Johanna Lojewski, Thomas Must & Jörg van Norden

Historisch denken Lernen braucht problemorientierte Gegenwartsbezüge

*Empirische Erkenntnisse zu narrativen Kompetenz bei Schüler*innen und Studierenden*

1. Einleitung

Fachspezifische Professionalisierung ist das Kerngeschäft universitärer Fachdidaktik. Mit BiProfessional bot die Qualitätsoffensive Lehrer*innenbildung auch in Bielefeld die Möglichkeit zu evaluieren, was geschichtsdidaktische Lehre leistet.[1] Konkret geht es in dem betreffenden Teilprojekt, das von Thomas Must und Jörg van Norden verantwortet wird, um das Praxissemester, in dem Studierenden ein halbes Jahr in der Schule tätig sind. Diese Zeit wird durch Seminare vorbereitet und begleitet. Dieses inzwischen fest etablierte Format ist durch das Lehrer*innenausbildungsgesetz von 2009 auf den Weg gebracht worden.[2] Es verschränkt universitäre Geschichtsdidaktik mit unterrichtlicher Praxis, sodass erstere auf den Prüfstand kommen und letztere innovative Konzepte aus Empirie und Theorie kennenlernt.

In der Geschichtsdidaktik besteht im Wesentlichen Konsens darüber, dass „Zeit" das proprium historischen Denkens ist und in Narrationen zum Ausdruck kommt. Epistemologisch gefasst, vertritt die Geschichtsdidaktik inzwischen einen narrativen Konstruktivismus.[3] Narrative Kompetenz spielt auch in Lehrplänen eine wichtige Rolle, ist aber nach unserer Beobachtung im Geschichtsunterricht selbst noch nicht recht angekommen. Das wollten wir im Zuge einer ersten Studie an Schulen überprüfen.[4] Wenn Lehrer*innen diese Kompetenz fördern und fordern wollen, müssen auch sie wissen, worum es sich dabei handelt und selbst narrativ kompetent sein. Universitäre Lehre will sie dazu qualifizieren. Ob ihr das in Bielefeld gelungen ist, war Gegenstand einer zweiten Studie, die sich auf die universitäre Geschichtslehrer*innenausbildung im Praxissemester bezog und im Rahmen der Qualitätsoffensive Lehrer*innenbildung durchgeführt worden ist. In beiden Studien ging es um die Lernprogression narrativer Kompetenz und es wurden vergleichbare Testinstrumente verwendet. Forschung in Schule und Universität gingen also Hand in Hand. Insofern werden die Ergebnisse der im Geschichtsunterricht durchgeführten Testung in diesen Beitrag mit einbezogen. Die Auswertung der erhobenen Daten erfolgte zunächst qualitativ.[5] Im Folgenden greifen wir die Kritik auf, die

[1] Siehe: https://www.uni-bielefeld.de/einrichtungen/biprofessional/teilprojekte-(tp)/tp-2-fachdidaktische-konk/tm_5.xml/.
[2] Rahmenkonzeption zur inhaltlichen und strukturellen Ausgestaltung des Praxissemesters im lehramtsbezogenen Masterstudiengang; verfügbar unter: https://www.zfsl.nrw.de/ENG/Praxissemster_alle_Lehraemter/Rechtlicher_Rahmen/Rahmenkonzeption-zur-strukturellen-und-inhaltlichen-Ausgestaltung-des-Praxissemesters-im-lehramtsbezogenen-Masterstudiengang-2010.pdf.
[3] *Jörg van Norden*: Was machst du für Geschichten. Didaktik eines narrativen Konstruktivismus. Freiburg 2011; *Andreas Körber*: Didaktische Perspektiven auf Reenactment als Geschichtsorte. In: *Sabine Stach/Juliane Tomann* (Hrsg.): Historisches Reenactment. Disziplinäre Perspektiven auf ein dynamisches Forschungsfeld (Medien der Geschichte, Bd. 4). Berlin, Boston 2021, S. 97–129, S. 104.
[4] *Jörg van Norden/Wanda Schürenberg* (Hrsg.): Lernprogression narrativer Kompetenz im Geschichtsunterricht. Ein Vergleich von Waldorf- und Regelschule. Frankfurt am Main 2019.
[5] Ebd; *Vanessa Neumann/Wanda Schürenberg/Jörg van Norden*: Wie entwickelt sich narrative Kompetenz im Geschichtsunterricht. Eine qualitative Studie. In: Zeitschrift für Geschichtsdidaktik 15, 2016.

bildungswissenschaftlicherseits in Bezug auf das Forschungsdesign geäußert wurde, und berichten die Ergebnisse einer Reanalyse der Daten mit Hilfe eines nichtparametrischen Tests für abhängige Stichproben (Friedman-Test, vgl. Kap. 4).

2. Theoretische Grundlagen

Ziel unserer Studien ist es, die Lehre in Universität und Schule nach vorn zu bringen, indem wir die narrative Kompetenz der Lernenden, hier der Studierenden und dort der Schüler*innen fordern und fördern. Narrative Kompetenz lässt sich einerseits produktiv und andererseits rezeptiv fassen. Ihre produktive Seite umfasst die Konstruktion eigener Narrationen, konkret die Verknüpfung der drei Zeitebenen Gegenwart, Zukunft und Vergangenheit. Anlass für diese Verknüpfung ist eine aktuelle Frage, die im Rückblick auf Erfahrungen, die man selbst oder andere gemacht haben, im Blick auf eine tragfähige Zukunft beantwortet wird. Der Rückgriff auf historisches Wissen liefert keine Rezepte, sondern Denkanstöße und die Einsicht, dass der Status quo nicht, wie so gern behauptet, alternativlos ist. Die Präsenz der Vergangenheit als des Befremdlichen, das uns in der Überlieferung entgegen tritt, wird zur befreienden Alteritätserfahrung.[6] Die rezeptive Seite narrativer Kompetenz ist der historisch-kritische Umgang mit Überlieferung. Im Blick auf Empirie war es notwendig, beide Seiten zu operationalisieren und ein entsprechendes Testinstrument zu entwickeln.

Im schulischen Bereich bewährte sich eine thematisch auf den jeweiligen Unterrichtsgegenstand abgestimmte Bildcollage. Sie zielte auf die produktive Seite narrativer Kompetenz ab. das heißt, die Proband*innen schrieben einen Text zu den Bildern. Eines der Bilder des Testinstruments stammt aus der Gegenwart, die anderen bis zu fünf Bilder entsprechen verschiedenen Zeitschichten in der Vergangenheit. In der Auswertung unterschieden wir vier Teilkompetenzen, die ihrerseits in ein basales, ein intermediäres und ein elaboriertes Niveau gestuft werden. Die erste Teilkompetenz bezeichnet die Fähigkeit, einen Gegenwartsbezug herzustellen. Die Stufung folgt der Erzähltypologie Jörn Rüsens. Auf dem basalen Niveau fehlt der Bezug, auf dem intermediären wird traditional oder kritisch und auf dem elaborierten genetisch erzählt. Wer die Werte und Normen früherer Zeiten für seine Gegenwart und Zukunft übernimmt, erzählt traditional, wer sich von dem, was früher galt, absetzt, kritisch und wer einen Mittelweg versucht, genetisch Um noch einmal auf die Niveaus und den damit verbundenen kognitiven Anspruch zurückzukommen: Das basale Niveau verknüpft Gegenwart und Vergangenheit nicht, es hat nonrelationalen Charakter, das intermediäre ist dagegen relational, weil es positiv oder negativ beide Zeitschichten auf einander bezieht. Das elaborierte Niveau geht einen Schritt weiter. Multirelational sucht es eine Synthese von dem, was früher galt, und dem, was heute rechtens ist. Diese Unterscheidung in drei verschiedene kognitive Bereiche greift auch bei den drei anderen Teilkompetenzen. Die zweite von ihnen ist wie die erste

[6] *Jörg van Norden*: Verlust der Vergangenheit. Historische Erkenntnis und Materialität zwischen Wiedererkennen und Befremden (Geschichtsdidaktik theoretisch, Bd. 2). Frankfurt am Main 2022.

zeitlicher Natur, bezieht sich also ebenfalls auf das proprium historischen Denkens.[7] Es handelt sich um die Chronologie: Auf dem basalen Niveau fehlt eine solche, auf dem intermediären wird eine zeitliche Reihenfolge vergangener Ereignisse hergestellt und auf dem elaborierten werden Zeitspannen ausgewiesen, also berücksichtigt, ob eine Entwicklung lang oder kurz gedauert hat. Die Unterscheidung zwischen Chronologie und Gegenwartsbezug entspricht übrigens der B- und der A-Reihe John McTaggerts. Er betont ihren Konstruktcharakter.[8] Die übrigen beiden Teilkompetenzen, Argumentation beziehungsweise K-Reihe, und Wissen beziehungsweise W-Reihe, sind nicht zeitlich verfasst, aber für Narrationen unabdingbar. Die K-Reihe orientiert sich an dem Geschichtenmodell der Bielefelder Germanistik.[9] Auf dem basalen Niveau fehlen Begründungszusammenhänge, auf dem intermediären werden Kausalitäten benannt und auf dem elaborierten Ambivalenzen aufgezeigt. Die W-Reihe orientiert sich an der Bilderreihe, die als Testinstrument in der Schule eingesetzt worden ist und auch Teil des universitären Setting war. Auf dem basalen Niveau werden besagte Bilder reproduziert, auf dem intermediären mit Hilfe von Kontextwissen reorganisiert und auf dem elaborierten wird die Entwicklung beurteilt, die in den Bildern angelegt ist. Diese Unterscheidung orientiert sich an Karl-Ernst Jeismanns Sach- und Werturteil.[10] Die Niveaus der W-Reihe korrelieren mit der Bloomschen Taxonomie, die für die Operationalisierung im schulischen Kontext Verwendung findet.[11] Die in Zusammenhang mit den Teilkompetenzen genannten Theorieansätze verbinden sowohl ihre Relevanz für narrative Kompetenz als auch die in ihnen angelegte Möglichkeit zur Graduierung.

3.1. Narrative Kompetenz in der Schule

Die Stichprobe der Schüler*innen umfasste 450 Proband*innen aus Waldorf- und Regelschule. Den insgesamt 21 Lerngruppen wurde vor und nach einer Unterrichtsreihe, die in der Regel sechs Doppelstunden umfasste, sowie sechs Wochen später eine Bilderreihe vorgelegt mit der Bitte, dazu einen zusammenhängenden Text zu schreiben. Die Bilderreihe war von Gruppe zu Gruppe auf das jeweilige Thema der Unterrichtsreihe ausgerichtet. Unterrichtsgegenstand war je viermal das antike Rom und die Stadt im Mittelalter, fünfmal Christen und Muslime sowie je sechsmal die Französische und zweimal die Industrielle Revolution. In der Waldorfschule wurde epochal unterrichtet, in der Regelschule einerseits expositorisch und

[7] *Jörg van Norden*: Geschichte ist Zeit. Historisches Denken zwischen Kairos und Chronos - theoretisch, pragmatisch, empirisch. Berlin 2014.

[8] *John McTaggart Ellis McTaggart*: Die Irrealität der Zeit. In: *Walther C. Zimmerli/Mike Sandbothe* (Hrsg.): Klassiker der modernen Zeitphilosophie. Darmstadt 1993, S. 67–86.

[9] *Dietrich Boueke/Frieder Schülein/Hartmut Büscher/Evamaria Terhorst/Dagmar Wolf*: Wie Kinder erzählen. Untersuchungen zur Erzähltheorie und zur Entwicklung narrativer Kompetenz. München 1995.

[10] *Karl-Ernst Jeismann*: Funktion und Didaktik der Geschichte. Begründung und Beispiele eines Lehrplans für den Geschichtsunterricht. In: *Joachim Rohlfes/Karl-Ernst Jeismann* (Hrsg.): Geschichtsunterricht. Inhalte und Ziele. Stuttgart 1974, S. 106–150, S. 132.

[11] *David R. Krathwohl*: Revising Bloom's Taxonomy. In: Theory into Practice, 2002, H. 4, S. 212–218; *Frank Schweppenstette/Anja Brolle/Kirsten Impekoven/Andreas Müller/Emma Thun/Elisabeth Wagner/Ralf Saal* (Hrsg.): Abitur. Original-Prüfungsaufgaben mit Lösungen. Gymnasium-Gesamtschule NRW (STARK-Verlag-Abitur-Prüfungen). München 15. Aufl. 2021; Professionalisierung in der Lehrerausbildung Anforderungsbearbeitung und Kompetenzentwicklung im Referendariat. Leverkusen-Opladen 2014; *Ministerium für Schule und Weiterbildung, Wissenschaft und Forschung*: Sekundarstufe II Gymnasium/Gesamtschule. Richtlinien und Lehrpläne. Geschichte. Düsseldorf 1999.

andererseits exploratorisch.[12] Die Auswertung der Essays zu den Bilderreihen orientierte sich an der qualitativen Inhaltsanalyse nach Mayring und erfolgt mit Hilfe von atlas.ti.[13] Jeder Essay erhielt dabei eine Markierung für jede der vier Kategorien Gegenwartsbezug, Chronologie, Argumentation und Wissen. Dabei wurde jeweils das höchste Niveau gewählt, dass in dem Essay identifiziert werden konnte. Die Identifizierung folgte einem Kodierleitfaden, der sich aus den oben skizzierten theoretischen Prämissen ergab. Der Leitfaden wurde so lange überarbeitet, bis Interkoderreliabilität (Krippendorfs α) hergestellt war. Die Forschungsfragen richteten sich auf die kategorienspezifische Lernprogression im Allgemeinen sowie auf den Einfluss des Alters, des Unterrichtsstils und -gegenstands auf die Entwicklung der Schüler*innen über die drei Testzeitpunkte. Außerdem wurden, so problematisch eine solche binäre Zuschreibung auch ist, auf Unterschiede zwischen Jungen und Mädchen sowie im ersten Essay hoch und niedrig kodierte Proband*innen geachtet.

Die qualitative Auswertung ergab, dass alle Lerngruppen im Verlauf der Unterrichtsreihe Fortschritte gemacht haben. Am erfolgreichsten waren sie in puncto Wissen. Die routinierte Klage, Schüler*innen lernten nichts mehr im Geschichtsunterricht, ist von daher zu relativieren. Das elaborierte Niveau der W-Reihe, das Werturteil, ist allerdings von diesem positiven Befund ausgenommen. Hier war keine Progression festzustellen. Die Fähigkeit, Aussagen argumentativ zu untermauern, entwickelte sich in erfreulichem Umfang. Die beiden genuin historischen Teilkompetenzen ließen sich dagegen kaum ausbauen. Die Chronologie schnitt dabei noch geringfügig besser ab als der Gegenwartsbezug. Geschichtsunterricht hat also besonders auf die A- und die B-Reihe zu achten. Ihr relativ schlechtes Abschneiden hat sicherlich damit zu tun, das sie weder in den Curricula noch in den Schulbüchern eine Rolle spielen und damit auch im herkömmlichen Geschichtsunterricht eher die Ausnahme sind. Argumentation und Wissen werden dagegen in allen Unterrichtsfächern besonders wertgeschätzt. Vergleicht man Daten über die drei Testzeitpunkte hinweg, drängt sich der Eindruck auf, dass sich das Sachurteil (W2) auf Kosten des Werturteils wie auch der A- und B-Reihe ausgebaut hat. Ihre schwache Performanz wurde erst in T3 deutlich, während sich in T2 noch ein durchaus positiveres Bild zeigt. Was das Werturteil angeht, war es häufig in T1 stärker ausgeprägt als im weiteren Verlauf. Hier manifestiert sich möglicherweise jene Verpflichtung auf weltanschauliche Neutralität, die als Erbe eines naiven Realismus trotz des Positivismusstreits der 1970er Jahre epistemologisch immer noch nicht überwunden zu sein scheint.[14] Die im Rahmen des vorliegenden Forschungsdesigns durchgeführten Unterrichtsreihen haben alle vier Teilkompetenzen gefördert, waren aber im Ganzen gesehen offensichtlich nur ein Tropfen auf dem heißen Stein. Das mag erklären, warum sich in T3 vieles verflüchtigt hatte, was in T2 noch Anlass zur Freude gab. Der jeweilige Unterrichtsstil hatte nur geringen Einfluss auf die Lernprogression. Es lässt sich allerdings vermuten, dass das Werturteil vom Epochalunterricht der Waldorfschule und den

[12] *Michael Zech*: Ideen und Praxis des Geschichtsunterrichts an den Waldorfschulen. In: *Jörg van Norden/Wanda Schürenberg* (Hrsg.): Lernprogression narrativer Kompetenz im Geschichtsunterricht. Ein Vergleich von Waldorf- und Regelschule. Frankfurt am Main 2019, S. 27–46; *Jörg van Norden*: Lob eines narrativen Konstruktivismus. In: Geschichte in Wissenschaft und Unterricht 60, 2009, H. 12, S. 734–741.

[13] *Philipp Mayring*: Qualitative Inhaltsanalyse: Grundlagen und Techniken. Weinheim Basel 12. Aufl. 2015.

[14] *Jörg van Norden*: Geschichte ist Einstellungssache. In: zeitschrift für didaktik der gesellschaftswissenschaften, 2012, H. 1, S. 54–75.

explorativ angelegten Unterricht der Regelschule stärker profitiert hat als von den lehrer*innenzentrierten Unterrichtsreihen. Was die Unterrichtsgegenstände betrifft, war „Stadt im Mittelalter“ förderlicher als „Industrialisierung“ und „Französische Revolution“. Das Alter der jeweiligen Proband*innen spielte keine Rolle. „Rom“ und „Christen und Muslime“ regten gleichermaßen das genetische Erzählen (A3) besonders an, obwohl letzterer Gegenstand mit älteren Schüler*innen erarbeitet wurde und sich der Gegenwartsbezug hier eingängiger darstellen lässt. Insgesamt gesehen wäre darüber nachzudenken, sich vom chronologischen Verfahren alter Schule zu verabschieden und stattdessen die Gegenstände altersgerecht und problemorientiert auszuwählen. Entsprechende Längsschnitte folgen nicht chronologisch auf einander, sind es aber in sich. Sie arbeiten durchaus mit einem Zeitlineal. Dieses Medium, das zeigen die Daten deutlich, fördert die Entwicklung der A- und der B-Reihe. Das binär zugeschriebene Geschlecht erwies sich als irrelevant für die Lernprogression. Was den Vergleich der im ersten Essay niedrig und hoch kodierten Proband*innen angeht, machten erstere deutlichere Lernfortschritte, auch wenn sie das elaborierte Niveau nur in seltenen Fällen erreicht haben. Letztere waren in diesem Punkt etwas erfolgreicher, ohne dass sie alles in allem besser abgeschnitten hätten. Auf ihre Förderung muss also geachtet werden.

Die skizzierten Ergebnisse zeigen nicht, wie Schüler*innen historisch denken. Lediglich die Performanz ihres Denkens kann erhoben werden. Sie wird von vielen unterschiedlichen Faktoren beeinflusst. Dazu zählen situationsbezogene Testmüdigkeit und Leistungsschwäche wie auch die jeweilige Lehrperson, Einflüsse außerschulischer Geschichtskultur und vieles mehr. Real existierender Geschichtsunterricht, in den sich empirische Wirksamkeitsforschung hinein begibt, bleibt ein Quasiexperiment mit allen damit verbundenen Konsequenzen für die Tragfähigkeit ihrer Aussagen. Immerhin kann das vorliegende Forschungsdesign Denkanstöße für die Pragmatik von Unterricht geben. Das war unser Ziel, um der Validität willen darauf verzichten zu müssen, keine Option.[15]

3.2 Narrative Kompetenz bei Studierenden

Die in Kapitel 3.1 vorgestellte Studie wurde als Anlass genommen, Fragen nach den entsprechenden Kompetenzen auch bei angehenden Lehrkräften zu stellen beziehungsweise danach, ob sich das in der Schule gezeigte Bild bis ins Studium angehender Geschichtslehrkräfte hält und bis zum Ende des Masters ein Lernfortschritt festzustellen ist. Da in der Bielefelder Lehramtsausbildung das Praxissemester mit seinem Fokus auf Forschendes Lernen als wesentliches Moment in der Kompetenzentwicklung gesehen wird, bot es sich an, besonders diese Phase, inklusive der jeweils vor- und nachbereitenden sowie begleitenden Veranstaltungen (Vorbereitungs- und Begleitseminare), zu untersuchen.[16] Die universitäre Studie übernahm

[15] *van Norden/Schürenberg*: Lernprogression narrativer Kompetenz im Geschichtsunterricht (Anm. 2).

[16] Siehe hierzu das Leitkonzept der Universität Bielefeld von 2011; verfügbar unter: http://www.bised.uni-bielefeld.de/praxisstudien/praxissemester/fo_le/bielefelder_ausgestaltung/Bielefelder_Leitkonzept/praxisstudien/praxisseme ster/fo_le/bielefelder_ausgestaltung/leitkonzept.pdf. Ferner *Renate Schüssler/Anke Schöning*: Forschendes Lernen im Praxissemester – Potential und Ausgestaltungsmöglichkeiten. In: *Renate Schüssler/Anke Schöning/Volker Schwier/Saskia Schicht/Johanna Gold/Ulrike Weyland* (Hrsg.): Forschendes Lernen im Praxissemester – Zugänge, Konzepte, Erfahrungen. Bad

die Bilderreihe, konkret die zur Industriellen Revolution, und ergänzte sie um ein Dilemmageschichte, die von den Studierenden bewältigt werden sollte: Mit einem Mörder konfrontiert, der um Hilfe bittet, kann man sich an ein uraltes Familienversprechen gebunden fühlen und den Bittsteller der Justiz entziehen, also traditional erzählen, das Versprechen zugunsten der aktuellen Rechtsnormen verwerfen und die Polizei verständigen, also kritisch erzählen, oder genetisch beides tun, indem man vor Gericht für den besten Rechtsbeistand sorgt.[17] Die produktive Seite narrativer Kompetenz wurde also jetzt im Testinstrument mehrfach abgedeckt. Die Auswertung folgte den selben Kategorien wie in der ersten Studie. Die zweite Studie berücksichtigte allerdings auch die rezeptive Seite, indem entsprechende offene Fragen in das Testinstrument mit aufgenommen wurden.[18] Die Studierenden sollten diesen Aufgabenkatalog jeweils zu Beginn (t1) und am Ende (t2) des Vorbereitungsseminars sowie nach dem Praxissemester (t3) bearbeiten.[19]

Diese Untersuchung wurde erstmalig vom Winter- 2016/17 bis zum Wintersemester 2017/18 durchgeführt (= Kontrollgruppe). Danach erfolgte eine Reflexion des Forschungsdesigns und der bisherigen Befunde. Als Konsequenz wurden mit dem Sommersemester 2018 zunächst pilotierend, dann ab dem Wintersemester 2018/19 explizite Inhalte in den Seminaren implementiert, um den Studierenden Möglichkeiten anzubieten, sich mit Perspektivität und Konstruktion von Narrationen sowie Quellen und Historiographie zu beschäftigen und diese Aspekte im Seminar zu reflektieren. Diese *Optimierung* sollte den Studierenden die wesentlichen konstruktivistischen Merkmale von Geschichte verdeutlichen, die im Bachelor eingeführt worden waren. Die Erwartung war, dass die Studierenden dadurch eine erkennbare Kompetenzentwicklung zeigen würden.[20] Die Untersuchung wurde folglich ein zweites Mal vom Winter- 2018/19 bis zum Wintersemester 2019/20 durchgeführt (= Experimentalgruppe).

Im Folgenden wird – mit Fokus auf den vergleichenden Blick zwischen Schüler*innen und Studierenden – die Auswertung lediglich des dritten Aufgabenteils, der Bilderreihe, beschrieben, denn hier ist das Vorgehen grundsätzlich mit dem in Kapitel 3.1 beschriebenen vergleichbar. [21] Die Befunde ähneln in vielerlei Hinsicht den bereits in der Schule generierten: Sowohl in der Kontroll- als auch in der Experimentalgruppe erzählten mehr als die Hälfte der Studierenden ohne einen Gegenwartsbezug, was sich auch über die drei Testzeitpunkte nicht änderte. Offensichtlich sahen die meisten Studierenden nicht die Notwendigkeit, den historischen Sachverhalt der Bilderreihe auf seine Bedeutung für die Gegenwart hin (A-Niveau 3) zu

Heilbrunn 2017, S. 39-50; *Anke Schöning*: Das Bielefelder Leitkonzept zum Forschenden Lernen im Praxissemester. In: PFLB – PraxisForschungLehrer*innenbildung 2019, H. 1.2, S. 10–17; https://doi. org/10.4119/pflb-1966

[17] *van Norden*: Was machst du für Geschichten (Anm. 3).

[18] *Jörg van Norden*: Students and Their „Idea of History" – A Theory Based Testing of Hermeneutical and Narrative Competences. In: *Friederike Neumann/Leah Shopkow* (Hrsg.): Teaching History, Learning History, Promoting History. Papers from the Bielefeld Conference on Teaching History in Higher Education. Frankfurt am Main 2018, S. 163–192;

[19] *Thomas Must*: Kompetenzentwicklung durch Forschendes Lernen? Überlegungen zur Funktion von Studienprojekten im Fach Geschichte anhand empirischer Befunde. In: HLZ – Herausforderung Lehrer*innenbildung 2018, H. 2.1, S. 299–314; https://doi. org/10.4119/hlz-2407 sowie *Jörg van Norden/Thomas Must*: Im Praxissemester historisch denken lernen? In: *Sebastian Barsch/Oliver Plessow* (Hrsg.): Universitäre Praxisphasen im Fach Geschichte – Wege zur Verbesserung der Lehramtsausbildung? Band 4 (Hochschulpädagogik). Berlin 2020, S. 195-217.

[20] *Thomas Must*: Kompetenzentwicklung im Praxissemester. Anspruch und Wirklichkeit im Fach Geschichte im empirischen Vergleich. In: PFLB - PraxisForschungLehrer*innenBildung 2020, Heft 2.1, S. 64-82. doi:10.4119/PFLB-3611

[21] Die Auswertung der aller Aufgabenteile wurde bereits an anderen Stellen publiziert, siehe etwa Anm. 19 und 20.

berücksichtigen. Ähnliches ist für die B-Reihe festzustellen: Die rein chronologische Darstellung (B–Niveau 2) überwiegt in beiden Gruppen und bleibt nahezu unverändert. Größere Entwicklungen zeichneten sich allerdings beim Umgang mit Wissen (W-Reihe) ab. Zu t2 hin gab es eine merkliche Zunahme von Sach- (W-Niveau 2) und Werturteilen (W-Niveau 3) auf Kosten des basalen Niveaus. Die Experimentalgruppe schnitt dabei erkennbar besser ab und entwickelte sich bis in t3 positiv, allerdings stärker zum Sach- als zum Werturteil hin. Was die Argumentation betrifft, zeigen sich die Studierenden über alle Testzeitpunkte mehrheitlich stabil auf dem Niveau 2, das heißt, sie begründen in der Regel ihre Aussagen. Kritischere Aussagen, die Ambivalenzen aufzeigen (K-Niveau 3), sind hingegen nur selten zu finden, wohl aber in der Kontrollgruppe. Im Vergleich der beiden Gruppen ist hinsichtlich der narrativen Kompetenz kein wesentlicher Unterschied festzustellen und auch die jeweilige Entwicklung ist bis auf geringe Ausnahmen nicht als signifikant auszuweisen. Nichtsdestotrotz fallen diese kaum ins Gewicht, so dass – auch unter Berücksichtigung der schon oben in Kapitel 3.1 benannten möglichen Störfaktoren – fraglich bleibt, ob das Praxissemester und das Forschende Lernen oder die Implementierung der oben skizzierte Seminarinhalten Wirkung auf die für diese Studien angenommenen Kompetenzbereiche zeitigen.[22] Unabhängig von der Lernprogression ist aber ebenfalls festzuhalten, dass lediglich die Perfomanz in der A-Reihe bedenklich ausfällt. Denn in der B- und W-Reihe zeigten die Studierenden bereits durchaus zufriedenstellende Qualitäten, wenn auch nur eingeschränkt auf dem Niveau 3.

4. Analysen zur Lernprogression in narrativer Kompetenz

Im Folgenden werden die obigen Ergebnisse der Analysen zur Lernprogression in der narrativen Kompetenz für die vier Teilkompetenzen Gegenwartsbezug (A-Reihe), Chronologie (B-Reihe), Argumentation (K-Reihe) und Wissen (W-Reihe) auf Basis einer Neuberechnung berichtet. Nach einer Darstellung der Analysemethoden werden zunächst ausgewählte Ergebnisse der Schüler*innen-Testung (Kapitel 4.2) und anschließend die der Studierenden-Testung (Kapitel 4.3) berichtet. Obgleich die beiden Proband*innen-Gruppen sich hinsichtlich verschiedener Merkmale unterscheiden (Alter, Bildungsniveau, Vorwissen, Thema bzw. Gegenstand, Testdurchführung und didaktisches Setting), werden die Befunde abschließend in einer vergleichenden Perspektive diskutiert (Kapitel 4.4).

4.1 Darstellung der Analysemethoden

Für die Analyse der Lernprogression wurden nicht-parametrische Testverfahren angewendet, die dem ordinalen Charakter der Daten gerecht werden und keine normalverteilten Daten erfordern. Insbesondere kam hier der Friedman-Test zu Anwendung, der ein nicht-parametrisches Äquivalent zu einer Varianzanalyse mit Messwiederholung darstellt. Anders als in der

[22] Anders sieht es bei der hermeneutischen Kompetenz aus. Hier besteht die Vermutung, dass die Implementierungen in den Seminaren zumindest geringfügig zu einer positiven Kompetenzentwicklung beitragen konnte: siehe *Must*: Kompetenzentwicklung im Praxissemester (Anm. 20).

Varianzanalyse kann im Rahmen des Friedman-Tests jedoch kein mehrfaktorielles Modell spezifiziert werden, sodass sich die Analysen auf ein einfaktorielles Modell beschränken. Damit ist es nicht möglich, potentielle Interaktionseffekte zwischen den verschiedenen Einflussvariablen zu untersuchen, ebenso wenig können Zwischensubjekteffekte ermittelt werden.
Für die Haupteffekte wurden Post-hoc-Tests und die Effektstärke ‚r' berechnet, wobei abweichend zu Cohen's d die folgenden Intervalle gelten: unter .10 = kein Effekt, .10 bis .20 = kleiner Effekt, .21 bis .39 = mittlerer Effekt, ab .40 = großer Effekt.[23] Zusätzlich wurde mit Hilfe von Kruskal-Wallis -H-Tests für mehr als zwei unabhängige Stichproben geprüft, ob es signifikante Gruppenunterschiede im Ausgangsniveau (t1) sowie zu den zwei folgenden Messzeitpunkten (t2 und t3) gibt.
Bei der nachfolgenden grafischen Darstellung der Ergebnisse werden aufgrund des begrenzten Wertebereiches (1 bis 3) die mittleren Ränge ausgewiesen (y-Achse), da sich damit die Unterschiede beziehungsweise die Entwicklung über die drei Messzeitpunkte anschaulicher darstellen lassen als mit Median-Box-Plots. Es ist wichtig darauf hinzuweisen, dass die abgebildeten Werte nicht den Messwerten selbst entsprechen, da die eingesetzten Testverfahren auf den ihnen zugewiesenen Rängen, das heißt auf der Ordnung der Messwerte (höher als, niedriger als) basieren und nicht auf den absoluten Abständen zwischen den Werten. Entsprechend der drei Messzeitpunkte können den Messwerten Ränge 1 bis 3 zugewiesen werden. Das bedeutet, dass ein mittlerer Rang von „2" nicht zwangsläufig dem intermediären Niveau der jeweiligen Teilkompetenz entspricht, sondern dem durchschnittlichen Rang der jeweiligen Messwerte innerhalb der drei Messzeitpunkte.[24]

4.2 Ausgewählte Ergebnisse der Schüler*innen-Testung

Im Folgenden werden ausgewählte Ergebnisse der Schüler*innen-Testung berichtet. Dabei stehen Vergleiche auf Jahrgangsebene im Fokus. Hierbei ist zu beachten, dass die 6. und 7. Jahrgänge nur Gymnasialklassen umfassen und die Jahrgänge 8, 9 und 11 nur aus Waldorfklassen bestehen. Damit sind die potentiellen Effekte nicht ausschließlich auf die Jahrgangszugehörigkeit beziehungsweise das Alter zurück zu führen, sondern diese sind konfundiert mit der Schulform, dem Unterrichtsstil und dem Unterrichtsgegenstand, sodass letztendlich nicht entschieden werden kann, welche dieser Variablen für die Unterschiede in der Lernprogression verantwortlich sind.[25] Hierfür wären mehrfaktorielle Auswertungsstrategien von Nöten, die für nicht-parametrische Testverfahren nicht implementiert sind (s.o.).

---

[23] Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. Advances in Methods and Practices in Psychological Science, 2, 156–168. doi:10.1177/2515245919847202
[24] Dabei erhält der niedrigste Wert innerhalb der Messreihe (t1 bis t3) den Rang 1 und der höchste den Wert 3. Kommt innerhalb der Testreihe ein Wert mehrfach vor, werden sogenannte verbundene Ränge gebildet, d.h. es wird ein Mittelwert aus den einzelnen Rängen gebildet. Zur Berechnung der mittleren Ränge werden über alle Messwerte eines Messzeitpunktes zunächst Rangsummen ermittelt und anschließend durch die Anzahl der Messwerte geteilt.
[25] Jürgen Bortz und Nicola Döring (2006): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Heidelberg, 4. Auflage, S. 524 -526.

Abbildung 1 zeigt zunächst die Lernprogression der fünf untersuchten Jahrgänge in der Teilkompetenz „Gegenwartsbezug“. Es zeigen sich zum Teil deutliche Unterschiede in der Lernprogression, insbesondere zwischen dem ersten und dem zweiten Messzeitpunkt.

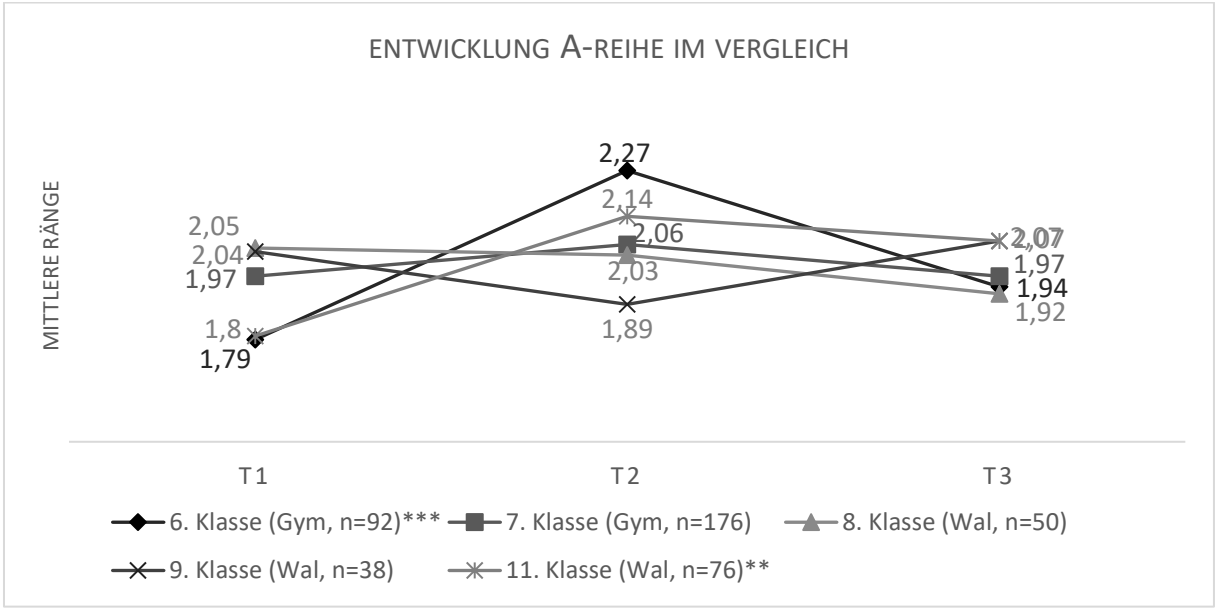

Abbildung 1: Ergebnisse nicht -parametrischer Varianzanalyse mit Messwiederholung nach Friedman für die A-Reihe. *Anmerkung:* *** = p ≤ .001, ** = p ≤ .01, * = p ≤ .05

Während in den 7., 8. und 9. Klassen keine signifikante Veränderung festzustellen ist, zeigt sich vor allem bei den 6. Klassen und bei den 11. Klassen ein signifikanter Effekt (6. Jg.: $\chi^2(2) = 21{,}912$, $p \leq 0.001$; 11. Jg.: $\chi^2(2) = 12{,}211$, $p = .002$). Post-hoc-Tests mit Bonferroni-Holm-Korrektur weisen für die 6. Klassen einen signifikanten Zuwachs zwischen t1 und t2 nach ($z_{t1\text{-}t2} = -3{,}207$, $p_{angepasst} = .003$, Effektstärke $r = .334$), der anschließende Abfall zwischen t2 und t3 verfehlt knapp die Signifikanzgrenze ($z_{t3\text{-}t2} = 2{,}212$, $p_{angepasst} = .054$). Es lässt sich dennoch vermuten, dass der Zuwachs in dieser Teilkompetenz in den getesteten 6. Klassen nicht sehr nachhaltig war.

Zwar zeigt sich für die 11. Klassen ein signifikanter Haupteffekt, doch fallen die Post-hoc-Analysen nicht signifikant aus, auch wenn sich in der Abbildung zwischen t1 und t2 ein sichtbarer Zuwachs erkennen lässt. Anders als bei den 6. Klassen zeigt sich hier kein merklicher Abfall zu t3.

Zu t3 zeigt sich in allen Jahrgängen ein ähnliches Niveau der Teilkompetenz „Gegenwartsbezug“. Dies lässt sich auch an den ähnlich hohen mittleren Rängen in Tabelle 1 ablesen.

Auffällig ist hier das vergleichsweise niedrige Ausgangsniveau (t1) der 11. Klassen in der A-Reihe, das sich insbesondere gegenüber dem mittleren Niveau in den 7. und 8. Klassen als signifikant erweist (vgl. Tab. 1), die Effektstärken liegen hier im mittleren Bereich. Daraus ließe sich folgern, dass diese Teilkompetenz nicht vom Vorwissen beziehungsweise Bildungsniveau abhängig ist.

Auch zum zweiten Messzeitpunkt (t2) unterscheiden sich die Jahrgänge hinsichtlich des durchschnittlichen Kompetenzniveaus. Post-hoc-Analysen weisen hier vor allem die Differenz zwischen dem 6. und 9. Jahrgang als signifikant aus (vgl. Tab. 1). In diesen beiden Jahrgängen zeigt sich zudem eine gegenläufige Tendenz zwischen dem ersten und dem zweiten Messzeitpunkt

(vgl. Abb. 1), das heißt während die Schüler*innen der 6. Klassen hier einen deutlichen Zuwachs aufweisen, nimmt das durchschnittliche Niveau im 9. Jahrgang zum zweiten Messzeitpunkt etwas ab.

Tabelle 1: Ergebnisse des Kruskal-Wallis-H-Tests für die A-Reihe, inkl. Post-hoc-Tests

		Md	Mittlere Ränge	df	χ^2	p	Sign. Unterschiede (Effektstärke r)
t1	6. Jahrgang	1	202,01				–
	7. Jahrgang	1	233,20				11. Jg.***(.304)
	8. Jahrgang	1,5	247,94	4	21,906	≤ .001	11. Jg.*** (.334)
	9. Jahrgang	1	216,34				–
	11. Jahrgang	1	177,39				–
t2	6. Jahrgang	2	249,28				9. Jg. ***(.333)
	7. Jahrgang	1	216,18				–
	8. Jahrgang	1	219,46	4	0,618	.002	–
	9. Jahrgang	1	167,54				–
	11. Jahrgang	1	202,95				–
t3	6. Jahrgang	1	219,43				–
	7. Jahrgang	1	218,91				–
	8. Jahrgang	1	214,96	4	0,272	.976	–
	9. Jahrgang	1	211,43				–
	11. Jahrgang	1	210,93				–

Weitere Analysen (nicht tabellarisch ausgewiesen) zeigen einen signifikanten Haupteffekt für den exploratorischen Unterrichtsstil ($\chi^2(2) = 21.691$, $p \leq .001$), der sich vor allem im stärkeren Kompetenzzuwachs zwischen t1 und t2 manifestiert, wie die Post-hoc-Analysen zeigen ($z_{t1\text{-}t2} = -2{,}724$, $p = .018$, $r = .221$). Das Geschlecht spielt für die Lernprogression in der Teilkompetenz „Gegenwartsbezug“ keine Rolle.

In Abbildung 2 wird die jahrgangsspezifische Entwicklung in der B-Reihe dargestellt, der Teilkompetenz „Chronologie“. Erneut zeigt sich hier vor allem für die 6. und 11. Klassen ein signifikanter Haupteffekt (6. Jg.: $\chi^2(2) = 31{,}484$, $p \leq 0.001$; 11. Jg.: $\chi^2(2) = 23{,}612$, $p = .002$).

Post-hoc-Analysen zeigen, dass in beiden Jahrgängen die Differenz zwischen t1 und t2 signifikant ausfällt, wobei es sich beim 6. Jahrgang um einen starken Effekt handelt (6. Jg.: $z_{t1\text{-}t2} = -3{,}981$, $p_{angepasst} \leq .001$, $r = .415$; 11. Jg.: $z_{t1\text{-}t2} = -2{,}596$, $p_{angepasst} = .027$, $r = .298$). Allerdings zeigt sich in beiden Jahrgängen auch ein deutlicher Abfall zwischen t2 und t3, der allerdings nur beim 11. Jahrgang signifikant ausfällt (6. Jg.: $z_{t3\text{-}t2} = 2{,}212$, $p_{angepasst} = .054$; 11. Jg.: $z_{t3\text{-}t2} = 2{,}275$, $p_{angepasst} = .045$, $r = .261$). Dies ließe sich ebenfalls so interpretieren, dass der Kompetenzzuwachs über die beobachtete Zeitspanne hinweg nicht von Dauer war.

Im Gegensatz dazu fällt auf, dass im 9. Jahrgang ein kontinuierlicher Anstieg zwischen t1 und t3 zu erkennen ist, selbiger erweist sich jedoch nicht als signifikant ($\chi^2(2) = 3{,}396$, $p = .183$).

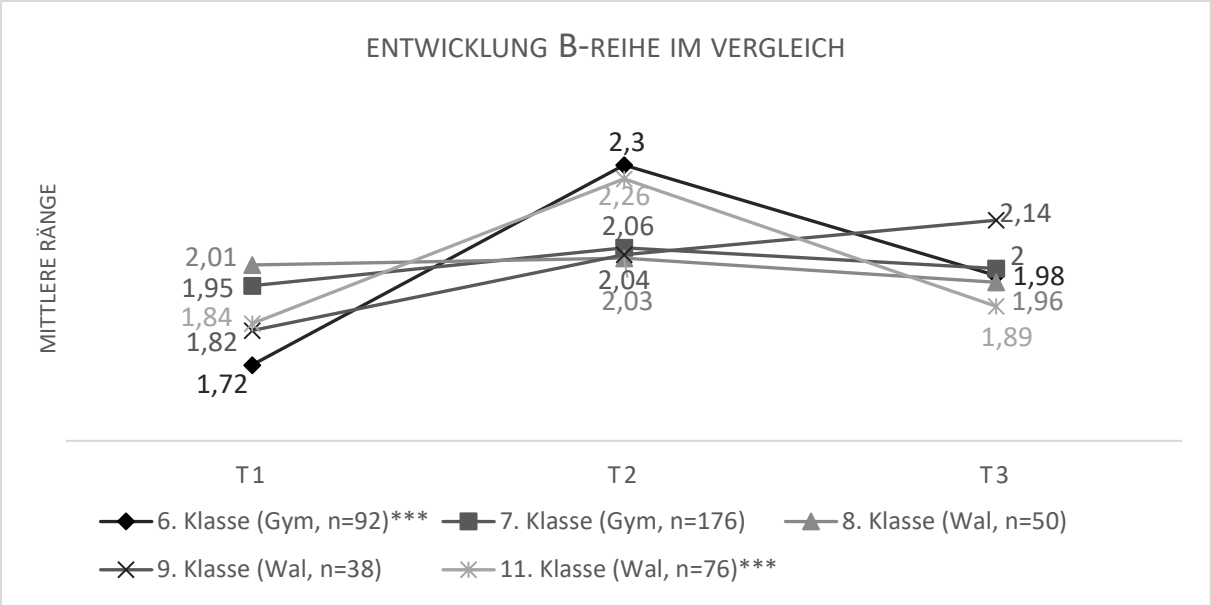

Abbildung 2: Ergebnisse nicht-parametrischer Varianzanalyse mit Messwiederholung nach Friedman für die B-Reihe. *Anmerkung:* *** = p ≤ .001, ** = p ≤ .01, * = p ≤ .05

In Tabelle 2 sind die Unterschiede zwischen den Jahrgängen hinsichtlich des Kompetenzniveaus zu den drei Messzeitpunkten dargestellt. Diese erweisen sich zu allen drei Messzeitpunkten als signifikant. Bereits an den mittleren Rängen lässt sich ablesen, dass der 11. Jahrgang zu allen Messzeitpunkten die niedrigsten Werte aufweist, die sich im Rahmen der Post-hoc-Analysen auch als signifikant gegenüber den anderen Jahrgängen erweisen. Das macht deutlich, dass die in Abbildung 2 sichtbare Entwicklung zwischen t1 und t2 auf einem vergleichsweise niedrigen Niveau stattfindet, das heißt trotz dieses Zuwachses kommt die Mehrheit der Schüler*innen des 11. Jahrgangs nicht über das basale Niveau hinaus. Auch der 6. Jahrgang weist zum ersten und dritten Zeitpunkt relativ niedrige mittlere Werte auf, zum zweiten Messzeitpunkt unterscheidet er sich jedoch nicht mehr von den übrigen Jahrgängen (mit Ausnahme des 11. Jahrgangs). Die Effektstärken liegen größtenteils im mittleren bis hohen Bereich. Insgesamt ist festzuhalten, dass die Schüler*innen der „mittleren Jahrgänge“ (7. bis 9.) durchgehend das intermediäre Niveau aufweisen.

Tabelle 2: Ergebnisse des Kruskal-Wallis-H-Tests für die B-Reihe, inkl. Post-hoc-Tests

		Md	Mittlere Ränge	df	χ^2	p	Sign. Unterschiede (Effektstärke r)
t1	6. Jahrgang	1	171,99				–
	7. Jahrgang	2	247,89				6.***(.324); 11.***(.370)
	8. Jahrgang	2	269,05	4	61,484	≤ .001	6.***(.416); 11.***(.438)
	9. Jahrgang	2	231,43				6.*(.243); 11.**(.310)
	11. Jahrgang	1	158,09				–
t2	6. Jahrgang	2	223,15				11.*(.229)
	7. Jahrgang	2	222,82				11.**(.210)
	8. Jahrgang	2	235,79	4	15,548	.004	11.*(.253)
	9. Jahrgang	2	233,37				11.*(.261)
	11. Jahrgang	1	172,68				–
t3	6. Jahrgang	1	194,43	4	48,212	≤ .001	11.*(.194)
	7. Jahrgang	2	236,06				6.*(.176); 11.***(.349)

8. Jahrgang	2	253,26	6.*(.250); 11.***(.401)
9. Jahrgang	2	267,89	6.**(.297); 11.***(.491)
11. Jahrgang	1	150,63	–

Weitere Analysen (nicht tabellarisch ausgewiesen) zeigen, dass der exploratorische Unterrichtsstil vor allem auf den Kompetenzzuwachs zwischen t1 und t2 einen Effekt hat, wie die Post-hoc-Analysen zeigen ($z_{t1\text{-}t2}$ = -2,667, $p_{angepasst}$ = .024, r = .216). Die Geschlechtszugehörigkeit erweist sich nicht als relevant für die Lernprogression in dieser Teilkompetenz.

Abbildung 3 stellt die Entwicklung in der K-Reihe, das heißt der Argumentationskompetenz, dar. Auch hier zeigt sich für die 6. Klassen ein signifikanter Haupteffekt ($\chi^2(2)$ = 26,057, p ≤ 0.001), der auf die Unterschiede zwischen dem ersten und den nachfolgenden Messzeitpunkten zurückzuführen ist ($z_{t1\text{-}t2}$ = -3,575, $p_{angepasst}$ ≤ .001, r = .373; $z_{t1\text{-}t3}$ = -3,059, $p_{angepasst}$ = .004, r = .319). Ein Kompetenzzuwachs lässt sich auch für die 7. Klassen feststellen, insbesondere zwischen dem ersten und dem zweiten Messzeitpunkt, der allerdings schwächer ausfällt (Haupteffekt: $\chi^2(2)$ = 10,437, p = 0.005; $z_{t1\text{-}t2}$ = -2,418, $p_{angepasst}$ = .048, r = .182). Auch im 9. Jahrgang zeigt sich ein signifikanter Haupteffekt ($\chi^2(2)$ = 7,357, p = 0.025), dieser ist jedoch vornehmlich auf den Anstieg zwischen t2 und t3 zurückzuführen, der allerdings das Signifikanzniveau knapp verfehlt ($z_{t2\text{-}t3}$ = -2,294, $p_{angepasst}$ = .066). Grundsätzlich zeigt sich im 9. Jahrgang eine im Vergleich zu den anderen Jahrgängen gegenläufige Entwicklung bezüglich der Teilkompetenz „Argumentation“: Während das durchschnittliche Kompetenzniveau der anderen Jahrgänge zu t2 ansteigt und zu t3 wieder leicht sinkt, fällt es beim 9. Jahrgang zunächst und steigt zum dritten Zeitpunkt deutlich an.

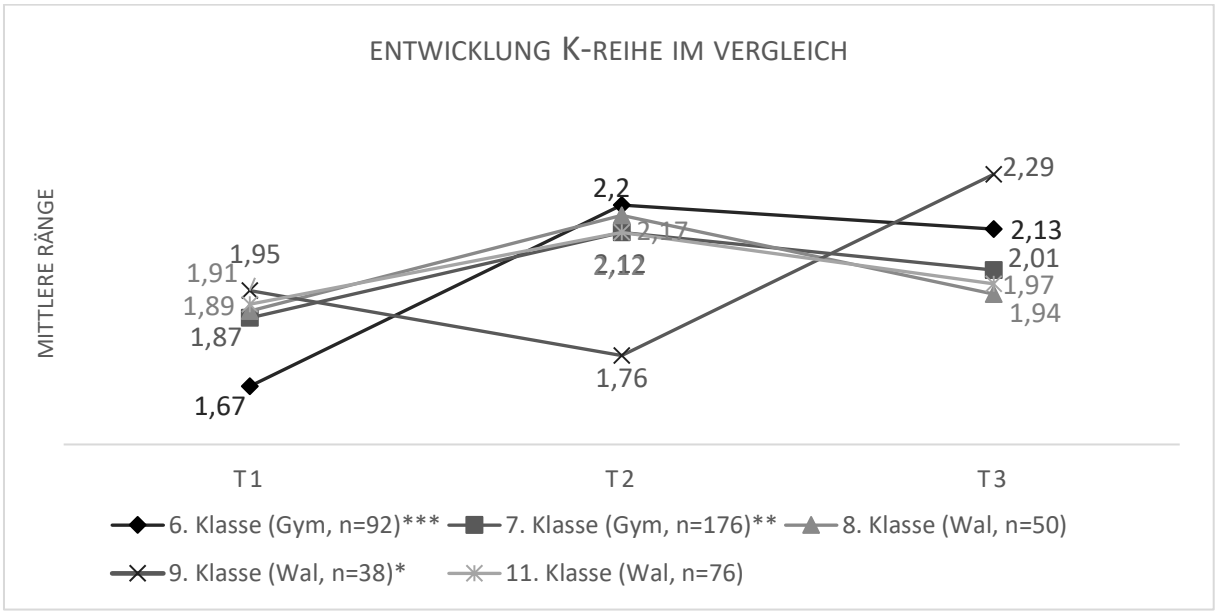

Abbildung 3. Ergebnisse nicht-parametrischer Varianzanalyse mit Messwiederholung nach Friedman für die K-Reihe*. Anmerkung:* *** = p ≤ .001, ** = p ≤ .01, * = p ≤ .05

In Tabelle 3 sind die Unterschiede zwischen den Jahrgangsgruppen im Hinblick auf die Kompetenzniveaus zu den drei Messzeitpunkten dargestellt.

Es lässt sich feststellen, dass die Schüler*innen zum ersten Messzeitpunkt über ein vergleichbares Kompetenzniveau verfügen, die Mehrheit bewegt sich hier auf dem intermediären

Niveau, dies ändert sich auch zum zweiten Messzeitunkt nicht wesentlich, mit Ausnahme des 9. Jahrgangs. Diese Differenz verfehlt jedoch knapp das Signifikanzniveau.
Zum dritten Messzeitpunkt weist der 9. Jahrgang ein signifikant höheres Kompetenzniveau auf, die Mehrheit der Schüler*innen erreicht das elaborierte Niveau. Hierin unterscheidet sich dieser Jahrgang insbesondere von dem 7., 8. und 11. Jahrgang, wobei die Effektstärken im mittleren Bereich liegen.

Tabelle 3: Ergebnisse des Kruskal-Wallis-H-Tests für die K-Reihe, inkl. Post-hoc-Tests

		Md	Mittlere Ränge	df	χ^2	p	Sign. Unterschiede (Effektstärke r)
t1	6. Jahrgang	2	218,13				–
	7. Jahrgang	2	212,68				–
	8. Jahrgang	2	212,54	4	1,590	.811	–
	9. Jahrgang	2	238,38				–
	11. Jahrgang	2	217,93				–
t2	6. Jahrgang	2	247,62				–
	7. Jahrgang	2	210,69				–
	8. Jahrgang	2	215,82	4	9,289	.054	–
	9. Jahrgang	2	188,86				–
	11. Jahrgang	2	209,48				–
t3	6. Jahrgang	2	245,06				–
	7. Jahrgang	2	203,26				6.*(.169)
	8. Jahrgang	2	192,45	4	19,888	≤ .001	–
	9. Jahrgang	3	273,97				7. **(.230); 8.**(.343);
	11. Jahrgang	2	202,69				9.*(.286)

Weitere Analysen zeigen einen mittelstarken Effekt des instruierten Unterrichtsstils auf die Lernprogression in der K-Reihe ($\chi^2(2) = 35{,}852$, $p \leq 0.001$), was bedeuten könnte, dass Schüler*innen diesbezüglich von einem lenkenden beziehungsweise lehrer*innenzentrierten Unterricht stärker und nachhaltiger ($p_{t1\text{-}t2} \leq .001$, $r = .423$; $p_{t1\text{-}t3} = .003$, $r = .284$) profitieren. Des Weiteren zeigt sich auch ein Effekt des Geschlechts ($\chi^2(2) = 33{,}490$, $p \leq 0.001$): Schülerinnen starten zwar von einem signifikant niedrigeren Niveau aus als Schüler, überholen diese aber zum zweiten Messzeitpunkt, das heißt ihre Lernkurve ist steiler als die der Schüler. Zum dritten Zeitpunkt zeigen sich keine Niveau-Unterschiede mehr.

Schließlich ist die Teilkompetenz „Wissen“ (W-Reihe) zu betrachten. Abbildung 4 ist zu entnehmen, dass sich insbesondere für die 6., 7. und 11. Klassen ein signifikanter Haupteffekt zeigt (6. Jg.: $\chi^2(2) = 43{,}565$, $p \leq 0.001$; 7. Jg.: $\chi^2(2) = 45{,}640$, p p ≤ 0.001; 11. Jg.: $\chi^2(2) = 26{,}470$, $p \leq 0.001$). Im Gegensatz dazu weisen die 8. und 9. Klassen über die drei Messzeitpunkte keine signifikante Veränderung ihres Kompetenzniveaus auf.

Gemäß der Post-hoc-Analysen erweist sich vor allem der Anstieg zwischen dem ersten und dem zweiten Messzeitpunkt als signifikant, wobei die Effektstärken im mittleren Bereich liegen (6. Jg.: $z_{t1\text{-}t2} = -3.354$, $p_{angepasst} = .003$, $r = .350$; 7. Jg.: $z_{t1\text{-}t2} = -4.690$, $p_{angepasst} \leq .001$, $r = .354$; 11. Jg.: $z_{t1\text{-}t2} = -3.244$, $p_{angepasst} = .004$, $r = .231$). Während das Niveau zum dritten Messzeitpunkt beim 6. und 11. Jahrgang stabil bleibt und damit signifikant über dem Ausgangniveau liegt,

sinkt es beim 7. Jahrgang wieder beinahe auf das Ausgangsniveau ($z_{t3\text{-}t2}$ = 3,065, $p_{angepasst}$ = .004, r = .231).

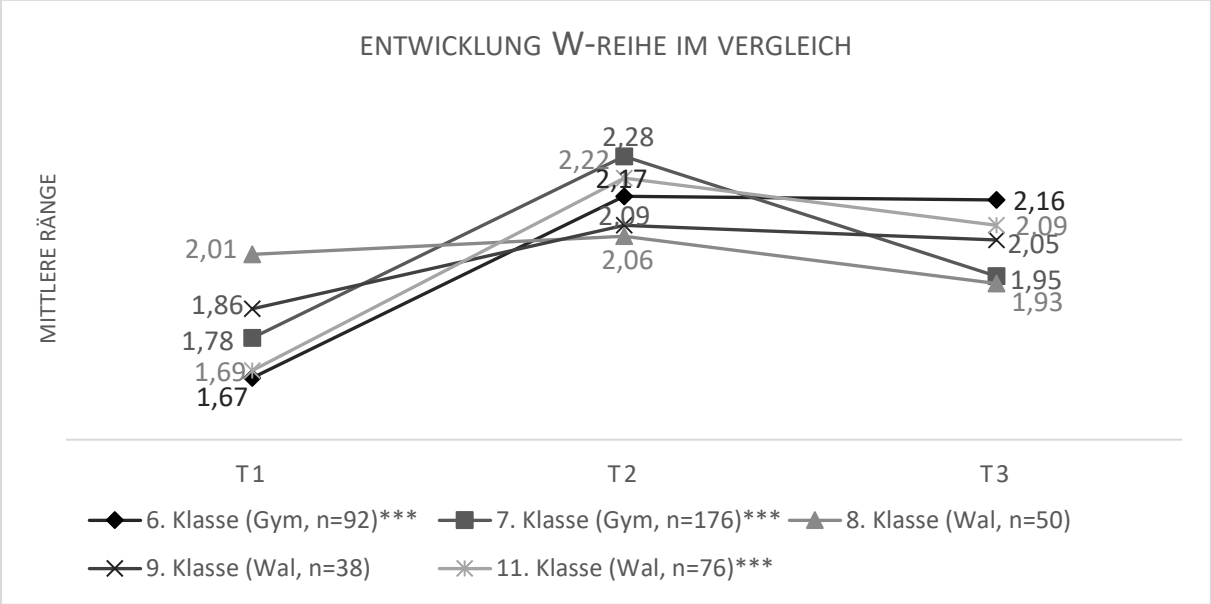

Abbildung 4: Ergebnisse nicht-parametrischer Varianzanalyse mit Messwiederholung nach Friedman für die W-Reihe. *Anmerkung:* *** = p ≤ .001, ** = p ≤ .01, * = p ≤ .05

Gruppenvergleiche mittels des Kruskal-Wallis-H-Tests zeigen (Tab. 4), dass die Schüler*innen der 8. und 9. Klassen auf einem vergleichsweise hohen Ausgangniveau starten und sich damit insbesondere von den Schüler*innen der 6. und 7. Klassen signifikant unterscheiden. Beim Blick auf die nachfolgenden Messzeitpunkte wird zudem deutlich, dass die 9. Klassen konstant die höchsten mittleren Ränge aufweisen, insofern relativiert sich der oben berichtete Befunde über die ausbleibende Entwicklung in der Teilkompetenz „Wissen“. Vor allem im Vergleich zu den 7. Klassen zeigt sich ein signifikanter Unterschied mittlerer Effektstärke zum dritten Messzeitpunkt: Im Gegensatz zu den Schüler*innen der 7. Klassen bleibt das durchschnittliche Niveau in der Teilkompetenz „Wissen“ bei den Schüler*innen der 9. Klassen hoch.

Tabelle 4: Ergebnisse des Kruskal-Wallis-H-Tests für die W-Reihe, inkl. Post-hoc-Test

		Md	Mittlere Ränge	df	χ^2	p	Sign. Unterschiede (Effektstärke r)
t1	6. Jahrgang	2	202,76				8.**(.267); 9.**(.290)
	7. Jahrgang	2	195,77				8.***(.259); 9.***(.268)
	8. Jahrgang	2	263,56	4	27,696	≤ .001	–
	9. Jahrgang	2	272,16				–
	11. Jahrgang	2	225,47				–
t2	6. Jahrgang	2	206,23				
	7. Jahrgang	2	206,11				
	8. Jahrgang	2	214,65	4	14,125	.007	*Post-hoc-Tests n.s.*
	9. Jahrgang	2	250,43				
	11. Jahrgang	2	240,22				
t3	6. Jahrgang	2	233,82				7. ***(.235)
	7. Jahrgang	2	184,72				–
	8. Jahrgang	2	218,71	4	37,119	≤ .001	–
	9. Jahrgang	2	272,50				7. ***(.338)
	11. Jahrgang	2	239,68				7. ***(.254)

Weitere Analysen weisen keine Effekte für die Geschlechtszugehörigkeit aus und auch im Hinblick auf den Unterrichtsstil zeigen sich keine eindeutigen Effekte für die Lernprogression in der W-Reihe.

Abschließend ist zu ergänzen, dass über die gesamte Schüler*innenstichprobe betrachtet in allen vier Teilkompetenzen ein Anstieg zum zweiten Messzeitpunkt zu verzeichnen ist, wobei der Effekt bei der Teilkompetenz Wissen (W-Reihe) erwartungsgemäß am höchsten ausfällt ($r_{t1\text{-}t2}$ =.303). Am schwächsten fällt er für die Teilkompetenz „Gegenwartsbezug“ aus (A-Reihe, $r_{t1\text{-}t2}$ =.130). Zum dritten Messzeitpunkt sinkt das Kompetenzniveau in allen vier Teilbereichen wieder und bewegt sich für die A- und B-Reihe auf dem Ausgangsniveau, während es bei der K- und W-Reihe ($r_{t1\text{-}t3}$ =.142 und $r_{t1\text{-}t3}$ =.178) signifikant höher als zum ersten Zeitpunkt ausfällt. Der Kompetenzzuwachs scheint also in Bezug auf Argumentation und Wissen etwas nachhaltiger zu sein. Mit Ausnahme der B-Reihe (Chronologie) zeigen sich für die Gymnasien etwas stärkere positive Effekte im Hinblick auf die Lernprogression als für die Waldorfschulen, insbesondere in der Teilkompetenz „Wissen“ (Waldorf: $r_{t1\text{-}t2}$ =.222; Gymnasien: $r_{t1\text{-}t2}$ =.352).

4.3 Ausgewählte Ergebnisse der Studierenden-Testung

Im Rahmen der folgenden Darstellung werden die in Kapitel 3.2 vorgestellten Studierendengruppen der Kontroll- und der Experimentalgruppe hinsichtlich der Lernprogression miteinander verglichen. Zusätzlich wird mit der Gruppe der Lehramtsstudierenden für die Primarstufe mit dem Schwerpunktfach Sachunterricht eine dritte Studierendengruppe berücksichtigt (nachfolgend SU-Gruppe). Da der Sachunterricht ebenfalls historisches Lernen in einer Teilperspektive berücksichtigt, müssten von den Studierenden ähnliche Kompetenzbereiche erwartet werden. Daher wurde eine Gruppe zumindest zur Erkundung und als zusätzliche Vergleichsgruppe herangezogen. Diese Studierendengruppe (n=24) wurde im WS 2016/17 befragt und ist daher hinsichtlich der Testbedingungen beziehungsweise Testaufgaben mit der Kontrollgruppe vergleichbar.

Abbildung 5 zeigt zunächst die Lernprogression der drei Gruppen in der A-Reihe, d.h. der Teilkompetenz „Gegenwartsbezug“.

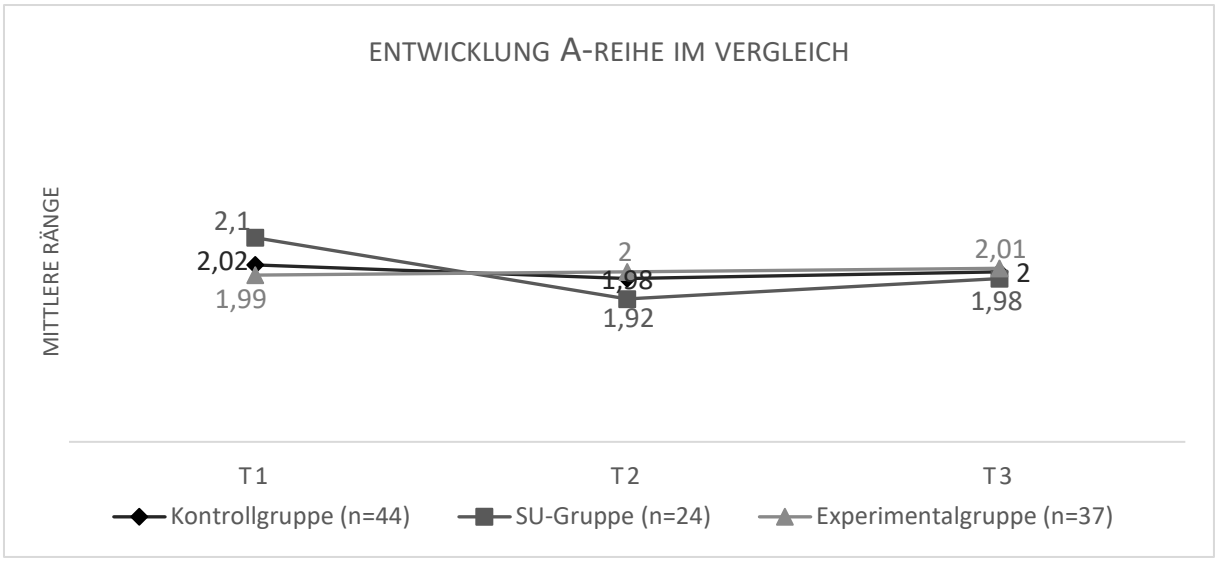

Abbildung 5: Ergebnisse nicht -parametrischer Varianzanalyse mit Messwiederholung nach Friedman für die A-Reihe

Für keine der drei Gruppen zeigt sich eine signifikante Veränderung in der Teilkompetenz „Gegenwartsbezug“ (Kontrollgruppe: $\chi^2(2) = 0.094$, $p = .954$; SU-Gruppe: $\chi^2(2) = 1{,}077$, $p = .584$; Experimentalgruppe: $\chi^2(2) = 0{,}024$, $p = .998$). Ebenso wenig unterscheiden sich die Gruppen signifikant hinsichtlich der Kompetenzniveaus zu den drei Erhebungszeitpunkten. Die entsprechenden Kennwerte und Ergebnisse des Kruskal-Wallis-Tests sind Tabelle 5 zu entnehmen.

Tabelle 5: Ergebnisse des Kruskal-Wallis-H-Tests für die A-Reihe

		Md	Mittlere Ränge	df	χ^2	p
t1	Kontrollgruppe	1	53,45			
	SU-Gruppe	1,5	53,50	2	0,059	.971
	Experimentalgruppe	1	52,14			
t2	Kontrollgruppe	1	53,55			
	SU-Gruppe	1	49,38	2	0,618	.734
	Experimentalgruppe	1	54,70			
t3	Kontrollgruppe	1	53,78			
	SU-Gruppe	1	50,50	2	0,272	.873
	Experimentalgruppe	1	53,69			

Demzufolge spiegeln die Ergebnisse der Berechnungen die bereits in Kapitel 3.2 beschriebenen Befunde wider: In allen drei Gruppen kommt die Mehrheit der Studierenden nicht über das basale Kompetenzniveau hinaus, dies gilt für alle drei Erhebungszeitpunkte.

Die nachfolgende Abbildung 6 stellt die Entwicklung in der B-Reihe dar. Wie bereits bei der A-Reihe, zeigen sich auch im Hinblick auf die Teilkompetenz „Chronologie“ für keine der drei betrachteten Studierendengruppen signifikante Veränderungen über die Zeit (Kontrollgruppe: $\chi^2(2) = 1.537$, $p = .464$; SU-Gruppe: $\chi^2(2) = 2{,}324$, $p = .313$; Experimentalgruppe: $\chi^2(2) = 0{,}304$, $p = .859$). Insbesondere die Experimentalgruppe bleibt weitgehend auf dem gleichen Niveau, während sich in der Kontroll- und SU-Gruppe eine geringfügige positive Entwicklung zwischen t1 und t2 abzeichnet.

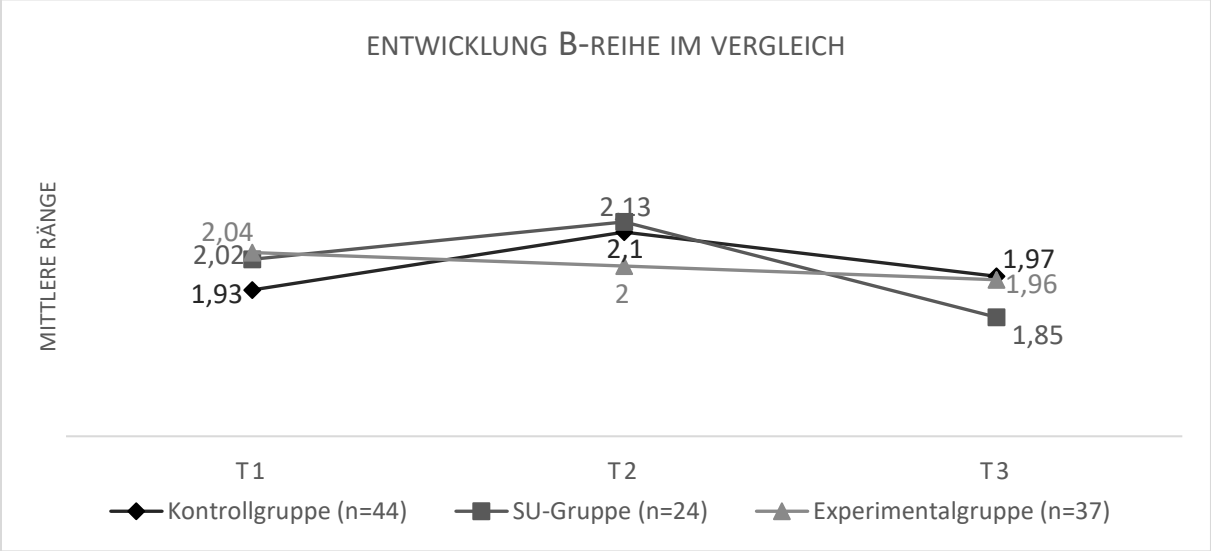

Abbildung 6: Ergebnisse nicht -parametrischer Varianzanalyse mit Messwiederholung nach Friedman für die B-Reihe

Wie auch bei der A-Reihe unterscheiden sich die Gruppen auch hier nicht signifikant hinsichtlich der Kompetenzniveaus zu den drei Erhebungszeitpunkten. Abweichend zur A-Reihe liegt das mittlere Niveau durchgängig auf Kompetenzstufe 2, das heißt die Studierenden sind mehrheitlich in der Lage eine zeitliche Reihenfolge herzustellen (vgl. Abschnitt 3.2). Die entsprechenden Kennwerte und Ergebnisse des Kruskal-Wallis-Tests sind Tabelle 6 zu entnehmen.

Tabelle 6: Ergebnisse des Kruskal-Wallis-H-Tests für die B-Reihe

		Md	Mittlere Ränge	df	χ^2	p
t1	Kontrollgruppe	2	53,14			
	SU-Gruppe	2	50,33	2	0,431	.806
	Experimentalgruppe	2	54,57			
t2	Kontrollgruppe	2	55,33			
	SU-Gruppe	2	49,79	2	1,043	.594
	Experimentalgruppe	2	52,31			
t3	Kontrollgruppe	2	54,80			
	SU-Gruppe	2	46,46	2	2,189	.335
	Experimentalgruppe	2	55,11			

In Bezug auf die Teilkompetenz „Argumentation“ (K-Reihe) zeigt sich ein etwas anderes Bild (Abb. 7).

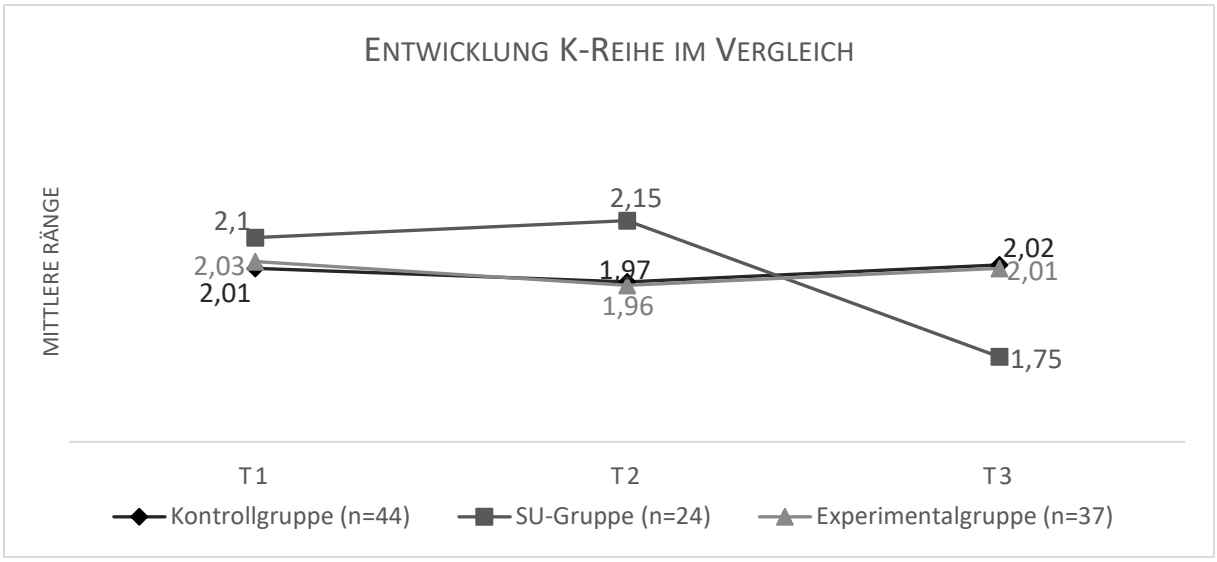

Abbildung 7: Ergebnisse nicht -parametrischer Varianzanalyse mit Messwiederholung nach Friedman für die K-Reihe

Während bei der Kontroll- und die Experimentalgruppe keine Veränderung über die Zeit festzustellen ist (Kontrollgruppe: $\chi^2(2) = 0.136$, $p = .934$; Experimentalgruppe: $\chi^2(2) = 0{,}192$ $p = .909$), zeichnet sich bei der SU-Gruppe ein deutlicher Abfall zu t3 ab, der jedoch die Signifikanzgrenze von $p < .05$ knapp verfehlt ($\chi^2(2) = 5.070$, $p = .079$), was vermutlich auf die kleine Stichprobengröße zurückzuführen ist.

Auch im Rahmen der Gruppenvergleiche mittels des Kruskal-Wallis-Tests (Tab. 7) zeigt sich, dass das mittlere Kompetenzniveau der SU-Studierenden zum dritten Erhebungszeitpunkt von der intermediären auf die basale Stufe fällt. Hierin unterscheiden sich diese Studierenden auch signifikant ($p = .017$) von den übrigen beiden Gruppen. Grundsätzlich lässt sich anhand der mittleren Ränge ablesen, dass die SU-Gruppe über alle Messzeitpunkte hinweg ein etwas niedrigeres Niveau aufweist als die Kontroll- und die Experimentalgruppe.

Tabelle 7: Ergebnisse des Kruskal-Wallis-H-Tests für die K-Reihe

		Md	Mittlere Ränge	df	χ^2	p
t1	Kontrollgruppe	2	53,65			
	SU-Gruppe	2	45,90	2	2,312	.315
	Experimentalgruppe	2	56,84			
t2	Kontrollgruppe	2	53,39			
	SU-Gruppe	2	48,46	2	0,916	.633
	Experimentalgruppe	2	55,49			
t3	Kontrollgruppe	2	55,16			
	SU-Gruppe	1	39,21	2	8,093	.017
	Experimentalgruppe	2	59,38			

Insgesamt ist die Mehrheit die Studierenden also über die gesamte Testreihe in der Lage, kausal zu argumentieren, eine Entwicklung in Richtung eines elaborierten Niveaus, bei dem auch Ambivalenzen aufgezeigt werden, zeichnet sich nicht ab.

Schließlich zeigt Abbildung 8 die Entwicklung in der W-Reihe, das heißt der Teilkompetenz, die den Umgang mit Wissen adressiert.

Wie bereits bei den anderen Teilkompetenzen zeigen sich auch für die W-Reihe keine bedeutsamen Veränderungen über die Zeit, wobei sich tendenziell für die Experimentalgruppe ein positiver Trend feststellen lässt (Kontrollgruppe: $\chi^2(2) = 2.456$, $p = .293$; SU-Gruppe: $\chi^2(2) = 0{,}040$, $p = .980$; Experimentalgruppe: $\chi^2(2) = 4{,}766$, $p = .092$).

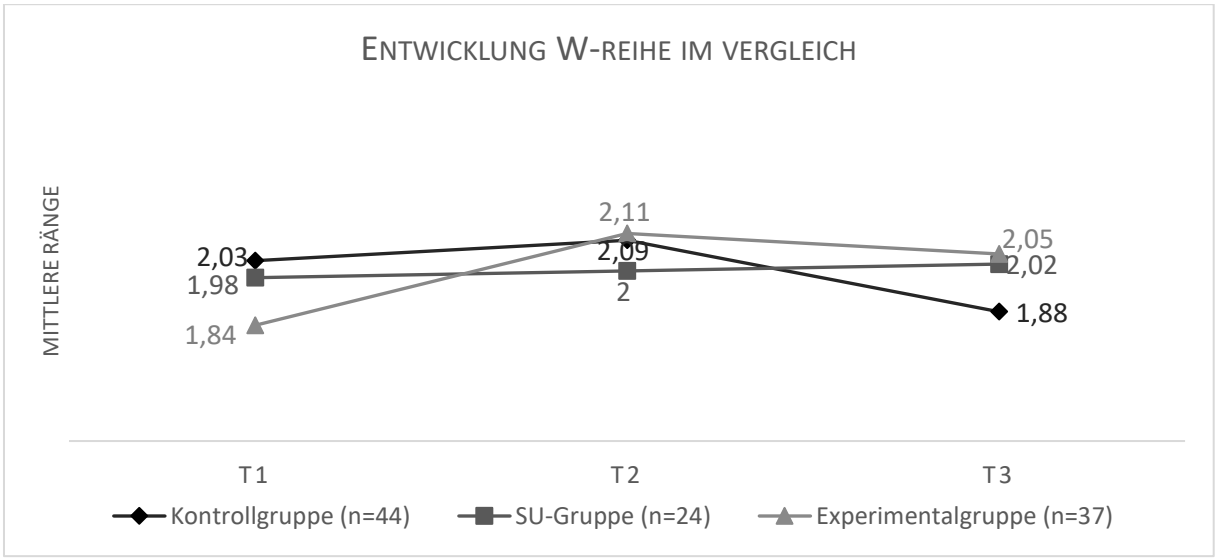

Abbildung 8: Ergebnisse der nicht -parametrischer Varianzanalyse mit Messwiederholung nach Friedman für die W-Reihe

Tabelle 8 ist zu entnehmen, dass in allen drei Gruppen durchgängig über die drei Messzeitpunkte das intermediäre Niveau (Ebene des Sachurteils) dominiert. Die an den mittleren Rängen ablesbaren Niveauunterschiede zwischen den drei Gruppen pro Messzeitpunkt erweisen sich nicht als signifikant, lediglich die Experimentalgruppe schneidet zu t3 tendenziell besser ab als die beiden anderen Gruppen ab ($p = .092$).

Tabelle 8: Ergebnisse des Kruskal-Wallis-H-Tests für die W-Reihe

		Md	Mittlere Ränge	df	χ^2	p
t1	Kontrollgruppe	2	54,78			
	SU-Gruppe	2	50,10	2	0,587	.746
	Experimentalgruppe	2	52,76			
t2	Kontrollgruppe	2	53,23			
	SU-Gruppe	2	45,27	2	3,907	.142
	Experimentalgruppe	2	57,74			
t3	Kontrollgruppe	2	49,18			
	SU-Gruppe	2	48,77	2	4,778	.092
	Experimentalgruppe	2	60,28			

Ingesamt gesehen können die Analysen die bereits oben in Kapitel 3.2 beschriebenen Ergebnisse stützen.

4.3 Vergleichende Betrachtung von Schüler*innen und Studierenden

Die oben referierten Befunde zeigen auf, dass die Lernprogression in den vier Teilkompetenzen im schulischen Setting eher zu gelingen scheint als im universitären Kontext beziehungsweise im Kontext des Praxissemesters. Dies ist insofern plausibel, als dass im schulischen Kontext gegenstands-, das heißt themenbezogen, gearbeitet wird, während es im Kontext der universitären Vorbereitung auf das Praxissemester um überfachliche, das heißt themenunabhängige Aspekte geht. Dementsprechend können die Ergebnisse der Testungen im schulischen Kontext auch als Lerneffekte eines inhaltsbezogenen didaktischen Settings interpretiert werden, was insbesondere auf die wissensbezogenen Teilkompetenz (W-Reihe) zutreffen dürfte. Andererseits ist es ebenso plausibel, dass die Studierenden über andere Voraussetzungen verfügen (zum Beispiel Vorwissen, schriftsprachliche Kompetenzen et cetera), so dass sie bereits auf einem höheren Niveau „starten“ und die Lernprogression allein aufgrund dessen „nach oben hin“ limitiert ist.

Um diese Annahme zu überprüfen, wurde ein Mann-Whitney-U-Test für zwei unabhängige Stichproben berechnet. Die Ergebnisse des MWU-Tests sind in Abbildung 9 dargestellt. Dabei markieren die schraffierten Balken signifikante Unterschiede zwischen der Gruppe der Schüler*innen (n = 433) und der Studierenden (n = 105).

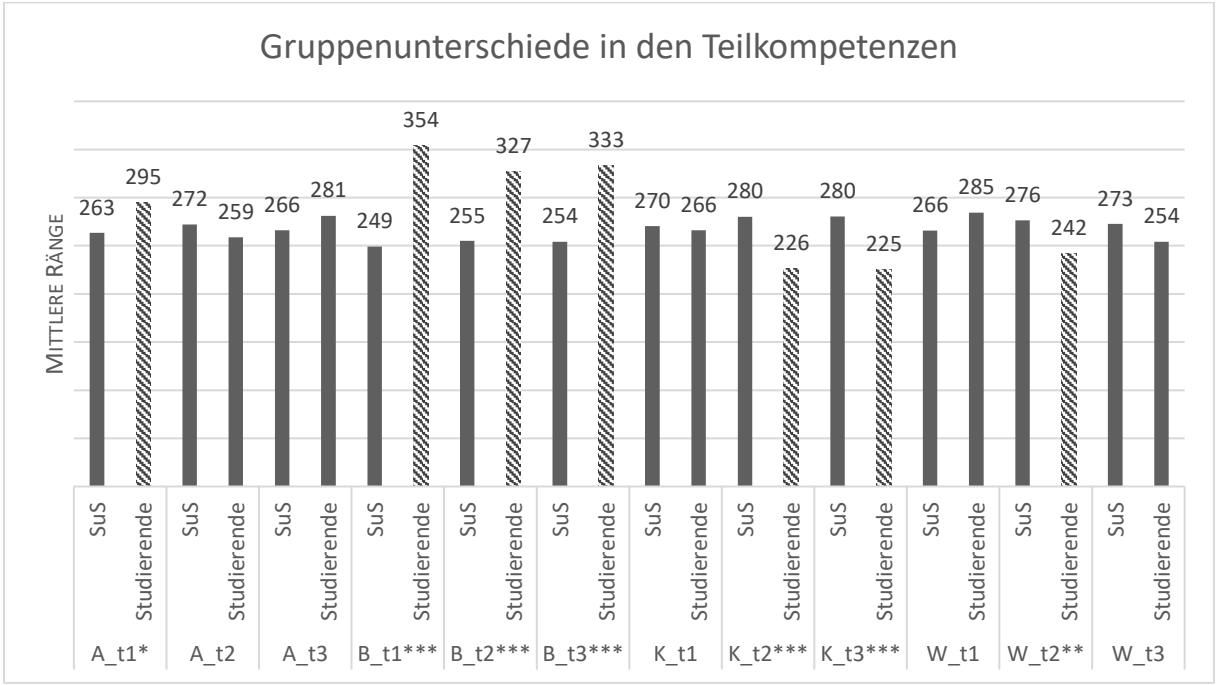

Abbildung 9: Ergebnisse des Mann-Whitney-U-Tests für zwei unabhängige Stichproben
Anmerkung: *** = p ≤ .001, ** = p ≤ .01, * = p ≤ .05

Es zeigen sich größere Unterschiede in der Teilkompetenz Chronologie (B-Reihe) in der erwarteten Richtung: die Gruppe der Studierenden verfügt bereits zum ersten Messzeitpunkt über ein höheres Kompetenzniveau als die Schüler*innen und behält diesen „Vorsprung“ über die nachfolgenden Messzeitpunkte. Die Effektgrößen bewegen sich hierbei im mittleren Bereich (r zwischen .213 und .297).

Auch in der Teilkompetenz „Gegenwartsbezug“ (A-Reihe) zeigt sich ein leichter Anfangsvorteil zugunsten der Studierenden, es handelt sich hierbei jedoch um einen zu vernachlässigenden Effekt (r = .096).
Eher erwartungswidrig verhält es sich bei der Teilkompetenz „Argumentation“ (K-Reihe). Bereits beim ersten Messzeitpunkt zeigt sich hier kein Kompetenzvorteil auf Seiten der Studierenden, zum zweiten und dritten Messzeitpunkt weisen die Schüler*innen ein signifikant höheres Niveaus auf, allerdings handelt es sich hier erneut um eher geringe Effekte (r < .21).
Die Ergebnisse der W-Reihe fallen erwartungsgemäß aus: während zum ersten Messzeitpunkt nur ein geringfügiger Unterschied zwischen Schüler*innen und Studierenden besteht, zeigt sich insbesondere zum zweiten Messzeitpunkt ein signifikanter Vorsprung der Schüler*innen, wenn auch nur in geringer Effektstärke (r= . 120). Hier lassen sich demnach die oben angesprochenen gegenstandsbezogenen Lerneffekte vermuten. Unterschiede in ähnlicher Größenordnung zeigen sich, wenn man die Gruppe der Studierenden mit den 8. Klassen vergleicht (nicht ausgewiesen), die bei der Testung ebenfalls das Thema „Industrialisierung“ bearbeitet haben. Hier zeigt sich allerdings bereits zu t1 ein signifikanter, wenn auch schwacher Unterschiede in der W-Reihe zugunsten der Schüler*innen (r = .161).

5. Ausblick

Von gewissen Nuancen abgesehen, zeigen beide Studien, dass in der Performanz historischen Denkens sowohl im Unterricht als auch in der Lehrer*innenbildung der Universität Bielefeld der Umgang mit Gegenwartsbezügen und die Bewertung von Vergangenheit oft nur eine geringe Rolle spielen. Vergangenes verbleibt offenbar als etwas faszinierend Fernes ohne Bedeutung für das Hier und Jetzt. Auch die statistische Neuberechnung stützt diesen Befund und verleiht den Studien zusätzliche Aussagekraft. Doch was bedeutet das für Geschichtsunterricht und die Ausbildung angehender Geschichtslehrkräfte? Offensichtlich muss der Gegenwartsbezug künftig stärker berücksichtigt werden und das nicht durch lediglich punktuelle Interventionen. Sie werden kaum längerfristige Wirkung zeitigen können. Vergleichbares gilt für die Bewertungsebene. Auch sie wird nur dann gefördert, wenn vor allem Fragen verhandelt werden, die solche Werturteile herausfordern. Beiden Aspekten trägt der problemorientierte Gegenwartsbezug Rechnung. Er ist eine geschichtsdidaktische Notwendigkeit. Sowohl Schüler*innen als auch Studierenden (im Fach Geschichte) muss bewusstwerden, dass die Gegenwart Beginn und Ziel kritisch-reflektierten historischen Denkens ist, Geschichte als Konstrukt wahrzunehmen und in seiner Orientierungsfunktion zu nutzen. Historisches Lernen in der Schule muss daher – um auch langjährigen geschichtsdidaktischen Forderungen eines zielführenden historischen Lernens konsequenter nachzukommen – auf diesen Bedarf ausgelegt und Lernprozesse so geplant werden, dass die Beschäftigung mit vergangenen Phänomenen stets aus einem gegenwärtigen Orientierungsbedürfnis resultiert und dorthin zurückkehrt. Als Konsequenz für das Studium und das Praxissemester bedeutet dies, dass die Studierenden einerseits intensivere Gelegenheiten zur Planung entsprechender historischer Lernprozesse

erhalten und andererseits im Rahmen Forschenden Lernens im Praktikum gezielter für die Bedeutung problemorientierter Gegenwartsbezüge sensibilisiert werden müssten.