Juristenausbildung

Messgenauigkeit und Fairness in Staatsprüfungen

Aktuelle Studien zeigen Gruppen-Unterschiede in juristischen Examina auf

Prof. Dr. Andreas Glöckner, Hagen und Prof. Dr. iur. Emanuel V. Towfigh, Wiesbaden

Die Prüfungsangst gehört zur klassischen Juristenausbildung in Deutschland dazu wie der Repetitor zum Jura-Studium. Noch immer wird der Note im ersten und im zweiten Staatsexamen eine besondere Bedeutung zugemessen – nicht zuletzt auch von den Absolventen. Doch wie aussagekräftig sind Noten in Staatsprüfungen überhaupt? Aktuelle Forschungsprojekte kratzen am Nimbus der Objektivität, Reliabilität und Validität. Die Autoren – ein Psychologe und ein Jurist – stellen ihre Forschungsvorhaben und Ergebnisse der Forschung vor. Auch wenn das empirische Ausleuchten der juristischen Staatsprüfungen erst begonnen hat, zeigen sich bereits heute Gruppen-Unterschiede in den Noten, deren Ursache weiter geklärt werden müssen. Die Forschung der Autoren könnte Auslöser für eine Reform des juristischen Prüfungswesens werden.

Wie gut – sprich messgenau und fair – sind die Benotungen in der ersten juristischen Staatsprüfung und im zweiten juristischen Examen? Erhalten Studierende, die sich vergleichbare juristische Fähigkeiten während ihres Studiums angeeignet haben, zumindest im Durchschnitt die gleichen Noten? Unterscheidet sich die Notengebung nach dem Geschlecht der Studierenden oder stehen sie vielleicht mit einem Migrationshintergrund im Zusammenhang?

Im Folgenden sollen aktuelle empirische Untersuchungen, die zur Beantwortung dieser Fragen beitragen, und ihre methodischen Hintergründe zusammengefasst werden. Zunächst werden die aus fachlich-diagnostischer Sicht relevanten Aspekte zur empirischen Bewertung der Präzision und Fairness akademischer Prüfungen im Allgemeinen erläutert (dazu I.). Anschließend stellen wir exemplarisch eine aktuelle französische Studie zur Geschlechterfairness in staatlichen Prüfungen vor, die mit Blick auf Methode und Ergebnisse maßstabsetzend ist (dazu II.). Im dritten Teil dieses Artikels gehen wir auf zwei empirische Analysen der Benotung im staatlichen Teil der Ersten Juristischen Prüfung in Nordrhein-Westfalen und Baden-Württemberg ein, die neben Geschlechterunterschieden auch weitere potenzielle Einflussfaktoren – etwa Migrationshintergrund und Studienort – untersuchen. Der Artikel schließt mit einer kurzen Zusammenfassung sowie einer Diskussion des weiteren Handlungsbedarfs zur Gewährleistung ausreichender Transparenz bezüglich der Messgenauigkeit und Fairness in juristischen Staatsprüfungen (dazu IV.).

I. Die Qualität von Prüfungen kann (und sollte) empirisch bestimmt werden

Lassen sich die in einem Studium erworbenen akademischen Fähigkeiten und die aus ihnen resultierenden Leistungen zuverlässig messen? Wie bei jeder anderen Messung lassen sich auch in diesem Bereich Messfehler – und seien sie noch so klein – nie ausschließen. Die psychologische Diagnostik, ein Teilgebiet der Psychologie, setzt sich unter anderem damit auseinander, wie die Güte von Messungen solcher und anderer Merkmale von Menschen mittels psychometrischer Tests empirisch untersucht – und wenn notwendig: verbessert – werden kann.[1] Dazu werden im Rahmen von Theorien (wie beispielsweise der klassischen Testtheorie) Annahmen darüber getroffen, in welchem Zusammenhang beobachtete Messwerte (beispielsweise die Leistung in der Prüfung), „wahre Werte" (beispielsweise die tatsächliche juristische Fähigkeit der Kandidatin oder des Kandidaten) und Messfehler zueinander stehen. Dies ermöglicht eine empirische Analyse der Güte psychometrischer Tests, unter die eben auch fachbezogene Staatsprüfungen subsumiert werden können. Die wichtigsten Anforderungen an solche Tests sind (1.) Objektivität, (2.) Reliabilität und (3.) Validität, welche auch als Hauptgütekriterien bezeichnet werden.

1. Objektivität

Die Objektivität bezeichnet die relative Unabhängigkeit des gemessenen Wertes vom Durchführenden der Messung (beispielsweise Prüferinnen und Prüfer oder Korrektorinnen und Korrektoren der Klausur). Objektivität kann durch einen hohen Grad der Standardisierung in Durchführung und Bewertung (beispielsweise durch Vorgabe eines Lösungsschlüssels) und durch multiple Beurteiler (beispielsweise den Einsatz mehrerer *unabhängiger* Prüferinnen und Prüfer, die also ohne Kenntnis der Bewertung der oder des anderen urteilen) gewährleistet werden. Der Zusammenhang der Einschätzungen mehrerer unabhängiger Beurteilerinnen und Beurteiler ermöglicht es, den Grad der Objektivität eines Tests zu messen. Eine hoher Grad an Übereinstimmung hinsichtlich verschiedener Kandidatinnen und Kandidaten, üblicherweise gemessen als Korrelationskoeffizient[2], sollte vorliegen, um (einen zentralen Aspekt) der Objektivität nachzuweisen.

1 Für eine ausführliche Darstellung siehe bspw. das Standard-Lehrbuch von *Lothar Schmidt-Atzert/Manfred Amelang*, Psychologische Diagnostik (Lehrbuch mit Online-Materialien), in dem auch die im Folgenden umrissenen Konzepte ausführlich erläutert sowie axiomatisiert und mathematisch hergeleitet werden. Vgl. auch *Towfigh/Petersen*, Ökonomische Methoden im Recht. Eine Einführung für Juristen, S. 201 ff.

2 Eine Korrelation bezeichnet einen statistischen Kennwert, der den Zusammenhang zweier Variablen auf einer Skala von -1 (perfekt negativer Zusammengang) bis 1 (perfekt positiver Zusammenhang) beschreibt. Null indiziert dabei, dass kein Zusammenhang zwischen den Variablen besteht. Mathematisch lässt sich die der Korrelationskoeffizient als Steigung einer linearen Funktion ausdrücken. Die Höhe der zu fordernden Objektivität ist dabei vom Messinstrument abhängig. Als grober Richtwert sollte eine Korrelation von r >= .80 gelten, um von einem robusten Zusammenhang ausgehen zu können.

2. Reliabilität

Die Reliabilität bezeichnet die Messgenauigkeit eines Tests. Diese wird bestimmt durch den Anteil der messfehlerbedingten Schwankungen an den Schwankungen der beobachteten Messwerte. Wie erwähnt, ist keine Messung fehlerfrei. Messfehler, die dazu führen, dass der Messwert – beispielsweise die Prüfungsleistung – nicht mit der tatsächlichen Merkmalsausprägung – beispielsweise der juristischen Fähigkeit – übereinstimmt, können dabei in der Person (beispielsweise Tagesform des Prüflings), der Situation (beispielsweise störende Hitze für alle Prüflinge) oder deren Zusammentreffen liegen. Letzteres kann beispielsweise auftreten, wenn sich ein Kandidat lediglich auf eine Aufgabe vorbereitet hat, die in der Prüfung zufällig aufgegriffen wird, so dass die allgemeine juristische Fähigkeit des Kandidaten tendenziell überschätzt wird. Die Reliabilität kann unter anderem dadurch bestimmt werden, dass die Ergebnisse von unabhängigen Teilen eines Tests miteinander korreliert werden (zum Beispielen Test-Hälften-Reliabilität) oder vergleichbare Tests wiederholt durchgeführt werden (zum Beispiel Test-Wiederholungs-Reliabilität). Die Reliabilität kann ebenso wie die Objektivität als Korrelationskoeffizient ausgedrückt werden, für den Nachweis der Messgenauigkeit werden hohe Korrelationen eines Tests gefordert.[3]

3. Validität

Die Validität eines Tests bezeichnet, ob dieser Test auch wirklich das Merkmal misst, das er zu messen vorgibt. Misst eine Prüfung tatsächlich die intendierten (beispielsweise juristischen) Fähigkeiten? Da die tatsächliche Fähigkeit üblicherweise nicht direkt zugänglich ist, müssen andere Kriterien herangezogen werden, um die Validität eines Tests nachzuweisen. Beispielsweise sollte die Leistung in der Prüfung mit anderen Tests für dieselbe Fähigkeit (beispielsweise Probeklausuren in der Examensvorbereitung oder den Vornoten), dem späteren Berufserfolg und auch mit der Einschätzung von Experten (beispielsweise betreuenden Professorinnen und Professoren) korrelieren. Üblicherweise wird eine Korrelation mittlerer Höhe mit verschiedenen solcher Validierungskriterien als Nachweis der sogenannten *konvergenten Validität* gefordert. Dabei ist mit etwas niedrigeren Zusammenhängen zu rechnen, da auch die zur Validierung herangezogenen Kriterien das tatsächliche Merkmal nur fehlerbehaftet und nicht vollumfänglich abbilden.[4]

Von großer praktischer Bedeutung für den hiesigen Zusammenhang ist darüber hinaus noch der Nachweis der sogenannten *diskriminanten Validität,* nämlich der Tatsache, dass der Test auch tatsächlich ausschließlich das Merkmal misst, das er zu messen vorgibt. Dies kann geprüft werden, indem man untersucht, ob ein Messwert auch mit anderen Merkmalen zusammenhängt (korreliert). Es ist beispielsweise hypothetisch möglich, dass eine Prüfung perfekt objektiv ist (alle Beurteilerinnen und Beurteiler kommen zur selben Note), darüber hinaus auch perfekt reliabel ist (die Beurteilung der Leistung in allen Teilaufgaben korrelieren perfekt), und dass auch konvergente Validität nachgewiesen werden kann (das Prüfungsergebnis korreliert hoch mit Berufserfolg und vergleichbaren Tests)– und dass trotzdem ein Problem durch eingeschränkte Validität vorliegt: Der Messwert könnte neben der tatsächlichen Fähigkeit noch durch Stereotype (beispielsweise die Vorstellung, ein Geschlecht sei leistungsstärker als das andere in einem bestimmten Fach) oder auch gesellschaftlich geteilte Ziele (beispielsweise, dass mehr Diversität und ein ausgeglichenes Geschlechterverhältnis erreicht werden sollen) beeinflusst werden. Die Testung diskriminanter Validität ist deshalb ausgesprochen wichtig, da Tests nur messen sollen, was sie zu messen vorgeben.

Aus fachlich-diagnostischer Sicht wird verlangt, dass jeder Test die Hauptgütekriterien Objektivität, Reliabilität und Validität erfüllt und dass die – üblicherweise nur imperfekte – Einhaltung dieser Kriterien auch empirisch nachgewiesen und für den Testteilnehmer transparent gemacht wird.

II. Benotung staatlicher Examen in Frankreich: Differenzierte Geschlechterdiskriminierung

Thomas Breda und *Mélina Hillion* haben vor wenigen Wochen in *Science* eine Studie publiziert, in der differenzierte Geschlechter-Effekte hinsichtlich der Benotung staatlicher Examina in Frankreich nachgewiesen werden, die vom Geschlechterverhältnis der in dem jeweiligen Fach bereits tätigen Personen abhängig ist.[5] Die Studie stellt einen wichtigen Nachweis der Verletzung der Forderung nach diskriminanter Validität in staatlichen Prüfungen dar, welche in Verlauf und Struktur große Ähnlichkeit mit entsprechenden staatlichen Examina in Deutschland aufweisen. In der Studie wurden die Noten von mehr als 100.000 Personen untersucht, die von 2006 bis 2013 an den fachspezifischen staatlichen Examina teilgenommen haben, die bei der Auswahl fast aller französischen Lehrer der Sekundar- und Oberstufe sowie von Professoren eingesetzt werden.[6] Der Vergleich der schriftlichen Noten aus den (jeweils aus mehreren Teil-Examina bestehenden) anonymen und deshalb „geschlechterblinden" schriftlichen Prüfungen mit den Noten in den mündlichen Prüfungen, in denen das Geschlecht naturgemäß ersichtlich war, ergab bei Letzteren eine systematische Bevorzugung der Personengruppe, die in dem jeweiligen Fachbereich personell unterrepräsentiert war. So wurden Frauen in Mathematik, Physik und Philosophie in der mündlichen Prüfung durchschnittlich 10 Prozent Rangplätze besser eingeschätzt als in der anonymen schriftlichen Prüfung. Der umgekehrte Effekt findet sich für die Beurteilung von Männern in Literatur und Fremdsprachen, wo Männer im Rahmen mündlicher Prüfungen um 3 Prozent bis 5 Prozent Rangplätze besser abschneiden.

3 Die genaue Höhe ist dabei abhängig davon, welche Anforderungen an den Test gestellt werden. Aber eine Reliabilität von r >= .70 bildet eine übliche Mindestanforderung.

4 Richtwerte für einen Nachweis der Validität sind mehrere Korrelationen mit r >= .40 für Außenkriterien und – abhängig von deren Ähnlichkeit – etwas höhere Korrelationen mit vergleichbaren Tests.

5 *Thomas Breda/Melina Hillion*, Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France, Science 353 (2016), S. 474.

6 CAPES und Agrégation. Die fachspezifischen Examen werden für 11 unterschiedliche Fächer durchgeführt. Weitere Analysen umfassten auch das CRPE.

Interessanterweise zeigt sich der Zusammenhang zwischen Geschlechteranteil in den Fächern und Benotung auch in einer zusätzlichen *fachunabhängigen* mündlichen Prüfung, in der die Prüflinge aller Fächergruppen dieselben Aufgaben bearbeiteten. Diese Übertragung der Bevorteilung des unterrepräsentierten Geschlechts auf einen eigentlich geschlechterneutralen mündlichen Test erlaubt es, die prominentesten Alternativerklärungen (unter anderem Erkennen von Handschriften, geschlechterspezifische Fähigkeitsunterschiede im jeweiligen Fach) auszuräumen und spricht stark für eine systematische Bevorzugung der unterrepräsentierten Geschlechtergruppe in den jeweiligen Prüfungen.

Zusammenfassend liefert die Studie von *Breda* und *Hillion* (2016) starke Evidenz für systematische, aber nach Fächergruppen differenzierte Geschlechterdiskriminierung in französischen Staatsexamen in 11 Fächern. Auch wenn die damit möglicherweise implizit intendierte Erreichung eines ausgeglichenen Geschlechterverhältnisses als gesellschaftlich erstrebenswert angesehen werden kann, ist eine damit verbundene Verunreinigung einer Messung aus fachlich-diagnostischer Perspektive kritisch zu beurteilen. Stattdessen sollten die Prozesse der Messung und der gesellschaftlich wünschenswerten Verarbeitung der Ergebnisse jedenfalls aus diagnostischer Sicht voneinander getrennt werden.

III. Juristische Staatsexamen in Nordrhein-Westfalen: Bessere Noten für Männer und Personen ohne Migrationshintergrund

In einer eigenen Studie[7] untersuchten wir unter anderem Unterschiede in der Benotung im staatlichen Teil der Ersten Juristischen Prüfung anhand von Noten für den Zeitraum von 09/2007–12/2010 (N = 2.217). Es zeigte sich, dass Frauen – gemessen an der Abiturnote[8]– mit den besseren Voraussetzungen in das Studium starten (arithmetisches Mittel [M] = 2,05 bei Frauen vs. M = 2,22 bei Männern [Schulnoten]), aber mit schlechteren Examensnoten abschließen (M = 7,33 vs. M = 7,62 Punkte). Weitere Analysen zeigten, dass bei Kontrolle (unter anderem) für die Unterschiede in der Abiturnote Männer im Schnitt um 0,7 Punkte (ca. 10 Prozent der Durchschnittsnote) bessere Examensnoten erzielten als Frauen. Dieser Effekt ist stärker in der mündlichen als in der unter Kennziffer (und somit anonym) abgelegten schriftlichen Prüfung. Ferner bleibt bei den Frauen selbst dann ein Malus von 0,24 Punkten erhalten, wenn man bei der Vorhersage der mündlichen Note für die schriftliche Note kontrolliert, also das Ergebnis der schriftlichen Note bei der Analyse der mündlichen Ergebnisse berücksichtigt. Anders ausgedrückt: Wenn man mittels der in unserem Datensatz vorhandenen Variablen (hypothetische) „statistische Zwillinge" bildet, die sich lediglich hinsichtlich ihres Geschlechts unterscheiden, sonst aber gleich alt sind, die gleiche Abiturnote erzielt haben, dieselben Klausuren schreiben, an derselben Uni studiert haben und dieselben Vornoten erzielt haben, so wird ein männlicher Kandidat in der mündlichen Prüfung, in der das Geschlecht ersichtlich ist, um 0,24 Punkte besser benotet als eine weibliche Kandidatin.

Ähnlich wie in der oben beschriebenen Studie in Frankreich zeigen sich also systematische Unterschiede in der Benotung von männlichen und weiblichen Prüflingen auch im ersten juristischen Examen in Deutschland. Allerdings ist darauf hinzuweisen, dass durch die deutlich kleinere Stichprobe und insbesondere durch die im Vergleich zur Studie in Frankreich eingeschränkteren statistischen Kontrollmöglichkeiten die *Ursachen* für diesen Unterschied nicht mit ausreichender Sicherheit bestimmbar sind. Die Untersuchung der Ursachen für diese Unterschiede wird im Rahmen eines durch das Justizministerium des Landes Nordrhein-Westfalen geförderten Projekts aktuell weiter vorangetrieben, welches unter anderem auch das zweite Examen und die Geschlechterzusammensetzung der Prüfungskommission mit in die Analyse einbezieht.

Neben den genannten Analysen untersuchten wir mittels des Datensatzes vom OLG Hamm ferner, ob ein etwaiger Migrationshintergrund einen Einfluss auf die Benotung hat. Als Indikation für einen potenziellen Migrationshintergrund wurde – unter strengen datenschutzrechtlichen Auflagen – eine onomastische Kodierung vorgenommen, die es erlaubt, die Kandidatinnen und Kandidaten in Personen mit deutscher beziehungsweise nicht-deutscher Namensherkunft zu unterteilen. Unter Verwendung dieser Kodierung zeigte sich, dass Personen, deren Name auf einen Migrationshintergrund schließen lässt, zwar ähnliche Abiturnoten aufwiesen wie Personen mit einer deutschen Namensherkunft (M = 2,11 vs. M = 2,04) aber schlechtere Examensnote erhielten als letztere (M = 7,01 vs. M = 7,74 Punkte). Bei „Konstanthaltung" der Abiturnote und weiterer Faktoren – also wiederum durch Erzeugung „statistischer Zwillinge"– blieb ein Notenunterschied bestehen, der wiederum in der mündlichen Prüfung größer war als in der schriftlichen Prüfung. Die Differenz in der mündlichen Note mit Sichtbarkeit des Merkmals (Name, der auf Migrationshintergrund schließen lässt) bei Kontrolle für die schriftliche Note beträgt dabei bis zu 0,43 Punkten.

Eine von *Hinz* und *Röhl*[9] vorgelegte Studie in Baden-Württemberg fand zunächst ebenfalls schlechtere Noten für Personen mit einem Namen, der auf einen Migrationshintergrund hinweist. Allerdings verschwand der Unterschied in der mündlichen Note bei Kontrolle für die schriftliche Note und erweiterte Kontrollfaktoren, wie zum Beispiel den sozioökonomischen Status. Die unterschiedlichen Ergebnisse könnten dadurch bedingt sein, dass Migrationshintergrund und die zusätzliche Kontrollvariable für sozioökonomischen Status untereinander korreliert sind, und der Effekt des Migrationshintergrunds dadurch unterschätzt wird, aber auch Unterschiede zwischen Bundesländern sind denkbar (siehe dazu auch die von *Hinz* und *Röhl* berichteten unterschiedlichen Effekte für unterschiedliche Studienorte). In jedem Fall

7 *Towfigh/Glöckner/Traxler*, Zur Benotung in der Examensvorbereitung und im ersten Examen, ZDRW 2014, S. 8 ff. Siehe auch *Glöckner/ Towfigh/Traxler*, The development of legal expertise, Instructional Science 2013, S. 989 ff. für eine ausführliche Analyse und Diskussion der Entwicklung rechtlicher Expertise im Klausuren-Kurs.

8 Wie bei dieser Annahme unterstellt und in verschiedenen Studien sehr gut nachgewiesen vgl. bspw. *Hinz/ Röhl*, Juristische Fakultäten in Baden-Württemberg: Wo studiert man am besten? VwBlBW 2016, S. 20 ff., bestand ein hoher Zusammengang zwischen der Abiturnote und der Abschlussnote mit einer Korrelation von r = -.45.

9 *Hinz/Röhl*, Juristische Fakultäten in Baden-Württemberg: Wo studiert man am besten?, VBlBW 2016, S. 20 ff.

scheint es zu früh für eine pauschale Entwarnung; weitere Analysen scheinen geboten. Außerdem wurden in der Studie von *Hinz* und *Röhl* ebenso wie in unserer erwähnten Studie Unterschiede in den Examensnoten für unterschiedliche Studienorte identifiziert, die sich zwar bei Kontrolle für die Zusammensetzung der Studierendenschaft verringerten, aber nicht komplett verschwinden.

IV. Zusammenfassung und Ausblick

Wie jede andere Art psychometrischer Tests, müssen sich auch staatliche Prüfungen akademischer Fähigkeiten der Herausforderung stellen, das interessierende Merkmal möglichst objektiv, reliabel und valide zu messen. Tests können nicht perfekt sein, sie sollten jedoch so gut, das heißt so präzise und fair wie möglich sein. Eine ausreichende Überprüfung der Hauptgütekriterien und die transparente Dokumentation der Ergebnisse sind notwendig, um Fairness und Präzision nachzuweisen und nachprüfbar zu dokumentieren. Wie bei der Einführung jedes anderen Tests auch, sollten diese aus fachlich-diagnostisch Sicht unstrittigen Anforderungen so weit wie möglich erfüllt werden, nicht zuletzt im Hinblick auf die oft mit erheblichen (unter anderem finanziellen) Konsequenzen verbundenen staatlichen Prüfungen.[10] Die Analysen sind leicht realisierbar. Im Sinne eines kritischen Realismus sollte die Angst vor Abweichungen von Perfektion abgelegt und das erreichte Ausmaß an Präzision und Fairness explizit benannt werden; das auch, um die Prüfverfahren verbessern zu können.

Bezüglich der Anforderung an die Fairness von Prüfungen demonstriert die französische Studie unter bestmöglicher statistischer Kontrolle, dass gesellschaftlich erwünschte Entwicklungen – wie der Ausgleich von Geschlechterverteilungen in verschiedenen Fächern – in die Benotung in staatlichen Examen einfließen können. Auch wenn dies zur Realisierung gesellschaftlich erwünschter Ziele beitragen mag, ist diese Art der Verletzung der diskriminanten Validität eines Tests bedenklich, da die Prüfung dann nicht mehr „ehrlich" ist und die staatliche Prüfung nicht mehr misst, was sie zu messen vorgibt. Wenn eine Ungleichbehandlung gesellschaftlich erwünscht ist, sollte diese offen diskutiert und transparent gemacht werden.

Unsere Untersuchung der Examensnoten zeigt auch in deutschen juristischen Staatsprüfungen erhebliche Geschlechterunterschiede sowie schlechtere Noten für Personen, deren Namen auf einen Migrationshintergrund hindeutet. Problematisch ist dabei insbesondere die Tatsache, dass ähnlich wie bei der Studie aus Frankreich ein Unterschied in der mündlichen Prüfung nachgewiesen werden kann, auch wenn man für die schriftliche Note, das Abitur und sonstige Faktoren kontrolliert. Allerdings erlauben die gegenwärtigen Analysen noch keine Schlussfolgerungen über die *Ursachen* der Unterschiede. Neben (unbewusster) Diskriminierung in der Prüfung sind auch weitere Ursachen vorstellbar, beispielsweise eine eingeschränkte Leistungsfähigkeit von Prüflingen durch die gedankliche Beschäftigung mit einem für die Prüfungssituation relevanten Stereotyp[11], eine denkbare Bevorzugung von Frauen oder Migranten im Abitur oder auch andere gruppenspezifische Reaktionen auf mündliche Prüfungen (beispielsweise erhöhte Prüfungsangst). Auch wurde der „Migrationseffekt" in der Studie in Baden-Württemberg nicht repliziert. In beiden Fällen sind weitere Untersuchungen notwendig, um die Erfüllung der Anforderung der Fairness des Tests nachzuweisen und – wenn notwendig – Maßnahmen zur weiteren Optimierung einzuleiten. Unabhängig davon sollten die Objektivität, Reliabilität und Validität staatlicher juristischer Prüfungen untersucht und transparent kommuniziert werden.

10 Eine aktuelle Studie weist bereits kurz nach dem Studium einen erheblichen Gehaltsunterschied von 14 % für Personen mit einem Prädikatsexamen nach im Vergleich zu Personen die kein Prädikatsexamen erreicht haben, was die Autoren unter Nutzung verschiedener Kontrollfaktoren auf die Signalwirkung des Prädikats zurückführen: *Freier u.a.*, The earnings returns to graduating with honors– Evidence from law graduates, Labor Economics 34 (2015), S. 39 ff.

11 Siehe beispielsweise die Ergebnisse von *Spencer u.a.*, Stereotype threat and women's math performance, Journal of Experimental Social Psychology 35 (1999), S. 4 ff. zum Effekt der empfundenen Stereotypen-Bedrohung auf Leistungen von Mädchen bei der Lösung von Mathematik-Aufgaben.

Prof. Dr. phil. Andreas Glöckner, Hagen
Der Autor ist Inhaber des Lehrstuhls für Allgemeine Psychologie: Urteilen, Entscheiden, Handeln an der FernUniversität in Hagen sowie Senior Research Fellow am Max-Planck-Institut zur Erforschung von Gemeinschaftsgütern in Bonn.

Leserreaktionen an anwaltsblatt@anwaltverein.de.

Prof. Dr. iur. Emanuel V. Towfigh, Wiesbaden
Der Autor ist Inhaber des Lehrstuhls für Öffentliches Recht, Empirische Rechtsforschung und Rechtsökonomik an der EBS Law School sowie Professor für Law & Economics an der EBS Business School, Wiesbaden und Research Affiliate am Max-Planck-Institut zur Erforschung von Gemeinschaftsgütern in Bonn.

Leserreaktionen an anwaltsblatt@anwaltverein.de.