Holger Thünemann,
Meik Zülsdorf-Kersting
(Hrsg.)

Methoden geschichtsdidaktischer Unterrichtsforschung

WOCHEN SCHAU GESCHICHTE

Holger Thünemann,
Meik Zülsdorf-Kersting (Hrsg.)

Methoden geschichtsdidaktischer Unterrichtsforschung

WOCHEN SCHAU GESCHICHTE

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://dnb.d-nb.de abrufbar.

www.wochenschau-verlag.de

Die Reihe „Geschichtsunterricht erforschen“
wird herausgegeben von
Monika Fenn
Peter Gautschi
Johannes Meyer-Hamme
Holger Thünemann
Meik Zülsdorf-Kersting

Umschlaggestaltung: Ohl Design
Gedruckt auf chlorfrei gebleichtem Papier
Gesamtherstellung: Wochenschau Verlag
ISBN 978-3-7344-0212-8 (Buch)
ISBN 978-3-7344-0213-5 (E-Book)

Inhalt

Vorwort

Geschichtsunterrichtsforschung ist wieder zu einem intensiv bearbeiteten Forschungsfeld der deutschsprachigen Geschichtsdidaktik geworden. Das ist wichtig für die Wissenschaftsdisziplin „Didaktik der Geschichte" und deren Bedeutung als Referenz für die Geschichtslehrerausbildung und Unterrichtspraxis. Die auf den Geschichtsunterricht bezogenen Studien der letzten zehn Jahre (z.B. von Borries u. a. 2005; Zülsdorf-Kersting 2007; Gautschi u. a. 2007; Gautschi 2009; Meyer-Hamme 2009; Martens 2010; Lange 2011; Meyer-Hamme/Thünemann/Zülsdorf-Kersting 2012; Köster 2013; Hodel u. a. 2013; Spieß 2014; Waldis u. a. 2015; Mathis 2015) bedienten sich unterschiedlicher Methoden, um Befunde zum Geschichtsunterricht bzw. zum historischen Denken von Schülerinnen und Schülern zu generieren. Dies ist eine wichtige Entwicklung, deren Fortsetzung unbedingt zu wünschen ist, um das komplexe Phänomen „Geschichtsunterricht" besser zu verstehen.

Es ist zugleich eine bemerkenswerte Entwicklung, weil die Ausbildung im Bereich der qualitativen, quantitativen wie triangulierten Forschungsmethoden immer noch kein genuiner Bestandteil des Geschichtsstudiums ist. Im Unterschied zu anderen Studienfächern wie der Psychologie, Soziologie oder auch der Erziehungswissenschaft dominieren im Geschichtsstudium weiterhin hermeneutische Methoden und seit dem *visual turn* zunehmend auch die Auseinandersetzung mit bildlichen Quellen und Darstellungen. Das Interesse an empirisch ausgerichteten Abschluss- und Qualifikationsarbeiten mag groß sein; die Realisierung steht aber nicht selten vor hohen methodischen Hürden. Die Einübung dieses Hürdenlaufs braucht Zeit, die die Studierenden in der finalen Phase ihres Studiums meist nicht haben. Nicht selten muss man möglichen Kandidatinnen und Kandidaten von interessanten Forschungsfragen abraten, weil die Defizite im Bereich der empirischen Methodenkompetenz zu groß sind.

Der Band „Methoden geschichtsdidaktischer Unterrichtsforschung" basiert auf der Methodenexpertise geschichtsdidaktischer (oder geschichtsdidaktisch interessierter) Empirikerinnen und Empiriker und verfolgt das Ziel, die weitere Geschichtsunterrichtsforschung zu befördern. Der vorliegende Band richtet sich an eine geschichtsdidaktische Leserschaft, die sich mit unterschiedlichen Methoden zur Erforschung des Geschichtsunterrichts vertraut machen möchte. Die Lektüre der Texte eignet sich für eine Erstbe-

gegnung mit i. d. R. sehr elaborierten Forschungsmethoden. Die einzelnen Beiträge können auch in der Ausbildung von Geschichtslehrerinnen und -lehrern Platz finden, da der Zugang zu den Forschungsmethoden über das Erkenntnisinteresse am Geschichtsunterricht erfolgt. Der vorliegende Band möchte keine weitere allgemeine Einführung in die Methoden der qualitativen Sozialforschung sein. Vielmehr sollen diese Methoden hinsichtlich ihres Potenzials zur Erforschung von Geschichtsunterricht und mithin in engem Bezug auf diesen geschichtsdidaktisch profiliert werden.

Die einzelnen Beiträge dieses Bandes folgen daher einem ähnlichen Aufbau – Vorstellung der Methode, geschichtsdidaktische Profilierung und exemplarische Anwendung. In einem *ersten* Schritt sollen die jeweilige Methoden wissenschaftstheoretisch verortet und die z.T. komplexen theoretischen Prämissen vorgestellt werden. In diesem methodologischen Teil erfolgen z. B. auch wissenschaftsgeschichtliche Ausführungen zur Herkunft und Entwicklung der betreffenden Methode. Da die Relevanz einzelner Methoden nur in Bezug auf bestimmte Forschungsfragen zu bestimmen ist, sollen die Methoden in einem *zweiten* Schritt als Reaktion auf spezifisch geschichtsdidaktische Forschungsfragen profiliert werden. Manche Methoden eignen sich primär, um die Prozesshaftigkeit von Geschichtsunterricht zu erforschen; andere bieten sich vor allem zur Erforschung der Wirkung von Geschichtsunterricht an. Der Konnex zwischen geschichtsdidaktischem Forschungsinteresse und empirischer Methode wird hier ausformuliert und begründet. Dabei wird deutlich, dass die Methoden nicht um ihrer selbst willen vorgestellt werden, sondern hinsichtlich ihrer Funktion für geschichtsdidaktische Forschungsanliegen. In einem *dritten* Schritt soll die exemplarische Anwendung die gerade angesprochene Funktionalität, aber auch die Grenzen der Methode demonstrieren. Die Beiträge sollen die jeweiligen Methoden somit auch durch die Analyse empirischer Daten aus dem Kontext der Geschichtsunterrichtsforschung (Unterrichtsvideo, Transkript, Schülertexte, Lehrerinterview etc.) gewissermaßen *in actu* vorstellen.

Manuel Köster unternimmt in seinem einleitenden Beitrag eine grundlegende Systematisierung des Tableaus unterschiedlicher Erhebungs- und Auswertungsmethoden im Hinblick auf deren Potenzial zur Erforschung von Geschichtsunterricht. Im Unterschied zu den dann folgenden Beiträgen ist der Fokus nicht auf eine einzelne Methode gerichtet, sondern auf die Vielfalt methodischer Ansätze. Die Beiträge von *Sebastian Barsch* (Qualitative Inhaltsanalyse), *Christiane Bertram* (Fragebogenkonstruktion), *Matthias Martens/Christian Spieß/Barbara Asbrand* (rekonstruktive Analyse von Unterrichtsvideos), *Christian Mehr* (Objektive Hermeneutik), *Johannes Meyer-Hamme* (Analyse von Fragebögen), *Doren Prinz/Holger Thünemann* (Mixed

Methods) und *Monika Waldis* (quantitative Analyse von Unterrichtsvideos) widmen sich dann in dem beschriebenen Dreischritt einzelnen Methodensettings der Geschichtsunterrichtsforschung.

Die meisten Beiträge gehen auf drei Tagungen des KGD-Arbeitskreises „Empirische Geschichtsunterrichtsforschung" aus den Jahren 2012 bis 2014 zurück. Wir danken den Autorinnen und Autoren dieses Bandes sehr für ihre Bereitschaft, sich auf das oben skizzierte Anliegen einzulassen, wie auch für ihr Engagement im genannten Arbeitskreis und den dort durchgeführten Methodenworkshops. Wir hoffen, die Einübung des Hürdenlaufs „Methodenschulung" mit diesem Band als eine reizvolle Herausforderung profilieren zu können.

Köln und Osnabrück, im Juli 2015 Die Herausgeber

Literatur

Borries, Bodo von u. a. 2005: Schulbuchverständnis, Richtlinienbenutzung und Reflexionsprozesse im Geschichtsunterricht. Eine qualitativ-quantitative Schüler- und Lehrerbefragung im deutschsprachigen Bildungswesen 2002. Neuried.

Gautschi, Peter 2009: Guter Geschichtsunterricht. Grundlagen, Erkenntnisse, Hinweise. Schwalbach/Ts.

Gautschi, Peter u. a. (Hrsg.) 2007: Geschichtsunterricht heute. Eine empirische Analyse ausgewählter Aspekte. Bern.

Hodel, Jan u. a. 2013: Schülernarrationen als Ausdruck historischer Kompetenz. In: Zeitschrift für Didaktik der Gesellschaftswissenschaften 4, 2, S. 121-145.

Köster, Manuel 2013: Historisches Textverstehen. Rezeption und Identifikation in der multiethnischen Gesellschaft. Berlin.

Lange, Kristina 2011: Historisches Bildverstehen oder Wie lernen Schüler mit Bildquellen? Ein Beitrag zur geschichtsdidaktischen Lehr-Lern-Forschung. Berlin.

Martens, Matthias 2010: Implizites Wissen und kompetentes Handeln. Die empirische Rekonstruktion von Kompetenzen historischen Verstehens im Umgang mit Darstellungen von Geschichte. Göttingen.

Mathis, Christian 2015: „Irgendwie ist doch da mal jemand geköpft worden". Didaktische Rekonstruktion der Französischen Revolution und der historischen Kategorie Wandel. Baltmannsweiler.

Meyer-Hamme, Johannes 2009: Historische Identitäten und Geschichtsunterricht. Fallstudien zum Verhältnis von kultureller Zugehörigkeit, schulischen Anforderungen und individueller Verarbeitung. Idstein.

Meyer-Hamme, Johannes/Thünemann, Holger/Zülsdorf-Kersting, Meik (Hrsg.) 2012: Was heißt guter Geschichtsunterricht? Perspektiven im Vergleich. Schwalbach/Ts.

Spieß, Christian 2014: Quellenarbeit im Geschichtsunterricht. Die empirische Rekonstruktion von Kompetenzerwerb im Umgang mit Quellen. Göttingen.

Waldis, Monika u. a. 2015: Material-Based and Open-Ended Writing Tasks for Assessing Narrative Competence among Students. In: Ercikan, Kadriye/Seixas, Peter (Hrsg.): New Directions in Assessing Historical Thinking. New York, London, S. 117-131.

Zülsdorf-Kersting, Meik 2007: Sechzig Jahre danach. Jugendliche und Holocaust. Eine Studie zur geschichtskulturellen Sozialisation. Berlin.

Manuel Köster

Methoden empirischer Sozialforschung aus geschichtsdidaktischer Perspektive

Einleitung und Systematisierung

In den vergangenen zehn Jahren war geschichtsdidaktische Forschung zu einem entscheidenden Teil empirische Forschung. Nachdem der deutschsprachigen Geschichtsdidaktik lange Zeit in fast schon formelhafter Weise ein Empiriedefizit vorgeworfen wurde (vgl. etwa Rohlfes 1996; 2005, 194; v. Borries 2002; Günther-Arndt/Sauer 2006; Beilner 2014), scheint heute die Zurückweisung dieses Vorwurfs gleichermaßen zum Topos zu werden. Ein Blick auf aktuelle Qualifikationsarbeiten, geschichtsdidaktische Zeitschriften, Tagungsbände (vor allem der Baseler Tagungsreihe „geschichtsdidaktik empirisch"; vgl. Hodel/Ziegler 2009; 2011; Hodel/Waldis/Ziegler 2013; Waldis/Ziegler 2015) und Forschungsüberblicke (Hasberg 2001; Beilner 2003; Günther-Arndt/Sauer 2006; Gautschi 2013; Köster/Thünemann/Zülsdorf-Kersting 2014; Barricelli/Sauer 2015) zeigt, dass empirische Forschung einen zentralen Bestandteil geschichtsdidaktischer Forschung ausmacht. Empirisch arbeitenden Geschichtsdidaktiker/-innen steht dabei kein genuin geschichtsdidaktisches Forschungsinstrumentarium zur Verfügung – ein solches scheint auch kaum vorstellbar. Der Einschätzung Saskia Handros und Bernd Schönemanns (2002, 4 f.; ähnlich auch Hasberg 2007, 25), das Proprium der Geschichtsdidaktik liege weniger in ihren Methoden als vielmehr in ihren Kategorien und Fragestellungen, ist auch mehr als zehn Jahre später uneingeschränkt zuzustimmen. Vielmehr bedienen sich empirisch arbeitende Geschichtsdidaktiker/-innen der Instrumente benachbarter Disziplinen. Vor allem die empirische Sozialforschung und die Psychologie fungieren hier als wichtige Bezugsdisziplinen.

In einem kürzlich publizierten Vortrag zum Stand des wissenschaftlichen Nachwuchses innerhalb der Geschichtsdidaktik vertrat Wolfgang Hasberg die Auffassung, die Geschichtsdidaktik „wandel[e] sich zunehmend zu einer empirischen Sozialwissenschaft mit psychologischen Anteilen" bei gleichzeitiger Gefahr des „methodischen Dilettantismus, weil sie nicht nur die Methodenarrangements, sondern zugleich die Erkenntnisinteressen ihrer Bezugsdisziplinen zu übernehmen" drohe (Hasberg 2014, 47 f.), während ihr

gleichzeitig „die Historiker abhanden" kämen (ebd., 48). Dieser Einschätzung muss man ebenso wenig zustimmen wie der Aussage, der Import empirischer Methoden aus den Nachbardisziplinen führe zu Arbeiten, bei denen die Frage, welchen „Beitrag sie zur Lösung genuin geschichtsdidaktischer Fragestellungen leisten, [...] häufig im Dunkeln" bleibe (ebd., 47). Hasbergs zugespitzter Kommentar zeigt aber deutlich, dass der bloße, unreflektierte Import sozialempirischer Methoden nicht automatisch zu innovativen geschichtsdidaktischen Fragestellungen und wegweisenden neuen Befunden führt. Entscheidend bleibt, auf der Grundlage geschichtsdidaktischer Erkenntnisinteressen geeignete Forschungsmethoden auszuwählen.

Eine Ausbildung in den Methoden empirischer Sozialforschung haben allerdings in der Regel nur diejenigen Geschichtsdidaktiker/-innen erfahren, die ein entsprechendes Zweit- oder Nebenfach studiert haben. Auf Forscher mit akademischem Hintergrund in den Geisteswissenschaften kann das theoretische, begriffliche und vor allem methodische Arsenal der genannten Bezugsdisziplinen bisweilen freilich recht hermetisch wirken. Der vorliegende Band möchte deshalb dazu beitragen, Methoden sozialempirischer und (häufig kognitions-)psychologischer Forschung geschichtsdidaktisch zu profilieren, um so das spezifische Erkenntnisinteresse und -potential verschiedener Verfahren *fachspezifisch* zu konkretisieren und bisher noch empirieunerfahrenen Geschichtsdidaktikerinnen und Geschichtsdidaktikern einen ersten Zugang zum differenzierten (und manchmal auch diffusen) Feld verschiedener Methoden zu eröffnen. Eine zumindest überblicksartige Kenntnis empirischer Methoden ist für Geschichtsdidaktiker aus zwei Gründen notwendig: Zum einen ist sie Voraussetzung für die Durchführung eigener Untersuchungen, sofern sie über die engeren Disziplingrenzen hinweg anschlussfähig sein sollen. Zum anderen bietet sie Orientierung bei der angemessenen Rezeption und Einschätzung anderer empirischer Studien (vgl. Hasberg 2007; einen ausführlichen pädagogisch-psychologischen Leitfaden liefert Rost 2013).

Die folgende Einleitung unternimmt den Versuch, das große Spektrum empirischer Zugänge zu systematisieren und die Beiträge dieses Bandes innerhalb dieser Systematik zu verorten. Insgesamt kann es freilich nur darum gehen, erste Schneisen zu schlagen und aus geschichtsdidaktischer Perspektive relevante Methoden exemplarisch vorzustellen.[1]

1 Für eine Diskussion ausgewählter qualitativer Erhebungsinstrumente vgl. Billmann-Mahecha (1998), für ausgewählte qualitative Auswertungsmethoden vgl. Kölbl (2006) und Martens (2012), für eine vorwiegend retrospektive Methodenreflexion der eigenen empirischen Praxis siehe auch v. Borries (2011).

1. Qualitative und quantitative Zugänge

Wie nicht zuletzt der Beitrag von *Doren Prinz* und *Holger Thünemann* in diesem Band zeigt, schließen qualitative und quantitative Zugänge einander keineswegs aus. Vielmehr liegen gerade in beide Ansätze verbindenden Mixed-Method-Zugängen besondere Potenziale, die unter anderem dazu geeignet sind, die Schwächen einer Methodenfamilie mit den Stärken der anderen auszugleichen. Die Frage „qualitativ oder quantitativ" sollte also idealiter nicht als Entweder-Oder-Frage aufgefasst werden, sondern allenfalls als Frage nach der jeweiligen Schwerpunktsetzung. Im Folgenden sollen – wenngleich etwas holzschnittartig – die Charakteristika qualitativer und quantitativer Methoden voneinander abgegrenzt werden.

Auch wenn aktuelle Handbuchartikel stets betonen, dass der Unterschied zwischen beiden Zugängen keineswegs so eindeutig ist, wie in älteren Publikationen bisweilen suggeriert wurde, verfolgen qualitative und quantitative Methoden doch tendenziell unterschiedliche Ziele und eignen sich damit für unterschiedliche Fragestellungen. Dies hängt auch mit der Neigung beider Methodenfamilien zu unterschiedlichen erkenntnistheoretischen Positionen zusammen: Dabei ist quantitativen Methoden eine gewisse Nähe zu positivistischen und (erkenntnistheoretisch) realistischen Positionen inhärent, indem sie davon ausgehen, dass eine Realität außerhalb des erkennenden Subjektes existiert, die dem erkennenden Subjekt zugänglich ist, und dass es Aufgabe empirischer Forschung ist, diese möglichst exakt zu beschreiben (vgl. Kromrey 2006, 24 ff.). „Möglichst exakt" wird in diesem Kontext – darin besteht die Nähe zu positivistischen Positionen – als analog zu naturwissenschaftlichen Verfahren verstanden (vgl. Pistrang/Barker 2012, 7). Die Grundannahme ist dabei, dass die „wirkliche Welt" inhärent regelgeleitet und strukturiert funktioniert, und zwar unabhängig vom jeweiligen Wirklichkeitsausschnitt. Daraus ergibt sich die Annahme, dass sich unterschiedliche empirische Wissenschaftsdisziplinen nur hinsichtlich ihres Gegenstandes unterscheiden, nicht aber hinsichtlich der prinzipiell anwendbaren Methoden zum Aufdecken dieser Gesetzmäßigkeiten (vgl. Kromrey 2006, 26).

Das Ziel quantitativer Verfahren ist es, Sachverhalte in Form von Zahlen auszudrücken und diese mittels mathematischer Verfahren zu verarbeiten. Die Datenerhebung geschieht zum Beispiel in Form von Fragebögen. Dabei besteht eine der größten Herausforderungen bei der Fragebogenkonstruktion in der Operationalisierung: komplexe theoretische Konstrukte in zahlenmäßig fassbare Items – das sind die einzelnen Fragen oder Statements im Fragebogen – zu verwandeln. Dies ist häufig nicht direkt möglich, da sich

der zu erforschende Sachverhalt nicht unmittelbar quantifizieren lässt. Die Aufgabe bei der Itemkonstruktion besteht dann darin, numerisch fassbare Indikatoren für das zu untersuchende Konstrukt zu finden. Die numerischen Beziehungen zwischen den erfassten Größen bilden dabei die Beziehung zwischen den dahinterstehenden Konstrukten ab. *Christiane Bertram* diskutiert in ihrem Beitrag die Konstruktion derartiger Messinstrumente aus geschichtsdidaktischer Perspektive. Speziell geht es um die Entwicklung eines auf dem FUER-Modell (Körber/Schreiber/Schöner 2007) basierenden Kompetenztests. Mit der Itemkonstruktion befasst sich auch *Johannes Meyer-Hamme*, der eine Re-Analyse der Items jüngerer Studien zum historischen Denken und Lernen unternimmt.

Quantifizierende Verfahren streben häufig Repräsentativität an: Das Sample, also die Gruppe der Befragten, steht stellvertretend für eine größere Gruppe, es bildet die Eigenschaften der sogenannten Grundgesamtheit stellvertretend ab. Quantitative Forschung, wie sie in der Geschichtsdidaktik in den 1990er Jahren vor allem durch Bodo von Borries (1995; 1999) betrieben wurde, bedarf daher großer Stichproben. Quantifizierende Verfahren weisen zudem häufig einen sehr hohen Grad an Standardisierung auf und zeichnen sich durch einen eher linearen Forschungsablauf aus: Die einzelnen Untersuchungsschritte werden auf vorher sehr genau festgelegte Weise nacheinander abgearbeitet. Im Gegensatz zu qualitativen Verfahren zielt quantitative Forschung auf die Prüfung von Hypothesen über Zusammenhänge und Regelmäßigkeiten ab. Auch hierin drückt sich ihre bereits beschriebene, analytisch-nomologische erkenntnistheoretische Grundhaltung aus (vgl. Kromrey 2006, 29). Die Hypothesenprüfung erfolgt dabei durch zum Teil recht anspruchsvolle Verfahren statistischer Datenauswertung.

Qualitativen Ansätzen geht es dagegen in aller Regel nicht um die Prüfung von Hypothesen, sondern vielmehr um die möglichst dichte und authentische Beschreibung des zu untersuchenden Feldes. Anstelle der Quantifizierung von Sachverhalten bildet eher das Verstehen individueller Perspektiven der Beforschten das Ziel qualitativer Verfahren. Zugespitzt ließe sich sagen: Während der einzelne Proband für die qualitative Forschung primär in seiner Individualität, als zu verstehender „Fall" interessant ist, dient er in der mit großen Stichproben arbeitenden quantitativen Forschung eher als „Datenlieferant", während das eigentliche Untersuchungsinteresse mehr im Allgemeinen als im Individuellen liegt. Damit weist qualitative Forschung häufig eine gewisse Nähe zu naturalistischen und konstruktivistischen Positionen auf (vgl. Flick 2014, 100 ff.). Wie Willig (2008; vgl. auch Kromrey 2006, 33) betont, sollten realistische und konstruktivistische Positionen allerdings eher als ein Kontinuum verstanden werden, bei dem qua-

litative und quantitative Ansätze tendenziell zu unterschiedlichen Positionen neigen, auf dem sie aber keineswegs an einem der Endpunkte festgelegt sind. Auch qualitative Forschung kann realistische Positionen beziehen, etwa wenn individuelle Sichtweisen an anderen Informationen gemessen werden, und auch quantifizierende Ansätze können sich der Erforschung individueller Konstrukte widmen.

Qualitative Forschung ist geprägt vom „Prinzip der Offenheit" (Mayring 2010a, 225), das sich einerseits darin ausdrückt, dass die Prüfung vorher festgelegter Hypothesen und Deutungen in der Regel als zu starke Einengung möglicher Erkenntnisse aufgefasst wird. Andererseits findet dieses Prinzip Ausdruck im häufig iterativen Forschungsablauf: Einzelne methodische Schritte werden mehrfach vollzogen, wobei im bisherigen Prozess gewonnene Erkenntnisse mit einfließen. Das liegt auch daran, dass qualitative Forschung die Vergleichbarkeit der einzelnen Fälle nicht über standardisierte Messinstrumente zu erreichen versucht, sondern über die Berücksichtigung möglichst aller Situationsbedingungen, auf die jeweils situationsspezifisch flexibel reagiert werden muss. Hypothesen über mögliche Zusammenhänge bilden daher eher den End- als den Ausgangspunkt qualitativer Forschung. Insgesamt weisen qualitative Methoden in aller Regel ein eher geringeres Maß an Standardisierung auf, sollten aber gerade deshalb im Sinne der Methodentransparenz begründungsintensiv verfahren.

Im Folgenden sollen zentrale quantitative und qualitative Methoden aus geschichtsdidaktischer Perspektive kurz profiliert werden. Spricht man von empirischen Methoden, so ist zunächst einmal zwischen solchen der Datenerhebung – Interviews, Fragebögen, Videographien, etc. – und solchen der Datenauswertung – mathematisch-statistische Verfahren, Qualitative Inhaltsanalyse, Dokumentarische Methode, usw. – zu unterscheiden. Zunächst gilt es jedoch, entsprechend der zu verfolgenden Fragestellung das Gesamtdesign der Studie – Experiment, Feldstudie, etc. – festzulegen. Die Grenzen sind hier nicht immer ganz eindeutig zu ziehen, zumal in der quantitativen, noch viel mehr aber in der qualitativen Forschung methodologische Entscheidungen in einem der drei Methodenbereiche Auswirkungen auf die anderen Bereiche haben. Entsprechend sind häufig mehre methodologische Entscheidungen parallel zu treffen. Hier sollen in erster Linie die jeweils zentralen Charakteristika vergleichend herausgearbeitet und geschichtsdidaktisch beleuchtet werden. Für eine vertiefte Diskussion sei auf die ausgesprochen umfangreiche sozialwissenschaftliche (z.B. Flick u. a. 1995; Kromrey 2006; Diekmann 2008; Schnell/Hill/Esser 2013; Flick 2014), erziehungswissenschaftliche (z.B. Helsper/Böhme 2008; Frieberts-

häuser/Langer/Prengel 2013) und psychologische (z.B. Mey/Mruck 2010a; Cooper u. a. 2012) Handbuchliteratur verwiesen.

Die einzelnen Beiträge dieses Bandes setzen auf unterschiedlichen der genannten Ebenen an. Während *Christiane Bertram* und *Johannes Meyer-Hamme* jeweils quantitative Erhebungsinstrumente diskutieren, rücken in *Monika Waldis* Beitrag mit der Videographie Datenerhebung und -auswertung gleichermaßen in den Blick. Hier werden sowohl quantitative wie auch qualitative Zugänge erörtert. Drei weitere Beiträge fokussieren qualitative Auswertungsmethoden: *Christian Mehr* erörtert Potenziale der Objektiven Hermeneutik, *Matthias Martens, Christian Spieß* und *Barbara Asbrand* diskutieren in ihrem Beitrag die Dokumentarische Methode als Auswertungsmethode für Unterrichtsvideos und *Sebastian Barsch* wirft einen geschichtsdidaktisch orientierten Blick auf die Qualitative Inhaltsanalyse. Der Beitrag von *Doren Prinz* und *Holger Thünemann* dagegen setzt auf Ebene des Gesamtdesigns empirischer Untersuchungen an: Hier werden unterschiedliche Varianten des Mixed Methods-Ansatzes, also der Verbindung qualitativer und quantitativer Zugänge innerhalb eines Forschungsprojektes, analysiert.

2. Quantitative Methoden

Entsprechend dem zugrundeliegenden analytisch-nomologischen Paradigma und der damit einhergehenden engen Orientierung an den Naturwissenschaften unterscheiden sich quantifizierende Ansätze – anders als qualitative Verfahren – hinsichtlich der Daten*auswertung* nicht grundsätzlich. Sie vertrauen vielmehr auf ein differenziertes, aber durchaus vergleichbares Repertoire statistischer Verfahren.[2] Unterschiede liegen eher in der Gesamtanlage der Studie, dem damit verbunden Erkenntnisinteresse sowie den Methoden der Daten*erhebung*. Die folgende, nach diesen methodischen Kriterien unterscheidende Darstellung folgt der auf Skowronek und Schmied (1977) beruhenden Systematisierung von Methodendesigns der Unterrichtsforschung bei Böhm-Kasper und Weishaupt (2008). Sie systematisiert quantitative Ansätze hinsichtlich ihres primären Forschungszwecks, der hier in *Erklären*, *Beschreiben* sowie *Bewerten* und *Verändern* unterteilt wird.

2 Selbstverständlich gibt es bei der quantitativen Datenauswertung durchaus unterschiedliche Verfahren, jedoch basieren diese alle grundsätzlich auf Varianten mathematisch-statistischer Operationen. Dabei sind unterschiedliche Auswertungsverfahren für unterschiedliche Fragestellungen und vor allem für unterschiedliche Datensätze geeignet.

2.1 Methodische Designs

2.1.1 Erklärende Ansätze

Bei auf die Erklärung von Sachverhalten bezogener quantifizierender Forschung lassen sich experimentelle und kausal-vergleichende Designs unterscheiden, wobei nur experimentelle Forschung zur Erklärung von Ursache-Wirkung-Beziehungen geeignet ist – wenngleich andere Designs solche Beziehungen durchaus nahelegen können. Experimentelle Designs zeichnen sich dadurch aus, dass das Sample aus mindestens zwei Gruppen besteht: aus einer Experimentalgruppe (auch Interventions- oder Treatmentgruppe genannt) und einer Kontrollgruppe. Ziel experimenteller Forschung ist es, herauszufinden, ob die Veränderung einer Variable (das ist die sogenannte unabhängige Variable) zu Veränderungen bei einer anderen Variable (der abhängigen Variable) führt. Es soll also ein statistischer Beleg für vermutete Ursache-Wirkung-Beziehungen erfolgen.

Wollte man etwa experimentell untersuchen, ob und wie ein Dokumentarfilm zum Nationalsozialismus das Geschichtsbewusstsein beeinflusst, so wäre das Anschauen der Dokumentation die unabhängige Variable, von der vermutet wird, dass sie die abhängige Variable, das Geschichtsbewusstsein, beeinflusst. Um den vermuteten Zusammenhang untersuchen zu können, müsste die Kategorie Geschichtsbewusstsein zunächst operationalisiert werden. Die Forschenden müssten also überlegen, an welchen Indikatoren sie Geschichtsbewusstsein festmachen wollen. Im vorliegenden Fall ließe sich zum Beispiel die Einschätzung unterschiedlicher historischer Sach- und Werturteile in einer Fragebogenerhebung als Ausdruck von Geschichtsbewusstsein ansehen. In einem experimentellen Design würde nun vor dem Setzen des Treatments, also dem Anschauen der Dokumentation durch die Experimentalgruppe, die Ausprägung der abhängigen Variable in einem Pretest erhoben. Anschließend würden die Probanden auf Experimental- und Kontrollgruppe verteilt. Hierbei stehen prinzipiell zwei Möglichkeiten zur Verfügung (vgl. Böhm-Kasper/Weishaupt 2008, 94 f.): die Parallelisierung und die Randomisierung. Beide Verfahren sollen dazu dienen, den Einfluss anderer Variablen, die die abhängige Variable ebenfalls beeinflussen können, zu kontrollieren. In unserem Beispiel könnte dies etwa das Vorwissen sein, aber auch bei anderen Faktoren wie dem methodischen und gattungsspezifischen Wissen – etwa zu Konstruktionsprinzipien historischer Dokumentarfilme – könnte ein Einfluss vermutet werden. Bei der Parallelisierung wird versucht, Personenpaare, die hinsichtlich der Ausprägung dieser Störvariablen vergleichbar sind, auf Experimental- und Kontrollgruppe zu verteilen. Unterschiedliche Ausprägungen der abhängigen Variable im Post-Test wären

dann allein auf das Treatment zurückzuführen. Dies kann jedoch mitunter zu großen Herausforderungen führen: Gerade dann, wenn von mehr als einer Störvariable auszugehen ist, kann das Bilden von Paaren schwierig oder unmöglich sein. Hinzu kommt, dass häufig nicht alle Störvariablen schon vor Untersuchungsbeginn bekannt sind. Bei der Randomisierung dagegen werden Personen allein nach dem Zufallsprinzip auf beide Gruppen verteilt. Die zufällige Verteilung sorgt dafür, dass die unterschiedlichen Störvariablen in ihren verschiedenen Ausprägungen gleichmäßig auf beide Gruppen verteilt werden. Dies setzt jedoch ein sehr großes Sample voraus, was in der geschichtsdidaktischen Forschung eher eine Ausnahme darstellt. Bei der Samplebildung wird hier stattdessen häufig mit bereits bestehenden Gruppen, etwa mit zwei Schulklassen, gearbeitet, die sich natürlich systematisch hinsichtlich etwaiger Störvariablen unterscheiden können (vgl. Klauer 2006, 78). Auch eine Parallelisierung ist hier also ausgeschlossen. In derartigen Fällen bleibt nur, externe Faktoren mittels komplexer multivariater statistischer Verfahren[3] zu kontrollieren (vgl. Diekmann 2008, 68). In solchen Fällen handelt es sich dann allerdings um keine „echten" Experimente mehr, sondern um quasi-experimentelle Studien (vgl. ebd. 39 f., 356 ff.). Diekmann (2008, 356) definiert daher Quasi-Experimente als „in der Hauptsache Experimente ohne Randomsierung."

Anders als in der Psychologie, bei der das Experiment gewissermaßen den methodischen Königsweg darstellt, sind „echte" experimentelle Designs mit randomisierter Verteilung der Probanden in der Geschichtsdidaktik kaum anzutreffen. Nicht nur die bereits beschriebenen Schwierigkeiten bei der Samplebildung und beim Ausschluss etwaiger Störvariablen sind hier als Ursache zu nennen, sondern auch die Tatsache, dass in vielen Fällen die unabhängige Variable keiner direkten Manipulation durch den Forscher zugängig ist. Will man etwa untersuchen, ob das historische Interesse Jugendlicher (abhängige Variable) von bestimmten Erfahrungen im Kindesalter (unabhängige Variable) abhängt, so sind diese Erfahrungen für die einzelnen Probanden, anders als die Treatments in einem (quasi-)experimentellen Design, nicht mehr veränderbar. Bei derartigen Designs handelt es sich um kausalvergleichende oder Ex-post-facto-Untersuchungen (Böhm-Kasper/Weishaupt 2008, 96). Die Kontrolle eventueller Störvariablen ist hier natürlich

3 Gerade im Bereich der Schul- und Unterrichtsforschung kommen daher bei der Datenanalyse sogenannte Mehrebenenanalysen zum Einsatz. Mithilfe dieser Verfahren ist es möglich, den Einfluss individueller Merkmale von dem sogenannter Aggregatmerkmale, also etwa den gemeinsamen Unterrichtserfahrungen einer Schulklasse, zu trennen und das Zusammenwirken dieser beiden Merkmalsgruppen zu untersuchen (vgl. Böhm-Kasper/Weishaupt 2008, 114 f.).

ebenfalls ein Problem: Parallelisierung und Randomisierung sind in der Regel nicht möglich, sodass Drittvariablen nur mittels Kovarianzanalyse kontrolliert werden können. Hier muss also mithilfe statistischer Verfahren versucht werden, die Ausprägung anderer, ebenfalls das Geschichtsinteresse beeinflussender Variablen konstant zu halten, um den Einfluss bereits erfolgter Erfahrungen im Kindesalter zu untersuchen. Bei Ex-post-facto-Designs ergibt sich durch die gleichzeitige Erhebung der abhängigen wie der unabhängigen Variable zudem häufig das Problem der ungeklärten Reihenfolge: Es ist nicht klar, welche der Variablen eigentlich die abhängige Variable ist. Im bislang diskutierten Beispiel würde sich dieses Problem nicht unbedingt stellen. Wollte man aber etwa untersuchen, ob das Geschichtsinteresse die Fachnote beeinflusst, wäre dies der Fall: Auch bei einer eindeutigen positiven Korrelation beider Konstrukte ließe sich die Richtung der Beeinflussung nicht beweisen. Führt das hohe Interesse am Gegenstand (möglicherweise über den Umweg einer Drittvariable) zur guten Note, oder hat möglicherweise die zunächst allein über Leistungsmotivation erreichte gute Schulnote das inhaltliche Interesse geweckt? Beide Antworten sind womöglich nicht gleichermaßen plausibel, eindeutige Antworten auf Fragen nach Ursache-Wirkung-Beziehungen können bloße Korrelationen aber nicht geben.

2.1.2 Beschreibende Ansätze

Wie die Bezeichnung nahelegt, geht es quantitativ-deskriptiven Untersuchungen nicht um die Prüfung von Hypothesen zu Wirkungsbeziehungen, sondern um die Beschreibung eines Phänomens. In der quantitativen Forschung stehen deskriptive Ansätze deshalb häufig am Beginn eines Forschungsprozesses, wenn ein noch neues Feld erschlossen werden soll. Wie Böhm-Kasper und Weishaupt (2008, 99) betonen, arbeitet die deskriptive Forschung in der Regel mit sehr großen Fallzahlen und repräsentativen Stichproben. In der geschichtsdidaktischen Empirie sind die bereits erwähnten großen Studien von Bodo von Borries (1995; 1999) typische Beispiele für derartige Designs. Hier wurde das Geschichtsbewusstsein Jugendlicher in Deutschland und im europäischen Vergleich mittels Fragebogenerhebung erfasst. In beiden Studien wurde dafür auf sehr große, wenngleich nicht immer (1995) repräsentative Samples zurückgegriffen. In seiner explorativen Ausrichtung und bezüglich seiner quantifizierenden Methoden kann auch das Schweizer Videoprojekt „Geschichtsunterricht heute“ (Gautschi u. a. 2007) als weiteres geschichtsdidaktisches Beispiel für derartige methodische Designs gelten. Anders als bei den Fragebogensurveys der 1990er Jahre wurde hier jedoch mit Videographien von Geschichtsunterricht gearbeitet, die

kodiert und quantitativ analysiert wurden. Die Fallzahlen waren daher naturgemäß deutlich kleiner.

Derartige deskriptive Designs können als Längs- oder Querschnittstudien angelegt sein (vgl. Böhm-Kasper/Weishaupt 2008, 99). Ein Beispiel für ein längsschnittliches Design (auch als Panelbefragung bezeichnet) bietet Georg Kanerts Dissertation (2014), für die Lehrpersonen des Faches Geschichte wiederholt mit dem gleichen Messinstrument befragt wurden, um zu erheben, wie diese die Wirksamkeit und Bedeutsamkeit in der Ausbildung erworbenen Wissens in unterschiedlichen berufsbiographischen Phasen einschätzen. Typische methodische Herausforderungen bestehen bei derartigen Designs neben der sogenannten Panelmortalität, also dem möglichen sukzessiven Rückgang der Rücksendequote durch Umzug, Krankheit oder Nichtinteresse, vor allem in der Attribuierung beobachteter Effekte: Sind etwaige Veränderungen im Antwortverhalten über die Zeit beruflichen Erfahrungen, veränderten subjektiven Theorien, neuen curricularen Rahmenbedingungen, Veränderungen in der Schülerschaft, gesellschaftlichen Diskursen oder anderen Faktoren zuzuschreiben? Bodo von Borries' Arbeiten aus den 1990er Jahren dagegen sind typische Vertreter querschnittlicher Designs, die versuchen, einen möglichst repräsentativen Ausschnitt der interessierenden Population abzubilden. Bisweilen werden derartige Designs auch als eine Art Pseudolängsschnitt verwandt, wenn die Befunde unterschiedlicher Altersgruppen miteinander verglichen werden und die Unterschiede auf Sozialisations- und Reifungsprozesse zurückgeführt werden (vgl. von Borries 1987; 2001). Dies kann freilich bestenfalls eine Behelfslösung darstellen, da ja nicht die gleichen Probanden in unterschiedlichen Lebensphasen befragt wurden.

Längs- wie Querschnittdesigns sind zudem häufig Gegenstand von Korrelationsuntersuchungen: Hier wird untersucht, ob die hohe Ausprägung eines Merkmals (zum Beispiel des Geschichtsinteresses) mit der hohen (positive Korrelation) oder niedrigen (negative Korrelation) Ausprägung eines zweiten Merkmals (zum Beispiel des historischen Inhaltswissens) korreliert. Wie bei den bereits erwähnten kausal-vergleichenden Untersuchungen sind auch hier keine Aussagen zu Ursache-Wirkung-Beziehungen möglich, wenngleich Hypothesen über derartige Beziehungen natürlich für spätere Studien generiert werden können. Werden, wie beim Geschichtsbewusstsein, komplexe Wechselwirkungsbeziehungen zwischen unterschiedlichen Faktoren angenommen, können Pfadanalysen, wie Bodo von Borries (2001) sie vorgelegt hat, Auskunft über mögliche Wirkungsrichtungen geben: Hier werden mehrere Alternativen der Richtung der gegenseitigen Beeinflussung verschiedener Faktoren mit statistischen Verfahren geprüft. Das Ergebnis der

Pfadanalyse gibt Auskunft darüber, welche Einflussrichtung anhand des Datenmaterials am wahrscheinlichsten ist (vgl. Böhm-Kasper/Weishaupt 2008, 116).

2.1.3. Bewertende und auf Veränderung zielende Ansätze

Evaluationsforschung kombiniert Aspekte der bereits genannten methodischen Designs mit dem Erkenntnisinteresse, die Güte und Eignung des Untersuchungsgegenstandes in Bezug auf einen bestimmten Aspekt zu untersuchen (vgl. Böhm-Kasper/Weishaupt 2008, 100). Bekanntestes Beispiel für derartige Designs sind sicherlich internationale Schulleistungsvergleiche wie TIMMS und PISA. *Christiane Bertrams* Beitrag in diesem Band stellt die Entwicklung eines vergleichbaren Testinstrumentes zur Diagnose historischer Kompetenzen dar. Ganz im Sinne der Logik von Evaluationsforschung vergleicht der methodische Ansatz ihrer Dissertation, auf die hier Bezug genommen wird, die Effektivität unterschiedlicher Formen der Arbeit mit Zeitzeugen im Geschichtsunterricht. Das beschriebene Testinstrument dient dem standardisierten Vergleich historischer Kompetenzen über mehrere Versuchs- und Kotrollgruppen. Da mit existierenden Schulklassen gearbeitet wurde, handelt es sich hier also um ein quasi-experimentelles Design mit dem Ziel der Bewertung verschiedener unterrichtlicher Maßnahmen hinsichtlich ihrer Effektivität in Bezug auf die Anbahnung historischer Kompetenzen.

2.2 Quantitative Erhebungsmethoden

Die Wahl der passenden Erhebungsmethode kann von einer Anzahl unterschiedlicher Faktoren abhängen. Entscheidend sind hier natürlich in erster Linie Fragestellung und Erkenntnisinteresse: Interessiere ich mich für Kommunikationsprozesse im Geschichtsunterricht, so sind beobachtende Methoden eine naheliegende Wahl. Interessieren mich dagegen die Einstellungen von Geschichtslehrern, so erscheinen (mündliche oder schriftliche) Befragungen naheliegender. Eng damit zusammen hängt die Frage der Operationalisierung – und damit auch der Quantifizierung – des Untersuchungsgegenstandes: Je nachdem, wie dieser sich in zählbaren Indikatoren fassen und erfassen lässt, liegen bestimmte Erhebungsmethoden näher als andere. Häufig liegt gerade in der Kombination unterschiedlicher Methoden besonderes Potenzial, wie *Doren Prinz* und *Holger Thünemann* in ihrem Aufsatz in diesem Band zeigen. Im Folgenden werden die drei üblicherweise (etwa bei Kromrey 2006; Diekmann 2008; Böhm-Kasper/Weishaupt 2008) unterschiedenen quantifizierenden methodischen Großformen beschrieben: die

Befragung, die Beobachtung und die Inhaltsanalyse. Dabei eignen sich laut Atteslander (2006, 48 f.) inhaltsanalytische Verfahren für die Analyse der *Produkte* menschlicher Tätigkeit,[4] während sich Beobachtung und Befragung bei der Untersuchung aktuellen menschlichen *Verhaltens* in „natürlichen Situationen" eignen. Atteslander grenzt – etwas unscharf – Experimente als eigene Gruppe hiervon ab, da diese der Erforschung des Verhaltens „in vom Forscher bestimmten Situationen" (49) dienten. Hier ließe sich jedoch durchaus einwenden, dass die Klassifizierung als experimentelles Design eher eine Aussage über das Gesamtdesign einer Untersuchung (Pre-Post-Design mit randomisierter Verteilung auf Treatment- und Kontrollgruppe) als über den Erhebungsort darstellt. Schließlich sind auch experimentelle Designs in „natürlichen Settings" vorstellbar (Feldexperimente).

2.2.1 Quantitative Befragungen

a) Persönliche und telefonische mündliche Befragungen

Grundsätzlich stehen der quantitativen Forschung Interviews als Erhebungsinstrumente zur Verfügung. Diese sind jedoch mit mindestens drei grundlegenden Schwierigkeiten verbunden: dem Organisationsaufwand, der Standardisierung der Fragen und der Kodierung der Antworten. Da quantifizierende Studien häufig mit großen Stichproben arbeiten, sind mündliche – dazu zählen auch die von Meinungsforschungsinstituten gern eingesetzten telefonischen – Befragungen häufig nicht praktikabel. Gespräche mit jedem einzelnen Probanden und jeder Probandin zu führen, sprengt bei der für avanciertere statistische Analysen notwendigen hohen Fallzahl die Kapazitäten vieler Projekte – zumindest in der Größenordnung, wie sie in der geschichtsdidaktischen Empirie üblich sind. Hier stellt sich zudem das Problem der Standardisierung, und zwar in doppelter Hinsicht: Einerseits müssen die Interviewfragen anhand eines strikten, geschlossenen Leitfadens abgearbeitet werden. Um Vergleichbarkeit zu gewährleisten, muss jede/r Beforschte jede Frage in identischer Formulierung und Reihenfolge gestellt bekommen. Dies lässt sich praktisch nur erreichen, indem die Fragen abgelesen werden. Die Antworten der Probanden wiederum müssen in zählbare Einheiten verwandelt werden. Eine Möglichkeit, um dies zu erreichen, ist es, mögliche Antwortalternativen vorzugeben. Wer einmal an einer telefonischen Meinungsumfrage teilgenommen hat, weiß, dass die Kombination aus

4 Inhaltsanalysen sind deshalb eigentlich keine Erhebungs-, sondern Auswertungsverfahren, werden jedoch in der einschlägigen Handbuchliteratur zu den Erhebungsverfahren gezählt. Da quantitative Auswertungsverfahren im Kontext dieses Kapitels nicht weiter diskutiert werden können, wird diese Konvention auch hier beibehalten.

vorgelesener Frage und zur Auswahl gestellter Antwort nicht nur zu einer künstlichen Gesprächssituation führt, sondern auch durchaus anspruchsvoll – die Probanden müssen sich mehrere Antwortalternativen und die dazugehörige Kennung merken und daraus auswählen, bevor sie antworten – und bisweilen geradezu frustrierend sein kann. Eine zweite Möglichkeit besteht darin, die Probanden frei antworten zu lassen und diese Antworten unter verschiedene Kategorien zu subsummieren. Derartige Kategorisierungen stellen freilich bereits einen komplexen Interpretationsakt dar. Selbst derartige, etwas offenere Formen standardisierter Interviews unterscheiden sich sehr deutlich von Alltagsgesprächen, wie Noelle-Neumann und Petersen (2000, 60) zugespitzt feststellen: Der Interviewer „stellt Fragen nach völlig privaten Dingen [...], wechselt sprunghaft die Themen, geht überhaupt nicht persönlich auf seine Gesprächspartner ein, sondern schert alle Gesprächspartner über einen Kamm, führt das Gespräch ‚nach Schema F' und verstößt dabei gegen alle Regeln einer gebildeten Unterhaltung." Neben den genannten Schwierigkeiten ist bei mündlichen Befragungen im Vergleich zu schriftlichen auch von einer höheren Reaktivität, also der Beeinflussung des Antwortverhaltens durch die Interviewsituation und die Anwesenheit des Forschenden, auszugehen. Stark standardisierte mündliche Befragungen bieten gegenüber schriftlichen allerdings den Vorteil, dass im Zweifel Rückfragen möglich sind. So sind systematische Verzerrungen aufgrund missverständlich formulierter Items hier unwahrscheinlicher.

b) Schriftliche Befragungen

Wie bereits erwähnt, besteht die zentrale Aufgabe bei schriftlichen Befragungen in der Operationalisierung der zu erhebenden Konstrukte. Die jeweiligen Indikatoren verschiedener Aspekte des zu erforschenden Konstruktes werden mit Items erhoben, die sich unterschiedlich klassifizieren lassen (vgl. Schnell/Hill/Esser 2011, 319-333; Diekmann 2008, 471-479). Hier sind zunächst einmal offene und geschlossene Formen zu unterscheiden: Während offene Fragen die Antwortformulierung dem Probanden überlassen, geben geschlossene Fragen Antworten vor. Offene Fragen haben den Vorteil, dass die Antworten tatsächlich dem Relevanzsystem der Befragten entsprechen. Geschlossene Formate dagegen können Antworten enthalten, an die die Befragten noch nie gedacht haben, die aber spontan plausibel erscheinen. Ebenso kann es der Fall sein, dass keine der Antwortmöglichkeiten den Befragten wirklich befriedigend erscheint, etwa weil diese anhand wissenschaftlicher Theorien kategorisiert sind, während die Befragten in Alltagstheorien und -kategorien denken. Dagegen bieten geschlossene Fragen für quantifizierende Untersuchungen den Vorteil höherer Standardisierung.

Nicht nur entfällt der interpretative Akt der Kategorisierung offener Antworten, sondern es kann auch sichergestellt werden, dass unterschiedliches Antwortverhalten nicht auf unterschiedliche Fähigkeiten beim Formulieren der Antworten zurückzuführen ist. Aber auch bei offenen Fragen geht es der quantitativen Forschung darum, die Antworten in ein (häufig a priori gebildetes) Kategorienraster einzuordnen. Wie Bodo von Borries (2013, 123) angesichts eigener Erfahrungen mit der Kategorisierung offener Items feststellt, gehen dabei „leider ausgesprochen innovative und alternative Ansätze der Lernenden, auch wenn sie nicht nur ‚teilrichtig', sondern sogar kreativ und überlegen sind, meist unter."

Fragen lassen sich zudem hinsichtlich des Frageinteresses klassifizieren (vgl. Schnell/Hill/Esser 2011, 319-333; Diekmann 2008, 471-479): So wird häufig zwischen Fragen nach Einstellungen, Überzeugungen, Verhalten und sozialstatistischen Merkmalen unterschieden. Fragen nach Einstellungen sind häufig in Form von Statements verfasst, zu denen sich die Befragten auf einer Ratingskala positionieren sollen (*„Die De-Konstruktion historischer Dokumentarfilme fördert die Einsicht in den Konstruktcharakter von Geschichte." „Stimme überhaupt nicht zu – stimme eher nicht zu – stimme eher zu – stimme voll zu"*), aber auch Rankingaufgaben, bei denen Items in eine Reihenfolge gebracht werden müssen, sind gebräuchlich. Im Gegensatz zu diesen Einstellungsfragen erheben Überzeugungsfragen die Überzeugung eines Probanden, dass etwas der Fall sei. Hierzu gehören also auch Wissensfragen. Neben offenen Formaten werden Überzeugungsfragen mittels Multiple-Choice-Verfahren (*„Auf welcher Ebene bewegt sich das Prinzip der Kontroversität in Bergmanns Konzeptionierung von Multiperspektivität?" „a. auf der Ebene der Quellen, b. auf der Ebene der Darstellungen, c. auf der Ebene der Schülerurteile"*) und dichotomer Ja/Nein-Fragen (*„Sinkt der Anteil geschichtsinteressierter Schülerinnen und Schüler an deutschen Gymnasien?"*) gestellt. Verhaltensfragen erheben Äußerungen zu Art, Häufigkeit und Dauer eigenen Verhaltens (*„Wie oft arbeiten Sie in ihrem Geschichtsunterricht mit dem Schulbuch?"*). Sozialstatistische Daten werden in aller Regel deshalb erhoben, weil man einen Zusammenhang zwischen diesen und dem zu erforschenden Konstrukt vermutet. So unterschieden sich etwa in einer (nicht repräsentativen) Studie im Mixed Method-Design, die das Textverstehen im Geschichtsunterricht untersuchte, die Einstellungen von Jugendlichen mit und ohne Migrationshintergrund zu einigen Aussagen bzgl. Nationalsozialismus und Holocaust (vgl. Köster 2013, 63 ff.) signifikant voneinander. Eine ausführliche Diskussion der Konstruktionsprinzipien und Itemformulierung bei schriftlichen Befragungen liefert *Christiane Bertrams* Aufsatz in diesem Band.

Im Unterschied zu Fragebögen beruhen Testverfahren auf einem Richtigkeitsstandard. Es gibt mehrere Vorschläge zur Differenzierung von Tests, etwa nach ihrer Funktion (allgemeine Intelligenztests, spezielle Intelligenz- und Begabungstests, Leistungstest, Persönlichkeitstests; vgl. Böhm-Kasper/Weishaupt 2008, 104) oder nach dem Maßstab, an dem individuelle Leistungen gemessen werden (vgl. ebd., 105): Während normorientierte Tests individuelle Leistungen an denen einer Normstichprobe (etwa einer repräsentativen Auswahl aller Achtklässler) messen, bildet bei kriterienorientierten Tests ein vorher definierter Leistungsstandard den Maßstab. Beide Testverfahren dienen unterschiedlichen Zielsetzungen: Normorientierten Tests geht es um die Analyse interindividueller Unterschiede, kriterienorientierten Tests dagegen um den Vergleich von individueller Leistung und vorab festgelegtem Leistungsziel.

Gerade aus geschichtsdidaktischer Perspektive erscheint hier erneut die Operationalisierung der zu erforschenden Konstrukte in Fragebögen bzw. Tests als zentrale Herausforderung. Die Umwandlung geschichtsdidaktischer Theorie in angemessene Fragebogenitems beruht nicht nur darauf, schlüssige Indikatoren für die entsprechenden Theorieelemente zu finden, sondern auch darauf, solche Antwortalternativen bereitzustellen, die die Komplexität des Untersuchungsgegenstandes wahren, ohne dabei aber die Befragten sprachlich zu überfordern, ihre Denkstrukturen mit der Struktur der zu erforschenden Konstrukte zu überformen oder – anders herum – die Struktur des zu erforschenden Konstruktes der Logik des Erhebungsinstrumentes zu unterwerfen. Weniger gelungene Fragebögen dringen unter Umständen nicht bis zu den Tiefenstrukturen historischen Denkens vor, sondern verbleiben an der Oberfläche, z. B. bei der Abfrage historischen Faktenwissens oder gesellschaftlich tradierter Deutungsmuster. *Johannes Meyer-Hamme* unternimmt in seinem Beitrag eine Re-Analyse zweier Studien unter dem Gesichtspunkt, ob die dort verwendeten Fragebögen tatsächlich die Konstrukte messen, die sie zu messen vorgeben. Es geht dort also um die Validität der Erhebungsinstrumente.

Bei der Messung historischer Kompetenzen hingegen wird bisweilen bezweifelt, ob sich überhaupt alle mit dem historischen Denken verbundenen Teilkompetenzen zuverlässig messen lassen (vgl. Körber 2008; Sander 2013) und ob dies ein zielführendes Unterfangen darstellt (vgl. VanSledright 2014; Thünemann 2015). Zudem liegt in der geschichtsdidaktischen Diskussion bislang kein Kompetenzmodell vor, welches die Frage der Graduierung historischer Kompetenzen auf empirisch operationalisierbare Weise gelöst hätte (vgl. Köster/Thünemann 2015). Die Möglichkeit der induktiven Identifizierung verschiedener Kompetenzniveaus mittels quantifizierender Verfah-

ren wird ebenfalls häufig als unwahrscheinliches Unterfangen betrachtet – gerade hinsichtlich solcher Kompetenzbereiche, die auf historische Orientierung der eigenen Lebenspraxis abzielen (vgl. ebd.; Körber 2007). Historisches Denken ist ein in hohem Maße individueller, situationsbezogener, perspektivengebundener und kreativer Prozess, der sich möglicherweise nur sehr schwer in Form standardisierter Antworten fassen lässt. Dies gilt in besonderem Maße für Tests: Der Konstruktcharakter von Geschichte, die Tatsache, dass eher Triftigkeit (vgl. Rüsen 1990, 77-105) als Richtigkeit den Maßstab zur Einschätzung historischer Narrationen bietet, lassen Ergebnisse historischen Denkens nur eingeschränkt mit schematischen Richtig/Falsch-Kategorisierungen vereinbar erscheinen, die die Basis von Tests – und damit auch der auf Grundlage von Tests ermittelten Kompetenzniveaus – darstellen. Diese Position ist innerhalb der Geschichtsdidaktik jedoch durchaus umstritten. *Christiane Betram* vertritt in ihrem Beitrag eine dezidierte Gegenposition.

Ein in der geschichtsdidaktischen Empirie noch neuer Zugang zur Kompetenztestung besteht in der Verbindung von videographiertem Material und standardisierter Befragung. Zur Messung der Kompetenzen von Geschichtslehrkräften wird momentan in zwei Projekten (Kanert/Resch 2014; Waldis u. a. 2014) auf sehr ähnlicher allgemeindidaktischer Theoriebasis (vgl. Shulman 1986) mit sogenannten Unterrichtsvignetten gearbeitet: Hier beurteilen Lehrerinnen und Lehrer videographierte Unterrichtssequenzen mithilfe standardisierter Instrumente. Den Richtigkeitsstandard bilden in beiden Fällen die Urteile universitärer Geschichtsdidaktiker. Gegenüber reinen paper-and-pencil-Verfahren bietet die Arbeit mit Unterrichtsvignetten den Vorteil, mit relativ authentischem, komplexem Material arbeiten zu können. Die der Messlogik inhärente Beurteilung anhand eines a priori definierten Standards von richtig und falsch bleibt freilich auch bei diesem Verfahren bestehen.

Ein eher fachunspezifisches Problem bei der Arbeit mit Fragebögen ist die Rücksendequote (vgl. Diekmann 2008, 516 ff.). Gerade dann, wenn sehr hohe Fallzahlen angestrebt werden, ist das Ausfüllen der Fragebögen in Gegenwart des Forschers – welches freilich den Vorteil der Möglichkeit zur Nachfrage bietet – nicht praktikabel. Werden Fragebögen dagegen per Post verschickt oder wird per Email um die Teilnahme an einer Onlinebefragung gebeten, stellt sich häufig das Problem, dass nur ein geringer Teil der Angeschriebenen antwortet. In der Forschung wurden mittlerweile verschiedene Techniken zur Erhöhung der Rücklaufquote identifiziert, aber trotzdem bleibt das Problem, dass die Rücklaufquote die Ergebnisse verzerrt, bestehen. So stellt sich etwa bei schriftlichen Lehrerbefragungen das Problem,

dass möglicherweise nur besonders erfolgreiche und motivierte Lehrpersonen den Fragebogen zurücksenden (vgl. Kanert 2013). Das Gegenteil ist aber auch vorstellbar: Möglicherweise beteiligen sich vor allem Lehrpersonen, die ihre Tätigkeit als Überforderung oder als wenig befriedigend empfinden, weil sie sich aus den Befunden der Studie einen Lösungsansatz versprechen. Als generell wirksames Verfahren zur Erhöhung der Rücklaufquote hat sich z. B. erwiesen, den Befragten die persönliche Relevanz der Studie in einem Begleitschreiben zu verdeutlichen: Sollen z. B. Lehrerinnen und Lehrer befragt werden, so sollte im Begleitschreiben darauf hingewiesen werden, in welcher Weise die Studie zur Verbesserung der Arbeitsbedingungen oder zur Erhöhung der Unterrichtsqualität beitragen kann. Wird bereits mit dem Fragebogen ein kleines Geschenk versandt, so erhöht dies die Rücklaufquote drastisch (vgl. Diekmann 2008, 519 f.). Jedoch dürfte dies für zahlreiche kleinere Projekte einen kaum zu bewältigen finanziellen Mehraufwand darstellen. Zahlreiche Untersuchungen gehen deshalb so vor, dass den Befragten die Teilnahme an einer Verlosung zugesagt wird. Dies ist jedoch in mehrfacher Hinsicht bedenklich. Zum einen erhöht „ein versprochenes Geschenk, das sozusagen erst nach getaner Arbeit [...] verschickt wird", die Rücklaufquote nicht (ebd., 520), und zum anderen kann dies zu noch deutlich größeren Verzerrungen führen: Einerseits spricht die Aussicht auf einen möglichen Preis unter Umständen eine bestimmte Personengruppe systematisch an, sodass diese dann im Sample überrepräsentiert ist, und andererseits besteht die Gefahr, dass im Fragebogen einfach irgendwie angekreuzt wird, um am Gewinnspiel teilnehmen zu können. Die Ergebnisse würden dann natürlich nicht die tatsächlichen Verhältnisse widerspiegeln.

2.2.2 Quantitative Beobachtungen

Wissenschaftliche Beobachtungen mit dem Ziel der Quantifizierung können in unterschiedlicher Form erfolgen (vgl. Atteslander 2006, 79 ff.): Der Beobachter kann das Geschehen zum Beispiel direkt oder per Videoaufzeichnung beobachten. Wichtiger als diese technische Differenzierung sind die Unterscheidungen anhand der drei Dimensionen der Strukturiertheit, der Offenheit und der Teilnahme. Dabei sind quantitative Beobachtungen häufig sowohl hinsichtlich der Wahrnehmung als auch hinsichtlich der Aufzeichnung in hohem Maße strukturiert: Anhand theoretischer Vorüberlegungen, häufig auch anhand zusätzlicher Pre-Tests, werden Beobachtungskategorien erarbeitet und Indikatoren für die jeweilige Merkmalsausprägung definiert. Dies ist vor allem dann wichtig, wenn der Forscher das Geschehen direkt beobachtet. Mithilfe strukturierter Beobachtungsschemata kann in solchen Fällen der Gefahr der selektiven Wahrnehmung des Forschers ent-