Mixed methods im Team am Beispiel eines Projekts zur langfristigen Entwicklung und Unterschieden von Hochschulexamensnoten

Volker Müller-Benedict, Elena Tsarouha, Thomas Gaens

Beitrag zur Ad-Hoc-Gruppe »Der Beitrag soziologischer Mixed Methods Forschung zur Untersuchung komplexer globaler und nationaler Entwicklungen«

In diesem Beitrag möchten wir unsere Erfahrungen über die Anwendung von mixed methods in einem Forschungsprojekt (Müller-Benedict, Grözinger 2017; DFG-Projekt MU1625/7) präsentieren. Daraus werden wir am Ende des Beitrags Einschätzungen über die Möglichkeiten und Grenzen der Methodik ableiten. Zunächst werden in Abschnitt 1 die Fragestellungen des Projekts und die sich daraus ergebenden Methoden dargestellt. Damit konnten am Anfang des Projekts Erwartungen an den Ertrag der Verwendung dieser Methoden formuliert werden, deren Einlösung im Projektverlauf zu überprüfen war. Im zweiten Abschnitt wird der quantitative Teil des Projekts mit Beispielen der Ergebnisse dargestellt, im dritten ebenso der qualitative Teil. Im vierten Abschnitt geht es dann um die Verknüpfung beider Teile, die zu weitergehenden Erklärungen führt. Im letzten Abschnitt leiten wir Überlegungen ab, welche Bedingungen in unserem Projekt die ertragreiche Anwendung von mixed methods einerseits begünstigt, andererseits erschwert haben.

1. Das Projekt und die Methoden

1.1 Vorüberlegungen und Erwartungen

Die grundsätzliche Fragestellung des Projekts war: „Wie, wie stark und warum unterscheiden sich die Examensnoten an Deutschlands Hochschulen leistungsunabhängig voneinander?“ Es setzte an der seit einiger Zeit auch in Deutschland aufkommenden Diskussion über „grade inflation“, also immer besseren Noten ohne eine dahinterstehende Verbesserung der Leistung an. Damit verbunden ist die Diskussion über gerechte Benotung, zum Beispiel bei der Verwendung von BA-Examensnoten für die Zulassung zum Masterstudium. Wie man schon an den vielen verschiedenen Frageworten sehen kann, ist hier ein Mix aus quantitativen und qualitativen Fragen vorhanden.

Zu Beginn des Projekts stellten wir einige Vorüberlegungen zu den möglichen Bezügen zwischen quantitativer und qualitativer Analyse an. Bei den Fragestellungen, so unsere Vermutung, gibt es wenig Anlass für Gemeinsamkeiten. Man lernt schon in den Einführungen in Forschungsmethoden:

Quantitative Forschung fragt nach was und wie viel, qualitative nach wie und warum. Diese Zweiteilung sieht man auch gut in unserer Forschungsfrage. Der Forschungsgegenstand, die Examensnoten, ist zwar derselbe für beide Forschungsrichtungen, aber unklar ist die Beziehung zwischen den Ergebnissen, die sich aus den unterschiedlichen Fragerichtungen ergeben. Allerdings sahen wir bei den aus den Forschungsfragen abgeleiteten Hypothesen (des quantitativen Bereichs) bzw. forschungsleitenden Vorüberlegungen (des qualitativen Bereichs) eher mögliche Gemeinsamkeiten. Ein Beispiel:

> *Hypothese:* Es gibt über den ganzen Zeitraum stabile Unterschiede im Notenniveau zwischen einzelnen Fächern, zum Beispiel Germanistik und Mathematik. Dazu passte die
>
> *forschungsleitende Vermutung*: Die Notenfindung spielt sich in den Kommissionen der Germanistik- und Mathematik-Prüfungen unterschiedlich ab, das heißt es könnten zum Beispiel unterschiedliche Kriterien eine Rolle spielen.

Zur Organisation der Forschung im Projekt orientierten wir uns an den verschiedenen Datengrundlagen. Wir hatten drei verschiedene Datensätze: erstens lange Zeitreihen von Noten an ausgewählten Hochschulen, die aus Archiven per Hand erhoben wurden (Datensatz ZA8622, erhältlich auf www.gesis.org/histat). Zweitens Vollerhebungen aller Noten an allen Hochschulen ab 1998 vom Statistischen Bundesamt (Statistisches Landesamt 2012). Drittens Texte aus Einzelinterviews und Gruppendiskussionen mit Prüferinnen und Prüfern an Hochschulen. Dafür bildeten wir drei Arbeitsgruppen, für die beiden ersten Datenbasen benutzten wir wie kaum anders möglich, quantitative Methoden, für die dritte Datenbasis qualitative. Die Gruppen analysierten ihre Bereiche jeweils autonom, es fand natürlich gegenseitige Information, aber keine Abstimmung oder Beeinflussung statt. Das heißt wir arbeiteten mit einem parallelen Design und hatten dabei natürlich die Hoffnung auf komplementäre Ergebnisse, also auf einen Zugewinn der Erkenntnis, wenn wir die Ergebnisse beider Analysen zusammenführen würden.

Die Erwartungen an mixed methods wurden im Forschungsantrag folgendermaßen formuliert: „Kann man die *Einflussgrößen auf der Makroebene* beziehen auf die in den Prüfungssituationen vorliegenden, von den *Prüfern wahrgenommenen Einflüsse*, gehen sie in dieselbe Richtung und weisen für verschiedene Situationen dieselben Unterschiede auf?“ (Projektantrag, S.16).

Die theoretischen Grundlagen für beide Herangehensweisen überschneiden sich. Das Konzept der „Fachkulturen“ und ihre Beschreibungen zum Beispiel bieten für die quantitative und qualitative Forschung dieselbe theoretische Grundlage für die Analyse der Fächerunterschiede, ebenso die pädagogisch-psychologische Testtheorie und ihre unterschiedlichen Konzepte von Bewertung und Benotung dieselbe Grundlage für die Analyse von Notenunterschieden. Aber das verhindert nicht einige Ergebnisse, die nur additiv nebeneinandergestellt, aber nicht aufeinander bezogen werden können.

Die im obigen Zitat genannten Einflussgrößen auf der Makroebene sind letztlich Differenzen oder Korrelationen von Variablen, hier einige Beispiele (Müller-Benedict, Grözinger 2017, S.28, 105, 104, 175):

- Die Fächer unterscheiden sich im Notenniveau erheblich, Jura hat die schlechteste, Biologie die beste Durchschnittsnote, Variable: Fach
- Weibliche Professoren geben leicht bessere Noten in vielen Fächern, Variable: Geschlecht
- Mehr Prüflinge führen oft zu schlechteren Durchschnittsnoten, Variable: Anzahl Examen pro Semester
- Diplom und Magister haben bessere Durchschnittsnoten als Staatsexamen, Variable: Abschlussart.

Für solche und andere Variablen schließen sich die Fragen an: Welche Bezüge zu den qualitativen Ergebnissen wird es geben? Gibt es Themen oder Schlüsselbegriffe oder Typologien, auf die man Variablen beziehen kann? Als ein Beispiel dafür, worauf es im Folgenden hinausläuft, betrachten wir die Variable Abschluss: Es stellte sich in den Gruppendiskussionen heraus, dass die Prüfenden in Prüfungen für Mathematik-Diplom deshalb ein höheres Notenniveau erwarteten, weil neben weiteren Aspekten die Bearbeitungszeit für die Diplomarbeit wesentlich länger ist als für eine schriftliche Staatsexamensarbeit in Mathematik (Tsarouha 2019, S.272f.).

1.2 Datenaufnahme

Bei der Datenaufnahme achteten wir darauf, Erhebungsorte möglichst so abzustimmen, dass die quantitativen Daten und die Prüfungserfahrungen der Gesprächsteilnehmenden zusammengebracht werden konnten.

Für die quantitative Stichprobe wurden acht Hochschulen und ein Landesarchiv (FU Berlin, TU Braunschweig, Göttingen, Heidelberg, Hildesheim (Archiv Lehramt), Münster, KIT Karlsruhe, Saarbrücken (nur Germanistik), Tübingen) ausgewählt, die die Überprüfung regionaler (zum Beispiel Bundesland politisch/Stadtstaat vs. Flächenstaat) sowie hochschulspezifischer Unterschiede (zum Beispiel Alter/Größe) zulassen und Bezüge zu den gewünschten Kontrastierungen in den Gruppendiskussionen erlauben. Die Auswahl der Studiengänge sollte das Fächergruppenspektrum abbilden und die Überprüfung von Unterschieden nach Zugangsvoraussetzungen, der Heterogenität der Studieninhalte (zum Beispiel standardisierter Studienverlauf versus große Auswahl) und nach Abschlussart (Staatsexamen versus Diplom/Magister) ermöglichen. Erforderlich war schließlich die Zustimmung der zuständigen Einrichtung. (Die Daten sind erhältlich bei der GESIS unter ZA8622.)

Im Laufe der Datenerhebung in den Archiven wurde eine gemeinsame Datenbasis erstellt: Das qualitative Teilprojekt hatte ständigen Zugriff auf ausgewählte Prüfungsakten, die quantitativen Rohdaten und auf aufbereitete Zeitreihen von Abschlussnoten, außerdem auf die im Zuge der Datenerhebungen ebenfalls erfassten Prüfungsordnungen aller im sample enthaltenen Studiengänge. Zudem wurden auf Bitte aus dem qualitativen Teilprojekt gelegentliche Auswertungen vorgenommen – zum Beispiel hinsichtlich von Notensprüngen, möglichen Tendenzen zur Mitte bei der Notenvergabe oder der Frage, ob die Noten normalverteilt sind.

Für die qualitative Erhebung wurden sechs Einzelinterviews und neun Gruppendiskussionen geführt. Die Einzelinterviews wurden teilweise thematisch zusammengefasst (Expertengespräche) oder inhaltsanalytisch (angelehnt an Mayring 2010) analysiert (problemzentrierte Interviews). Die Einzelinterviews dienten unter anderem der Vorbereitung der Gruppendiskussionen. Die Gruppendiskussionen wurden mittels der dokumentarischen Methode (Bohnsack 2014) analysiert. Fünf Gruppendiskussionen erfolgten mit Professorinnen und Professoren und vier mit ministerial berufenen Prüfungsvorsitzenden in Ersten Staatsexamen. Als Rekrutierungsbedingung von Prüfenden wurden Prüfungserfahrungen in den Bundesländern Baden-Württemberg und Niedersachsen festgelegt. Es standen Prüfungserfahrungen an mehreren Universitäten in den Disziplinen Mathematik und Germanistik im Fokus, beschränkt auf die Studiengänge Mathematik Diplom, Germanistik auf Magister sowie Lehramt an Gymnasien für die Unterrichtsfächer Mathematik und Deutsch. Im Zusammenhang mit dem Studiengang Lehramt an Gymnasien gibt es wesentliche bundeslandspezifische Unterschiede hinsichtlich des Einsatzes der ministerial berufenen Prüfungsvorsitzenden: in Niedersachsen wurden diese fächerübergreifend eingesetzt und in Baden-Württemberg bis über den Zeitpunkt der Erhebung hinaus fachspezifisch, gemäß der eigenen Facultas Docendi.

Die Bundesländer, die Disziplinen Mathematik und Germanistik und die Abschlüsse wurden für die qualitative Datenerhebung in dieser Weise ausgewählt, um gegebenenfalls Auswirkungen gegebener Unterschiede kontrastieren und Gemeinsamkeiten in den Prüfungspraktiken zu identifizieren. Darüber hinaus wurden die genannten Bundesländer auch ausgewählt, um diese Ergebnisse auf die aus Universitätsarchiven erhobenen Noten beziehen zu können.

Bei der Zusammensetzung der Diskussionsgruppen (drei in Niedersachsen und sechs in Baden-Württemberg) wurde darauf geachtet, dass möglichst homogene Gruppen zusammengesetzt wurden, zum Beispiel nur ministerial berufene Prüfungsvorsitzende aus demselben Bundesland, idealerweise mit Prüfungserfahrungen in einer gemeinsamen Disziplin oder Prüferinnen und Prüfer der Mathematik in einem Bundesland und häufig mit Prüfungserfahrungen an derselben Universität, sowohl mit Prüfungen des Ersten Staatsexamens als auch mit Diplomprüfungen. In den Gruppendiskussionen mit geteilten disziplinspezifischen Prüfungserfahrungen konnte ein gemeinsamer Erfahrungsraum aufgespannt werden, in dem sich die Prüfenden gegenseitig ergänzt und bestärkt, aber auch widersprochen haben. Zwei der Gruppendiskussionen mit ministerial berufenen Prüfungsvorsitzenden wurden mit Prüfungsvorsitzenden mit unterschiedlichen Facultas Docendi durchgeführt. In diesen Gesprächen wurden disziplinspezifische Prüfungspraktiken stärker kontrastiert.

2. Ergebnisse im quantitativen Teilprojekt

Die vorrangige Forschungsfrage im quantitativen Teilprojekt lautete im Anschluss an die aktuelle Debatte um eine Noteninflation an deutschen Hochschulen: Wie entwickeln sich die Noten im Zeitverlauf? Um Unterschiede und Gemeinsamkeiten zwischen Fächern, zwischen Hochschulen und zwischen Abschlüssen aufdecken zu können, wurden auf Studiengang- und Hochschulebene aggregierte Durchschnittsnoten untersucht (zu den folgenden Ergebnissen siehe ausführlicher Gaens 2018).

Es zeigt sich eine langfristig herrschende Notenhierarchie: Die Rangfolge beginnt mit Biologie als Studiengang mit den besten Noten, dicht gefolgt von Psychologie. In den ersten juristischen Staatsexamen werden im Durchschnitt die schlechtesten Noten vergeben, sie sind mehr als doppelt so hoch wie in Biologie. In den beiden wirtschaftswissenschaftlichen Studiengängen BWL und VWL sind die schlechtesten Durchschnittsnoten innerhalb der in der Stichprobe enthaltenen Diplomstudiengänge (Betriebswirtschaftslehre, Biologie, Chemie, Maschinenbau, Mathematik, Psychologie, Volkswirtschaftslehre) zu finden, die beiden Lehramtsstudiengänge und Germanistik liegen zwischen diesen Polen. Damit bestätigen sich die Eindrücke bisheriger Erhebungen, nach denen in den Naturwissenschaften bessere Noten als in den Geisteswissenschaften vergeben werden, während in ingenieur- und wirtschaftswissenschaftliche Studiengängen und vor allem in den Rechtswissenschaften die schlechtesten Noten erteilt werden.

Entgegen dem vielfach öffentlich vermittelten Eindruck entwickeln sich die durchschnittlichen Abschlussnoten nicht nur konstant zum Besseren, sondern in zwei unterschiedlichen Verlaufsformen: In der Mehrzahl (8) der untersuchten Studiengänge verbessern sie sich tatsächlich langfristig, dann allerdings begleitet durch zyklische Schwankungen, und in drei Studiengängen verlaufen sie zyklisch auf einem relativ stabilen Notenniveau. Die Diplomstudiengänge zählen alle zur ersten Gruppe, ebenso beide erfasste Lehramtsstudiengänge. Die beiden Magisterstudiengänge Germanistik und Soziologie sowie das erste juristische Staatsexamen zählen zu Letzteren. Völlig konstante Notenniveaus sind über alle Prüflinge gemittelt nicht zu finden, auch langfristige Verschlechterungen der Noten sind auf diesem Aggregatniveau nicht zu beobachten.

In den Studiengängen mit zyklischem Verlauf auf relativ stabilem Notenniveau verlaufen die Noten in unterschiedlichen Schwankungsbreiten, die Zyklen dauern aber wie auch in den Studiengängen mit Verbesserung circa 20 Jahre. Auf Hochschulebene gibt es durchaus Abweichungen vom Studiengangtrend, vereinzelt sind hier auch langfristige Verschlechterungen des Notenniveaus zu finden – dies ist aber die absolute Ausnahme. Häufig verlaufen die Noten an den Hochschulen gemäß dem Trend im Studiengang und die (Rang-) Differenzen zwischen den Studiengängen sind im Zeitverlauf relativ stabil.

Schließlich zeigen die Daten abschlussspezifische Muster, wie in den Abbildungen 1 bis 4 zu sehen ist. Dort abgebildet sind die Fächer und Bundesländer, die Gegenstand der Gruppendiskussionen waren: Mathematik und Germanistik in Baden-Württemberg und Niedersachen.

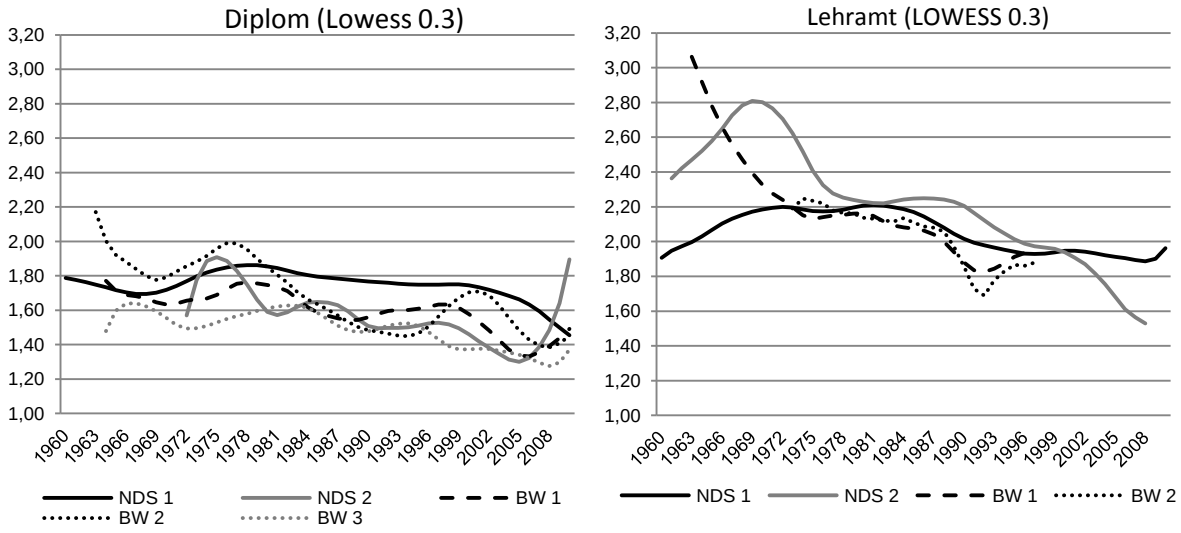

Abbildung 1/2: Durchschnittliche Abschlussnoten an den Hochschulen in Mathematik im Zeitverlauf

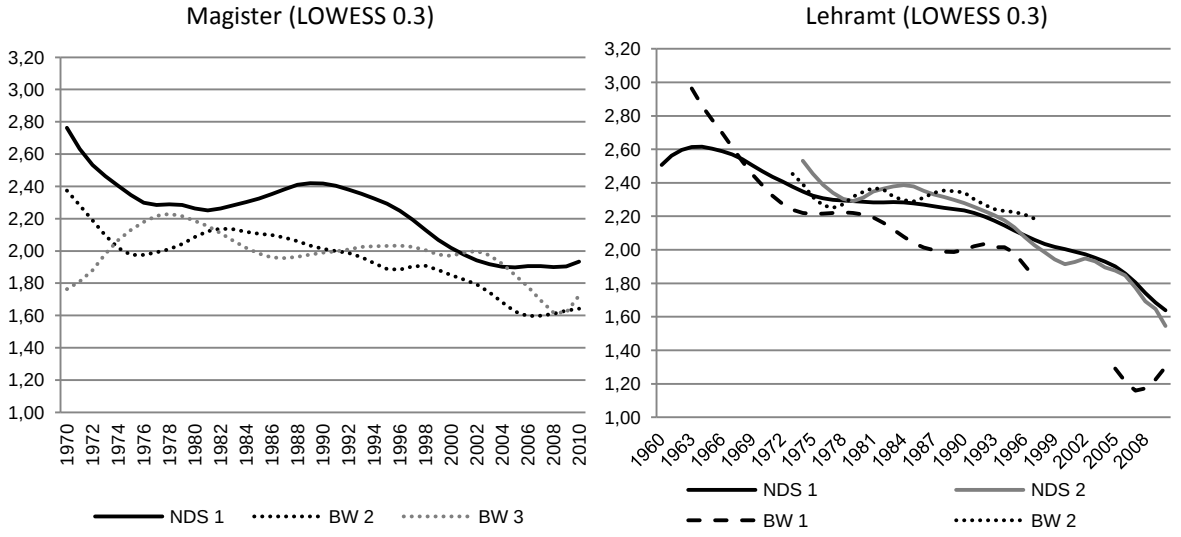

Abbildung 3/4: Durchschnittliche Abschlussnoten an den Hochschulen in Germanistik im Zeitverlauf

Um kurzfristige Semesterschwankungen auszugleichen, sind die hier abgebildeten Zeitreihen mit der LOWESS Technik geglättet, einem Anpassungsverfahren für Daten, das die grafische Analyse des Zusammenhangs von Variablen erleichtert. Die Glättung der Daten wird durch ein Regressionsmodell erreicht.

Als erstes springen die zyklisch verlaufenden langfristigen Verbesserungen ins Auge. Wie weiterhin zu sehen ist, werden in Mathematik Lehramt über die gesamte Zeit hinweg betrachtet schlechtere Noten vergeben als im entsprechenden Diplomstudiengang, wobei im Zeitverlauf durch die im Lehramt deutlich stärkere Verbesserung eine Angleichung stattfindet. In Deutsch ist der Unterschied zwischen den unterschiedlichen Abschlussarten nicht so groß wie in Mathematik, auch hier zeigt sich jedoch über die Jahre im Lehramt eine stärkere Verbesserung als im Magister.

Während das Notenniveau in Germanistik Magister zu den häufigsten Messzeitpunkten deutlich über dem in Mathematik Diplom liegt und damit schlechter ist, zeigt sich innerhalb des Abschlusses Lehramt keine derart deutliche Differenz zwischen den beiden Fächern. Auffällig sind außerdem die parallelen Entwicklungen der Noten an den baden-württembergischen Hochschulen in Mathematik Lehramt, sowie an den niedersächsischen Hochschulen in Deutsch Lehramt.

Zusammengefasst ergeben diese Beobachtungen eine Reihe von stabilen und langfristigen Notenunterschieden, die offenbar wenig mit unterschiedlichen Leistungen der Geprüften zu erklären sind. Deshalb stellt sich die Frage: Gibt es Erklärungen in der Praxis der Notengebung für diese hier beispielhaft aufgeführten Differenzen und Dynamiken?

3. Ergebnisse im qualitativen Teilprojekt

Trotz einer detaillierten Rechtsgrundlage von Abschlussprüfungen gibt es auch nicht definierte Räume, die basierend auf dem Verständnis von Prüfungen seitens der Professorenschaft und der ministerial berufenen Prüfungsvorsitzenden unterschiedlich ausgefüllt werden können. Die Handlungsspielräume können grundsätzlich vielseitig determiniert sein, wie im weiteren Verlauf an Hand von Beispielen dargelegt wird. Es gibt Einflüsse, die sich aufgrund formaler Bestimmungen im Prüfungsgeschäft ergeben und solche, die dem subjektiven Ermessen der Prüfenden oder des Prüfungsausschusses bzw. der Prüfungskommission zuzuordnen sind.

Vor diesem Hintergrund lautet die Forschungsfrage der qualitativen Untersuchung: *Gibt es systematische nicht-leistungskonforme Einflussgrößen auf die Notengebung in den Abschlussprüfungen an deutschen Hochschulen*? (für eine ausführliche Darstellung der Ergebnisse dieser Untersuchung siehe Tsarouha 2019).

Die Forschungsfrage zielt auf komplexe Einflussgrößen auf die Notengebung, vorrangig in mündlichen Prüfungen. Basierend auf einer Vorstudie (Müller-Benedict, Tsarouha 2011) und konzeptionellen Überlegungen wurden anfänglich drei Strukturmerkmale für systematische Notenunterschiede herangezogen. Diese Strukturmerkmale sind: Disziplinen, Fächer (hier gemeint als standortgebundene Besonderheiten) und Studiengänge. Bei der konzeptionellen Ausdifferenzierung wurde eine Vielzahl bestehender Literatur zum Thema ‚Fachkulturen' berücksichtigt, mit dem Ziel eine konsistente Begriffsverwendung und begriffliche Differenzierung der Einflussgrößen zu ermöglichen.

Aus dem qualitativen Datenmaterial heraus haben sich, angelehnt an die genannten Strukturmerkmale, Typen an Einflussgrößen identifizieren und weiter differenzieren lassen. Durch die weitere Differenzierung der Einflussgrößen haben sich innerhalb des Projektes die verwendeten Begrifflichkeiten (Strukturmerkmale/Typen/Einflussgrößen) in der qualitativen Forschung weiterentwickelt. Die

festgestellten Einflussgrößen auf die Notengebung sind vielseitig. Sie können unabhängig voneinander nebeneinander existieren oder in Zusammenhang miteinander stehen, sich kumulieren oder gegenseitig verstärken. Darüber hinaus können sich verschiedene Einflussgrößen aber auch aufgrund ihrer gegenteiligen Wirkungsweisen ausbalancieren bzw. nivellieren. Insgesamt wurden sechs Typen an Einflussgrößen aus dem Datenmaterial heraus in einem zirkulären Prozess mit einer parallel verlaufenden konzeptionellen Aufbereitung bestehender Literatur gewonnen. Diese sechs Typen sind:

Disziplin-, fach-, kommission-, bundesland-, studiengang- und abschlussspezifische Typen von Einflussgrößen (Tsarouha 2019, S.205ff; 2017, S.127ff). Nachfolgend werden drei der genannten Typen an Einflussgrößen erläutert:

(1) Disziplinspezifische Einflussgrößen am Beispiel der *fachlichen Erwartungen*
Die disziplinspezifischen Einflussgrößen sind typisch für eine bestimmte Disziplin. Sie existieren demnach über verschiedene Universitäten und teilweise über die beiden in dieser Studie gegenübergestellten Studiengänge (Lehramt versus Diplom/Magister) hinweg.
Für die Disziplin Mathematik lassen sich ebenfalls spezifische Einflüsse feststellen. Einige Befragte gaben an, dass es Professorinnen und Professoren mit der Meinung gäbe, *„die Guten machen Diplom und wenn es dafür nicht reicht, dann macht er halt Lehramt“* (BW_SE_M, S.63). Das bedeutet, dass Prüfende in der Mathematik an das jeweilige Studentenklientel unterschiedliche Erwartungen bezüglich des Leistungsniveaus haben können, die zum Beispiel in den Ersten Staatsexamen möglicherweise zu einem negativen Einfluss auf die Notengebung führen. Dies ist spezifisch für die Mathematik und wird von den Prüfenden der Germanistik nicht bestätigt. Vielmehr äußerten einige Probanden, dass sie eher keine Unterschiede hinsichtlich des Leistungsniveaus zwischen Prüflingen in Magisterprüfungen und Prüflingen der Ersten Staatsexamen erwarten würden (Tsarouha 2019, S.218ff.).

(2) Fachspezifische Einflussgrößen am Beispiel der *Notenskala (für Diplom/Magister)*
Fachspezifische Einflussgrößen sind typisch für einzelne Fächer bestimmter Universitäten. Sie können auf historisch gewachsene strukturelle Gegebenheiten und Institutionalisierungsprozesse der jeweiligen Universität zurückgeführt werden.
Einzelne Personen äußerten, dass die Noten immer besser würden, wenn die Notenskala kleinschrittiger sei. Eine mögliche Erklärung dafür könnte sein, dass Prüfende den Wunsch haben, dass sich in den Noten auch Unterschiede der erzielten ‚Punkte‘ zeigen. Entsprechend verändert sich das Notenniveau, wenn zum Beispiel vier Punkte pro Notenschritt vergeben werden und dabei einmal die Notenschritte 1, 1,5, 2 etc. zur Verfügung stehen und im Vergleich dazu die Notenschritte 1, 1,3, 1,7, 2 etc. Der Einfluss der verwendeten Notenskala ist nicht-leistungskonform, da vergleichbare Leistungen zu unterschiedlichen Notenergebnissen führen.
Für die Studiengänge des Lehramts an Gymnasien wirkt der Einfluss der Notenskala ebenfalls nicht-leistungskonform. Allerdings ist dieser Einfluss dann nicht fach-, sondern bundeslandspezifisch (Tsarouha 2019, S.234f, 258).

(3) Studiengangspezifische Einflussgrößen am Beispiel des *Betreuungsverhältnisses*
Hierunter werden Einflussgrößen verstanden, welche spezifisch für bestimmte Studiengänge sind, die aus einer Disziplin und einem gegebenen Abschluss bestehen. Diese Einflussgrößen wirken für diese Studiengänge über Universitäten hinweg.
In der Mathematik erwähnten sowohl Befragte der Professorenschaft als auch der ministerial berufenen Prüfungsvorsitzenden, dass es eine studiengangspezifische Diskrepanz gäbe. Bei der Erstellung der Diplomarbeit würde eine intensive Betreuung stattfinden, während nur wenige Staatsexamenskandidaten und Staatsexamenskandidatinnen ihre Zulassungsarbeit im Unter-

richtsfach Mathematik anfertigen würden. Zusätzlich würden häufig die Betreuerinnen und Betreuer der Diplomarbeiten auch die mündlichen Prüfungen abnehmen. Die Intensität der Betreuung der Abschlussarbeit wird im Kontext der Diplomprüfungen als nicht-leistungskonformer Einfluss eingestuft, der sich positiv auf die Noten auswirken könnte. Eine geringere Intensität der Betreuung bei der Erstellung der Abschlussarbeit im Studiengang Lehramt an Gymnasien für das Unterrichtsfach Mathematik könnte zu einer objektiveren Bewertung führen, so dass unter anderem aufgrund fehlender Loyalitätsempfindungen und einer vergleichsweise geringeren Empathie insgesamt schlechtere, aber leistungskonforme Bewertungen resultieren könnten (Tsarouha 2019, S.265f).

4. Zusammenführung der quantitativen und qualitativen Ergebnisse

Die Zusammenführung der Ergebnisse der quantitativen und qualitativen Forschungsschwerpunkte soll nachfolgend an Beispielen dargestellt werden. Dabei werden Differenzierungen, Erweiterungen bzw. Erklärungen in beiden Wirkungsrichtungen, von qualitativ auf quantitativ und umgekehrt, dargestellt.

4.1 Qualitative Ergebnisse erweitern quantitative Ergebnisse

Zunächst soll am Beispiel der Mathematik durch die Kontrastierung der Prüfungspraktiken des Diplomstudiengangs und des Studiengangs Lehramt an Gymnasien für das Unterrichtsfach Mathematik eine Ergebniszusammenführung erfolgen. Hierfür können exemplarisch einige qualitativ ermittelte Einflussgrößen zur Klärung der durchschnittlichen Notenunterschiede zwischen der Diplomprüfung und dem Ersten Staatsexamen herangezogen werden. Die bereits angesprochenen Einflüsse der *Bearbeitungszeit* (abschlussspezifischer Einfluss, siehe oben) und des *Betreuungsverhältnisses* (studiengangspezifischer Einfluss) können sich tendenziell positiv auf die Abschlussnoten im Diplom auswirken.

Weitere Einflüsse sind *Fächerwahl* und *Formalität*. Die Fächerwahl ergibt sich im Kontext des studiengangspezifischen Typs. Einige befragte Personen gaben an, dass Mathematik als Unterrichtsfach seitens der Studierenden tendenziell eher aus strategischen Gründen als aus persönlicher Neigung für die Disziplin gewählt würde. Dies könnte dazu führen, dass durch eine geringere Neigung und Motivation eine geringere Leistung erbracht wird und leistungskonform schlechtere Noten resultieren (Tsarouha 2019, S.267f).

Im Kontext des Typs kommissionsspezifischer Einflussgrößen ist die Formalität genannt (Tsarouha 2019, S.239ff). In Ersten Staatsexamen würde eine erhöhte Formalität in der mündlichen Prüfung aufgrund der Anwesenheit der oder des ‚fremden' ministerial berufenen Prüfungsvorsitzenden wahrgenommen, die zu einer erhöhten Nervosität führen könne. Dieser Einfluss könnte sich im Studiengang Lehramt an Gymnasien für das Unterrichtsfach Mathematik negativ auf das Notenniveau auswirken, da manche Prüflinge, nach Angaben von Befragten, unter ihrem Leistungspotential bleiben würden, und ist demnach als nicht-leistungskonform einzustufen. Einige Prüfende der Germanistik in Baden-Württemberg gaben an, dass die erhöhte Formalität in den Ersten Staatsexamen dazu führe, dass sich die Prüflinge im Vergleich zu den Magisterprüflingen besser vorbereiten würden. Dieser Einfluss könnte sich somit studiengangspezifisch unterschiedlich auswirken.

Das zweite Beispiel bezieht sich auf die Notenunterschiede zwischen den Studiengängen Mathematik Diplom und Germanistik Magister. Aus den Gruppendiskussionen geht hervor, dass in der Mathe-

matik aufgrund der *Struktur des Wissens* eine höhere studienbegleitende Selektion stattfinden würde, so dass leistungskonform bessere Abschlussnoten erzielt würden. Dieser Einfluss ist disziplinspezifisch (Tsarouha 2019, S.221). Das Vordiplom in Mathematik sorge zusätzlich für eine hohe *Selektion*, wodurch nur bessere, also leistungsstärkere Studierende in den Abschlussprüfungen vertreten seien. Dieser studiengangspezifische Einfluss ist ebenfalls leistungskonform (Tsarouha 2019, S.262f).

Manche Befragte äußerten, dass für die Prüflinge im Studiengang Germanistik Magister trostlose *Berufsaussichten* gegeben seien. Diese könnten zu einer nicht-leistungskonformen milderen Bewertung in Prüfungen führen. Es handelt sich um einen disziplinspezifischen Einfluss (Tsarouha 2019, S.224). Dadurch, dass in der Abschlussprüfung des Mathematik Diploms häufig nur eine prüfende und eine beisitzende Person anwesend seien, sei die Atmosphäre sehr *familiär*. Es ist die Frage zu stellen, inwiefern die Kommissionszusammensetzung in diesen Prüfungen zu nicht-leistungskonform milderen Bewertungen führen kann (Tsarouha 2019, S.240).

Anhand der angeführten, aus der qualitativen Untersuchung validierten, Prüfungspraktiken lassen sich Erklärungsansätze für die quantitativ belegten Unterschiede in den Notenniveaus ableiten. An den Beispielen ist die Tendenz erkennbar, dass in Abschlussprüfungen des Studiengangs Mathematik Diplom sowohl leistungskonform als auch nicht-leistungskonform positive Einflussgrößen gegeben sind, die ein besseres durchschnittliches Notenniveau erklären könnten.

Es wird aber ebenfalls deutlich, dass die verschiedenen Einflüsse sehr komplex sind und keine kohärente Argumentation für gegebene Notenunterschiede erlauben. So gibt es zum Beispiel ebenso positive Einflussgrößen, die im Kontext der Germanistik wirken können und dabei das durchschnittliche Notenniveau nicht unter das der Mathematik absenken. Es zeigt sich, dass durchschnittliche Notenunterschiede mit den derzeitigen Analysen nicht auf wenige richtungsweisende Einflüsse reduziert und ausreichend begründet werden können.

4.2 Quantitative Ergebnisse erklären qualitative Ergebnisse

Hier stehen beispielhaft umgekehrt zwei Zitate aus den Gruppendiskussionen, die mit Hilfe der quantitativen Analysen vertieft und differenziert werden konnten, also auch die umgekehrte Erklärungsrichtung war möglich:

> „[…] Hat man diesen Dreierschritt den es jetzt gibt […] [1,0]. 1,3. 1,7 und dann 2 oder hat man wie es früher im Staatsexamen war 1,0. 1,5 oder 2. Wenn man so viele Zwischenschritte hat werden […] [die] Noten besser […]“ (BW_Uni_D_RB1, S.9).

> „[…] Nur (2) Physik und Mathe sind die schicksalhaftesten mündlichen Prüfungen […] Weil es keine Klausuren drunter gibt“ (BW_SE_G, S.52).

Die Zitate können mit Hilfe der quantitativen Analysen in der umgekehrten Erklärungsrichtung vertieft und differenziert werden. In der Tabelle 1 abgebildet sind aus Platzgründen aus multiplen Regressionsmodellen entnommene Effekte der verwendeten Notenskala sowie der Anteile der mündlichen Prüfungen an den Prüfungen insgesamt. Diese Untersuchung konnten wir nur für solche Studiengänge durchführen, bei denen auch genügend unterschiedliche Prüfungsvorschriften in den Hochschulen vorhanden waren. Es zeigt sich, dass die Gesamtnote stets umso besser wird, je geringer die Anzahl weiterer mündlicher Prüfungen zusätzlich zu den Abschussarbeiten oder -klausuren ist. Wenn dann nur eine einzige Prüfung vorhanden ist, wie in dem zweiten Zitat, wirkt das erst recht notenverbessernd, weil sie dann natürlich eine sehr schicksalhafte Bedeutung hat.

Es zeigt sich aber auch, dass das erste Zitat so generell nicht zutrifft, sondern disziplinspezifisch unterschiedlich gilt. Ein stärker differenziertes Notensystem weist zwar in Biologie, Chemie und VWL

einen notenverbessernden, in Psychologie aber einen notenverschlechternden Effekt auf. Diese studiengangspezifischen Effekte zeigen, dass die Aussagen der Prüfenden stets an ihren konkreten fachspezifischen Prüfungskontext gebunden sind.

Tabelle 1: Effekte der verwendeten Notenskala sowie der Anteile der mündlichen Prüfungen

	Beta
Chemie Diplom (n=6 087; r^2(adj)=0,11)	
Notensystem[a]: 0,3/0,7 Differenzierung	-0,270***
Anteil Arbeit vs. mündlich=1/3[c]	-0,072***
Biologie Diplom (n=5 375; r^2(adj)=0,03)	
Notensystem[b]: Ganze Noten	0,126***
Psychologie Diplom (n=8 690; r^2(adj)=0,08)	
Anteil Klausur(en) vs. mündlich	0,096***
Notensystem[b]: Ganze Noten	-0,284***
Notensystem[b]: Halbe Noten	-0,162***
VWL Diplom (n=6 066; r^2(adj=0,26)	
Anteil Klausur(en) vs. mündlich	-0,133***
Notensystem[a]: 0,3/0,7 Differenzierung	-0,295***
BWL Diplom (n=26 242; r^2(adj)=0,12)	
Anteil Arbeit vs. mündlich	-0,034***
Referenzkategorien: [a] Notensystem: keine 0,3/0,7 Differenzierung; [b] Notensystem: 0,3/0,7 Differenzierung; [c] Anteil Arbeit vs. mündlich = 1/4	

5. Wie (gut) hat mixed methods funktioniert?

Wenn wir unsere Erfahrungen zusammenfassen,

(1) können wir positiv feststellen, dass wir dadurch, dass wir sich überschneidende Erhebungseinheiten gewählt haben, sicherstellen konnten, dass die Teilnehmenden der Gruppendiskussionen auch über das reden, was in der quantitativen Analyse analysiert wurde, genauer: einige der Noten, über deren durchschnittliche Differenzen sie geredet haben, haben sie mit einiger Wahrscheinlichkeit selbst gegeben. Ihre Benotung erfolgte dabei im Rahmen der jeweils gültigen Prüfungsordnungen, die wiederum als Einflussgrößen in die quantitativen Analysen eingegangen sind.

(2) erscheint uns als günstige Voraussetzung für mixed methods, dass drei zentrale Variablen der quantitativen Analyse in der qualitativen Analyse ebenfalls als vorstrukturierende Merkmale be-

rücksichtigt wurden. Dadurch sind beide Analysen hinsichtlich der Variablen bzw. Einflussgrößen Disziplin, Standort bzw. Fach und Abschlussart bzw. Prüfungskommission anschlussfähig.

(3) ergab es sich, dass wichtige Strukturmerkmale des Forschungsfelds an sich, wie die Prüfungsordnungen und die studiengangmäßige Organisationsform nach Abschlüssen, die als Kombination von Variablen der quantitativen Analyse vorhanden waren, sich auch als Typen in der qualitativen Analyse herausstellten und deshalb aufeinander bezogen werden konnten.

(4) zeigte sich, dass die parallele Bearbeitung eines gemeinsamen Forschungsgegenstands aus verschiedenen methodischen Zugängen heraus, wenn auch mit jeweils unterschiedlichem Fokus durchgeführt, gegenseitige Inspiration für die Teilvorhaben zu schaffen vermag, da die gemeinsame Expertise im Austausch immer wieder zwangsläufig durch einen Perspektivwechsel begleitet wird.

Darüber hinaus gibt es eine Reihe von Ergebnissen, zu denen es keine Entsprechungen zwischen beiden Methoden gibt. Dazu zählen vor allem die wichtige Frage der langfristigen Entwicklung, die mit den konjunkturell veränderlichen Anzahlen der Prüflinge zusammenhängt oder die Frage der genderspezifischen Unterschiede der Benotung. Diese Themen erschienen in den Gruppendiskussionen nicht.

Insofern wäre kritisch zu hinterfragen, ob hier das parallele, unabhängige Design geeignet war. Durch ein sequentielles Design, in dem der qualitative Teil an die Ergebnisse des quantitativen Teils stärker gekoppelt gewesen wäre, hätte sich aber vielleicht eine weniger reichhaltige Übersicht über die Einflussgrößen bei der Notengebung ergeben als durch das parallele Design, bei dem die Breite der qualitativen Ergebnisse allein aus dem Textmaterial hervorging. Außerdem ist ein rein sequenzielles Design in einem zeitlich begrenzten Drittmittelprojekt alleine aufgrund des beschränkten Förderzeitraums schwierig umzusetzen – es sollte nicht vergessen werden, dass die Konzeption und Anwendung von mixed methods nicht nur methodischer Natur ist, sondern in einer immer stärker auf Drittmittelfinanzierung angelegten Forschungslandschaft auch forschungspraktischen Zwängen unterliegt.

Literatur

Bohnsack, Ralf. 2014. *Rekonstruktive Sozialforschung. Einführung in qualitative Methoden.* 9. überarbeitete und erweiterte Auflage. Opladen: Barbara Budrich.

Gaens, Thomas. 2018. *Der Einfluss leistungskonformer und leistungsexterner Prüfungsbedingungen auf die Notengebung an deutschen Hochschulen. Eine empirische Untersuchung der langfristigen Entwicklung von Examensnoten*. Flensburg. https://www.zhb-flensburg.de/fileadmin/content/spezial-einrichtungen/zhb/dokumente/dissertationen/gaens/dissertation-thomas-gaens.pdf

Mayring, Philipp. 2010. *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. 11. Auflage. Weinheim, Basel: Beltz.

Müller-Benedict, Volker und Gerd Grözinger (Hrsg.). 2017. *Noten an Deutschlands Hochschulen. Analysen zur Vergleichbarkeit von Examensnoten 1960 bis 2013*. Wiesbaden: Springer VS

Müller-Benedict, Volker und Elena Tsarouha. 2011. Können Examensnoten verglichen werden? Eine Analyse von Einflüssen des sozialen Kontextes auf Hochschulprüfungen. *Zeitschrift für Soziologie* 40:388–409.

Statistisches Landesamt Schleswig-Holstein, Forschungsdatenzentrum. 2012. *Hochschulprüfungsstatistik 1995–2010*. Kiel.

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

Tsarouha, Elena. 2017. Typologie der Einflussgrößen auf die Notengebung. In *Noten an Deutschlands Hochschulen,* Hrsg. Müller-Benedict, Volker und Gerd Grözinger, 117–169. Wiesbaden: Springer VS.