Jede Korrektur eine andere Note: Quantitative Untersuchung der Objektivität juristischer Klausurbewertungen

Clemens Hufeld[*]

Zusammenfassung

In einem Experiment wurden dieselben 15 Anfängerklausuren von 23 Personen insgesamt 230 Mal korrigiert. Jede Klausur hat entweder 15 oder 16 Benotungen bekommen, jede Note von einer anderen Person. Die Unterschiede wurden statistisch analysiert, mit dem Ergebnis, dass die durchschnittliche Abweichung zwischen niedrigster und höchster gegebener Note bei 6,47 liegt und statistisch erwartbar nur 42% der vergebenen Noten pro Klausur ±1 um den Durchschnitt liegen. Theoretisch werden juristische Prüfungen als psychometrischer Test begriffen und das Strukturgleichungsmodell des juristischen Staatsexamens um die latente Variable „juristisches Können" wird beschrieben. Das Experiment untersucht das erste der drei Hauptgütekriterien psychometrischer Tests – die Objektivität.

A. Einleitung

Mit der Notengebung im juristischen System wird versucht, Absolventinnen und Absolventen rational nach ihrem Können zu unterteilen.[1] Nach wie vor hängt ein großer Teil der persönlichen Lebens- und Karriereplanung von diesen Noten ab. In Hinblick auf die gravierenden Implikationen für die einzelne Person, ist ein hohes Maß an Objektivität für das System juristischer Notengebung erstrebenswert. Gem. § 16 Abs. 1 S. 2 BayJAPO soll die Erste Juristische Prüfung feststellen, ob die Bewerberinnen und Bewerber das Ziel des rechtswissenschaftlichen Studiums erreicht haben und für das Referendariat fachlich geeignet sind. Die Objektivität der Examensnote soll durch mindestens sechs geschriebene, fünfstündige Klausuren, die innerhalb von zwei oder drei Wochen als Gutachten, Urteile oder Schriftsätze verfasst werden, gewährleistet werden. Hierbei gilt der prüfungsrechtliche Grundsatz der Chancengleichheit nach Art. 12 Abs. 1 i.V.m. Art. 3 Abs. 1 GG, nach dem für vergleichbare Prüflinge so weit wie möglich vergleichbare Prüfungsbedingungen und Bewertungsmaßstäbe gelten sollen.[2]

Innerhalb der juristischen Ausbildungswelt ist es jedoch eine seit Langem bekannte und diskutierte Binsenweisheit, dass die Chancengleichheit bei Bewertungen schriftlicher Klausuraufgaben oder Hausarbeiten nicht ideal ist. Geschichten, dass

* Clemens Hufeld hat an der LMU München Rechtswissenschaften, Linguistik, Informatik und Umweltstudien studiert. Er promoviert bei Prof. Dr. Jens Kersten an der LMU München und ist Referendar in Sachsen-Anhalt. Er dankt Akad. Oberrat Dr. Martin Heidebach für die Unterstützung und das Ermöglichen des Experiments, allen teilnehmenden Studierenden und Korrektor:innen sowie dem Herausgeberkreis und den Gutachtern für ihre hilfreichen Kommentare. Kontakt: chufeld.lehre@gmail.com; https://orcid.org/0000-0002-9428-3456.

1 *Church*, in: Wis.L.Rev. 1991, 825 (828).

2 NVwZ-RR 2015, 858 (859) = BVerwG, Beschl. v. 30.6.2015 – 6 B 11/15 (VGH Mannheim).

Freunde im Studium fast identische Hausarbeiten abgeben, aber sehr unterschiedliche Noten bekommen, sind wohl einem Großteil der juristischen Welt bekannt. Selten werden solche Geschichten publiziert und diskutiert.[3] Bisher existieren jedoch kaum systematische Untersuchungen, inwieweit diesen Anekdoten auch Evidenz zukommt. Dieser Aufsatz gibt ein Forschungsprojekt wider, dass sich mit der Frage befasst, ob Bewertungsunterschiede aufgrund der Korrekturperson tatsächlich bestehen und wenn ja, in was für einem Rahmen sich diese halten. Hierzu wurden von 23 Personen insgesamt 230 Korrekturen zu denselben 15 Klausuren angefertigt und die Ergebnisse analysiert.

B. Stand der Forschung zum Thema Objektivität juristischer Prüfungen

Zu dem Themengebiet Evaluation juristischer Notengebung ist überraschend wenig empirische Forschung publiziert. Ein Experiment wie dieses ist mir jedenfalls formlos bekannt: Beispielhaft soll bei Repetitorien Vergleichbares versucht und während der Pandemie sollen manche Klausuren an mehrere Korrektorinnen und Korrektoren verschickt worden sein. Im anglo-amerikanischen System gab es vereinzelt empirische Berichte. So hat *Lawrence Church* der University of Wisconsin Law School berichtet, dass er in einem strafrechtlichen Seminar zu Übungszwecken die Aufsätze der Studierenden von allen anderen Studierenden hat benoten lassen. Die gegebenen Noten lagen alle zwischen 75 und 93 Punkten aus 100 mit durchschnittlichen Abweichungen pro Aufsatz von 11,9 Punkten. Er zieht den Schluss: „*There is no way to explain the degree of disparity except to note the obvious: different graders react differently to the same legal essay.*“[4]

Im deutschsprachigen Raum besteht zu diesem Themengebiet bisher keine Forschung, zumindest aber eine empirische Untersuchung der Klausurpraxis von *Towfigh* et al., die zeigen, dass Geschlecht, Fakultät, Anzahl der Probeklausuren, ob Studierende Migrationshintergrund haben und der Examenstermin signifikante Effekte auf die Examensnote haben.[5]

Weitere Forschung setzt sich mit der Art und Weise der Notengebung verschiedener Law Schools auseinander[6] oder benennt mangelnde Objektivität zwar als Problem und setzt sich für größere Objektivität ein, untersucht aber nicht empirisch, inwieweit Differenzen gegeben sind.[7] Allgemeine Forschung zu Benotung an sich stellt grundsätzlich fest: „‚Benotung‘ ist ein Versuch zur Messung; diese Messung ist im Bereich des Menschlichen aber in sich illusorisch.“[8] Über diese

3 So aber *Walter*, in: BayVBl 2023, S. 689 ff., wo der Stichentscheid 6 Punkte unter der Erstkorrektur lag.

4 *Church*, in: Wis.L.Rev 1991, 825 (830 f.).

5 *Towfigh/Traxler/Glöckner*, in: ZDRW 1 (2014), S. 8.

6 *Kaufman*, in: Journal of Legal Education, 44(3) (1994), S. 415.

7 *Crane*, in: 34 New Engl.L.Rev. 2000, 785 (785).

8 *Eckstein*, in: Schütz/Skowronek/Thieme (Hrsg.), S. 38.

Aussagen hinaus sind konkrete Erhebungen zum juristischen Bereich jedoch kaum zu finden.

Selbst der Bericht von *Church*, der dieser Studie methodologisch am nächsten kommt, befasst sich mit Aufsätzen, was eine andere Prüfungsform als die strukturierte juristische Falllösungsklausur ist. Die weitere Forschung hat entweder andere Forschungsfragen oder andere Methodologien. Insgesamt besteht ein bemerkenswertes Defizit datengetriebener Analyse und Selbstreflektion in Bezug auf das juristische Prüfungswesen.

C. Eine statistische und psychologische Perspektive auf die juristische Klausurpraxis

§ 16 Abs. 1 S. 3 BayJAPO sagt, dass die Bewerberinnen und Bewerber in der juristischen Staatsprüfung zeigen sollen, das Recht mit Verständnis erfassen und anwenden zu können und über die hierzu erforderlichen Kenntnisse in den Prüfungsfächern zu verfügen. Kurz: Juristische Prüfungen dienen dem Erkenntnisgewinn über das juristische Können der geprüften Person. Juristisches Können ist jedoch nicht direkt messbar, da es nicht eine simple Größen- oder Temperaturangabe, sondern eine abstrakte Fähigkeit darstellt. Um dieses Konzept des juristischen Könnens einschätzen zu können, versucht man von anderen, messbaren Werten darauf zu schließen. Ein solches abstraktes Konzept, das erst durch andere Messungen sichtbar gemacht wird, nennt sich in statistischer Terminologie eine latente Variable.[9]

Ein Beispiel hierfür ist das Konzept „Zufriedenheit", das nicht als direkt messbare Größe mit eigener Einheit existiert, sondern durch Messungen etwa von Selbstaussagen, physiologischen Werten, dem Familienleben, erreichten Zielen oder sozialen Beziehungen erkennbar gemacht werden kann.

Die Verknüpfung messbarer Werte (genannt manifeste Variablen) mit der nicht direkt messbaren latenten Variable ist ein Prozess, der sich Operationalisierung nennt. Das Resultat ist die Bildung eines sogenannten latenten Variablenmodells. Abbildung 1 zeigt, wie verschiedene messbare Werte X, die jeweils einem Messfehler e unterliegen und eine spezifische Gewichtung λ haben, von der latenten Variable *LV* bestimmt werden. Abbildung 1 stellt ein Modell für eine sogenannte reflektive latente Variable dar, in der das Konstrukt *LV* existiert und die messbaren Werte X beeinflusst (siehe Richtung der Pfeilspitzen).

9 *Bühner*, Einführung in die Test- und Fragebogenkonstruktion, S. 23.

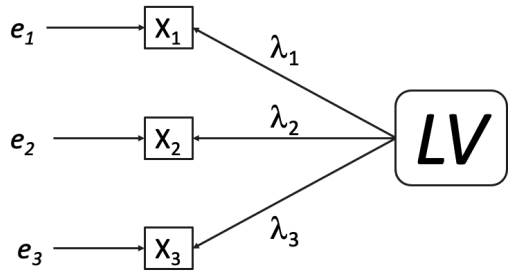

Abbildung 1 - Schematische Darstellung eines latenten Variablenmodells.[10] LV ist die latente Variable, λ die Gewichtung der jeweiligen Messung X und e der Fehler bei der Messung X.

In der juristischen Prüfungspraxis wird angenommen, dass eine Ursache-Wirkungsbeziehung zwischen hohem juristischem Können bzw. Ausbildungsgrad und hohen Noten besteht. Damit das juristische latente Variablenmodell funktionieren kann, muss eine Korrelation zwischen dem Ergebnis der Prüfung und juristischem Können bestehen. So stellt jede juristische Prüfung eine Überprüfung der Hypothese dar, dass ein hoher Ausbildungsstand zu einer guten Note führt und diese Note indikativ für juristisches Können ist. Die eingangs gestellte Frage nach der Messung juristischen Könnens als sogenannte latente Variable wird in der Statistik mit einem Strukturgleichungsmodell beantwortet, d.h. einem statistischen Modell, das aus einem Messmodell und einem Strukturmodell besteht, um das dahinterliegende Konzept sichtbar zu machen.

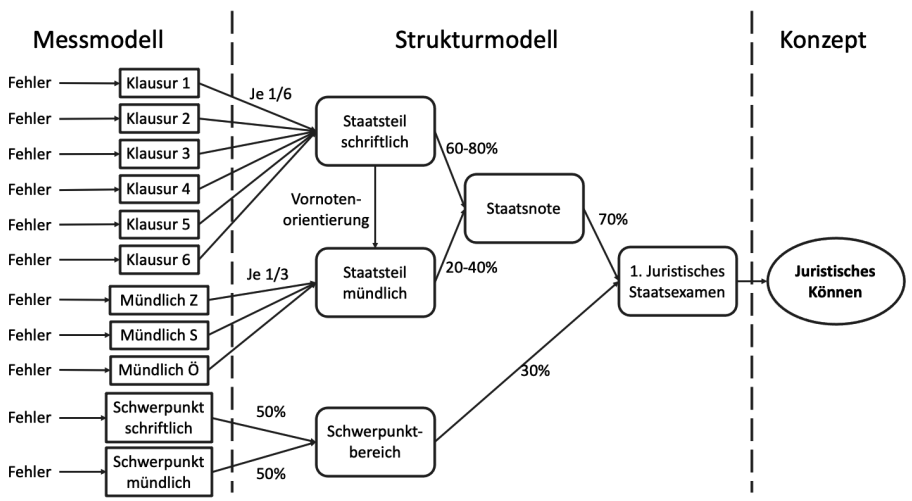

Abbildung 2 - Strukturgleichungsmodell des ersten juristischen Staatsexamens.

10 Angepasst von *Geiser*, Datenanalyse mit Mplus, S. 42.

Abbildung 2 zeigt das Strukturgleichungsmodell des Ersten Juristischen Staatsexamens, also wie juristisches Können aktuell operationalisiert wird. Das Messmodell basiert hierbei auf der Messung verschiedener Prüfungsergebnisse, die im Strukturmodell in verschiedenen Gewichtungen zusammengeführt werden und so letztendlich das juristische Können indizieren. Man spricht bei dieser Wirkrichtung von einem Modell mit einer formativ latenten Variable.[11] Im Rahmen der anfänglichen Studienphasen, wie sie in diesem Projekt untersucht wurde, sähe dieses Modell simpler aus. Es würde lediglich aus drei Teilen bestehen, nämlich

Probeklausur → Resultat → Kenntnisstand zumVorlesungsstoff

Juristische Prüfungen stellen psychologische Tests dar.[12] Damit die grundlegende Annahme der kausalen Verbindung zwischen Noten und juristischem Können aufrechterhalten werden kann, muss das juristische Testmodell – wie jedes andere Testmodell im Rahmen der Forschung zu psychologischen Tests – drei Hauptgütekriterien erfüllen: Objektivität, Validität und Reliabilität. Unter Objektivität versteht man den Grad, in dem die Ergebnisse eines Tests unabhängig von der untersuchenden Person sind.[13] Die Reliabilität gibt den Grad der Zuverlässigkeit eines Messwerts an.[14] Die Validität gibt an, ob der Messwert eines Tests auch wirklich misst, was er zu messen beansprucht.[15] Die Kriterien bauen aufeinander auf. Ein Test, bei dem die Ergebnisse von der messenden Person abhängig sind, kann einen Messwert nicht zuverlässig angeben. Ein Test, der den Messwert nicht zuverlässig angibt, kann auch nicht wirklich das messen, was er beansprucht.

Jenseits der Hauptgütekriterien gibt es noch sieben Nebengütekriterien: die Normierung, Vergleichbarkeit, Ökonomie, Nützlichkeit, Zumutbarkeit, Fairness sowie Nicht-Verfälschbarkeit eines Tests.[16] Man kann jedes dieser Gütekriterien auf ihr Vorliegen in Bezug auf juristische Prüfungen erforschen. Alle Gütekriterien haben mit dem in Abbildung 2 dargestellten Messmodell zu tun, also dem Verhältnis zwischen dem wahrgenommenen Wert (Klausurnote) und dem Messfehler. Darüber hinaus muss man, um die Sinnhaftigkeit juristischer Prüfungen abschließend zu bewerten, zum einen die Klarheit des Konzepts des juristischen Könnens hinterfragen. Zum anderen muss die Gewichtung verschiedener Leistungen im Strukturmodell als Indikationsgrundlagen für die latente Variable kritisch betrachtet werden. In dieser Untersuchung liegt der Fokus auf dem ersten und grundlegendsten Kriterium des Messmodells, der Objektivität der Notenvergabe. Diese ist Bestandteil des Messfehlers der manifesten Variablen, der in Abbildung 1 als *e* und in Abbildung 2 als Kasten mit „Fehler“ gezeigt ist.

11 *Bühner*, Einführung in die Test- und Fragebogenkonstruktion, S. 22.
12 *Bühner*, Einführung in die Test- und Fragebogenkonstruktion, S. 15.
13 *Bühner*, Einführung in die Test- und Fragebogenkonstruktion, S. 568.
14 *Bühner*, Einführung in die Test- und Fragebogenkonstruktion, S. 598.
15 *Bühner*, Einführung in die Test- und Fragebogenkonstruktion, S. 601.
16 Siehe Kap. 5 der DIN 33430, zitiert in *Bühner*, Einführung in die Test- und Fragebogenkonstruktion, S. 617.

Hierbei ist zu beachten, dass der Begriff der Objektivität, insbesondere in der juristischen Welt, stark belegt ist und häufig genutzt wird. In diesem Aufsatz wird der Begriff im Sinne der psychometrischen Gütekriterien[17] genutzt. Der Begriff ist hier nicht als Aussage über Allgemeingültigkeit zu verstehen, sondern ist näher an einer Intersubjektivität verschiedener bewertender Personen orientiert. Abbildung 3 visualisiert das Verständnis der Objektivität als Gütekriterium.

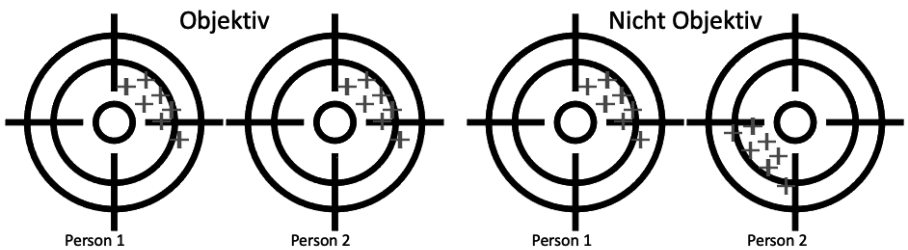

Abbildung 3 - Visualisierung der Objektivität als Gütekriterium eines Tests. Zwei unterschiedliche Prüfer müssen hierfür bei der Bewertung eines Tests ähnliche Ergebnisse erzielen.

Die Verordnung über eine Noten- und Punkteskala für die erste und zweite juristische Prüfung aus dem Jahre 1981 regelt zwei verschiedene Skalen zur Bewertung juristischer Klausuren. Die erste ist eine metrische, diskrete Intervallskala und enthält 19 Bewertungsmöglichkeiten für eine Klausur (0-18 Punkte). Die zweite ist ordinal und enthält sieben Stufen, die in absteigender Reihenfolge sehr gut, gut, vollbefriedigend, befriedigend, ausreichend, mangelhaft sowie ungenügend lauten. Klausuren werden mit beiden Skalen bewertet, indem eine Zahl vergeben wird, die unter eine der Kategorien der ordinalen Skala subsumiert wird. Damit kann die Untersuchung der Objektivität zum einen an der Notenzahl und zum anderen an der Notenstufe festgemacht werden.

D. Methodik

Für die vorliegende Untersuchung wurden 15 echte (d.h. studentische) Bearbeitungen derselben Klausur in drei Gruppen von je 10 Klausuren aufgeteilt. Die 23 Korrektorinnen und Korrektoren wurden auf die drei Gruppen verteilt, sodass die Bearbeitungen in Gruppen 1 und 2 von jeweils acht, die Bearbeitungen in Gruppe 3 von sieben Personen korrigiert wurden. Jede Person hat zehn Klausuren korrigiert, sodass insgesamt 230 Korrekturen erhoben wurden. Siehe Tabelle 1 für die Aufteilung.

17 *Bühner*, Einführung in die Test- und Fragebogenkonstruktion, S. 22.

I. Auswahl der Klausuren sowie Korrektorinnen und Korrektoren

Die Klausur, die Gegenstand der Studie war, ist eine verwaltungsrechtliche Probeklausur für das dritte Semester an der LMU München. Sie wird im Rahmen des Tutoriums Verwaltungsrecht gestellt – die begleitende Übung zur Grundlagenvorlesung Allgemeines Verwaltungsrecht und Verwaltungsprozessrecht. Die Klausur wird im Wintersemester kurz vor Weihnachten bearbeitet. Zu diesem Zeitpunkt haben die Studierenden etwa acht Wochen Verwaltungsrecht gehört. Es handelt sich also um eine Anfängerklausur, die darauf ausgelegt ist, Grundstrukturen abzuprüfen. Inhaltlich geht es um eine Anfechtungsklage gegen den Widerruf einer gaststättenrechtlichen Genehmigung. Die Schwerpunkte der Klausur sind die Auslegung des klägerischen Begehrens mit Bestimmung der statthaften Klageart, die Fristberechnung der Anfechtungsklage, die in Bayern problematische Passivlegitimation des Freistaates, wenn das Landratsamt tätig wird, der Umgang mit der fehlenden Anhörung und in der materiellen Rechtmäßigkeit die saubere Prüfung der Zuverlässigkeit gem. § 4 Abs. 1 Nr. 1 GastG.

Die in dieser Studie benutzten Bearbeitungen sind im Wintersemester 21/22 entstanden. Alle hier genutzten Bearbeitungen habe ich im Rahmen meiner Lehrtätigkeit für das Tutorium selbst korrigiert. Um vernünftigerweise davon ausgehen zu können, dass die ausgewählten Klausuren eine Spannbreite der Notenskala abdecken statt in hohen bzw. niedrigen Notenstufen gesammelt zu sein, habe ich meine eigenen Bewertungen als Grundlage der Auswahl genommen. Dadurch soll nicht gesagt sein, dass meine Bewertungen als „richtig" anzusehen seien. Um die konkreten Bearbeitungen zu nutzen habe ich von allen 15 Studierenden per E-Mail eine Einwilligung zur Nutzung ihrer Klausur zu Forschungszwecken eingeholt. Die Klausur wurde wegen pandemiebedingter Einschränkungen zuhause und, bis auf eine, am Computer geschrieben. Die Bearbeitungen wurden alle derart anonymisiert, dass weder im Dokument noch in den Metadaten der Name zu finden war.

Im Anschluss habe ich Korrektorinnen und Korrektoren angesprochen. Diese wurden nach universitätsüblichen Kriterien ausgewählt: eine Promotion ist begonnen bzw. abgeschlossen oder die Person könnte eine solche aufnehmen. Bis auf eine Person, die sich durch besondere Korrekturerfahrung qualifiziert hat, promovieren alle 23 Korrektorinnen und Korrektoren in Rechtswissenschaften oder haben diese bereits abgeschlossen oder haben in mindestens einem Examen die Note „vollbefriedigend" erreicht. Korrekturerfahrung war bewusst kein Kriterium, da es bei fast allen Grundkursklausuren üblich ist, einige erstmalige Korrektorinnen und Korrektoren zu beschäftigen. Nur erfahrene Korrektorinnen und Korrektoren für die Untersuchung zu engagieren, hätte vielmehr ein Bias[18] eingeführt. Für dieses Projekt standen keine Forschungsgelder bereit, also habe ich alle Korrektorinnen

18 „Bias" ist eine Gewichtung für oder gegen einen Forschungsgegenstand. Ein Bias verringert die Validität als Gütekriterium eines Tests, da man nicht mehr nur einen Zusammenhang misst, sondern aus verschiedenen Gründen voreingenommen ist. Eine Forschung ohne Bias ist nicht möglich, aber eine Reduktion ratsam. Siehe *Delgado-Rodriguez*, Bias.

und Korrektoren über mein Forschungsvorhaben informiert, woraufhin diese die Korrekturen aus gutem Willen und Interesse an dem Forschungsgegenstand übernommen haben.

II. Durchführung der Korrekturen

Die 15 Klausuren wurden in drei Gruppen von je 10 Klausuren aufgeteilt. Die Gruppen habe ich so aufgeteilt, dass der Durchschnitt meiner eigenen Korrekturen in allen drei Gruppen sehr dicht beieinander war. Die Durchschnitte in den Gruppen 1, 2 und 3 waren nach meinen Korrekturen respektive 6,9, 6,7 und 7,2 Punkte. Die Aufteilung der bearbeiteten Klausuren auf die drei Gruppen erfolgte dergestalt, dass jeweils 5 Klausuren in zwei der drei Gruppen zugeordnet werden: Gruppe 1 enthält Bearbeitungen 1-10, Gruppe 2 enthält Bearbeitungen 1-5 und 10-15 und Gruppe 3 enthält Bearbeitungen 5-15. Dadurch konnten mehr Klausurbearbeitungen untersucht werden und Biases aufgrund einzelner Klausuren überprüft werden.

Die Verteilung geschah online per geteilter Ordner. Jeder Ordner enthielt den Sachverhalt, den Lösungsvorschlag, ein Excel-Dokument zum Eintragen der Noten und die der Gruppe zugehörigen 10 Klausurbearbeitungen. Einen Link zum jeweiligen Ordner habe ich per E-Mail an die Korrektorinnen und Korrektoren verschickt, mit der Bitte die Klausuren so zu korrigieren, wie sie das gewohnt sind, und mir die Noten in dem Excel-Dokument zurückzusenden. Es wurden absichtlich keine Angaben zu Art und Weise der Korrektur gemacht. In der E-Mail wurde darum gebeten, die Ergebnisse bis zu einem gewissen Zeitpunkt, der mindestens drei Wochen in der Zukunft lag, zu melden. Von 30 angefragten Korrektorinnen und Korrektoren haben 23 ihre Ergebnisse zurückgeschickt.

III. Durchführung der Analyse

Die Analyse habe ich mit Python3 durchgeführt, insbesondere mit den Paketen pandas (Version 1.5.2) zur Bearbeitung und Manipulation des Datensatzes, scipy (Version 1.7.3) und numpy (Version 1.23.5) für diverse mathematische Operationen. Die Visualisierung der Ergebnisse habe ich mit dem Paket matplotlib (Version 3.6.3) und dem darauf aufbauenden seaborn (Version 0.11.2) vorgenommen.

Um die Objektivität der Bewertung juristischer Klausuren zu überprüfen, wird mithilfe des Pakets pinguoin (Version 0.5.3) der Intraclass Correlation Coefficient berechnet, der das Ausmaß der Übereinstimmung der Ergebnisse verschiedener Beobachter angibt. Durch diese Metrik wird die Abhängigkeit der Bewertung vom einzelnen Bewerter ersichtlich.

Der Link zu den Daten und dem Analysecode kann in meinem GitHub Repo gefunden werden.[19]

19 https://github.com/chufeld/juristischeNotengebung.

E. Ergebnisse

In Tabelle 1 sind die Ergebnisse aller Korrekturen zu sehen. Die Spalten (von oben nach unten) sind mit Buchstaben benannt, die jeweils für eine Person steht, die korrigiert hat. Der Wert in einer Zelle zeigt die Note, die die jeweilige Person der jeweiligen Klausur gegeben hat. Die leeren Stellen in jeder Spalte rühren daher, dass jede Person nur 10 der 15 Klausuren korrigiert hat. Die Zeilen (von links nach rechts) zeigen je eine Klausur, benannt „B“ für Bearbeitung 1 bis 15. Sie enthalten also alle Noten, die von den verschiedenen Personen für dieselbe Klausur gegeben wurden. Die Durchschnittszeile und -spalte (markiert mit „ø”) zeigen den Klausurdurchschnitt (für jede Zeile) und den Durchschnitt der korrigierenden Person (für jede Spalte). Die Spalte max diff zeigt die Differenz zwischen der höchsten und der niedrigsten Note, die für die jeweilige Klausur gegeben wurde. Man bemerke, dass eine max diff von 5 bedeutet, dass 6 Punkte auf der Notenskala erreichbar waren. Bei Klausur B1 etwa wurden Punkte zwischen 2 und 7 vergeben. Tabelle 2 enthält in den Zeilen die Werte der Standardabweichung und Varianz pro Klausur.

I. Korrekturergebnisse und Auswertung

Die Analyse kann zum einen pro Klausur und zum anderen pro korrigierender Person durchgeführt werden.

Zunächst zur Analyse der Klausuren. Hierzu zeigt Abbildung 4 die Boxplots aller Klausurergebnisse. Bei den Boxplots liegen innerhalb der Box die Hälfte der Ergebnisse. Die nach oben und unten weggehenden Whisker zeigen die jeweils anderen 25%. Die als Diamanten dargestellten Punkte stellen Ausreißer dar.

Der Gesamtdurchschnitt über alle Klausuren war 6,83 Punkte. Die Klausuren zeigen Abweichungen zwischen niedrigster und höchster Note zwischen vier Punkten pro Klausur und elf Punkten pro Klausur. Klausur B5 war dabei die unter den Korrektorinnen und Korrektoren umstrittenste Klausur mit Abweichungen zwischen 4 und 14 Punkten. Der Gesamtdurchschnitt dieser Klausur lag bei 8,88 Punkten, also eine durchaus als gelungen anzusehende Klausur. Die Klausur, bei der sich die Korrektorinnen und Korrektoren am ehesten einig waren, war B8 mit Noten zwischen 2 und 5 Punkten. Der Durchschnitt dieser Klausur war 3,00 Punkte.

Tabelle 1 - Ergebnisse der Korrekturen. Zeilen sind einzelne Klausuren (B1-B15), Spalten sind einzelne Korrektorinnen und Korrektoren (A-W). ø = Durchschnitt; max diff = Differenz zwischen niedrigster und höchster Note der Klausur.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	ø	max diff
B1	2	4	5	2	5	7	4	4	3	2	3	4	3	4	5	3	-	-	-	-	-	-	-	3,75	5
B2	2	5	3	5	3	6	5	6	5	6	4	3	5	5	3	2	-	-	-	-	-	-	-	4,25	4
B3	8	11	7	7	6	8	15	7	10	7	7	9	9	8	9	7	-	-	-	-	-	-	-	8,44	9
B4	6	6	9	7	6	7	9	7	7	5	6	7	7	10	6	5	-	-	-	-	-	-	-	6,88	5
B5	7	10	14	12	7	9	12	9	7	4	8	8	8	11	8	8	-	-	-	-	-	-	-	8,88	10
B6	9	9	10	7	9	10	15	10	-	-	-	-	-	-	-	-	12	11	8	8	7	9	10	9,60	8
B7	14	16	15	16	10	14	17	14	-	-	-	-	-	-	-	-	13	15	13	16	13	11	14	14,07	7
B8	3	4	2	2	4	3	5	2	-	-	-	-	-	-	-	-	2	3	3	5	2	3	2	3,00	3
B9	4	3	4	2	2	3	6	3	-	-	-	-	-	-	-	-	6	2	7	6	3	4	3	3,87	5
B10	11	8	15	12	9	12	14	9	-	-	-	-	-	-	-	-	8	10	9	9	6	10	10	10,13	9
B11	-	-	-	-	-	-	-	-	2	6	3	7	7	8	4	4	4	3	7	5	3	2	8	4,87	6
B12	-	-	-	-	-	-	-	-	7	8	4	10	10	10	7	4	7	4	9	11	3	6	8	7,20	8
B13	-	-	-	-	-	-	-	-	2	3	2	4	4	5	2	3	5	2	5	7	2	3	4	3,53	5
B14	-	-	-	-	-	-	-	-	8	11	11	7	9	9	11	6	8	8	11	10	4	9	9	8,73	7
B15	-	-	-	-	-	-	-	-	3	7	3	6	6	6	5	5	5	6	5	7	4	3	9	5,33	6
ø	6,6	7,6	8,4	7,2	6,1	7,9	10	7,1	5,4	5,9	5,1	6,5	6,8	7,6	6	4,7	7	6,4	7,7	8,4	4,7	6	7,7	6,83	6,47

Tabelle 2 - Standardabweichung (σ) und Varianz (σ2) der Klausurbewertungen.

	B1	*B2*	*B3*	*B4*	*B5*	*B6*	*B7*	*B8*	*B9*	*B10*	*B11*	*B12*	*B13*	*B14*	*B15*	*ø*
Std	1,34	1,39	2,19	1,41	2,45	2,03	1,91	1,07	1,64	2,36	2,13	2,57	1,51	2,02	1,68	1,85
Var	1,80	1,93	4,80	1,98	5,98	4,11	3,64	1,14	2,70	5,55	4,55	6,60	2,27	4,07	2,81	3,60

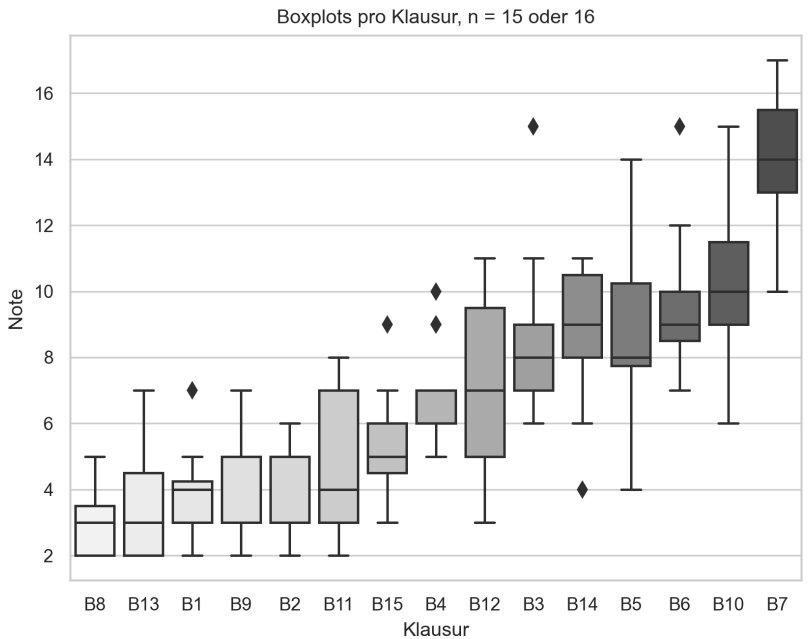

Abbildung 4 - Boxplots der Ergebnisse pro Klausur, sortiert nach Durchschnittsnote (aufsteigend).

Klare Grenzen, dass Klausuren, die über oder unter eine gewissen Notengrenze liegen, gleichmäßiger als andere bewertet wurden, waren nicht erkenntlich. Es scheint einen Trend zu geben, dass die Varianz bei Klausuren, die im Durchschnitt durchgefallen waren, niedriger als bei hoch bewerteten Klausuren ist. Ein möglicher Grund hierfür ist, dass im durchgefallenen Notenbereich unter 4 Punkten kein Unterschied besteht, ob nun 0, 1, 2 oder 3 Punkte gegeben werden, sodass eine gewisse Einigkeit über die generelle Region der Klausur auf der Notenskala herrscht und sich die Noten zwischen 2 und 5 Punkten aufhalten. Aus Gründen der Remonstrationspraxis gegen Klausurkorrekturen könnten hier 2 Punkte häufiger als drei Punkte sein, das bedürfte aber weiterer Forschung. Die Durchschnitte der durchgefallenen Klausuren (B1, B8, B9, B13) befanden sich alle zwischen 3,00 und 4,00. Hier war jedoch bemerkenswert, dass 27 aus den gesammelten 61 Bewertungen dieser Klausuren (44% der Bewertungen) diese Klausuren hätten bestehen lassen, teilweise mit 7 Punkten. Auch bei den Klausuren, die im Durchschnitt knapp über den 4 Punkten liegen (B2 und B11), hätten elf aus 31 Bewertungen (35%) die Klausuren durchfallen lassen. Acht der 15 Klausuren haben außerdem von mindestens einer Person eine Note im durchgefallenen Bereich bekommen, auch wenn bei fünf davon der Durchschnitt über 4 Punkten lag. Insbesondere Klausur B12 wäre von einer Person mit 2 Punkten, von einer anderen mit 11 bewertet worden. Es herrscht also insbesondere bei der Bestehensgrenze kein klarer Maßstab.

Ein solcher Maßstab besteht auch nicht bei hoch oder sehr hoch bewerteten Klausuren. Die Durchschnittsdifferenz zwischen höchster und niedrigster Note, wenn man nur die Klausuren betrachtet, die im Durchschnitt über 7 Punkten bewertet wurden, beträgt 8,29.

Man sieht anhand von Tabelle 2, dass die Standardabweichung im Durchschnitt 1,85 beträgt. Da die Standardabweichung nach oben oder unten eintreten kann, ist selbst bei nur einer Standardabweichung mit Differenzen von 3,7 (also Noten etwa von 4 bis 6,7) zu rechnen, mithin eine gesamte Notenstufe. Im Einzelfall ist für eine Person, deren Klausur bewertet wird, jedoch nicht ein statistischer Wert ausschlaggebend, sondern der tatsächliche Wert, der auch über diese eine Standardabweichung hinausgehen kann. Da jede Person realistischerweise die Klausuren „in echt" hätte korrigieren können, sind alle Werte die aufgetreten sind – auch wenn sie statistisch als Ausreißer gelten – nicht von der Hand zu weisen.

Das wichtigste Ergebnis des Experiments ist der Durchschnittswert der max-diff-Spalte. Der durchschnittliche Unterschied zwischen höchster und niedrigster gegebener Note über alle Klausuren hinweg beträgt 6,47 Punkte. Das bedeutet, dass eine Klausur im Durchschnitt – je nachdem wer die Klausur korrigiert – mit einer Notenspannweite von 7 Punkten zu rechnen hat, oder knapp 40% der möglichen Skala.

Hierbei ist zu bedenken, dass die Benotungspraxis dazu führt, dass die realistisch zu erreichende Skala den möglichen Punktebereich noch weiter einschränkt. 0 Punkte sind wohl nur möglich, wenn man ein leeres Blatt abgibt. Der Unterscheid zwischen einem und zwei Punkten ist rein theoretischer Natur, sodass die Punkte 0 und 1 äußerst selten vorkommen. Ebenso verhält es sich mit 18 und 17 Punkten, die zwar nicht unmöglich sind, dann aber doch mit einer sehr geringen Wahrscheinlichkeit erreichbar sind.

Abbildung 5 zeigt die Boxplots der einzelnen Korrektorinnen und Korrektoren, sortiert nach deren Durchschnittsergebnis. Die Plots sind in drei Gruppen eingeteilt, Gruppe 1 hat Klausuren B1-B10 korrigiert, Gruppe 2 B1-B5 sowie B10-B15 und Gruppe 3 B5-B15.

Es scheint nicht der Fall zu sein, dass Korrektorinnen und Korrektoren entweder insgesamt sehr gute oder sehr schlechte Noten geben. Mit Ausnahme von P und G, die jeweils relativ niedrige oder relativ hohe Noten vergeben haben, zeigt die Höhe der Boxplots, dass die Skala in gewissem Maße genutzt wurde. D hat die Skala mit Noten zwischen 2 und 16 am meisten ausgenutzt, P mit Noten zwischen 2 und 8 am wenigsten. Durchschnittlich wurden 10,2 Skalenpunkte genutzt (also etwa Noten zwischen 2 und 12,2).

Es stechen wieder die Mindest- und Maximalnoten heraus. Zwei Drittel der Korrektorinnen und Korrektoren haben in mindestens einer Klausur nur zwei Punkte vergeben. Es hat aber in insgesamt 230 Korrekturen niemand die Punkte 0, 1 oder 18 vergeben. Bezeichnend ist auch, dass die Minimalnote nicht von der

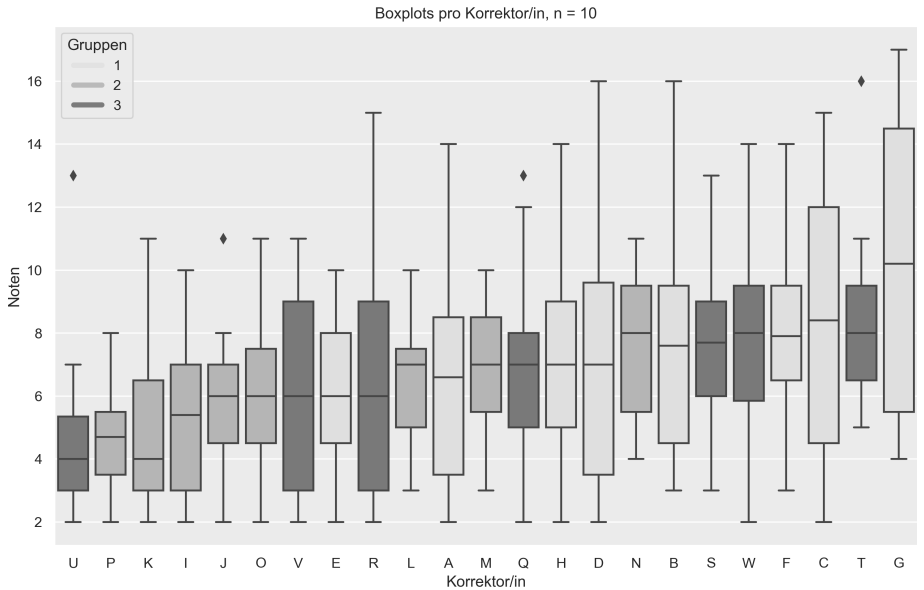

Abbildung 5 - Boxplots der Klausurergebnisse je Korrektor*in, sortiert nach Durchschnittsnote (aufsteigend). Farbgebung nach den drei Korrekturgruppen.

Gruppe abzuhängen scheint, sodass davon ausgegangen werden kann, dass eine klar durchgefallene Klausur 2 Punkte erhält, unabhängig davon, was die anderen Klausurergebnisse im Stapel der zu korrigierenden Klausuren sind. Die gegebenen Maximalnoten sind stark von einer Klausur abhängig, die als besonders gelungen herausstach, so sehr, dass von den Korrektorinnen und Korrektoren Bedenken zu deren Legitimität angemerkt wurde. Die Schwankung in den Maximalnoten ist also – zumindest bei Gruppe 1 und 3 – als die individuelle Schwankung bei einer einzigen Klausur, B7, zu sehen.

Wenn man die Kombination aus Korrektorinnen und Korrektoren und Klausuren betrachtet, kann man nicht erkennen, das einzelne Personen für die großen Spannweiten verantwortlich waren. So hat zum Beispiel Korrektor/in J in zwei Fällen die niedrigste Note gegeben (B4, B5), aber auch bei einer Klausur die höchste Note (B14). Ähnlich hat auch C bei B8 die niedrigste, bei B10 aber die höchste Note gegeben. Es gibt Personen, die im Durchschnitt tendenziell höher oder niedriger benoten, so zum Beispiel U und T, die ähnliche Spannbreiten an Noten gegeben haben, aber um 4 Punkte verschoben. Es ist jedoch nicht so, dass manche Korrektorinnen und Korrektoren ausschließlich die höchste oder niedrigste Note geben. Auch bei ähnlichen Durchschnittswerten pro Korrektor/in ist immer noch ein sehr gemischtes Bewertungsbild bezüglich jeder Klausur zu erkennen.

Diese Uneinigkeit kann mit der Interrater Reliabilität gemessen werden, die den Grad der Übereinstimmung unterschiedlicher Beobachter ausdrückt. Als Korre-

lationsmaß habe ich den sogenannten Intraclass Correlation Coefficient (ICC) gewählt, der anders als andere Korrelationsmaße nicht paarweise Vergleiche vornimmt, sondern Gruppen vergleicht. So können alle Noten eines Korrektors oder einer Korrektorin als Gruppe genommen werden und mit den Ergebnissen anderer Korrektorinnen und Korrektoren verglichen werden, die dieselben Klausuren korrigiert haben.

Tabelle 3 - Intraclass Korrelationskoeffizienten für die drei Klausurgruppen. F = f-statistic, df = degrees of freedom, CI = Konfidenzintervall.

	ICC2	F	Df1	Df2	p-value	CI95%
Grp. 1	0,620	35,48	4	60	0,000	[0,34-0,93]
Grp. 2	0,861	124,14	4	56	0,000	[0,67-0,98]
Grp. 3	0,50	29,46	4	56	0,000	[0,23-0,9]

Der ICC nimmt einen Wert zwischen 0 und 1 an. 1 bedeutet hierbei volle Übereinstimmung der verschiedenen Bewerter, 0 bedeutete absolute Abweichung. Es gibt verschiedene Varianten des ICC. In Tabelle 3 berichte ich das Ergebnis des ICC2, in dem eine Teilmenge aller möglicher Bewerter (Korrektorinnen und Korrektoren) jeweils dieselben Bearbeitungen bewerten.[20] Man geht in der Psychometrie davon aus, dass ein Übereinstimmungswert ab 0,8 gut ist, um ein hohes Maß an Übereinstimmung einerseits und geringe Redundanzen des Tests andererseits anzuzeigen. Die Werte für die erste und dritte Gruppe zeigen Übereinstimmung weit unter einem für Tests akzeptablen Niveau an. Die zweite Gruppe ist demgegenüber in einem guten Niveau der Übereinstimmung, was dafür spricht, dass es klausurabhängig ist, wie sehr Korrektorinnen und Korrektoren übereinstimmen. Manche Klausuren scheinen umstrittener zu sein als andere. In Hinblick auf die Konfidenzintervalle in der rechten Spalte von Tabelle 3 zeigt sich jedoch, dass diese Werte mit Vorsicht zu genießen sind. Aufgrund der geringen Zahl an Beobachtungen liegt hier keine klare Aussage des ICC in einem eng eingeschränkten Bereich vor.

II. Analyse und Implikationen für die juristische Welt

Die Ergebnisse zeigen, dass die vorliegend gemessenen Bewertung in sehr geringem Maß objektiv sind, sondern vielmehr von der korrigierenden Person abhängen. Es gibt zum einen Korrektorinnen und Korrektoren, die im Durchschnitt generell höhere oder niedrigere Noten vergeben. Zum anderen gibt es Klausuren, die manchen Korrektorinnen und Korrektoren – unabhängig von deren sonstigem Korrekturverhalten – besonders gut oder besonders schlecht gefallen.

Es fällt schwer, im Feld der juristischen Notengebung einen Hypothesentest durchzuführen, da für eine Null-Hypothese eine grundlegend akzeptierte Abweichung

20 *Shrout/Fleiss*, in: Psychological Bulletin 86(2) (1979), S. 420.

vom „wahren" Klausurergebnis angenommen werden muss. Eine als Basis anzunehmende akzeptierte Abweichung besteht nicht. Hierfür wäre wesentlich mehr datenbasierte Forschung und wissenschaftliche Diskussion nötig. Als Anhaltspunkt kann die Praxis der Drittkorrektur – oder Stichentscheid – bei Uneinigkeit im Examen genommen werden. Eine Drittkorrektur wird etwa in Bayern eingeholt, wenn die Zweitkorrektur mindestens drei Punkte von der Erstkorrektur abweicht (vgl. § 30 Abs. 1 S. 4 BayJAPO), was bedeutet, dass eine Abweichung von zwei Punkten zwischen Erst- und Zweitkorrektur (bspw. 5 in der Erstkorrektur, 3 in der Zweitkorrektur) nach oben oder unten vom Prüfungsamt als erwartet und akzeptabel angesehen wird. Im Fall der Abweichung von bloß zwei Punkten wird der Durchschnitt als Bewertung angenommen (§ 30 Abs. 1 S. 3 BayJAPO). Dieser Durchschnitt soll den wahren Wert annähern. Das bedeutet, dass eine Abweichung der gegebenen Bewertung vom wahren Ergebnis in Höhe von ± 1 Punkt als akzeptable Abweichung angesehen wird.

Um das „wahre Ergebnis" einer Klausur festzustellen, kann man im Rahmen der Classical Test Theory[21] davon ausgehen, dass es einen wahren Wert *T* gibt, der keinen Messfehler enthält. Dieser Wert kann dadurch approximiert werden, dass derselbe Test unendlich häufig angewendet wird.[22] Im hiesigen Fall kann daher der Durchschnitt aller 15 oder 16 Korrekturen für eine Klausur als beste Annäherung des wahren Wertes angenommen werden. Wenn man die Durchschnittsnote zur nächsten ganzen Note rundet, bekommt man damit eine Skala von drei Punkten als akzeptable Ergebnisspanne. Beispielsweise hat Klausur B1 einen Durchschnitt von 3,75 Punkten. Damit wäre der wahre Wert 4 Punkte, akzeptable Noten wären 3, 4 oder 5 Punkte.

Würde diese Methodik zugrunde gelegt werden, dann wäre eine Null-Hypothese, dass die Standardabweichung aller möglicher Korrekturen einer Klausur nur maximal so hoch ist, dass ein überwiegender Anteil der Klausurkorrekturen, etwa 90%, nicht mehr als ein Punkt nach oben oder unten vom „wahren" Ergebnis abweichen ($H_0 = \sigma \leq x \mid (P(-z < y < z) \geq 0{,}9 \wedge z = \pm 1 / x)$). Die Hypothese, die hier zu testen ist, wäre demnach, dass weniger als 90% der Klausurkorrekturen innerhalb ± 1 Punkt des wahren Ergebnisses liegen ($H_1 = \sigma > x \mid (P(-z < y < z) \geq 0{,}9 \wedge z = \pm 1 / x)$). Um diese Hypothese zu überprüfen, muss ein statistischer Hypothesentest genutzt werden. Der Test ergibt einen Wahrscheinlichkeitswert (sog. p-value / probability value), der angibt, mit was für einer Wahrscheinlichkeit die Standardabweichung der gemessenen Ergebnisse den Konditionen der H_0 entspricht.[23]

Ich habe für alle Einzelkausuren mit scipy.stats.normaltest einen Test der Normalverteilung durchgeführt, der bei allen außer zwei Klausuren (B3 und B6) eine Normalverteilung ergeben hat. Es gibt allerdings nach einer visuellen Kontrolle der Histogramme auch bei diesen beiden Klausuren keinen Grund zur Annahme,

21 *Novick*, in: Journal of Mathematical Psychology 3(1) (1966), S. 1 ff.

22 *Cappelleri/Lundy/Hays*, in: Clinical Therapeutics 36.5 (2014), S. 648 ff.

23 *Dahiru*, in: Annals of Ibadan postgraduate medicine 6.1 (2008), S. 21 ff.

dass nicht von einer Normalverteilung auszugehen ist. Die Ergebnisse in Hinblick auf B3 und B6 sind wohl auf die geringe Zahl an Beobachtungen zurückzuführen. Es wird also insgesamt von einer Normalverteilung von Klausurergebnissen um einen wahren Wert ausgegangen. Ich nehme außerdem zur Analyse beispielhaft die Durchschnittswerte aller Klausurkorrekturen, also Gesamtdurchschnitt von 6,83 und eine Standardabweichung von 1,85.

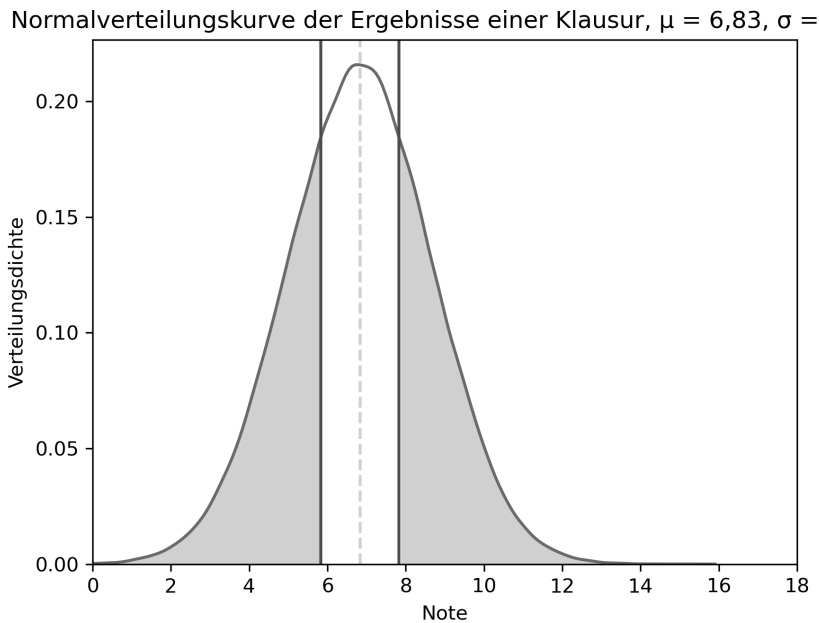

Abbildung 6 - Normalverteilungskurve der Ergebnisse einer standardisiert angenommenen Klausur mit einem Durchschnittsergebnis (gleichzeitig wahres Ergebnis; gestrichelte Linie) von 6,83 Punkten und einer Standardabweichung von 1,85 Punkten.

Abbildung 6 ist eine Darstellung der Normalverteilung der Ergebnisse einer Klausur mit einem Durchschnitt von 6,83 Punkten und der experimentell herausgefundenen durchschnittlichen Standardabweichung von 1,85. Die vertikale, gestrichelte Linie zeigt das wahre Ergebnis der Klausur, die vertikalen, durchgezogenen Linien bei 5,83 und 7,83 Punkten markieren den akzeptablen Abweichungsraum. Diese beiden Punkte haben bei genanntem Durchschnitt und Standardabweichung einen z-score von ± 0,55, was bedeutet, dass sich die Linien 0,55 Standardabweichungen vom wahren Wert entfernt befinden. Aufgrund dieser z-scores kann man die weiße Fläche unter der Kurve zwischen den durchgezogenen Linien berechnen, die angibt, wie viele der Ergebnisse innerhalb der akzeptablen Grenzen liegen. Bei den angenommenen Parametern liegen nur 42% aller Korrekturen innerhalb einer Abweichung von ± 1 Punkt vom wahren Wert der Klausur, also innerhalb eines 3-

Punkte Bereiches. 58% der Korrekturen zeigen größere Abweichungen vom wahren Wert, 28% der Korrekturen zeigen sogar Abweichungen jenseits ± 2 Punkten vom wahren Wert. Diese Berechnung wurde anhand einer standardisierten Klausur vorgenommen. Um die obere und untere Grenze der gemessenen Ergebnisse zu zeigen, werden hier noch die Wahrscheinlichkeiten für die Klausuren mit der niedrigsten und höchsten gemessenen Standardabweichung berichtet. Die Klausur mit der geringsten Standardabweichung ist B8 (σ = 1,07). Bei dieser lägen immerhin 65% der Ergebnisse innerhalb des ± 1 Punkt Rahmens. Bei B12, der Klausur mit der höchsten Standardabweichung (σ = 2,57), lägen nur 30% der Ergebnisse im Rahmen. Das heißt, dass selbst bei der Klausur mit der geringsten gemessenen Abweichung in 35% der Fälle die Korrekturen im inakzeptablen Bereich jenseits ± 1 Punkt liegen würden.

Damit 90% der Ergebnisse innerhalb ± 1 Punkt liegen, dürfte die Standardabweichung nur σ = 0,606 betragen. Ein Levene-Test,[24] der vergleicht, ob die gemessenen Werte einerseits und Werte im H_0-Fall andererseits aus einer Population mit gleicher Varianz stammen, ergibt einen p-Wert von 0,0000. Die H_0 kann daher ohne weiteres und selbst unter Annahme der strengsten α-Werte verworfen werden.

Korrekturergebnisse für dieselbe Klausur lagen nur bei durchschnittlich 42% der Korrekturen innerhalb derselben Notenstufe. In § 12 Abs. 1 S. 1 BayJAPO heißt es: „Erweist sich, dass das Prüfungsverfahren mit Mängeln behaftet war, die die Chancengleichheit erheblich verletzt haben, so kann der Prüfungsausschuss (…) anordnen, dass (…) die Staatsprüfung oder einzelne Teile derselben zu wiederholen sind." Dieser Satz zeigt, dass der Landesgesetzgeber die Chancengleichheit gewährleisten möchte und er nicht davon ausgeht, dass seine Prüfungen stets fehlerfrei ablaufen. Dass aber das Prüfungssystem selbst ein Problem für die Chancengleichheit sein könnte, wird nicht in Betracht gezogen. Das BVerwG meint, der Erfolg in Klausuren hänge von „Faktoren wie der individuellen Begabung, dem persönlichen Lerneifer und der Intensität der Vorbereitung ab".[25] Nach den hier präsentierten Ergebnissen wird man darüber nachdenken müssen, die korrigierende Person in diese Liste mit aufzunehmen.

Wenngleich die Bestimmung der Ursachen weitergehender Forschung bedarf, können hier bereits abstrakt zwei Ursachen für die Differenzen bei der Bewertung der Klausur benannt werden. Zum einen die Unklarheit über den Prüfungsgegenstand („Was bedeutet juristisches Können?") und zum anderen, Unklarheit über den Prüfungsmaßstab („Welcher Klausurinhalt ist „vollbefriedigend", welcher „ausreichend", etc.?"). Der zweite Grund könnte leichter als der erste kontrollierbar werden. Er kann auf die Pluralität verschiedener Ansichten zurückgeführt werden. Der Wettstreit verschiedener Meinungen, der definierend für die juristische Welt ist und ein wichtiges Element der Rechtsstaatlichkeit darstellt, soll aber nicht

24 *Levene*, in: Olkin et al. (Hrsg.), S. 278 ff.; *Brown/Forsythe* (1974), in: Journal of the American Statistical Association 69 (1974), S. 364 ff.

25 NVwZ-RR 2015, 858 (859) = BVerwG, Beschl. v. 30.6.2015 – 6 B 11/15 (VGH Mannheim).

Rechtfertigungsgrund dafür sein, dass das Interesse an einer guten Prüfung außer Acht gelassen wird. Ein Argument, nach dem zu einer juristischen Leistung zwingend verschiedene Meinungen bestehen und daher auch eine Klausur unterschiedlich bewertet wird, mag der Realität entsprechen, lässt ein Prüfungssystem aber als undurchsichtig und unfair dastehen. Hierbei ist der öffentlich-rechtliche Beurteilungsspielraum zu nennen, der die Unbestimmtheit des Prüfungsmaßstabs verwaltungsrechtlich schützt. Als gerichtlich nicht voll überprüfbares Element auf Tatbestandsseite liegt die Letztentscheidungsbefugnis für den Prüfungsmaßstab (sog. prüfungsspezifische Wertungen) bei der korrigierenden Person.[26] In welcher Form dieser Spielraum der Interessenabwägung zwischen korrigierenden Personen, Prüfungsämtern und den Studierenden in Hinblick auf die möglicherweise defizitäre Objektivität am ehesten gerecht wird, sollte in weiterer Forschung überprüft werden.

F. Was nun?

Das jetzige Prüfsystem besteht bereits seit über 200 Jahren.[27] Die Bundesrepublik Deutschland ist nach dem World Justice Project im Rule of Law Index auf Platz 6 von 140 Ländern gelistet.[28] Das Examenssystem leistet mit seiner langen Ausbildung und schwierigen Abschlussprüfungen einen wichtigen Beitrag zur hohen Rechtsstaatlichkeit Deutschlands. Diese Studie soll daher auch nicht so verstanden werden, dass das Bewertungssystem willkürlich sei, zumal der Forschungsgegenstand lediglich eine Klausur in einem kleinen Teil des Prüfungswesens ist. Dennoch sind die Ergebnisse zur Objektivität bedenklich und genauer zu untersuchen. Wenn es also der Anspruch des juristischen Prüfungssystems ist, die verfassungsrechtlich aufgetragene Chancengleichheit zu gewährleisten, sollte den Ergebnissen dieser Studie weiter nachgegangen werden.

Rufe nach einer Veränderung des Systems wären verfrüht: die Ergebnisse müssen in anderen Rechtsgebieten, Abschnitten des Studiums sowie Arten der Korrekturen wiederholt werden und bei einem alternativen Prüfsystem müsste feststehen, dass es tatsächlich besser als das jetzige ist. Den Begriff „besser“ handhabbar zu machen ist zum einen eine politische Entscheidung, die sich durch die Landesgesetzgeber in den Prüfungsordnungen niederschlagen muss, und zum anderen ein wissenschaftlicher Forschungsauftrag, um der Politik eine taugliche Entscheidungsgrundlage liefern zu können. Ohne die wissenschaftliche Untermauerung mit Daten zur aktuellen Prüfungsform und alternativen Prüfsystemen, wäre jede politische Entscheidung ein Stochern im Nebel auf Kosten der Studierenden und der Rechtsstaatlichkeit. Als Hauptergebnis dieser Studie ergibt sich für die akademi-

26 *Decker,* in: BeckOK VwGO, Posser/Wolff (Hrsg.), VwGO, § 114 Rn. 36a; Vgl. etwa NJW 1959, 1842 = BVerwGE, Urteil vom 24. 4. 1959 - BVerwG VII C 104/58 (Münster); BVerfGE 84, 34 (52 ff.).

27 *Hattenhauer*, in: JuS 1989, 513 (514 f.).

28 *Botero/Agrast/Ponce*, World Justice Report 2022, S. 22.

schen Rechtswissenschaften Deutschlands ein Auftrag, weitere Daten zu erheben, zu analysieren und Lösungsansätze für geringere Schwankungsbreiten bei der Bewertung von Klausuren zu entwickeln.

Die notwendige Forschung ist vielfältig. Abbildung 2 zeigt die verschiedenen Stufen des juristischen Strukturgleichungsmodelles. Diese Studie prüft auf Ebene des Messmodells eines von zehn Gütekriterien bei einer zweistündigen Anfängerklausur im Verwaltungsrecht. Aufgrund der hohen Relevanz des Staatsexamens ist erste Priorität festzustellen, inwieweit die Ergebnisse auf Ebene des Staatsexamens bestätigt werden können. Danach können Ursachen der Abweichungen ergründet werden, etwa Stimmung der korrigierenden Person, Uhrzeit der Korrektur, Position im Korrekturstapel, hohe oder niedrige Noten in der unmittelbar vorangehenden Klausur, etc. Insbesondere sollte die Auswirkung eines Korrekturbogens, der Schwerpunkte der Klausur – bis hin zu einer Bewertung mit vorgegebenen Bewertungseinheiten pro Gliederungspunkt[29] – als Erwartungshorizont angibt, empirisch untersucht werde. Solche klareren Vorgaben könnten vielversprechend sein.

Neben der Forschung zur Objektivität sollten die Reliabilität und die Validität als die beiden anderen Hauptgütekriterien untersucht werden. Auch hinsichtlich aller Nebengütekriterien bestehen Forschungsmöglichkeiten. Parallel zur Untersuchung des Strukturmodells sollte jedenfalls das Konzept des juristischen Könnens, das aus den Juristischen Ausbildungs- und Prüfungsordnungen der Länder als Prüfungsziel hervorgeht, näher definiert werden[30] und daraus Schulungen für (Examens-)Korrektorinnen und Korrektoren entwickelt werden, was in Klausuren wie zu bewerten ist.[31] Forschung zum Strukturmodell könnte etwa von den Ergebnissen der Forschung zu den Hauptgütekriterien ausgehen und anhand des dort festgestellten Messfehlers durch Korrektorinnen und Korrektoren errechnen, wie viele Leistungen benötigt werden, um diesen Fehler im Mittel in einem akzeptierten Rahmen zu halten. Parallel zur Erforschung des bestehenden Systems könnte an Universitäten mit alternativen Klausurformen experimentiert werden, um festzustellen, ob bei diesen die Objektivität eher gegeben ist und das Ziel des Abprüfens juristischen Könnens gleichzeitig noch gewährleistet wird.

29 *Heidebach*, in: ZDRW 3 (2015), S. 205 ff. spricht sich hierfür aus. Es ist anzumerken, dass *Heidebach* das Tutorium Verwaltungsrecht der LMU und die hier verwendete Klausur organisiert. Meine ursprünglichen Klausurkorrekturen habe ich mit genau diesen Bewertungseinheiten angefertigt. Die Erfahrungen sind sowohl bei Studierenden als auch bei Korrektorinnen und Korrektoren positiv. Konkrete Zahlen zu den Auswirkungen auf die Objektivität fehlen auch hier.

30 Vgl. *Pilniok*, in: Krüper (Hrsg.), S. 211, der die Situation direkt ausdrückt: „Juristisches Wissen ist bisher weitgehend eine Blackbox".

31 S. hierzu den Werkstattbericht zu einer Korrektorenschulung an der FAU Erlangen-Nürnberg S. 84 ff.

G. Diskussion zu den Grenzen der Studie

Die vorliegende Studie unterliegt in ihrer Durchführung und in Bezug auf ihre Aussagekraft einigen Limitationen, die hier anhand von möglichen Einwänden angesprochen werden.

I. Die Klausuren bzw. die Korrektorinnen und Korrektoren sind nicht repräsentativ.

Das Auswahlkonzept für die Klausuren ist für das, was in dieser Studie erforscht wird, von geringer Relevanz. Man könnte mit 15 Klausuren jenseits der zehn Punkte Unterschiede zwischen Korrektoren ebenso aufzeigen, wie mit 15 Klausuren unter vier oder mit einer Mischung aus jeder Notenstufe. Forschungsgegenstand hätte auch eine Klausur aus einem beliebigen anderen Rechtsgebiet sein können. Ob die Unterschiede zwischen Rechtsgebieten ähnlich sind, gilt es jetzt mit diesen Ergebnissen als Hintergrund zu untersuchen.

Repräsentativität ist demgegenüber ein statistisch unklar definierter Begriff, der eine Teilmenge einer Population beschreibt, die Rückschlüsse auf die Gesamtmenge zulässt.[32] Ausschlaggebend ist also, dass die gewählten Klausuren und die gewählten Korrektorinnen und Korrektoren eine Gesamtmenge repräsentieren, auf die von dem Experiment geschlossen werden soll. Die Klausuren sind im engsten Sinn nur für andere Bearbeitungen genau dieser Klausur als Population repräsentativ. Eine Abstraktion auf Anfängerklausuren generell ist erst möglich, wenn untersucht wird, inwieweit die Objektivität vom Rechtsgebiet bzw. der konkreten Klausur abhängt. Eine Abstraktion auf Ergebnisse des Staatsexamens ist nicht möglich. Wie im vorherigen Abschnitt beschrieben, ist dies der wichtigste nächste Forschungsschritt. In diesem Aufsatz möchte ich daher ausdrücklich keine Aussage zur Objektivität im Staatsexamen treffen und warne die Leserinnen und Leser auch davor, die Ergebnisse auf das Staatsexamen zu übertragen. Dieser Aufsatz trifft daher lediglich eine beschränkte Aussage für eine Probeklausur, wie sie in jeder Universität in Deutschland hätte gestellt werden können.

Die Klausuren sind für die Untersuchung besonders wertvoll, da sie computergeschrieben vorliegen, was den Faktor Handschrift – der aufgrund der fortschreitenden Verbreitung des getippten Examens ohnehin hinfällig zu untersuchen wäre – ausschließt. Die Klausuren wurden anhand einer breiten Notenverteilung ausgewählt. Innerhalb der Classic Test Theory ist die wahre Benotung einer Aufgabe nie messbar, sondern jede beobachtete Note besteht aus dem wahren Ergebnis plus einen Messfehler. Ich habe die Klausuren so ausgewählt, dass ich von einer gewissen Spannweite der Ergebnisse ausgehen konnte, ohne dabei davon auszugehen, dass meine Ergebnisse besonders nah am „wahren" Wert seien.[33] Keine Auswahl von Klausuren ist frei von Biases und die Auswahl soll auch in gewissem Maß

32 *Kaptein/van den Heuvel*, Statistics for Data Scientists, S. 43.

33 Meine Korrekturen habe ich mit Bewertungseinheiten angefertigt. Damit sind sie nur beschränkt mit freien Bewertungen vergleichbar. Ich habe meine Bewertungen auch nicht bei der Analyse berücksichtigt.

willkürlich sein. Jede Klausur ist eine echte, ernsthaft angefertigte Bearbeitung der Klausur, als solche ist jede gewählte Klausur ein tauglicher Bestandteil der Menge plausibler Weise zu korrigierender Klausuren.

Die Auswahl der Korrektorinnen und Korrektoren ist demgegenüber repräsentativ für die Gruppe der korrigierenden Personen auf Ebene von Anfängerklausuren. Ich habe mich dagegen entschieden, die Gruppe der Korrektorinnen und Korrektoren besonders homogen zu halten. Hätte man etwa nur Männer zwischen 20 und 30 Jahren genommen, die an der LMU München im öffentlichen Recht promovieren und bereits mindestens in einem Durchgang korrigiert haben, wäre die Gruppe zwar homogener, diese Gruppe würde jedoch nur in geringem Maß der Praxis entsprechen. Bei Grundkursklausuren mit bis zu 1000 Bearbeitungen müssen Lehrstühle auf eine breite Gruppe an Personen zurückgreifen, die je zwischen 10 und 100 Klausuren korrigieren. Hierbei sind zu jeder Zeit sowohl erstmalige als auch sehr erfahrene Korrektorinnen und Korrektoren tätig mit unterschiedlichen Examensnoten. Eine besondere Auswahl anhand von Alter, Korrekturerfahrung, Geschlecht oder gesteigerter Kenntnisse des Rechtsgebietes würden ein Entfernen von der Gesamtpopulation bedeuten.

II. Die Anzahl an Klausuren und Korrektorinnen und Korrektoren ist nicht hoch genug, um belastbar zu sein.

Eine höhere Anzahl an Korrektorinnen und Korrektoren und Klausuren wäre wünschenswert gewesen, war aber für den Sinn und Zweck der hier getroffenen Aussagen nicht nötig. In Ermangelung empirischer Referenzpunkte, wie hoch Abweichungen zwischen Korrektorinnen und Korrektoren sein könnten, hätte eine geringere Zahl Klausuren mit weniger Korrekturen bereits Unterschiede aufzeigen können. Wenn mehr Forschung zu dem Thema vorliegt, kann im Rahmen einer Metaanalyse eine abstrakte Basislinie der Abweichung bei juristischen Bewertungen als generalisierbarer *status quo* der Objektivität eventuell herausgearbeitet werden. Diese Basislinie könnte dann mit größeren Studien verfeinert werden. Die Unterschiede zwischen 23 verschiedenen Korrektoren reichen als erster Anhaltspunkt jedenfalls aus, um weitere Forschung zu Abweichungen anzuregen.

Die geringe Zahl an Bewertungen bewirkt aber, dass die in diesem Experiment ermittelte Wahrscheinlichkeit, mit der Noten um ein wahres Ergebnis herum vergeben werden nicht auf andere Klausuren abstrahierbar sind. Nach dem Gesetz der kleinen Zahlen[34] werden bei einer niedrigen Zahl an Beobachtungen gewisse eingetretene Ergebnisse mit höherer Wahrscheinlichkeit getroffen als bei einer hohen Zahl von Beobachtungen. Wenn die juristische Notenskala 19 mögliche Stufen hat, aber nur 16 Korrekturen vorliegen, sind zwangsläufig mindestens drei Punkte nicht vergeben worden, also mit einer Wahrscheinlichkeit von 0 belegt, während andere Ergebnisse, weil sie einmal vorgekommen sind, wahrscheinlicher erscheinen

34 *Rabin*, in: The Quarterly Journal of Economics, 117(3) (2002), S. 775 ff.

als sie das tatsächlich sind. Wenn bei Klausur B5 etwa eine 4 und eine 14 vergeben wurde, erscheinen diese Möglichkeiten im Verhältnis zu anderen Ergebnissen mit einer höheren Wahrscheinlichkeitsdichte als zu dominant. Bei B5 wurde etwa keine 6 oder keine 11 vergeben, was aber nicht bedeutet, dass diese statistisch nicht erwartbar seien. Erst bei einer erheblich größeren Menge an Korrekturen kann dieser Messfehler nivelliert werden und eine abstrakte Aussage über die wahre Verteilung von Ergebnissen bei jeder Bearbeitung juristischer Klausuren schlechthin getroffen werden. Somit kann nicht ausgeschlossen werden, dass die Ergebnisse hier – auf die gesamte juristische Korrekturpraxis bezogen – zufällig sind. Ob dies der Fall ist, gilt es in weiteren Experimenten, vorzugsweise mit höherer Teilnehmerzahl, zu erforschen.

III. Die Ergebnisse sind beeinflusst worden, da die Korrektorinnen und Korrektoren vorher von dem Experiment wussten.

Dieses Experiment wurde ohne Forschungsgelder betrieben. Der Anreiz der Korrektorinnen und Korrektoren bestand aus persönlicher Verbundenheit, Interesse an der Wissenschaft und eine Auswertung ihrer eigenen Korrekturleistung, die ich jeder teilnehmenden Person habe zukommen lassen. Um den Grund der Korrektur hinreichend zu erklären, musste ich jeder Person den Hintergrund des Experiments erläutern. Die Klausurbearbeitungen wurden ohne Wissen des Experiments verfasst. Bei den Korrektorinnen und Korrektoren lag durch die Erklärung aber keine blinde Studie mehr vor. Dies ist aus einem Grund unerheblich und verbessert die Ergebnisse aus einem anderen Grund sogar. Zum einen wurde jede Person gleichmäßig beeinflusst. Da ich allen dieselben Informationen zum Experiment gegeben habe, kann man bei allen Personen von einer gleichmäßigen Beeinflussung ausgehen, die dadurch vernachlässigbar wird. Zum anderen bewirkt die Art der Beeinflussung, nämlich das Bewusstsein, dass ihre Korrekturleistung untersucht wird, dass die Personen – wenn überhaupt anders – dann gewissenhafter korrigieren würden. Als Ergebnis der Beeinflussung sollten die Abweichungen also sinken und die Objektivität steigen. Alle Abweichungen, die sich trotz des Wissens über das Experiment zeigen, sind Anzeichen von ungewollten Unterschieden, die man selbst bei explizitem Steigern des Objektivitätsbewusstseins nicht willentlich ändern kann.

IV. In einem anderen Rechtsgebiet / im Staatsexamen / bei erfahreneren Korrektorinnen und Korrektoren wären die Unterschiede nicht so hoch.

Dieser Einwand wurde verstärkt als Reaktion auf die vorläufigen Ergebnisse vorgebracht und ist nicht unplausibel. Hierzu kann zum jetzigen Zeitpunkt von niemandem – außer vielleicht mancher LJPA mit unveröffentlichten Ergebnissen – eine empirisch belegte Aussage getroffen werden. Indiziert ist dies durch selten ver-

öffentlichte Einzelfälle[35] und § 13 Abs. 3 S. 6 JAVO Schleswig-Holstein, in dem einem Regelungsbedarf für Differenzen zwischen Korrektoren jenseits von 6 Punkten nachgekommen wird. Es bedarf für eine empirische Aufarbeitung einer Wiederholung des Experiments mit einer fünfstündigen Examensklausur, die von Personen korrigiert wird, die auch im Staatsexamen korrigieren. Bei diesen ist aufgrund der Erfahrung zu hypothetisieren (und zu hoffen), dass die Ergebnisse näher beieinander liegen als in der vorliegenden Studie.

Aus Studierendenperspektive ist diese Möglichkeit freilich nicht tröstlich, da man – wenn diese Annahme stimmt – erst im Examen, lediglich potentiell und bloß im Durchschnitt mit objektiveren Korrekturen rechnen kann. Das wäre aus Sicht des Systems zwar weniger bedenklich, aus Sicht einer bestmöglichen Ausbildung aber weiterhin suboptimal. Aus Perspektive der Korrektorinnen und Korrektoren erscheint eine höhere Objektivität bis zum Examen auch fraglich. Der universitäre Betrieb ist auf viel Unterstützung angewiesen, um die Masse an Korrekturen zu bewältigen. Eine Erfahrungsvoraussetzung würde schnell eine restriktiv hohe Einstiegshürden für neue Korrektorinnen und Korrektoren darstellen, zumal diese, um Erfahrung zu sammeln, auch an echten Klausuren üben müssen.

H. Ergebnis: mehr Forschung erforderlich

Die Ergebnisse sind besorgniserregend. Bei diesem Experiment betrug die durchschnittliche Abweichung zwischen niedrigster und höchster Note 6,47 Punkte. Bei durchschnittlicher Standardabweichung liegen nur 42% der gegebenen Noten innerhalb derselben Notenstufe.

Die Studie ist ein erster Ansatzpunkt für mehr Forschung, der auf einer isolierten Betrachtung einer einzigen Anfängerklausur als Teils des Prüfungsgeschehens basiert. Es sei daher nochmal davor gewarnt, die Ergebnisse auf andere Korrekturen zu verallgemeinern. Die Ergebnisse indizieren aber Probleme mit der Chancengleichheit bei juristischen Klausurbewertungen. Dieses Indiz sollte ernst genommen werden, ohne es dabei in seiner Tragweite zu überschätzen. Über eine Indizwirkung hinaus können Aussagen über das gesamte Prüfungssystem zum jetzigen Zeitpunkt nicht getroffen werden, insbesondere nicht, ob diese Unterschiede bei Examensklausuren auch bestehen.

Ziel dieser Forschung ist das im Grundsatz gute und erprobte Prüfungssystem zu verbessern, indem mögliche Defizite aufgezeigt werden. Eine Verbesserung kann erreicht werden, wenn die Rechtswissenschaft als Ergebnis dieser Studie ihr eigenes Prüfungssystem genauer erforscht und dadurch herausfindet, ob ein Handeln in Form von Anpassungen der Prüfungspraxis geboten ist. Veränderungen des Prüfungssystems sind so lange nicht geboten, bis die folgenden Punkte weiter erforscht sind:

35 Etwa *Walter*, in: BayVBl 2023, 689 ff., wo der Stichentscheid 6 Punkte unter der Erstkorrektur lag.

1. Ist die mangelnde Objektivität ein isoliertes Problem oder tritt dies in verschiedenen Rechtsgebieten und Stadien der Bewertung bis hin zum Staatsexamen auf?
2. Kann der hier gefundene Umfang der Abweichungen in anderen Erhebungen bestätigt werden?
3. Was sind mögliche Ursachen dieser geringen Objektivität?
4. Was ist das Ziel einer juristischen Prüfung? Insbesondere: Wie kann eine Prüfung juristisches Können abprüfen?
5. Wie können wir das Prüfsystem objektiver machen? Reicht ein Anpassen der Art und Weise des Benotungssystems oder müssen alternative Prüfungsformen genutzt werden?
6. Wie können beim Anpassen des jetzigen Systems Übergangskosten für Studierende und die juristische Arbeitswelt minimiert werden?

Um diese Fragen schnellstmöglich zu klären, ermuntere ich auch andere Juristinnen und Juristen, sich vielfältiger, auch empirischer Forschungsmethoden zu bedienen und damit gemeinsam einen Beitrag zur Verbesserung des juristischen Prüfungssystems zu leisten.

Diese Studie hatte noch einen interessanten Nebeneffekt. Die 15 Bearbeitungen können nun als Gradmesser für Strenge oder Milde einzelner Korrektorinnen und Korrektoren genutzt werden. Dieser Ansatz bietet Entwicklungsmöglichkeiten: Feedback zur Korrektur, Korrekturtraining und ein System zur Prüfung von Korrekturleistung.

Literaturverzeichnis

Botero, Juan Carlos/Agrast, Mark David/Ponce, Alejandro, The World Justice Project Rule of Law Index 2022, Washington D.C. 2022. https://worldjusticeproject.org/rule-of-law-index/downloads/WJPIndex2022.pdf (18.12.2023).

Brown, M. B./Forsythe, A. B., Robust Tests for the Equality of Variances, in: Journal of the American Statistical Association 69 (1974), S. 364-367.

Bühner, Markus, Einführung in die Test- und Fragebogenkonstruktion, Bloomington 2021.

Cappelleri, Joseph C./Lundy, J. Jason/Hays, Ron D., Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures, in: *Clinical Therapeutics* 36.5 (2014), S. 648-662.

Church, W. Lawrence, Forum – Law School Grading, in: Wis.L.Rev. (1991), S. 825-833.

Crane, Linda R., Grading Law School Examinations: Making a Case for Objective Exams to Cure What Ails Objectified Exams, in: 34 NEW ENG. L. REV. (2000), S. 785-808.

Dahiru, Tukur, P-value, a true test of statistical significance? A cautionary note., in: *Annals of Ibadan postgraduate medicine* 6.1 (2008), S. 21-26.

Delgado-Rodriguez, Miguel/Llorca, Javier, Bias, in: Journal of Epidemiology & Community Health 58.8 (2004), S. 635-641.

Eckstein, Brigitte, in: Schütz, Mathias / Skowronek, Helmut / Thieme Werner (Hrsg.), Prüfungen als hochschuldidaktisches Problem Ergebnisse und Materialien eines Expertenseminars in Hamburg-Rissen vom 31.1-2.2.1969. Bertelsmann, Bielefeld 1969.

Geiser, Christian, Datenanalyse mit Mplus: eine anwendungsorientierte Einführung. 2. Aufl., Wiesbaden 2011.

Hattenhauer, Hans, Juristenausbildung – Geschichte und Probleme, in: JuS (1989), S. 513-520.

Heidebach, Martin, Prüfen im rechtswissenschaftlichen Studium: Die Korrektur juristischer Hausarbeiten anhand eines verbindlichen Bewertungseinheiten-Systems, in: ZDRW 3 (2015), S. 205-214.

Kaptein, Marits/van den Heuvel, Edwin, Statistics for Data Scientists. An Introduction to Probability, Statistics and Data Analysis. Cham 2022.

Kaufman, Nancy H., A Survey of Law School Grading Practices, in: Journal of Legal Education, 44(3) (1994), S. 415-423.

Levene, H., Robust tests for Equality of Variances, in: I. Olkin et al. (Hrsg.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Redwood City 1960, S. 278-292

Novick, M.R., The axioms and principal results of classical test theory, in: Journal of Mathematical Psychology 3(1) (1966), S. 1-18.

Pilniok, Arne, Strukturen juristischen Wissens, in: Krüper (Hrsg.), Rechtswissenschaft lehren, Tübingen 2022, S. 184-213.

Posser, Herbert/Wolff, Heinrich Amadeus / Decker, Andreas, BeckOK VwGO, 64. Edition, Stand: 01.04.2023.

Shrout, Patrick E./Fleiss, Joseph L., Intraclass correlations: uses in assessing rater reliability, in: Psychological bulletin 86(2) (1979), S. 420-428.

Rabin, Matthew, Inference by Believers in the Law of Small Numbers, in: The Quarterly Journal of Economics 117(3) (2002), S. 775-816.

Towfigh, Emanuel/Traxler, Christian/Glöckner, Andreas, Zur Benotung in der Examensvorbereitung und im ersten Examen - Eine empirische Analyse, in: ZDRW 1 (2014), S. 8-27.

Walter, Christian, Zur rechtswidrigen Praxis des Stichentscheids in der Ersten Juristischen Staatsprüfung in Bayern, in: BayVBl (2023), S. 689-691.