichtiger für die Gesamtnote sind als in Münster.

5. Geschlechtseffekte

Interessante Ergebnisse ergeben sich auch bei der Betrachtung des Abschneidens der Geschlechter. Zunächst sind die Abiturnoten der Studentinnen in unserem Datensatz (also der Frauen, die sich für ein Jura-Studium in Bielefeld, Bochum oder Münster entscheiden), hochsignifikant besser (M = 2,05 bei Frauen, 2,22 bei Männern – Schulnoten). Der Effekt kehrt sich allerdings um, wenn man die Examensnote betrachtet: hier schneiden die Frauen etwa knapp 0,3 Punkte schlechter ab (7,33 Punkte Frauen vs. 7,62 Punkte Männer). Kontrolliert man nun für die Abiturnote (die ja bei Frauen besser ist und damit für eine bessere Note sprechen würde) sowie Alter, Studienort, „Abschichter", Prüfungsjahr und -monat, so steigt der Unterschied auf 0,7 Punkte (ebenfalls hochsignifikant), was ein schlechteres Abschneiden um knapp 10 % bedeutet. Dieser Geschlechtseffekt tritt unabhängig von der Universität und bei allen Prüfungsleistungen auf; er ist bei den zivilrechtlichen Klausuren stärker als bei den öffentlich-rechtlichen und bei diesen stärker als bei den strafrechtlichen. Im mündlichen Prüfungsabschnitt tritt der Effekt stärker auf als im schriftlichen. Erstaunlicherweise schneiden Frauen in der mündlichen Prüfungen auch dann noch schlechter ab (-0,24 Punkte, hochsignifikant), wenn man für die (im Schnitt ja bereits schlechteren) schriftlichen Noten von Frauen kontrolliert. Mit anderen Worten: Wenn man nach den in unserem Datensatz vorhandenen Variablen zwei „statistische Zwillinge" vergleicht, die sich nur im Geschlecht unterscheiden, ansonsten aber die gleiche Abiturnote mitbringen, gleich alt sind, an derselben Uni studiert haben, dieselben Klausuren schreiben und gleichen Vornoten erzielt haben, die sich also lediglich in ihrem Geschlecht unterscheiden, so wird eine weibliche Kandidatin dennoch schlechter benotet als ein männlicher Kandidat.

6. What's in a name?

Für die Studierenden von der Universität Münster haben wir, wie oben erwähnt, anhand ihrer Namen eine Herkunftszuordnung vorgenommen, um auch „Migrationseffekte" untersuchen zu können. Die Abiturnoten der Kandidaten mit nichtdeutschen Namen sind dabei ein wenig schlechter als die jener Kandidaten mit deutschem Namen, der Unterschied ist aber nicht statistisch signifikant (M = 2,11 vs.

2,04 – Schulnoten). Auch hier ist aber der „blanke" Unterschied im Examensergebnis zwischen diesen beiden Gruppen hochsignifikant (*M* = 7,74 Punkte bei Kandidaten mit deutschen Namen, 7,01 Punkte bei den übrigen); kontrolliert man mit den genannten Variablen, sinkt der Unterschied auf ca. 0,5 Punkte. Das Resultat ist robust und unabhängig davon, ob man die onomastisch bearbeitete Klassifizierung oder die manuelle Kodierung durch die studentischen Hilfskräfte zugrunde legt. Die Studierenden mit nicht-deutschen Namen bekommen über alle Prüfungsleistungen hinweg schlechtere Noten. Kontrolliert man bei den mündlichen Noten zusätzlich für die schriftlichen Examensnoten, verbleibt immer noch ein (zusätzlicher) Effekt zwischen 0,25 Punkten (Onomastik-Verfahren, insignifikant) und 0,43 Punkten (manuell kodiert, signifikant), um den die mündliche Prüfungsleistung durchschnittlich noch einmal schlechter bewertet wird als die schriftliche.

Zwei Herkunftsregionen weisen besondere Merkmale auf. Zum einen erreichen diejenigen Jura-Studierenden, deren Namen eine Herkunft aus dem Gebiet der ehemaligen UdSSR nahelegt, eine um 0,46 Schulnoten hochsignifikant bessere Abiturnote als die übrigen Jura-Studierenden; im Examen schneiden sie aber etwa 0,99 Punkte schlechter ab (signifikant, im Wesentlichen getrieben durch die Klausuren). Zum anderen schneiden die Examenskandidaten, deren Namen onomastisch auf eine Herkunft aus dem Mittleren Osten hindeutet, wenn man für ihr Abitur kontrolliert, im schriftlichen Teil des Examens nur insignifikant schlechter ab, aber im Mündlichen erreichen sie lediglich eine um 1,25 Punkte schlechtere Note als ihre Kommilitonen (hochsignifikant). Kontrolliert man bei der mündlichen Note wiederum für die Leistungen im schriftlichen Teil, verbleibt ein starker Effekt von einem Punkt (hochsignifikant); im Gesamtergebnis wirkt sich dies mit einem Manko von 0,75 Punkten (hochsignifikant) aus.

7. „Abschichter"

Schließlich haben wir Regressionen für „Abschichter" gerechnet. Dabei zeigt sich, dass nach Kontrolle für Hochschule, Abiturnote, Alter, Geschlecht, Prüfungsjahr und -monat „Abschichter" eine um 0,46 Punkte bessere Gesamtnote erreichen (hochsignifikant); dieser Effekt findet sich in allen mündlichen und schriftlichen Teilbereichen; am stärksten ist er im Strafrecht, am schwächsten im Zivilrecht. Auch hier ist zu betonen, dass es sich bei der Analyse lediglich um eine Korrelation handelt. Ob die besseren Noten tatsächlich durch das Abschichten getrieben sind oder ob nur die motivierteren oder ehrgeizigeren Studierenden eher abschichten, kann nicht mit Sicherheit gesagt werden. Immerhin weisen die Daten aus dem Münsteraner Klausurenkurs darauf hin, dass sich jedenfalls in der Examensvorbereitung bei besseren Kandidaten stärkere fachspezifische Lerneffekte einstellen; für diese Studierenden könnte es also lohnend sein, die Examensvorbereitung durch ein Abschichten der Klausuren stärker zu gliedern.

8. Weitere Faktoren

a) Saisonal

Die Gesamtnoten sind über den beobachteten Zeitraum stabil und nicht signifikant unterschiedlich. Allerdings sind relativ zum Examensmonat Januar die schriftlichen Noten im September in allen Fächern um 0,5 bis 1,2 Punkte deutlich und hochsignifikant besser. Ein ähnlicher Effekt, allerdings weniger stark und lediglich für das Zivil- und Strafrecht signifikant, zeigt sich bei den Prüfungen im April. Bei den mündlichen Noten sind solche Effekte nicht sichtbar. Bei den saisonalen Effekten – zumal in den beiden genannten Monaten – liegt die Überlegung nahe, dass diese vor allem durch die „Freischüssler" getrieben werden (§ 25 JAG NRW), für die diese beiden Monate regelmäßig die letztmöglichen Examenstermine sind, die deshalb in diesen Terminen verstärkt anzutreffen sind und die im Allgemeinen als stärkere Kandidaten bezeichnet werden. Mit unserem Datensatz lässt sich diese Hypothese allerdings nicht überprüfen, so dass die Beantwortung dieser Frage späteren Forschungsarbeiten anheimfällt.

b) Klausursteller

Bei den Examensprüfungen sind über Ländergrenzen hinweg agierende „Tauschringe" üblich geworden, die zeitgleich identische Klausuren stellen. Nimmt man die Klausuren der Referenzgruppe Berlin zum Maßstab, so ergibt sich, dass sich die Noten für die unterschiedlichen Klausuren der unterschiedlichen Klausursteller im Wesentlichen nicht signifikant unterscheiden (Hamburg, Hessen, Mecklenburg-Vorpommern, Niedersachen, Nordrhein-Westfalen, Rheinland-Pfalz, Saarland), mit Ausnahme von Bremen (0,67 Punkte bessere Ergebnisse, signifikant), Thüringen (0,51 Punkte besser, signifikant) und Sachsen (-0,28 Punkte schlechter, signifikant).

C. Diskussion

I. Vorbereitung durch Probeklausuren

Zunächst einmal ist festzuhalten, dass das Schreiben von Probeklausuren in der Examensvorbereitung sich lohnt, auch wenn der Effekt jeder einzelnen Klausur bescheiden ist. Einerseits übt das Klausurenschreiben und trägt für sich genommen zu einem allgemeinen und einem fachspezifischen Lernfortschritt bei; andererseits ist es ein zuverlässiges Maß für die Einschätzung der Lernentwicklung. Unsere Ergebnisse zu den Motivationszyklen unterstreichen die Bedeutung von Pausen in der aufreibenden Examensvorbereitung, wobei ein Rhythmus von sieben bis acht Wochen empfehlenswert zu sein scheint. Schließlich kann man sagen, dass jedenfalls beim schwächsten Drittel der Studierenden der Lerneffekt vom Lösen von Probeklausuren etwa nach 25 Klausuren deutlich nachlässt; bei allen weiteren beobachten wir bis zur 40. Klausur durchgehend Steigerungen. Danach steigen in den Klausurenkursen zu viele Kandidaten aus, um noch statistisch robuste Aussagen machen zu können; Veränderungen können dann vorwiegend endogen getrieben sein, also etwa dadurch, dass systematisch stärkere (die beispielsweise genug Übung zu haben meinen) oder

schwächere Kandidaten (die beispielsweise schwächer motiviert sind) aufhören, Klausuren zu schreiben. Die in unserem Datensatz verzeichneten Studierenden haben über durchschnittlich 43 Wochen hinweg durchschnittlich 24 Klausuren geschrieben.

II. Strategische Notenvergabe

Bei der Bewertung der Leistung im Examen sieht man deutliche Effekte an den Notenschwellen. Diese sind in der mündlichen Prüfung besonders stark und systematisch ausgeprägt. Die Prüfer berücksichtigen offenbar die schriftlichen Noten der Kandidaten und richten die Beurteilung an den Notenschwellen aus. „Knappen" Kandidaten (± 1 Punkt von einer Notenschwelle) gelingt übermäßig häufig eine „Punktlandung" exakt am Schwellenwert oder knapp über die Hürde; knapp „schlechtere" Ergebnisse (< 0,5 Punkte unter einer Schwelle) treten jedoch kaum auf und werden offensichtlich von den Prüfern systematisch vermieden – wer es nicht schafft, schafft es „deutlich" nicht.

Wir haben mit einer Reihe von Prüfern im ersten Staatsexamen gesprochen, die allesamt von diesem Befund nicht überrascht waren; die Frage des „Anhebens" werde in den Beratungen der Prüfungskommission regelmäßig offen erörtert. Den Kandidaten solle das Gefühl erspart bleiben, knapp an einer wichtigen Schwelle gescheitert zu sein; ferner könne durch eindeutige Ergebnisse die Wahrscheinlichkeit von Rechtsbehelfen reduziert werden. Bisweilen wollen Prüfer mit dieser Benotung auch Zufälligkeiten, Unbilligkeiten und Unschärfen des schriftlichen Teils der Prüfung abmildern und die Übergänge zwischen den willkürlich gesetzten Notenschwellen, die für die spätere berufliche Laufbahn der Kandidaten erhebliche Konsequenzen haben, etwas deutlicher gestalten. Unbewusste Prozesse der Akzentuierung und der Versuch der Einordung in klar abgrenzbare Kategorien (z.B. die Kategorisierung der Kandidaten als „vollbefriedigend" vs. „befriedigend" usw.) könnten dabei ein wichtige Rolle spielen.[9] Spricht man dagegen mit nicht prüfenden Juristen, dann vernimmt man oftmals Unverständnis ob dieser Benotungspraxis. Durch diese Beeinflussung könnten – bewusst oder unbewusst – Quellen der Willkür ins Verfahren gelangen. Das ist, insbesondere wenn man die systematisch schlechteren mündlichen Ergebnisse von Frauen und von Kandidaten mit Migrationshintergrund sieht, nicht auszuschließen. Hinzu kommt, dass von diesem Effekt nur jene Kandidaten betroffen sind, die um eine Notenschwelle herum gruppiert sind – nur sie werden über die Schwelle gehoben oder unter die Schwelle gedrückt. Die Vergleichbarkeit der Punkt-Noten wird damit reduziert.

Ob man den Notenschwellen-Effekt für problematisch hält, hängt wohl maßgeblich damit zusammen, ob man in der Staatsprüfung allein ein Verfahren der Leistungsmessung sieht oder ob man auch die biographische Bedeutung des Examens zu be-

9 Die zentrale Rolle dieser unbewussten Prozesse der Kohärenzbildung bei rechtlichen Urteilen wird bspw. in den folgenden Quellen nachgewiesen und diskutiert: *Simon*, in: University of Chicago Law Review 71, S. 511 ff.; *Glöckner/Engel*, in: Journal of Empirical Legal Studies, 10, S. 230 ff.

rücksichtigen bereit ist. Was bedeutet es, mit 0,02 Punkten an der nächsten Notenschwelle zu scheitern? Will man ein schon verhältnismäßig objektives Prüfungsverfahren weiter optimieren, oder akzeptiert man mit Blick auf die Berufslaufbahn einen Rest Dezisionismus?[10]

Wollte man eine strategische Notenvergabe in Bezug auf Notenschwellen unterbinden, so würde es wohl genügen, den Prüfern der mündlichen Prüfung nicht die genauen Vorpunkte der Kandidaten im schriftlichen Teil der Examensprüfung mitzuteilen, sondern lediglich die Notenstufe (also etwa „ausreichend" oder „vollbefriedigend"), so dass ihnen auch eine Orientierungsmöglichkeit für den anzustrebenden Schwierigkeitsgrad der Prüfung bleibt.

III. Universitäre Unterschiede

Die Analysen haben ferner gezeigt, dass der Hochschulort der Kandidaten sich erheblich im Examensergebnis niederschlägt. Rund 20 % der Varianz lassen sich mit dem Abitur erklären, was sich mit Ergebnissen vorangegangener Studien für andere Fachbereiche wie beispielsweise Medizin und Psychologie deckt.[11] Soweit darüber hinaus Unterschiede sichtbar sind, kann ihre Ursache nicht restlos aufgeklärt werden. Sicher ist, dass eine Reihe von Faktoren eine Rolle spielt, die in unseren Daten unbeobachtet bleiben. So wäre es möglich, dass manche Universitäten systematisch Studierende mit besonderen Profilen anziehen und dass sich diese Eigenschaften auch auf den Examenserfolg auswirken. Es liegt aber auch nahe zu vermuten, dass die Ausbildung einen Effekt hat, weil die Unterschiede in der Leistung im schriftlichen Teil des Examens über die Fächer hinweg variieren. Genauere Aussagen würden aber eine intensivere Untersuchung erfordern, etwa der – sich ebenfalls nach Fächern unterscheidenden – Studierendenbiographien und des universitären Teils der ersten juristischen Prüfung.

IV. Diskriminierung?

1. Geschlechtseffekte

Überraschend für uns war, dass wir deutliche Geschlechtseffekte bei der Datenanalyse gefunden haben und zwar sowohl im universitären Klausurenkurs als auch im

10 Siehe zur Bedeutung des Prüfungsverfahrens in einem größeren Zusammenhang und allgemein zu diskutierten Optimierungen *Oebbecke*, in: JR 2003, S. 397 ff.

11 Studien zeigen (um Reliabilitätseinschätzungen und Varianzeinschränkungen korrigierte) Korrelationen zwischen Abiturnote und Studienerfolg (Noten) vorwiegend im Bereich 0,40-0,60 und mithin eine Varianzerklärung zwischen 20 % und 36 %. Einen Überblick über vorliegende Befunde bieten *Formazin et al.*, in: Psychologische Rundschau 62, S. 221 ff. Ebenda wird statistisch nachgewiesen, dass diese Zusammenhänge dadurch entstehen, dass sowohl Studienleistung als auch Abitur durch Intelligenz (speziell der Fähigkeit zu schlussfolgerndem Denken sowie durch Wissen) getrieben werden.

Examen selbst.[12] Die sich hier aufdrängende Frage ist, ob Frauen im Examen diskriminiert werden. Diese Frage lässt sich mit unseren Daten nicht beantworten. Für eine Diskriminierung würde zunächst einmal sprechen, dass es keine offensichtlichen Gründe gibt, weshalb Frauen – die auch bessere Abiturnoten aufweisen – schlechtere Juristen sein sollten, dass sie aber gleichwohl bereits bei den verhältnismäßig diskriminierungsunanfälligen Klausuren signifikant schlechter abschneiden. Zwar werden die Klausuren unter einer Kennziffer geschrieben, so dass die Prüfer nicht über die Namen auf das Geschlecht der Kandidaten schließen können; allerdings könnte etwa die Handschrift entsprechende Hinweise liefern und unterbewusst wirken. Gegen eine Diskriminierung in diesem Bereich spricht eventuell, dass beim „Üben" im Klausurenkurs Frauen auch geringere Lernfortschritte (d.h. eine geringere Steigerung je Klausur) erzielen. Erstaunlicherweise gibt es aber sogar einen über den im schriftlichen Teil der Examensprüfung hinausgehenden negativen Effekt, wenn man die mündlichen Prüfungen betrachtet: Hier schneiden Frauen durchschnittlich *noch* schlechter ab. Im Gespräch mit Prüfern haben wir immer wieder gehört, dass die schwächere Bewertung von Studentinnen in der mündlichen Prüfung damit zusammenhänge, dass sich diese aufgrund allgemein beobachteter geringerer Selbstsicherheit weniger aktiv am Prüfungsgespräch beteiligten und seltener (non-verbal) signalisierten, dass sie eine Frage beantworten wollen.

Interessanterweise zeigen Untersuchungen mit Studierenden sowohl der Rechtswissenschaften als auch der Betriebswirtschaftslehre an der Harvard University ähnliche Befunde: Frauen sind bei Studienbeginn gleich gut, machen aber am Ende schlechtere Abschlüsse.[13] Weil dieser Effekt bei unterschiedlichen Fächern auftritt, ist nicht anzunehmen, dass es sich um einen fachspezifischen Effekt handelt. Dass die von Prüfern im Staatsexamen angeführte weniger aktive Prüfungsbeteiligung von Frauen eine Rolle spielt, legt auch ein als Reaktion auf die Befunde an der Harvard University durchgeführtes Pilotprojekt nahe. So konnte unter anderem gezeigt werden, dass eine stärkere Motivation zur aktiven Mitarbeit nachteilige Effekte reduzieren kann. Aber was sind die Ursachen dieses zurückhaltenden und mit Blick auf die Benotung nachteiligen Verhaltens? Psychologische Untersuchungen legen nahe, dass wahrgenommene Stereotype ein solches Verhalten bedingen können. Speziell konnte gezeigt werden, dass die vom Prüfling empfundene „Bedrohung" durch wahrgenommene

12 Einer aktuellen amerikanischen Studie zufolge zeigt sich auch im Anwaltsberuf zwischen den Geschlechtern ein erheblicher Unterschied in der Leistung (gemessen an den Mandanten berechneten Arbeitsstunden und am Akquisevolumen), der aber möglicherweise mit motivationalen Effekten zu begründen ist: Den Daten zufolge streben Frauen deutlich seltener die Partnerschaft in einer Sozietät an. Vgl. *Ferrer/Azmat*, Gender Gaps in Performance: Evidence from Young Lawyers, Working Paper, Stand: März 2012, http://ssrn.com/abstract=2050037 (18.12.2013).

13 Eine Zusammenfassung der Befunde sowie umfangreichen Debatten zu dem Thema finden sich unter: http://ontheculture.com/discrimination-the-women-of-harvard-law-school/ sowie unter http://www.nytimes.com/2013/09/08/education/harvard-case-study-gender-equity.html?_r=0 (beide 18.12.2013). Zum aktuellen Diskussionsstand zu Diskriminierungen in der Ausbildung siehe auch *Hanna/Linden*, in: American Economic Journal: Economic Policy 4 (2012), S. 146 ff.; *Breda/Ly*, Do Professors Really Perpetuate the Gender Gap in Science? Evidence from a Natural Experiment in a French Higher Education Institution, Centre for the Economics of Education (LSE, London), Discussion Papers #0138 (2012).

Stereotype beispielsweise bezüglich systematischer Geschlechterunterschiede zu einer Verschlechterung der tatsächlichen Leistung führt, da kognitive Ressourcen durch die Beschäftigung mit der Bedrohung gebunden und somit von der eigentlichen Aufgabe abgezogen werden.[14]

§ 2 Abs. 2 JAG NRW zufolge soll die Prüfung zeigen, „dass der Prüfling das Recht mit Verständnis erfassen und anwenden kann und über die hierzu erforderlichen Rechtskenntnisse in den Prüfungsfächern mit ihren [...] Bezügen, ihren rechtswissenschaftlichen Methoden sowie [...] Grundlagen verfügt. Dies schließt Grundkenntnisse über Aufgaben und Arbeitsmethoden der rechtsberatenden Praxis ein." Zwar mögen Auftreten, Präsentation und Sprache bei der späteren Berufsausübung wichtige Erfolgsfaktoren sein. Gemessen am Maßstab des JAG scheinen „selbstsicheres Auftreten" und eine aktive Beteiligung am Prüfungsgespräch allerdings sachfremde Erwägungen zu sein, die eine systematische Diskriminierung von Frauen im Prüfungsgespräch nahelegen. Dem halten erfahrene Prüfer allerdings entgegen, dass das Ablegen der Staatsexamina einer konkreten Berufsqualifikation (zum Richteramt) diene, so dass man die für die Berufsausübung erforderlichen Fähigkeiten in der Prüfung nicht unberücksichtigt lassen dürfe. Es gebe eben keine Rechtskenntnisse „an sich", sondern nur Kenntnisse in einer bestimmten Präsentationsform, und die schnelle und selbstsichere Reaktion auf ein auftauchendes Rechtsproblem sei eine wichtige richterliche Fähigkeit, die daher auch im Rahmen von § 2 Abs. 2 JAG NRW geprüft werden dürfe.

2. Namensherkunft

Auch die bei der Datenanalyse beobachteten Herkunftseffekte verursachen ein Unbehagen. Bei den Kandidaten mit einem Namen, der eine Herkunft aus dem Gebiet der ehemaligen UdSSR nahelegt, sind trotz deutlich besserer Abiturnoten (im Vergleich zu allen übrigen Jura-Studenten von der Universität Münster in unserem Datensatz) die schriftlichen Klausurergebnisse schlechter; eine zusätzliche Verschlechterung in der mündlichen Prüfung ist nicht zu beobachten. Man mag spekulieren, dass dies gegebenenfalls mit sprachlichen Defiziten zusammenhängen könnte, die möglicherweise in der Abiturprüfung (eventuell. aufgrund entsprechender Fächerwahl) nicht voll durchschlagen, oder mit unterschiedlichen Lernstilen, die im Abitur günstige, im Examen aber ungünstige Folgen haben.

Gravierender scheinen die Ergebnisse bei denjenigen Examenskandidaten zu sein, deren Name eine Herkunft aus dem Mittleren Osten vermuten lässt. Bei diesen lassen sich im Abitur und bei den Examensklausuren keine signifikanten Unterschiede zur übrigen Stichprobe feststellen; erst in der mündlichen Prüfung gibt es massive Abstriche. Wenn schriftlicher und mündlicher Prüfungsteil im Wesentlichen demselben Zweck dienen und wie bei den Frauen Auftreten und Präsentation eigentlich keine

14 Arbeiten zu dem Phänomen der Bedrohung durch wahrgenommene Stereotype auf Leistung von Frauen und Minoritäten in verschiedensten Kontexten finden sich bspw. in *Steele*, in: American Psychologist 52, S. 613 ff., sowie in *Spencer et al.*, in: Journal of Experimental Social Psychology 35, S. 4 ff.

Rolle spielen sollten, dann ist dieser Befund nur schwer zu erklären. Hier liegt es nahe, eine Diskriminierung anzunehmen.

Sowohl bei dem Geschlechts- als auch beim Herkunftseffekt können wir eine Diskriminierung weder mit der notwendigen Gewissheit ausschließen noch sie nachweisen. Vielmehr scheinen hier weitere empirische Untersuchungen lohnend, weil sie interessante und praktisch relevante Einsichten erwarten lassen. Einiges spricht dafür, dass es sich nicht notwendigerweise um eine bewusste Diskriminierung handelt, dass vielmehr die subjektive Wahrnehmung von Stereotypen die tatsächliche Leistungsfähigkeit von Stereotypen betroffener Kandidaten reduziert; für Frauen wurde dieser Effekt, wie oben erörtert, bereits andernorts gezeigt, und auch für Ausländer scheint er nicht unplausibel.

Literaturverzeichnis

Breda, Thomas/Ly, Son Thierry, Do Professors Really Perpetuate the Gender Gap in Science? Evidence from a Natural Experiment in a French Higher Education Institution, Centre for the Economics of Education (LSE, London), Discussion Papers #0138 (2012).

Formazin, Maren/Schroeders, Ulrich/Köller, Olaf/Wilhelm, Oliver/Westmeyer, Hans, Studierendenauswahl im Fach Psychologie, in: Psychologische Rundschau 62 (2011), S. 221-236.

Glöckner, Andreas/Engel, Christoph, Can We Trust Intuitive Jurors? Standards of Proof and the Probative Value of Evidence in Coherence-Based Reasoning, in: Journal of Empirical Legal Studies 10 (2013), S. 230-252.

Hanna, Rema N./Linden, Leigh L., Discrimination in Grading, in: American Economic Journal: Economic Policy 4 (2012), S. 146-168.

Humpert, Andreas, Erfahrungen mit Personennamen zur Bildung von Stichproben für Betriebsbefragungen, in: ZUMA-Nachrichten 54 (2004), S. 141-153.

Humpert, Andreas/Schneiderheinze, Klaus, Stichprobenziehung für telefonische Zuwandererumfragen – Praktische Erfahrungen und Erweiterung der Auswahlgrundlage, in: Gabler/Häder (Hrsg.), Telefonstichproben – Methodische Innovationen und Anwendungen in Deutschland, Münster 2002, S. 187-208.

Humpert, Andreas/Schneiderheinze, Klaus, Stichprobenziehung für telefonische Zuwandererumfragen – Einsatzmöglichkeiten der Namenforschung (Onomastik), in: ZUMA-Nachrichten 47 (2000), S. 36-63.

Oebbecke, Janbernd, Juristenausbildung zwischen Staat und Hochschule, in: JR 2003, S. 397-400.

Simon, Dan, A third view of the black box: cognitive coherence in legal decision making, in: University of Chicago Law Review 71 (2004), S. 511-586.

Spencer, Steven J./Steele, Claude M./Quinn, Diane M., Stereotype threat and women's math performance, in: Journal of Experimental Social Psychology 35 (1999), S. 4-28.

Steele, Claude M., A threat in the air: how stereotypes shape intellectual identity and performance, in: American Psychologist 52 (1997), S. 613-629.

Glöckner, Andreas/Towfigh, Emanuel/Traxler, Christian/, Development of Legal Expertise, in: Instructional Science 41 (2013), S. 989-1007.