Erhöhte Vergleichbarkeit von Noten durch Steuerungsmöglichkeiten bei Abschlussprüfungen an Hochschulen - am Beispiel kommissionsspezifischer Einflussgrößen[1]

Elena Tsarouha

Abstract

Prüfungen an Hochschulen haben mehrere Funktionen und dienen u. a. dazu, eine Bestenauslese der Studierenden zu gewährleisten. Dabei werden die erzielten Noten als Leistungsindikator verwendet. Das bedeutet, dass Zugänge im Hochschulwesen und im Berufsleben über Noten reguliert werden, obwohl die Notengebung in Deutschland bereits seit der Bildungsreform in den 1970er Jahren kritisiert wird und nationale und internationale Studien systematische Notenunterschiede in den Hochschulen bestätigen. Diese systematischen Notenunterschiede können durch verschiedene Prüfungspraktiken beeinflusst sein. Auf der Datenbasis von vier problemzentrierten Einzelinterviews, zwei Experteninterviews und neun Gruppendiskussionen wurde mittels der dokumentarischen Methode eine Typologie an Einflussgrößen erstellt. Im vorliegenden Beitrag werden ausgewählte Einflussgrößen des Typs der kommissionsspezifischen Einflussgrößen vorgestellt. Anhand der gewählten Beispiele werden Prüfungspraktiken beschrieben und deren potentieller Einfluss auf die Leistungsbeurteilung und Leistungsbewertung in Abschlussprüfungen an deutschen Hochschulen dargestellt. Anschließend werden Steuerungsmöglichkeiten der Hochschulen aufgezeigt und diskutiert, die zu einer höheren Vergleichbarkeit von Noten beitragen können. Im Fazit werden Grenzen der Einflussnahme angesprochen. Vor dem Hintergrund einer beschränkten Aussagekraft und Vergleichbarkeit von Prüfungsnoten werden Maßnahmen präsentiert, die eine Interpretation von erzielten Noten unterstützen.

[1] Der vorliegende Beitrag basiert auf einer Dissertationsschrift, welche im Springer VS Verlag unter dem Titel *Prüfungspraktiken an deutschen Hochschulen. Eine empirische Studie zu systematischen Einflussgrößen auf die Notengebung in Abschlussprüfungen* erschienen ist (Tsarouha 2019).

1 Darstellung des Forschungsstands und der Fragestellung

Die Notengebung wird im deutschen Hochschulkontext bereits im Zuge der Bildungsreform in den 1970er Jahren in Frage gestellt und damit auch die Aussagekraft und Vergleichbarkeit von Noten angezweifelt (Kvale 1972; Lämmert 1981; Hitpass u. Trosien 1987). Seitdem ist das Forschungsgebiet Prüfungen und Noten im nationalen Hochschulkontext kaum erforscht worden (Wissenschaftsrat 2003, 2007, 2012). Anders hingegen werden die Themen Noten und *grade inflation* vor allem in US-amerikanischen Studien seit den 1960er Jahren erforscht (Juola 1976; Kuh u. Hu 1999; Hu u. Kuh 2003; Johnson 2003). Im Rahmen des DFG-Projekts *Die Notengebung an Hochschulen in Deutschland* wurde das nationale Forschungsdesiderat aufgegriffen (Müller-Benedict u. Grözinger 2017).

Noten haben mehrere Funktionen, wie z. B. die Selektion, Rekrutierung, Prognose zukünftiger Leistungen und eine didaktische/pädagogische Funktion (Kvale 1972; Prahl 1995). Trotz aller Kritik werden Noten weiterhin als Leistungsindikator und Ausschlusskriterium verwendet. Noten regulieren demnach Zugänge im Bildungswesen, wie die Aufnahme und Fortsetzung eines Studiums. Darüber hinaus bestimmen Hochschulabschlussnoten auch berufliche Laufbahnen und Berufspositionen. Dies wird bei der Besetzung von Richterämtern besonders deutlich. Das Amt der Richterin und des Richters bleibt, etwa in Baden-Württemberg, Absolventinnen und Absolventen vorbehalten, deren herausragende Leistung im Ersten und im Zweiten Staatsexamen durch eine Abschlussnote von jeweils mindestens 8,0 Punkten (befriedigend) sichtbar wird[2] (https://www.mit-recht-in-die-zukunft.de/richterstaatsanwalt/die_bewerbung/, Stand 09.10.2019). Vor dem Hintergrund, dass Noten keine objektiven Bewertungen darstellen (Kalthoff 1996; Lüders 2001) und es systematische Notenunterschiede gibt, ist diese Form der Rekrutierung zu kritisieren.

Der Wissenschaftsrat dokumentiert für das deutsche Hochschulsystem, dass die Durchschnittsnoten in den einzelnen Disziplinen weit auseinanderfallen: Während in den Rechtswissenschaften Durchschnittsnoten von 3,3 erzielt werden, erreichen Absolventinnen und Absolventen in der Biologie durchschnittliche Noten von 1,3 (Wissenschaftsrat 2003). Darüber hinaus existieren studiengangspezifische Unterschiede innerhalb einer Disziplin: In Prüfungen des Staatsexamens für das Lehramt werden häufig schlechtere Noten als in Diplom- und Magisterstudiengängen vergeben (Müller-Benedict u. Tsarouha 2011; Wissenschaftsrat 2012). Universitätsspezifische Unterschiede der Durchschnittsnoten im Prüfungsjahr 2010 stellt der Wissenschaftsrat (2012) fest: Die durchschnittlichen Abschlussnoten für den Studiengang Lehramt an Gymnasien für das Unterrichtsfach Deutsch liegen im Vergleich von 43 untersuchten Hochschulen zwischen 1,3 und 2,6. In Abschlussprüfungen des Magisterstudiengangs Germanistik werden an 47 Hochschulen Notendurchschnitte von 1,6 bis 2,2 erzielt. Gaens und Müller-Benedict (2017) weisen langfristige, stabile durchschnittliche Notenunterschiede zwischen Studiengängen für den Zeitraum 1961-2010 nach: In Biologie und Psychologie werden die besten Noten erreicht und in VWL, BWL und Jura erzielen Absolventinnen und Absolventen die schlechtesten Noten.

Ob es sich dabei um systematische Notenunterschiede handeln könnte, haben Müller-Benedict und Tsarouha analysiert (2011). Ausschließen konnten sie hierbei, dass besonders schlaue Personen z. B. Psychologie studieren, während weniger intelligente Personen z. B. eher Jura wählen würden (Müller-Benedict u. Tsarouha 2011). Die Autoren zeigen, dass die Durchschnittsnoten im Abitur der Absolventinnen und Absolventen verschiedener Abschlussexamensarten nur geringfügige Unterschiede bezüglich der Startbedingungen in das Studium aufzeigen und damit keine Erklärungskraft für die erzielten durchschnittlichen Notenunterschiede in den Abschlussprüfungen bieten. Auch die Annahme, dass Universitäten z. B. aufgrund einer hohen Reputation nur eine bestimmte Klientel an Studierenden aufnehmen und dadurch besse-

[2] Die Einstellungsvoraussetzungen können in verschiedenen Bundesländern variieren. Beispielsweise werden in Schleswig-Holstein zwei mit Prädikat (mindestens 9 Punkte) abgeschlossene Staatsexamina vorausgesetzt (https://www.schleswig-holstein.de/DE/Fachinhalte/K/karriere/Juristen/juristen_Richter_Staatsanwaelte.html, Stand 09.10.2019).

re durchschnittliche Abschlussnoten erzielen würden, kann für einige untersuchte und durch einen Numerus Clausus (NC) zulassungsbeschränkte Studiengänge ausgeschlossen werden (Müller-Benedict u. Tsarouha 2011).

Es stellt sich somit die Frage, worauf systematische Notenunterschiede zurückzuführen sind. Hochschulprüfungen unterliegen weitreichenden Reglementierungen und Rahmenbedingungen, die dazu beitragen sollen, Prüfungen zu standardisieren. Dadurch sollen Prüfungen objektiviert und vergleichbar werden. Dennoch eröffnen sich im Prüfungswesen Handlungsspielräume für die Prüfenden. Diese werden anhand der Weisungsfreiheit von Prüfenden deutlich. Die Weisungsfreiheit umfasst Aufgabenstellung, Prüfungsmethode und -technik, Fehlergewichtung, Einstufung des Schwierigkeitsgrades, Bewertung der Darstellungsweise und Benotung (Hartmer 2012). Der Aspekt der Benotung beinhaltet u. a. das Verständnis der Notenskala bzw. des Notensystems und von einzelnen Notenniveaus. Die Weisungsfreiheit der Prüfenden verdeutlicht, dass Prüfungen abhängig von den einzelnen Prüfenden sein können (Brückel et al. 2000). Gleichwohl ist es durch das Prüfungsrecht legitimiert, dass Prüfungen nicht gleich, sondern vergleichbar sein müssen (basierend auf Art. 3 Abs. 1 GG).
Innerhalb der Handlungsspielräume, sogenannter Dispositionsspielräume (Lüders 2001), wirken unterschiedliche Einflussgrößen. Diese können auf unterschiedliche Wirkungskontexte zurückgeführt werden. Die vielfältigen Einflussgrößen lassen sich strukturiert aufbereiten und als Typologie der Einflussgrößen auf die Notengebung darstellen (Tsarouha 2017, 2019). Vor diesem Hintergrund wird die Forschungsfrage des Beitrags diskutiert: *Welche Steuerungsmöglichkeiten gibt es für Hochschulen, um Abschlussnoten vergleichbar(er) zu machen?* Nachfolgend wird das Vorhaben exemplarisch für den Typ der kommissionsspezifischen Einflussgrößen an ausgewählten Beispielen dargelegt.

2 Methoden und Daten

Im Rahmen des DFG-geförderten Projekts *Notengebung an Hochschulen in Deutschland*[3] wurden Noten hinsichtlich systematischer Notenunterschiede erforscht (Müller-Benedict u. Grözinger 2017). Das Projekt verfolgte einen Mixed-Methods-Ansatz. Für die quantitativen Analysen wurden Daten in acht Hochschulen sowie in Landesarchiven (FU Berlin, TU Braunschweig, Göttingen, Heidelberg, Hildesheim (Archiv Lehramt), Münster, KIT Karlsruhe, Saarbrücken (nur Germanistik), Tübingen) erhoben. Zusätzlich wurden Prüfungsstatistiken aus dem Statistischen Bundesamt für den Zeitraum von 1995 bis 2013 für die Analysen herangezogen. Anhand dieser Daten wurden Längsschnittanalysen über mehrere Jahrzehnte zur Entwicklung der Noten (Gaens u. Müller-Benedict 2017) und Querschnittsanalysen z. B. zu Einflüssen von individuellen und studentischen Merkmalen oder universitären und regionalen Dimensionen durchgeführt (Grözinger 2017). Der vorliegende Beitrag stellt Ergebnisse der weiteren Säule im DFG-Projekt vor, durch welche die Studie um qualitative Analysen zu Prüfungspraktiken in Hochschulabschlussprüfungen ergänzt wurde. Die qualitative Erhebung umfasste vier problemzentrierte Einzelinterviews, zwei Experteninterviews und neun Gruppendiskussionen. Die Einzelinterviews wurden inhaltsanalytisch (angelehnt an Mayring 2010) analysiert und die Experteninterviews wurden thematisch zusammengefasst. Die Gruppendiskussionen wurden mittels der dokumentarischen Methode (Bohnsack 2014) analysiert. Die Einzelinterviews unterstützten u. a. die Entscheidung über die Auswahl der zu berücksichtigenden Disziplinen und Studiengänge in den Gruppendiskussionen. Die Experteninterviews dienten dazu, die bundeslandspezifischen Prüfungsbedingungen und deren Entwicklung offenzulegen. Die Gruppendiskussionen und Einzelinterviews wurden mit Prüfenden mit langjährigen Prüfungserfahrungen geführt.

[3] DFG-gefördertes Projekt (MU1625/7), *Notengebung an Hochschulen in Deutschland*, Leitung: Prof. Müller-Benedict und Prof. Grözinger (EUF), Laufzeit 2012-2015

Fünf der Gruppendiskussionen setzten sich aus Professorinnen und Professoren zusammen und vier Gruppendiskussionen erfolgten mit ministerial berufenen Prüfungsvorsitzenden des Ersten Staatsexamens für das Lehramt an Gymnasien. Dabei waren die ministerial berufenen Prüfungsvorsitzenden hochschulexterne Personen, z. B. aus dem Schulkontext. Prüfungserfahrungen sollten vor allem in den Disziplinen Mathematik und Germanistik vorliegen und in den Studiengängen Diplom, Magister und Lehramt an Gymnasien. Ferner sollten Prüfungserfahrungen in den Bundesländern Baden-Württemberg oder Niedersachsen gegeben sein. Diese Auswahl stützte sich auf die Annahme, dass strukturelle Merkmale wie Disziplin, Studiengang sowie standortspezifische Gegebenheiten Einfluss auf die Prüfungspraktiken nehmen könnten. An den Gruppendiskussionen haben zwischen zwei und bis zu fünf Prüferinnen und Prüfer teilgenommen. Die Zusammensetzung der Gruppendiskussionen wurde möglichst homogen gestaltet. Professorinnen und Professoren diskutierten getrennt von den ministerial berufenen Prüfungsvorsitzenden und die Gruppendiskussionen wurden weitgehend so zusammengesetzt, dass Prüfende mit Prüfungserfahrungen in derselben Disziplin, desselben Studiengangs und teilweise auch an derselben Universität vertreten waren. Zwei Diskussionen mit ministerial berufenen Prüfungsvorsitzenden wurden disziplinübergreifend geführt. Das Ergebnis der qualitativen Analysen ist eine Typologie der Einflussgrößen auf die Notengebung in Abschlussprüfungen an deutschen Hochschulen (Tsarouha 2017, 2019). Die Typologie bezieht sich hauptsächlich auf Einflüsse, die in der mündlichen Prüfung wirken.

3 Typ kommissionsspezifische Einflussgrößen

Die kommissionsspezifischen Einflussgrößen lassen sich von fünf weiteren Typen unterscheiden: disziplin-, fach-[4], bundesland-, studiengang- und abschlussspezifische Einflussgrößen (Tsarouha 2019).
Die kommissionsspezifischen Einflussgrößen entfalten sich abhängig von der personellen Zusammensetzung der Prüfungskommission in mündlichen Prüfungen (zu den folgenden Ergebnissen siehe ausführlicher Tsarouha 2019):

- eine Prüferin/ein Prüfer und eine beisitzende Person
- mehrere Hochschulprofessorinnen und -professoren in kollegialen Prüfungen
- mehrere Hochschulprofessorinnen und -professoren in kollegialen Prüfungen mit einer/einem zusätzlichen ministerial berufenen Prüfungsvorsitzenden

Prüfungen mit nur einer prüfenden und einer beisitzenden Person sind an einzelnen Standorten in den Magisterprüfungen und standortübergreifend für das Mathematik Diplom üblich. Magisterprüfungen in der Germanistik werden häufig durch Prüfungskommissionen mit mehreren Professorinnen und Professoren durchgeführt. Ministerial berufene Prüfungsvorsitzende werden z. B. in Ersten Staatsexamina für das Lehramt an Gymnasien eingesetzt. Dabei existieren bundeslandspezifische Vorgaben bezüglich des Personenkreises der ministerial berufenen Prüfungsvorsitzenden (Tsarouha 2017, 2019).

In Abbildung 1 wird die Struktur innerhalb der Typen der Einflussgrößen aufgezeigt. Ein Typ lässt sich weiter differenzieren in verschiedene sinngenetische Faktoren, welche thematisch zugehörige Einflüsse bündeln. Die aus den Gesprächen identifizierten Einflüsse können spezifisch für eine bestimmte Prüfungszusammensetzung sein oder derselbe Einfluss kann sich unterschiedlich in verschiedenen Kommissionszusammensetzungen der mündlichen Prüfung auswirken. In den nachfolgenden ausgewählten Beispielen des kommissionsspezifischen Typs wirkt der Einfluss 1 sowohl in Kommissionen mit mehreren Prüfenden und ministerial berufenen Prüfungsvorsitzenden als auch in Prüfungen mit einer prüfenden und einer beisitzenden

[4] Zur begrifflichen Definition von disziplin- und fachspezifischen Unterschieden siehe Tsarouha (2017, 2019).

Person. Somit gibt es spezifische Typiken für die jeweilige Prüfungszusammensetzung (Tsarouha 2019). Im Vergleich dazu wirkt Einfluss 2 ausschließlich in Prüfungen mit mehreren Prüfenden und ministerial berufenen Prüfungsvorsitzenden.

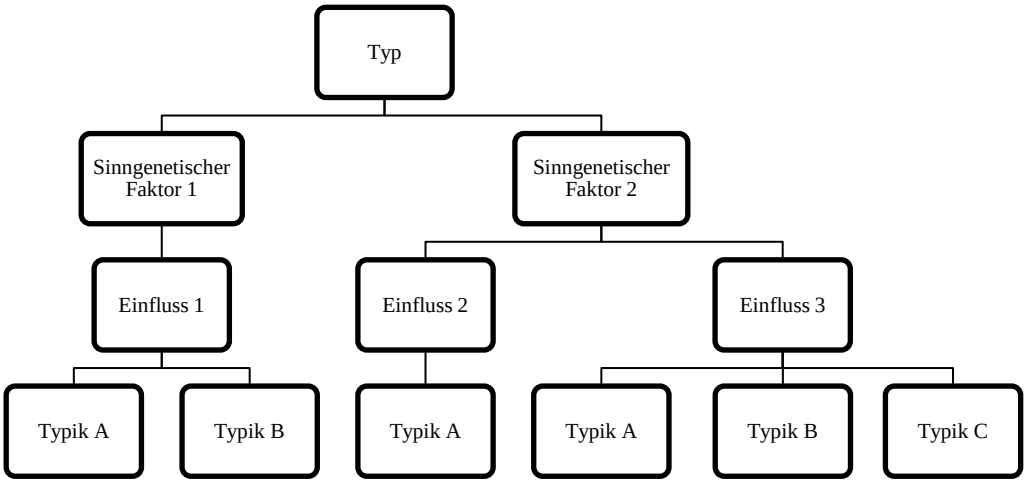

Abbildung 1: Typ – sinngenetische Faktoren – Einflüsse – Typiken
Erläuterungen zur Darstellung: Die unterste Ebene beinhaltet Typiken, die sich spezifisch in den jeweiligen Kommissionszusammensetzungen zeigen: Typik A = Mehrere Prüfende als kollegiale Prüfung mit einer/einem ministerial berufenen Prüfungsvorsitzenden; Typik B = Eine prüfende und eine beisitzende Person, Typik C = Mehrere Prüfende als kollegiale Prüfung (vgl. Tsarouha 2019, S. 207)

In der Abbildung 2 wird die Struktur aus Abbildung 1 mit Beispielen angereichert. Diese ausgewählten Beispiele werden im Folgenden schrittweise erläutert.

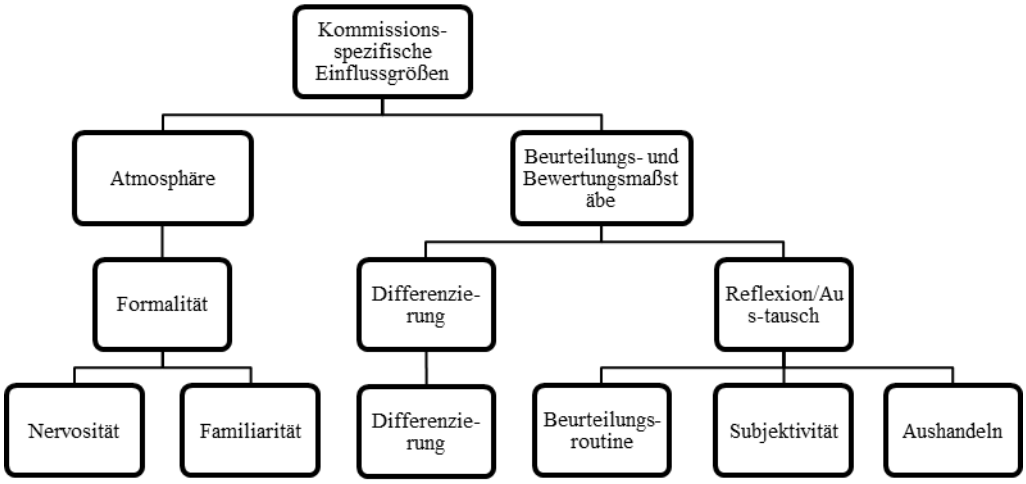

Abbildung 2: Auswahl kommissionspezifischer Einflüsse
(Darstellung angelehnt an Tsarouha 2019, S. 207)

Der Typ kommissionsspezifischer Einflussgrößen umfasst insgesamt drei sinngenetische Faktoren und 13 Einflussgrößen und Typiken (Tsarouha 2019). Im Beitrag wird der Typ kommissionsspezifischer Einflussgrößen exemplarisch an den sinngenetischen Faktoren *Atmosphäre* und *Beurteilungs- und Bewertungsmaßstäbe* beschrieben.

Der sinngenetische Faktor Atmosphäre wird anhand des Einflusses der *Formalität* einer mündlichen Prüfung dargestellt. Eine erhöhte Formalität in Prüfungen des Ersten Staatsexamens werde laut den Befragten durch die Anwesenheit ministerial berufener Prüfungsvorsitzender in den mündlichen Prüfungen sichtbar. Ein Einfluss zeige sich in Prüfungen mit mehreren Prüfenden und ministerial berufenen Prüfungsvorsitzenden durch die *Nervosität* der Studierenden. In Prüfungen mit nur einer prüfenden und einer beisitzenden Person sei die Formalität geringer und die Atmosphäre werde als *familiär* empfunden. Für die untersuchten Studiengänge bedeute dies, dass in mündlichen Prüfungen des Ersten Staatsexamens für das Lehramt an Gymnasien eine hohe Formalität gegeben sei, wodurch Prüflinge eine vergleichsweise höhere Nervosität verspüren könnten. Die Befragten gaben an, dass diese Nervosität in den untersuchten Unterrichtsfächern Germanistik und Mathematik unterschiedlich wirke. Während Prüfende der Mathematik äußerten, dass die Prüflinge des Ersten Staatsexamens aufgrund der hohen Nervosität unter ihren Möglichkeiten bleiben würden, wurden seitens der Prüfenden der Germanistik weitere Konsequenzen genannt. Manche Professorinnen und Professoren der Germanistik waren der Meinung, dass sich die Staatsexamenskandidatinnen und -kandidaten durch die erhöhte Formalität im Gegensatz zu Prüflingen des Magisters besser auf die Prüfung vorbereiten und dadurch bessere Leistungen erzielen würden:

„Person 2: Es gibt nämlich einerseits die angespannten es gibt aber auch und das ist mir auch relativ häufig passiert
Person 1: [Die zu entspannten mhmm ja ja]
Person 2: [Es gibt die genau diese […]- die gehen da] rein und plaudern dann [mit einem]
Person 1: [Ja ja]

> *Person 2: Und so und man merkt die sind nicht konzentriert[,] die sind nicht fokussiert und die sind sich des Ernstes dieser Prüfungssituat- also die nehmen es nicht als Prüfungssituation wahr[,] sondern die sagen sich[: ,]unterhalten wir uns ein bisschen über […] das Seminar[‘]“*[5] *(BW_Uni_D_RB2, S. 90f.)*

Prüfende der Germanistik berichteten auch, dass sie eine starke Nervosität der Prüflinge bei der Leistungsbeurteilung berücksichtigen würden und dadurch schlechtere Bewertungen erfolgen könnten.
Bezüglich der untersuchten Mathematik Diplomprüfungen teilten befragte Prüfende mit, dass Prüfungen mit einer prüfenden und einer beisitzenden Person eine geringere Formalität besitzen und die mündlichen Prüfungen als familiär empfunden würden. Durch diese Atmosphäre würden Prüflinge ihr Leistungspotential eher offenlegen können. Zusätzlich zu der familiären Atmosphäre aufgrund der Abwesenheit von fremden Prüfenden sei in den Diplomprüfungen ein enges Betreuungsverhältnis zwischen Prüferin oder Prüfer und Prüfling gegeben. Dadurch kann ggf. eine objektive Bewertung erschwert sein, so dass aufgrund von Empathie oder Sympathie positiver als angemessen bewertet wird.

Der zweite angeführte sinngenetische Faktor wird als Beurteilungs- und Bewertungsmaßstäbe bezeichnet. Dieser wird anhand von zwei unterschiedlichen Einflüssen veranschaulicht: *Differenzierung* und *Reflexion/Austausch*. Der Einfluss Differenzierung wurde ausschließlich für die Prüfungen mit mehreren Prüfenden und ministerial berufenen Prüfungsvorsitzenden festgestellt. Der sinngenetische Faktor Reflexion/Austausch kann in allen drei Kommissionszusammensetzungen unterschiedlich wirken. Für die untersuchten Studiengänge bedeutet dies, dass in Prüfungen des Ersten Staatsexamens ggf. eine stärkere Differenzierung der Leistungsbeurteilung und -bewertung erfolgen kann. Dies wurde seitens der Befragten auf die Anwesenheit der ministerial berufenen Prüfungsvorsitzenden zurückgeführt, die aufgrund ihrer Benotungspraxis aus der Schule eine stärkere Differenzierung bei der Beurteilung und Bewertung der mündlichen Prüfungsleistung bewirken könnten. Die gegebenen Leistungs- und Notenniveaubeschreibungen in den Prüfungsordnungen scheinen unzureichend und vage zu sein. Ministerial berufene Prüfungsvorsitzende gaben im Vergleich zu Professorinnen und Professoren an, zusätzlich auf Handreichungen aus ihrem Berufsalltag zurückgreifen zu können. In einer Gruppendiskussion mit ministerial berufenen Prüfungsvorsitzenden aus verschiedenen Disziplinen wird geäußert:

> *„Und das finde ich ganz schön schwierig[,] […]wenn man bei der Note Fünf sagt[: ,]eine Leistung die den Anforderungen nicht entspricht[,] jedoch erkennen lässt[,] das[s] die notwendigen Grundkenntnisse vorhanden sind[.‘] [I]ch habe dann immer die Notenbildungsverordnung für die Noten an der Schule noch so nebenher in der Tasche“ (BW_SE_G, S. 65)*

Manche Professorinnen und Professoren teilten mit, dass sie aufgrund unzureichender Differenzierungsmöglichkeiten bei der Beurteilung von Prüfungsleistungen diese milder bewerten würden. Es sei nicht einfach, Benotungen auf Zehntelnoten genau zu begründen. Eine mögliche Konsequenz daraus könnte auch sein, dass sehr gute und sehr schlechte Leistungen unterschieden werden, aber die mittleren Noten aufgrund der Unsicherheit weniger häufig vergeben werden. Folglich könnten Prüfungskommissionen ohne ministerial berufene Prüfungsvorsitzende zu positiv verzerrten Prüfungsergebnissen führen. In Prüfungen mit ministerial berufenen Prüfungsvorsitzenden könnte diesen Verzerrungen entgegengewirkt werden, so dass das Notenspektrum möglicherweise besser ausgeschöpft wird und die Noten ggf. leistungskonform schlechter ausfallen.

Der Einfluss Reflexion/Austausch wirkt je nach Kommissionszusammensetzung über eine *Beurteilungsroutine*, einer erhöhten *Subjektivität* oder durch ein *Aushandeln* zwischen den Prüfenden. In Prüfungen des Ersten Staatsexamens wurde seitens der Befragten die Beurteilungsroutine der ministerial berufenen Prüfungsvorsitzenden hervorgehoben. Diese führe zu einer differenzierten und angemessenen Beurteilung und Bewertung von Prüfungsleistungen

[5] Alle Zitate wurden zugunsten der Lesefreundlichkeit überarbeitet.

durch die ministerial berufenen Prüfungsvorsitzenden. Ferner würde diese Prüfungsroutine den Professorinnen und Professoren fehlen, wodurch Prüfungsleistungen weniger differenziert und milder benotet würden. In diesem Fall könnten die Prüfungsnoten in Prüfungen ohne ministerial berufene Prüfungsvorsitzende positiv verzerrt sein.

In Prüfungen mit nur einer prüfenden und einer beisitzenden Person, wie es z. B. für einzelne Standorte in den Magisterprüfungen oder standortübergreifend für das Mathematik Diplom üblich ist, scheint die Reflexion bzw. der Austausch innerhalb der Kommission tendenziell eingeschränkt. Das bedeutet, dass die Prüferin oder der Prüfer darüber entscheidet, inwieweit die Einschätzungen der beisitzenden Person zur Rate gezogen werden und in die Beurteilung und Bewertung einfließen. Entsprechend könnten die Prüfungsergebnisse in dieser Kommissionskonstellation stärker von der Einschätzung einer einzelnen Person abhängig und somit stärker subjektiv gefärbt sein. Zusätzlich kann die alleinige Verantwortung für die Prüfungsbeurteilung und -bewertung dazu führen, dass eine positiv verzerrte Note resultiert. In einer Gruppendiskussion mit Professorinnen und Professoren für die Disziplin Germanistik wird geäußert:

> *„Und ich glaube es liegt einfach auch daran[,] wenn man zu mehreren ist[,] dann gibt man leichter eine schlechte Note[,] weil man dann das Gefühl hat das ist jetzt keine persönliche […] Schwierigkeit oder man hat einen schlechten Tag oder man hat was überhört oder so […]“(BW_Uni_D_RB1, S. 9)*

Gemäß dieser Aussage würde in kollegialen Prüfungen durch eine geteilte Verantwortung tendenziell leistungskonformer bewertet werden. Auch konnte in kollegialen Prüfungen und Prüfungen mit ministerial berufenen Prüfungsvorsitzenden in der Germanistik ein stärkerer Austausch zwischen den Prüfenden bei der Notenfindung identifiziert werden, der sich z. B. in einem Aushandeln der Note widerspiegelte. Beim Aushandeln der Note müssen Prüfende ihre Vorschläge ggf. gegenüber den Kommissionsmitgliedern begründen. Dies könnte dazu führen, dass stärker über die jeweiligen Beurteilungs- und Bewertungsmaßstäbe reflektiert wird. Als Folge dessen könnten Prüfende ggf. nicht-leistungskonforme Einflüsse identifizieren und die erbrachte Prüfungsleistung stärker differenzieren oder weniger milde beurteilen und bewerten. Für Prüfungen im Lehramt an Gymnasien für das Unterrichtsfach Mathematik wurden in verschiedenen Gruppendiskussionen unterschiedliche Vorgehensweisen bei der Notengebung beschrieben (Tsarouha 2019). Beispielsweise erklärten manche ministerial berufene Prüfungsvorsitzende aus Baden-Württemberg, dass jede Prüferin und jeder Prüfer die jeweils eigenen Prüfungsteile beurteilen und bewerten würde. Die Gesamtnote würde dann über die einzelnen Noten gemittelt. Es ist möglich, dass aus der beschriebenen geringeren Auseinandersetzung mit weiteren Prüfenden eine geringere Reflektion über die eigenen Prüfungspraktiken und die Prüfungspraktiken der anderen resultiert.

4 Diskussion der Steuerungsmöglichkeiten

Der Beitrag zeigt, dass es vielfältige Einflüsse auf die Notengebung in mündlichen Abschlussprüfungen an deutschen Hochschulen gibt. Die Einflussgrößen des vorgestellten kommissionsspezifischen Typs sind strukturell bedingte Einflussgrößen, die sich durch die Vorgaben der Prüfungsordnungen ergeben, denn in den Prüfungsordnungen wird festgelegt, wie sich die Prüfungskommission zusammensetzt. Dies bedeutet auch, dass die Universitäten bzw. die Ministerien über Steuerungsmöglichkeiten verfügen, um einer potentiellen Notenverzerrung entgegenzuwirken und Abschlussnoten vergleichbarer zu machen.

4.1 Ministerial berufene Prüfungsvorsitzende

Bei der Prüfungsatmosphäre wurde die unterschiedlich empfundene Formalität hervorgehoben. Diese wurde insbesondere beim Vergleich der Ersten Staatsexamen und den Diplomprüfungen deutlich. Manche Prüfende der Germanistik teilten die Erfahrung, dass dies zu einer verbesserten Vorbereitung und damit einer verbesserten Prüfungsleistung der Staatsexamenskandidatinnen und -kandidaten führen könne, wohingegen Prüfende der Mathematik eine erhöhte Nervosität betonten, wodurch schlechtere Leistungen in den Ersten Staatsexamen resultieren könnten. Es bleibt unklar, ob die vergleichsweise hohe Nervosität der Lehramtsstudierenden des Unterrichtsfachs Mathematik ausschließlich auf die Anwesenheit der fremden, ministerialberufenen Prüfungsvorsitzenden zurückführbar ist. Die hohe Nervosität könnte sich auch dadurch ergeben, dass das Fach nach Aussagen einiger Befragten nicht ausschließlich aus persönlicher Neigung, sondern aus (berufs-) strategischen Gründen von manchen Studierenden gewählt würde. Da in den Gruppendiskussionen angegeben wurde, dass ministerial berufene Prüfungsvorsitzende tendenziell über eine geringe Einflussnahme in der Prüfung verfügen, sollten die möglichen Vor- und Nachteile ihres Einsatzes kritisch reflektiert werden (Tsarouha 2019).

4.2 Kollegiale Prüfungen

Die genannte Familiarität in Prüfungen mit einer prüfenden und einer beisitzenden Person könnte zu einer positiveren Beurteilung der Prüfungsleistung führen. Die Familiarität in dieser Kommissionszusammensetzung und die beschriebenen Auswirkungen auf das Prüfungsergebnis scheinen nicht exklusiv für den Studiengang Mathematik Diplom zu sein. Der mit der familiären Atmosphäre einhergehende Aspekt des engen Betreuungsverhältnisses aufgrund der Abschlussarbeit wurde für den Diplomstudiengang Mathematik hervorgehoben. Jedoch wird auch in den Gesprächen mit Prüfenden der Germanistik deutlich, dass sich das Verhältnis zu Kandidatinnen und Kandidaten, die ihre Abschlussarbeit bei den Prüfenden der mündlichen Prüfung erstellt haben, von anderen Kandidatinnen und Kandidaten unterscheide (Tsarouha 2019). Demnach scheint grundsätzlich in Frage gestellt, ob Betreuende der Abschlussarbeit objektive Prüferinnen und Prüfer in den mündlichen Prüfungen sein können. Durch weitere Prüfende im Sinne kollegialer Prüfungen könnten positive Verzerrungen der Leistungsbeurteilung und -bewertung verringert oder vermieden werden.

Einige Ergebnisse des dritten Einflusses *Austausch/Reflexion* könnten disziplin-, fach- und studiengangübergreifend wirksam sein. Prüfungen mit einer prüfenden und einer beisitzenden Person scheinen gegenüber kollegialen Prüfungen (mit und ohne ministerial berufenen Prüfungsvorsitzenden) einerseits stärker subjektiv geprägt zu sein und anderseits eine positivere Beurteilung und Bewertung zu begünstigen. In Prüfungen mit mehreren Prüfenden werden Einflüsse auf die Leistungsbewertung und -beurteilung, die sich z. B. aus der Weisungsfreiheit der Prüfenden ergeben, wie etwa der Fehlergewichtung oder dem Verständnis der Notenskala, ausgeglichen. Zusätzlich kann sich eine geteilte Verantwortung zwischen mehreren Prüfenden dahingehend auswirken, dass leistungskonformer bewertet wird und ggf. schlechtere Noten vergeben werden. Außerdem bieten kollegiale Prüfungen die Möglichkeit, dass die einzelnen Prüfenden ihre eigenen Prüfungspraktiken im Vergleich mit Kolleginnen und Kollegen reflektieren und möglicherweise korrigieren können. Eine Reflektion über die eigenen Prüfungspraktiken könnte dazu führen, dass Prüfungen bei ein und demselben Prüfenden objektiver und vergleichbarer werden. Eine weitere Folge kann eine Auseinandersetzung mit dem Verständnis der Notenskala und eine daraus resultierende differenziertere Benotung sein. Eine Neugierde auf die Prüfungspraktiken von Kolleginnen und Kollegen wurde in der Germanistik geäußert. Hinweise aus den Gesprächen mit Vertretenden der Mathematik lassen offen, ob kollegiale Prüfungen automatisch eine Auseinandersetzung mit den Prüfungspraktiken von anderen Prüfenden gewährleisten. Eine gemeinsame Notenfindung könnte von weiteren Faktoren abhängen, beispielsweise von der Bereitschaft des Einzelnen oder einer vergleichbaren Fachexpertise

aller Prüfenden zu allen Prüfungsteilen. Dennoch sprechen die Ergebnisse dafür, dass mündliche Abschlussprüfungen als kollegiale Prüfungen durchgeführt werden sollten.

4.3 Schulungsmaßnahmen für Prüfende

Manche Prüfende der Germanistik berichteten, dass sie die Nervosität der Prüflinge bei der Bewertung negativ berücksichtigen würden. Das bedeutet, dass die Leistung des Prüflings durch einzelne Merkmale wie beispielsweise einem nervösen Auftreten negativ beeinflusst werde. Ob eine gegebene Nervosität als Kriterium bei der Leistungsbeurteilung herangezogen werden sollte, ist kritisch zu reflektieren. Inwiefern bewusste und unbewusste Einflüsse in Prüfungen möglich sind, könnte z. B. im Rahmen von institutionalisierten Schulungsmaßnahmen an Prüfende vermittelt werden.

4.4 Differenzierte Leistungsniveau- und Notenniveaubeschreibungen

Der Aspekt einer differenzierten Leistungsbeurteilung und -bewertung, der in den Ersten Staatsexamensprüfungen über die ministerial berufenen Prüfungsvorsitzenden in die Prüfungen eingebracht werde, könnte durch differenzierte Leistungsniveau- und Notenniveaubeschreibungen auf universitärer Ebene für Professorinnen und Professoren in anderen Prüfungskommissionen wirksam werden. Aus einer differenzierteren Beurteilung kann eine stärkere Ausnutzung der Notenskalen und somit eine Spreizung der Abschlussnoten resultieren. Neben leistungskonformen Prüfungsergebnissen für einzelne Absolventinnen und Absolventen könnte eine weitere Konsequenz darin bestehen, dass unterschiedliche Leistungsniveaus der Absolventinnen und Absolventen stärker sichtbar werden. Dabei scheint eine disziplinspezifische Verständigung über verschiedene Universitäten hinweg, unrealistisch, da innerhalb einer Disziplin z. B. verschiedene Paradigmen gegeben sein können, die jeweils spezifische Beurteilungskriterien erfordern. Vielfältige Paradigmen können auch an einer Fakultät einer Universität gegeben sein. Dennoch scheint eine Verständigung zumindest auf universitärer Ebene unter der Berücksichtigung disziplin- und fachspezifischer Unterschiede und Gemeinsamkeiten zu einer höheren Vergleichbarkeit von Abschlussnoten beitragen zu können.

5 Fazit

Die genannten Steuerungsmöglichkeiten zielen auf Veränderungen auf der Mikroebene in mündlichen Prüfungen ab. Es sollen nicht-leistungsbezogene Einflussgrößen ausgeschlossen oder zumindest minimiert werden. Angestrebt werden Bedingungen, die eine Verzerrung bei der Leistungsbewertung und -beurteilung verhindern oder zumindest einer solchen entgegenwirken. Es ist davon auszugehen, dass Prüfungen durch Steuerungsmöglichkeiten objektiver und dadurch vergleichbarer werden können. Gleichzeitig sind (mündliche) Prüfungen immer soziale Interaktionen, die durch subjektive Wahrnehmungen und Interpretationen geprägt sind, so dass eine absolute Vergleichbarkeit nicht hergestellt werden kann.

Alternative Leistungsindikatoren wie z. B. Praktika, Auslandserfahrungen, Gutachten etc. scheinen eher zusätzliche Indikatoren und kein adäquater Ersatz für die Verwendung von Noten zu sein. Noten sind weiterhin Leistungsindikator und Ausschlusskriterium. Auch wenn quantitative Studien signifikante durchschnittliche Notenunterschiede „nur“ im Zehntelnotenbereich zeigen und die Notenunterschiede vermeintlich gering sind, ist ein damit verbundener, möglicher Ausschluss zu weiteren Qualifikationsmöglichkeiten oder beruflichen Positionen zu

kritisieren. Mit dem Wissen, dass Noten nicht auf Zehntelnoten genau Leistungsniveaus widerspiegeln, jedoch auf Zehntelnoten genau z. B. geregelt ist, ob ein Studium fortgeführt werden darf, sind Bemühungen um weitere Lösungen gerechtfertigt und notwendig. Vor diesem Hintergrund sollten neben den genannten (und weiteren) Steuerungsmöglichkeiten zusätzliche Maßnahmen ergriffen werden, welche die Interpretation erzielter Noten unterstützen und Alternativen zur Verwendung der Noten als Leistungsindikator bieten. Eine Forderung besteht darin, Prüfungen und erzielte Prüfungsleistungen transparenter zu gestalten. Das kann einerseits durch eine Offenlegung der Anforderungen und Beurteilungskriterien der jeweiligen Prüfungen gegenüber den Studierenden erfolgen. Anderseits können Noten auf Zeugnissen ergänzt werden, indem die Verortung des Einzelnen innerhalb der Prüfungskohorte erfolgt (Müller-Benedict u. Grözinger 2017). Hierbei besteht jedoch die Gefahr, dass Studierende stigmatisiert werden, die z. B. nicht zu den besten 50 Prozent gehören. Dadurch, dass der Referenzrahmen das Leistungsniveau einer Kohorte ist, kann dieselbe Note einmal zu den besten 25 Prozent einer Kohorte gehören und im Folgejahr in einem schlechteren Viertel verortet sein. Zusätzlich löst die Einordung der einzelnen Abschlussnote in das Notenniveau einer Prüfungskohorte nicht das nachfolgend geschilderte Problem bei der Rekrutierung: Wie ist zu entscheiden, wenn Absolvent A an der Universität A eine Abschlussnote von 1,7 hat und sich im Viertel der besten Abschlüsse befindet, während Student B an der Universität B mit einer Abschlussnote von 1,5 zum zweitbesten Viertel der Absolventinnen und Absolventen gehört? Da Notenniveaus der Hochschulen keine exakten Leistungsniveaus widerspiegeln, können die Noten verschiedener Universitäten nicht 1:1 verglichen werden (Tsarouha 2019). Aufgrund der genannten Einschränkungen der Aussagekraft von Noten ist zu empfehlen, dass Zugänge zumindest anteilig auch über Losverfahren geregelt werden.

Literatur

Bohnsack, R. (2014). Rekonstruktive Sozialforschung Einführung in qualitative Methoden. 9. überarbeitete und erweiterte Auflage. Opladen: Barbara Budrich.

Brückel, F., Holtgrewe, H., Konopka, T., Landmann, U., Macke, G., Nennstiel, C., Raether, W., Rapp, S., Schumacher, S., Simen, J. & Weingart, V. (2000). Mündliche Hochschulprüfungen Vorbereitung – Durchführung – Bewerten – Beraten. In Arbeitsgruppe Hochschuldidaktische Weiterbildung an der Alberts-Ludwigs-Universität Freiburg i. Br (Hrsg.), Besser Lehren. Praxisorientierte Anregungen und Hilfen für Lehrende in Hochschule und Weiterbildung, Heft 10. Weinheim: Deutscher Studien-Verlag.

Gaens, T. & Müller-Benedict, V. (2017). Die langfristige Entwicklung des Notenniveaus und ihre Erklärung. In Müller-Benedict, V & Grözinger, G. (Hrsg.), Noten an Deutschlands Hochschulen (S. 17-78). Wiesbaden: VS Verlag für Sozialwissenschaften.

Grözinger, G. (2017). Einflüsse auf die Notengebung: eine Analyse ausgewählter Fächer auf Basis der Prüfungsstatistik. In Müller-Benedict, V. & Grözinger, G. (Hrsg.), Noten an Deutschlands Hochschulen (S. 79-116). Wiesbaden: VS Verlag für Sozialwissenschaften.

Grundgesetz für die Bundesrepublik Deutschland in der im Bundesgesetzblatt Teil III, Gliederungsnummer 100-1, veröffentlichten bereinigten Fassung, das zuletzt durch Artikel 1 des Gesetzes vom 28. März 2019 (BGBl. I S. 404) geändert worden ist.

Hartmer, M. (2012). Der Prüfer. Seminar Prüfungsrecht an Hochschulen. Deutscher Hochschulverband. Arbeitsmaterial 24.01.2012.

Hitpass, J. & Trosien, J. (1987). Leistungsbeurteilung in Hochschulabschlussprüfungen innerhalb von drei Jahrzehnten. Bad Honnef: K. H. Bock.

https://www.mit-recht-in-die-zukunft.de/richterstaatsanwalt/die_bewerbung/, Stand 09.10.2019.

https://www.schleswig-holstein.de/DE/Fachinhalte/K/karriere/Juristen/juristen_Richter_Staatsanwaelte.html, Stand 09.10.2019.

Hu, S. & Kuh, G. D. (2003). Maximizing What Students Get Out of College: Testing a Learning Product. Journal of college student development: JCSD ACPA. Band 44 (29), S. 185-203.

Johnson, V. E. (2003). Grade Inflation: A Crisis in College Education. New York: Springer Verlag.

Juola, A. E. (1976). Grade Inflation in Higher Education: What Can Or Should We Do? Paper presented at the Annual Meeting of National Council on Measurement in Education in San Francisco, California.

Kalthoff, H. (1996). Das Zensurenpanoptikum. Zeitschrift für Soziologie, Jg. 25, Heft 2, S. 106-124.

Kuh, G. D. & Hu, S. (1999). Unraveling the Complexity of the Increase in College Grades from the Mid-1980s to the Mid-1990s. Educational Evaluation and Policy Analysis, Vol. 21 (3), S.297-321.

Kvale, S. (1972). Prüfung und Herrschaft. Weinheim: Beltz.

Lämmert, E. (1981). Diplomprüfungen im Widerstreit Die Funktion von Hochschulabschlussprüfungen für das Studium und für den Beruf. Symposium am 29. und 30.April 1981, Berlin 1981, S. 9-15.

Lüders, M. (2001). Dispositionsspielräume im Bereich der Schülerbeurteilung. Auch ein Beitrag zur Professions- und Organisationsforschung. Zeitschrift für Pädagogik Jg. 47, Heft 2, S. 217-234.

Mayring, P. (2010). Qualitative Inhaltsanalyse. Grundlagen und Techniken, 11. Auflage, Weinheim und Basel: Beltz.

Müller-Benedict, V & Grözinger, G. (Hrsg.) (2017). Noten an Deutschlands Hochschulen. Wiesbaden: VS Verlag für Sozialwissenschaften.

Müller-Benedict, V. & Tsarouha, E. (2011). Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen. Zeitschrift für Soziologie, 40 (5), S. 388–409.

Prahl, H.-W. (1995). Prüfungen. In Huber, L. (Hrsg.) Ausbildung und Sozialisation in der Hochschule, In Lenzen, D. (Hrsg.) unter Mitarbeit von Schründer, A., Enzyklopädie Erziehungswissenschaft, Band 10, (S. 438-450). Stuttgart: Ernst Klett Verlag für Wissen und Bildung.

Tsarouha, E. (2017). Typologie der Einflussgrößen auf die Notengebung. In Müller-Benedict, V & Grözinger, G. (Hrsg.), Noten an Deutschlands Hochschulen (S.117-169). Wiesbaden: VS Verlag für Sozialwissenschaften.

Tsarouha, E. (2019). Prüfungspraktiken an deutschen Hochschulen. Eine empirische Studie zu systematischen Einflussgrößen auf die Notengebung in Abschlussprüfungen. Wiesbaden: Springer VS.

Wissenschaftsrat (2003). Prüfungsnoten an Hochschulen 1996, 1998 und 2000 nach ausgewählten Studienbereichen und Studienfächern – Arbeitsbericht. Hrsg. v. d. Geschäftsstelle des Wissenschaftsrats, Drucksache 5536–03.

Wissenschaftsrat (2007). Prüfungsnoten im Prüfungsjahr 2005 an Universitäten (einschließlich KH, PH, TH) sowie an Fachhochschulen (einschließlich Verwaltungsfachhochschulen) nach ausgewählten Studienbereichen und Studienfächern – Arbeitsbericht. Hrsg. v. d. Geschäftsstelle des Wissenschaftsrats, Drucksache 7769–07.

Wissenschaftsrat (2012). Prüfungsnoten an Hochschulen im Prüfungsjahr 2010 - Arbeitsbericht mit einem wissenschaftspolitischen Kommentar des Wissenschaftsrates. Hrsg. v. d. Geschäftsstelle des Wissenschaftsrats, Drucksache 2627-12.