Wissenschaftliche Beiträge

Geschlechts- und Herkunftseffekte bei der Benotung juristischer Staatsprüfungen

*Emanuel V. Towfigh / Christian Traxler / Andreas Glöckner**

In kaum einer anderen akademischen Disziplin sind Unterschiede in der Benotung der Examina von zentralerer Bedeutung für die Karrieremöglichkeiten und den Karriereerfolg der Absolventinnen und Absolventen wie in der Rechtswissenschaft. So wurde 2015 in einer empirischen Studie nachgewiesen, dass bereits kurz nach dem Studium ein erheblicher Gehaltsunterschied (14%) zwischen Personen mit einem Prädikatsexamen und solchen, die kein Prädikat erreichen konnten, besteht.[1] Die Autoren führen dies — unter Berücksichtigung verschiedener Kontrollfaktoren — sowohl auf die Signalwirkung des Prädikats wie auch auf die damit verbundenen Karrieremöglichkeiten zurück. Die akademische Ausbildung an juristischen Fakultäten, die Prüfungsvorbereitung sowie die Rahmenbedingungen und Bewertungsprozesse in den Prüfungen sind auch deshalb kontinuierlich dahingehend zu überprüfen, ob sie den Ansprüchen an eine objektive und faire Ausbildung und Benotung entsprechen. Gerade das föderale System mit der Zuständigkeit der Länder für die juristischen Prüfungen wird dadurch herausgefordert, gleichwertige Bedingungen für Absolventinnen und Absolventen zu schaffen (vgl. § 5 d Abs. 1 S. 2 DRiG). Dass dies bislang nur unzureichend gelingt, zeigt auch eine aktuelle Untersuchung von *Kähler, Engel* und *Ritter,* die anhand einer Analyse der Examensergebnisse von Referendarinnen und Referendaren, welche das Bundesland gewechselt hatten, herausfanden, dass es bei den untersuchten zehn Bundesländern Unterschiede beim Schwierigkeitsgrad der zweiten juristischen Staatsprüfung gibt.[2] In einem Umfeld, das einen besonderen Fokus auf die formale Qualifikation legt, können diese Unterschiede Auswirkungen auf das berufliche Fortkommen haben. Ist aber die konkrete Examensnote dermaßen entscheidend für Karriere und Einkommen, so sollte das auf diese Note hinführende Examen in besonderem Maße objektiver Gradmesser der tatsächlichen Fähigkeiten von Absolventinnen und Absolventen sein und gewährleisten, das gleiche Voraussetzungen auch zu gleichen Chancen auf gute Noten führen.

* Prof. Dr. *Emanuel V. Towfigh* ist Inhaber des Lehrstuhls für Öffentliches Recht, Empirische Rechtsforschung und Rechtsökonomik an der EBS Universität Law School, Wiesbaden; Prof. Dr. rer. pol. *Christian Traxler* ist Professor für Ökonomie an der Hertie School of Governance, Berlin; Prof. Dr. phil. *Andreas Glöckner* ist Professor für Psychologie an der FernUniversität in Hagen. — Die Autoren danken Dr. *Katharina Towfigh* für hilfreiche Anmerkungen zum Manuskript. — Für weitere Einzelheiten und Details der hier im folgenden dargestellten Befunde sei verwiesen auf den Abschlussbericht „Empirische Untersuchung zur Benotung in der staatlichen Pflichtfachprüfung und in der zweiten juristischen Staatsprüfung in Nordrhein-Westfalen von 2006 bis 2016“, abrufbar über die Website des Ministeriums der Justiz NRW www.justiz.nrw.de.

1 *Freier/Schumann/Siedler*, in: Labour Economics 34(C) (2015), S. 39 ff.

2 *Kähler/Engel/Ritter*, in: ZfRS 2017, S. 133 ff.

Nicht nur deshalb lag eine Überprüfung der 2014 in der ZDRW veröffentlichten empirischen Befunde zur Benotung in der Examensvorbereitung und im ersten Examen[3] auf Basis einer breiteren Datengrundlage, weiterer und besserer Kontrollvariablen sowie unter Berücksichtigung nicht nur des ersten, sondern auch des zweiten Examens nahe. Das Ministerium der Justiz des Landes Nordrhein-Westfalen, welches insbesondere im Nachgang zur oben genannten Studie und der Debatte, die sich politisch und medial angeschlossen hatte,[4] ein Interesse an weiterer Sachverhaltsaufklärung hatte, stellte einen entsprechenden Datensatz mit rund 18.000 Ergebnissen der ersten und zweiten juristischen Staatsprüfungen in NRW aus den Jahren 2006 bis 2016 zur Verfügung. Im Fokus stand die Frage, ob sich in den Prüfungen (systematische) Geschlechts- und Herkunftseffekte zeigen, ob sich dafür Ursachen benennen lassen (insbesondere ob sich bewusste oder unbewusste Diskriminierung als Ursache belegen oder ausschließen lässt) und ob sich den empirischen Beobachtungen Anhaltspunkte für Ansätze zur Verbesserung der Prüfungen entnehmen lassen, mit denen das Ministerium ggf. bestehenden strukturellen Benachteiligungen bestimmter Gruppen im Prüfungsverfahren aktiv entgegenwirken könnte.

A. Geschlechts- und Herkunftseffekte in der empirischen Forschung

I. Studien mit juristischem Fokus

In unserer oben erwähnten, 2014 an dieser Stelle veröffentlichten empirischen Studie zur Benotung in der Examensvorbereitung und im ersten Examen hatte sich u.a. gezeigt, dass Frauen und Kandidaten, deren Namen auf einen Migrationshintergrund schließen lässt, im Examen systematisch schlechter abschneiden.

Unter Nutzung von Daten aus der staatlichen Pflichtfachprüfung der ersten juristischen Staatsprüfung im Bezirk des Oberlandesgerichts Hamm für den Zeitraum 2007 – 2010 konnten wir zeigen, dass Frauen – gemessen an der Abiturnote – mit besseren Voraussetzungen in das Studium starten, am Ende aber mit einer um 0,7 Punkte schlechteren Examensnote abschließen (ca. 10% der Gesamtnote).[5] Dieser Unterschied ist in der mündlichen Prüfung ausgeprägter als in der anonymisierten schriftlichen Prüfung. So bleibt eine Differenz von 0,24 Punkten in der mündlichen

3 *Towfigh/Traxler/Glöckner*, in: ZDRW 2014, S. 8 ff.

4 LT-Drucks. NRW 16/6657; LT-Drucks. NRW 16/6922; LT-Drucks. Niedersachsen 17/3571; LT-Drucks. Bayern 17/5972; *Baron von Lijnden,* Benotung von Übungsklausuren und Staatsexamen, Frauen und Migranten im Nachteil, Freischüssler vorn, in: Legal Tribune Online, 16. 04. 2014, http://www.lto.de/persistent/a_id/11720/ (09.04.2018); *Lüpke-Narberhaus,* Im Zweifel für den Mann, in: Spiegel Online, 14.04.2014, http://www.spiegel.de/lebenundlernen/uni/jura-examen-fraue n-und-auslaender-schneiden-schlechter-ab-a-963081.html (09.04.2018); *Senol,* „Hier liegt es nahe, eine Diskriminierung anzunehmen", in: MiGAZIN, 04.04.2014, http://www.migazin.de/2014/04/04 /hier-liegt-es-nahe-eine-diskriminierung-anzunehmen/ (09.04.2018); *Martenstein,* Über Frauen, Juristen und Chihuahuas, ZeitMagazin Nr. 20/2014; dpa-Meldung v. 14.04.2014; *Haunhorst,* Sexistische Juristerei? in: Jetzt.de, 15.04.2014, https://www.jetzt.de/redaktionsblog/sexistische-juristerei-586215 (09.04.2018); Diskriminierung im Jura-Staatsexamen – Kampf gegen die Mauern in den Köpfen muss vorangehen, Presseerklärung der AG Migration und Vielfalt in der SPD v. 15.04.2014.

5 *Towfigh/Traxler/Glöckner*, in: ZDRW 2014, S. 8 ff.

Prüfung bestehen, selbst wenn man zusätzlich für die schriftliche Note kontrolliert, d.h. selbst wenn man berücksichtigt, dass Frauen schon in der schriftlichen Prüfung systematisch schlechter abschneiden, verschlechtern sie sich in (und aufgrund) der mündlichen Prüfung noch einmal. Anders ausgedrückt: Vergliche man eine Examenskandidatin und einen Examenskandidaten mit gleichem Abitur und gleichen schriftlichen Noten, so schnitt die Frau im Vergleich zu ihrem männlichen Kollegen um durchschnittlich 0,24 Punkte schlechter ab.

Bei diesen Geschlechtseffekten handelt es sich statistisch-methodisch um Korrelationen, d.h. es ist festzustellen, dass schlechtere Noten überzufällig häufig (oder systematisch) gemeinsam mit dem Merkmal „weiblich" auftreten. Zwar lassen sich für ein schlechteres Abschneiden von Frauen belastbare kausale Aussagen kaum treffen. Wenn man aber davon ausgeht, dass Talent und Fähigkeit grundsätzlich gleichverteilt sind, sollte dies auch über das Differenzierungsmerkmal „Geschlecht" hinweg gelten, d. h., man sollte bei Berücksichtigung des Merkmals eigentlich keine Unterschiede feststellen können. Lassen sich Unterschiede feststellen, so erscheint also naheliegend, dass (bewusste oder — wahrscheinlicher — unbewusste) Diskriminierung vorliegt, und zwar entweder auf individueller Ebene (etwa seitens der Prüferinnen oder Prüfer) oder strukturell (d.h. im Ausbildungssystem). Naheliegend erscheint ferner, dass diskriminierende Strukturen schon sehr viel früher als in der Prüfungssituation selbst wirksam sind, also etwa schon in der Schule oder an der Universität, und dass diese dann in der Prüfungssituation fortwirken.

Anders als etwa *Hinz* und *Röhl* formulieren, kann auf Grundlage der Befunde Diskriminierung mitnichten kategorisch ausgeschlossen werden. Die Erklärungsversuche der beiden Autoren für das unterschiedliche Abschneiden von Männern und Frauen sind wohl eher Ausdruck einer Form struktureller Diskriminierung (auch an den Universitäten) und stützen damit eher die Annahme einer Diskriminierung.[6]

Ähnliche Unterschiede bei den Prüfungsergebnissen zeigten sich zwischen Personen, deren Name (aufgrund einer onomastischen Analyse) auf einen Migrationshintergrund schließen ließ, im Vergleich zu Personen mit untechnisch gesprochen „traditionell deutschen" Namen. Trotz vergleichbarer Abiturleistungen zeigte sich ein Notenunterschied von 0,73 Punkten. Der Unterschied in der mündlichen Note war wiederum stärker als in der schriftlichen, und es blieb ein Unterschied von bis zu 0,43 Punkten bestehen, wenn zusätzlich für die in der schriftlichen Note bereits anonym abgefragten juristischen Kenntnisse und Fähigkeiten kontrolliert wurde. Dieser Unterschied zwischen Personen mit einem Namen, der auf einen Migrati-

6 *Hinz/Röhl*, in: JZ 2016, S. 874 (879). — Mit Blick auf die Frage, inwiefern man bei handschriftlichen Klausuren davon ausgehen kann, dass sie Rückschlüsse etwa auf Geschlecht und Herkunft nicht zulassen (wie *Hinz* und *Röhl* vertreten), verweisen wir auf unsere Ausführungen in der vorangegangenen Studie. Wir sind nach wie vor der Auffassung (und fühlen uns aus zahllosen Gesprächen mit Kollegen in unserer Auffassung bestätigt), dass die Handschrift in gewissem Umfang Rückschlüsse auf Geschlecht und Herkunft zulässt. Unser Vorschlag, auch diese Frage wissenschaftlich zu untersuchen, wurde vom Ministerium der Justiz NRW leider nicht aufgegriffen, so dass hier belastbare Evidenz nach wie vor fehlt.

onshintergrund hindeutet, wurde in einer 2016 erschienenen Studie für eine Stichprobe aus Baden-Württemberg zunächst bestätigt.[7] Allerdings verschwand dort der Unterschied in der mündlichen Note bei Kontrolle für die schriftliche Note und unter Berücksichtigung weiterer Kontrollvariablen, die den sozioökonomischen Status greifen sollten. Da sozioökonomischer Status und Migrationshintergrund üblicherweise korreliert sind, deutet letzteres Ergebnis tendenziell auf eine Überschätzung des reinen Effekts eines Migrationshintergrunds hin. Die baden-württembergischen Befunde sollten zum Anlass genommen werden, sie mit überzeugenden Indikatoren für den sozioökonomischen Hintergrund empirisch zu überprüfen.

II. Studien mit nicht-juristischem Fokus

Eine weitere aktuelle Studie, die sich mit Geschlechtseffekten in Prüfungsverfahren befasst, stammt aus Frankreich: In einer Analyse von über 100.000 Benotungen einer staatlichen Prüfung, auf die bei der Auswahl fast aller französischen Lehrer der Sekundar- und Oberstufe sowie von Professoren zurückgegriffen wird, konnten *Breda* und *Hillion* Effekte einer differenziellen Geschlechterdiskriminierung nachweisen.[8] Der Vergleich der anonymen schriftlichen Noten mit den naturgemäß nicht anonymen mündlichen Bewertungen ergab, dass die Personengruppe, die in einem Fach unterrepräsentiert war, bei Bekanntheit des Geschlechts jeweils tendenziell bessere Noten erhielt, als Personen des anderen Geschlechts. So zeigt die Analyse über elf Fächer u.a., dass Frauen in den Fächern Mathematik, Physik und Philosophie in der mündlichen Prüfung durchschnittlich um 10 %-Rangplätze besser eingeschätzt wurden als in der anonymen schriftlichen Prüfung. Dies indiziert eine *positive* Diskriminierung von Frauen in Fächern, in denen weniger Frauen als Männer tätig sind. Der umgekehrte Effekt ließ sich für die Fächer Literatur und Fremdsprachen nachweisen, in denen Männer in den mündlichen Prüfungen um ca. 3 – 5 % Rangplätze besser abschnitten als in den anonymen schriftlichen Prüfungen.

Auch trotz des Umstandes, dass es sich um eine staatliche Prüfung handelt, ist unklar, ob sich diese Befunde über Fächergrenzen, Länder und Prüfungsformate hinweg übertragen lassen. Die Studie ist jedoch in verschiedener Hinsicht bemerkenswert: So zeigt sie (a), dass auch unter bestmöglicher statistischer Kontrolle eine Ungleichbehandlung von Geschlechtern in staatlichen Examina nachgewiesen werden kann; dass es sich (b) nicht immer um eine Benachteiligung von Frauen handeln muss, sondern dass (c) das übergeordnete Ziel der Gleichbehandlung paradoxerweise vielleicht gerade dadurch verletzt wird, dass Prüferinnen und Prüfer — wahrscheinlich unbewusst und wohl vor allem bei der Benotung in mündlichen Prüfungen — von ihnen wahrgenommene Nachteile auszugleichen versuchen.[9]

7 *Hinz/Röhl*, in: VBlBW 2016, S. 20 ff.

8 *Breda/Hillion*, in: Science 353 (2016), S. 474 ff.

9 Vgl. dazu auch *Glöckner/Towfigh*, in: AnwBl 2016, S. 706 ff.

B. Datensatz

Für die neue, im Folgenden vorgestellte Studie bildet ein vom Landesjustizprüfungsamt NRW (LJPA) anonymisiert bereitgestellter Datensatz aller elektronisch erfassten Noten der zweiten juristischen Staatsprüfung der Abschlussjahrgänge 2006 bis 2016 in NRW die Grundlage der Analyse. Vor der Anonymisierung im LJPA wurde der Datensatz, wie in der vorangegangenen Studie, einer onomastischen Analyse unterzogen,[10] mit deren Hilfe eine differenzierte Kodierung der Vor- und Nachnamen nach deren Herkunftsregionen erfolgte. Der Datensatz deckt dabei mehrere (auch erfolglose) Prüfungsversuche ab und beinhaltet teilweise – nämlich dann, wenn beide Examina in NRW abgelegt wurden – die Noten der ersten juristischen Staatsprüfung. Insgesamt umfasst der Datensatz Noten von 19.883 Personen, die 26.342 Prüfungsversuche im zweiten Examen durchliefen. 5.208 Prüfungsversuche wurden nicht bestanden.

Die Versuche enthalten reguläre Versuche, Wiederholungsversuche sowie Notenverbesserungsversuche. Die zentralen empirischen Analysen für das zweite Examen fokussieren auf den für die Forschungsfrage „relevanten" Versuch, d.h. entweder den direkt bestandenen Versuch oder — falls dadurch eine Notenverbesserung erzielt wurde — den relevanten Wiederholungs- bzw. Verbesserungsversuch. Dadurch wird sichergestellt, dass für jede Person nur ein Versuch analysiert und eine Doppelbetrachtung bzw. eine höhere Gewichtung von Wiederholungs-/Verbesserungsversuchen vermieden wird. Die im Folgenden besprochenen Effekte sind jedoch robust und tendenziell stärker ausgeprägt, wenn Wiederholungen und freiwillige Verbesserungsversuche ausgeschlossen werden, d.h., wenn nur die jeweils ersten Versuche untersucht werden.

Die Analyse erfolgte vorwiegend unter Nutzung linearer Regressionsmodelle, in denen überprüft wird, ob unter Berücksichtigung verschiedener Kontrollvariablen der Effekt eines Faktors wie bspw. Geschlecht oder Migrationshintergrund auf die Bewertung „statistisch bedeutsam" (d.h. statistisch signifikant unterschiedlich von einem Null-Effekt) bleibt. Das Signifikanzniveau gibt dabei an, wie (un)wahrscheinlich es ist, dass ein solch starker Einfluss beobachtet werden könnte, obwohl tatsächlich kein Unterschied besteht (das Ergebnis gleichsam „zufällig", also ohne eine zugrundeliegende Regelmäßigkeit zustande kommen könnte). Nach einer verbreiteten Konvention spricht man bei Wahrscheinlichkeiten von unter 1%, unter 5% bzw. unter 10% von einem hoch-signifikanten, einem signifikanten respektive einem schwach- oder marginal-signifikanten Effekt. Die Nutzung sog. Kontrollvariablen erlaubt es dabei, konzeptionelle Vergleichsgruppen zu erzeugen — also Personen, die in den gemessenen Dimensionen (wie bspw. Alter, der Abiturnote oder

10 Durchgeführt durch die Firma Humpert & Schneiderheinze Sozial- und Umfrageforschung, siehe zu den methodischen Grundlagen die Ausführungen in der vorangegangenen Studie (ZDRW 2014, S. 9 und Fn. 3).

dem Zeitpunkt der Prüfung) vergleichbar sind, weil der partielle Einfluss dieser Variablen statistisch „kontrolliert" wird.

Tabelle 1

Variable	Mittelwert	St.Abw.	Min	Max	N
A. Erstes Examen					
Gesamtnote erste Prüfung	7,772	1,985	4,000	15,940	10.042
• staatliche Pflichtfachprüfung (70%)	7,315	1,946	3,700	14,200	4.390
• universitäre Prüfung (30%)	9,146	2,355	0,400	16,833	4.351
B. Zweites Examen					
Gesamtnote zweites Staatsexamen	7,495	1,993	3,270	15,020	17.971
Note schriftliche Prüfungsteile	6,169	2,105	0,250	14,625	18.958
Note mündliche Prüfungsteile	9,200	2,439	2,000	17,750	17.970
Note Zivilrecht	6,164	2,389	0,500	15,250	9.696
Note Strafrecht	5,811	2,465	0,000	15,000	9.696
Note öffentliches Recht	6,077	2,652	0,000	16,500	9.696
C. Kontrollvariablen					
Abiturnote	2,263	0,605	0,900	3,900	4.592
Alter	30,026	2,914	21,052	67,584	19.865
Frauen	0,529	0,499	0	1	19.865
Staatsangehörigkeit (nicht-deutsch)	0,028	0,166	0	1	9.784
Geburtsort (nicht-deutsch)	0,091	0,290	0	1	9.784
Onomastik (nicht-deutsch)	0,138	0,345	0	1	9.784
Migrationsindikatoren: 1. Generation	0,061	0,240	0	1	9.784
Migrationsindikatoren: ≥ 2. Generation	0,079	0,270	0	1	9.784
Frau in Prüfungskommission	0,354	0,478	0	1	17.970

Die zentralen Analysen basieren auf maximal 17.971 beobachteten Gesamtnoten (von ebenso vielen Personen; siehe *Tabelle 1 Panel B*). Für die schriftlichen Teilnoten liegen maximal 18.958 Beobachtungen vor.[11] Teilnoten für die Rechtgebiete Öffentliches Recht, Strafrecht und Zivilrecht (im zweiten Examen) liegen für maximal 9.696 Fälle vor. Für die Analysen der Noten aus dem ersten Examen stehen für 10.042 Kandidatinnen und Kandidaten – nahezu allen Personen, die zwischen

11 Die im Vergleich zu den Gesamtnoten höhere Beobachtungszahl ist damit zu erklären, dass hier einige Kandidaten (vor allem in Fällen mit Noten unterhalb der 3,5-Punkte-Schwelle) im Beobachtungszeitraum nicht zur mündlichen Prüfung angetreten sind, so dass keine Gesamtnote vorliegt.

dem Wintersemester 2010 und dem Sommersemester 2016 einen Prüfungsversuch im zweiten Examen unternommen haben – die Gesamtnoten auch der ersten Prüfung zur Verfügung (*Tabelle 1 Panel A*). Separate Daten für den staatlichen und universitären Teil des ersten Examens liegen für 4.390 bzw. 4.354 Personen vor. Die basalen Informationen zu den wichtigsten Kontroll- und Vorhersagevariablen sind in *Tabelle 1 Panel C* zusammengefasst.

Für alle Beobachtungen, bei denen Information über die mündlichen Prüfungsteile im zweiten Examen vorliegen, wird auch verzeichnet, ob zumindest eine Frau Teil der Prüfungskommission war; in 35,4% der mündlichen Prüfungen war dies der Fall.

Für etwa 10.000 Beobachtungen liegen zusätzlich Informationen zur Staatsangehörigkeit und zum Geburtsort vor: 2,8% der Kandidatinnen und Kandidaten haben eine nicht-deutsche Staatsangehörigkeit und 9,1% wurden im Ausland geboren. Die onomastische Analyse zeigt auf, dass die Namen von 13,8% der Kandidatinnen und Kandidaten auf einen nicht-deutschen Ursprung schließen lassen.[12] Unter Berücksichtigung aller drei Indikatoren kann hier eine weitere Differenzierung vorgenommen werden. In unserer Analyse vergleichen wir jene *Migranten* („erster Generation"), die im Ausland geboren wurden und über keine deutsche Staatsbürgerschaft verfügen (etwa 6%), und jene, die zwar in Deutschland geboren wurden, aber entweder einen Namen ausländischen Ursprungs oder eine nicht-deutsche Staatsbürgerschaft haben. Etwa 8% verfügen über einen solchen *Migrationshintergrund* der „zweiten oder späteren Generation".[13]

C. Forschungsfragen

Inhaltlich knüpfen die Forschungsfragen der zweiten Studie unmittelbar an die Ergebnisse der Vorgängerstudie an und befassen sich zum einen mit der Frage, ob die beobachteten Unterschiede für Geschlecht und Migrationshintergrund auch bei einer durch den größeren Datensatz nun möglichen differenzierteren Betrachtung des ersten Examens und im erstmalig untersuchten zweiten Examen bestehen bleiben. Zum anderen wurden die bereits in der ersten Studie beobachteten Diskontinuitäten in der Notenverteilung rund um die Notenschwellen (mangelhaft/ausreichend, ausreichend/befriedigend, befriedigend/vollbefriedigend, vollbefriedigend/gut, gut/sehr gut) einer genaueren Analyse unterzogen und dabei auch speziell die Unterschiede in der *Wahrscheinlichkeit* für Personen unterschiedlicher Personengruppen (insbesondere Frauen und Migranten) mit ansonsten gleichen Vorausset-

12 Der Vergleich mit den zwei „härteren" Indikatoren für einen Migrationshintergrund (Geburtsort und Staatsangehörigkeit) zeigt dabei, dass der Onomastik-Indikator eine sehr hohe Treffgenauigkeit aufweist.

13 Die Summe aus den so gebildeten Indikatoren für Migrationshintergrund der ersten oder höheren Generationen ergibt nicht den ausschließlich Onomastik-basierten Indikator. Dies liegt daran, dass bei der Definition der generationsspezifischen Migrationsindikatoren neben den Onomastik-Werten auch weitere Variablen (Staatsangehörigkeit, Geburtsort) genutzt werden, sodass hier eine marginal breitere Definition entsteht.

zungen, die nächste Stufe zu erreichen, analysiert. Dabei wurde auch der mögliche Einfluss der Zusammensetzung von Prüfungskommissionen, vor allem die Frage, ob sich ein potenzieller Geschlechtseffekt bei Kommissionen mit Beteiligung von Frauen verringert, untersucht.[14]

D. Ergebnisse

I. Geschlechtseffekte

1. Erstes Examen

Auch bei differenzierter Betrachtung der umfangreicheren Datengrundlage zeigt sich für das erste Examen ein substantieller negativer Effekt in der Gesamtbenotung von Frauen im Vergleich zu Männern. Die Analyse der Abschlussnote des ersten Examens liefert ein klares Ergebnis: Frauen erzielen durchschnittlich um 0,29 Notenpunkte (entspricht 3,6%) schlechtere Note als Männer (*Tabelle 2, Modell 1*). Der Unterschied steigt auf 0,38 Punkte, wenn für die Abiturnote kontrolliert wird (*Modell 2*).

Tabelle 2: Analyse von Geschlechterunterschieden in der ersten Prüfung

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variable	Gesamtnote (erste Prüfung)			staatliche Pflichtfachprüfung		universitäre Prüfung	
Frauen	-0,287***	-0,377***	0,031*	-0,225***	-0,510***	0,147**	-0,161**
	[0,000]	[0,000]	[0,090]	[0,000]	[0,000]	[0,044]	[0,018]
Abiturnote		-1,568***	-0,259***		-1,554***		-1,675***
		[0,000]	[0,000]		[0,000]		[0,000]
Note staatl.			0,857***				
Pflichtfachpr			[0,000]				
Konstante	7,927***	11,590***	2,164***	7,438***	11,112***	9,058***	13,017***
	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]
N	10.042	4.596	4.252	4.390	4.254	4.354	4.252
R^2	0,005	0,261	0,898	0,003	0,231	0,001	0,179

Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1

Neu ist die Erkenntnis, dass dieser Notenunterschied hauptsächlich durch den staatlichen Teil der Prüfung getrieben wird. Im universitären Teil der Prüfungen fallen die Geschlechtsunterschiede qualitativ anders aus. Kontrolliert man nicht für das Abiturergebnis, so erhalten Frauen hier sogar um 0,15 Punkte *bessere* Noten als Männer (*Modell 6*). Dieser Unterschied kehrt sich jedoch wieder um, wenn für

14 Eine Analyse des vergleichbaren Effekts für Kommissionen mit Beteiligung von Personen mit Migrationshintergrund konnte aufgrund der sehr geringen Anzahl von Prüferinnen und Prüfern mit Migrationshintergrund nicht erfolgen.

die Abiturnote kontrolliert wird: Nach Berücksichtigung dieser Variablen schneiden Frauen im universitären Prüfungsteil durchschnittlich 0,16 Punkte schlechter ab als Männer (*Modell 7*). Unter Berücksichtigung der Abiturnote finden wir also auch im universitären Teil überzufällig schlechtere Noten von Frauen – wenn auch in deutlich geringerem Ausmaß als in der staatlichen Pflichtfachprüfung. Bei Kontrolle für die Note im staatlichen Teil sowie für die Abiturnote erzielen Frauen sogar marginal (0,03 Notenpunkte) bessere Gesamtnoten als Männer. D.h., dass die Geschlechtsunterschiede in der Gesamtnote vorwiegend in der staatlichen Pflichtfachprüfung entstehen und im universitären Teil nicht weiter verstärkt, sondern sogar marginal abgeschwächt werden.

2. Zweites Examen

Zum Ersten Mal wurde anhand der verfügbaren Daten auch das zweite Examen in den Blick genommen.

Tabelle 3: Geschlechtsunterschiede im zweiten Examen: Gesamtnote

	(1)	(2)	(3)	(4)
Variable:	Gesamtnote (zweites Examen)			
Frauen	-0,143***	-0,446***	-0,060**	-0,070*
	[0,000]	[0,000]	[0,033]	[0,074]
Gesamtnote			0,729***	0,695***
erstes Examen			[0,000]	[0,000]
Abiturnote				-0,257***
				[0,000]
Konstante	7,571***	23,569***	9,289***	7,515***
	[0,000]	[0,000]	[0,000]	[0,000]
Weitere Kontrollvariablen	Nein	Ja	Ja	Ja
N	17.971	17.971	9.086	4.251
R^2	0,001	0,095	0,582	0,597

Modelle (2) – (4) kontrollieren für das Alter zum Prüfungszeitpunkt (linearer und quadratischer Term) sowie für Abschlussmonat-spezifische Effekte. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

a. Gesamtnote

Die Analyse der Gesamtnoten im zweiten Examen ergibt, dass Frauen um 0,14 Notenpunkte schlechtere Noten (1,9%) erzielen als Männer (Tabelle 3, Modell 1). Dieser Notenunterschied bleibt – in abgeschwächter Form – auch nach Kontrolle für die Note aus dem ersten Examen sowie der Abiturnote qualitativ bestehen. Allerdings variieren die Geschlechtsunterschiede in den Noten zwischen Referendarinnen und Referendaren, die an unterschiedlichen Universitäten studiert haben. Deutlich stärker ausgeprägt (und auch statistisch signifikant stärker als die Refe-

renzgruppe Köln) sind die Geschlechtsunterschiede unter Referendarinnen und Referendaren, die an der Universität Bonn studiert haben; deutlich weniger stark ausgeprägt (aber statistisch nicht signifikant unterschiedlich zur Referenzgruppe Köln) sind die Geschlechtsunterschiede von Studierenden aus Münster.[15]

b. Differenzierung nach mündlicher und schriftlicher Note

Separate Analysen der (durchschnittlichen) mündlichen und schriftlichen Noten zeigen, dass Geschlechtsunterschiede in beiden Prüfungsteilen auftreten. Der Unterschied beträgt 0,116 Notenpunkte (1,9%) in den schriftlichen Noten (*Tabelle 4, Modell* 1) und 0,225 Notenpunkte (2,4%) in den mündlichen Noten (*Tabelle 5, Modell 1*). Bei Kontrolle für das Alter der Prüflinge bzw. für zeitspezifische Faktoren verstärken sich diese Unterschiede deutlich (*Tabellen 4 und 5, Modell 2*).

Tabelle 4: Durchschnittliche schriftliche Note im zweiten Examen

	(1)	(2)	(3)	(4)
Frauen	-0,116***	-0,441***	-0,045	-0,024
	[0,000]	[0,000]	[0,124]	[0,574]
Note 1.Examen			0,750***	0,718***
			[0,000]	[0,000]
Abiturnote				-0,215***
				[0,000]
Weitere Kontrollvariablen	Nein	Ja	Ja	Ja
Konstante	6,230***	22,260***	9,148***	6,274***
	[0,000]	[0,000]	[0,000]	[0,000]
N	18.958	18.958	9.969	4.579
R^2	0,001	0,126	0,581	0,575

Modelle (2) – (4) kontrollieren für das Alter zum Prüfungszeitpunkt (linearer und quadratischer Term) sowie für Abschlussmonats-spezifische Effekte. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

Wird nun weiter für Vornoten (die Abiturnote und/oder die Noten aus dem ersten Examen) kontrolliert, so ist kein signifikanter Geschlechtsunterschied bei den schriftlichen Noten festzustellen (*Tabelle 4, Modelle 3 und 4*). Dies gilt jedoch nicht für die mündlichen Teilnoten. In den mündlichen Prüfungsteilen finden wir – selbst nach Kontrolle für die Vornoten aus den schriftlichen Teilen, aus dem ersten Examen oder aus dem Abitur – einen (gleichsam zusätzlichen) hoch-signifikanten Geschlechtsunterschied (*Tabelle 5, Modelle 4 bis 6*): Vergleicht man also „gleich gute" Referendarinnen und Referendare – gemessen anhand einer identischen Note

15 Siehe dazu Abschlussbericht für das Justizministerium, verfügbar über www.justiz.nrw.de.

im schriftlichen Prüfungsteil des zweiten Examens oder im ersten Examen – so erzielt eine Frau durchschnittlich eine um etwa 0,2 Notenpunkte schlechtere Note in den mündlichen Prüfungsteilen.

Tabelle 5: Durchschnittliche mündliche Note im zweiten Examen

	(1)	(2)	(3)	(4)	(5)	(6)
Frauen	-0,225***	-0,538***	-0,546***	-0,238***	-0,132***	-0,205***
	[0,000]	[0,000]	[0,000]	[0,000]	[0,006]	[0,003]
Frauen x FPK			-0,006	0,024	0,027	0,097
			[0,902]	[0,518]	[0,614]	[0,185]
Männer x FPK			-0,028	-0,023	0,064	0,021
			[0,613]	[0,579]	[0,270]	[0,801]
Note schriftl.				0,823***		
Prüfungsteile				[0,000]		
Note 1.Examen					0,778***	0,733***
					[0,000]	[0,000]
Abiturnote						-0,294***
						[0,000]
Konstante	9,319***	25,741***	25,753***	7,561***	9,298***	7,983***
	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]	[0,001]
Weitere Kontrollvariablen	Nein	Ja	Ja	Ja	Ja	Ja
N	17.970	17.970	17.969	17.969	9.084	4.251
R^2	0,002	0,078	0,078	0,487	0,449	0,450
F-test			0,767	0,395	0,634	0,482

Die Variable „FPK“ ist ein Indikator für die Beteiligung von mindestens einer Frau in der Prüfungskommission. In der Zeile F-test werden p-Werte für Tests der Null-Hypothese „Frauen x FPK = Männer x FPK“ (also der Hypothese, dass die Präsenz einer Frau in der Prüfungskommission den gleichen Effekt auf Männer und Frauen hat) berichtet. Modelle (2) – (6) kontrollieren für das Alter zum Prüfungszeitpunkt (linearer und quadratischer Term) sowie für Abschlussmonatsspezifische Effekte. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

c. Zusammensetzung der Prüfungskommission

Eine Analyse der durchschnittlichen Noten in der mündlichen Prüfung zeigt hinsichtlich der Beteiligung von Frauen in der Prüfungskommission keinen Effekt: Ob zumindest eine Prüferin Teil der Kommission ist, hat keinen Einfluss auf die durchschnittliche Note im mündlichen Prüfungsteil – weder für die Note von weiblichen

noch für jene von männlichen Studierenden. Das bedeutet, dass der Geschlechtsunterschied in der durchschnittlichen mündlichen Note nicht durch die Zusammensetzung der Prüfungskommission beeinflusst zu werden scheint (*Tabelle 6, Modelle 3-6*).

Allerdings verbirgt der Fokus auf die durchschnittlichen mündlichen Noten einen systematischen Effekt, der erst bei einer differenzierteren Analyse sichtbar wird. Eine Betrachtung aller Kandidatinnen und Kandidaten die nach den schriftlichen Noten in einem hinreichend engen Bereich um eine der relevanten Notenstufen liegen, zeigt, dass sich die Wahrscheinlichkeit, die nächsthöhere Notenstufe zu erreichen, dann verändert, wenn eine Frau in der Prüfungskommission ist (s. unten D.III.3.).

II. Herkunftseffekte

Zur Analyse der Effekte eines potenziellen Migrationshintergrunds auf die Notengebung verwenden wir als direkte und indirekte Indikatoren einen nicht-deutschen Geburtsort, das Nicht-Vorhandensein einer deutschen Staatsangehörigkeit, sowie eine nicht-deutsche Herkunftszuordnung der Namen auf Basis eines Onomastikverfahrens. Unter Berücksichtigung aller drei Indikatoren differenzieren wir auch Personen mit Migrationserfahrung aus der ersten bzw. der zweiten oder späteren Generationen (s. oben B.). Da die primären Indikatoren für die Migrationsanalyse nicht für alle Personen vorhanden sind, reduziert sich die Größe der verfügbaren Stichprobe für die Hauptanalysen des zweiten Examens auf ca. 8.700 relevante Beobachtungen. Der Fokus der Analysen liegt wiederum auf der finalen Note, ggf. nach Wiederholungs- und Verbesserungsversuch.

1. Erstes Examen

Für alle drei Migrations-Indikatoren zeigt sich ein starker Einfluss auf die Gesamtnote im ersten Examen. Die Effekte der „harten" direkten Migrations-Indikatoren Geburtsort und Staatsangehörigkeit bleiben bei Kontrolle für Abiturnote und Geschlecht bestehen. Für Migrantinnen und Migranten der ersten Generation (Geburt im Ausland) sind die Unterschiede deutlich stärker und robuster. Ein durchschnittlicher „deutscher" Prüfling (d.h. mit Null-Werten in allen drei Indikatoren) erzielt im untersuchten Datensatz im Durchschnitt eine Gesamtnote von 7,93 Punkten. Studierende mit einer nicht-deutschen Namensherkunft, einem nicht-deutschen Geburtsort und ohne deutsche Staatsangehörigkeit erreichen im Durchschnitt lediglich eine Gesamtnote von 6,51. Die letztere Gruppe erzielt im ersten Examen damit eine um etwa 18% (1,42 Punkte) schlechtere Abschlussnote. Für Studierende mit Migrationshintergrund, die bereits in Deutschland geboren wurden, sind die Notenunterschiede geringer. Die Evidenz deutet daher auf einen positiven integrativen Effekt hin, der zu einer Verminderung der Notennachteile zu führen scheint.[16]

16 Siehe dazu Abschlussbericht für das Justizministerium, verfügbar über www.justiz.nrw.de.

2. Zweites Examen

a. Gesamtnote

Analysen der Gesamtnoten im zweiten Examen zeigen für die direkten Indikatoren, dass die durchschnittliche Gesamtnote im zweiten Examen bei nicht-deutscher Namensherkunft um 0,82 Notenpunkte schlechter, bei einem nicht-deutschen Geburtsort um 0,49 Notenpunkte schlechter ist. Bezüglich der Staatsangehörigkeit zeigt sich kein statistisch signifikanter Unterschied für die Gesamtnote. In Summe ergibt sich aus den beiden hoch-signifikanten Indikatoren eine – im Vergleich zu „deutschen" Referendarinnen und Referendaren (Durchschnittsnote 7,74) – um bis zu 17% (1,31 Punkte) schlechtere Note. Die Effektgrößen gehen nur marginal zurück, wenn für Alter, Geschlecht und Abschlusszeitpunkt-spezifische Effekte kontrolliert wird (*Tabelle 6, Modell 2*). Wird auch für die Abschlussnote aus dem ersten Examen kontrolliert, so schrumpfen die Effekte erheblich (auf 0,32 – bei Blick auf die Namensherkunft bzw. 0,13 Notenpunkte – bei Blick auf den Geburtsort, *Modell 3*), bleiben jedoch statistisch signifikant. Vergleicht man also Referendarinnen und Referendare mit exakt gleichen Noten im ersten Examen, so erzielen jene mit einem nicht-deutschen Namen bzw. einem nicht-deutschen Geburtsort im zweiten Examen weiterhin systematisch schlechtere Noten.[17]

Trennt man nach unterschiedlichen Herkunftsregionen (*Tabelle 6, Modell 4*), so zeigt sich ein hohes Maß an Heterogenität innerhalb der Gruppe der „nicht-deutschen" Referendarinnen und Referendare. Für Referendarinnen und Referendare mit osteuropäischen Wurzeln (MI: Region 4) zeigt sich kein oder nur ein kleiner negativer Effekt, während für solche aus dem Nahen und Mittleren Osten (MI: Region 3) die größten Unterschiede festzustellen sind.[18] Bei der Interpretation dieser regionalen Herkunfts-Indikatoren ist jedoch außerordentliche Vorsicht angebracht, zumal auch innerhalb der einzelnen regionalen Gruppen ein hohes Maß an Heterogenität vorliegt und diese Subgruppen relativ klein sind. Abschließend wird noch für die Abiturnote kontrolliert (*Tabelle 6, Modell 5*). Damit verliert man zwar die Hälfte der Beobachtungen (und damit auch die Präzision, d.h., das Signifikanzniveau der geschätzten Koeffizienten), die relevanten Effektgrößen ändern sich aber kaum.

17 Gleichzeitig zeigt sich ein *positiver* Koeffizient für den dritten Indikator (nicht-deutsche Staatsangehörigkeit), der – auch aufgrund der hohen Korrelation mit den anderen beiden Indikatoren – nur schwer zu interpretieren und auch nur schwach-signifikant ist.

18 In dieser Spezifikation muss der Effekt aus den ersten drei Indikatoren (I1-I3) mit dem jeweiligen Herkunftsland-Indikator summiert werden, um einen gruppenspezifischen Effekt bewerten zu können.

Tabelle 6: Effekte der direkten Migrationsindikatoren auf die Gesamtnote im zweiten Examen

	(1)	(2)	(3)	(4)	(5)
I1: Onomastik	-0,819***	-0,678***	-0,315***	-0,454**	-0,431
	[0,000]	[0,000]	[0,000]	[0,012]	[0,136]
I2: Geburtsort	-0,487***	-0,299***	-0,134**	-0,238***	-0,282***
	[0,000]	[0,000]	[0,016]	[0,000]	[0,000]
I3: Staatsangeh.	-0,012	0,047	0,165*	0,188*	0,242*
	[0,923]	[0,703]	[0,077]	[0,056]	[0,062]
Note 1.Examen			0,723***	0,721***	0,689***
			[0,000]	[0,000]	[0,000]
Abiturnote					-0,263***
					[0,000]
MI: Region 1				0,378	0,546
				[0,219]	[0,268]
MI: Region 2				0,240	0,165
				[0,268]	[0,623]
MI: Region 3				-0,075	0,057
				[0,695]	[0,851]
MI: Region 4				0,420**	0,261
				[0,029]	[0,396]
MI: Region 5				0,052	0,098
				[0,793]	[0,747]
MI: Region 6				0,259	0,119
				[0,253]	[0,733]
Weitere Kontrollvariablen	Nein	Ja	Ja	Ja	Ja
Konstante	7,741***	22,156***	8,440***	8,404***	6,834***
	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]
N	8.757	8.757	8.504	8.474	4.239
R^2	0,032	0,117	0,584	0,585	0,601

Modelle (2) – (5) kontrollieren für das Geschlecht, Alter (linear und quadratischer Term) sowie für Abschlussmonat-spezifische Effekte. MI Regionen: (1) Afrika, (2) Süd/Ost- und Zentral-Asien, (3) West-Asien, (4) Ost-Europa, (5) Süd-Europa, (6) Andere/Unklare Herkunft, Referenzkategorie: Zentral- und Nord-Europa. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

b. Differenzierung nach schriftlicher und mündlicher Prüfung

Differenziert man in der Analyse nach schriftlicher und mündlicher Prüfung, so zeigt die Untersuchung der schriftlichen Noten konsistente und starke Effekte bei den Indikatoren nicht-deutsche Namensherkunft und nicht-deutscher Geburtsort. Die durchschnittliche schriftliche Note ist bei Referendarinnen und Referendaren mit nicht-deutscher Namensherkunft um 1,11 Notenpunkte, bei nicht-deutschem Geburtsort um 0,52 Notenpunkte schlechter. Selbst ohne Berücksichtigung des negativen (aber insignifikanten) Effektes des dritten Indikators (Staatsangehörigkeit) ergibt sich damit ein Notenunterschied von bis zu 1,63 Notenpunkten, was – relativ zur schriftlichen Durchschnittsnote von 6,25 Punkten bei „deutschen" Referendarinnen und Referendaren – einen Unterschied von 26% ausmacht. Unter Berücksichtigung von Kontrollvariablen gehen diese Unterschiede teilweise recht deutlich zurück, bleiben jedoch über alle Spezifikationen hinweg statistisch signifikant (*Tabelle 7, Modelle 2-5*).

Auch für die durchschnittliche Note aus den mündlichen Prüfungsteilen zeigt sich ein vergleichbares Bild. Auch hier weisen die beiden ersten Indikatoren wieder quantitativ starke, statistisch hoch-signifikante Notenunterschiede auf (*Tabelle 8, Modell 1*), die sich auf 1,45 Notenpunkte addieren. In absoluten Zahlen ist diese Effektgröße damit ähnlich zu jenem Notenunterschied, der für die schriftlichen Noten festgestellt wurde. Relativ zu der besseren mündlichen Durchschnittsnote (9,52 Punkte), beläuft sich der Unterschied jedoch auf „nur" 15% – und fällt damit deutlich geringer aus, als die 26%, die für die schriftlichen Noten gemessen wurden. Die Migrationseffekte spielen relativ betrachtet eine stärkere Rolle in der schriftlichen als in der mündlichen Prüfung, wobei zu berücksichtigen ist, dass das Gewicht der mündlichen Prüfung für die Gesamtnote relativ geringer ist.

Nach Kontrolle für alters-, geschlechts- und Abschlusszeitpunkt-spezifische Effekte gehen die Effekte etwas zurück, bleiben dabei aber hoch-signifikant (*Tabelle 8, Modell 2*). Wird auch für die schriftlichen Teilnoten kontrolliert (*Modell 3*), so sinken die Effekte weiter und der Indikator nicht-deutscher Geburtsort ist statistisch insignifikant. Gleichwohl bleibt der Effekt für den Onomastik-Indikator hoch-signifikant: Prüflinge aus dieser Gruppe schneiden in der mündlichen Prüfung noch einmal schlechter ab als in der schriftlichen, verschlechtern sich – relativ zu den Kolleginnen und Kollegen mit einem deutschen Namen – in der mündlichen Prüfung somit weiter. Ein ähnliches Bild ergibt sich, wenn anstatt der schriftlichen Teilnoten für die Note aus dem ersten Examen kontrolliert wird (*Modell 4*): in beiden Fällen behält aber der Indikator „nicht-deutsche Namensherkunft" einen hoch-signifikaten negativen Effekt.

Tabelle 7: Effekte der Migrationsindikatoren auf die durchschnittliche schriftliche Note im zweiten Examen

	(1)	(2)	(3)	(4)	(5)
I1: Onomastik	-1,106***	-0,832***	-0,334***	-0,459**	-0,476*
	[0,000]	[0,000]	[0,000]	[0,020]	[0,099]
I2: Geburtsort	-0,519***	-0,265***	-0,148***	-0,277***	-0,302***
	[0,000]	[0,000]	[0,009]	[0,000]	[0,000]
I3: Staatsangeh.	-0,031	-0,010	0,054	0,028	0,040
	[0,815]	[0,935]	[0,570]	[0,774]	[0,790]
Note 1.Examen			0,745***	0,743***	0,710***
			[0,000]	[0,000]	[0,000]
Abiturnote					-0,222***
					[0,000]
MI: Region 1				0,460	0,593
				[0,124]	[0,191]
MI: Region 2				0,207	0,147
				[0,356]	[0,655]
MI: Region 3				-0,161	0,034
				[0,431]	[0,912]
MI: Region 4				0,464**	0,292
				[0,025]	[0,341]
MI: Region 5				0,251	0,324
				[0,239]	[0,290]
MI: Region 6				0,202	0,101
				[0,390]	[0,769]
Weitere Kontrollvariablen	Nein	Ja	Ja	Ja	Ja
Konstante	6,251***	22,842***	8,423***	8,369***	5,799***
	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]
N	9.744	9.744	9.387	9.355	4.566
R^2	0,047	0,188	0,586	0,588	0,580

Modelle (2) – (5) kontrollieren für das Geschlecht, Alter (linear und quadratischer Term) sowie für Abschlussmonat-spezifische Effekte. MI Regionen: (1) Afrika, (2) Süd/Ost- und Zentral-Asien, (3) West-Asien, (4) Ost-Europa, (5) Süd-Europa, (6) Andere/Unklare Herkunft, Referenzkategorie: Zentral- und Nord-Europa. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

Tabelle 8: Effekte der Migrationsindikatoren auf die durchschnittliche mündliche Note im zweiten Examen

	(1)	(2)	(3)	(4)	(5)
I1: Onomastik	-0,952***	-0,807***	-0,326***	-0,405***	-0,273
	[0,000]	[0,000]	[0,000]	[0,000]	[0,288]
I2: Geburtsort	-0,495***	-0,282***	-0,032	-0,114	-0,218**
	[0,000]	[0,008]	[0,696]	[0,174]	[0,020]
I3: Staatsangeh.	0,026	0,096	0,084	0,207	0,230
	[0,879]	[0,566]	[0,533]	[0,127]	[0,115]
Schriftliche Note			0,810***		
(2.Examen)			[0,000]		
Note 1.Examen				0,770***	0,767***
				[0,000]	[0,000]
MI: Region 1					0,071
					[0,872]
MI: Region 2					0,093
					[0,770]
MI: Region 3					-0,355
					[0,198]
MI: Region 4					0,134
					[0,629]
MI: Region 5					-0,372
					[0,180]
MI: Region 6					0,078
					[0,813]
Weitere Kontrollvariablen	Nein	Ja	Ja	Ja	Ja
Konstante	9,523***	23,257***	5,918***	8,454***	8,464***
	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]
N	8.756	8.756	8.756	8.503	8.473
R^2	0,027	0,092	0,483	0,448	0,448

Modelle (2) – (5) kontrollieren für das Geschlecht, Alter (linear und quadratischer Term) sowie für Abschlussmonat-spezifische Effekte. MI Regionen: (1) Afrika, (2) Süd/Ost- und Zentral-Asien, (3) West-Asien, (4) Ost-Europa, (5) Süd-Europa, (6) Andere/Unklare Herkunft, Referenzkategorie: Zentral- und Nord-Europa. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

III. Erreichen der nächsthöheren Notenstufe

Die Analysen haben gezeigt, dass Geschlechts- und Herkunftseffekte bei den nicht anonymen mündlichen Prüfungen besonders stark ausgeprägt sind (und zusätzlich zu den Unterschieden in schriftlichen Noten und sonstigen Vornoten auftreten). Vor dem Hintergrund dieser klaren Evidenz stellt sich nun die Frage, was dies für die Gleichbehandlung von Prüflingen mit gleichen Voraussetzungen oder Vornoten bedeutet bzw. ob alle Referendarinnen und Referendare die gleiche Wahrscheinlichkeit haben, die nächste Notenstufe zu erreichen.

1. Bedeutung der Vornoten bei der Notenvergabe

Bei einer Betrachtung der Notenverteilung zeigen sich sowohl bei den Gesamtnoten für das erste Examen als auch für das zweite Examen (hier noch stärker ausgeprägt), deutliche Häufungen von Noten über den Notenschwellen und fehlender Beobachtungsmasse unter den jeweiligen Schwellenwerten.

Abbildung 1: Notenverteilung im zweiten Examen: Durchschnitt aus schriftlichen Teilnoten.

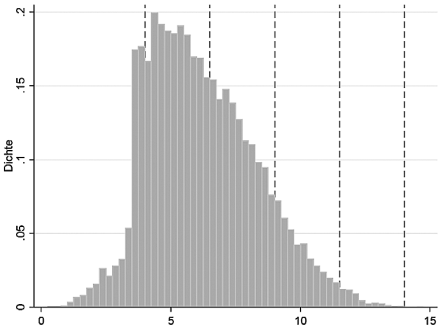
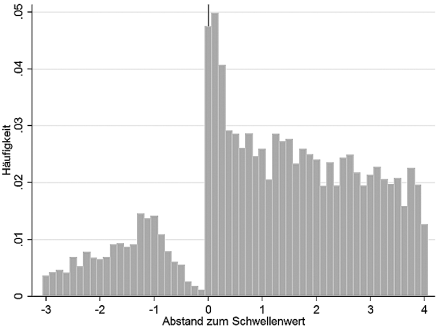


Abbildung 2: Verteilung der mündlichen Noten relativ zum Schwellenwert.



Auffallend ist, dass diese Diskontinuität in der Verteilung der Gesamtnoten bei der Verteilung der schriftlichen Noten nicht zu beobachten ist. Dort zeigen sich weder „unnatürliche“ Häufungen über noch eine fehlende Beobachtungsmasse unter den relevanten Schwellenwerten (s. *Abbildung 1*). Dies legt nahe, dass die Anomalien in der Verteilung durch eine vornotenorientierte Notengebung in den mündlichen Teilen des zweiten Examens produziert werden. Das Wissen über die schriftlichen Vornoten scheint von den Mitgliedern der Prüfungskommissionen berücksichtigt zu werden, um über Anpassungen der mündlichen Noten „eindeutige“ Gesamtnoten herzustellen, die nicht knapp unter den jeweiligen Notenstufen liegen. Diese Interpretation wird durch einen genaueren Blick auf die Verteilung der mündlichen Noten bestätigt.

Im Fall einer von der Vornote unabhängigen Notenvergabe in der mündlichen Prüfung sollten wir eine Verteilung beobachten, die auch an den Notenstufen stetig streut. *Abbildung 2* zeigt indessen ein deutlich anderes Bild auf: Fälle, in denen

Studierende knapp (genauer: zwischen 0,1 und 0,5 Notenpunkte) unter dem jeweiligen „Zielwert" liegen, der für das Erreichen der nächsten Notenschwelle notwendig wäre (in der Abbildung durch den Null-Punkt abgebildet), treten überzufällig selten auf. Gleichzeitig zeigt sich eine deutliche Häufung von Fällen, in denen der jeweilige Schwellenwert knapp oder genau erreicht wird. Dieses Muster reflektiert sehr deutlich die vornotenorientierte Vergabe der mündlichen Noten: Die Häufung der Werte oberhalb der Schwelle in Kombination mit den fehlenden Werten unterhalb der Schwelle spricht für „vornotenorientiertes Anheben". Dies gilt nicht nur für die Notenstufe befriedigend/vollbefriedigend, das Muster lässt sich vielmehr bei allen vier Schwellenwerten beobachten.[19]

2. Wahrscheinlichkeit des Erreichens der nächsthöheren Notenstufe

a. Erstes Examen

Mit Blick auf die Abschlussnoten von Frauen im ersten Examen zeigt die Analyse, dass Frauen eine um 5,3%-Punkte niedrigere Rate an „besser als 9,0"-Noten haben (*Tabelle 9, Modell 1*). Relativ zu Männern haben Frauen eine etwa 17% geringere Rate an vollbefriedigenden oder besseren Noten. Bei Kontrolle für die Abiturnote steigt der Unterschied auf 7,3%-Punkte bzw. 23% an (*Modell 2*).[20] Wird für die Note aus dem staatlichen Prüfungsteil kontrolliert, verschwindet jedoch der Geschlechterunterschied (*Modell 3*). Das bedeutet, dass der Geschlechtsunterschied vor allem vom staatlichen Teil des ersten Examens getrieben wird (und nicht von der durch die Universitäten verantworteten Teil der Prüfung). Diese Aussage wird durch die Resultate in den *Modellen 4 bzw. 5* unterstützt, die aufzeigen, dass Frauen eine um 2,7 bzw. 7,3%-Punkte geringere Wahrscheinlichkeit haben, in diesem Teil der Prüfung eine Note von 9,0 (oder mehr) zu erreichen.

Eine detaillierte Analyse für die Notenschwelle 9,0 (vollbefriedigend) zeigt einen starken Einfluss der drei Indikatoren Namensherkunft, Geburtsort und Staatsangehörigkeit auf die Wahrscheinlichkeit, die Note von 9,0 oder mehr zu erreichen. Für Migrantinnen und Migranten der ersten Generation ist die Wahrscheinlichkeit, die Stufe des *vollbefriedigend* zu erreichen um 19,2%-Punkte geringer, für solche der zweiten und höheren Generation immer noch um 12,6%-Punkte niedriger (*Tabelle 10, Modell 1*). Interessant ist dabei, dass nach Kontrolle für die Abiturnote letzterer Migrationsindikator insignifikant wird (Modell 2). Der erste Indikator bleibt jedoch hoch-signifikant und bildet einen starken negativen Unterschied (von weiterhin 16,8 %-Punkte) ab. Die Ergebnisse legen also nahe, dass Studierende mit einem Migrationshintergrund bei gleicher Abiturnote eine gleich hohe Chance haben, ein Prädikatsexamen zu erreichen, *sofern sie bereits in Deutschland geboren wurden*. Für Migrantinnen und Migranten der ersten Generation gilt dies jedoch

19 Siehe dazu Abschlussbericht für das Justizministerium, verfügbar über www.justiz.nrw.de.

20 Die hier und im Folgenden präsentierten Resultate für binäre abhängige Variablen wurden mit dem linearen Wahrscheinlichkeitsmodel (*linear probability model, LPM*) geschätzt. Dies vereinfacht die Interpretation der Koeffizienten für Dummy-Indikatoren und Interaktionsterme. Schätzungen basierend auf nicht-linearen Modellen (*Probit*) liefern qualitativ und quantitativ vergleichbare Ergebnisse.

nicht: Selbst bei gleichen Abiturnoten werden gute Examensnoten unterdurchschnittlich häufig erreicht.[21] Völlig unabhängig von der Modellspezifikation zeigen *alle* Schätzungen (s. Ergebnisse der F-tests) signifikante Unterschiede in den Effekten für Migranten der ersten bzw. der zweiten und höheren Generation, die konsistent mit einer integrativen Wirkung sind.

Tabelle 9: Wahrscheinlichkeit, die Notenschwelle von 9,0 zu erreichen (erstes Examen)

	(1)	(2)	(3)	(4)	(5)
Variable:	Indikator: Note ≥ 9,0 (Gesamtnote)			Note ≥ 9,0 (staatl. Pflichtfachpr.)	
Frauen	-0,053***	-0,070***	0,013	-0,027**	-0,073***
	[0,000]	[0,000]	[0,156]	[0,036]	[0,000]
Abiturnote		-0,310***	-0,054***		-0,269***
		[0,000]	[0,000]		[0,000]
Note staatl. Pflichtfachpr.			0,168*** [0,000]		
Konstante	0,315***	1,014***	-0,835***	0,253***	0,887***
	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]
N	10.042	4.596	4.252	4.390	4.254

Schätzungen mit linearem Wahrscheinlichkeitsmodell (LPM). Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

Tabelle 10: Effekte der indirekten Migrationsindikatoren auf Stufenerreichung 9 Punkte im ersten Examen

	(1)	(2)	(3)	(4)	(5)
	Gesamtnote über Schwellenwert (9,0)			Note in staatl. Pflichtfachprüfung über Schwellenwert (9,0)	
MI: 1. Generation	-0,192***	-0,168***	-0,035*	-0,127***	-0,123***
	[0,000]	[0,000]	[0,073]	[0,000]	[0,000]
MI: ≥ 2. Generation	-0,126***	-0,022	0,020	-0,058**	-0,007
	[0,000]	[0,311]	[0,205]	[0,014]	[0,734]
MI: Dt./Geb.Ausland	-0,098***	-0,077**	-0,035	-0,097***	-0,067**
	[0,000]	[0,018]	[0,195]	[0,005]	[0,045]
Abiturnote		-0,310***	-0,056***		-0,269***
		[0,000]	[0,000]		[0,000]
Note staatl. Pflichtfachpr.			0,168*** [0,000]		

21 Unklar bleibt hier, inwiefern die An- und Über-Rechnung von in Drittstaaten erzielten Abiturnoten zu einer Verzerrung der Befunde führt.

	(1)	(2)	(3)	(4)	(5)
	Gesamtnote über Schwellenwert (9,0)			Note in staatl. Pflichtfachprüfung über Schwellenwert (9,0)	
Weitere Kontrollvariablen:	Nein	Ja	Ja	Nein	Ja
Konstante	0,310***	1.023***	-0,827***	0,250***	0,893***
	[0,000]	[0,000]	[0,000]	[0,000]	[0,000]
N	9.430	4.596	4.252	4.296	4.254
F-test	0,001	0,000	0,022	0,029	0,000

Schätzungen mit linearem Wahrscheinlichkeitsmodell (LPM). Abhängige Variable ist ein Indikator der angibt, ob die Gesamtnote (bzw. ob die Note in der staatlichen Pflichtfachprüfung) über dem Wert von 9,0 liegt. Modelle (2), (3) und (5) kontrollieren für das Geschlecht, Alter (linear und quadratischer Term) sowie für Abschlussmonat-spezifische Effekte. Die Zeile F-test dokumentiert die p-Werte für Tests der Null-Hypothese, dass sich die Effekte für MI: 1. Generation und MI: ≥ 2. Generation nicht unterscheiden. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

b. Zweites Examen

Eine Analyse der Wahrscheinlichkeit der Erreichung der Notenstufen 9,0 im zweiten Examen zeigt, dass Frauen eine um 3,6%-Punkte geringere Rate an Prädikatsexamina als Männer aufweisen (*Tabelle 11, Modell 1*). Nimmt man den Anteil von Prädikatsexamina bei Männern als Grundlage (29,3%), so entspricht das einem Geschlechtsunterschied von über 12%. Dieser Unterschied steigt bei Kontrolle für Alter und Abschlusszeitpunkt deutlich an (*Modell 2*). Wird jedoch für die Gesamtnote im ersten Examen kontrolliert (in der es ja ebenfalls einen deutlichen Geschlechtsunterschied gibt), so zeigt sich kein *zusätzlicher* Unterschied in der hier untersuchten Wahrscheinlichkeit (*Modell 3*). Dieser Befund ändert sich nicht, wenn zusätzlich auch für die Abiturnote kontrolliert wird (*Modell 4*).

Werden also ein durchschnittlicher Jurist und eine durchschnittliche Juristin gleichen Alters und mit gleichen Noten im ersten Examen miteinander verglichen, so ist *kein* statistisch signifikanter Unterschied hinsichtlich der Wahrscheinlichkeit festzustellen, mit der beide die Notenschwelle 9,0 erreichen. Das Resultat legt also nahe, dass der Geschlechtsunterschied – der in der Rate der Prädikatsexamina im zweiten Examen klar zu beobachten ist (siehe Modell 1 und 2) – im Wesentlichen bereits in den Notenunterschieden im ersten Examen zum Ausdruck kommt, und darüber hinaus *durchschnittlich* kein *weiterer* Unterschied zu beobachten ist (Modell 3 und 4).

Tabelle 11: Wahrscheinlichkeit zur Erreichung der Notenschwellen von 9,0 (zweites Examen)

	(1)	(2)	(3)	(4)
Variable:	Indikator: Gesamtnote ≥ 9,0 (zweites Examen)			
Frauen	-0,036***	-0,083***	-0,010	-0,018
	[0,000]	[0,000]	[0,194]	[0,113]
Gesamtnote			0,141***	0,136***
erstes Examen			[0,000]	[0,000]
Abiturnote				-0,060***
				[0,000]
Konstante	0,293***	2,871***	0,561**	1,057**
	[0,000]	[0,000]	[0,017]	[0,017]
Weitere Kontrollvariablen	Nein	Ja	Ja	Ja
N	17.971	17.971	9.086	4.251

Schätzungen mit linearem Wahrscheinlichkeitsmodell (LPM). Modelle (2) – (4) kontrollieren für das Alter zum Prüfungszeitpunkt (linearer und quadratischer Term) sowie für Abschlussmonats-spezifische Effekte. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

Auch die unterschiedlichen Migrationsindikatoren haben starken Einfluss auf die Wahrscheinlichkeit, eine Note von 9,0 oder höher zu erreichen. Der Einfluss dieser Indikatoren sinkt nach Kontrolle für Alter, Geschlecht, Abschlusszeitpunkt und für die Note in der ersten Prüfung bzw. die Abiturnote; dennoch sind signifikante Notenunterschiede auch nach Kontrollen für diese Noten festzustellen.

Konsistent mit den Ergebnissen für die Gesamtnoten zeigt sich auch hinsichtlich der Wahrscheinlichkeit, eine Note von 9,0 oder höher zu erreichen, ein starker Einfluss der unterschiedlichen Migrationsindikatoren (*Tabelle 12*). Der Einfluss der Indikatoren sinkt nach Kontrolle für Alter, Geschlecht, Abschlusszeitpunkt und für die Note in der ersten Prüfung bzw. die Abiturnote; dennoch sind signifikante Notenunterschiede auch nach Kontrollen für diese Noten festzustellen. D.h., dass bei Referendarinnen und Referendaren mit Migrationshintergrund die Wahrscheinlichkeit, die nächsthöhere Notenstufe zu erreichen, geringer ist, als für „deutsche" Referendarinnen und Referendare.

Tabelle 12: Effekte der direkten Migrationsindikatoren auf Stufenerreichung 9 Punkte im zweiten Examen

	(1)	(2)	(3)	(4)	(5)
I1: Onomastik	-0,137***	-0,113***	-0,047***	-0,119**	-0,124*
	[0,000]	[0,000]	[0,000]	[0,018]	[0,080]
I2: Geburtsort	-0,082***	-0,050***	-0,020	-0,030*	-0,033
	[0,000]	[0,002]	[0,143]	[0,058]	[0,131]
I3: Staatsangeh.	-0,023	-0,014	0,015	0,031	0,029
	[0,344]	[0,565]	[0,511]	[0,187]	[0,396]
Note 1.Examen			0,139***	0,139***	0,135***
			[0,000]	[0,000]	[0,000]
Abiturnote					-0,063***
					[0,000]
MI: Region 1				0,126*	0,070
				[0,054]	[0,433]
MI: Region 2				0,127**	0,087
				[0,032]	[0,316]
MI: Region 3				0,073	0,079
				[0,163]	[0,295]
MI: Region 4				0,085	0,020
				[0,104]	[0,791]
MI: Region 5				0,046	0,050
				[0,391]	[0,506]
MI: Region 6				0,040	0,010
				[0,509]	[0,914]
Konstante	0,309***	3,005***	0,435*	0,435*	0,898**
	[0,000]	[0,000]	[0,071]	[0,070]	[0,042]
N	8.757	8.757	8.504	8.474	4.239

Schätzungen mit linearem Wahrscheinlichkeitsmodell (LPM). Abhängige Variable ist ein Indikator der angibt, ob die Gesamtnote über 9,0 liegt. Modelle (2) – (5) kontrollieren für Alter, (linearer und quadratischer Term), Geschlecht und Abschlussmonat-spezifische Effekte. MI Regionen: (1) Afrika, (2) Süd/Ost- und Zentral-Asien, (3) West-Asien, (4) Ost-Europa, (5) Süd-Europa, (6) Andere/Unklare Herkunft, Referenzkategorie: Zentral- und Nord-Europa. Robuste p-Werte in eckigen Klammern; Signifikanz-Niveau: *** p<0,01; ** p<0,05; * p<0,1.

3. Zusammensetzung der Prüfungskommission

Wie bereits oben ausgeführt, treten die beobachteten Geschlechtseffekte in der *durchschnittlichen* Note des mündlichen Prüfungsteils unabhängig von der Zusam-

mensetzung der Prüfungskommission auf (D. I. 2. c.). Eine differenziertere Betrachtung derjenigen Kandidatinnen und Kandidaten, die aufgrund ihrer schriftlichen Vornoten in einem hinreichend engen Bereich um eine der relevanten Notenstufen liegen, offenbart hingegen einen systematischen Effekt, der bei der Betrachtung allein der durchschnittlichen Noten verborgen blieb[22] und über den wir an anderer Stelle im Detail berichten[23].

Wir untersuchen dabei, wie hoch die Wahrscheinlichkeit ist, dass Referendarinnen und Referendare mit bestimmten Vornoten in der mündlichen Prüfung einen Schwellenwert überspringen und damit die nächste Notenstufe erreichen. Dabei kontrollieren wir für den jeweiligen Abstand zum nächsten Schwellenwert (der sich aus der schriftlichen Note ergibt), der für die Wahrscheinlichkeit der Schwellenerreichung natürlich maßgeblich ist. Dies bedeutet, dass eine mögliche „Quelle" für Geschlechtsunterschiede – nämlich die im Vergleich zu Männern schlechteren schriftlichen Noten von Frauen – bereits absorbiert wird. Anders ausgedrückt zeigt die Analyse mögliche Geschlechtsunterschiede zwischen Studierenden auf, die *mit gleicher schriftlicher Durchschnittsnote* in die mündliche Prüfung gehen.

Die Resultate zeigen, dass es – bei Betrachtung aller vier Schwellenwerte – bei Frauen durchschnittlich eine um 1,3%-Punkte weniger wahrscheinlich ist, die nächste Notenstufe zu erreichen als bei Männern (M: 84,5% F: 83,3%). Durchschnittlich bedeutet hier auch, dass dieses Ergebnis für eine „durchschnittliche Prüfungskommission" gilt. Differenziert man jedoch zwischen ausschließlich mit männlichen Prüfern besetzten Kommissionen und solchen, an denen zumindest eine Prüferin beteiligt ist, so ergibt sich ein deutlich anderer Befund: Sind alle Prüfer männlich, so finden wir einen größeren Geschlechtsunterschied von 2,3%-Punkten. Ist jedoch zumindest eine Frau Teil der Prüfungskommission, so verschwindet der Geschlechtsunterschied: Relativ zu einer rein männlich besetzten Kommission erhöht sich die Wahrscheinlichkeit bei Frauen, die nächste Notenschwelle zu erreichen marginal, während die Wahrscheinlichkeit bei Männern fällt. In Prüfungskommissionen unter Beteiligung mindestens einer Prüferin verbleiben damit *keine* quantitativ relevanten oder statisch signifikanten Geschlechtsunterschiede hinsichtlich der Wahrscheinlichkeit, die Notenschwelle zu überspringen. Diese Befunde bleiben bei Kontrolle für zusätzliche Faktoren (etwa die Note aus dem ersten Examen) sowohl qualitativ wie auch quantitativ weitgehend unverändert. Weitere Analysen deuten darauf hin, dass der „Nivellierungs-Effekt" nicht mit der Zahl der Frauen in der Kommission steigt.[24]

22 Genauer betrachten wir dazu wieder Fälle mit schriftlichen Noten, die -1,5 bis +0,5 um die Schwellenwerte liegen. Weder die grafische Evidenz aus Abbildung 2 noch die im Folgenden besprochenen Ergebnisse sind sensitiv hinsichtlich der exakten Definition dieses Bereichs rund um die Schwellenwerte.

23 *Traxler/Glöckner/Towfigh* (Manuskript in Vorbereitung). Gender composition in exam committees and gender gaps in student achievement: Evidence from German Law exams.

24 Die Aussagekraft dieser Analyse ist jedoch stark eingeschränkt, da kaum Beobachtungen mit mehr als einer Frau pro Kommission vorliegen: 65% der Prüfungen werden von rein männlich besetzten,

Bei einer Betrachtung über die Zeit ist zu beobachten, dass der Anteil der Prüfungen mit mindestens einer Frau in der Kommission lange bei etwa einem Drittel lag: nur in jeder dritten Kommission war zumindest eine Prüferin dabei. In den letzten Jahren – insbesondere in den Jahren 2015 und 2016 – ist diese Zahl jedoch deutlich angestiegen (2016 auf 44%). Dieser vor dem Hintergrund der hier präsentierten Befunde begrüßenswerte Trend geht auf konkrete Anstrengungen des LJPAes zurück, die Prüfungskommissionen mit Prüferinnen zu besetzen.

E. Schlussfolgerungen, Empfehlungen und Ausblick

Zunächst ist festzuhalten, dass die beobachteten Unterschiede zwischen Männern und Frauen sowie zwischen Personen mit bzw. ohne Migrationshintergrund in juristischen Examina durch sehr unterschiedliche Ursachen hervorgerufen werden können, die im Prüfling selbst, im Ausbildungssystem, in der Organisation der Prüfung oder auch in den Personen der Prüferinnen und Prüfer zu verorten sein können. Die Analysen können nur teilweise mögliche Ursachen für Unterschiede beleuchten; eine Aussage zu kausalen Zusammenhängen, die zu den beobachteten systematischen Notenunterschieden führen, erfordert eine andere Datenlage. Für den Großteil der Befunde lassen sich daher keine Aussagen über Kausalzusammenhänge treffen.

I. Diskriminierung?

Die Befunde aus der Vorgängerstudie von 2014 haben sich auch angesichts der erweiterten Datenbasis als robust erwiesen und bestätigt, so dass nach wie vor naheliegt, dass es sich um (ggf. unbewusste) Diskriminierungseffekte handelt, die in der Ausbildung und in den Prüfungsverfahren verankert sind. Dafür spricht etwa der Umstand, dass die Zusammensetzung der Prüfungskommission Auswirkung auf die Benotung hat (D.III.3.), als auch die Beobachtung, dass bestimmte Effekte nur im staatlichen, nicht aber im universitären Teil der Prüfung zu beobachten sind (oben D.I.1.), bzw. umgekehrt dass im Referendardienst die Herkunftsuniversität sich auf die beobachtete Stärke des Geschlechtseffekts auswirkt (oben D.I.2.a — was u.a. an einer spezifischen Zusammensetzung der Studierenden der Herkunftsuniversität liegen kann oder auch an besonderen Form der Sozialisierung an den unterschiedlichen Hochschulen). Jedenfalls kann Diskriminierung als Faktor für die beobachteten Bewertungsunterschiede aufgrund Geschlecht oder Herkunft nach wie vor nicht ausgeschlossen werden.

II. Zusammensetzung der Prüfungskommissionen

Für die Zusammensetzung der Prüfungskommissionen lassen sich jedoch, wie an anderer Stelle im Detail besprochen wird (siehe Fn. 23), die Ergebnisse kausal interpretieren: die Anwesenheit einer Prüferin in der Kommission führt dazu, dass

30% von Kommissionen mit genau einer Prüferin, 5% von Kommissionen mit zwei Prüferinnen und nur 0,3% von Kommissionen mit drei Prüferinnen abgenommen.

sich die Geschlechtsunterschiede hinsichtlich der Wahrscheinlichkeit des Erreichens der nächsthöheren Notenstufe nivellieren. Eine diversere Besetzung der Prüfungskommissionen könnte somit einen wichtigen Beitrag zur Chancengleichheit der Prüflinge darstellen.

Gleichzeitig könnte eine ausgeglichenere Zusammensetzung der Prüfungskommissionen einer potenziellen Ungleichbehandlung von Prüflingen entgegenwirken. Frauen sind in Prüfungskommissionen üblicherweise unterrepräsentiert, selbiges gilt in einem noch deutlicheren Ausmaß für Prüferinnen und Prüfer mit Migrationshintergrund. Eine diversere Besetzung der Prüfungskommissionen könnte zum einen dazu beitragen, dass einer potenziellen Diskriminierung direkt entgegengewirkt wird. Darüber hinaus — und vermutlich praktisch bedeutsamer — ist ein positiver Effekt auf die Prüflinge zu erwarten. Empirische Studien legen die Vermutung nahe, dass die Existenz positiver Rollenmodelle die nachgewiesenen negativen Effekte von Stereotypen — speziell der wahrgenommenen Bedrohung durch Stereotype (*stereotype threat*) — reduziert werden.[25]

III. Abhängigkeit von Vornoten

Wie oben dargelegt, basiert die beobachtete Diskontinuität der Notenverteilung rund um die Notenstufen maßgeblich auf der vornotenorientierten Punktvergabe in der mündlichen Prüfung. In der Praxis überrascht dieser Effekt nicht, völlig unabhängig von den nun vorliegenden empirischen Befunden ist die „strategische" Notenvergabe zur Erreichung bestimmter Notenstufen gängige und allgemein bekannte Praxis in mündlichen Prüfungen. Berücksichtigt man die eingangs angesprochene Bedeutung bestimmter Notenstufen für Karriere und Verdienst, ist das negative Signal, welches von einer knapp verpassten Notenstufe für mit dem Prüfungsprozess vertraute Personaler ausgeht, nicht gering zu schätzen. Häufig wird die nächsthöhere Notenstufe mit einer „Punktlandung" erreicht, knapp verpasst wird sie nur selten. Kommt es dennoch zu einer solchen Situation, so wird gemeinhin angenommen, dass die Prüfungskommission der Auffassung war, diese Kandidatin bzw. dieser Kandidat habe die nächste Notenstufe nicht „verdient"; im Nichterreichen der Notenschwelle liegt also ein eigenständiges Signal.

Dieses Signal ist indessen nicht frei von Willkür, es hängt in hohem Maße von der konkreten Zusammensetzung der Prüfungskommission ab und enthält ein starkes dezisionistisches Element: Die Kommission kann in Kenntnis der Vornoten die Vergabe der mündlichen Noten davon abhängig machen, ob der jeweilige Prüfling ihrem Bild von einem „vollbefriedigenden", „guten" oder „sehr guten" Juristen entspricht; sie kann die Bewertung der mündlichen Leistung also auf die schriftli-

25 Bspw. *Spencer/Steele/Quinn*, in: Journal of Experimental Social Psychology 35 (1999), S. 4 ff.; *Lusher/Campbell/Carrell*, in: Journal of Public Economics 159 (2018), S. 203 ff.; allerdings konnte eine aktuelle Studie, in der die Beteiligung von Frauen in Berufungskommissionen an einer deutschen Hochschule untersucht wurde, keinen generellen positiven Effekt eines höheren Frauenanteils in den Kommissionen auf den Frauenanteil unter den Neuberufungen bzw. Berufungsvorschlägen finden: *Auspurg/Hinz/Schneck*, in: Forschung & Lehre 2017, S. 770 ff.

chen Noten konditionieren. Damit kann die Kommission jedenfalls in Notenrandbereichen die Bedeutung der schriftlichen Noten relativieren und die Bewertung durch ihr eigenes, in einer recht kurzen persönlichen Begegnung gebildetes und möglicherweise von eigenen Vorlieben gefärbtes Urteil gleichsam ersetzen. Hinzu kommt, dass es sich bei der Beurteilung der mündlichen Prüfungsleistungen nicht mehr um unabhängige Beobachtungen handelt, weil die Kenntnis der Vornoten die Bewertung der mündlichen Beiträge beeinflussen kann. Gerade die Aggregation *unabhängiger* Bewertungen führt aber statistisch zu einem Ausgleich von Fehlern und damit zu zuverlässigeren (valideren) Bewertungen.

Bedenkt man neben alledem die ungleichen Wahrscheinlichkeiten für das Erreichen einer Notenstufe, so bekommt diese Benotungspraxis einen besonders schalen Beigeschmack. Eine Benotung der mündlichen Prüfung ohne vorherige Kenntnis der Ergebnisse aus der schriftlichen Prüfung könnte hingegen frei von strategischen Überlegungen der Prüfungskommission erfolgen und damit ein objektiveres Bild der Fähigkeiten der Kandidatin und des Kandidaten abgeben. Eine Normalverteilung der Examensergebnisse insgesamt wäre dann der „Normalzustand", das knappe Erreichen oder Verfehlen einzelner Notenstufen wäre Ergebnis eines Zufalls, keine bewusste Entscheidung einer Prüfungskommission, und würde damit den eigenständigen Signalwert einbüßen; das Ungleichgewicht zwischen Prüferinnen und Prüfern im schriftlichen und mündlichen Teil der Prüfung würde ausgeglichen und die Prüfung auch in dem Sinne objektiver als dezisionistische Elemente und damit Vorstellungen der Kommission, was eine „gute" Juristin ausmacht, ausgemerzt würden.

IV. Integrationseffekte

Über potentielle Integrationseffekte lassen sich aus der vorliegenden Untersuchung keine verlässlichen Aussagen treffen. Sowohl für das erste wie auch für Teilergebnisse zum zweiten Examen zeigt sich zwar im Vergleich zwischen Migrantinnen und Migranten der ersten und zweiten (bzw. höheren) Generation schwache Evidenz für positive Integrationseffekte (s.o. D.II1. und D.III.2.a). Insgesamt ist das Bild aber gemischt und wenig systematisch, was auch mit der eingeschränkten Fallzahl an Beobachtungen in den verschiedenen Subgruppen mit unterschiedlichen Migrationshintergründen zu erklären sein könnte.

V. Anpassungen des Prüfungsverfahrens

Auch wenn – mit wenigen Ausnahmen – nur sehr eingeschränkt Aussagen über die Kausalität der Befunde möglich sind, so zeigt die Studie dennoch systematische Unterschiede zwischen gesellschaftlich relevanten Gruppen. Um eine größere Chancengleichheit zwischen den Gruppen zu garantieren, könnten Anpassungen des Prüfungsverfahrens (s. E.II und E.III) mit überschaubarem Aufwand große Wirkung entfalten.

So sollten die positiven Auswirkungen von Prüferinnen in Prüfungskommissionen zum Anlass genommen werden, ihren Einsatz in verstärktem Maße anzustreben. Ziel sollte es sein, jede Kommission mit mindestens einer Frau zu besetzen. NRW ist hier auf einem guten Weg, in den letzten Jahren ist die Zahl der Prüferinnen kontinuierlich gestiegen; die bisherigen Bemühungen sollten fortgesetzt werden.

Die Zahl der Prüferinnen und Prüfer mit Migrationshintergrund ist bisher so gering, dass statistische Aussagen nicht möglich sind. Damit lassen sich auch keine empirisch fundierten Empfehlungen aussprechen. Die Ergebnisse zum Einfluss von Prüferinnen in den mündlichen Prüfungsteilen legen jedoch nahe, dass über einen häufigeren Einsatz von Prüferinnen und Prüfern mit Migrationshintergrund diskutiert werden sollte.

Eine weitere mit nur geringem Aufwand umsetzbare Veränderung betrifft unmittelbar das mündliche Prüfungsverfahren. Zur Vermeidung der Diskontinuität der Notenverteilung und um auch der übersteigerten Bedeutung der Notenstufen entgegenzuwirken, sollte in Betracht gezogen werden, die Noten aus den schriftlichen Prüfungen den Kommissionen der mündlichen Prüfungen nicht vorab mitzuteilen. Dies führt zu unabhängigeren Bewertungen der konkreten Prüfungsleistung in der mündlichen Prüfung und zu einer im Schnitt valideren Gesamtnote durch den Ausgleich potenzieller Fehler in den unterschiedlichen Prüfungen.

VI. Zukünftige Forschungsansätze

Im Interesse eines objektiven Prüfungsverfahrens und des Erreichens größtmöglicher Chancengleichheit liegt es nahe, den Ursachen einer (bewussten oder unbewussten) Diskriminierung auf den Grund zu gehen und zu ergründen, inwieweit unerwünschte Einflüsse auf die Benotung wirken. Dies könnte zum Beispiel dadurch erreicht werden, dass Stichproben von Prüferinnen und Prüfern „virtuelle Prüflinge“ auf der Basis der realen Anforderungen beurteilen (z.B. durch Korrektur einer identischen Klausur in unterschiedlichen Handschriften). Ein solches Verfahren könnte Aufschluss geben über die Reliabilität der Messung der Leistung durch Prüfungsnoten, aber auch helfen, die Frage der Kausalität von Diskriminierung zu beantworten.

Gleichzeitig ist eine detaillierte Untersuchung von Einflussfaktoren auf Seiten der Prüflinge von Interesse, die es ermöglichen könnten, Aussagen zu kognitiven Fähigkeiten, Persönlichkeit aber auch der Einstellung zur Prüfung und Einschätzung der Prüfungssituation (z.B. Bedrohung durch Stereotype) zu treffen.

Solche Untersuchungen sind mit den vorhandenen Daten noch nicht möglich, sondern bedürfen einer maßgeschneiderten Datenerhebung und der Beteiligung sowohl von Prüflingen als auch von Prüferinnen und Prüfern.