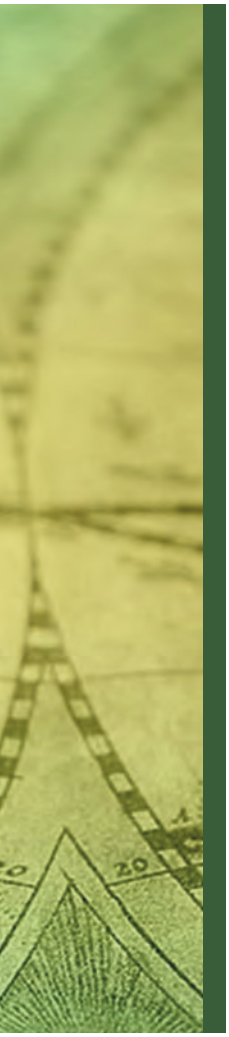
Trautwein, Ulrich; Bertram, Christiane; von Borries, Bodo; ...

Kompetenzen historischen Denkens erfassen. Konzeption, Operationalisierung und Befunde des Projekts "Historical Thinking - Competencies in History" (HiTCH)

Münster ; New York : Waxmann 2017, 144 S.

HiTCH
Historial Thinking
Competencies in History



U. Trautwein, C. Bertram, B. von Borries,
N. Brauch, M. Hirsch, K. Klausmeier, A. Körber,
C. Kühberger, J. Meyer-Hamme, M. Merkt,
H. Neureiter, S. Schwan, W. Schreiber, W. Wagner,
M. Waldis, M. Werner, B. Ziegler, A. Zuckowski

Kompetenzen historischen Denkens erfassen

Konzeption, Operationalisierung und Befunde des Projekts „Historical Thinking – Competencies in History“ (HiTCH)

WAXMANN

Ulrich Trautwein, Christiane Bertram, Bodo von Borries,
Nicola Brauch, Matthias Hirsch, Kathrin Klausmeier,
Andreas Körber, Christoph Kühberger, Johannes Meyer-Hamme,
Martin Merkt, Herbert Neureiter, Stephan Schwan,
Waltraud Schreiber, Wolfgang Wagner, Monika Waldis,
Michael Werner, Béatrice Ziegler, Andreas Zuckowski

Kompetenzen historischen Denkens erfassen

Konzeption, Operationalisierung und Befunde des Projekts „Historical Thinking – Competencies in History“ (HiTCH)

Waxmann 2017
Münster • New York

Bibliografische Informationen der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Print-ISBN 978-3-8309-3598-8
E-Book-ISBN 978-3-8309-8598-3

www.waxmann.com
info@waxmann.com

Umschlaggestaltung: Inna Ponomareva, Jena
Umschlagfoto: © tomertu – Fotolia.com
Satz: Stoddart Satz- und Layoutservice, Münster
Druck: Hubert & Co., Göttingen

Gedruckt auf alterungsbeständigem Papier,
säurefrei gemäß ISO 9706

Inhalt

1 Einleitung ... 9

2 Theoretische Grundlagen: Historische Kompetenzen, Kompetenzmodelle und Kompetenzmessung im Fach Geschichte ... 14
2.1 Die narrativistische Geschichtstheorie – Grundlage aktueller Geschichtsdidaktik ... 15
2.2 Didaktische Wendungen ... 17
2.2.1 Geschichtsbewusstsein ... 17
2.2.2 Didaktische Konsequenz der Perspektivität: Multiperspektivität ... 18
2.2.3 Historisches Denken – der Prozess hinter einem narrativen Geschichtsverständnis ... 20
2.3 Kompetenzen historischen Denkens: Das FUER-Modell ... 25
2.3.1 Grundlagen ... 25
2.3.2 Kompetenzbereiche und Kernkompetenzen – inhaltliche Beschreibung ... 32
2.3.3 Überlappungsbereiche ... 35
2.3.4 Graduierungslogik im FUER-Modell ... 35
2.4 Das FUER-Modell im nationalen und internationalen Kontext ... 38
2.4.1 Weitere deutschsprachige Konzeptionen ... 38
2.4.2 Englischsprachige Konzeptionen ... 40
2.4.3 Die Wahl des FUER-Modells als theoretische Grundlage für den HiTCH-Test – Zusammenschau der Gründe ... 44
2.5 Forschungsübersicht: Standardisierte Testungen und empirische Studien zu historischem Denken ... 45
2.5.1 Deutschsprachige Studien ohne Bezug zur Kompetenzorientierung ... 46
2.5.2 Kompetenzorientierte Studien im deutschsprachigen Raum ... 47
2.5.3 Kompetenzorientierte Studien im englischsprachigen Raum ... 51
2.5.4 Large-Scale-Assessements in Frankreich und den USA: Die Entwicklung der standardisierten Tests CEDRE und NAEP ... 54
2.5.5 Der Stand (inter-)nationaler Ansätze zum Assessment historischen Denkens ... 55

3 Entwicklung eines historischen Kompetenztests für Large-Scale-Assessments ... 56
3.1 Herausforderungen, die mit der Operationalisierung historischer Kompetenzen für Testungen in Large-Scale-Assessments verbunden sind ... 56
3.1.1 Der Konstruktcharakter von Geschichte und die Konsequenzen für die Testentwicklung ... 56
3.1.2 Prozedurale und kategorisierende Kompetenzen historischen Denkens und die Testentwicklung ... 57
3.1.3 Kontext- und Themengebundenheit der Aufgaben ... 57
3.1.4 Umgebende Geschichtskultur als Hintergrundvariable historischen Denkens ... 58

3.2 Idealtypische Schritte bei der Testentwicklung 59
3.2.1 Auswahl eines theoretischen Modells 59
3.2.2 Operationalisierung der theoretischen Konstrukte: empirisches und numerisches Relativ 60
3.2.3 Auswahl der Aufgabenformate 63
3.2.4 Weitere Maßnahmen zur Sicherung der Validität der Messung 63
3.3 Beispiele aus dem HiTCH-Itempool 64
3.3.1 Beispiel-Aufgaben zur historischen Sachkompetenz 64
3.3.2 Beispiel-Aufgaben zur historischen Fragekompetenz 67
3.3.3 Beispiel-Aufgaben zur historischen Methodenkompetenz (Re- und De-Konstruktion) 70
3.3.4 Beispiel-Aufgaben zur historischen Orientierungskompetenz 74
3.4 Der Prozess der Aufgabenentwicklung für den nunmehr vorliegenden HiTCH-Test 75
3.4.1 Adressieren des gesamten Denkprozesses in themenbezogenen Testheften 75
3.4.2 Cognitive Labs 76
3.4.3 Pilotierung I 77
3.4.4 Pilotierung II 78
3.4.5 Haupterhebung 80

4 Haupterhebung: Design, Datenerhebung, Methoden der Datenauswertung 82
4.1 Stichprobe 82
4.2 Testdesign 82
4.3 Validierungsinstrumente 83
4.4 Analysestrategien im HiTCH-Projekt 83

5 Ergebnisse 89
5.1 Itemauswahl und Reliabilität des HiTCH-Instruments 89
5.2 Überprüfung der Ein- bzw. Mehrdimensionalität 94
5.3 Vergleichbarkeit der Itemschwierigkeiten in Subgruppen 98
5.4 Vergleichbarkeit der Itemschwierigkeiten nach Testheftversion 100
5.5 Testleistung und Anzahl nicht bearbeiteter Items 103
5.6 Kriteriumsbezogene und diskriminante Validität 106
5.7 Allgemeine Lesekompetenz vs. historische Kompetenz: Vertiefende Analysen 110
5.8 Historische Kompetenz als Prädiktor für die erfolgreiche Nutzung multipler historischer Dokumente 112

6 Diskussion und Ausblick 116
6.1 Das HiTCH-Projekt: Ein kurzes Fazit 116
6.2 Was misst der HiTCH-Test – und was misst er nicht? 117
6.2.1 Abdeckung des zugrundeliegenden Kompetenz-Konstrukts im HiTCH-Test 118
6.2.2 HiTCH-Test – zu herausfordernd für die Schülerinnen und Schüler? 119
6.2.3 Verhältnis von Wissen und Kompetenzen im HiTCH-Test 120
6.2.4 Graduierungen der Kompetenz 121

6.2.5 HiTCH-Test als Ersatz für schulische Leistungskontrollen? 123
6.3 Ausblick: Die Bedeutung von HiTCH für Forschung und Schule 124
6.3.1 Anregungen für die Forschung 124
6.3.2 Anregungen für den Geschichtsunterricht 125
6.3.3 Weiterarbeit im HiTCH-Projekt 126

Literatur 129

Appendix 144

1 Einleitung

„Warnung!
Geschichte kann zu Einsichten führen
und verursacht Bewusstsein!“

Ein Schild mit dieser Aufschrift begrüßt die Besucherinnen und Besucher am Eingang des zeitgeschichtlichen Forums in Leipzig. Ironisch warnt es vor eigentlich recht wünschenswerten Folgen einer Befassung mit der Vergangenheit. Ganz so leicht ist es aber doch nicht: Weder sind Einsichten und Bewusstsein alleinige oder gar notwendige Folgen jeglicher Beschäftigung mit Geschichte, noch lassen sich Geschichte und Bewusstsein so einfach als Ursache und Wirkung voneinander trennen.

Anders als in Diktaturen, in denen Geschichte unhinterfragt zur Legitimation der politischen Realität benutzt wird, besteht der Anspruch freiheitlicher Gesellschaften darin, ihre Mitglieder zu kritischem Selber-Denken zu befähigen und ihr historisches Bewusstsein zu schärfen. Diesem Ziel dient insbesondere auch kompetenzorientierter Geschichtsunterricht in der Schule. An den jeweils betrachteten Themen sollen Methoden und Einsichten erarbeitet werden, mit deren Hilfe den Lernenden ein reflektierter und (selbst-)reflexiver Umgang mit der Geschichte und Vergangenheit möglich wird.

Die meisten Geschichtsdidaktikerinnen und -didaktiker sind sich darin einig, dass im Geschichtsunterricht nicht primär „Wissen über die Vergangenheit“ vermittelt werden soll, sondern dass es vielmehr darum geht, grundlegende Kompetenzen für den Umgang mit Vergangenheit/Geschichte aufzubauen (vgl. z.B. für die Vereinigten Staaten: Stearns, 1998; Wineburg, 2001; Mandell, 2008; VanSledright, 2014; für Kanada: Seixas, 2008; Seixas & Morton, 2013; für Europa: Erdmann & Hasberg, 2011; für Australien: Taylor & Young, 2003). Trotz unterschiedlicher Bildungstraditionen ist „Geschichte denken statt pauken“ (Schreiber & Mebus, 2005) das verbindende Element eines Großteils der nationalen und internationalen Diskussion über die Ziele des historischen Lernens bzw. der „history education“ (Köster, Thünemann & Zülsdorf-Kersting, 2014).

Während hinsichtlich der Ziele weitgehende Einigkeit in der Geschichtsdidaktik herrscht, bestimmen Divergenzen die Diskussion, wenn es um die Überprüfung geht, inwieweit Schülerinnen und Schüler Kompetenzen historischen Denkens tatsächlich erwerben. Zur Klärung müssen verschiedene mögliche Zwecke dieser Überprüfung unterschieden werden. Um nur einige zu nennen: Sollen die Kompetenzentwicklungen Einzelner erfasst werden (Diagnostik auf Einzelfallebene) oder interessiert die Verteilung von Kompetenzausprägungen in und zwischen Gruppen (Assessment-Studien)? Sollen in evaluativer Absicht Ergebnisse konkreter Lernprozesse untersucht oder sollen Hinweise für weiterführende Entscheidungen gewonnen werden? Sollen in den Aufgaben zur Bestimmung der Kompetenzausprägungen vornehmlich die im

Unterricht thematisierten Lerninhalte aufgegriffen werden oder vor allem neue Inhalte und Probleme, um so vorrangig transferable Fähigkeiten in den Blick zu nehmen?

Notwendig ist auch eine Positionierung zur Frage, inwiefern historische Kompetenzen im Rahmen von vergleichenden Schulleistungsstudien überhaupt erfasst werden können. Während z.B. in Deutschland der Verband der Geschichtslehrer dies prinzipiell in Frage stellt (Verband der Geschichtslehrer, 2011), werden im internationalen Rahmen Möglichkeiten von sogenannten *Large-Scale-Assessments* (LSA) auch im Fach Geschichte erprobt (z.B. Ercikan & Seixas, 2015; Kühberger, 2013). Einzuordnen sind solche Vergleichsstudien in den grundlegenden Paradigmenwechsel der Bildungspolitik, der auch den deutschsprachigen Raum erfasste. Die einflussreiche „Klieme-Expertise" (Klieme et al., 2003), mit der in Deutschland auf den sogenannten PISA-Schock, also die schlechten Ergebnisse deutscher Schülerinnen und Schüler im internationalen Vergleich, reagiert wurde, schlug die Einführung nationaler Bildungsstandards vor. Sie sollen die föderalen Bildungssysteme darauf verpflichten, „den Aufbau von Kompetenzen, Qualifikationen, Wissensstrukturen, Einstellungen, Überzeugungen, Werthaltungen […]" zu unterstützen, um so „die Basis für ein lebenslanges Lernen zur persönlichen Weiterentwicklung und gesellschaftlichen Beteiligung" (S. 12f.) zu legen. Unter „Kompetenz" werden dabei die „bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten" bezeichnet, die nötig sind, „um bestimmte Probleme zu lösen", „sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können" (Weinert, 2001, S. 27–28).

Bildungsstandards und (kompetenzorientierte) Bildungspläne zu setzen, ist letztlich ein politischer Akt, welcher normative Entscheidungen über gesellschaftliche Ziele von Bildung und Erziehung ebenso umfasst wie die Verpflichtung, die für ihre Erreichung nötigen Rahmenbedingungen und Ressourcen bereit zu stellen. Allerdings sind solche Entscheidungen nur dann verantwortlich zu treffen, wenn die angesprochenen Konzepte tragfähig sind. In anderen Worten: Es bedarf solider Kompetenzmodelle und der Möglichkeit, die darin beschriebenen Kompetenzen reliabel und valide zu erfassen. Dies wiederum ist eine Aufgabe der Wissenschaft. Da in Deutschland gemäß den Vorgaben der Kultusministerkonferenz (KMK) die Bildungsstandards auch den Besonderheiten fachlichen Lernens Rechnung tragen sollen, sind insbesondere die Fachdidaktiken gefordert, auf Grundlage der domänenspezifischen Erkenntnisse tragfähige Kompetenzmodelle zu entwickeln.

Zunächst wurden, finanziert von der KMK und unter Federführung des Instituts zur Qualitätssicherung im Bildungswesen (IQB), Bildungsstandards sowie, damit verbunden, Kompetenzmodelle und Instrumente zu deren Überprüfung nur für die „PISA-Fächer" (Unterrichtssprache, 1. Fremdsprache, Mathematik, Naturwissenschaften) entwickelt. Getragen von der Sorge, ohne derartige Konzeptionen zu Fächern „zweiter Güte" zu werden (Sachse, 2005), fand in einigen weiteren Fächern die Entwicklung von Kompetenzmodellen und Bildungsstandards ohne diesen zentralen Auftrag und ohne die damit verbundene Finanzausstattung statt. Für das Fach Ge-

schichte entstanden in der Folgezeit im Rahmen von geschichtsdidaktischer Reflexion und Forschung, aber auch im Rahmen von Verbandstätigkeit oder Lehrplanarbeit der Bildungsverwaltungen eine Reihe von z.T. deutlich unterschiedlich ausgerichteten Kompetenzmodellen (Vergleiche finden sich in Barricelli, Gautschi & Körber, 2012; Körber, 2007b; Mayer, 2014; für die Entwicklung im anglophonen Bereich siehe Eisele-Brauch, 2010).

Eine forschungspolitische Entscheidung war es, im Rahmen von zentralen Programmen der Deutschen Forschungsgemeinschaft (DFG) oder des Bundesministeriums für Bildung und Forschung (BMBF) zur Kompetenzmodellierung und -erfassung in der Folgezeit auch Fächer zu adressieren, die in PISA und vergleichbaren Untersuchungen nicht vertreten waren. Hinter diesen Programmen steht als übergeordnetes Ziel der Bildungspolitik, generalisierbares Wissen über Bildungsprozesse und ihre Rahmenbedingungen zu gewinnen, um darauf aufbauend bildungsrelevante Entscheidungen und Reformen im Bildungssystem treffen zu können (vgl. hierzu das Rahmenprogramm zur Förderung der Empirischen Bildungsforschung (www.empirische-bildungsforschung-bmbf.de) und das Programm zur „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor" des BMBF (http://www.kompetenzen-im-hochschulsektor.de/) sowie das Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" der DFG).

Ähnliche Programme gab es auch für die Schweiz und für Österreich. In der Schweiz wurden im Kontext der gesetzlich geforderten „Harmonisierung" der Bildungsziele für die Fächer Schulsprache, erste Fremdsprache, Mathematik und Naturwissenschaften von der Schweizerischen Erziehungsdirektorenkonferenz (EDK) Forschungsprojekte finanziert, um sogenannte Basiskompetenzen zu erarbeiten (http://www.edk.ch/dyn/12930.php). In Österreich kam es ähnlich wie in Deutschland zu breit angelegten Bildungsstandardüberprüfungen in Deutsch, Englisch und Mathematik (www.biefie.at). Allerdings blieben auch die weiteren Fächer(-kombinationen) aufgrund einer Arbeit an neuen kompetenzorientierten Maturaformaten und Lehrplänen sowie an einer neuen Leistungsprüfungsverordnung im bildungspolitischen Fokus.

Im Kontext „empirischer Kompetenzmodellierung und -erfassung" ist das hier vorgestellte interdisziplinäre Projekt „HiTCH: Historical Thinking – Competencies in History" zu verorten. Das Ziel des HiTCH-Projekts bestand darin, einen Test zu entwickeln, mit dem sich die Kompetenzen historischen Denkens von Schülerinnen und Schülern in Large-Scale-Assessments erfassen lassen. Damit verbunden war die Absicht, eine Grundlage für die Berücksichtigung der Domäne Geschichte in großen Schulleistungsstudien zu legen. Im Anschluss an die allgemeine Kompetenzdefinition Weinerts (2001) und an den Diskussionsstand der Geschichtsdidaktik (Barricelli et al., 2012) werden dabei unter Kompetenzen historischen Denkens domänenspezifische, nicht aber an konkrete, zuvor unterrichtete Gegenstände gebundene Fähigkeiten, Fertigkeiten und Bereitschaften verstanden.

Ein wesentlicher Fokus der Arbeit im HiTCH-Projekt lag auf der Entwicklung entsprechender Aufgabenformate für die Kompetenztestung. Die Herausforderung bestand darin,

- dass die Aufgaben üblichen psychometrischen Gütekriterien entsprechen müssen, also objektiv, reliabel und valide sein müssen;
- dass sie fachliche, geschichtsdidaktische und geschichtstheoretische Anforderungen erfüllen müssen, dass sie also z.B. die zentralen Kompetenzbereiche adressieren müssen, dass ihre Konstruktionslogik auf ganz unterschiedliche historische Gegenstände anpassbar sein muss und dass sie nicht mechanisch oder durch „Ersatzhandlungen", wie etwa nur durch sinnentnehmendes Lesen, bewältigt werden können;
- dass die Aufgaben sowohl in einem überschaubaren zeitlichen Rahmen zu bearbeiten als auch auszuwerten sein müssen.

Die Testentwicklung erforderte eine enge interdisziplinäre Zusammenarbeit von Expertinnen und Experten der Empirischen Bildungsforschung und der Geschichtsdidaktik.[1] Grundlage war die Berücksichtigung der international anerkannten Standards Empirischer Bildungsforschung, bei gleichzeitiger Fundierung des Tests auf einer theoretisch begründeten Definition historischen Denkens und der dazu nötigen Kompetenzen. Aufgrund der Anschlussfähigkeit an die nationale und internationale Auseinandersetzung mit historischen Kompetenzen wurde das Kompetenz-Strukturmodell der FUER-Gruppe (FUER-Modell) als Theoriebasis gewählt (vgl. Kapitel 2.3)

Die für die Testentwicklung nötige Forschungs- und Entwicklungsarbeit wäre ohne die finanzielle Unterstützung durch das Bundesministerium für Bildung und Forschung[2] an die Universitäten Tübingen (Ulrich Trautwein), Eichstätt-Ingolstadt (Waltraud Schreiber), Hamburg (Bodo von Borries, Andreas Körber) sowie das Leibniz-Institut für Wissensmedien Tübingen (Stephan Schwan) nicht möglich gewesen. Als Konsortialpartnerinnen und -partner schlossen sich Nicola Brauch (Universität Bochum), Johannes Meyer-Hamme (Universität Paderborn), Christoph Kühberger (Pädagogische Hochschule Salzburg) sowie Béatrice Ziegler (Pädagogische Hochschule Nordwestschweiz) dem HiTCH-Projekt an. Christiane Bertram, Matthias Hirsch, Kathrin Klausmeier, Martin Merkt, Herbert Neureiter, Wolfgang Wagner, Monika Waldis, Michael Werner und Andreas Zuckowski haben als wissenschaft-

1 Der Begriff der „Empirischen Bildungsforschung" hat in den vergangenen Jahren unterschiedliche Konnotationen erhalten. So wird die Empirische Bildungsforschung als ein interdisziplinäres Forschungsfeld konzipiert, in das sich etablierte Disziplinen wie die Erziehungswissenschaft, Psychologie, Soziologie sowie die Fachdidaktiken einbringen, ohne ihre angestammte Disziplin zu verlassen. Darüber hinaus gibt es jedoch auch eine wachsende Zahl von Wissenschaftlerinnen und Wissenschaftlern, die sich primär als „Empirische Bildungsforscherinnen oder -forscher" definieren und damit den Kern einer neuen Disziplin bilden. Wenn im Folgenden von der Zusammenarbeit zwischen „Empirischer Bildungsforschung" und Fachdidaktik die Rede ist, so liegt dem das letztgenannte Verständnis von Empirischer Bildungsforschung zugrunde.

2 Die Förderung erfolgte im Rahmen des BMBF-Programms zur Förderung von Forschungsvorhaben in Ankopplung an Large-Scale-Assessments unter der Fördernummer LSA006.

liche Mitarbeiterinnen und Mitarbeiter einen wesentlichen Anteil am Gelingen des Projekts. Bewusst wollen wir auch die sehr engagierten wissenschaftlichen Hilfskräfte namentlich aufführen, die zum Teil sehr eigenständige Beiträge in die Arbeit eingebracht haben: Anna Benning, Hanna Großmann, Lisa Henke, Johannes Hoffmann, Tobias Mall, Franziska Meis, Juliana Schwerdtfeger, Franziska Singh und Julia Wittig. Das Projektteam von insgesamt sieben Hochschulen in Deutschland, Schweiz und Österreich wurde zwischenzeitlich überführt in ein Konsortium, das die Arbeit an und mit dem HiTCH-Test weiterführen wird (Informationen hierzu: www.hitch-projekt.de).

Für das HiTCH-Projektteam markiert die vorliegende Veröffentlichung einen ersten Meilenstein. Sie beschreibt die theoretischen Prämissen, die methodische Vorgehensweise und die Ergebnisse der geleisteten Entwicklungsarbeit für einen Large-Scale-Test in Geschichte, diskutiert die Bedeutung für die Empirische Bildungsforschung und die Geschichtsdidaktik, markiert weiteren Forschungsbedarf, aber auch die Relevanz für die Weiterentwicklung des Geschichtsunterrichts. Der vorliegende Band richtet sich an Bildungsforscherinnen und -forscher sowie Kolleginnen und Kollegen der Fachdidaktik, aber auch an die Lehrerbildung in allen Phasen, an Fachleute in der Bildungsadministration, an Lehrerinnen und Lehrer sowie an Studierende. Weil der HiTCH-Test in interdisziplinärer Kooperation entstanden ist, wird die Fachterminologie beider Disziplinen verwendet. Manches wird der einen Lesergruppe stärker und der anderen weniger erläuterungsbedürftig erscheinen. Wir haben uns bemüht, die Konzepte zugleich verständlich und fachlich präzise darzustellen.

In diesem Band geht es um die Vorstellung des Weges hin zu einem Test zur quantitativen Erfassung historischer Kompetenzen. Dafür werden in Kapitel 2 ausgehend von den geschichtstheoretischen Prämissen die grundlegenden Ziele des Geschichtsunterrichts und die dort zu erlernenden Kompetenzen skizziert. Anschließend werden die Herausforderungen bei der empirischen Erfassung historischer Kompetenzen herausgearbeitet und bisherige Vorgehensweisen sowie der Forschungsstand skizziert. Kapitel 3 beschreibt auf dieser Grundlage die Entwicklung eines historischen Kompetenztests für Large-Scale-Assessments. Die durchlaufenen Phasen des HiTCH-Tests werden kurz skizziert. Die Methodendarstellung (Kapitel 4) rückt die in der Haupterhebung eingesetzten Instrumente und die gewählten Analysestrategien ins Zentrum. Die Ergebnisse (Kapitel 5) werden dann im Hinblick auf die interne Struktur der Daten und die Validität des Instruments in Abgrenzung zu verwandten Konstrukten (beispielsweise Lesekompetenz) betrachtet. Die Publikation schließt mit einer Diskussion der Ergebnisse sowie mit einem Ausblick auf Forschungsdesiderate (Kapitel 6).

2 Theoretische Grundlagen: Historische Kompetenzen, Kompetenzmodelle und Kompetenzmessung im Fach Geschichte

Kompetenztests sollen objektiv, reliabel und valide sein. Das bedeutet, dass das Ergebnis des Tests unabhängig davon sein soll, wer den Test durchführt bzw. auswertet und auch andere Rahmenbedingungen keine Rolle spielen (Kriterium der Objektivität). Zudem sollen die Tests auch bei wiederholter Messung unter nur leichten, den Rahmen des Üblichen nicht verlassenden Abweichungen nicht nur verlässlich zu den gleichen Ergebnissen kommen (Kriterium der Reliabilität), sondern tatsächlich auch das erfassen, was gemessen werden soll (Kriterium der Validität). Deshalb müssen die entsprechenden Kompetenzen zuvor theoretisch definiert und begründet sein. Das gilt auch für die Beziehung der zu messenden Größe zu anderen, verwandten oder zu unterscheidenden Größen wie auch für die Beziehung von zu messenden Aspekten zueinander.

In der deutschsprachigen Debatte hat sich für die Definition einer Kompetenz sowie für die Darstellung der Zusammenhänge der Kompetenzen das Format „Kompetenzmodell" durchgesetzt, wobei allerdings die einzelnen Kompetenzmodelle in ihrem Anspruch, ihrer Reichweite und ihrem Elaborationsgrad sehr unterschiedlich ausfallen. Oft graphisch unterstützt, begründen und beschreiben Kompetenzmodelle theoretisch und systematisch, welchen spezifischen Beitrag zu einer allgemeinen Bildung ein Fach leisten soll, welche besonderen Fähigkeiten, Fertigkeiten und Bereitschaften in ihm gefördert und wie Qualitätsausprägungen solcher Kompetenzen unterschieden werden sollen. Auch die Anteile allgemeiner („generischer") Kompetenzen in den fachspezifischen Anforderungen werden in ihnen reflektiert (Körber, 2007a, 2012). Ein Kompetenzmodell historischen Denkens soll idealerweise definieren, was unter *historischem Denken* zu verstehen ist, wozu dieses dient, welche Operationen und Teilschritte dabei zu unterscheiden sind und welche Fähigkeiten, Fertigkeiten und Bereitschaften zu seinem Vollzug nötig sind.

Im Folgenden werden Grundlagen der geschichtsdidaktischen Kompetenzmodellierung herausgearbeitet. Dies geschieht unter Bezugnahme auf das FUER-Kompetenzmodell (Körber, Schreiber & Schöner, 2007), wobei Anschlüsse zu vergleichbaren bzw. partiell verwandten Modellen diskutiert werden. Ein sehr grundlegender Unterschied des FUER-Modells zu den meisten Modellen ist, dass diese von einer auf schulisches Lernen ausgerichteten Form historischen Denkens ausgehen und das Verfügen-Können über ein curricular vorgegebenes Geschichtsverständnis oder ein Denken-Können im schulischen Rahmen fokussieren. Das FUER-Modell dagegen geht von der grundsätzlichen Orientierungsfunktion aus, die historisches Denken für Individuen und Gesellschaften hat. In diesem größeren Kontext wird Schule als der Ort angesehen, an dem das Historisch-Denken-Lernen explizit gefördert wird. Die narrativistische Geschichtstheorie (vgl. das nachfolgende Kapitel 2.1) findet im FUER-Modell eine explizitere Beachtung als in anderen Modellen, wobei auch diese sich über ein narratives Geschichtsverständnis fundieren. Auf die Narrativitätstheorie aufbauend wird in Kapitel 2.2.3 historisches Denken näher erschlossen, u.a. auch in der

spezifischen Fokussierung des historischen Lernens als Historisch-Denken-Lernen (Kapitel 2.2.3.7).

2.1 Die narrativistische Geschichtstheorie – Grundlage aktueller Geschichtsdidaktik

Die narrativistische Geschichtstheorie ist ein Ergebnis wissenschaftsphilosophischer wie fachdidaktischer Reflexionen und Forschungen zur Funktion von Geschichte für das Leben der Individuen und menschlicher Gesellschaften, zu den Bedingungen und Möglichkeiten, Prinzipien und Verfahren historischer Erkenntnis und zu den Formen und Funktionen historischen Wissens. Die Analytische Philosophie der Disziplin Geschichtswissenschaft, vornehmlich in der Variante von Arthur C. Danto (1965; USA), Paul Ricœur (1988/1983; Frankreich) sowie Hans-Michael Baumgartner (1975, 1997) und Jörn Rüsen (1982, 1983, 1989, 1994, 2008, 2013; jeweils Deutschland) identifizierte die Funktion von Geschichte für Individuen und Gesellschaften in der zeitlichen Orientierung lebensweltlicher Identitäten und Handlungen und arbeitete die hierfür geltenden erkenntnistheoretischen Rahmenbedingungen heraus. Im Zuge dieser Theorieentwicklung wird zwischen *Vergangenheit* und *Geschichte* unterschieden. Vergangenheit steht für die Wirklichkeit früherer Zeiten, die nicht als solche erfassbar ist, und Geschichte für diejenige grundsätzlich narrative Form, in welcher Vergangenes, als Ergebnis historischer Denkprozesse, allein dargestellt werden kann. Dies bedeutete zum einen eine Absetzung von der älteren Vorstellung von Geschichte als eines feststehenden, wenn auch erst allmählich im Zuge von Forschung immer besser bekannten Bestandes von Wissen über die Vergangenheit. Zum anderen verbindet sich damit die grundsätzlich hinterfragende Haltung gegenüber jeglicher Art verbindlicher, identitätsstiftender Masternarrative, wie sie lange Zeit in der Nationalgeschichtsschreibung vorherrschend waren und zum Teil noch immer sind. An deren Stelle trat die Anerkennung einer prinzipiellen Mehrzahl konkurrierender und dennoch triftiger Narrationen (Jörn Rüsen) über gleiche Gegenstände bei gleichzeitiger Ablehnung eines theorie- und methodenlosen Relativismus. Dem entsprach die Anerkennung der auch gesellschaftlich-kulturellen Gebundenheit und Funktion allen historischen Denkens. Hieraus wurde der Bedarf nach strukturellen Qualitätskriterien für historisch triftige Aussagen abgeleitet (u.a. Rüsen 1983, 1986, 1989, 2013).

Als konstitutive Merkmale aller „Geschichten", also als epistemologische Prinzipien, sind identifiziert worden:

- *Narrativität,* d.h. dass in einer historischen „Narration", ausgehend von einer Fragestellung, mindestens zwei verschiedene, zeitlich differente Gegebenheiten oder Begebenheiten sinnhaft miteinander verknüpft werden, sodass eine sprachlich vermittelte Verlaufsstruktur entsteht (Brauch, 2015; Rüsen, 1983, 2013). Narrativität wird also strukturell verstanden und so gegenüber belletristischen, wesentlich fiktionalen bzw. fiktiven Erzählungen abgegrenzt.

- *Perspektivität,* d.h. die Standortgebundenheit sowohl des Geschichtsdenkenden, also auch der Urheberin und des Urhebers von Quellen und der Rezipientin und des Rezipienten historischer Narration sowie die damit verbundene Perspektivierung. Damit ist die jeweilige Bedeutungszuschreibung gemeint, die aus dem Zusammenhang von Ereignissen eine orientierungsbietende Narration macht. Die einzelnen Perspektiven sind grundsätzlich partikular. Deshalb steigert es die Qualität historischen Denkens, jeweils mehrere Perspektiven zu berücksichtigen und in Bezug zu einander zu setzen, wobei eine Vollständigkeit aller Perspektiven nicht denkbar ist.
- *Retrospektivität,* d.h., die auf vergangene Ge- und Begebenheiten *rückblickende* Entstehung jeglicher historischen Narration. Retrospektivität ist gekennzeichnet von einer größeren (zeitlichen) Distanz vom jeweiligen Ereignis- und Zusammenhangskomplex und der damit einhergehenden prinzipiellen Informiertheit über spätere Entwicklungen. Die dabei eingenommene Retro-Perspektive unterscheidet sich damit unhintergehbar von den Perspektiven der Zeitgenossen; ebenso unhintergehbar ist ihre eigene Gegenwartsgebundenheit. Im gegenwartsgebundenen und retrospektiv gewonnenen Wissen über Ge- und Begebenheiten, über Zusammenhänge der Vergangenheit, über die Perspektiven damaliger Zeitgenossinnen und Zeitgenossen ist zugleich die Möglichkeit angelegt, daraus Orientierung für Gegenwart und Zukunft zu gewinnen (vgl. auch Ventzke, 2016).
- die Prinzipien der *Partikularität* und der *Selektivität*, d.h. die Unmöglichkeit, eine wie auch immer definierte vergangene Wirklichkeit *in toto* in einer Geschichte zu fassen. Partikularität steht dabei für fehlende Spuren vergangenen Geschehens ebenso wie für die Lücken der Überlieferung, die durch zufällige Verluste wie durch systematische Filterung, nicht zuletzt bei der Archivierung, entstanden sein können. Selektivität ist abhängig vom historisch Denkenden und ist z.B. bestimmt durch dessen Fragestellung, die jeweiligen individuellen wie kollektiven Perspektiven oder das jeweilige inhaltliche, methodische, theoretische Vorwissen.
- *Konstruktivität,* d.h. die unhintergehbare Notwendigkeit, dass Geschichte immer von jemandem hergestellt wird, dass also auf Überlieferung beruhende Einzelheiten („Vergangenheitspartikel“ [3]) aktiv zu einem stimmigen narrativen Zusammenhang gefügt werden. Die Bindung jeder historischen Narration an ihre Konstrukteurin und ihren Konstrukteur betont die Abhängigkeit z.B. von Fragestellungen, von individuellen und gesellschaftlichen Zusammenhängen, von Interessen an der Bildung historischen Sinns. Zugleich verweist das Prinzip der Konstruktivität auf den Bedarf an Kriterien für die Qualität historischer Aussagen und Orientierungen sowie der dazugehörigen Prozesse (vgl. unten Triftigkeiten, Kapitel 2.2.3.6).

3 Die Anführungszeichen bei „Vergangenheitspartikel“ verweisen darauf, dass diese nicht eine vergangene Wirklichkeit abbilden, sondern notwendig perspektivisch gebrochen sind, etwa durch die Autorin oder den Autor der Quelle oder den sich mit der Quelle Befassenden.

2.2 Didaktische Wendungen

2.2.1 Geschichtsbewusstsein

Aus dem Fundament der narrativistischen Geschichtstheorie folgt die Berücksichtigung nicht-akademischen historischen Denkens und außerwissenschaftlicher historischer Erzählungen als relevante Ausprägungen des grundlegenden Prozesses historischer Orientierung. Dies aber zieht notwendig geschichtsdidaktisch relevante Konsequenzen nach sich. In Deutschland wurde, u.a. von Karl-Ernst Jeismann, Jörn Rüsen, Hans-Jürgen Pandel und Bodo von Borries in den 1970er Jahren das Konzept des *Geschichtsbewusstseins* entwickelt, verstanden als Komplex von Dispositionen, Prozessen und Fähigkeiten im und zum „Umgang" mit Geschichte. Auf der Basis der skizzierten narrativistischen Geschichtstheorie wurden zum einen geschichtsdidaktische Prinzipien zur Erforschung des Geschichtsbewusstseins von Individuen und Gruppen in der Gesellschaft formuliert, zum anderen Prinzipien eines auf ein entwickeltes, reflektiertes und (selbst-)reflexives Geschichtsbewusstsein zielenden Lernens. Unter anderem über die Rezeption der englischsprachigen Arbeiten von Jörn Rüsen (zusammengefasst in Rüsen, 2005) wurde das Konzept der *Historical Consciousness* auch im angelsächsischen (Seixas, 2004) und skandinavischen Raum verbreitet (Jensen, 2003; Eliasson, Alvén, Axelsson Yngvéus, & Rosenlund, 2015).[4] Das betrifft auch die Implementierung in Schulcurricula der westlichen Welt (Brauch, in press).

Nicht erst aus der narrativistischen Geschichtstheorie, sondern bereits aus ihrem historistischen Vorläufer stammt die in der deutschen Geschichtswissenschaft zentrale Unterscheidung von *Quelle* und *Darstellung* als zweier grundsätzlicher, hinsichtlich ihres erkenntnistheoretischen Status zu unterscheidender Materialgattungen. Sie wurde im Rahmen der narrativistischen Geschichtstheorie sowohl verfeinert als auch erweitert, so etwa mit Blick auf „narrative Quellen" oder auf Zeitzeugenaussagen, die auf spezielle Weise Aspekte von Quelle und Darstellung vereinen (Schreiber & Árkossy, 2009).

In einem kompetenzorientierten Geschichtsunterricht sollen Schülerinnen und Schüler einerseits verschiedene Gattungen von Quellen und Darstellungen zu unterscheiden lernen (vgl. u.a. die Ansätze bei Pandel, van Boxtel, van Drie oder Wineburg), andererseits sollen sie verstehen, dass es bei einer Reihe von Materialien von der Fragestellung abhängt, ob sie als Quellen oder Darstellungen genutzt werden. Dies wird insbesondere im FUER-Modell betont. Es macht also keinen Sinn, eine quasi-dichotome, idealtypische Unterscheidung zwischen diesen Materialkategorien zu treffen. Es geht vielmehr darum, die Einsichten, die in dieser Unterscheidung und den idealen Quellen- und Darstellungstypen gefasst werden, im Erkenntnispro-

4 Es steht dort aber, insbesondere in Bezug auf schulisches Lernen in Konkurrenz zu anderen Ansätzen der Operationalisierung „historischen Denkens", etwa aus dem englischen *Schools Council History Project* (SCHP), der US-amerikanischen Stanford History Education Group (Wineburg, 2001) und der kanadischen Gruppe um Peter Seixas (Seixas, 2011, 2015).

zess des historischen Denkens nutzbar zu machen (Ziegler, 2007; Seixas, 2015; Körber, 2016).

2.2.2 Didaktische Konsequenz der Perspektivität: Multiperspektivität

Perspektivität als Prinzip historischer Erkenntnis ist bereits angesprochen worden. Didaktisch ist aus ihr die Konsequenz gezogen worden, aus mindestens zwei Gründen immer die Berücksichtigung mehrerer relevanter Perspektiven zu fordern: Zum einen wird so am eindeutigsten der Gefahr einer Indoktrination oder gar Überwältigung – und somit quasi einem Verstoß gegen den „Beutelsbacher Konsens" (Schiele & Schneider, 1977; Buchstein, Frech & Pohl, 2016) entgegengewirkt. Gemeint ist, dass Lernende nicht zur einfachen Übernahme oder Konstruktion von alternativlos scheinenden Deutungen und Wertungen verleitet werden, zu denen sie sich selbst nicht mehr kritisch-distanzierend verhalten können. Multiperspektivität in diesem Sinne ist die Voraussetzung dafür, Geschichte selber denken zu können (Bergmann, 2000). Zum anderen erfordert das Selber-Denken-(Können) auch, die Unhintergehbarkeit der Perspektivität als Charakteristikum des Historischen zu berücksichtigen, Perspektivität also nicht als ‚leider unvermeidbare Verzerrung' anzusehen, sondern gewissermaßen als Teil des historischen Universums selbst.

Aus beiden Gründen ist es sinnvoll, drei „Ebenen" zu differenzieren, in denen sich Perspektivität unterschiedlich findet (u.a. Borries, 2004 oder Stradling, 2004) und die im Folgenden beschrieben werden.

2.2.2.1 Perspektivität auf der Ebene der Quellen

Alle Quellen repräsentieren bestimmte Perspektiven, u.a. weil sich in ihnen beispielsweise der kulturelle, soziale und politische Standort der Verfasserin oder des Verfassers und dessen Intentionen widerspiegeln. Daraus ergibt sich die Forderung, historische Sachverhalte anhand jeweiliger Quellen aus den als relevant herausgearbeiteten unterschiedlichen Perspektiven zu erforschen („Multiperspektivität im engeren Sinne"; Borries, 2004). Aus diesem Grunde dürfen die Perspektivitäten der Wahrnehmungen, Prägungen und Bedeutungszuweisungen zu vergangenen Be- und Gegebenheiten, wie sie sich aus den Quellen erschließen lassen, nicht als äußerliche ‚Verzerrungen' einer dahinter liegenden „Wahrheit" missverstanden werden. Sie sind ihrerseits ein zu erforschender Aspekt der nur auf der Basis von Quellen zugänglichen „damaligen Situation".

2.2.2.2 Perspektivität auf der Ebene der historischen Darstellungen

Die Einsicht in den perspektivischen – d.h. partikular-einschränkenden, aber auch bedeutungs-konstituierenden – Charakter aller historischen Darstellungen schafft die Grundlage für den de-konstruierenden Umgang mit Narrationen fremder Urheberinnen und Urheber aus der aktuellen wie aus vergangenen Gegenwarten. Die Einsicht hat zugleich auch Rückwirkungen auf die Erarbeitung und Gestaltung eigener historischer Narrationen.

Erst auf dieser Grundlage kann ein historisch Denkender historische Darstellungen de-konstruierend sowohl auf die Fragestellungen, den Quellenbezug, die Interessen und deutungsleitenden Normen und Konzepte des Autors befragen als auch ihr Verhältnis zu den eigenen Fragen, Interessen und Normen etc. klären. Auch hier gilt, dass die Zusammenschau und der Vergleich voneinander abweichender („kontroverser"; Borries, 2004) Perspektiven historischer Darstellungen weder Vollständigkeit in Bezug auf das betrachtete Vergangene ermöglicht, noch einen allgemeinverbindlichen Anspruch in Bezug auf die zeitspezifischen Deutungen und Sinngebungen.

2.2.2.3 Perspektivität auf der Ebene der Rezipientinnen und Rezipienten

Schließlich folgt aus der Einsicht in die allgemeine Perspektivität historischen Denkens auch diejenige in die Perspektivität des heutigen Fragens, Re- und De-Konstruierens, Deutens, Wertens und also Orientierens, des eigenen und des aller anderen (vgl. Borries, 2004: Pluralität der sinnbildenden Schlussfolgerungen wie der Wertungen über vergangene Ge- und Begebenheiten).

Qualitätssteigernd für das historische Denken ist damit die Einsicht, dass es weder eine einzige Geschichte zu einem beliebigen historischen Zusammenhang geben kann, noch dass es auch nur sinnvoll wäre, sich für die sozial, kulturell, politisch und individuell verschieden positionierten Menschen auf eine Geschichtsdeutung mit universalem Anspruch (Masternarrativ) und ihre Vermittlung zu einigen. Ebenso ergibt sich aus der Einsicht in das Prinzip der Perspektivität auf Rezipientinnen- und Rezipientenebene, dass es bedeutsam und möglich ist, sich über die unterschiedlichen („pluralen") Geschichtsdeutungen und ihre Orientierungsleistungen auszutauschen – ohne gegenseitige Abwertung oder gar Gewalt. Aus der „Pluralität" historischen Erkennens und der „Kontroversität historischer Narrationen" folgt aber auch nicht die Notwendigkeit einer umstandslosen Anerkennung jeglicher Deutungen und Sinnbildungen, sondern die gesellschaftliche und auch fachliche Notwendigkeit ihrer Diskussion. Notwendig ist also der Austausch über gemeinsame Anforderungen, etwa hinsichtlich der „Triftigkeiten" der jeweiligen Narrationen (vgl. Kapitel 2.2.3.6).

2.2.3 Historisches Denken – der Prozess hinter einem narrativen Geschichtsverständnis

2.2.3.1 Historisches Denken: anthropologisch-funktional

Die anthropologisch-funktionale Begründung historischen Denkens besteht in der immer wieder neu zu leistenden zeitlichen Orientierung, die der Bewältigung erfahrener Kontingenz dient (Rüsen, 1983, 2013). Jörn Rüsen hat die entsprechenden theoretischen Konzepte und Prinzipien zunächst am Beispiel der wissenschaftlichen Geschichtsforschung entwickelt und im zugehörigen zirkulären Modell der *disziplinären Matrix* verdeutlicht.

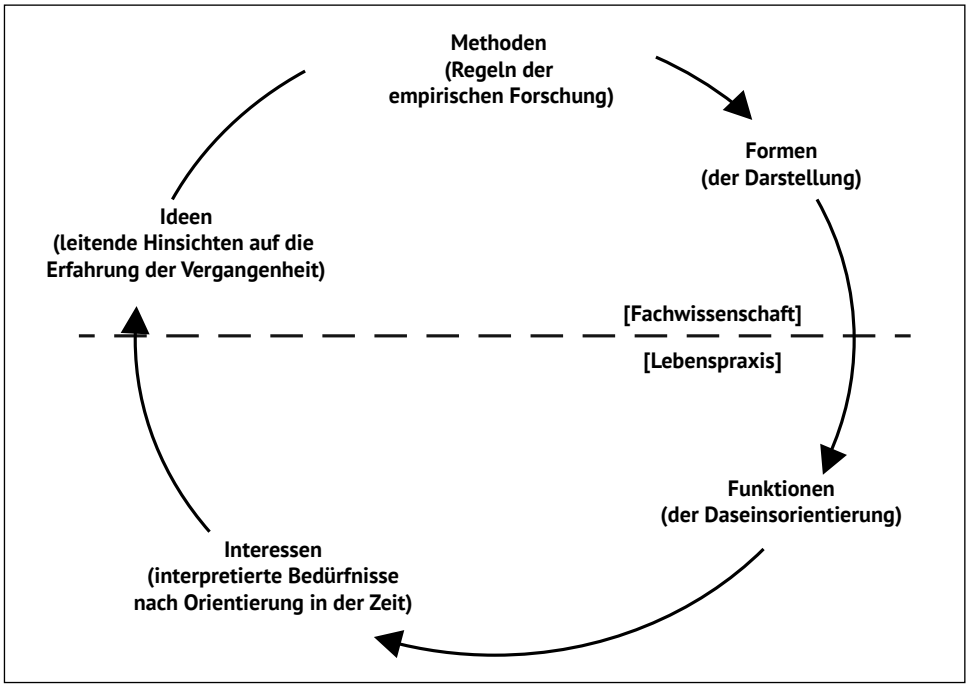


Abbildung 1: Rüsen, J.: Historische Vernunft, 1983, S. 26.

Die Matrix wurde bald auch für schulisches Geschichtslernen (Rüsen, 2001; Seixas, 2016) und generell für historisches Denken (Körber, 2007a; Hasberg & Körber, 2003) adaptiert und ist eine wesentliche Grundlage für die Theorie historischen Denkens, die ihrerseits die Basis der Kompetenzmodellierung bildet, wie sie u.a. von der FUER-Gruppe ausdifferenziert wurde (vgl. Kapitel 2.3).

2.2.3.2 Historisches Denken: Theoriebildung

Historisches Denken wird verstanden als eine von allen Menschen notwendig vollzogene Operation, die z.B. zur Klärung der eigenen individuellen und kollektiven Identität und der Orientierung des eigenen Handelns in zeitlicher Hinsicht dient. Dem historischen Denken liegt ein Komplex an Fähigkeiten, Fertigkeiten und Bereitschaften zugrunde, die der Wissenschafts- und „Alltagswelt" gemeinsam und allen verfügbar sind. Gleichzeitig sind die Kompetenzen historischen Denkens aber einer Entwicklung und Rationalisierung zugänglich, so dass Kompetenzen in sehr unterschiedlichen Ausprägungen vorliegen können.

Damit verbunden ist die Forderung nach der Anerkennung der Möglichkeit und Wirklichkeit unterschiedlicher Verarbeitungen. In pluralen Gesellschaften geht es um die gegenseitige Anerkennung unterschiedlicher, von eigenen Positionen abweichenden Verarbeitungen und Resultate, sofern diese Triftigkeit (vgl. Kapitel 2.2.3.6) für sich beanspruchen können.

2.2.3.3 Historisches Denken im gesellschaftlichen Kommunikationszusammenhang

Geschichtsdidaktisch gewendet wird auch die Einsicht, dass alles historische Denken nicht allein individuell ist, sondern sowohl in einem gesellschaftlichen Rahmen als auch in einer gesellschaftlichen Kommunikation stattfindet (Röttgers, 1982). Mit dem Konzept der *Geschichtskultur* wird in der didaktischen Theoriebildung der gesellschaftlich feststellbare Umgang mit Geschichte bezeichnet, der die Sinnbildungsprozesse der Individuen rahmt und dem historischen Denken der Einzelnen sowohl Denkräume eröffnet als auch deren Wahrnehmung und Äußerungen beeinflusst.

Aus der Eigenschaft historischer Orientierungsleistungen, nicht nur für das jeweilige Individuum Geltung und Orientierungskraft zu besitzen, sondern in gesellschaftlicher Kommunikation auch für spezifische Gruppen Relevanz zu gewinnen, ergibt sich sowohl ein Interesse des Einzelnen an (zumindest partieller) Übereinstimmung der eigenen orientierenden Geschichtssicht mit derjenigen der anderen, als auch ein Anspruch der Gesellschaft als Ganzer an einem Mindestmaß an Übereinstimmung. Es muss dabei nicht um konkrete Orientierungen gehen, zumindest aber um auf den Prozess des historischen Denkens bezogene Fähigkeiten und kategoriale und begriffliche Grundlagen. Sie sind die Bedingung dafür, miteinander über historische Deutungen und Sinnbildungen kommunizieren zu können. In freiheitlichen Gesellschaften ist beispielsweise die verfassungsrechtlich verbriefte Möglichkeit und Erwünschtheit eines offenen und fachlich informierten gesellschaftlichen Diskurses unter Partizipation möglichst vieler Bürgerinnen und Bürger ein solches Mindestmaß an Übereinstimmung.

2.2.3.4 Historisches Denken: Sinnbildungsmuster

Historisches Denken besteht, idealtypisch betrachtet, in der Erarbeitung narrativ formulierter Zusammenhänge zwischen Ge- und Begebenheiten unterschiedlicher historischer Zeiten (Re-Konstruktion) und der kritisch-reflektierenden Analyse bereits verbal vorliegender historischer Narrationen oder nonverbal ausgedrückter Kontinuitätsvorstellungen (De-Konstruktion). Die re-konstruierende Schaffung von Deutungen und zeitübergreifenden Sinnbildungen wie die de-konstruierende Auseinandersetzung mit vorliegenden Narrationen und ihren Kontinuitätsvorstellungen erfolgt notwendig aus einer jeweils gegenwärtigen Perspektive. Solche Kontinuitätsvorstellungen können sehr vielfältig und unterschiedlich sein. Gleichwohl lassen sich in ihnen immer wiederkehrende Muster finden, die den historisch Denkenden als Grundmuster bei der individuellen Sinnbildung zur Verfügung stehen.

Jörn Rüsen hat in diesem Sinne eine Typologie unterschiedlicher, historisch wie logisch aufeinander aufbauender Erzähltypen entwickelt, die er ihrer orientierenden Funktion wegen als *Sinnbildungsmuster* bezeichnet (Rüsen, 1982, 2008, 2013). Unterschieden werden traditionale, exemplarische, genetische und kritische Sinnbildungsmuster. In der traditionalen Sinnbildung wird die zeitübergreifende Geltung eines Zustandes bzw. eines Wertes allein durch Identifikation und Erzählen dessen Ursprungs und somit durch „Stillstellung" aller seitherigen Zeit begründet und gesichert. Das exemplarische Sinnbildungsmuster stellt, angesichts der Erkenntnis fortgesetzter Veränderungen, Orientierung durch die Identifikation und Formulierung zeitübergreifend gültiger und somit übertragbarer Regeln her. Das genetische Sinnbildungsmuster geht von der Erkenntnis der Wandelbarkeit auch der Regeln und Logiken menschlichen Leidens und Handelns aus; es ermöglicht temporale Orientierung durch die Identifikation und Formulierung zeitabhängiger und damit auch in die Zukunft extrapolierbarer Entwicklungen. Das kritische Sinnbildungsmuster fügt den bestätigenden und erweiternden Dimensionen eine kritische, sich distanzierende bei. Dieses Muster hat Rüsen in der neuesten Darstellung unter Berücksichtigung der Ergänzungen des Modells durch von Borries (1988) und Körber (2013) nunmehr als mehrfache, jeweils das hierarchisch vorangehende Muster delegitimierende, d.h. seine Orientierungsfähigkeit bezweifelnde Übergangsform ausgearbeitet (Rüsen, 2013).

2.2.3.5 Historisches Denken: Kategorien, Konzepte, Prinzipien

In der Geschichtsdidaktik wird der Blick darauf gewendet, geschichtstheoretische, aber auch inhalts- und methodenbezogene Konzepte nicht als etwas der Geschichte und dem historischen Denken und Lernen Äußerliches, sondern als deren konstitutive Bestandteile zu fassen – und zwar sowohl für die akademische Disziplin, für die Lehrerbildung, wie für das schulische Geschichtslernen.

Maßgeblich für historische Narrationen sind nicht nur die Perspektiven und Interessen der die Geschichte Verfassenden und die Quellen Hinterlassenden, sondern

auch die Bedeutungen der Begriffe und Konzepte, mit deren Hilfe die jeweiligen Perspektiven überhaupt erst erfasst und kommuniziert werden können. Sie prägen die Möglichkeiten der narrativen Konstruktion von Sinn mit. Solche Konzepte und Begriffe nicht einfach implizit anzuwenden und für Eigenschaften der Vergangenheit oder historischer Orientierung zu halten, kennzeichnet höher entwickelte Ausprägungen historischer Kompetenz. Begriffe und Konzepte können dann als Instrumente des eigenen (und fremden) historischen Denkens explizit reflektiert werden.

Im Kompetenzmodell von FUER ist diese Fähigkeit der begrifflich-konzeptuellen Erfassung im Kompetenzbereich *Sachkompetenz(en)* modelliert. Die Kenntnis von – mit Jörn Rüsen als „historische Eigennamen" zu bezeichnenden – Daten, Orten, Ereignissen und deren Bezeichnung ist ehrenwertes Lernziel, jedoch aufgrund der fehlenden Übertragbarkeit auf neue, andere Fälle, kein Bestandteil von historischer Sachkompetenz. Diese erfordert ihre Systematisierung in Konzepte. Diese können inhalts-, methoden-, theorie- oder subjektbezogen sein (Schöner, 2007) und haben unterschiedliche Reichweiten. Exemplarisch werden im Folgenden Möglichkeiten der Kategorisierung am Inhalts- und am Theoriebezug verdeutlicht. „Historisch" sind inhaltsbezogene Konzepte dann, wenn sie prototypische Merkmale ebenso umfassen, wie deren Zeitspezifik bzw. ihren Wandel in der Zeit. Die Konzepte können sich nach dem Grad ihrer Abstraktion unterscheiden (vgl. „Burg", „König", „Krieg", „Handel" im Vergleich zu „Revolution", „Macht", „Herrschaft" oder zu „Politik", „Gesellschaft", „Gender", „Performanz"). Letztere werden auch als inhaltsbezogene Kategorien bezeichnet, weil sie überzeitliche Systeme darstellen, die erst durch Konzepte verzeitlicht und an Fällen bzw. Situationen konkretisiert werden (vgl. Bräuer, Lehmann & Werner, 2016; Schöner, 2007; Schreiber, 2012; Ventzke, 2012).

Auch die theoriebezogenen Kategorisierungen lassen sich nach Reichweiten unterscheiden. Im Anschluss an die englischsprachige Forschung können Konzepte mittlerer Abstraktion als *second order concepts* bezeichnet werden. Sie betreffen den Prozess des historischen Denkens und Urteilens („Periodisierung", „Ursache" und „Wirkung"), auf der höchsten Abstraktionsebene steht die begriffliche Fassung der epistemologischen Prinzipien (Kühberger, 2012; s.o. Kapitel 2.1).

2.2.3.6 Historisches Denken: Triftigkeiten als Qualitätskriterien für historische Narrationen

Aus den Geltungsansprüchen aller Geschichten ergibt sich ein Bedarf an strukturellen Kriterien für die Qualität historischer Aussagen und Orientierungen sowie der dazugehörigen Prozesse. Naive Wahrheits- und Objektivitätskonzepte sind dafür – gerade in stark heterogenen Gesellschaften – nicht ausreichend. Mit dem Konzept der *Triftigkeiten* – 2013 bezeichnet er sie auch als „Plausibilitäten" – hat Jörn Rüsen hierfür Überlegungen vorgelegt, die sowohl die Absicherung der Geltungsansprüche eigener Geschichten ermöglicht als auch die Überprüfung von Geltungsansprüchen, die andere erheben (Rüsen, 1983, 2013). Rüsen unterscheidet

- die *empirische Triftigkeit/Plausibilität* als Kriterium, inwiefern eine Geschichte grundsätzlich bzw. transparent nachvollziehbar auf der Überlieferung von Erfahrungen aus der Vergangenheit beruht. Diese Eigenschaft kann durch Benennung der Quellen als Grundlagen hergestellt werden, wobei im Idealfall deren Auswahl offen gelegt und nach Vollständigkeit gestrebt wird und die Relevanz und Aussagekraft der Quellen nicht nur behauptet, sondern argumentativ dargelegt wird.
- die *normative Triftigkeit/Plausibilität* als Kriterium für die Zustimmungsfähigkeit zu den in der Geschichte gewählten Relevanz-, Auswahl- sowie Urteilskriterien bei den jeweiligen Adressatinnen bzw. Adressaten/ Rezipientinnen bzw. Rezipienten. Dabei geht es auch um die Art und Weise, wie der Zusammenhang zwischen in der Vergangenheit geltenden und aktuellen bzw. für die Zukunft antizipierten Normen hergestellt wird. Ein Gütekriterium historischer Narrationen in einer heterogenen Gesellschaft ist, die Zustimmung der adressierten Rezipientinnen und Rezipienten dadurch zu ermöglichen, dass unterschiedliche Zielgruppen mit unterschiedlichen Perspektiven adressiert werden und die Geschichte so erzählt wird, dass diese unterschiedlichen Perspektiven in ihr berücksichtigt werden („Perspektivenerweiterung").
- die *narrative Triftigkeit/Plausibilität* als Kriterium für die Nachvollziehbarkeit der narrativen Konstruktion durch das Publikum. Sie wird sichergestellt durch passende Anwendungen allgemein anerkannter Regeln für die Darstellung von Erklärungen und Zusammenhängen. Zur Verdeutlichung können die jeweils dahinterstehenden Ideen und wissenschaftlichen Theorien expliziert werden.

Unterschieden wird zudem zwischen Narrationen, die Behauptungen setzen, ohne diese zu belegen, aber dennoch zustimmungsfähig sind (einfache Triftigkeit), und Narrationen, die ihre Quellen ausweisen, ihre Normen in den Sinnbildungsangeboten offenlegen und den Zusammenhang von Quellenaussagen und Sinnbildungsangeboten argumentativ begründen (gesteigerte Triftigkeit). Die Narrationen unterscheiden sich also in allen drei Dimensionen hinsichtlich des Grades, in dem der Geltungsanspruch der erzählten Geschichte gesichert wird (Körber, 2016).

2.2.3.7 Historisches Denken: Lernen

In der Summe muss als Konsequenz der narrativistischen Geschichtstheorie nicht nur Geschichte als je perspektivisches Konstrukt betrachtet werden und historisches Denken als der Prozess, in dem es entsteht, sondern auch historisches Lernen als Historisch-Denken-Lernen. Lernen kann also nicht mehr als die Übernahme eines grundsätzlich (wenn auch nicht vollständig) objektiven, für alle gleichermaßen gültigen Wissensbestandes gedacht und organisiert werden. Vielmehr zielt Lernen auf die Befähigung zu eigenständigem Vollzug der Operationen historischen Denkens. Erworben werden die dafür notwendigen Kompetenzen an begründet ausgewählten Inhalten und Themen (u.a. Mierwald & Brauch, 2015).

Es geht beim historischen Lernen in der Konsequenz der narrativistischen Geschichtstheorie also darum, zu historischem Denken zu befähigen, d.h. zur eigenständigen und verantwortlichen Orientierung in der Zeit. Dies umfasst in synthetischer Hinsicht die Prozesse der Re-Konstruktion von Vergangenem, in analytischer Hinsicht die Prozesse der De-Konstruktion vorliegender historischer Narrationen und schließt die Befähigung zur immer wieder neuen Reflexion und Erweiterung der historischen Orientierung und Identitätsbildung ein (Bräuer & Schreiber, 2016; Körber et al., 2007).

2.3 Kompetenzen historischen Denkens: Das FUER-Modell

Die Entwicklung und Weiterentwicklung des FUER-Modells erfolgte und erfolgt in einem internationalen Konsortium, der FUER-Gruppe. Sie besteht aus deutschen, österreichischen und Schweizer Geschichtsdidaktikerinnen und -didaktikern und wurde von Geschichtslehrkräften aus diesen Ländern und zusätzlich aus Belgien, Ungarn, Rumänien und Südtirol unterstützt. FUER ist ein Akronym und steht für das Programm der Gruppe, die **F**örderung **u**nd **E**ntwicklung eines **r**eflekierten und (selbst-) reflexiven Geschichtsbewusstseins.

Die Beschäftigung mit historischen Kompetenzen setzte 2003 ein, führte 2006/2007 zur Erarbeitung des Kompetenz-Strukturmodells der FUER-Gruppe (Schreiber et al., 2006; Körber et al., 2007) und ist bis heute nicht abgeschlossen, weil Ausdifferenzierungen, Vertiefungen und Erweiterungen vorgenommen werden. Aus den Resultaten der unterschiedlichen Arbeitsphasen werden drei Ergebnisse vorgestellt, die für das HiTCH-Projekt von besonderer Relevanz sind: aus den Grundlagen (1) die Herausarbeitung der Basisoperationen der Re- und De-Konstruktion in ihren Fokussierungen (Kapitel 2.3.1.1) und (2) die Dynamisierung des Konzepts Geschichtsbewusstsein (Kapitel 2.3.1.2), und in Bezug auf (3) das Kompetenz-Strukturmodell die definierten Kompetenzbereiche (vgl. Kapitel 2.3.2), die Überlappungsbereiche (vgl. Kapitel 2.3.3) und die Überlegungen zur Graduierungslogik (vgl. Kapitel 2.3.4).

2.3.1 Grundlagen

2.3.1.1 Basisoperationen historischen Denkens und „Fokussierungen"

Auf Grundlage der unter Kapitel 2.2.1 dargestellten Unterscheidung zwischen Quellen und Darstellungen werden die Basisoperationen der vergangenheitsbezogenen Re-Konstruktion und der auf bereits vorliegende historische Narrationen bezogenen De-Konstruktion unterschieden.

Bei der *Re-Konstruktion* wird synthetisierend vorgegangen. Aufgrund einer historischen Fragestellung werden neue Narrationen geschaffen, wobei alle drei Triftig-

keitskriterien (vgl. Kapitel 2.2.3.6) Beachtung finden. Bei der Theoriebildung konnte dabei auf die historisch-kritische Methode zum Umgang mit Quellen zurückgegriffen werden, aber auch auf die Unterteilung des historischen Denkens in drei Schritte durch Karl-Ernst Jeismann (1977, 1980), nämlich in „Sachverhaltsanalyse", „Sachurteil" und „Werturteil", bzw. auf die damit vergleichbare Unterscheidung zwischen „historischer Wahrnehmung", „historischer Deutung" und „historischer Orientierung" bei Jörn Rüsen (1983). Schließlich konnten auch Erfahrungen guter Praxis historischer Forschung einbezogen werden, die – u.a. über die Reflexion des Forschungsstands – neben den Quellen immer auch die bereits vorliegende Historiografie zum im Blick stehenden Thema berücksichtigt (Ziegler, 2007).

Bei der Realisierung der Basisoperation *De-Konstruktion* werden vorliegende Narrationen auf die in ihnen inhärenten Quellenbezüge (nach Jeismann: Sachverhaltsanalysen), auf zeitspezifische Zeit- und Sinnbildungen (nach Jeismann: Sachurteile) und auf Orientierungsangebote (nach Jeismann: Werturteile) hin untersucht. Es werden also analysierende Operationen vollzogen.

In der sogenannten *Sechs-Felder-Matrix* (Schreiber, 2002; Schreiber et al., 2006) hat die FUER-Gruppe die Basisoperationen des Re-Konstruierens und De-Konstruierens präzisiert, indem, aufbauend auf den Ansätzen von Jeismann und Rüsen, drei Fokussierungen unterschieden wurden, in denen re- und de-konstruiert wird. Es handelt sich um die Fokussierung auf Vergangenes, dann auf die in jeweiligen Gegenwarten geschaffenen historischen Narrationen und schließlich auf die Orientierungsleistungen für Gegenwart und Zukunft.

Die drei *Fokussierungen* strukturieren die Sechs-Felder-Matrix in der vertikalen Richtung, die beiden Basisoperationen in horizontaler Richtung.

- Die *Fokussierung auf Vergangenheit* zielt auf die in der Disziplin der Geschichtswissenschaft methodisch kontrolliert erarbeiteten und deshalb als ‚gesichert' angenommenen Einzelheiten über die jeweilige Vergangenheit. Als „Vergangenheitspartikel" (man beachte die Anführungszeichen!) sind sie in jeder Narration enthalten. Die Berücksichtigung des Kriteriums der empirischen Triftigkeit (s.o. Kapitel 2.2.3.6) im Vollzug der beiden Basisoperationen Re- und De-Konstruktion in der Fokussierung auf Vergangenes zielt darauf, die empirische Begründetheit der jeweils verwendeten „Vergangenheitspartikel" sicherzustellen. Unstrittig bleibt, dass damit nicht deren vergangene Tatsächlichkeit behauptet werden kann, sondern lediglich eine empirische Plausibilität dafür, bestärkt insbesondere durch die Anwesenheit und Prüfbarkeit von Quellen unterschiedlicher Gattungen.
- Die *Fokussierung auf Geschichte* konzentriert die Aufmerksamkeit auf die Konstruktion von (synchronen wie diachronen) Zusammenhängen, auf den deutenden Umgang mit Vergangenem und die sinnbildende Nutzung von Vergangenem für die historische Orientierung. In re- wie in de-konstruierenden Denkprozessen kommt der interpretierenden und deutenden Konstruktion, in welcher Vergangenheitspartikel zueinander in Bezug gesetzt werden, entscheidende Bedeutung zu. Als Raster für die Analysen können die zugrunde gelegte Fragestellung, die kon-

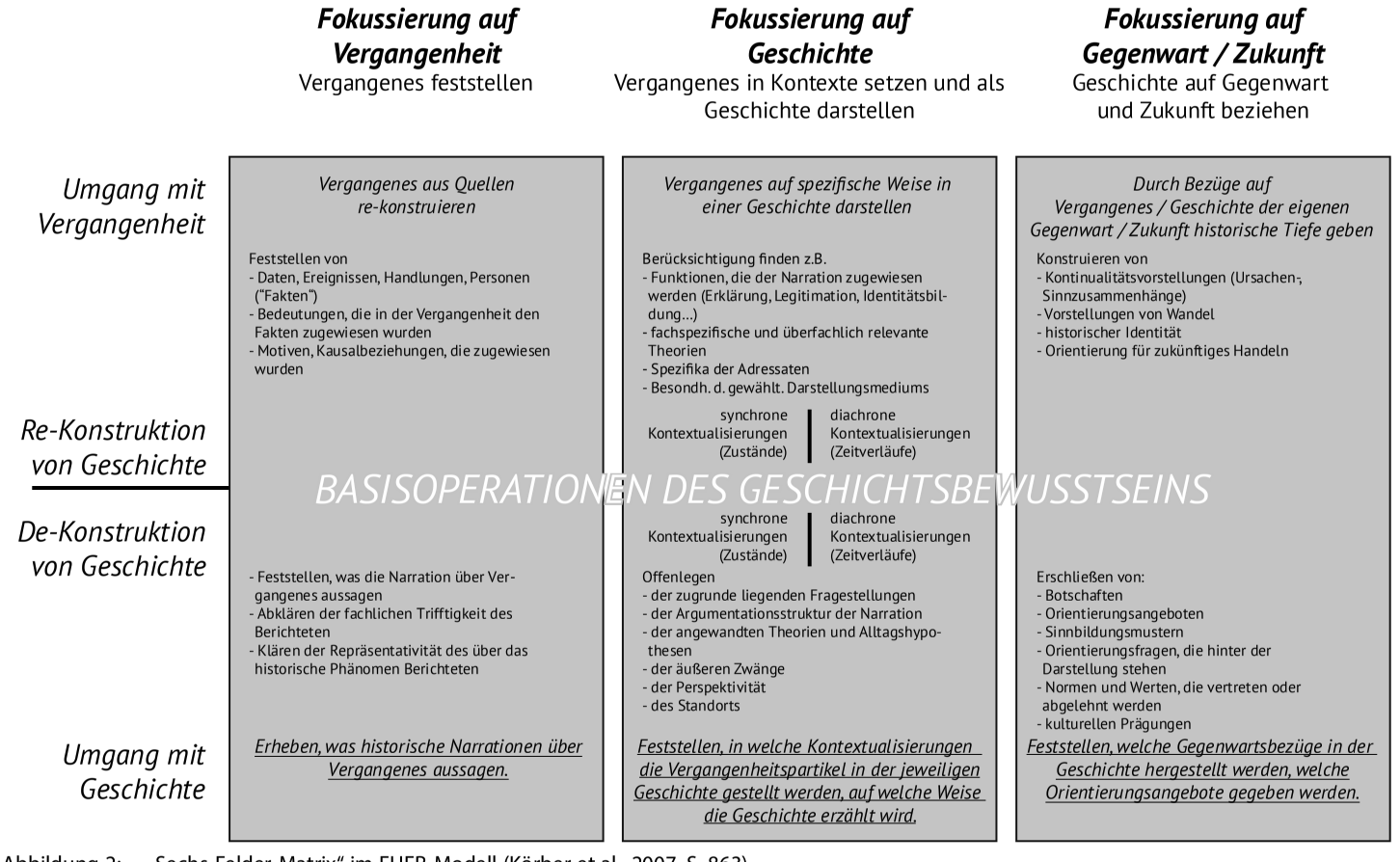


Abbildung 2: „Sechs-Felder-Matrix“ im FUER-Modell (Körber et al., 2007, S. 863)

textualisierenden Einordnungen, die dahinter stehenden theoretischen Konzepte, die narrative Struktur der Darstellung, die Auswirkungen, die die Wahl des Mediums für die Darstellung hat, etc. genutzt werden.
- Idealtypisch davon getrennt werden in der *Fokussierung auf Gegenwart/Zukunft* die zumindest indirekt enthaltenen Bezüge auf die eigene Gegenwart und die Orientierungsangebote für die eigene Zukunft. Hier kommen Vorstellungen von Kontinuität und Wandel zum Ausdruck, die den Rezipientinnen und Rezipienten angeboten werden. Im Zentrum der Analysen stehen damit die konkret vorgeschlagenen Sinnkonstruktionen bzw. die Anregungen zur eigenständigen Sinnkonstruktion, zudem jeweils deren normative und theoretische Verortung.

Die drei Fokussierungen kennzeichnen also aufeinander bezogene Elemente jeder Narration. Die Pfeile in Abbildung 3 deuten dies an.

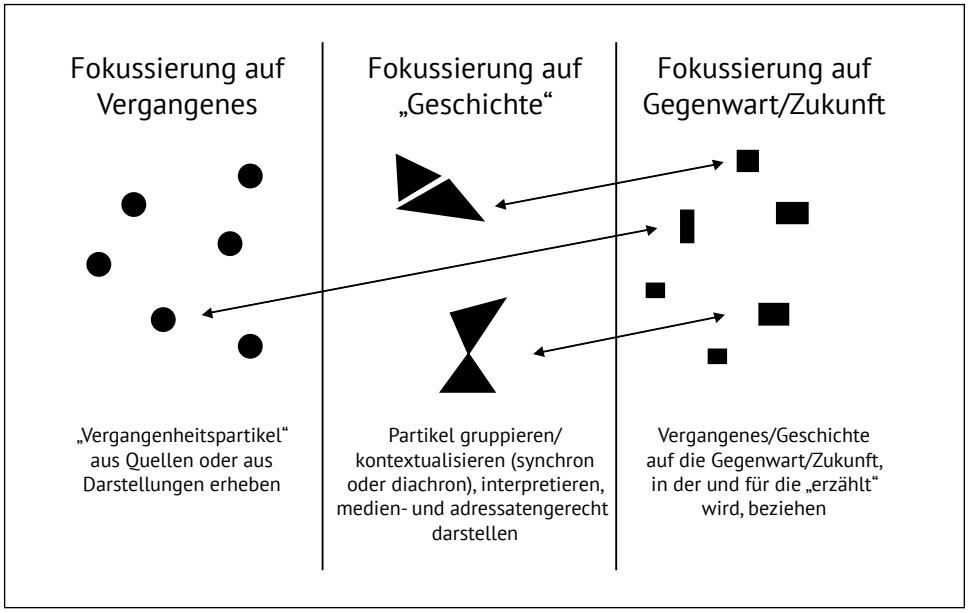


Abbildung 3: „Fokussierungen“ im FUER-Modell (Schreiber & Árkossy, 2009, S. 19)

2.3.1.2 Dynamisierung des Konzepts Geschichtsbewusstsein

Mit der Dynamisierung des Konzepts Geschichtsbewusstsein ist eine zweite Phase der Arbeit an der FUER-Gruppe gekennzeichnet. Den Ausgangspunkt bildet das Konzept einer disziplinären Matrix, die Jörn Rüsen 1983 vorgelegt hat (Rüsen, 1983; vgl. Kapitel 2.2.3.1, dort auch die Abbildung der Matrix). Wolfgang Hasberg und Andreas Körber, zwei tragende Mitglieder der FUER-Gruppe, haben Rüsens Matrix der Geschichtswissenschaft unter dem Titel „Geschichtsbewusstsein dynamisch“ zu einem Prozessmodell historischen Denkens ausdifferenziert (Hasberg & Körber, 2003).

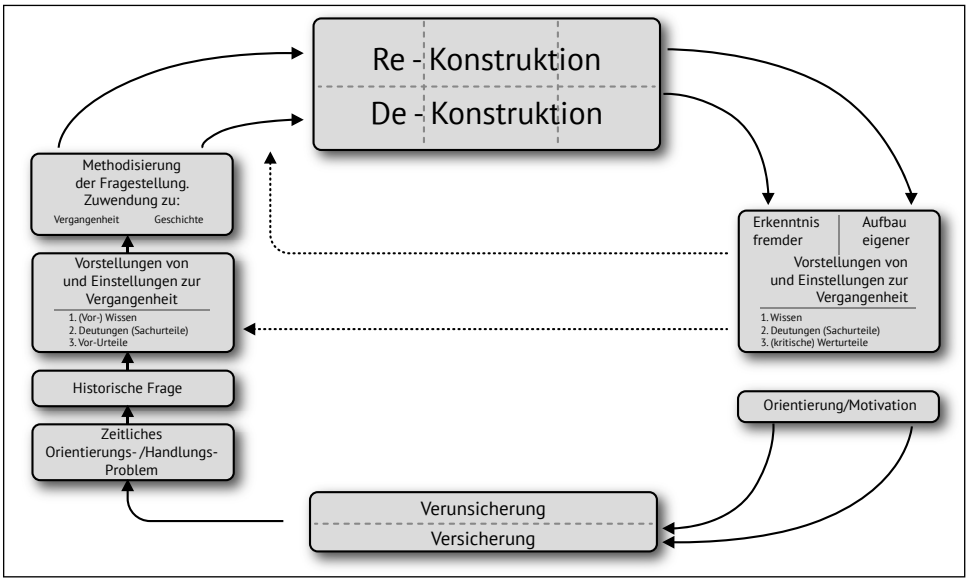


Abbildung 4: „Geschichtsbewusstsein dynamisch“ (Hasberg & Körber, 2003, S. 187)

Die Modellierung geht von einem Orientierungsbedürfnis aus, das in einer „Verunsicherung“ besteht, die die bisherige „Orientiertheit“ (Bräuer & Schreiber, 2016) in Frage stellt. Das zeitliche Orientierungs- und Handlungsproblem wird in eine historische Frage überführt. Bei der Antwortsuche wirken sich Vorstellungen von und Einstellungen zur Vergangenheit/Geschichte aus.

Aufgrund der Fragestellung erfolgt entweder eine re-konstruierende Auseinandersetzung mit Vergangenheit oder eine de-konstruierende Auseinandersetzung mit vorliegenden historischen Narrationen. Das Ergebnis der (im Idealfall methodengeleiteten) Re- und De-Konstruktion ist der Aufbau einer eigenen Narration oder die Prüfung fremder Deutungen und Sinnbildungen. Die Ergebnisse können durch die Beantwortung der anfangs gestellten Frage eine Orientierung geben, also die Verunsicherung auflösen. Ergebnis können aber auch neue Verunsicherungen sein, die einen weiteren historischen Denkprozess initiieren.

Dieses Prozessmodell „Geschichtsbewusstsein dynamisch“ wurde von der FUER-Gruppe als Ausgangspunkt für die Bestimmung der zentralen Kompetenzbereiche historischen Denkens genommen. Daher ist im Hintergrund der grafischen Darstellung des FUER-Modells (siehe Abbildung 5) das Prozessmodell schemenhaft zu erkennen.

2.3.1.3 Die Kompetenzbereiche, Kernkompetenzen, Einzelkompetenzen – formale Unterscheidung

Bei der Entwicklung des FUER-Modells wurden Kompetenzbereiche und Kernkompetenzen unterschieden und definiert. Idealtypisch können ihnen alle Einzelkompetenzen zugeordnet werden, die im konkreten Denkprozess vollzogen werden.

Kompetenzbereiche historischen Denkens umfassen Gruppen verwandter Kompetenzen, die sich aus der Systematik des historischen Denkens ergeben (Schreiber et al., 2006, Glossar, S. 56). Aus dem eben dargestellten Prozessmodell „Geschichtsbewusstsein dynamisch" wurden drei *prozedurale Kompetenzbereiche* abgeleitet, die historischen Frage-, Methoden- und Orientierungskompetenzen. Der vierte Kompetenzbereich ist ein *kategorisierender* („historische Sachkompetenzen"). Er umfasst die Fähigkeit, Fertigkeit, Bereitschaft, begrifflich zu strukturieren und bezieht sich auf Prinzipien, Kategorien, Konzepte, Skripts und andere Strukturierungsschemata. Zwischen den prozeduralen Kompetenzbereichen und der historischen Sachkompetenz bestehen notwendig Zusammenhänge, die in der graphischen Darstellung als *Überlappungsbereiche* visualisiert werden.

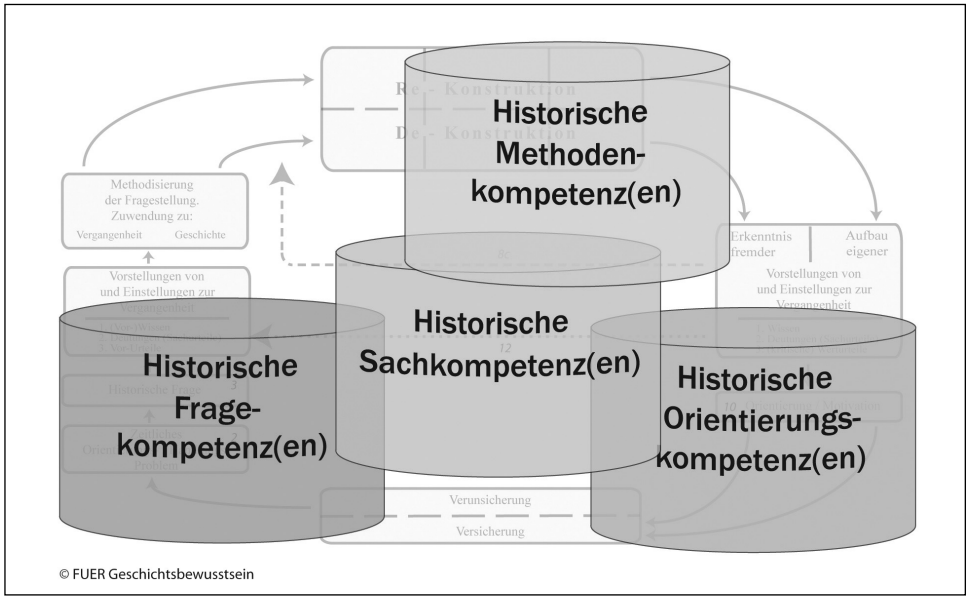


Abbildung 5: FUER-Modell „Kompetenzen historischen Denkens" (Schreiber et al., 2006, S. 30)

Die Kompetenzbereiche werden in den so genannten *Kernkompetenzen* operationalisiert. Diese „strukturieren den jeweiligen Kompetenzbereich, sind systematisch abgeleitet und deshalb eindeutig zuordenbar (Schreiber et al., 2006, Glossar, S. 58).

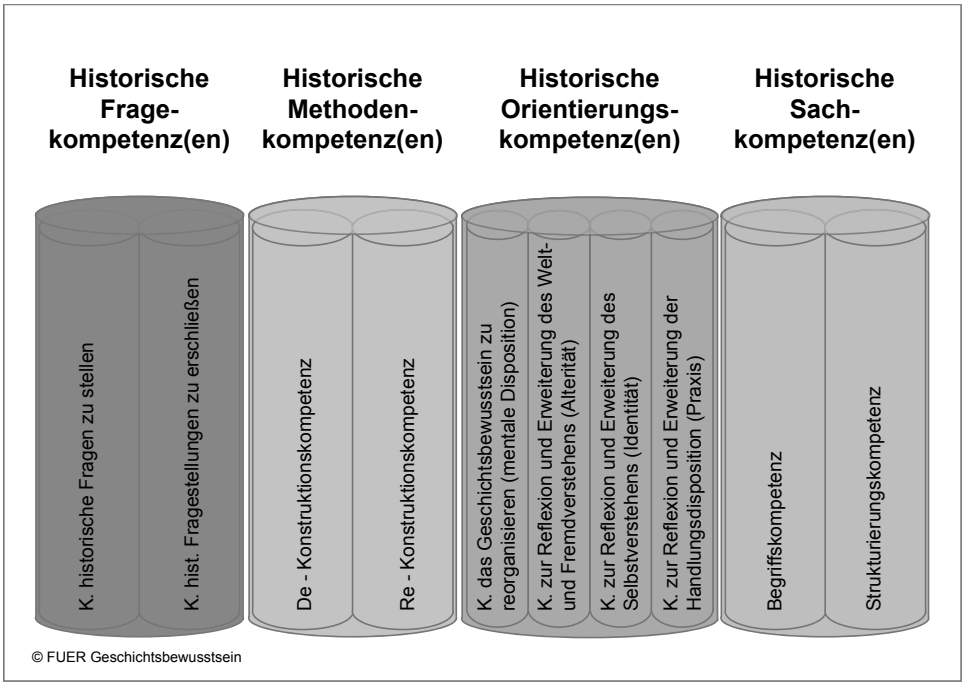


Abbildung 6: Kernkompetenzen zur Operationalisierung der Kompetenzbereiche des FUER-Kompetenz-Strukturmodells (Schreiber et al., 2006, Glossar, S. 58)

Neben den theoretisch begründeten und idealtypisch modellierten Kompetenzbereichen und Kernkompetenzen stehen Einzelkompetenzen, die sich im konkreten Vollzug historischen Denkens zeigen. Einzelkompetenzen können entweder eindeutig einer Kernkompetenz zugeordnet sein oder sich auf die Überlappungsbereiche zwischen mehreren Kernkompetenzen beziehen. Ihr Ausweis kann deduktiv erfolgen, z.B. unter Bezug auf Kompetenzbereiche und Kernkompetenzen oder induktiv ausgehend vom konkreten Vollzug historischen Denkens (Schreiber et al., 2006, Glossar, S. 59).[5]

5 Die Quellengattung „historische Karikatur" erkennen können, wäre eine der Sachkompetenz eindeutig zugeordnete Einzelkompetenz. Demgegenüber bezieht sich die Einzelkompetenz „Überprüfen können, inwiefern ein Autor eine zitierte Karikatur in seiner Narration triftig genutzt hat", auf mehrere Kernkompetenzen: Es gehen ein das Klassifizieren-Können von Bildern als historische Karikatur (Sachkompetenz), re-konstruierende Methodenkompetenzen für den Umgang mit Karikaturen und de-konstruierende Methodenkompetenz für die Analyse historischer Narrationen.

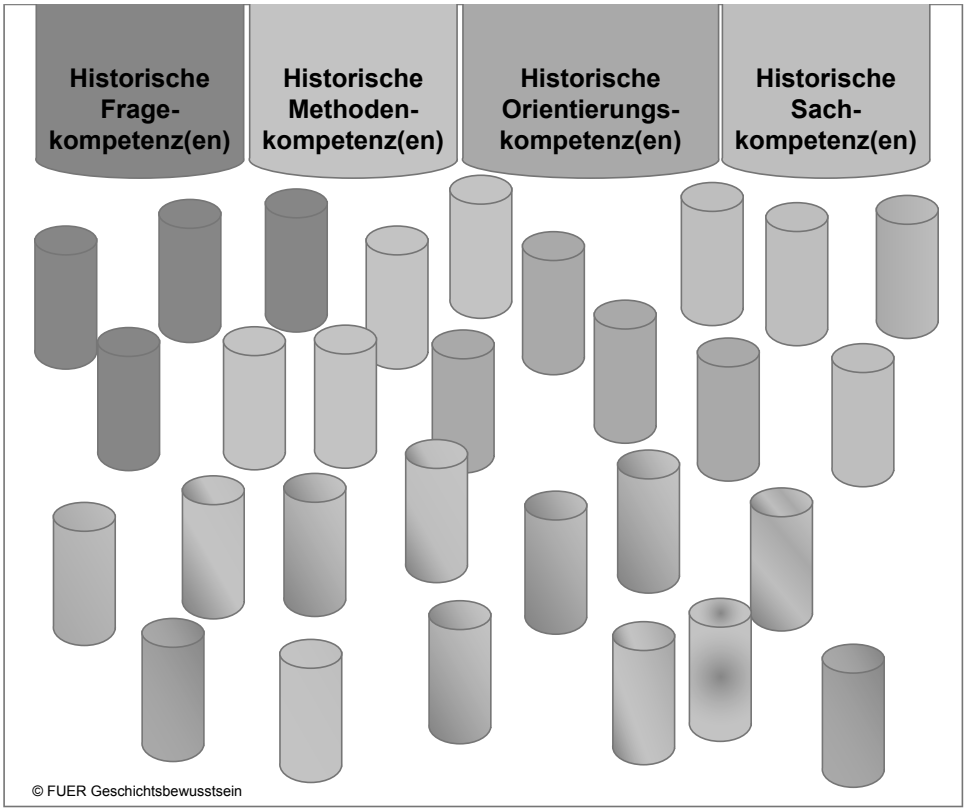


Abbildung 7: Einzelkompetenzen im FUER-Kompetenz-Strukturmodell (Schreiber et al., 2006, Glossar, S. 59)

2.3.2 Kompetenzbereiche und Kernkompetenzen – inhaltliche Beschreibung

2.3.2.1 Historische Fragekompetenzen

Indem historische *Fragekompetenzen* als eigener Kompetenzbereich definiert werden, erweitert das FUER-Modell die Kompetenzdefinition Weinerts (2001), die das Problemlösen in neuen Situationen betont, indem es dem Aspekt des „Problemformulierens" eine wichtige Funktion zuweist. Wie bereits dargestellt, geht die Modellierung der FUER-Gruppe davon aus, dass der Prozess historischen Denkens durch gegenwärtige „Verunsicherungen" ausgelöst wird, die bisherige, historische Sinnbildungen in Frage stellen. Verunsicherungen können auftreten, wenn Entscheidungen für morgen getroffen werden oder vorliegende Deutungen und Sinnbildungen in ihrer Triftigkeit und/oder Orientierungsrelevanz eingeschätzt werden müssen. Irritationen und Interesse können aber auch durch eine zufällige Begegnung mit Unerwartetem oder durch eine arrangierte Lernumgebung ausgelöst werden (Bräuer & Schreiber, 2016; Kühberger, 2016; Meyer-Hamme, 2014; Schreiber, 2007). Entscheidend ist, dass eine

Bündelung der Verunsicherungen/Irritationen in einer historischen Fragestellung erfolgt, die historisches Denken initiiert: „Ohne historische Fragen keine Geschichte“ (Schreiber, 2007, S. 156). Fragen an den Anfang des Denkprozesses zu stellen, bedeutet zugleich, dass die gestellten Fragen lenken, weil sie bestimmte Aspekte ins Zentrum rücken, bestimmte Standpunkte einnehmen, auf eine bestimmte Auswahl von Materialien zielen (Droysen, 1857/1977, 1881/1974).

Als erste Kernkompetenz der Fragekompetenz wird definiert, historische Fragen an die Vergangenheit selbst zu formulieren, um so auf die aufgetretenen „Verunsicherungen“ zu reagieren (Brauch, 2016). Die zweite Kernkompetenz umfasst die Fähigkeit, Fertigkeit und Bereitschaft, historische Fragestellungen anderer zu erschließen und mit „Verunsicherungen“ in Zusammenhang zu bringen. Die erste Kernkompetenz kann der Basisoperation der Re-Konstruktion zugeordnet werden, die zweite der Basisoperation der De-Konstruktion.

2.3.2.2 Historische Methodenkompetenzen (Re- und De-Konstruktion)

Wenn zur Beantwortung einer Frage an die Vergangenheit Quellen und Darstellungen untersucht werden, wird ein methodisch geregelter *Re-Konstruktionsprozess* mit dem Ziel, zu einer eigenen Narration zu kommen, in Gang gesetzt. Eine Frage hingegen, die die Struktur einer vorliegenden Narration in den Blick nimmt, initiiert einen *De-Konstruktionsprozess*. Re- und De-Konstruktionsprozesse werden, wie unter Kapitel 2.3.1.1 verdeutlicht, als die zwei Basisoperationen des historischen Denkens verstanden. Die dafür notwendigen methodischen Fähigkeiten werden im FUER-Modell im Bereich der *Methodenkompetenz* (*Re- und De-Konstruktion*) verortet. Die Kernkompetenzen der Methodenkompetenz sind somit die Fähigkeiten, Fertigkeiten und Bereitschaft zu methodisch regulierter Re- und De-Konstruktion. Der Re-Konstruktionskompetenz zugeordnet werden können die Schritte einer methodisch geleiteten Quellenarbeit, von der Heuristik, über die innere und äußere Quellenkritik, die Interpretation der Quellen zur kontextualisierenden Narrativierung der Ergebnisse. Als historisch-kritische Methode werden diese Schritte in den Einführungen in die historische Forschung beschrieben (z.B. Lingelbach & Rudolph, 2005). Der De-Konstruktionskompetenz können die Schritte der kritischen Analyse historischer Narrationen zugeordnet werden, indem z.B. von der Deskription der Elemente der Oberflächenstruktur ausgehend, in mehreren Schritten die darunter liegenden Tiefenstrukturen erfasst und auf die Beachtung der Triftigkeits- bzw. Plausibilitätskriterien (Schreiber & Gruner, 2010) hin überprüft werden.

2.3.2.3 Historische Orientierungskompetenzen

Der Kompetenzbereich der *Orientierungskompetenz* adressiert die Fähigkeit, Fertigkeit und Bereitschaft, die Erkenntnisse und Einsichten, die durch die Re- und De-Konstruktionsprozesse gewonnen wurden, explizit als historische Sinnbildung auf die eigene Gegenwart und Zukunft zu beziehen. Damit können Zukunftsentscheidungen auch unter Bezug auf vergangene Erfahrungen begründet werden. Darüber hinaus geht es darum, vorhandene Orientierungsangebote auf die eigene Person und Lebenswelt bzw. die eigene Weltsicht zu beziehen. Dies verlangt Selbstreflexion und Fähigkeiten zur Selbsteinschätzung (Körber et al., 2007; Kühberger, 2012).
Als Kernkompetenzen wurden definiert,

- die Kompetenz, das eigene historische Welt- und Fremdverständnis zu reflektieren und zu erweitern. Es geht dabei darum, mit der „Alterität" des Vergangenen im Vergleich zu Gegenwartserfahrungen umgehen und Zeitlichkeit von Entwicklung und Veränderung denken zu können.
- die Kompetenz, das eigene historische Selbstverständnis zu reflektieren und zu erweitern; damit sind Prozesse der Identitätsbildung und -reflexion angesprochen.
- die Kompetenz, das eigene Geschichtsbewusstsein zu reorganisieren; damit ist gemeint, die mentale Disposition für das Denken von Geschichte als lebenslangen Prozess zu verstehen.
- die Kompetenz, historisch fundierte Dispositionen gegenwärtiger und zukünftiger Handlungen zu reflektieren und zu erweitern. Die Fähigkeit, Fertigkeit und Bereitschaft, in der Praxis historisch fundiert zu agieren, ist angesprochen.

2.3.2.4 Historische Sachkompetenzen

Wie oben ausgeführt, unterscheiden sich die historischen Sachkompetenzen von den drei prozeduralen Kompetenzen. Sie bezeichnen die Fähigkeit, Fertigkeit und Bereitschaft zur begrifflichen Kategorisierung (Schöner, 2007). Sie äußern sich darin, in immer wieder neuen Zusammenhängen und Situationen über domänenbezogene Begrifflichkeiten verfügen zu können (vgl. Kapitel 2.2.3.5). Kategorisiert werden kann z.B. mit Hilfe der epistemologischen Prinzipien, inhaltsbezogener Kategorien und Konzepte, methodenbezogener Skripts für das Re- und De-Konstruieren oder orientierungsbezogener Zeit- und Sinnbildungsmuster.

Erst durch die Verfügung über solche Kategorisierungen ist es möglich, sich z.B. über historische Orientierungen auszutauschen, Unterschiede und Gemeinsamkeiten zu bezeichnen, raum-, zeit- und kulturabhängige Veränderungen zu beschreiben, sowohl über historisches Denken selbst als auch über Ergebnisse eines Denkprozesses (Deutungen, Interpretationen, Schlussfolgerungen) mit anderen zu kommunizieren. Historisches Denken erfolgt so nicht nur individuell-vereinzelt, vielmehr werden Vorschläge zur historischen Orientierung gesellschaftlich anschlussfähig, Triftigkeiten bzw. Plausibilitäten können geprüft und zur Grundlage von Austauschprozessen ge-

macht werden. Als Kernkompetenzen der historischen Sachkompetenz wurden Begriffs- und Strukturierungskompetenz definiert.

2.3.3 Überlappungsbereiche

Die Kompetenzbereiche lassen sich idealtypisch klar trennen. Zwischen den Bereichen bestehen aber systematische Zusammenhänge und Überlappungen. Den Zusammenhang zwischen den prozeduralen Kompetenzen bestimmen die Abläufe des historischen Denkprozesses (Entwicklung von auf Re- oder De-Konstruktion zielenden Fragestellungen; ihre methodisch regulierte Bearbeitung einschließlich der Narrativierung der Ergebnisse; in Bezugsetzung zu bisher vorhandenen historischen Orientierungen). Der reale Denkprozess kann abweichend von der idealtypischen Modellierung Vor- und Rückwärtsbewegungen enthalten. Die „Verunsicherungen", die ihn auslösen, entstehen im realen Denkprozess nicht nur durch lebensweltlich auftretende Orientierungsbedürfnisse. Irritationen können z.B. auch bei der Auseinandersetzung mit Quellen und Darstellungen auftreten, bei der Konfrontation mit unterschiedlichen Orientierungsangeboten oder wenn aufgrund der Quellenlage die empirische Triftigkeit von plausibel erscheinenden Narrationen nicht nachgewiesen werden kann.

Zusammenhänge und Überlappungen bestehen auch zwischen den prozeduralen Kompetenzen und der historischen Sachkompetenz. Die vorhandenen Ausprägungen der Sachkompetenzen beeinflussen die Qualität des Denkprozesses in den unterschiedlichen Phasen. Umgekehrt können historische Sachkompetenzen in jedem Denkprozess weiterentwickelt werden. Was zuvor fragmentiertes, nicht kategorisierbares Wissen war, kann im Prozess konkreten historischen Denkens verortet und eingeordnet, konzeptualisiert werden etc.

Die Zusammenhänge und Überlappungen zwischen den Kompetenzbereichen sind Folgen der in Kapitel 2.1 und 2.2 dargestellten Komplexität historischen Denkens. Die Einzelkompetenzen, die aktiviert werden, um historische Denkprozesse zu vollziehen, sind nicht nur in den Kompetenzbereichen, sondern auch in Überlappungsbereichen verortet (vgl. Kapitel 2.3.1.3) und können untereinander in Beziehung stehen.

2.3.4 Graduierungslogik im FUER-Modell

Im Unterschied zu anderen Kompetenzmodellen historischen Denkens wurde im Rahmen des FUER-Modells auch ein Graduierungskonzept entwickelt, das es erlaubt, unterschiedliche Ausprägungen im historischen Denken zu unterscheiden (Körber, 2007c, 2012).

Für die Ausdifferenzierung der Kompetenzniveaus wird der Grad der Verfügung über gesellschaftlich relevante Konventionen für historisches Denken herangezogen.

Solche Konventionen sind notwendig, um in einer Gesellschaft oder in einer Gruppe zu Vergangenheit/Geschichte/historische Orientierungen überhaupt kommunikationsfähig zu sein (vgl. Kapitel 2.3.1.1, Fokussierungen). Dem *con-venire*, d.h. der Einigung auf bestimmte Konventionen, haften notwendig normative Züge an.

In pluralen Wissensgesellschaften wäre als Konvention z.B. eine Einigung auf ein narrativistisch-konstruktivistisches Geschichtsverständnis und auf die von Jörn Rüsen definierten Triftigkeitskriterien (vgl. Kapitel 2.2.3.6) für historische Narrationen denkbar bzw. auch auf Operationalisierungen, wie sie z.B. im FUER-Modell vorgelegt wurden. Konventionen müssten dann bezogen auf alle Kompetenzbereiche vereinbart werden. Es gäbe also konventionelle Formen für das Stellen und Erkennen historischer Fragen, für re- und de-konstruierende Methoden, für historische Orientierung beim Umgang mit der Welt, mit sich und den anderen. Bei diesen Operationen würden Kategorien, Konzepte und Strukturierungsschemata angewandt, die in der Gesellschaft bzw. der jeweiligen sozialen Gruppe anerkannt sind (= Konventionen zum Kompetenzbereich der historischen Sachkompetenzen).

Grundsätzlich spielen Schule, Universität und andere formale und nonformale Bildungsinstitutionen für das systematische Erlernen und Unterscheiden der je gültigen oder ausgehandelten Konventionen eine bedeutsame Rolle. Insofern können die Schülerinnen und Schüler einer Schulstufe als „soziale Gruppe" verstanden werden und die über Bildungspläne oder Bildungsstandards definierten Kompetenzziele und die Modi der Kompetenzförderung als dort geltende Konventionen.[6] Folgt man dem Kompetenzverständnis der FUER-Gruppe, müssen die in der Schule erworbenen Kompetenzen auf lebensweltlich relevante Orientierungsfragen bzw. auf außerschulisch angetragene Orientierungsangebote anwendbar sein.

Generell darf nicht übersehen werden, dass es sich bei Konventionen um zeit-, raum-, kultur- und gruppenspezifische Konstruktionen handelt. Sie können für ganze Gesellschafen gelten oder nur für einzelne Gruppen innerhalb einer Gesellschaft. In der Graduierungslogik des FUER-Modells ist es ein differenzierendes Kriterium, *ob* der/die historisch Denkende den Konstruktcharakter von Konventionen erfasst und *inwiefern* sich dies auf sein historisches Denken auswirkt. Der Grad der „Konventionsverfügung" und damit das Niveau für Kompetenzausprägungen bestimmt sich somit durch die Art und Weise des Umgangs mit Konventionen, die dem Denkenden möglich sind, und nicht daran, über welche konkreten Konventionen er/sie verfügt. Daraus ergeben sich folgende Niveauunterscheidungen:

- Das Kompetenzniveau wird als *konventionell/intermediär* bezeichnet, wenn der/die historisch Denkende über gesellschaftliche Konventionen verfügen kann, indem er/sie diese im historischen Denken anwendet.
- Das Kompetenzniveau wird als *transkonventionell/elaboriert* bezeichnet, wenn es der/dem historisch Denkenden möglich ist, die Konventionen in ihrem Konstrukt-

6 Zentral ist dafür auch die Erkenntnis, dass normative Konventionen wie Lehrpläne fluide gesellschaftliche Konstrukte und Ergebnisse entsprechender Aushandlungsprozesse bezüglich der Philosophie des Lehrplans sowie der Auswahl der Gegenstände oder des Kompetenzverständnisses sind (Brauch, 2015, 2016b).

charakter zu erkennen und zu reflektieren. Er/Sie kann also die für die jeweilige Gruppe und Gesellschaft geltenden Konventionen als zeit-, raum-, kulturabhängig und damit veränderbar wahrnehmen, ist in der Lage, diese für sich selbst zu modifizieren oder ganz zu verändern, ggfs. dabei über Möglichkeiten nachzudenken, mit diesen Einsichten in die eigene Gruppe bzw. in andere mit ihr in Bezug stehende Gruppen hinein zu wirken.

- Als *a-konventionell/basal* wird das Kompetenzniveau bezeichnet, wenn der/die sich mit Geschichte/Vergangenheit Befassende sich situativ und in wechselnden Situationen auch völlig unterschiedlich verhält, ohne dabei systematisch auf die Konventionen seiner Gruppe zurückgreifen zu können.

Die dreistufige Differenzierung der Niveaus (konventionell/intermediär, transkonventionell/elaboriert und a-konventionell/basal) werden als *Niveau-Bereiche* definiert, die idealtypisch voneinander abgrenzbar sind. Innerhalb der Niveau-Bereiche bestehen notwendig zahllose Abstufungen der Kompetenzausprägungen. Im a-konventionell/basalen Niveau können historisch Denkende z.B. bezogen auf unterschiedliche Kompetenzbereiche einige/mehrere Konventionen kennen, ohne diese regelmäßig zu nutzen. Im konventionell/intermediären Niveau kann der/die historisch Denkende z.B. erkennen, dass in der Schule anders mit Geschichte umgegangen wird, als in seinen peer-Groups, ohne daraus Konsequenzen zu ziehen. Im transkonventionell/elaborierten Niveau kann der/die historisch Denkende für sich Alternativen zu den Konventionen in seiner Gruppe finden, ohne darüber nachzudenken, inwiefern diese Einsichten kollektiviert werden können/sollen.

Nicht nur innerhalb der Niveau-Bereiche bestehen Abstufungen, die – beispielsweise in Bezug auf eine empirische Erfassung – auszudifferenzieren wären. Auch die Übergangsstufen zwischen den Kompetenzniveaus können näher bestimmt werden. Im Übergang vom basalen zum intermediären Niveau etwa handelt es sich um erste schwache Ausprägungen des Verfügens über Konventionen (Lehmann, Werner & Zabold, 2016; Meyer-Hamme, 2007), im Übergang vom konventionell/intermediären zum transkonventionell/elaborierten Niveau um tastende Versuche der Konventionsreflexion (Werner & Schreiber, 2015).

Die hinter der Graduierungslogik des FUER-Modells stehende Komplexität ist beträchtlich, u.a. deshalb, weil der Einzelne unterschiedlichen Gruppen angehört, in denen in der Regel auch unterschiedlich mit Geschichte/Vergangenem umgegangen wird, oder weil Konventionen Veränderungsprozessen unterliegen, die zum Teil schwierig zu erfassen sind. Es ist also zu betonen, dass die Graduierungslogik des FUER-Modells vorerst ganz grundsätzlich verschiedene Niveaus historischen Denkens unterscheidet, ohne eine Fokussierung auf historisches Denken in schulischen Zusammenhängen vorzunehmen. Es gibt bislang lediglich erste Ansätze für systematische Kriterien zur Beschreibung von Niveau-Unterschieden innerhalb der drei Niveaubereiche.

2.4 Das FUER-Modell im nationalen und internationalen Kontext

Neben dem FUER-Modell wurden in den letzten Jahren weitere deutschsprachige und internationale Modellierungen historischen Denkens entwickelt und teilweise auch für pragmatische und empirische Ansätze genutzt. International ist die den deutschsprachigen Modellen weitestgehend gemeinsame Kompetenzkonzeption nach Weinert (2001) und Klieme et al. (2003) nicht verbreitet, weshalb z.B. für den englischsprachigen Bereich mit Seixas und Ercikan (2015) allgemeiner von „Kognitionsmodellen" historischen Denkens als einer Vergleichsebene gesprochen werden sollte. Im Folgenden wird ein Überblick über deutsch- und englischsprachige Konzeptionen gegeben.

2.4.1 Weitere deutschsprachige Konzeptionen

1) Die „Einheitlichen Prüfungsanforderungen für die Abiturprüfung" (EPA) der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (KMK, 1989/2005) strukturieren im Unterricht zu entwickelnde und von Abiturientinnen und Abiturienten zu erwartende Fähigkeiten mittels „Operatoren". Diese bereits aus der Zeit vor der Kompetenzorientierung stammende und fachübergreifend gültige, fachlich aber fokussierte Konzeption ist insofern als Referenz- und Vergleichsebene für die Erstellung fachspezifischer Kompetenzkonzeptionen und -tests bedeutsam, als sie ebenfalls Tätigkeiten von Schülerinnen und Schülern modelliert (ohne jedoch die ihnen zugrunde liegenden Fähigkeiten kompetenztheoretisch zu begründen). Dafür werden drei Anforderungsbereiche definiert und auf die Fächer bezogen: Der Anforderungsbereich I („Reproduktion") umfasst das Wiedergeben von Sachverhalten aus einem abgegrenzten Gebiet und im gelernten Zusammenhang mittels eines reproduktiven Benutzens eingeübter Arbeitstechniken (Operatoren: „nennen", „aufzählen", „zusammenfassen", „wiedergeben"). Operatoren für Leistungen im Anforderungsbereich II („Reorganisation und Transfer") fokussieren auf das selbstständige Erklären, Bearbeiten und Ordnen bekannter Inhalte und das angemessene Anwenden gelernter Inhalte und Methoden auf andere Sachverhalte (Operatoren: „analysieren", „untersuchen", „herausarbeiten", „erläutern", „einordnen"). Der Anforderungsbereich III („Reflexion und Problemlösung") umfasst den reflexiven Umgang mit neuen Problemstellungen, mit den eingesetzten Methoden und mit gewonnenen Erkenntnissen, um zu eigenständigen Begründungen, Folgerungen, Deutungen und Wertungen zu gelangen (Operatoren: „bewerten", „Stellung nehmen", „diskutieren", „prüfen", vergleichen"). Die in der Abstufung der Anforderungsbereiche gedachte „Niveau"-Differenzierung unterscheidet sich von der im FUER-Modells vorgelegten Graduierung insofern fundamental, als innerhalb jedes Anforderungsbereichs Aufgaben in unterschiedlich reflektierter Weise gelöst werden kön-

nen (Körber & Meyer-Hamme, 2007; Körber, Meyer-Hamme & Schreiber, 2007; Körber, 2007a, 2012).

2) Ein frühes, von Michael Sauer vorgeschlagenes „pragmatisches Kompetenzmodell“ (Sauer, 2006) formuliert als Kompetenzbereiche Fähigkeiten historischen Denkens von Schülerinnen und Schülern mit einem spezifischen Fokus auf Medien und Methoden (Methodenkompetenz), beschränkt diese aber auf im Geschichtsunterricht zu fördernde Ausprägungen. Darüber hinaus werden unter *Deutungs- und Reflexionskompetenz* fachspezifische Kompetenzen für den Umgang mit Vergangenheit und Geschichte beschrieben. Ein anderer Bereich konzeptualisiert vornehmlich die Verfügung über konkrete, nicht-übertragbare Wissens- und Deutungsbestände; er wird dennoch als Kompetenzbereich bezeichnet („Sachkompetenz“). Der Geschichtslehrerverband (Verband der Geschichtslehrer Deutschlands, 2007, 2011) und einige Rahmenlehrpläne haben das Modell aufgegriffen. Mit dem FUER-Modell bestehen nur geringe Überschneidungen, schon wegen der Beschränkung auf das Feld „Schule“, vor allem aber wegen des hohen Anteils an nicht-übertragbaren Wissens- und Deutungsbeständen und der insgesamt geringen Bedeutung von Transfer.[7]

3) Hans-Jürgen Pandel (2005) geht in seiner Kompetenzkonzeption von außerschulischen „Bewährungsfeldern“ der Kompetenzen historischen Denkens aus und somit, wie die FUER-Gruppe, über schulisches Geschichtslernen hinaus. Er unterscheidet vier, in einem Kreislaufmodell entlang der Zeitachse verortete Kompetenzbereiche: Die Gattungskompetenz als die Fähigkeit, mit den verschiedenen Textgattungen, die die Grundlage der Befassung mit Geschichte bilden, umgehen zu können; die Interpretationskompetenz als die Fähigkeit, aus den Gattungen historisches Wissen und historischen Sinn zu entnehmen; die geschichtskulturelle Kompetenz als Voraussetzung zur Befragung und Bearbeitung in der Geschichtskultur aufzufindender Sinnangebote; die narrative Kompetenz als die Fähigkeit, aus zeitdifferenten Ereignissen durch Sinnbildung eine Geschichte herzustellen. Dieses Modell folgt somit zwar einer anderen Strukturierung, bezieht sich aber wesentlich auf auch im FUER-Modell gefasste Operationen und auf ihnen zugrunde liegende Tätigkeiten historischen Denkens. Insofern besteht eine deutlich höhere Anschlussfähigkeit zum FUER-Modell als bei der Konzeption Sauers (Einbeziehung außerschulischen Geschichtslernens; Kreislaufmodell; re- und de-konstruktive Elemente). Insgesamt ist die Theoriefundierung, der Zusammenhang zwischen den Kompetenzbereichen und der Zusammenhang zwischen Wissen und Kompetenz aber weniger systematisch angelegt als beim FUER-Modell.

---

7 In Rahmenlehrplänen und ministeriellen Vorgaben finden sich immer wieder auch eigene Kompetenzdefinitionen und -modellierungen – in unterschiedlich enger Anlehnung an eines der genannten, in eigenständiger Kombination der Elemente mehrerer von ihnen oder durch Einbezug von Elementen fachübergreifender (generischer) oder nachbarfachlicher Modellierungen. Sowohl in der Definition wie auch in der Benennung, besonders aber in der Begründung dieser Abgrenzungen und Definitionen zeigen sich dabei sowohl Gemeinsamkeiten wie Unterschiede – nicht selten auch die Verwendung identischer Terminologien für unterschiedlich gefasste Aspekte.

4) Wiederum anders strukturiert Peter Gautschi (2009) sein Kompetenzmodell. Wie das FUER-Modell geht er von einem zirkulären Prozess aus, beschränkt sich aber auf die synthetischen (in der FUER-Terminologie: „re-konstruktiven") Prozesse. Zudem fokussiert er nur das schulische Lernen. Bei der Unterteilung von Dimensionen bzw. Operationen historischen Denkens bezieht er sich ausschließlich auf Jeismann (u.a. 1977, 2000: Sachanalyse, Sachurteil, Werturteil), nicht z.B. auch auf Rüsen. Der von Gautschi skizzierte Regelkreis historischen Lernens ist geeignet, grundlegende Anforderungen an die Kompetenzförderung im Geschichtsunterricht zu beschreiben. Mit den ausdifferenzierten Aspekten besteht wie beim Modell Pandels eine klare, wenn auch anders gelagerte Anschlussfähigkeit zum FUER-Modell. Eine Reihe von zentralen Elementen des FUER-Modells (u.a. Einbinden des schulischen Lernens in das historische Denken; die gesamte Basisoperation der De-Konstruktion, Theorieentwicklungen zur historischen Orientierung) bleiben bei Gautschi allerdings ausgeblendet.

Insgesamt zeigt sich, dass das FUER-Modell Anschlussstellen für die in anderen Modellen fokussierten Überlegungen zur Kompetenzmodellierung bietet. Es wird aber auch deutlich, dass das FUER-Modell spezifische Setzungen enthält. Zunächst steht das eigenständige, ergebnisoffene – aber nicht beliebige – historische Denken und Urteilen nicht neben eher reproduktiven, auf Fördermaßnahmen bezogene Fähigkeiten, sondern im Zentrum. Hiermit korrespondiert ein differenziertes Verständnis von Wissen: Insofern es um die Kenntnis von vergangenen „Tatsachen" bzw. „Fakten" geht (*Know What*), gilt dieses im FUER-Modell nicht als Teil der Kompetenz, sondern (lediglich) als dasjenige Substrat, auf welches die Kompetenzen angewendet und woran sie erworben werden. Demgegenüber ist das Aufbauen und Verfügen-Können über begrifflich-kategoriales Wissen, etwa in Form von Konzepten, Kategorien und Prinzipien, Strukturierungs- und Ordnungssystemen konstitutiver Teil der Kompetenzmodellierung (Kompetenzbereich: historische Sachkompetenzen). Das begrifflich-kategoriale Wissen ermöglicht die Übertragung und Anwendung auf neue Fragen, Fälle und Probleme (Bräuer et al., 2016; Kühberger, 2012).

2.4.2 Englischsprachige Konzeptionen

Mit Blick auf die Prozesse und Räume inter- und transnationaler historischer Orientierung und Kommunikation stehen Konzeptionen für historisches Denken und Lernen vor der Anforderung, nicht nur vor dem jeweils „eigenen" nationalen Hintergrund hinsichtlich der jeweiligen Vergangenheit, der vorfindlichen Geschichts- und Erinnerungskultur sowie des herrschenden geschichtswissenschaftlichen wie erziehungswissenschaftlich-didaktischem Verständnisses begründet zu sein, sondern auch anschlussfähig zu werden an Konzeptionen im internationalen Rahmen. Im Folgenden werden daher einige Modellierungen historischen Denkens im englischen Sprachraum skizziert.

Ähnlich der Entwicklung im deutschsprachigen Raum wurden auch dort eher strukturelle Konzeptionen zu historischem Denken und Können erarbeitet (Keirn, 2016). Sie stehen neben den vornehmlich für schuladministrative Erfolgskontrollen entwickelten Entwürfen zur Überprüfung von Wissen und Können. In den USA, Kanada, Großbritannien und den Niederlanden entstanden Konzeptionen, welche die Überzeugungen der Lernenden zu der Natur von Geschichte und historischem Denken charakterisieren sollen (Lee, 1983; Lévesque, 2009; Seixas & Morton, 2013; Stearns, 1998). Im Zentrum stehen metatheoretische Einsichten, konkret das Verfügen über theorieförmige Konzepte „zweiter Ordnung", die nicht *in* der Geschichte auffindbar sind, sondern mit denen *über* Geschichte gedacht und kommuniziert werden kann. So forderte etwa Stearns (1998), dass Schülerinnen und Schüler einen „historical habit of mind" erwerben sollen, der es ihnen ermöglicht, „different magnitudes of historical change, different examples of conflicting interpretations, and multiple kinds of evidence" einzuschätzen (Stearns, 1998, S. 3). Daher sollten die Lernenden über „structural second order concepts" verfügen, also über Begriffe wie „evidence", „cause", „empathy", „change" und „time" (siehe auch Lee, 1983; Lévesque, 2009). Die konkreten Ausformungen dieser second order concepts variieren in den verschiedenen Ansätzen:

1. *Kanada:* Das *historical thinking*-Konzept der kanadischen Arbeitsgruppe um Peter Seixas (Ercikan, Seixas, Lyon-Thomas & Gibson, 2015; Seixas & Morton, 2013; Seixas, Gibson & Ercikan, 2015) fokussiert auf die Verfügung über sechs als besonders bedeutsam angesehene theoretische Meta-Konzepte (die sogenannten *Big Six*: *Historical Significance; Evidence; Continuity and Change; Cause and Consequence; Historical Perspectives; The Ethical Dimension*) und deren Berücksichtigung im Prozess des historischen Denkens (Seixas & Morton, 2013).
2. *Großbritannien:* In Großbritannien wird seit den 1980er-Jahren über "the key notion of evidence for or against a belief" (Rogers, 1980, S. 6) diskutiert. In Abgrenzung zum faktenorientierten Geschichtsunterricht ist „The New History" auf die second order concepts fokussiert, welche als „evidence skills" (Lee, 2014, S. 173) für die Konstruktion und Einordnung historischen Wissens genutzt werden sollen. Aktuell wird unter dem Zugriff „distracting dichotomies and restricted options" (Cain & Chapman, 2014, S. 112) über das Aufheben der Dichotomie zwischen *content* und *skills* – d.h. der Trennung von Fachwissen im Sinne von propositionalem Fakten- bzw. Deutungswissen über Vergangenes einerseits und einem Methodenkönnen im Sinne von Arbeitstechniken andererseits – zugunsten eines „disciplinary understanding" diskutiert.
3. *Niederlande:* Van Drie und van Boxtel (2008) stellen in ihrem Modell des *historical reasoning* mehrere für die argumentative Erstellung historischer Aussagen nötige Tätigkeiten bzw. Fähigkeiten vor und gewinnen hieraus Förderungs-, Forschungs- und Assessment-Fragen. Als Facetten definieren sie „describe change", „compare" und „explain". Dem werden folgende Operationen zugeordnet: "asking historical questions, use of meta-concepts, use of sources, use of substantive concepts, contextualization und argumentation".

4. *USA:* Die Forschung in den USA wurde wesentlich durch Wineburg (1991) inspiriert, der im Vergleich von Arbeitsweisen von Historikerinnen und Historikern und Laien Prozesse identifizierte, die sowohl generell für das Arbeiten mit "multiplen Dokumenten" relevant sind, als auch spezifisch historische Operationen adressieren. Zum einen kennzeichnet es die Arbeit von Historikerinnen und Historikern, dass sie sich nicht auf Informationen aus einzelnen Dokumenten verlassen, sondern ihre gewonnenen Informationen mit weiteren Dokumenten abgleichen und hierbei Unterschiede und Gemeinsamkeiten identifizieren, auf deren Grundlage bestimmte Aussagen belegt oder widerlegt werden (*corroboration*) (Rouet, Britt, Mason & Perfetti, 1996; Rouet, Favart, Britt & Perfetti, 1997; Wineburg, 1991). Bei der Interpretation und dem Vergleich unterschiedlicher Dokumente beachten Historikerinnen und Historiker Informationen zu den jeweiligen Autoren (*sourcing*). Während diese beiden Operationen beim Umgang mit multiplen Dokumenten generell eine wichtige Rolle spielen, bezieht sich die dritte von Wineburg identifizierte Operation (*contextualization*) auf die Einbettung der vermittelten Informationen in einen möglichst sinnvollen historischen Kontext. Die Fähigkeit zur Anwendung historischer Expertise auf neue Materialien spielt eine wichtige Rolle (Wineburg, 1991).

In den USA kann eine weitere Gruppe von konzeptionellen Ansätzen identifiziert werden, die sich aus der Auseinandersetzung mit dem erzählenden, schulbuchzentrierten Vorgehen in amerikanischen Schulen ergibt. Beginnend in den 1960er und 1970er Jahren setzte eine theoretische Diskussion über die Eigenständigkeit historischen Lernens (historical literacy) gegenüber einer übergreifenden, sozialwissenschaftlichen, durch überfachliche Kompetenzen gekennzeichneten Ausrichtung ein (Lee, 2005a, 2005b; Nokes, 2010a, 2010b, 2011; Perfetti, Britt & Georgi, 1995). Inhaltsbezogen wurden Diskussionen um Änderungen im Curriculum geführt, wobei konzeptuell einerseits die Frage der In- bzw. Exklusion verschiedener Bevölkerungsgruppen und ihrer Perspektiven im Zentrum standen (Keirn 2016) und andererseits – bezogen auf die Ergänzung der nationalen Curricula durch Weltgeschichte die Frage nach den dafür zu wählenden Konzeptionen (Zentrierung auf den Westen der Welt oder eine stärker interkulturelle Ausrichtung; Keirn 2016).

Beide Diskussionsstränge ließen „Historisches Denken" stärker in den Fokus rücken, vor allem die Prinzipien wissenschaftlicher Auseinandersetzung etwa mit Primärquellen und Darstellungen (Wineburg, 2001; Wineburg, Martin & Monte-Sano, 2013). Das Programm *Teaching American History* (TAH) etwa, in dem Lehrerfortbildungen zum Umgang mit Quellen im Geschichtsunterricht angeboten werden, wird seit 2001 von der Regierung der Vereinigten Staaten finanziell gefördert (Westhoff, 2009). Dies führte auch zu neuen Ansätzen in der Leistungsmessung. Das online verfügbare Angebot *Beyond the Bubble* der Stanford History Education Group (SHEG) bietet in enger Zusammenarbeit mit der Library of Congress den Lehrkräften quellenbasierte Schüleraufgaben (*History Assessments of Thinking,* HATs) an, die im Unterricht eingesetzt werden können und/oder ein formatives Assessment ermöglichen

(Breakstone, 2013; Breakstone, Smith & Wineburg, 2013). Die Schülerantworten werden im Hinblick auf das „historical knowledge" (z.B. Erfassung der Fakten), auf die „evaluation of evidence" (hierunter werden die drei Strategien sourcing, corroboration und contextualisation gefasst) und hinsichtlich der Argumentationsqualität (z.B. Kohärenz; Smith & Breakstone, 2015) ausgewertet.

Auch wenn im englischsprachigen Raum entwickelte bzw. in englischer Sprache publizierte „Frameworks" – „historical thinking" (z.B. Mandell, 2008) oder „historical reasoning" (z.B. van Drie & van Boxtel, 2008) – nicht als „Kompetenzmodelle" im Sinne der deutschen Bildungsdebatte und -forschung zu verstehen sind und auch nicht auf die Kompetenzdefinition von Weinert (2001) zurückgreifen, funktionieren sie ebenfalls als „models of cognition and learning in historical thinking" (Seixas & Ercikan, 2015, S. 1). Ähnlich wie in den Kompetenzmodellen der deutschsprachigen Forschung werden Zwecke, Ziele und Konzepte historischen Lernens definiert und als Grundlage für empirische Leistungsfeststellungen herangezogen. Die Diskussion um Anknüpfungs- und Übersetzungsmöglichkeiten wie auch Annäherungen zwischen den verschiedenen Ansätzen kennzeichnen die aktuelle Forschungslandschaft (Körber, 2016; Monte-Sano & Reisman, 2016; Seixas, 2016; Seixas & Ercikan, 2015).

Dass Anknüpfungspunkte zwischen englischsprachigen Konzepten und Aspekten des FUER-Modells bestehen, wurde deutlich. Die second order concepts z.B. sind im FUER-Modell in der Sachkompetenz verortet (vgl. Kap. 2.3.2.4), als epistemologische Prinzipien des Fachs. „Asking historical questions" im Sinne der Fragekompetenz nach FUER wird von niederländischer Seite und in Historical-Thinking-Konzepten betont. Hierzu heißt es in den *Historical Thinking Standards* (USA): Die Lernenden (1) „should raise questions, compare and contextualize information learned out of different sources to create a historical narrative by themselves", (2) „have to analyze the assumptions from which the narrative was created, and should assess the validity of the evidence presented" (National Center for History UCLA, 1996). Um die Förderung der Methodenkompetenz (Re- und De-Konstruktion), allerdings vorrangig in Bezug auf den re-konstruktiven Umgang mit Vergangenem, geht es auch in den an Wineburg, die Stanford-Group oder van Drie und van Boxtel anschließenden Ansätzen.

Offenkundig wird aber auch, dass in zwei Aspekten das FUER-Modell über die englischsprachigen Konzeptionen hinausgeht: mit dem als „Orientierung" konzeptualisierten Aspekt der Nutzung historischen Denkens für die eigene Lebenspraxis zum einen und mit der systematischen Ausdifferenzierung des Umgangs mit vorliegenden historischen Narrationen (De-Konstruktion) zum anderen.

2.4.3 Die Wahl des FUER-Modells als theoretische Grundlage für den HiTCH-Test – Zusammenschau der Gründe

Das FUER-Modell weist eine Reihe Charakteristika auf, die für die Wahl dieses Modells als theoretische Grundlage des HiTCH-Tests den Ausschlag gaben.

1) Das FUER-Modell gründet sich auf ein modellhaft ausgearbeitetes und an die narrativistische Geschichtstheorie zurückgebundenes Konzept des historischen Denkens, welches die gegenwärtige Orientierungsfunktion zum Ausgang nimmt und hieraus die für den Vollzug dieses Denkprozesses nötigen prozeduralen Kompetenzbereiche ableitet und in ihrer gegenseitigen Funktion beschreibt (Hasberg & Körber, 2003; Rüsen, 1983). Diese prozeduralen Kompetenzbereiche werden zusammengeführt mit dem kategorisierenden Verfügen-Können über Konzepte und Strukturen. Das FUER-Modell unterscheidet sich von den anderen deutschsprachigen wie internationalen Modellen im Hinblick auf die Breite des Konstrukts und die Tiefe der Begründung (vgl. Kapitel 2.4.2). Die theoretische Fundierung erleichtert es, einen reliablen und validen Test zu entwickeln; die Breite ist eine Voraussetzung dafür, dass ein auf dem Modell gegründeter Kompetenztest nicht nur einen Ausschnitt des Historisch-Denken-Könnens modelliert und abdeckt, sondern historisches Denken umfänglich messen kann.
2) Das FUER-Modell verfügt über ein theoretisch ausgearbeitetes Konzept unterschiedlicher Kompetenzniveaus und ist somit mit einem Graduierungsmodell verbunden. Basierend auf dieser Graduierungslogik können Förderperspektiven innerhalb der Kompetenzbereiche und bezogen auf historisches Denken als Ganzes eröffnet werden. Es ist somit geeignet als Grundlage auch für Weiterentwicklungen des HiTCH-Tests, in welchem Niveauunterschiede historischen Denken-Könnens erfasst werden sollen.
3) Das FUER-Modell wurde von einer größeren Gruppe von Expertinnen und Experten aus der deutschsprachigen Fachdidaktik wie auch von Schulpraktikerinnen und -praktikern aus insgesamt sieben Ländern entwickelt. Damit wurde das Modell schon im Entstehungsprozess in vielfacher Hinsicht „ausargumentiert" und „abgeschliffen". Zudem erweiterte sich der Blick auf Anforderungen unterschiedlicher Schulsysteme. Dies schafft eine günstige Ausgangslage für eine internationale Weiterentwicklung des HiTCH-Tests.
4) Vom FUER-Modell ausgehend wurden in den letzten Jahren Unterrichtskonzepte ausgearbeitet (vgl. die Reihe „FUER Geschichtsbewusstsein Themenhefte Geschichte" und das Projekt digital-multimediales Schulbuch (mBook) für die Sekundarstufe I und II des Gymnasiums und für heterogene, inklusive Klassen der Sekundarstufe I (Schreiber, Sochatzy, Ventzke, 2013; Sochatzy, 2016; Ventzke, Sochatzy & Schreiber, 2013, 2014). Zudem wurde in den vergangenen Jahren das FUER-Modell einigen Bildungsplänen komplett (z.B. Thüringen, Deutschsprachige Gemeinschaft in Belgien) oder in wesentlichen Teilen (z.B. Österreich, Sachsen, Baden-Württemberg, Bayern) zugrunde gelegt. Die praktische Realisierbar-

keit und die Akzeptanz durch Lehrkräfte sowie Schülerinnen und Schüler ist eine weitere Bedingung für eine gelingende Testentwicklung.

5) Mehrfach hat das FUER-Modell sich bereits als Grundlage für die Empirie bewährt (Bertram, Wagner & Schaser, 2015; Bertram, Wagner & Trautwein, 2013, 2014; Borries, Fischer, Leutner-Ramme & Meyer-Hamme, 2005; Körber et al., 2007; Kühberger & Mellies, 2009; Lehmann et al., 2016; Meyer-Hamme, 2007, 2009; Schreiber, Schöner & Sochatzy, 2012; Sochatzy, 2016; Sochatzy & Merkt, 2015; Ventzke, 2016).
6) Perspektivisch ist das FUER-Modell über die kognitiven Anteile hinaus offen für die ästhetischen und emotionalen Dimensionen des historischen Denkens und Lernens (Borries & Meyer-Hamme, 2014; Lehmann, 2015). Diese Dimensionen sind nicht nur pragmatisch bedeutsam (Lehmann, 2015; Schreiber, Kraus, Lehmann & Zabold, 2015); sie können im Hinblick auf alle Kompetenzbereiche theoretisch modelliert werden (Lehmann et al., 2016).[8] Dies lässt in Zukunft auch die empirische Überprüfung möglich erscheinen.

2.5 Forschungsübersicht: Standardisierte Testungen und empirische Studien zu historischem Denken

Standardisierte Testungen im Fach Geschichte sind in Deutschland nicht etabliert (Köster et al., 2014; für Deutschland darin Bracke, Flaving, Köster & Zülsdorf-Kersting, 2014). Hierzu trägt auch eine weithin gehegte (und auch von den Autorinnen und Autoren des HiTCH-Tests lange Zeit geteilte) Skepsis bei, mit solch standardisierten Aufgaben könne der Spezifik historischen Denkens und des diesbezüglichen Könnens nicht Rechnung getragen werden (Körber, Borries, Pflüger, Schreiber & Ziegler, 2008).[9]

Um den Forschungsstand mit Ansatzpunkten für die Entwicklung von Testaufgaben darzustellen bzw. zu verdeutlichen, wo Neuland in der Aufgabenentwicklung zu beschreiten war (vgl. Kapitel 3), werden im Folgenden in der gebotenen Kürze zu-

8 Z.B. können sich in der Sachkompetenz unterschiedliche Emotionskonzepte (Scham, Trauer ...) zeigen, in der Frage- und Orientierungskompetenz drücken sich die emotionalen Anteile der eigenen Orientierungsbedürfnisse und Interessen aus, in den Methodenkompetenzen (Re- und De-Konstruktion) zeigt sich die Wahrnehmung historischer Alterität und Entwicklung auch in emotionaler und ästhetischer Hinsicht und in der Orientierungskompetenz drückt sich die Anerkennung und Reflexion der auch affektiven und emotionalen Dimension von historischer Selbst- und Fremdwahrnehmung (Zugehörigkeit, Abgrenzungsbedürfnisse etc.) aus.

9 Allerdings konnten etwa Schönemann, Thünemann und Zülsdorf-Kersting (2010) an Hand von Abiturklausuren der Leistungskurse in Nordrhein-Westfalen zeigen, dass schulpraktisch weit verbreitete Formen der Leistungserfassung das selbstständige historische Denken von Geschichtslernenden ebenfalls nur unzureichend erfassen. In den Abiturarbeiten dominierte die Wiedergabe historischer Kenntnisse, wogegen die historisch-kritische Auseinandersetzung mit Quellen und Darstellungen bzw. die eigenständige Auseinandersetzung mit Fragestellungen nebensächlich blieben. Trotz dieser mangelnden Fokussierung auf die in den Einheitlichen Prüfungsanforderungen für die Abiturprüfung (vgl. Kapitel 2.4.1) durchaus geforderten historischen Denkleistungen, konnten Schülerinnen und Schüler mit solchen reproduktiven Leistungen zum Teil sehr gute Noten erzielen.

nächst deutschsprachige Ansätze aus der Zeit vor bzw. nach dem Einsetzen der Kompetenzorientierung charakterisiert, bevor der Blick auf außereuropäische Studien, vornehmlich wieder aus dem englischsprachigen Raum, gerichtet wird.

2.5.1 Deutschsprachige Studien ohne Bezug zur Kompetenzorientierung

2.5.1.1 Paradigma 1: Standardisierte Tests zu historischem Wissen und historischen Kenntnissen

Die ersten Versuche, die Ergebnisse des schulischen Geschichtslernens mit offenen wie auch standardisierten Testformaten der Messung zugänglich zu machen, standen in Deutschland im Zusammenhang mit der Schulreform der 1970er Jahre („Curriculumdebatte"). Dabei wurden meist auf einer sehr konkreten Ebene Inhaltskenntnisse abgefragt (Feiks, Laubert & Rothermel, 1975; Ingenkamp & Mielke, o.J.; Marz, Arnold & Reischmann, 1987; Mielitz, 1969; Oehler, 1969). Davon setzte sich der Ansatz von Borries (1974) ab, in dem fachspezifische Auswertungs- und Anwendungsaufgaben im Hinblick auf historische Materialien (Quellen oder Darstellungen in Form von z.B. Texten, Bilder, Grafiken) formuliert wurden, die Transferübungen und Methodenzugriffe umfassten, so dass man diese Aufgaben als Vorläufer der heutigen Kompetenztests betrachten kann.

Auch in jüngerer Zeit wurde versucht, inhaltliches Wissen mit geschlossenen Aufgaben zu erheben, um daraus Einschätzungen zum aktuellen Geschichtsunterricht und Konsequenzen für seine Weiterentwicklung abzuleiten. Großes Medienecho erreichten die Studien von Deutz-Schroeder und Schroeder (2007) sowie Schroeder, Deutz-Schroeder, Quasten und Schulze Heuling (2012), weil sie Mängel bei zeitgeschichtlichen Kenntnissen und Urteilen bezogen auf die NS-Zeit, die DDR und die Bundesrepublik Deutschland aufzeigten. Die teilweise heftige Kritik an den Studien wie an ihrer politischen Rezeption machte dabei deutlich, welche theoretischen und methodischen Probleme Versuche mit sich bringen, punktuelle Zustimmungen zu komplexen, voraussetzungsvollen Aussagen, die zudem Kenntnisse, Sach- und Werturteile vermengen, als valide und reliable Messungen von (orientierungsrelevantem) historischem „Wissen" zu interpretieren (Borries, 2007a, 2013; Meyer-Hamme, 2016; Siebeck, 2013). Eine Erfassung eigenständigen kritischen Könnens im Bereich des historischen Denkens war mit den beiden Studien gar nicht erst geplant.

Mit Blick auf die empirische Untersuchung historischer Kompetenzen liegt die Bedeutung solcher Studien und vornehmlich der Diskussion um sie darin, die Notwendigkeit einer sauberen konzeptionellen Definition unterschiedlicher Wissensformen (vgl. Kapitel 2.2.3.5) und die Fokussierung auf transferables Wissen verdeutlicht zu haben, wie es den Kompetenzbereich der „Sachkompetenz" des FUER-Modells charakterisiert (vgl. Kapitel 2.3.2.4).

2.5.1.2 Paradigma 2: Standardisierte Tests zu Geschichtsbewusstsein

Einen anderen Strang eröffneten ab den 1980er Jahren Studien, die sich auf das Konzept Geschichtsbewusstsein bezogen. Geschichtsbewusstsein wurde anfangs weniger mit historischem Denken in Verbindung gebracht, über das die Schülerinnen und Schüler – vermittelt durch den Geschichtsunterricht – lernen sollten zu verfügen. Vielmehr wurden darunter Haltungen, Einstellungen (Vorlieben, Überzeugungen, Selbstdefinitionen) und Dispositionen als Ergebnisse historischer Sozialisation und Enkulturation in Familien, Massenmedien, Schule verstanden (u.a. Borries, 1988; Borries, Pandel & Rüsen, 1991; Borries & Rüsen, 1994).

Erst in den folgenden, maßgeblich von von Borries getragenen, zum Teil international vergleichend angelegten Studien[10] wurden auch „Fähigkeiten" (Konzeptbeherrschung, Transferleistungen, Denkstrategien) in den Blick genommen (Angvik & Borries, 1997; Borries, 1995; Borries et al., 2005). Insbesondere in den 1997 und 2005 publizierten Untersuchungen wurden Aufgaben vorgelegt, die die Vorstufe zu Kompetenzaufgaben zu historischem Denken darstellen, z.B. der Vergleich von Schulbuchauszügen oder von widersprüchlichen Darstellungen zum gleichen Thema (Methodenkompetenz: De-Konstruktion), Aufgaben zu einem „experimentellen Perspektivenwechsel" (Methodenkompetenz: Re-Konstruktion) oder zu zentralen historischen Begriffen (Sachkompetenz). Allerdings fehlte noch eine einordnende kompetenztheoretische Grundlegung.

Dies gilt auch in Bezug auf die Aufgaben zur historischen Orientierung. Ausgehend von der Unterscheidung Jeismanns (1977) sollten die Jugendlichen z.B. vorgegebene Statements im Hinblick auf Vergangenheitsdeutungen, Gegenwartswahrnehmungen und Zukunftserwartungen einschätzen (Angvik & Borries, 1997). In der Summe ging es um die Frage, inwiefern historischen Phänomenen und den mit ihnen verbundenen oder zu verbindenden Sinnbildungen (vgl. Kapitel 2.2.3.4) Bedeutung für die eigene Lebenspraxis zugewiesen wird. Es ging somit um historische Orientierung, nicht aber um Orientierungskompetenz, also um die Fähigkeit, im Prozess des historischen Denkens eine historische Orientierung in der Zeit herzustellen.

2.5.2 Kompetenzorientierte Studien im deutschsprachigen Raum

2.5.2.1 Studien mit dem FUER-Modell als theoretische Grundlage

Das Kompetenzstrukturmodell der FUER-Gruppe hat bereits an verschiedenen Stellen als Empirie-Grundlage gedient. Diese sollen – aufgeschlüsselt nach Kompetenzbereichen – im Folgenden kurz vorgestellt werden.

10 Nach dem Fall des Eisernen Vorhangs in Europa (1989–1991) wurde in der internationalen Studie „Youth and History" (Angvik & Borries, 1997), die in 27 europäischen Ländern und mit einer Stichprobe von mehr als 31.500 Schülerinnen und Schüler durchgeführt wurde, das Geschichtsbewusstsein und die politischen Einstellungen europäischer Jugendlicher (15-Jährige) untersucht.

1) Bereits während der Entwicklungsphase des FUER-Modells wurden in Hamburger Studien zum Schulbuchverständnis und zu Reflexionsprozessen im Geschichtsunterricht (Borries et al., 2005) Methodenkompetenzen (Re- und De-Konstruktion) von Schülerinnen und Schülern empirisch erfasst. Dabei kamen vorrangig geschlossene Aufgaben zum Einsatz, die anhand eines Vergleichs von drei Schulbuchauszügen de-konstruktive Prozesse fokussierten und zugleich am thematischen Beispiel der Christianisierung Mitteleuropas zu re-konstruktiven Denk-Operationen aufforderten. Es konnte festgestellt werden, dass beträchtliche Widersprüche zwischen den Schulbuchauszügen (für die 7. Jahrgangsstufe) nur von einer Minderheit der Schülerinnen und Schüler erkannt wurden, so dass der Eindruck erhärtet wurde, dass die Schulbuchauszüge in ihrer historisch argumentativen Tiefenstruktur kaum verstanden wurden.
2) In einer ergänzenden Studie von Meyer-Hamme (2005) konnten – basierend auf einer quantifizierten qualitativen Kodierung von 415 Schüleressays – re- und de-konstruktive Prozesse durch eine Faktorenanalyse getrennt werden, so dass empirisch erhärtet werden konnte, dass de-konstruktive Prozesse den Lernenden unter den damaligen Bedingungen (d.h. ohne explizite Kompetenzorientierung im Unterricht) deutlich schwerer fielen als re-konstruierende.
3) In einer Studie von Bertram, Wagner und Schaser (2015) wurden ebenfalls Methodenkompetenzen (Re- und De-Konstruktion) von Schülerinnen und Schülern untersucht (N = 190). Offen formulierte Aufgaben zielten auf re- und de-konstruierende Leistungen unter Bezug auf Quellen und Darstellungen. Mit Hilfe einer theoriebezogenen Kodierung konnten in Faktorenanalysen und Korrelationsberechnungen Zusammenhänge zwischen den in den offenen Instrumenten erfassten Methodenkompetenzen (Re- und De-Konstruktion) und den in standardisierten Instrumenten erfassten Sachkompetenzen der Lernenden aufgezeigt werden.
4) Die Basisoperation der De-Konstruktion und das Fördern von historischen Kompetenzen standen im Zentrum eines Projekts zur Schulbuchanalyse (Schreiber et al., 2012), bei dem auf dem Weg kategorisierender qualitativer Inhaltsanalysen ein umfassendes Codebuch zur Analyse historischer Narrationen erarbeitet wurde. Für die Analysen wurden u.a. die Fokussierungen auf Vergangenheit, Geschichte und Gegenwart/Zukunft genutzt (vgl. Kapitel 2.3.1.1). Zur abschließenden Beurteilung der Narrationen wurde auf die Triftigkeitskriterien nach Rüsen (vgl. Kapitel 2.2.3.6) Bezug genommen.
5) Darauf fußend stand in einem internationalen Schülerprojekt die Förderung von Methodenkompetenz (De-Konstruktion) im Mittelpunkt. Hierfür wurde das unter (4) vorgestellte Analyseraster elementarisiert und von den Schülerinnen und Schülern für vergleichende Schulbuchanalysen genutzt, die auf De-Konstruktion zielten (Schreiber & Gruner, 2010). Es konnte belegt werden, dass Schülerinnen und Schüler durch De-Konstruktionsleitfäden dabei unterstützt werden, zu eigenständigen Einschätzungen der zu vergleichenden Schulbücher zu kommen.

6) Auch Kühberger (2013) untersuchte die Methodenkompetenz (De-Konstruktion) von Lernenden anhand ihres Umgangs mit filmischen Darstellungen. 155 Essays von Schülern der 7. Schulstufe wurden unter anderem im Hinblick auf die Nutzung von empirischen, narrativen, normativen oder metareflexiven Konzepten untersucht. Das letztgenannte Konzept dient zur Klärung, ob „theoretische Überlegungen [wie] z.B. Quellenkritik, methodischer Umgang mit Quellen/Darstellungen" (S. 45) für die Einschätzung des Konstruktcharakters des Films genutzt werden. Dabei stellte sich heraus, dass empirische Kriterien (z.B. Kritik an der Kleidung der Darsteller), gefolgt von narrativen (z.B. Anmerkungen zur Schauspielerei) am häufigsten genutzt, während die metareflexive Ebene selten und normative Aspekte gar nicht angesprochen wurden.
7) Im Bereich der Unterrichtsforschung lässt sich ein Projekt von Annemarie Kraus (2013) verorten. Sie überprüfte in einer qualitativen Wirksamkeitsstudie, inwiefern mit Hilfe eines Unterrichtskonzepts zur Zeitzeugenbefragung im Kontext Mauerfall, die Weiterentwicklung von Fragekompetenz gefördert werden konnte (Kraus, 2013). Effekte konnten sowohl auf die Kernkompetenzen „historische Fragen stellen können", als auch auf die Kompetenzen „Fragestellungen erschließen können" nachgewiesen werden.
8) Bertram et al. (2014, 2015, im Druck) schließlich entwickelten im Rahmen einer randomisierten Interventionsstudie zur Wirksamkeit von Zeitzeugenbefragungen im Geschichtsunterricht einen standardisierten Kompetenztest, der die Einsicht der Lernenden in die epistemologischen Prinzipien von Geschichte erfasst. Dieses Instrument hat sich in mehreren Erhebungen mit Schülerinnen und Schülern der oberen Sekundarstufe I (N = 311 und N = 962) und Studierenden (N = 360 und N = 544) bewährt. Exploratorische Faktorenanalysen ergaben in allen Erhebungen ein sehr ähnliches Verteilungsmuster mit drei Faktoren, die benannt wurden als: Einsicht, (1) dass die Geschichte eine Re-Konstruktion der Vergangenheit ist, Einsicht, (2) dass Darstellungen de-konstruiert werden müssen, und Einsicht, (3) dass Zeitzeugenaussagen Charakteristika von Quellen und Darstellungen zugleich aufzuweisen.[11]

11 Nur nebenbei angemerkt sei, dass das FUER-Modell inzwischen auch einigen Arbeiten zum Kompetenzerwerb im Geschichtsstudium zugrunde gelegt wird. Mit Bezug auf die Fragekompetenz ist dies der Fall in laufenden Arbeiten von Nicola Brauch. Zur empirischen Erfassung und Messung geschichtswissenschaftlicher Fragekompetenz liegen bislang theoretische Beiträge vor (Brauch, 2015, 2016) sowie erste exploratorische empirische Studien. Studierende eines Proseminars zur mittelalterlichen Geschichte der Universität Kiel (N = 43) erhielten im Verlauf des Semesters Einführungen in die Erarbeitung geschichtswissenschaftlicher Fragestellungen. Die nach dem Semester verfassten Hausarbeiten wurden inhaltsanalytisch in den Einheiten Einleitung und Schlussteil untersucht, um zu Indikatoren zur Identifizierung von Qualitätsmerkmalen der Fragestellungen zu gelangen. Den Daten ließen sich erste Hinweise auf einen Zusammenhang zwischen der Eigenständigkeit der Quellenauswahl für die Hausarbeit und der Qualität der Fragestellung entnehmen. Diese Hinweise werden seit Januar 2016 im Dissertationsprojekt von Lena Behrendt (Bochum) im Lehr-/Lernsetting des Schülerlabors Geisteswissenschaften der Ruhr-Universität Bochum weiterverfolgt. Des Weiteren haben in Hamburg vorbereitende Arbeiten zu Ermöglichung eines Vergleichs von Schülerinnen und Schüler- und Studierendenleistungen stattgefunden (Sahm, 2015).

2.5.2.2 Weitere deutschsprachige Studien

Neben den das FUER-Modell nutzenden Untersuchungen gibt es eine Reihe von Untersuchungen, welche historisches Denken-Können bzw. ausgewählte Aspekte davon mit Hilfe anderer Konzeptionen untersuchen. Einige davon seien kurz skizziert:

1) Der Projektverbund *narratio* (Jan Hodel, Holger Thünemann, Monika Waldis-Weber, Meik Zülsdorf-Kersting) hat auf Basis von 186 Schülernarrationen der 9. und 10. Jahrgangsstufe untersucht, zu welchen historischen Denkprozessen diese in der Lage sind und inwiefern die Qualität historischen Denkens von der Themenwahl und der Textgattung für die Narration abhängen. Dabei zeigen sich u.a., dass beim geschichtskulturell präsenten Thema Nationalsozialismus („Judenboykott") normative Aussagen überwogen und kaum Bezug zu den vorgelegten Materialien hergestellt wurde, während bei einem weitgehend unbekannten Thema („Japanische Handelsbeziehungen im 18. Jh.") signifikant mehr Materialbezüge und Aussagen zu Sachverhalten gemacht wurden. Darüber hinaus wurden Unterschiede in Materialbezügen und Werturteilen auch in Abhängigkeit zum für die Narration gewählten Textgattung (wie Blog-Eintrag oder Artikel in Schülerzeitung) sichtbar (Hodel, Waldis, Thünemann, Zülsdorf-Kersting & Ziegler, 2015; Hodel, Waldis, Zülsdorf-Kersting & Thünemann, 2013).[12]
2) Der Auseinandersetzung mit epistemologischen Prinzipien zuzuordnen ist die Untersuchung von Hartmann (2008), die historische Perspektivenübernahme mit standardisierten Aufgabenformaten erfasst hat. Ausgehend von einem psychologischen Entwicklungsstufenmodell, wurden standardisierte Instrumente zur Untersuchung der historischen Perspektivenübernahme – verstanden als eine Kompetenz historischen Verstehens – entwickelt. Schülerinnen und Schüler der 6. und 10. Jahrgangsstufe des Gymnasiums (N = 375) beurteilten auf der Basis von Likert-Skalen historische Szenarien, die drei Aspekte historischer Perspektivenübernahme abbildeten: Gegenwartsfixierung, Rolle des historischen Akteurs und historische Kontextualisierung. Die zentralen Ergebnisse dieser Arbeit zeigen beispielsweise eine altersabhängige Fähigkeit zur Perspektivenübernahme und plausible korrelative Zusammenhänge mit Außenkriterien (z.B. Geschichtsinteresse oder Perspektivenübernahme im Alltag).
3) In ihrer qualitativen Untersuchung zur Sinnbildung Lernender auf der Basis von Bildquellen hat Kristina Lange (2011) herausgefunden, dass Schülerinnen und Schüler in Antagonismen denken. Bei der Übertragung eines dualistischen Gesellschaftsmodells des „oben" und „unten" oder „reich" und „arm" auf alle Zeiten wurde der Unterschied zwischen verschiedenen historischen Epochen nivelliert,

12 Zum selben Schluss kommt Hartung, welcher mittels eines hermeneutischen Zugangs die Abhängigkeit der Schreibleistungen vom Textgenre anhand einer umfangreichen Datenbasis von 229 Texten zur Weimarer Republik nachwies (Hartung, 2013, 2015). Auch die Erhebungen bei angehenden Geschichtslehrpersonen (Nitsche & Waldis, 2016; Waldis, Marti & Nitsche, 2015) verweisen auf den Einfluss des Textgenres sowie der Auswahl und Anordnung der Quellen auf das historische Erzählen bzw. Argumentieren.

wobei die Lernenden die Bilder in der Regel „als Abbild der historischen Vergangenheit“ (Lange, 2011, S. 266) verstanden.

4) Zum methodenbezogenen Konzept „Quellen und Darstellungen“ liegen ebenfalls mehrere Studien vor. Sie befassen sich insbesondere damit, inwiefern Schülerinnen und Schüler den grundlegenden Unterschied zwischen Quellen und Darstellungen überhaupt verstanden haben, ausgehend von der Annahme, dass sie deren Aussagemöglichkeiten und -grenzen nur dann angemessen einschätzen können. Sowohl Martens (2010), bezogen auf Lernende der 6., 8. und 10. Jahrgangsstufe, als auch Schönemann, Thünemann und Zülsdorf-Kersting (2010), im Hinblick auf Abiturientinnen und Abiturienten im Zentralabitur 2008 in NRW, stellten fest, dass die Lernenden oft schon an der Unterscheidung von Quellen und Darstellungen scheiterten.

2.5.3 Kompetenzorientierte Studien im englischsprachigen Raum

Vielen Studien aus dem angloamerikanischen Raum liegen Konzeptionen des *historical thinking* oder *historical reasoning* zugrunde (vgl. Kapitel 2.4.2). Sie fokussieren *second order concepts*, u.a. mit Methodenbezug. Eine Sonderrolle nehmen die in englischer Sprache publizierten niederländischen Studien zu Fragekompetenz ein.

1) Wineburg (1991, 2001) und Kolleginnen und Kollegen (Baron, 2012; Rouet et al., 1996) haben die Arbeitstechniken von Forschenden in der Geschichtswissenschaft als Ausgangspunkt genommen, um in einem Laien-Experten-Vergleich die zentralen kognitiven Prozesse bei der Analyse und der Interpretation von historischem Material herauszuarbeiten. Nach einer Inhaltsanalyse Analyse wurden die Arbeitsstrategien des *sourcing*, der *corroboration* und *contextualisation* mit Hilfe von faktorenanalytischer Dimensionsuntersuchungen voneinander getrennt (vgl. Kapitel 2.4.2). Ausgehend von Wineburgs Ergebnissen untersuchten Rouet et al. (1996) in einer Experimentalstudie, wie Studierende als „historische Novizen“ ($N = 24$) eine Sammlung von Dokumenten zu einer historischen Kontroverse mit und ohne Quellen in einem Essay verarbeiteten. Die drei von Wineburg identifizierten Strategien konnten bestätigt werden. Es gelang den Schülerinnen und Schülern meist, die kontroversen Historikeraussagen entsprechend der corroboration-Strategie materialbasiert zu vergleichen und zu beurteilen. Hierbei hatten die Quellen und Erlebnisberichte den höchsten Glaubwürdigkeitsstatus inne.[13]

13 Baron (2012) ging in ihrer Studie zur Erforschung eines historischen Ortes durch Historikerinnen und Historiker ebenfalls von den drei Strategien Wineburgs aus und modifizierte und erweiterte diese. Die zusätzliche Strategie *supposition* meint, dass historisch Forschende bei einem Fehlen bestimmter Informationen in einem kontrollierten Vorgehen nach plausiblen Erklärungen suchen. Die Strategie *empathic insight* hingegen spricht die Fähigkeit von Historikerinnen und Historikern an, in der Reaktion auf den physisch erfahrenen Ort das soziale, emotionale und intellektuelle Erleben der Menschen, die früher an diesem Ort agierten, zu imaginieren. Basierend auf einem umfassenden Wissen über die vergangenen Lebensbedingungen scheinen professionelle Historikerinnen und Historiker zu einer experimentellen Perspektivenübernahme und einem systematischen Perspektivenwechsel fähig zu sein (Baron, 2012).

2) Weniger ein abgeschlossenes Forschungs- als ein auch auf Unterrichtspraxis bezogenes Entwicklungsprojekt mit Forschungsrelevanz ist das ebenfalls von Wineburg geleitete Projekt Beyond the bubble (https://beyondthebubble.stanford.edu/) an der Stanford University in den USA. Dort werden unter anderem für die Lehrpersonen geeignete diagnostische Aufgabenformate entwickelt, die sie bei ihrer Unterrichtsplanung wie bei der Leistungsüberprüfung unterstützen sollen. Den fachdidaktisch Forschenden eröffnen die Texte einen tieferen Einblick in die historischen Denk- und Arbeitsprozesse von Jugendlichen. Lernende legen ihre Denkschritte über die think-aloud Methode offen oder verfassen – durch kompetenzorientierte Fragen angeleitet – Kurzessays zu zentralen historischen Quellen und Darstellungen, die dann inhaltsanalytisch ausgewertet werden.
3) Studien zur historischen Fragekompetenz wurden in den letzten Jahren überwiegend in Lehr-/Lernsettings durchgeführt. Dabei sind u.a. die Arbeiten des niederländischen Geschichtsdidaktikers Albert Logtenberg (2012) von Interesse, der in seiner Dissertationsschrift auch den Bezug zum FUER-Modell herstellt und mögliche Anschlüsse an das niederländische Modell aufzeigt (Logtenberg, 2012). Er stützte seine Forschung auf das Framework of Historical Reasoning (Van Drie & Van Boxtel, 2008), das – ähnlich wie das FUER-Modell – die Fähigkeit „asking historical questions" als eine zentrale kognitive Anforderung begreift und sich dabei auch auf die Theorie Jörn Rüsens bezieht. Aufgegriffen werden auch Diskurse der Kognitionspsychologie (Chin & Chia, 2004; King & Kitchener, 1994) und der Theoriebildung in den USA (v.a. Sam Wineburg, Bruce VanSledright), in Kanada (v.a. Peter Seixas) und in England (v.a. Peter Lee). Van Drie und van Boxtel unterscheiden Fragen, deren Antwort eine historische Beschreibung bzw. eine Erklärung, einen Vergleich oder eine Bewertung erforderlich machen. Die drei Studien Logtenbergs zum historischen Fragen konzentrierten sich auf die Erforschung des (Geschichts-)Unterrichts mit 15 bis 16-jährigen Lernenden. Ein Befund ist z.B., dass problematisierende und erzählende Texte die Schülerinnen und Schüler stärker zu auf Vergleich zielenden Fragen motivierten als sachliche Texte. Es wurde etwa dieselbe Menge an *higher-order* (lange Antwort erforderlich) wie *lower-order Fragen* (kurze Antwort erforderlich) generiert. Darunter waren die meisten Fragen des higher-order Typs solche, die eine Erklärung oder Beschreibung als Antwort erwarteten. Das Ergebnis ist für künftige Studien deshalb interessant, weil sich das Anspruchsniveau der Fragetypen wie folgt ableiten ließe: am leichtesten waren auf Beschreibung zielende Fragen (469/729: 64,3%), gefolgt von solchen, die eine Erklärung (26,2 %), Bewertung (3,8%) oder einen Vergleich (2,2%) erwarteten.
4) Ein inzwischen vielfach eingesetztes und adaptiertes Instrument zur Erfassung epistemologischer Überzeugungen zu Geschichte hat 2010 Liliana Maggioni vorgelegt: „Beliefs about History Questionnaire (BHQ)" (Maggioni, 2010; Maggioni,

Alexander & VanSledright, 2004; Maggioni, VanSledright & Alexander, 2009).[14] Auf einer sechsstufigen Likert-Skala wird dabei die Zustimmung von Schülerinnen und Schülern, Studierenden und Expertinnen sowie Experten zu verschiedenen Aussagen zur Wissengenese und -beschaffenheit in Geschichte erhoben und die sich abbildenden Muster zur Bildung epistemologischer Typen auf der Basis faktorenanalytischer Dimensionsuntersuchungen genutzt: Unterschieden wurden auf diese Weise „Objektivisten", „Subjektivisten" und „Kriterialisten". Verschiedene Studien, welche den BHQ in Interventionssettings einsetzten, verweisen auf eine beträchtliche Inkonsistenz des Antwortverhaltens der Schülerinnen und Schüler in Abhängigkeit zur Situation und zum Aufgabenkontext (vgl. Maggioni et al., 2009). Dies ist auch dann der Fall, wenn die Intervention darauf ausgerichtet war, Schülerinnen und Schüler beim Erwerb von solideren epistemologischen Überzeugungen zu unterstützen (VanSledright & Reddy, 2014). Auf das Vorliegen inkonsistenter und situationsabhängiger Überzeugungen verweisen auch die Ergebnisse im Projekt Chata (z.B. Lee, 2004; Lee & Ashby, 2000; Lee & Shemilt, 2003), in welchem Kinder und Jugendliche zum Argumentieren über historische Erzählungen bzw. zum Nachweis von Evidenz und Ursächlichkeit/Begründung angeleitet wurden.

5) In Bezug auf inhalts-, methoden-, theoriebezogene Begriffs- und Strukturierungskonzepte wird in der englischsprachigen Literatur insbesondere diskutiert, inwiefern sie zu anspruchsvollerem historischen Argumentieren bei Lernenden führen. Beispielsweise zeigten Halldén und andere (1997), dass Schülerinnen und Schüler keinen Blick für strukturelle Ursachen haben, wenn sie Ursachen als „Grund und Anknüpfungspunkt für Handlungen" verstehen. Rouet et al. (1997) konnten nachweisen, dass Schülerinnen und Schüler, die über ein Verständnis von meta-historischen Konzepten verfügen, elaboriertere Argumentationsstrategien verwendeten und sich auf den vertiefenden Vergleich von möglichen Interpretationen einließen. In einer kürzlich veröffentlichten experimentellen Studie von Stoel, van Drie und van Boxtel (2015) ergab sich eine signifikante Korrelation zwischen der Qualität des kausalen historischen Argumentierens (*causal historical reasoning*) in Aufsätzen und der bewussten Kenntnis von Strukturierungskonzepten zum Aspekt der „historischen Ursache".

14 Im deutschsprachigen Raum wurde der BHQ in jüngster Zeit faktorenanalytisch und auf Zusammenhänge mit historischem Argumentieren (Mierwald, Seiffert, Lehman & Brauch, 2016) und Ausprägungen geschichtsdidaktischer Kompetenzen untersucht (Mierwald et al., 2016; Nitsche & Waldis, 2016).

2.5.4 Large-Scale-Assessements in Frankreich und den USA: Die Entwicklung der standardisierten Tests CEDRE und NAEP

Neben einigen Messungen von Schülerleistungen in Large-Scale-Formaten, die vor allem die Verfügung über Wissen und Deutungen sowie Anwendungen historischer Denkformen im Rahmen engerer curricularer Vorgaben erfassen, gibt es auch Large-Scale-Messungen, welche – zumindest dem Anspruch nach – Kompetenzen adressieren.

Dies gilt etwa für Frankreich, wo die Leistungen von Schülerinnen und Schülern im Rahmen des CEDRE (Le cycle des évaluations disciplinaires réalisées sur échantillon) regelmäßig überprüft werden. Auch wenn wegen der Geheimhaltung der Testaufgaben keine Beispielitems publiziert werden, kann anhand der Beschreibung der Leistungen der Gruppe 3 (überdurchschnittliche, aber nicht brillante Schülerinnen und Schüler) ein Eindruck von der Art der Testung gewonnen werden, die sich ausdrücklich auf „compétences" bezieht. Diese Analyse ergibt, dass unter Kompetenzen einerseits eine Mischung curriculumkonformer Reproduktionsleistungen sowie eher fachunspezifischer Arbeitstechniken (z.B. Informationsentnahme, Textvergleich) und Materialumgangsweisen (z.B. Kartenbenutzung, Organigramme) verstanden werden. Andererseits werden Kompetenzen unter dem Stichwort der „Methodenorientierung" jenseits der Curriculumthemen („n'étant pas au programme") gefasst. Das Vorgehen beim Testen scheint sich an PISA zu orientieren. Erfasst wird nur ein Teil der Kompetenzbereiche nach der breiten Definition des FUER-Modells.

In den Vereinigten Staaten werden im Rahmen von NAEP (National Assessment of Educational Progress) seit den 1960er Jahren regelmäßig Lernende verschiedener Jahrgangsstufen (4., 8. und 12. Klasse) bundesweit getestet. Dabei sollen nicht nur ihre Kenntnisse zur US-amerikanischen Geschichte, sondern auch die Denkoperationen des *knowing* und *thinking* history erfasst werden. Während unter *knowing* (genauer: *historical knowledge and understanding*) historisches Wissen, aber auch Einsichten verstanden werden, werden unter *thinking* (genauer: *historical analysis and interpretation*) Kompetenzen historischen Denkens gefasst, wie z.B. die Herstellung von Kausalbeziehungen oder das Abwägen von Beweisen (http://nagb.org/publications/frameworks.htm/). Bei einer genaueren Analyse dieser Aufgaben zeigt sich aber, dass trotz der anders lautenden Ansprüche überwiegend eher Daten- und Faktenwissen abgefragt werden, so dass die meisten Aufgaben keinen Einblick in das historische Denken von Jugendlichen ermöglichen (Bertram & Wagner, in Druck). VanSledright (2014) kritisiert diese Form des Assessments deutlich: „In fact, I suggest that the typical testing approaches we have used for decades do precious little to make sense of student learning" (S. IX).

2.5.5 Der Stand (inter-)nationaler Ansätze zum Assessment historischen Denkens

Aus der Zusammenschau der empirischen Forschungen zum historischen Denken wird deutlich, dass die Entwicklung des HiTCH-Tests in einigen Hinsichten durchaus anschlussfähig an empirische Forschungen und Assessment-Konzepte sowohl im deutschsprachigen als auch im internationalen Rahmen ist.

Innerhalb dieses Spektrums besetzt er allerdings eine besondere Position. Dies hängt mit seiner Orientierung an einem Kompetenzbegriff zusammen, der die tatsächliche Denk- und Orientierungstätigkeit der Schülerinnen und Schüler ohne einen expliziten und notwendigen curricularen Bezug und somit fokussiert auf die transferablen Fähigkeiten erfasst. Er ersetzt als Test somit nicht die jeweiligen unter besonderen administrativen Bedingungen und vor dem Hintergrund unterschiedlicher geschichts- wie erziehungswissenschaftlicher Traditionen entwickelten anderen Ansätze und Formate. Aber er ergänzt sie um eine neue Komponente, die mit Blick auf die Herausforderungen der globalisierten, post-traditionalen Gesellschaften an Bedeutung gewinnen dürfte. Die Aufarbeitung der nationalen und internationalen Forschungslage bestätigt auch, dass bisher kein standardisierter Test vorliegt, der die Kompetenzen historischen Denkens in einer wünschenswerten Breite erfasst. An dieser Stelle setzt die Entwicklung des HiTCH-Tests an.

3 Entwicklung eines historischen Kompetenztests für Large-Scale-Assessments

Der HiTCH-Test wurde auf der Basis des FUER-Modells mit dem Ziel entwickelt, bei Schülerinnen und Schülern in der 9. Jahrgangsstufe aller Schularten historische Kompetenzen objektiv, reliabel und valide mithilfe standardisierter Testaufgaben zu erfassen. Der Entwicklung des HiTCH-Tests liegt die Entscheidung zugrunde, bezogen auf die Graduierungslogik des FUER-Modells (vgl. Kapitel 2.3.4) ein intermediäres, konventionelles Kompetenzniveau zu adressieren, das junge Menschen befähigt, auch ohne weiteren Geschichtsunterricht historisch zu denken, also eigenständig Fragen an die Vergangenheit nachzugehen, vorgelegte Narrationen zu überprüfen, Orientierungsangebote kritisch zu beurteilen, auch in Bezug auf die Bedeutung für die eigene Identitätsbildung. Basale/a-konventionelle Kompetenzausprägungen werden daran sichtbar, dass die Aufgaben des HiTCH-Tests nicht oder nur ansatzweise gelöst werden können. Der HiTCH-Test macht es aber nicht möglich, auch transkonventionelle Niveaus zu erfassen.[15]

In Kapitel 3.1 werden zu ausgewählten Herausforderungen, die mit der Operationalisierung historischer Kompetenzen für Testungen in Large-Scale-Assessments verbunden sind, grundlegende Lösungsstrategien skizziert, Kapitel 3.2 stellt übliche Schritte der Testentwicklung und deren Realisierung bei der Erarbeitung des HiTCH-Test vor. Kapitel 3.3 zeigt Beispiele aus dem Itempool und Kapitel 3.4 skizziert den konkreten Prozess der Aufgabenentwicklung für den nunmehr vorliegenden HiTCH-Test.

3.1 Herausforderungen, die mit der Operationalisierung historischer Kompetenzen für Testungen in Large-Scale-Assessments verbunden sind

3.1.1 Der Konstruktcharakter von Geschichte und die Konsequenzen für die Testentwicklung

Dem Konstruktcharakter von Geschichte als einem der aus dem narrativistischen Geschichtsverständnis folgenden epistemologischen Prinzip (vgl. Kapitel 2.1) und einem kompetenzorientierten Konzept historischer Bildung ist geschuldet, dass den Fragen, die historisches Denken und historische Orientierung operationalisieren, nicht umstandslos *eine* richtige Antwort zugeordnet werden kann. Vergangenheitsbezogene Deutungen und orientierende Sinnbildungen hängen immer auch von den historisch Denkenden ab.

15 Zu überprüfen, ob und wie das transkonventionelle Niveau, insbesondere mit geschlossenen Aufgaben, gemessen werden kann, wäre ein eigenes Projekt.

Quantitative Testverfahren, die auf einem Richtigkeitsstandard (Eid, Gollwitzer & Schmitt, 2010) beruhen, scheinen dieser konstruktivistischen Verfasstheit von Geschichte zu widersprechen. Weil aber zugleich gilt, dass die erzählte(n) Geschichte(n) nicht beliebig sind, sondern Triftigkeitskriterien gelten (vgl. Kapitel 2.2.3.6), schließen Konstruktcharakter und Testung sich nicht prinzipiell aus. Es muss in einem Test darum gehen, nicht die historische Orientierung Einzelner und historische Identitätskonstruktion zu erfassen, sondern die dafür notwendigen Fähigkeiten historischen Denkens.

3.1.2 Prozedurale und kategorisierende Kompetenzen historischen Denkens und die Testentwicklung

Über welche Ausprägungen historischen Denkens Schülerinnen und Schüler verfügen, wird *in actu* am deutlichsten, wenn sie die beiden Basisoperationen historischen Denkens vollziehen (vgl. Kapitel 2.3.1.1), sich also re-konstruierend mit Vergangenem befassen oder de-konstruierend mit vorhandenen historischen Narrationen umgehen. Es ist bei der Konstruktion des Tests zu berücksichtigen, dass das Verfügen über prozedurale Kompetenzen wie über Kategorien und Konzepte erfasst wird, die beim re- und de-konstruierenden, historische Orientierungen ermöglichenden historischen Denken zum Tragen kommen. Die Herausforderung besteht dabei darin, Aufgaben zu entwickeln, die zu aktivem historischem Denken auffordern. Ihre Bearbeitung muss es ermöglichen, das Verfügen über ausgewählte und relevante Aspekte der historischen Kompetenzbereiche zu messen. Zugleich gilt: Auch wenn einzelne Aufgaben nur einzelne Aspekte des historischen Denkens adressieren, der Test insgesamt muss den Gesamtprozess im Blick behalten.

3.1.3 Kontext- und Themengebundenheit der Aufgaben

Eine Kernaussage der Lehrpläne der deutschsprachigen Länder ist, dass Geschichtsunterricht über alle Themen hinweg die Entwicklung eines reflektierten und (selbst-) reflexiven Geschichtsbewusstseins unterstützen soll. Das heißt, dass Schülerinnen und Schüler ihre Kompetenzen sowohl an Themen zeigen können müssten, die sie im Unterricht behandelt haben, als auch an Themen, die ihnen fremd sind.

Es scheint naheliegend, in Kompetenztests einerseits durch den Unterricht bekannte und andererseits unbekannte Themen aufzugreifen, um Kompetenzausprägungen zu erfassen. Der zweite Blick macht aber klar, dass andere Wege gesucht werden müssen: Lehrpläne legen zwar Inhalte fest, an denen die Kompetenzentwicklung gefördert werden soll; innerhalb des deutschsprachigen Bildungswesens gibt es hierbei auch deutliche Schnittmengen. Dennoch bestehen gravierende Unterschiede bei den Konkretisierungen der Inhalte auf der Ebene der Schulen und Schulklassen. In schulinternen Curricula und bezogen auf die einzelnen Unterrichtsstunden treffen

die jeweiligen Lehrkräften unterschiedliche Auswahlentscheidungen und Profilierungen. Dazu kommt, dass die in den Lehrplänen gewählten Themen in der öffentlichen Geschichtskultur der beteiligten Länder unterschiedlich präsent sind.

Die Konsequenz, die für den HiTCH-Test daraus gezogen wurde, ist, dass im Zentrum theoretisch abgeleitete prozedurale und kategorisierende Fähigkeiten stehen (vgl. Kapitel 2.3.2: Kompetenzen nach dem FUER-Modell), die an unterschiedlichen Themen erworben und ausdifferenziert werden können. Die gemessene Leistung muss aber, auch in Hinblick auf einen länder- und schulartübergreifenden Einsatz des Tests, unabhängig davon sein, ob bzw. wie bestimmte Inhalte zuvor im Unterricht behandelt worden sind. Die Testaufgaben werden deshalb zwar in historische Kontexte eingebettet. Die zur Bearbeitung notwendigen Informationen werden aber, z.B. durch die beigefügten Materialien oder die Itemformulierung, geliefert.

Zudem wird darauf geachtet, die historischen Kompetenzen an unterschiedlichen Themen zu erfassen. Damit soll erreicht werden, dass in summa nicht die Auseinandersetzung mit einem historischen Gegenstand, sondern das Verfügen über von Themen losgelöste Fähigkeiten getestet wird. Zudem wird auf diese Weise erreicht, dass bei den für die Operationalisierungen gewählten Themen für Schülerinnen und Schüler unterschiedlicher (Bundes-)Länder jeweils sowohl curriculumsnahe wie -ferne Themen berücksichtigt sind.

Darüber hinaus werden auch themenungebundene Testaufgaben entwickelt. Dies ist z.B. bezogen auf historische Sachkompetenzen möglich, wenn das Verständnis zu den epistemologischen Prinzipien oder zu zentralen Konzepten, Kategorien und Strukturierungsschemata erfasst werden soll, die auf unterschiedliche historische Situationen und Fälle angewandt werden können (vgl. Kapitel 2.2.3.5).

Aus der Entscheidung, die notwendigen thematischen Erläuterungen in die Aufgaben zu integrieren, ergibt sich aber eine neue Herausforderung: Der Lese- und Zeitaufwand darf nicht zu hoch werden. Zugleich ist darauf zu achten, dass die Schülerinnen und Schüler auf Basis der fachspezifischen Materialauswertung ihre Fähigkeit zu historischem Denken nachweisen können.

3.1.4 Umgebende Geschichtskultur als Hintergrundvariable historischen Denkens

Die umgebenden Geschichtskulturen haben Einfluss auf die Einstellungen und das Vorwissen junger Menschen. Während es sehr wohl Ziel des HiTCH-Tests ist zu eruieren, inwiefern Schülerinnen und Schüler sich des Einflusses von Geschichtskulturen bewusst sind, ist es nicht die Aufgabe, derartige Einflüsse zu erfassen oder zu klassifizieren. Dazu kommt, dass in unterschiedlichen (Bundes-)Ländern unterschiedliche Themen geschichtskulturell eher präsent bzw. eher vernachlässigt sind. Als Konsequenz für die Testentwicklung ergibt sich, solche historischen Themen und Fragestel-

lungen als Beispiele heranzuziehen, bei denen nicht geschichtskulturell dominierende Deutungen die Auseinandersetzung mit den konkreten Aufgaben erschweren.[16]

3.2 Idealtypische Schritte bei der Testentwicklung

Die Entwicklung des HiTCH-Tests orientierte sich an den üblichen Standards in der Schulleistungsforschung (u.a. Hartig, Klieme & Leutner, 2008; Moosbrugger & Kelava, 2012; Pellegrino, Chudowsky & Glaser, 2001). Da nicht alle Leserinnen und Leser mit diesen Methoden vertraut sind, wird zunächst jeweils das idealtypische Vorgehen bei der Entwicklung eines Kompetenztests skizziert und anschließend vorgestellt, wie bei der HiTCH-Entwicklung vorgegangen wurde.

Generell werden in der Literatur zur Testentwicklung *Speed-* oder *Geschwindigkeitstests* von sogenannten *Power-* oder *Niveautests* unterschieden. Bei ersteren geht es darum, (eher) einfache Aufgaben zu lösen; die Leistungsdifferenzierung erfolgt über die Begrenzung der Bearbeitungszeit. Letztere enthalten auch schwierigere Aufgaben, die selbst bei unbegrenzter Zeit nicht von allen Schülerinnen und Schülern richtig gelöst werden können. Die Differenzierung der Leistungen erfolgt über das Schwierigkeitsniveau der Aufgaben. Vorzugsweise werden solche Tests zur Feststellung komplexer kognitiver Fähigkeiten genutzt (Jonkisz, Moosbrugger & Brandt, 2012). Der HITCH-Test zählt zur zweiten Testart.

3.2.1 Auswahl eines theoretischen Modells

Kompetenzen sind nicht direkt messbar. Sie können nur aus messbaren Sachverhalten (den „Indikatoren") geschlossen werden. Um zu diesen Indikatoren zu kommen, müssen nach Klieme et al. (2003) zunächst – theoriebasiert – die Kernkonstrukte geklärt werden, die untersucht werden sollen, denn ein Leistungsvergleich orientiert sich nicht „an einer willkürlichen Auswahl von Aufgabenstellungen […], sondern an Kompetenzen und Kompetenzmodellen" (Klieme et al., 2003, S. 87).

Die Klärung der Kernkonstrukte ist in Kapitel 2 ausführlich dargestellt worden. Ausgehend vom heute in der westlichen Theoriediskussion vorherrschenden narrativistischen Geschichtsverständnis erfolgte die Entscheidung, sich dabei auf das FUER-Modell zu beziehen (vgl. Kapitel 2.4.3). Die dort ausgewiesenen Kompetenzbereiche und Kernkompetenzen wurden dargestellt (vgl. Kapitel 2.3), auch im Vergleich mit anderen nationalen und internationalen Kompetenzmodellen (vgl. Kapitel 2.4).

16 Vergleichend sei auf das Ergebnis der *narratio*-Studie verwiesen, nach der die Themenwahl deutliche Auswirkungen auf das historische Denken nach sich zieht (Waldis et al., 2015; vgl. Kapitel 2.5.2.2).

3.2.2 Operationalisierung der theoretischen Konstrukte: empirisches und numerisches Relativ

Für quantitative Erhebungen müssen theoretische Konstrukte „messbar" gemacht werden. Durch das Messen wird der Ausprägungsgrad bestimmter Merkmale („empirisches Relativ") durch die Angabe von Zahlen („numerisches Relativ") repräsentiert. Damit wird möglich, dass bestimmte mathematische Operationen durchgeführt werden können. Die numerisch gefassten Ergebnisse lassen Aussagen über die Verhältnisse im empirischen Merkmalsbereich zu (Eid et al., 2010). Die theoretisch definierten Konstrukte werden also in direkt beobachtbare Indikatoren (= Bearbeitung von Testaufgaben) übersetzt, für deren Lösung die im Kompetenzmodell beschriebenen Fähigkeiten notwendig sind.

Im HiTCH-Test werden die Indikatoren unter Bezug insbesondere auf die Kernkompetenzen der Kompetenzbereiche (wie Begriffskompetenz im Kompetenzbereich historische Sachkompetenzen oder De-Konstruktionskompetenz im Kompetenzbereich historische Methodenkompetenz) gewonnen (vgl. Kapitel 2.3.1.3). Die Testaufgaben fokussieren einzelne Aspekte dieser Kernkompetenzen (epistemologische Prinzipen; Sinnbildungen in Narrativen); die Antworten der Schülerinnen und Schüler sind die direkt beobachtbaren Indikatoren für die Kompetenzausprägungen, über die sie verfügen können, weil in ihnen die im Kompetenzmodell beschriebenen Fähigkeiten zum Ausdruck kommen (Körber & Meyer-Hamme, 2015; Meyer-Hamme, 2015).

3.2.2.1 Vermeidung von *construct underrepresentation*

Für die Entwicklung des HiTCH-Tests, der historische Kompetenzen in der ganzen Bandbreite reliabel und valide erfassen soll, wurden zur Vermeidung von *construct underrepresentation* (Messick, 1995) alle vier Kompetenzdimensionen mit ihren sie operationalisierenden Unterfacetten (Kernkompetenzen) adressiert (zur formalen Unterscheidung von Kompetenzbereichen, Kernkompetenzen und Einzelkompetenzen vgl. Kapitel 2.3.1.3). Zu diesem Zweck wurden zunächst eine Vielzahl von Aufgabenblöcken bzw. Items generiert, die für sich genommen einzelne Facetten des Konstrukts erfassen und in der Summe alle Facetten des historischen Denkens repräsentieren sollten.

Zu bedenken ist, dass bei der konkreten Bearbeitung von Aufgaben vielfach Kompetenzen aktiviert werden, die ihrerseits Aspekte mehrerer der idealtypisch unterscheidbaren Kompetenzbereiche/Kernkompetenzen umfassen (zu den theoretisch begründeten Zusammenhängen zwischen den Kompetenzbereichen vgl. die Ausführungen zu den Überlappungsbereichen, Kapitel 2.3.3 und zu Kern- und Einzelkompetenzen, Kapitel 2.3.1.3). Testitems erfassen dann auch mehrere Kompetenzdimensionen in unterschiedlichen Anteilen (vgl. ausführlich Blum, Drüke-Noe, Hartung & Köller, 2006). Auch wenn bei der Aufgabenkonstruktion für den HiTCH-Test versucht wurde, Aufgabenblöcke zu generieren, die sich möglichst eindeutig auf eine

Kernkompetenz beziehen, haben in konkreten Denkprozessen möglicherweise Fähigkeiten, die idealtypisch anderen Kompetenzbereichen zuzuordnen sind, einen Einfluss auf die Bearbeitung der Aufgaben. Dies lässt sich an den vier Kompetenzbereichen verdeutlichen:

- Der Bereich der historischen Fragekompetenzen umfasst im FUER-Modell den Teil des historischen Denkens, in dem auf Grund von Orientierungsbedürfnissen historische Fragestellungen entwickelt werden. Die verfolgte Fragestellung hat Konsequenzen für die Materialauswahl, die ihrerseits ein Teilaspekt der Methodenkompetenz (Re- und De-Konstruktion) ist. Im Prozess historischen Denkens besteht also ein logischer Zusammenhang zwischen Orientierung, Fragestellungen und Methoden. Bei der Entwicklung von Aufgaben wurde jeweils versucht, auf relevante Aspekte des Kompetenzbereichs historische Fragekompetenz zu fokussieren, u.a. indem Materialvorgaben oder Itemformulierungen Denkprozesse aus benachbarten Kompetenzbereichen abnahmen und die Schülerinnen und Schüler sich auf den Umgang mit historischen Fragen konzentrieren konnten (vgl. auch die Itembeispiele zu Fragekompetenz in Kapitel 3.2.3). Es wurden u.a. Aufgaben entwickelt, bei denen sie unterschiedliche Ausrichtungen von historischen Fragen identifizieren müssen, bzw. Aufgaben, bei denen sie plausible Zusammenhänge zwischen in der Aufgabe vorgegebenen Orientierungsbedürfnissen, historischen Fragen und vorgegebenen Materialien herstellen müssen.
- Der Bereich der historischen Methodenkompetenzen (Re- und De-Konstruktion) schließt den Umgang mit und die Auswertung von Materialien – Quellen und Darstellungen – ein und umfasst, darauf bezogen, auch die Erstellung einer historischen Narration. Unterschieden werden die auf Re-Konstruktion gerichteten synthetisierenden und auf De-Konstruktion gerichteten analytischen Fähigkeiten (Methodenkompetenz: Re- und De-Konstruktion). Derartige methodische Kompetenzen sind nicht unabhängig z.B. von epistemologischen Einsichten in historisches Denken und von der inhaltsbezogenen Kategorisierungsfähigkeit, mit anderen Worten von den historischen Sachkompetenzen. Aufgabenstellungen zur Methodenkompetenz (Re- und De-Konstruktion) umfassen im bisher vorliegenden HiTCH-Test z.B. Materialauswertungen (Vergleich multiperspektivischer Quellen bzw. kontroverser Darstellungen); die Schülerinnen und Schüler werden dabei von der Sachkompetenz voraussetzenden Wahl des Fokus für den Vergleich entlastet. Andere Aufgaben zielen auf die Fähigkeit zur Synthetisierung von Informationen in eine plausible historische Ordnung. Hier werden mögliche Bausteine für historische Narrationen vorgegeben, so dass die Aufgabe in der narrativ triftigen (vgl. Kapitel 2.2.3.6) Auswahl und Anordnung besteht.
- Historische Orientierungskompetenzen umfassen den Bereich historischen Denkens, in dem aus vergangenen Be- und Gegebenheiten Sinn für die Gegenwart und Zukunft abgeleitet wird. Im Fokus einer Kompetenzmessung steht die Fähigkeit, solche Sinnbildungen zu erkennen und herstellen zu können, nicht aber die Frage, welche Sinnbildungen und Identitätskonstruktionen die Lernenden selbst vertreten. Überlappungen mit Sach- und Methodenkompetenzen liegen in Bezug auf die

Fähigkeit zur kategorisierenden Auswahl, bzw. in Bezug auf prozedurale methodische Fähigkeiten vor. Im vorliegenden HiTCH-Test gibt es Aufgaben, in denen die Schülerinnen und Schüler solche Sinnbildungen selbstständig herstellen müssen. Es gibt aber auch Aufgaben, in denen sie Muster hinter vorgegebenen Sinnbildungen erkennen müssen, so dass erfassbar wird, inwiefern sie über die dafür notwendigen Konzepte verfügen. In beiden Fällen entlasten die Aufgabenstellungen von inhaltlichen Entscheidungen, damit nicht Sach- sondern tatsächlich Orientierungskompetenz gemessen wird.

- Der Kompetenzbereich der historischen Sachkompetenzen umfasst das Verfügen über Konzepte und Kategorien historischen Denkens. Bei Aufgaben, die diesen Kompetenzbereich adressieren, werden Aussagen zu ausgewählten Konzepten (z.B.: „Herrschaft“) oder Kategorien (z.B. aus den epistemologischen Einsichten) den Schülerinnen und Schülern vorgelegt, mit dem Ziel zu überprüfen, inwiefern sie über Fähigkeiten zu kategorisieren verfügen. Dazu müssen sie etwa Begriffe typischen Merkmalen zuordnen oder die Plausibilität von Aussagen einschätzen.

Die Auflistung ist nicht abschließend. Deutlich werden soll daran, dass zwar eine klare Fokussierung der Kompetenzbereiche, nicht aber eine vollständige Trennung der Aufgaben in unterschiedliche Dimensionen möglich ist. Entscheidend ist für den HiTCH-Test, dass alle Kompetenzbereiche im Test berücksichtigt sind.

3.2.2.2 Berücksichtigung der „breiten Zielgruppe“ bei der Operationalisierung

Die Entscheidung, die Aufgabenblöcke für die einzelnen Kompetenzbereiche des FUER-Modells an verschiedenen Themen und Situationen zu konkretisieren, zielt nicht nur, wie eben dargestellt, darauf, die Breite des Konstrukts historischen Denkens angemessen abzubilden und in testfähige Einzelaspekte zu gliedern. Sie berücksichtigt zugleich, dass die Zielgruppe des HiTCH-Tests sehr heterogen ist: Schülerinnen und Schüler des 9. Schuljahres aller Schularten (außer Förderschulen) in Deutschland wie auch in Österreich und der deutschsprachigen Schweiz wurden angesprochen.

Aus der Breite der Zielgruppe wurde eine weitere Konsequenz gezogen: Bezogen auf die einzelnen Kompetenzbereiche mussten die Operationalisierungen leichte wie schwere Items enthalten. „Je breiter die Zielgruppe, desto mehr müssen die Aufgaben über einen breiteren Schwierigkeitsbereich streuen und ggf. auch inhaltlich breiter gefächert sein, um möglichst viele Merkmalsausprägungen abdecken zu können“ (Jonkisz et al., 2012, S. 33). Bei der Entwicklung der Aufgaben und Aufgabenformate wurde der Schwierigkeitsgrad berücksichtigt. In die Pilotstudien und der Haupterhebung wurde zudem die Schwierigkeit der Aufgaben empirisch geprüft. Dabei wurde auch untersucht, inwiefern es in Subgruppen auffällige Differenzen bei der Lösung von Aufgaben (DIF-Effekte) gibt, die auf weitere Einflussfaktoren zurückschließen lassen (vgl. Kapitel 4).

3.2.3 Auswahl der Aufgabenformate

Zu unterscheiden sind zunächst Aufgaben mit einem offenen und Aufgaben in einem geschlossenen Antwortformat. In Aufgaben mit offenem Antwortformat formuliert die Testperson die Antworten selbst, z.B. in Kurzessays oder Ergänzungsaufgaben; in geschlossenen Aufgaben wählen die Schülerinnen und Schüler aus vorgegebenen Antwortmöglichkeiten aus. In der HiTCH-Testentwicklung spielten offene bzw. halboffene Aufgabenformate eine nur untergeordnete Rolle. Die für die erste Pilotierung entwickelten Aufgaben mit offenem Format fanden wegen der strengen Zeitvorgaben und des begrenzten Budgets in der weiteren Testentwicklung keine Berücksichtigung. Die Erweiterung auf halboffene und offene Formate soll in weiteren Projektphasen erprobt werden.

Hinsichtlich der geschlossenen Aufgaben wurden bei der Entwicklung des HiTCH-Tests in einem Brainstorming, abgestimmt auf die adressierten Kompetenzbereiche, eine Vielzahl von Antwortformaten diskutiert und erarbeitet. Für die Festlegung gaben die Forschungsübersicht über standardisierte Testungen und empirische Studien zu historischem Denken (vgl. Kapitel 2.5) wichtige Impulse. Als standardisierte Aufgabenformate wurden schließlich Ordnungsaufgaben (Zuordnung oder Umordnung) und Auswahlaufgaben (z.B. dichotome Aufgaben, in denen zwischen richtig oder falsch gewählt werden soll, Multiple-Choice- oder Single-Choice-Aufgaben) ausgewählt.

Generell wurde bei der Wahl des Aufgabenformats die Einfachheit oder Komplexität der Testanweisung, der Zeitaufwand und die Ratewahrscheinlichkeit bei der Lösung der Aufgabe, wie auch der Auswertungsaufwand gegeneinander abgewogen (Jonkisz et al., 2012).

3.2.4 Weitere Maßnahmen zur Sicherung der Validität der Messung

Faktoren wie Lesekompetenz, Kenntnisse von, Vorstellungen über, Einstellungen zu oder Interesse an bestimmten Themen können im Sinne von *construct-irrelevant variance* (Messick, 1995) bei den Schülerinnen und Schülern einen Einfluss auf die Beantwortung der Testaufgaben haben. Dies lässt sich nicht komplett vermeiden (gerade in solchen Fällen, in denen eine natürliche Überlappung unterschiedlicher Konstrukte vorhanden ist), aber grundsätzlich ist bei der Konstruktion eines Tests darauf zu achten, die konstrukt-irrelevanten Einflüsse gering zu halten. Bei der Entwicklung des HiTCH-Tests wurde deshalb u.a. darauf geachtet, eine Variation des Themenbezugs zu garantieren sowie allzu lange (und sprachlich komplexe) Texte zu vermeiden. Zudem wurden zusätzlich die Lesefähigkeiten und die allgemeinen kognitiven Fähigkeiten der Schülerinnen und Schüler erhoben, um die Frage empirisch beantworten zu können, inwiefern mit dem Test etwas spezifisch Historisches und nicht eine Mischung aus Lesekompetenzen und kognitiven Fähigkeiten bei der Lösung von Testaufgaben erhoben wird (vgl. die Hinweise in Kapitel 3.4.5).

3.3 Beispiele aus dem HiTCH-Itempool

Im Rahmen der Testentwicklung wurden mehr als 250 Aufgabenblöcke und knapp 1500 Items entwickelt und in drei groß angelegten Erhebungen (vgl. die Übersicht in Kapitel 3.4) eingesetzt. Unter dem Begriff *Aufgabenblock* verstehen wir einen Teilbereich des Testinstruments, der zu einem Themenbereich, meist auf der Grundlage von vorgegebenen Materialien einen Kompetenzbereich adressiert und mehrere Items umfasst. Ein *Item* wiederum stellt die kleinste Aufgabeneinheit dar, also z.B. eine einzelne Ankreuzaufgabe, die von den Schülerinnen und Schülern korrekt oder falsch gelöst werden kann.[17] Im Folgenden werden die Prototypen der eingesetzten Testaufgaben vorgestellt. Zur Wahrung der Testsicherheit werden keine Original-Items wiedergegeben. Die hier vorgestellten Aufgaben wurden „nachgebaut". Nicht alle Aufgabentypen, die im HiTCH-Test vorhanden sind, können hier vorgestellt werden.[18]

3.3.1 Beispielaufgaben zur historischen Sachkompetenz

Die Sachkompetenzen adressieren das Verfügen-Können über Begriffs- und Strukturierungskonzepte mit einem Inhalts-, Verfahrens-, Theorie-, Orientierungsbezug (vgl. Kapitel 2.3.2.4). Bei den Aufgabenblöcken und Items zur Sachkompetenz wurden sowohl inhaltsbezogene Aufgaben (etwa zum Konzept Staatsformen), als auch theoriebezogene Aufgaben (etwa zu den epistemologischen Prinzipien) sowie methodenbezogene Aufgaben (etwa zur Verfügung über einen Quellen- und Darstellungsbegriff) formuliert.

Aufgabenbeispiel 1: Umgang mit historischen Begriffskonzepten (Sachkompetenz)
Die Aufgabe setzt sich aus zwei Teilen zusammen: Im ersten Teil wird getestet, inwieweit die Schülerinnen und Schüler in der Lage sind, abstrakte Begriffskonzepte konkreten Beschreibungen zuzuordnen. Getestet wird in dem Beispiel das Verfügen über grundlegende Konzepte zur Kategorie „Wirtschaft", die in der deutschsprachigen Wirtschaftsgeschichte, aber auch in deren Präsenz im Geschichtsunterricht und der außerschulischen Geschichtskultur von Relevanz sind. Im zweiten Teil der Aufgabe wird getestet, inwieweit diese Konzepte typischen Beispielen, wie sie auch im Geschichtsunterricht vorkommen, zugeordnet werden können. Als Aufgabenformat wird die Zuordnung gewählt. In solche Aufgaben gehen auch erlernte (Detail-) Kenntnisse mit ein. Es kann also nicht sicher gesagt werden, ob die Schülerinnen und Schüler über konzeptuelle Kompetenzen verfügen oder ob sie die Beispiele (nur)

17 Im Zuge der Auswertungsprozeduren kann als „Item" auch die Zusammenfassung mehrerer Antworten bezeichnet werden.
18 Die Original-Items können nach Absprache mit dem HiTCH-Konsortium eingesehen und unter Einhaltung von Testsicherheitsbestimmungen in Studien eingesetzt werden. Zum Prozedere vgl. die Website www.hitch-projekt.de.

gelernt haben und sie reproduzierend wiedergeben, ohne auch über transferable und kategorisierende Fähigkeiten zu verfügen.

Aufgabenbeispiel 1

In der Geschichte hat es verschiedene Organisationsformen für „Wirtschaft" und „Produktion" gegeben, die sich im Laufe der Zeit verändert haben. Dennoch gibt es ein paar typische Formen, die zu bestimmten Zeiten besonders bedeutsam waren.

Begriffe der Wirtschaftsgeschichte	
1) Produktion durch sich selbst optimierende Computersysteme	4) Standardisierte Massenproduktion
2) Großgrundbesitz/Latifundien	5) Lehnswesen/Lehnswirtschaft
3) Subsistenzwirtschaft/Bedarfswirtschaft	6) (Erste) Industrielle Revolution

Ordne bitte die Nummern der Begriffe der Wirtschaftsgeschichte den typischen Merkmalen zu.

Typische Merkmale	**Nummer**
Herstellung von Gütern durch Roboter und andere Maschinen	(1)
Selbstversorgung kleinerer Gruppen. Sicherung des Lebensunterhaltes	(3)
Entwicklung von Fabriken. Betrieb von Maschinen mit Dampfkraft	(6)
Anbau z.B. von Getreide, Oliven, Trauben, Gemüse, u.a. durch den Einsatz von Sklaven	(2)
Verleihung von Land und Leuten vom König an Gefolgsleute; Abgaben, Waffendienst und Treue als Gegenleistung	(5)
Arbeitsteilung bei Herstellung von Konsumgütern, häufig Fließbandarbeit	(4)

Ordne nun bitte die Nummern der Begriffe der Wirtschaftsgeschichte einem typischen historischen Beispiel zu.

Historisches Beispiel	**Nummer**
England um 1850	(6)
USA um 2015	(1)
Steinzeit	(3)
Römisches Reich, um 100 v. Chr.	(2)
Deutschland um 1960	(4)
Frankreich im Mittelalter	(5)

Aufgabenbeispiel 2: Umgang mit epistemologischen Prinzipien von Geschichte (Sachkompetenz)

Das zweite Aufgabenbeispiel fokussiert theoriebezogene Begriffskompetenzen. Es erhebt, inwiefern die Schülerinnen und Schüler Konsequenzen aus den hier in Umschreibungen definierten epistemologischen Prinzipien für die Konstruktion triftiger Narrationen ziehen können. Im Multiple-Choice-Format ist auszuwählen, ob eine Folgerung nicht („darf man nicht"), möglich („kann man") oder zwingend („muss man") aus einem epistemologischen Prinzip abzuleiten ist. Damit werden die Prinzipien Perspektivität, Partikularität und Narrativität (vgl. Kapitel 2.1) umschrieben.

Mit der Frage danach, welche Reaktionen auf epistemologische Prinzipien – hier der Partialität historischer Überlieferung – plausibel sind, werden diese in den Prozess historischen Denkens gestellt und müssen reflektiert werden. Mit der Verfügung über epistemologische Annahmen steht ein Aspekt historischer Sachkompetenz im Vordergrund dieser Aufgabe und nicht die Messung der epistemologischen Überzeugungen der Lernenden selbst. Operationalisiert wird die Aufgabe im Überlappungsbereich zwischen historischer Sach- und Methodenkompetenz. In der Aufgabenstellung wird das Prinzip der Partialität umschrieben, es muss also nicht gelernt worden sein. Der Methodenkompetenz (Re- und De-Konstruktion) zuzuordnende Handlungsweisen und Einsichten sind ausformuliert.

Aufgabenbeispiel 2

Kreuze an, wie man mit Vergangenheit entweder nicht umgehen darf, umgehen kann oder umgehen muss wenn man eine überzeugende Geschichte über die Vergangenheit verfasst.

Wegen …	… darf man nicht…	… kann man…	… muss man …	… folgendermaßen reagieren.
Weil nicht alles überliefert ist, was früher geschah…	☐	x	☐	… versuchen, sich selbst ein Bild von der Vergangenheit zu machen, indem man viele Informationsquellen benutzt (Filme ansieht, Bücher liest oder mit Experten spricht).
	☐	☐	x	… selbst erschließen, was nicht wortwörtlich in Quellen berichtet wird, aber dennoch zu den überlieferten Quellen passt.
	x	☐	☐	… annehmen, dass man mit weiterer Quellensuche alles über die Vergangenheit erfahren könnte.
	x	☐	☐	… davon ausgehen, dass Experten für Geschichte alles behaupten können, ohne dass es überprüft werden kann.

3.3.2 Beispiel-Aufgaben zur historischen Fragekompetenz

Unter der Fragekompetenz wird im Sinne des FUER-Modells (vgl. Kapitel 2.3.2.1) die Fähigkeit verstanden, in Folge zeitlicher Orientierungsbedürfnisse historische Fragen selbst zu stellen (zuzuordnen zur Basisoperation der Re-Konstruktion) oder die Fragestellungen anderer zu verstehen (zuzuordnen zur Basisoperation der De-Konstruktion).

Aufgabenbeispiel 3: Fragen verstehen (Fragekompetenz)
Im ersten Aufgabenbeispiel zur Fragekompetenz steht im Vordergrund, ob die Schülerinnen und Schüler unterschiedliche Typen von Fragen an die Vergangenheit unterscheiden können. Im Sinne der oben dargelegten narrativistischen Geschichtstheorie sind historische Fragen im vollen Sinne dadurch gekennzeichnet, dass sie einen narrativen Zusammenhang zwischen Prozessen in der Vergangenheit und ihrer Bedeutung für die Zukunft herstellen. Davon abzugrenzen sind solche Fragen, die nur auf die Veränderungen in der Vergangenheit zielen und solche, die nur nach Einzelheiten in der Vergangenheit fragen. Im Hintergrund dieser Aufgabenstellung ist die Unterscheidung der drei Fokussierungen, wie sie in der Sechs-Felder-Matrix systematisiert ist (Hasberg & Körber, 2003). Die Aufgabe zielt darauf ab, ob die Schülerinnen und Schüler unterscheiden können zwischen Fragen, die auf Daten und „Fakten" zur „Vergangenheit", auf die historischen Zusammenhänge in der erzählten „Geschichte" bzw. auf die Bedeutung für „Gegenwart und Zukunft" fokussieren (vgl. hierzu Kapitel 2.3.1.1, Fokussierungen). Dabei werden die drei Fokussierungen umschrieben und konkretisiert; es wird also nicht davon ausgegangen, dass sie im Unterricht erarbeitet worden sind. Es kommt nicht darauf an, ob die Jugendlichen die Fragen beantworten können, sondern ob sie in der Lage sind, die unterschiedlichen Reichweiten und Zielsetzungen der Fragen zu erkennen.

Aufgabenbeispiel 3

Wenn man sich für die Hexenverfolgungen früher interessiert, kann man ganz unterschiedliche Fragen stellen. Die unten stehenden Fragen zum Thema „Hexen" sollt ihr nicht selbst beantworten. Ihr sollt lediglich einordnen, um was für einen Typ von Frage es sich jeweils handelt:

Handelt es sich um eine Frage, die nach **Einzelheiten** in der Vergangenheit fragt (z.B. Namen, Ereignisse)? Oder ist es eine Frage, die nach **Zusammenhängen** in der Vergangenheit fragt (z.B. Entwicklungen)? Oder handelt es sich um eine Frage, in der direkt nach der **Bedeutung** der Vergangenheit **für heute** gefragt wird?

Bitte setze nur ein Kreuz pro Zeile – dort, wo der Schwerpunkt der Frage liegt!

	Frage zielt vor allem ab auf …		
	Einzelheiten in der Vergangenheit	Zusammenhänge in der Vergangenheit	Bedeutung für heute
In welchem Jahr wurde die letzte Hexe in Deutschland verbrannt?	x	☐	☐
Wie kam es dazu, dass in der Frühen Neuzeit so viele angebliche Hexen verfolgt wurden?	☐	x	☐
Welche Erkenntnisse über den Umgang mit Außenseitern können aus der Hexenverfolgung gezogen werden?	☐	☐	x
Inwiefern haben sich die Argumente, mit denen die „Hexen" verfolgt wurden, vom 14. bis zum 17. Jahrhundert verändert?	☐	x	☐
Sind die Gegner der Hexenverfolgung Helden, die uns ein Vorbild sein können?	☐	☐	x
Wer war der Autor des Buches der „Hexenhammer"?	x	☐	☐

Aufgabenbeispiel 4: Fragestellungen im Prozess historischen Denkens (Fragekompetenz)

Das zweite Aufgabenbeispiel zur Erfassung der Fragekompetenz bezieht sich auf den Gesamtprozess des historischen Denkens. Die Zuordnungsaufgabe adressiert Fragekompetenz als Fähigkeit, die Schritte von einem Orientierungsbedürfnis zu einem möglichen Orientierungsangebot zu steuern.

Das Aufgabenbeispiel ist so konstruiert, dass Ausgangs- und Endpunkt des historischen Denkprozesses vorgegeben sind: Angesetzt wird an einer in der Gegenwart verorteten Verunsicherung; als Ergebnis wird ein historisch begründetes Orientierungsangebot vorgeschlagen. Die Schülerinnen und Schüler sollen die Lücken zwischen den Vorgaben füllen, indem sie a) passende Zugriffe für Fragen an die Vergangenheit auswählen und b) das zur Fragestellung passende historische Quellenmaterial bestimmen. Dafür stehen ihnen insgesamt sechs Textbausteine zur Verfügung. Sie sollen sie so auswählen und in Zusammenhang setzen, dass eine plausible Aussage entsteht. Bei der korrekten Auswahl der Textbausteine ergibt sich a) der Zusammenhang zwischen einer möglichen und vorgegebenen Verunsicherung in der

Aufgabenbeispiel 4

A	Texte, die in der Zeit des aufkommenden Buchdrucks hergestellt wurden, könnten Informationen liefern.
B	Es stellt sich die Frage, welche Rechte und Pflichten Menschen in Gesellschaften hatten, die sehr stark oder sehr wenig religiös geprägt waren.
C	Bildungsmaßnahmen in der Reformationszeit könnten darauf hin untersucht werden.
D	Die Gründe für die Religionskriege der Reformationszeit wären als Beispiel zu untersuchen.
E	Flugblätter eignen sich dazu; es gibt Beispiele, die Gewalt rechtfertigen und andere, die Frieden fordern.
F	Dazu können Gesetzestexte der Reformationszeit mit Gesetzestexten weniger religiöser Zeiten verglichen werden.

Ordne die Buchstaben der Sätze (A-F) so in die untere Tabelle ein, dass ein sinnvoller Weg zu erkennen ist, um ein in der Gegenwart sichtbar werdendes Problem durch Geschichte zu klären. In der ersten Zeile findest du ein Beispiel.

Gegenwärtige Beobachtung	Klärendes Beispiel aus der Vergangenheit	Geeignetes Material für die Erschließung	Mögliches Ergebnis
Beispiel *Es wird oft diskutiert, wie viel Toleranz Religion braucht und verträgt.*	*Für diese Fragestellung wäre der Umgang mit der Spaltung der katholischen Kirche durch die Reformation interessant.*	*Die theologischen Streitschriften katholischer Gelehrter gegen die Reformatoren könnten untersucht werden.*	*Andersdenkende zu unterdrücken ist wirkungslos und zudem auch rückständig. Die Forderung nach Toleranz gegenüber anderen Religionen geht keinesfalls mit einer Schwächung der eigenen Religion einher.*
Manche behaupten, persönliche Freiheit werde nur durch abnehmende Gläubigkeit gestärkt	B	F	Zwar wurde Menschen mit den Bürgerrechten mehr Freiheit eingeräumt. Aber Unterdrückung gab (und gibt) es immer noch, auch ohne religiöse Begründung.
Viele Menschen fragen sich, wie es aus religiösen Gründen zu Gewalt kommen kann.	D	E	Religion muss bei toleranten Menschen kein Zankapfel sein. Aber im Kampf um Macht können religiöse Ziele verbunden mit anderen Interessen sogar zu Kriegen führen.
Oft wird die Meinung vertreten, dass strenge Gläubigkeit und schlechte Bildung immer zusammen hängen.	C	A	Dass mit der Reformation auch viele Bücher verbreitet wurden und die Bibel aus dem Lateinischen in die Volkssprachen übersetzt wurde, ist ein Beispiel für das Zusammenwirken von Bildung und Gläubigkeit.

Gegenwart und einer hierzu passenden Fragestellung an die Vergangenheit und b) einem aufgrund dieser Fragestellung für die Beantwortung ausgewählten Quellenmaterials, das ein (vorformuliertes) historisch fundiertes Orientierungsangebot begründen kann. Die erste auszufüllende Lücke verweist also auf die Fähigkeit, geeignete auf die Vergangenheit bezogene Fragestellungen zu entwickeln; die zweite Lücke ist ein Indikator für die Fähigkeit, aufgrund dieser Fragestellung geeignetes Material auswählen zu können. Die Aufgabe erfasst also, inwieweit die Schülerinnen und Schüler in der Lage sind, plausible Zusammenhänge zwischen einem Orientierungsbedürfnis, einer historischen Frage und der Auswahl eines Materialkorpus herzustellen. Damit adressiert die Aufgabe auch den Überlappunsgbereich zwischen Frage- und Methodenkompetenz (vgl. Kapitel 2.3.3).

3.3.3 Beispiel-Aufgaben zur historischen Methodenkompetenz (Re- und De-Konstruktion)

Die historischen Methodenkompetenzen werden durch die Kernkompetenzen zur Re- und De-Konstruktion operationalisiert (vgl. Kapitel 2.3.2.2). In der Re-Konstruktion geht es darum, auf der Basis eines kompetenten Umgangs mit Quellen und Darstellungen eine eigene Narration bezogen auf eine vergangenheitsbezogene Fragestellung zu erarbeiten. Einige Schritte hiervon – die Quellenkritik, -analyse und -interpretation – stehen seit den 1970er Jahren im Zentrum des (quellenorientierten) Geschichtsunterrichts (Schneider, 2010). De-Konstruieren zu lernen wird dagegen erst seit einigen Jahren als Ziel des Unterrichts wahrgenommen.

Aufgabenbeispiel 5: Historische Aussagen durch Quellen stützen oder widerlegen (Methodenkompetenz: Re-Konstruktion)
Die Zielsetzung des folgenden Aufgabenbeispiels dürfte den Schülerinnen und Schülern aus dem Unterricht vertraut sein. Sie sollen zwei kurze Textquellen daraufhin untersuchen, ob mit diesen Quellen vorgegebene Aussagen (hier zu Kreuzzügen) widerlegt oder belegt werden können bzw. ob auf Grundlage dieser beiden Quellen über Behauptungen entschieden werden kann.

Aufgabenbeispiel 5

📖 Mat. 1: Albert von Aachen berichtet im Jahr 1101 über die Ursachen des Ersten Kreuzzugs (1096–1099)

Ein Priester, Peter aus Frankreich, hat als Erster mit aller Leidenschaft, die er besaß, zu diesem Zug aufgefordert. Dieser Priester hatte zuvor eine Pilgerreise nach Jerusalem unternommen. Damals musste er in der Kirche des Heiligen Grabes, ach, Dinge sehen, so sündhaft und böse, dass sein Herz voller Trauer aufseufzte und er Gott zur Rache aufrief. Da war der Himmel von Finsternis bedeckt und Peter ging an das Heilige Grab, um dort zu beten und dort erschien ihm Jesus und sprach: „Peter, eile so rasch du kannst zurück nach Frankreich und erzähle dort, was mein Volk und die heiligen Stätten hier zu ertragen haben, und entflamme die Herzen der Gläubigen, Jerusalem und die heiligen Orte zu säubern und die Heiligtümer zu befreien." Peter kehrte zurück nach Frankreich und überbrachte dem Papst die Botschaft, die Gott ihm aufgetragen hatte. Daraufhin gelobten nun die Bischöfe, Herzöge, Grafen und viele weitere aus ganz Frankreich, einen Zug nach dem Heiligen Grabe zu unternehmen. Seinem Aufruf folgten Kirchenleute, die Fürsten verschiedener Reiche und endlich die ganze Menge des Volkes. Die ganze Christenheit, ja selbst das weibliche Geschlecht, eilte froh, vom Geist der Buße getrieben, zur Teilnahme an diesem Zug

📖 Mat. 2: Anna Komnene (1083–1154) berichtet über die Ursachen des Ersten Kreuzzugs (1096–1099)

Ein Franzose mit dem Namen Peter hatte sich auf die Pilgerreise zum Heiligen Grab begeben, musste von den Türken und Sarazenen viel Ungemach erleiden und war nur mit Mühe und Not in seine Heimat nach Frankreich zurückgekehrt. Dass er sein Ziel nicht erreicht hatte, nahm er nicht so einfach hin, sondern wollte sich wieder auf denselben Weg machen. Es war ihm aber klar, dass er nicht einfach erneut die Pilgerreise zum Heiligen Grab aufnehmen konnte, wenn ihm nicht noch etwas Schlimmeres zustoßen sollte. Daher entwickelte er einen schlauen Plan. Er wollte in allen Ländern Westeuropas verkünden: „Eine göttliche Stimme befiehlt mir, allen Baronen [Adligen] in Frankreich zu predigen, sie sollten ihre Heimat verlassen, sich auf Pilgerfahrt zum Heiligen Grab begeben und sich mit ganzer Kraft und ganzem Herzen bemühen, Jerusalem aus der Hand der Muslime zu befreien." Und das setzte er auch in die Tat um. Er senkte gleichsam eine göttliche Stimme in aller Herzen und brachte die Menschen in sämtlichen Ländern dazu, sich mit Waffen und Pferden zu versammeln. So waren sie voller Bereitschaft und Begeisterung, und alle Straßen waren voll von ihnen.

Welche der folgenden Aussagen über den Ersten Kreuzzug kann man mit diesen (nur diesen!) Textquellen (Mat. 1, Mat. 2) unterstützen (belegen), welche entkräften (widerlegen)?

Setze bitte nur ein Kreuz pro Zeile!

	Mit diesen Quellen (Mat. 1 und Mat. 2)…		
	… zu entkräften (widerlegen)	… nicht zu entscheiden	… zu unterstützen (belegen)
Der Papst musste gedrängt werden, die Kreuzzugsaufrufe zu unterstützen.	☐	x	☐
Am ersten Kreuzzug nahmen nur Ritter teil.	x	☐	☐
Peter behauptete, den Auftrag zur Kreuzzugspredigt von einer höheren Macht erhalten zu haben.	☐	☐	x
Peter handelte aus egoistischen Gründen, als er zum Kreuzzug aufrief.	☐	x	☐
…	…	…	…

Aufgabenbeispiel 6: Zeitliche Reihenfolgen herstellen (Methodenkompetenz: Re-Konstruktion)

In der folgenden Aufgaben steht ein anderer Aspekt der Methodenkompetenz (Re-Konstruktion) im Zentrum: Die sinnhafte Verknüpfung von mindestens zwei zeitlich differenten Ereignissen als Grundmuster einer Narration (vgl. Kapitel 2.1) erfordert, Zeitlichkeit in der Abfolge, Gleichzeitigkeit, Dauer und im Abstand von Ereignissen wahrnehmen zu können.

In der Aufgabe zur Herstellung einer Chronologie werden kurze Narrationen (im Nachbau zum Ersten Weltkrieg und der Russischen Revolution) angeboten, verbunden mit der Aufgabe, die berichteten Ereignisse chronologisch zu sortieren.

Aufgabebeispiel 6

📖 Mat. 1: Der Sturz des Zaren

Nach der Niederlage Russlands im Krieg gegen Japan (1904/1905) häuften sich die Unruhen im Land. Der Ruf nach politischen Reformen wurde immer lauter und radikale Kräfte wie die Bolschewiki mit ihrem Führer Lenin planten den Sturz des Zaren. […] Im Februar 1917 demonstrierten Tausende in der russischen Hauptstadt. Anfänglich forderten sie nur „Brot", später wurde daraus „Schluss mit dem Krieg" und „Nieder mit der Zarenherrschaft". Die Demonstrationen entwickelten sich zum Generalstreik. Die Armee stellte sich auf die Seite der Streikenden und verweigerte den Gehorsam. Daraufhin dankte Zar Nikolaus II. am 2. März 1917 ab […]

📖 Mat. 2: Der Erste Weltkrieg

Aber schon bald zeigte sich, dass der Krieg zu einer langwierigen mörderischen Materialschlacht geworden war, die keine Seite gewinnen konnte. Erst das Eingreifen der USA 1917 brachte die Entscheidung. Nach vier Jahren Krieg bat Deutschland 1918 um einen Waffenstillstand.

In der folgenden Tabelle findest du ein paar Ereignisse, die in den beiden Texten benannt werden

A	Februarrevolution in Russland
B	Ende des Ersten Weltkriegs
C	Russisch-Japanischer Krieg
D	Eintritt der USA in den Ersten Weltkrieg

Bringe diese Ereignisse in die passende zeitliche Reihenfolge und trage die Buchstaben der Ereignisse in den Zahlenstrahl ein.

Aufgabenbeispiel 7: Interpretationen von Geschichte analysieren und vergleichen (Methodenkompetenz: De-Konstruktion)

Das nachfolgende Aufgabenbeispiel zielt auf die Kernkompetenz der De-Konstruktion. Wenn die Struktur einer vorliegenden Narration untersucht wird, geht es u.a. darum, die in den Narrationen vorgenommenen Schlussfolgerungen für die Gegenwart und Zukunft überhaupt zu erkennen. Ob die Lernenden diese Teiloperation beherrschen, wurde erfasst, indem sinnbildende Interpretationen nebeneinander gestellt wurden – im Nachbau mit Bezug zu den Gallienfeldzügen von Cäsar. Die Schülerinnen und Schüler sollten entscheiden, welche Interpretationen in welchem Text zu finden sind.

Aufgabenbeispiel 7

Gaius Iulius Caesar ist auch heute noch – nicht zuletzt durch die Comicreihe "Asterix" – eine der bekanntesten Persönlichkeiten der Antike. Alle Historiker sind sich einig, dass Caesars Eroberungen in Gallien mit enormen Opferzahlen unter der einheimischen Bevölkerung verbunden waren. Die Bedeutung der Feldzüge Caesars wird in der historischen Forschung unterschiedlich beurteilt.

Lies die Interpretationen der Historiker aufmerksam durch. Du sollst sie anschließend vergleichen.

🕮 **Mat. 1: T. Mumson:**

"Caesars Gallienfeldzüge forderten Blutzoll, und ja: Caesar handelte zum Großteil eigenmächtig, eigensinnig und arrogant. Aber ohne Caesars Eroberung des Westens und die hierauf folgende Verbreitung römischer Lebensweisen (*Romanisierung*) hätte die große Völkerwanderung der Spätantike bereits 400 Jahre früher stattgefunden; die westliche Zivilisation, wie wir sie kennen, wäre nie entstanden. Dass das griechisch-römische Fundament des modernen Europa erhalten geblieben ist, ist somit Caesars Werk, auch wenn er hiervon natürlich nichts wissen konnte."

🕮 **Mat. 2: F. Dumjanzki:**

"Caesar wusste, dass die römische Republik am Ende war und dass in der römischen Politik nun nur noch das Recht des Stärkeren galt. Hierfür brauchte er jedoch Ressourcen (Machtmittel) und ihm ergebene Truppen. Die Eroberung Galliens war hierfür das ideale Betätigungsfeld: Caesar konnte das Gebiet wie eine Schatzkammer plündern und es gleichzeitig als Truppenübungsplatz nutzen. Sein Vorgehen mag man als kaltblütig bewerten, aber beim Eroberungsdrang der Römer wäre über kurz oder lang ein anderer Feldherr gekommen, der genauso gehandelt hätte wie Caesar. Caesar war schlichtweg schneller, schlauer und konsequenter."

🕮 **Mat. 3: B. Kirnon:**

"Caesars brutales Vorgehen in Gallien ist durch nichts zu entschuldigen. Sein Vorgehen sorgte selbst in der römischen Oberschicht für Widerspruch, die für ihre Geltungs- und Eroberungssucht berüchtigt war. Dass man dort forderte, Caesar an feindliche germanische Stämme auszuliefern, beweist: Er ging selbst für die Maßstäbe seiner eigenen Zeit massiv zu weit. Mehr als 400.000 Menschen in einem einzigen Feldzug niederzumetzeln (so Caesars eigene Angaben über den Feldzug 56/55 v. Chr. gegen die Usipeter und Tenkterer) ist Völkermord, damals wie heute."

Zu welchen Einschätzungen kommen die Autoren jeweils?
Kreuze für jeden der Autoren „nein“ oder „ja“ an.

Für diese Autoren sind Caesars Feldzüge ein Beispiel dafür, dass…	Mat. 1		Mat. 2		Mat. 3	
	Nein	Ja	Nein	Ja	Nein	Ja
… die Menschenrechte überzeitliche Gültigkeit besitzen.	x	☐	x	☐	☐	x
… in der Geschichte die Taten Einzelner zu Entwicklungen führen, die diese gar nicht beabsichtigt hatten.	☐	x	x	☐	x	☐
… es in der Weltgeschichte Situationen gibt, in denen militärisches Vorgehen gegen fremde Stämme und Völker als „in Ordnung“ eingeschätzt wird.	x	☐	☐	x	☐	x
… auch negative Taten langfristig positive Folgen haben können.	☐	x	x	☐	x	☐

3.3.4 Beispiel-Aufgaben zur historischen Orientierungskompetenz

Die Kompetenz, sich historisch orientieren zu können (vgl. Kapitel 2.3.2.3), mit standardisierten Items zu überprüfen, ist eine Herausforderung, weil die Fähigkeiten historischen Denkens und die Einstellungen zu den in den Items enthaltenen Sinnbildungen sehr stark aufeinander bezogen sind. Im HiTCH-Test haben wir, wie bereits erläutert, bewusst darauf verzichtet, individuelle Orientierungen zu thematisieren. Es wurden vielmehr Aufgabenblöcke formuliert, in denen vorgegebene historische Orientierungsangebote reflektiert werden sollten.

Die nachgebaute Beispielaufgabe (Beispielaufgabe 8) testet, inwiefern die Schülerinnen und Schüler in der Lage sind, narrative Aussagen mit expliziten Vergangenheits- und Zukunftsbezügen zu systematisieren und unterschiedlichen Sinnbildungsmustern zuzuordnen (zu Sinnbildungsmustern nach Rüsen vgl. Kapitel 2.2.3.4).

Aufgabenbeispiel 8

Gib an, welche Absicht am besten zu der jeweiligen Aussage passt. Setze bitte nur ein Kreuz pro Zeile.

	Absicht			
	Verweist auf Bewährtes und will so Sicherheit geben.	Durch vergangene Beispiele soll eine grundsätzliche Regel gezeigt werden.	Durch Erklärung vergangener Entwicklungen sollen Veränderungen abgeschätzt werden.	Kritisiert Vergangenes und will Bestehendes nicht weiterführen, sondern verändern.
An der Beschäftigung mit früheren Kriegen wird deutlich, dass Kriege viele Menschen so erschüttern, dass sie danach nicht mehr gut weiterleben können.	☐	x	☐	☐
Der Fortschrittsgedanke von höher, schneller, weiter wie bei den Olympischen Spielen hat nicht zu schöneren Wettkämpfen geführt. Der Kampf um Aufmerksamkeit und Sponsoren schadet dem Sport.	☐	☐	☐	x
Seit die Verfassung der USA 1787 in Kraft trat, sind die Rechte und Freiheit der amerikanischen Bürger garantiert.	x	☐	☐	☐
China hat es in den letzten Jahrzehnten geschafft, sich von einem armen Land zu einer konkurrenzfähigen Wirtschaftsmacht zu entwickeln. Schon bald wird China vermutlich die größte Wirtschaftsmacht der Welt sein.	☐	☐	x	☐

3.4 Der Prozess der Aufgabenentwicklung für den nunmehr vorliegenden HiTCH-Test

3.4.1 Adressieren des gesamten Denkprozesses in themenbezogenen Testheften

Die jetzige Version des HiTCH-Tests entstand als Produkt eines iterativen Prozesses, bei dem auf der Grundlage des FUER-Modells Testaufgaben entwickelt bzw. optimiert wurden und daraufhin empirisch auf ihre psychometrische Eignung geprüft wurden. Den allgemeinen Startpunkt der weiteren Testentwicklung bildeten Themenhefte, in denen der ganze Regelkreis des historischen Denkens an jeweils einem Thema („Pest", „Nürnberger Prozesse", „Nationen", „Hexenverfolgung", „Untergang Roms", „Ostrakismos", „USA", „DDR") und mit Aufgaben, die alle Kompe-

tenzbereiche adressierten, durchlaufen wurde. Dieses Vorgehen hatte zum Ziel, den einzelnen Standorten die Entwicklung tragfähiger Aufgabenkonzepte, die historisches Denken adressieren, zu erleichtern und zu gewährleisten, dass die historischen Kompetenzen in wünschenswerter Breite erfasst werden.

Zu jedem Thema wurden jeweils Aufgaben entwickelt, die überprüften, ob bzw. inwieweit die Lernenden eigene Fragen stellen bzw. ob und inwieweit sie vorgegebene Fragestellungen erfassen können. Es wurde multiperspektivisches Quellenmaterial vorgelegt, zu dem re-konstruktive Leistungen eingefordert wurden. Ausgewählte darstellende Texte wurden genutzt, um die Methodenkompetenz (De-Konstruktion) zu erfassen. Anhand des eingesetzten Materials wurde zudem – als eine Facette der Sachkompetenz – die Kompetenz im Umgang mit dem Quellen- und Darstellungsbegriff ermittelt. Die Fähigkeit und Bereitschaft, die so gewonnenen Erkenntnisse für das Verstehen der Gegenwart und für Orientierungen in der Zukunft zu nutzen, wurde in weiteren Items erfasst. Darüber hinaus wurden Aufgaben entwickelt, die den Bereich der Sachkompetenz adressierten, ohne eine bestimmte Epoche oder ein konkretes Thema anzusprechen (z.B. Einsicht in epistemologische Prinzipien, Verständnis zentraler Begriffskonzepte, wie z.B. Staats- und Herrschaftsformen).

3.4.2 *Cognitive Labs*

Alle neu entwickelten Aufgabenblöcke und Items wurden mit einzelnen Testpersonen oder in kleinen Gruppen aus der Zielpopulation erprobt, um die Verständlichkeit der Materialen und Itemformulierungen sowie die Schwierigkeiten der Aufgaben zu evaluieren.

Darüber hinaus wurden für ausgewählte Aufgaben sogenannte *Cognitive Labs* durchgeführt (vgl. Werner & Schreiber, 2015). Es handelt sich dabei um Interviews während und unmittelbar nach der Bearbeitung der Testaufgaben. Spontane Äußerungen der Schülerinnen und Schüler (Methode des „lauten Denkens“) wurden durch situative Nachfragen („prompts“) vertieft, welche sich unter anderem auf das Verständnis der Aufgabe, auf Lösungsstrategien, auf das vorgängige Begriffsverständnis oder die Begründung, warum eine bestimmte Antwortalternative gewählt wurde, bezogen (Willis, 2005).

Ziel war es, die Denkprozesse der Schülerinnen und Schüler, die durch geschlossene Aufgaben ausgelöst wurden, besser zu verstehen. Dazu wurde zum einen untersucht, inwieweit die Denkprozesse während der Testbearbeitung auf Kompetenzen historischen Denkens hinweisen, zum anderen wurde die Auswirkung der gewählten Themen auf die Denkprozesse adressiert. Daher wurden zusätzlich Informationen zu Interesse, Vorwissen und geschichtskultureller Erfahrung der Schülerinnen und Schüler zu den jeweiligen historischen Themen erhoben und bei der Auswertung berücksichtigt.

Die Auswertung der für die Fragestellungen relevanten Schüleräußerungen erfolgte inhaltsanalytisch und computerunterstützt (mit Hilfe der Software MAXQDA).

Der zentrale Befund der Analysen war, dass durch geschlossene Aufgaben tatsächlich historische Denkprozesse ausgelöst werden können (Werner & Schreiber, 2015).

Die Informationen aus den Cognitive Labs wurden genutzt, um die Testitems zu präzisieren und zu verbessern. Beispielsweise war wiederholt zu beobachten, dass die Einstellungen und Vorstellungen zu Museen die Bearbeitung von Items, welche epistemologische Einsichten am Beispiel einer historischen Ausstellung thematisierten, beeinflussten. Um diese Einflüsse auszuschließen, wurde bei der weiteren Testentwicklung darauf geachtet, Erkenntnisprinzipien des Faches Geschichte ohne Bezug auf geschichtskulturelle Medien zu erfassen.

Auch für die Abkehr von der Gestaltung der Tests als geschlossene Themenhefte trugen die Cognitive Labs bei: Sie erbrachten z.B., dass Schülerinnen und Schüler die eingangs zum Themenkomplex gegebenen Informationen nicht während der gesamten Bearbeitung präsent hielten, dass sie Rückbezüge auf zuvor bearbeitete Aufgaben herstellten, die zu Missverständnissen von Aufgabenstellungen führten, oder dass irreführende Nutzungen vorheriger Lösungen erfolgten.

3.4.3 Pilotierung I

Die in der ersten Runde der Itemgenerierung entwickelten Aufgaben wurden in der Pilotierung I im Jahr 2013 an insgesamt 1701 Schülerinnen und Schülern aus 9. Klassen aller Schulformen getestet (Gymnasium: 59,8%, andere Schulformen: 40,2%; Durchschnittsalter: 14,89 Jahre; Mädchen: 51,2%; Baden-Württemberg: 25,2%, Bayern: 31,1%, Hamburg: 30,3%; Nordrhein-Westfalen: 6,9% und Sachsen: 6,5%). In dieser Pilotierung wurden insgesamt 233 Aufgabenblöcke und 1308 Items mithilfe eines sogenannten Multi-Matrix-Designs, bei dem jeder Schüler bzw. jede Schülerin nur eine Teilmenge der Aufgaben erhält, eingesetzt. Jede Schülerin und jeder Schüler bearbeitete eines von acht Themenheften, die jeweils ein Thema vertieften, so dass jedes Themenheft von jeweils ca. 10,0% bis 13,6% der Schülerinnen und Schüler bearbeitet wurden ($170 \leq N \leq 231$). Ein Teil der Aufgabenblöcke zur Testung der Sachkompetenz (z.B. Begriffskompetenz) wurde von allen bearbeitet und wurde in den Analysen für eine gemeinsame Verankerung verwendet. Ein weiterer Teil von Aufgabenblöcken zur Sachkompetenz (z.B. zur Einsicht in epistemologische Prinzipien) lag in zwei Varianten vor und wurde von jeweils der Hälfte der gesamten Stichprobe ausgefüllt. Neben den Kompetenzaufgaben wurde der persönliche Hintergrund der Lernenden erfasst (z.B. sozio-ökonomischer Hintergrund, Schulnoten).

Ziel der Datenauswertung war die Itemauswahl und -optimierung aufgrund statistischer Verfahren. Daher wurden die Items deskriptiv anhand ihrer Mittelwerte und Streuung beschrieben, was u.a. ermöglichte, die Schwierigkeit der Items einzuschätzen. In Reliabilitätsanalysen wurden die interne Konsistenz der Skalen sowie die Itemtrennschärfe und die korrelativen Zusammenhänge zwischen den Aufgabenblöcken bezogen auf die einzelnen Testhefte überprüft. Zudem wurde untersucht, welchen Beitrag die einzelnen Aufgabenblöcke bzw. Items zu einem Gesamtmaß histori-

scher Kompetenzen leisten können. Hierfür wurde mit den – auf der Skalenebene im Hinblick auf ihre interne Konsistenz reliablen – Aufgabenblöcken eine Faktorenanalyse mit nur einem Faktor gerechnet, der den Globalfaktor „historische Kompetenz" abbildete. Aus der Ladung auf diesen einen Faktor ergab sich eine Reihung der eingesetzten Aufgabenblöcke mit akzeptabler Reliabilität im Hinblick auf ihren jeweiligen Beitrag zur Messung des Globalfaktors „historische Kompetenz".

Da für jede Kompetenzdimension (Kompetenzbereiche/Kernkompetenzen) pro Testheft nur wenige Aufgabenblöcke vorlagen und zudem die Testhefte thematisch fokussiert waren (was zu einer Konfundierung der Kompetenzen mit dem Thema führen könnte), erschien eine mehrdimensionale Betrachtung der Items designbedingt nicht sinnvoll. Daher wurde in mehreren einfachen und multiplen logistischen Regressionen ein Globalfaktor zur (groben) Beurteilung der psychometrischen Eigenschaften der Aufgabenblöcke herangezogen. Zum einen wurde jedes Item im Hinblick auf die Ladung auf dem Globalfaktor geschätzt, zum anderen wurde überprüft, ob ein Item im Gymnasium vs. Nicht-Gymnasium oder im Osten oder Westen der Bundesrepublik unterschiedliche Messeigenschaften aufwies (sogenannte DIF-Effekte).

Die Pilotierung I erbrachte folgende zentrale Befunde: (1) Trotz einer eher schwachen Verlinkung der Testitems als Konsequenz des verwendeten Multi-Matrix-Designs fanden sich Belege dafür, dass sich mithilfe (einer Auswahl) der standardisierten Testaufgaben eine gemeinsame Kompetenz („historische Kompetenz") messen lässt. (2) Die Schwierigkeiten (= Mittelwerte) und Werteverteilungen (= Standardabweichungen) einer großen Zahl von Items lagen in einem wünschenswerten Bereich. (3) Nur bei wenigen Items fanden sich deutliche Indizien dafür, dass sie für bestimmte Schülergruppen besonders einfach oder schwierig wären (sogenannte DIF-Effekte). Auf Basis der Befunde der Pilotierung I wurden auf Basis der psychometrischen Kennwerte sowie unter Berücksichtigung fachdidaktischer Überlegungen (u.a. Passung zum Kompetenzmodell, Breite der gemessenen Kompetenz) Aufgabenblöcke sowie Items beibehalten, optimiert oder ausgeschlossen.

3.4.4 Pilotierung II

In der Pilotierung II sollte zusätzlich zur weiteren Optimierung des Itempools auch geprüft werden, ob jenseits des Globalfaktors weitere Dimensionen der „historischen Kompetenzen" nachgewiesen werden können, also ob sich etwa die theoretisch definierten Kompetenzbereiche oder die Basisoperationen der Re- und De-Konstruktion unterscheiden lassen.

Aus bereits erprobten Aufgabenblöcken aus der Pilotierung I wurde eine Aufgabenauswahl gebildet, die allen teilnehmenden Schülerinnen und Schülern vorgelegt wurde. Dieser gemeinsame Itempool umfasste pro Kompetenzbereich mindestens zwei Aufgabengruppen. Die Aufgaben bezogen sich auf unterschiedliche Themenbereiche, womit der Einfluss von Vorwissen, Interesse an spezifischen Themen und ge-

schichtskultureller Präsenz auf die Ausprägung der Leistung bei dieser Itemgruppe reduziert wurde. Über die allen Schülerinnen und Schülern vorgelegten Aufgaben hinaus wurden neu entwickelte Aufgaben, z.T. zu weiteren Themengebieten, eingesetzt. An der Entwicklung weiterer Aufgaben beteiligten sich auch assoziierte Partner der Universitäten bzw. Hochschulen Aarau (Schweiz), Bochum (Deutschland), Salzburg (Österreich). Damit erweiterte sich das Themenspektrum bspw. um Aufgaben zur japanischen Geschichte, König Artus oder zu Jugendkrawallen in der Schweiz in den 1980er Jahren. Zudem wurden anhand von kleinen Materialauszügen De-Konstruktionsaufgaben an verschiedenen Themen und Materialarten durchgespielt (z.B. Comics, Karikaturen, Denkmäler). Es wurde erneut ein Multi-Matrix-Design realisiert, indem die Aufgaben in insgesamt fünf unterschiedlichen Testheften zusammengestellt wurden, die jeweils dieselben Ankeraufgaben umfassten (s.o.).

Die zweite Pilotierung wurde im Jahr 2014 durchgeführt. Es nahmen 1295 Schülerinnen und Schüler der 9. Jahrgangsstufe in Deutschland, Österreich und der Schweiz teil (Gymnasium: 52,6 %, andere Schulformen; 47,4%; Durchschnittsalter: 14,80 Jahre; Mädchen: 52,0%; Bayern: 22,6%, Hamburg und Schleswig-Holstein: 31,0%; Nordrhein-Westfalen: 21,0%, Schweiz: 19,1%, Österreich: 6,3%).

Mithilfe von Faktorenanalysen wurde bestimmt, ob eingesetzte Items auf demselben Faktor laden und somit die gleiche Kompetenzdimension erfassen. Als Grundüberlegung steht hinter den Faktorenanalysen, dass das eigentlich interessierende Merkmal (oder die „latente Variable"), z.B. die Einsicht der Lernenden in epistemologische Prinzipien, nicht direkt messbar ist, sondern anhand mehrerer Items operationalisiert wird (vgl. Kapitel 3.2.2). Werden verschiedene Items in ähnlicher Weise gelöst, dann ist dies ein Hinweis darauf, dass diese Items inhaltlich zusammengehören und dass es sich um eine abgrenzbare Kompetenzdimension handelt. Die Faktorenanalysen geben zusätzlich eine Auskunft darüber, wie gut die einzelnen Items das zugrunde liegende Konstrukt abbilden. Wenn ein Item in der Faktorenanalyse eine „hohe Ladung" hat, d.h. einen starken Zusammenhang mit der latenten Variable aufweist, dann kann dieses Item als ein relevanter Indikator für die zugrunde liegende latente Variable betrachtet werden. Niedrige Ladungen hingegen bedeuten, dass sich ein Item als Indikator zur Messung der latenten Variablen weniger gut eignet.

Als Ergebnis der Faktorenanalysen zeigte sich, dass ein Ein-Faktor-Modell die Struktur der Daten am besten beschrieb. Auf diesem Faktor luden Items aus allen vier Kompetenzbereichen (Fragekompetenz, Methodenkompetenz (Re- und De-Konstruktion), Orientierungskompetenz, Sachkompetenz) des FUER-Modells. Dieser Faktor kann als historische Kompetenz interpretiert werden; er diente zugleich dazu, die Auswahl der Items für die Haupterhebung zu begründen.

Eine weitergehende Differenzierung der Items zu unterschiedlichen Kompetenzbereichen in unterschiedliche Dimensionen (Faktoren) wurde geprüft, konnte aber auf Basis der Daten der Pilotierung II nicht bestätigt werden. Mögliche Gründe hierfür könnten sein, dass (1) die prozeduralen Kompetenzbereiche und der kategorisierende Kompetenzbereich theoretisch so eng zusammenhängen (vgl. Überlappungsbereiche, Kapitel 2.3.3), dass sie empirisch schwer zu separieren sind, (2) eine etwaig

vorhandene mehrfaktorielle Struktur der historischen Kompetenz vom Generalfaktor „historische Kompetenz“ oder von itemspezifischen Effekten überlagert wird, (3) die Separierung der Kompetenzbereiche und ihrer Kernkompetenzen in den verwendeten Items noch nicht ausreichend geglückt ist, (4) die Items Einzelkompetenzen aktivieren, die mit mehreren Kompetenzbereichen zusammenhängen.

Neben der Überprüfung der Dimensionalität und der Erprobung neuer Aufgaben hatte die zweite Pilotierung auch das Ziel, die psychometrisch geeignetsten Aufgaben für die Haupterhebung (siehe Kapitel 4) zu identifizieren. Bereits in dieser Phase wurden komplexe statistische Analyseverfahren verwendet, die jedoch erst in den Kapiteln 4 und 5 genauer beschrieben werden.[19] Die aufgrund psychometrischer Überlegungen besten Aufgaben gehören 15 Aufgabenblöcken mit bis zu 16 Items an und umfassen insgesamt 143 Items.[20] Die Auswahl wurde für die Haupterhebung um einige aus fachdidaktischen Gründen unverzichtbare Aufgaben auf insgesamt 152 Items ergänzt.[21] Dieser Pool an Aufgaben musste wegen der angezielten Bearbeitungszeit noch reduziert und an wenigen Stellen sprachlich und strukturell optimiert werden. Als Grundlage dafür wurden in einer internen Expertenberatung fachdidaktisch ausgewiesene Wissenschaftlerinnen und Wissenschaftler der HiTCH-Standorte gebeten, alle Aufgaben zu lösen und die Aufgabenschwierigkeit bzw. Verständlichkeit und die eigenen Erfahrungen während der Aufgabenlösung zu kommentieren. Die Ergebnisse flossen in die Auswahl und Gestaltung der Aufgaben für die Haupterhebung mit ein.

3.4.5 Haupterhebung

Die Haupterhebung diente der Überprüfung der Reliabilität und Validität des HiTCH-Instrumentariums. Sie wird im nachfolgenden Kapitel 4 vertieft beschrieben. An dieser Stelle werden die Ziele nur kurz skizziert:

In der Hauptstudie sollte insbesondere untersucht werden, (1) ob das HiTCH-Instrument eine hinreichende Homogenität für eine eindimensionale Skalierung aufweist (was als Hinweis darauf verstanden werden kann, dass die verwendeten Aufgaben ein- und dieselbe Kompetenz erfassen), (2) ob die Messeigenschaften über verschiedene Subgruppen hinweg ähnlich sind (keine DIF-Effekte und somit eine „faire“ Erfassung der Kompetenz) und (3) ob sich ein plausibles Muster von Korre-

19 Insbesondere kamen eindimensionale Rasch-Modelle wie auch konfirmatorische Nested-Factor-Ansätze zur Anwendung. Die eindimensionale Rasch-Analyse in ConQuest (Wu, Adams, Wilsons & Haldane, 2007) zeigte eine hohe Score-Reliabilität (WLE-Person Separation Reliability: .88). 39% der Varianz ließ sich durch die Klassenzugehörigkeit erklären, die – wie üblich bei Schulleistungstests in mehrgliedrigen Schulsystemen (z.B. Baumert, Trautwein & Artelt, 2003) – auf Schulartunterschiede zurückgeführt werden kann.

20 Genauer gesagt handelt es sich um 86 Items sowie 57 Subitems, von denen jeweils drei zu einem Item zusammengefasst wurden (Complex Multiple Choice), und neun Subitems einer Reihenfolgeaufgabe, die zu einem (Partial Credit-)Item zusammengefasst wurden.

21 Die Zusammenfassung von Einzelinformationen aus Subitems zu Items in der Pilotstudie und der Haupterhebung war nicht identisch. Aufgrund dieser Unterschiede wird hier zunächst die Itemanzahl im Sinne der Variablenanzahl im Datensatz (also Variablen für Items bzw. Subitems) genannt.

lationen mit Außenkriterien (und damit Hinweise auf eine hohe Validität des Tests) finden lassen würde. Die Validität des Instruments wurde hinsichtlich der Abgrenzung von benachbarten Kompetenzbereichen (Lesekompetenz) und weiteren Kriterien (z.B. Intelligenz, Schulform, Schulnoten) untersucht.

Neben den in die Haupterhebung integrierten Validierungsbestandteilen wurden weitere Validierungsstudien durchgeführt, deren Befunde teilweise im vorliegenden Band berichtet werden:

1) In fünf Klassen an zwei baden-württembergischen Gymnasien (N = 111) wurde, unter Bezug auf die Arbeiten von Wineburg (vgl. Kapitel 2.4.2), überprüft, ob die Leistung im HiTCH-Test prädiktiv ist für die Anwendung adäquater Strategien beim Umgang mit historischem Quellenmaterial.
2) An bayerischen Mittel-, Realschulen und Gymnasien (N = 500) wurde neben dem HiTCH-Test der Test zur Erfassung Literarästhetischer Urteilskompetenzen (LUK; Frederking, 2008; Frederking & Brüggemann, 2012; Frederking, Brüggemann & Hirsch, 2016; Frederking, Meier, Brüggemann, Gerner & Friedrich, 2011; Frederking, Meier, Stanat & Dickhäuser, 2008; Frederking, Roick & Steinhauer, 2011) eingesetzt. Es sollte nicht nur die Abgrenzung zur allgemeinen Lesefähigkeit, sondern auch zum literarischen Verstehen geprüft werden.
3) In Hamburg und Paderborn wurde das HiTCH-Instrument an 146 Studierenden erprobt. Ziel war die Auslotung des Potenzials der HiTCH-Aufgaben für junge Expertinnen und Experten historischen Denkens (Meis & Zuckowski, in Druck; Meyer-Hamme & Körber, in Vorb.).

4 Haupterhebung: Design, Datenerhebung, Methoden der Datenauswertung

4.1 Stichprobe

An der Haupterhebung nahmen 2.853 Schülerinnen und Schüler aus insgesamt 52 Schulen der 9. Jahrgangsstufe teil: 49,2% weiblich, Durchschnittsalter: 14,41 Jahre, Gymnasium: 53,7%, Gesamtschulen (hier wurden auch Stadtteilschulen, Gemeinschaftsschulen etc. zugeordnet): 22,4%, Realschulen: 15,9%, Hauptschulen: 7,9%, Baden-Württemberg: 22,9%, Bayern: 4,1%, Hamburg/Schleswig-Holstein: 25,3%, Nordrhein-Westfalen: 17,0%, Thüringen: 0,6%, Schweiz: 9,5% und Österreich: 20,6%. Die Schulen wurden – nach Einholung der Genehmigung der zuständigen Schulbehörden – von Mitgliedern des HiTCH-Projekts direkt angesprochen und für die Teilnahme gewonnen. Bei der Gewinnung der Stichprobe wurde sowohl auf eine hinreichende geographische Differenzierung als auch auf eine Berücksichtigung unterschiedlicher Schulformen geachtet. Eine Repräsentativität der Stichprobe wurde jedoch nicht angestrebt, da die Untersuchungsziele der Stichprobe diese nicht erforderlich machten und keine Mittelwertvergleiche zwischen (Bundes-)Ländern und Schulformen im Sinne einer Leistungsvergleichsstudie angestrebt waren. Von daher wurde auch durchgängig auf eine Gewichtung der Stichprobe verzichtet.

4.2 Testdesign

Allen Schülerinnen und Schülern wurde ein gemeinsames Heft vorgelegt, das die für den finalen HiTCH-Test vorgesehenen Aufgaben (15 Aufgaben mit 106 Items) enthielt, die nach den beiden Pilotierungen auf der Basis psychometrisch und fachdidaktisch begründeter Auswahlprozeduren ausgewählt worden waren. Dieses Testheft enthielt darüber hinaus in Form eines Schülerfragebogens einige wenige Angaben zum familiären Hintergrund der Schülerinnen und Schüler, ihrer schulischen Motivation sowie zu ihrem Leistungsstand (Schulnoten). Darüber hinaus wurde eines von sechs zusätzlichen Testheften bearbeitet. 451 Schülerinnen und Schülern führten einen Test zur Erfassung der kognitiven Grundfähigkeiten durch und 616 einen Lesekompetenztest. Zudem wurden drei Testhefte mit zusätzlichen historischen Kompetenzaufgaben vorgelegt (Zusatz-Kompetenztest 1: $N = 453$; Zusatz-Kompetenztest 2: $N = 432$; Zusatz-Kompetenztest 3: $N = 588$), die in künftigen Weiterentwicklungen des HiTCH-Tests ihren Platz finden sollen. Daneben gab es ein Zusatzheft zu Geschichtsbewusstsein bzw. -sozialisation ($N = 309$). Vier Schülerinnen und Schüler haben kein Zusatzheft bearbeitet. Im Folgenden wird nur über die Zusammenhänge des HiTCH-Tests mit den ersten beiden Zusatzheften berichtet (Lesen, Intelligenz).

4.3 Validierungsinstrumente

Ein zentrales Ziel der Haupterhebung war die Überprüfung der diskriminanten und konvergenten Validität des HiTCH-Instruments. Hierzu wurden neben Informationen aus dem Schülerfragebogen (insbesondere Geschlecht, Interesse und schulische Leistungen) insbesondere die Testleistungen in zwei der sechs zusätzlich administrierten Testhefte (kognitive Grundfähigkeiten und Lesekompetenz) herangezogen.

Kognitive Grundfähigkeiten. Zur Erfassung der allgemeinen kognitiven Fähigkeiten wurden zwei Skalen (figural und verbal) aus dem KFT4-12+R (Heller & Perleth, 2000) für die 9. Jahrgangsstufe eingesetzt. In der vorliegenden Stichprobe ergaben sich auf Basis einer Rasch-Skalierung akzeptable Score-Reliabilitäten für beide Facetten (WLE PSR: .85 für figurale, .62 für verbale kognitive Grundfähigkeiten). Der KFT wird auch in Schulleistungsstudien wie PISA (Baumert, Stanat & Demmerich, 2001) häufig als Kontrollvariable eingesetzt.

Lesekompetenzen. Die Lesekompetenz der Schülerinnen und Schüler wurde mit der „Lesetestbatterie" für die Klassenstufe 8-9 (Bäuerlein, Lenhard & Schneider, 2012) geprüft, die die basale Lesekompetenz sowie das tiefergehende Textverständnis erfasst. Zur Überprüfung der basalen Lesekompetenz sind aus einer Liste kurzer, einfacher Sätze innerhalb von drei Minuten möglichst viele zu lesen und auf die inhaltliche Richtigkeit hin zu beurteilen. Zur Erfassung des Textverständnisses werden ein expositorischer und ein narrativer Text mit jeweils 19 Multiple-Choice-Verständnisfragen vorgelegt, wobei sich die Fragen auf unterschiedliche Ebenen des Textverständnisses bzw. verschiedene Formen der Textrepräsentation beziehen. Alle Schülerinnen und Schüler erhielten den Test zur Erfassung der basalen Lesekompetenz, jeweils die Hälfte den Test zum Verständnis des expositorischen bzw. des narrativen Tests. Die beiden Tests zum Textverständnis wurden Rasch-skaliert, wobei sich in der vorliegenden Stichprobe akzeptable Score-Reliabilitäten ergaben (WLE PSR: .72 für den Test mit expositorischer, .68 für den mit narrativer Textgrundlage).

Testmotivation. Nach Bearbeitung der Testinstrumente wurden den Schülerinnen und Schülern fünf Items zur Erfassung der Testmotivation vorgelegt: „Ich habe die Aufgaben sorgfältig bearbeitet", „Ich war konzentriert", „Ich habe mich angestrengt", „In einer Klassenarbeit hätte ich mich mehr angestrengt" (umgepolt) sowie „Die Bearbeitung des Testhefts hat mir Spaß gemacht". Die resultierende Skala wies eine interne Konsistenz (Cronbachs Alpha) von .76 auf.

4.4 Analysestrategien im HiTCH-Projekt

Trotz der weiten Verbreitung der Klassischen Testtheorie (KTT) ist es inzwischen aufgrund gut bekannter Grenzen der KTT (z.B. nicht überprüfbare Voraussetzungen wie Unidimensionalität, Intervallskalenniveau der Daten; Moosbrugger & Kelava, 2012) in der Schulleistungsforschung üblich, komplexere, aber weniger voraussetzungsreiche Verfahren zu verwenden. Für die Skalierung von (Kompetenz-)Tests

mit kategorialem Antwortformat (dichotome oder ordinale Daten, also z.B. falsche – teilrichtige – richtige Lösung) haben sich Modelle der Item Response Theorie (IRT) etabliert. Im Folgenden werden die für die nachfolgenden Auswertungen zentralen Konzepte kurz erläutert. Für eine gründliche Einführung in die Item Response Theorie, die im Rahmen dieses Bandes natürlich nicht zu leisten ist, sei auf die Arbeiten von de Ayala (2009), Moosbrugger und Kelava (2012) sowie Rost (2004) verwiesen. Da die exakte Darstellung der verwendeten Verfahren und der dabei resultierenden psychometrischen Kennwerte für die kritische Rezeption durch das Fachpublikum unverzichtbar ist, sind in dieser Hinsicht auch eher „technisch" anmutende Darstellungen, wie sie den vorliegenden Abschnitt kennzeichnen, notwendig. Um den Ergebnisteil (Kapitel 5) jedoch gleichzeitig einem möglichst breiten Rezipientenkreis zu erschließen, werden dort die zentralen psychometrischen Befunde jeweils auch in ihrer inhaltlichen Bedeutung dargestellt.

Im Rahmen von IRT-Modellen werden latente Variablen (also nicht direkt beobachtbare Merkmale, wie z.B. historische Kompetenz) und deren Zusammenhänge anhand von beobachteten Daten (gelöste, nicht gelöste Items) geschätzt. Bei Leistungstests geht man üblicherweise von – einer oder mehreren – quantitativen latenten Variablen aus, die als Dimensionen bezeichnet werden. So lassen sich beispielsweise Testleistungen aus unterschiedlichen Domänen (z.B. Kompetenzen in den Bereichen Mathematik und Naturwissenschaften) oder Subfacetten innerhalb von domänenspezifischen Kompetenzen unterscheiden (z.B. PISA-Subskalen der Lesekompetenz: Informationen suchen und extrahieren, textbezogenes Kombinieren und Interpretieren, Reflektieren und Bewerten; Klieme et al., 2010). Die latenten Variablen werden dabei als Prädiktor für die Lösungshäufigkeit auf den einzelnen Items (den Indikatoren der latenten Variablen) betrachtet. Je höher die Ausprägungen auf einer Dimension (also z.B. hohe historische Kompetenz), desto höhere Lösungswahrscheinlichkeiten sind bei den entsprechenden Items zu erwarten. Dabei kann dieser Zusammenhang prinzipiell stärker oder schwächer sein, was sich in höheren oder niedrigeren Faktorladungen ausdrückt (in der IRT werden diese Ladungen als Diskriminationsparameter bezeichnet; Glockner-Rist & Hoijtink, 2003). Außerdem können Items generell „leichter" oder „schwieriger" als andere sein, also durchschnittlich höhere oder niedrigere (relative) Lösungshäufigkeiten (Itemschwierigkeiten) aufweisen.

Der Zusammenhang zwischen einer latenten Variable und den Item-Responses (also den Antworten auf die Testitems) kann nun auf unterschiedliche Weise anhand von nonlinearen Funktionen spezifiziert werden. Üblicherweise werden entweder Logit- oder Probit-Funktionen verwendet. Dabei approximieren die vorhergesagten Lösungswahrscheinlichkeiten mit abnehmender bzw. zunehmender Ausprägung auf der latenten Variable die Werte 0 bzw. 1 (bei einer linearen Funktion könnten nicht definierte Lösungswahrscheinlichkeiten – also kleiner 0 oder größer als 1 – resultieren). Für sehr leistungsstarke Schülerinnen und Schüler würde man demnach erwarten, dass viele Items korrekt gelöst werden, für sehr leistungsschwache Schülerinnen und Schüler, dass nur wenige Items korrekt gelöst werden.

Im einfachsten Fall, dem eindimensionalen einparametrischen IRT-Modell für dichotome Indikatoren (wenn eine Logit-Funktion verwendet wird, so bezeichnet man dieses Modell oft als einparametrisches logistisches Modell, kurz 1PL oder „Rasch-Modell"), hängt die Lösungswahrscheinlichkeit ausschließlich von einer einzigen latenten Variable ab, und die Items unterscheiden sich lediglich in ihren Schwierigkeiten (für jedes Item wird ein Parameter für die Schwierigkeit geschätzt). In Abbildung 8 sind die sogenannten itemcharakteristischen Kurven dreier Items dargestellt, also die Lösungswahrscheinlichkeiten über das Spektrum der Personenfähigkeiten. Die Items 1 und 2 weisen identische Diskriminationsparameter (jeweils $a_1 = a_2 = 1$) auf und unterscheiden sich lediglich in ihren Itemschwierigkeiten ($b_1 = -1$, $b_2 = 1$). Mit zunehmender Fähigkeit steigen die Lösungswahrscheinlichkeiten der beiden Items, bis sie sich dem Wert 1 nähern. Dabei bleiben die Lösungswahrscheinlichkeiten für das schwierigere Item 2 jeweils niedriger als bei Item 1.

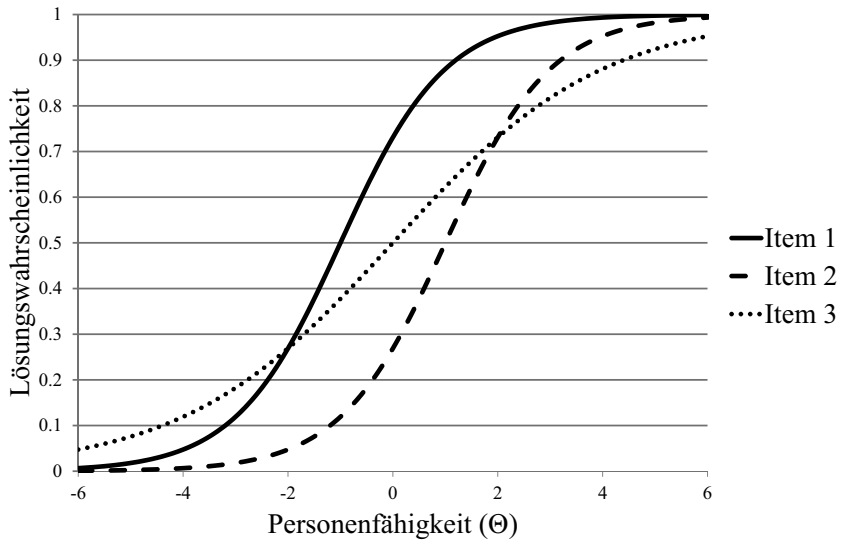


Abbildung 8: Lösungswahrscheinlichkeit in Abhängigkeit von Personenfähigkeit, Itemschwierigkeit und Itemdiskrimination (Logit-Link)

Im zweiparametrischen Modell (kurz 2PL bei Logit-Modellen) wird für jedes Item zusätzlich ein Diskriminationsparameter geschätzt, der sich analog zu einer Faktorladung interpretieren lässt. Zur Identifikation des Modells muss dabei die Metrik der latenten Variable durch Festlegung der Varianz und des Mittelwerts definiert werden – z.B. $\mathrm{Var}(\Theta) = 1$, $\mathrm{M}(\Theta) = 0$. Während die Reihenfolge der Lösungshäufigkeiten von Items bei identischen Diskriminationsparametern entlang der Kontinuums der Personenfähigkeiten konstant bleibt, verändert sich diese, wenn unterschiedliche Diskriminationsparameter vorliegen. In Abbildung 8 ist dies durch Item 3 verdeutlicht, das einen von Item 1 bzw. Item 2 (jeweils $a_1 = a_2 = 1$) abweichenden Diskriminationsparameter aufweist ($a_3 = 0.5$). Hier ändert sich die Reihenfolge der Lösungswahrschein-

lichkeiten abhängig von der Personenfähigkeit: Während im untersten Fähigkeitsbereich ($\Theta < -2$) die Lösungswahrscheinlichkeit für Item 3 die für Item 1 übersteigt, kehrt sich dieses Bild oberhalb dieser Grenze um. Gleiches gilt für den Vergleich von Item 1 und Item 3, wobei hier der kritische Punkt, an dem sich die itemcharakteristischen Kurven schneiden, bei $\Theta = 2$ liegt.

Solche Ladungsunterschiede stellen in einparametrischen Modellen eine Verletzung der Modellannahmen dar, da dort lediglich unterschiedliche Itemschwierigkeiten, aber konstante Diskriminationsparameter bzw. Ladungen angenommen werden. Abweichungen von Items bezüglich ihrer Ladungen können in solchen Modellen durch Itemfit-Statistiken wie die *unweighted-mean-square-* bzw. „Outfit"-Statistik (Wright & Masters, 1982) aufgedeckt werden (Wu & Adams, 2013). Besonders relevant für die Beurteilung des Itemfits ist die sogenannte *weighted-mean-square-* oder „Infit"-Statistik (Wright & Masters, 1982). Dieser Index reagiert weniger sensitiv auf nicht modellkonforme Item-Antworten von Personen, deren Fähigkeit sich stark von der Itemschwierigkeit unterscheidet (also z.B. leistungsschwache Schülerinnen und Schüler bei einem recht anspruchsvollen Item). Bei einer perfekten Modellpassung würden die mean-square-Statistiken sämtlicher Items den Wert 1.0 aufweisen (abgesehen von Abweichungen durch Stichprobenfehler). Abweichungen von diesem Wert verweisen auf Fehlanpassungen, die ggf. durch Itemausschluss reduziert werden können, wobei Werte kleiner 1 im Sinne einer besonders guten Diskrimination eher toleriert werden als Werte größer 1.

Im Falle mehrdimensionaler Modelle lassen sich dann – analog zu faktorenanalytischen Verfahren mit kontinuierlichen Variablen – Modelle mit Einfachladungen, bei denen jeder Indikator ausschließlich Ladungen auf eine latente Variable aufweist (*between-item multidimensionality*; Adams, Wilson & Wang, 1997), von Modellen mit Mehrfachladungen, bei denen Indikatoren auf mehrere latente Variablen laden (*within-item multidimensionality*; Adams et al., 1997) unterscheiden. Aus theoretischer Perspektive kann häufig davon ausgegangen werden, dass bei der Lösung von Testitems mehrere zugrundeliegende (Kompetenz-)Dimensionen beteiligt sind (z.B. Lesekompetenz und mathematische Kompetenz bei einer Textaufgabe in Mathematik). Eindimensionalität zeigt sich in solchen Fällen empirisch nur dann, wenn sämtliche Items in (nahezu) gleicher Weise von den verschiedenen zugrunde liegenden Dimensionen abhängen (Gustafsson & Åberg-Bengtsson, 2010).

Mehrfachladungen lassen sich auch häufig im Rahmen von Testlet-Modellen (Wang & Wilson, 2005) nachweisen, bei denen Items zu einem gemeinsamen Stimulus (z.B. einem Textabschnitt) als abhängig von einer zusätzlichen latenten Variable – neben der zentralen „Zieldimension" – betrachtet werden. Dabei können beispielsweise bestimmte Vorkenntnisse zu einem Themenbereich (z.B. Kenntnis spezifischer Begriffe aus der jeweiligen Epoche) zusätzlich zur Lösungswahrscheinlichkeit eines kompetenzbezogenen Items beitragen.

Allgemeiner spricht man hier auch von *Nested-Factor*-Modellen (Gustafsson & Åberg-Bengtsson, 2010), bei denen Itemlösungen durch mehrere orthogonale (also unkorrelierte) Faktoren vorhergesagt werden. Für das HiTCH-Instrument wurde u.a.

ein solches Modell geschätzt, bei dem die Itemlösungen als abhängig von einem allgemeinen historischen Kompetenzfaktor, zusätzlichen spezifischen Faktoren für die einzelnen Teilkompetenzbereiche (Re-Konstruktion, De-Konstruktion etc.) sowie aufgabenblockspezifischen Faktoren („Testlets", also Items, die jeweils eine gemeinsame Textgrundlage o.ä. aufweisen) betrachtet wurden. Die Berechnung der resultierenden Varianzkomponenten wird in Appendix veranschaulicht.

Bei einem perfekt eindimensionalen Test zur Erfassung historischer Kompetenzen würden sämtliche Items in Abbildung 9 (sofern die Annahme identischer unstandardisierter Ladungen, also identischer Diskriminationsparameter nicht verletzt ist) lediglich Varianzanteile aufgrund des Faktors historische Kompetenz sowie Residualvarianz aufweisen. Sämtliche andere Varianzkomponenten (und somit auch die entsprechenden Faktorvarianzen) wären gleich Null. Abweichungen von diesem psychometrischen „Idealfall" würden sich ergeben, wenn Items jeweils auch Varianzanteile aufweisen, die auf Kompetenzfacetten bzw. auf Testlet-Effekte (aufgabenspezifische Varianzanteile) zurückgehen.

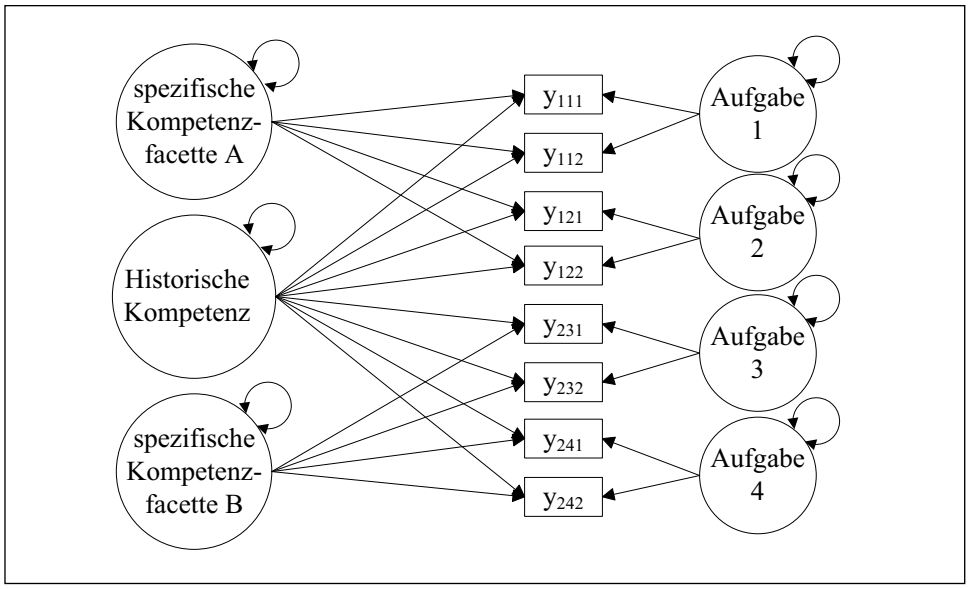


Abbildung 9: Nested-Factor-Modell mit latenten Variablen (Faktoren) für allgemeine sowie spezifische historische Kompetenzen und aufgabenspezifische („Testlet"-) Faktoren (Faktorladungen jeweils fixiert auf $\lambda = 1$).

Neben zusätzlichen latenten Variablen können auch manifeste Variablen (z.B. Geschlecht) itemspezifische Effekte über die latente Variable hinaus aufweisen. So ist es beispielsweise möglich, dass in einem eindimensionalen Modell der angenommene (lineare) Zusammenhang zwischen der Ausprägung der latenten Fähigkeitsvariable über verschiedene Ausprägungen einer beobachtbaren (kategorisierten) Variable variiert, was als *Differential Item Functioning* (DIF) bezeichnet wird. So könnte etwa

ein bestimmtes Item in einem einparametrischen Modell – bei gleichen Ausprägungen auf der latenten Variable (also z.B. bei identischer historischer Kompetenz) – von Mädchen deutlich häufiger gelöst werden als von Jungen. Das würde im Umkehrschluss bedeuten, dass man aus der Ausprägung auf der latenten Variable nicht mehr unmittelbar die Lösungshäufigkeit auf diesem Item ableiten kann. Solche Unterschiede zwischen Gruppen können aus Itemmerkmalen resultieren, die im Hinblick auf das zu erfassende Konstrukt entweder als relevant oder als irrelevant zu betrachten sind. Wird die Lösungswahrscheinlichkeit eines Items durch ein im Hinblick auf das zu erfassende Konstrukt irrelevantes Merkmal beeinflusst (z.B. englischsprachiger Text in einem deutschsprachigen historischen Kompetenztest, den Mädchen besser verstehen als Jungen), dann liegt ein sogenannter *Item-Bias* vor (Zumbo, 1999). Von einem *Impact* hingegen spricht man, wenn die Itemlösung von einem zusätzlichen relevanten Merkmal beeinflusst wird (Zumbo, 1999), also etwa der Verfügbarkeit bestimmter Konzepte (z.B. Revolution).

5 Ergebnisse

In diesem Kapitel werden die zentralen Befunde der psychometrischen Prüfung des HiTCH-Instruments vorgestellt. Wie in Kapitel 4.4 dargestellt, erfolgte diese Prüfung auf der Basis der Item-Response-Theorie (IRT) und reflektiert damit den *state of the art* in modernen Schulleistungsstudien wie PISA, TIMSS und TOSCA. Im Bereich der Forschung zu historischen Kompetenzen ist die Verwendung dieser Verfahren bisher (noch) unüblich, wofür es gute Gründe gibt. So sind für die Beantwortung vieler Forschungsfragen qualitative Ansätze angemessener als quantitative. Die Basis der quantitativ angelegten Analysen bildet wiederum in der Regel die Klassische Testtheorie (KTT), deren Verwendung bei den jeweiligen Projekten gegenstandsangemessen sein kann. Gerade bei der Testentwicklung von standardisierten Kompetenztests bietet die Anwendung der IRT jedoch eine Reihe von spezifischen Vorteilen, auf die im Rahmen der HiTCH-Testentwicklung nicht verzichtet werden sollte (vgl. die Erläuterungen in Kapitel 4). Die Verwendung von IRT-Verfahren bringt jedoch zwangsläufig auch die Verwendung einer Terminologie mit sich, die denjenigen Leserinnen und Lesern, die über keinen Hintergrund in der modernen Testentwicklung verfügen, nicht unmittelbar geläufig sein dürfte.

Die folgende Ergebnisdarstellung, die sich an unterschiedliche Adressatengruppen richtet, folgt zwei Prämissen: Einerseits sind die gängigen Standards moderner Testentwicklung sowie die zugehörige Terminologie (vgl. de Ayala, 2009; Moosbrugger & Kelava, 2012; Rost, 2004) zu verwenden, da dies die Grundlage einer kritischen Rezeption bei der entsprechenden wissenschaftlichen Community darstellt. Andererseits bedarf es für eben diese kritische Rezeption durch die Vertreterinnen und Vertreter von Geschichtsdidaktik, Geschichtswissenschaft und angrenzender Fächer auch einer erklärenden, interpretierenden Ergänzung. Die hierbei resultierende Textform des Ergebnisteils ist gekennzeichnet von einem absichtlichen Nebeneinander und Miteinander von kompakter, methodisch orientierter Darlegung auf der einen Seite und einer inhaltlichen Erläuterung und Deutung, die auch ohne Kenntnisse der verwendeten Methoden und Auswertungsverfahren verständlich sein sollte, auf der anderen Seite.

5.1 Itemauswahl und Reliabilität des HiTCH-Instruments

Die Itemauswahl wurde auf der Basis von inhaltlichen und psychometrischen Überlegungen vorgenommen. Bei allen Items wurden also sowohl die psychometrischen Kennwerte zu Rate gezogen als auch die inhaltliche Bedeutung der jeweiligen Items für einen inhaltlich validen Test berücksichtigt, da dieser die Komplexität des historischen Denkens in hinreichender Breite adressieren soll.

Das in der Haupterhebung verwendete Instrument umfasste ursprünglich 15 Aufgabenblöcke mit insgesamt 106 Items, die teilweise ein sogenanntes Complex Multiple Choice-Format aufweisen (drei Aufgabenblöcke mit insgesamt 19 Items, die jeweils auf drei Subitems basieren, sowie einer Reihenfolgeaufgabe, die auf neun Subitems

basiert). Bei den Complex Multiple Choice-Formaten wurde ein Item als korrekt gelöst gewertet, wenn jeweils alle drei Subitems korrekte Lösungen aufwiesen. Die Itemanzahl der einzelnen Aufgabenblöcke variiert von einem Item (Reihenfolgeaufgabe) bis zu 16 Items. Zunächst wurden alle 106 Items des HiTCH-Instruments mithilfe eines zweiparametrischen logistischen (2PL) Modells skaliert. Hierbei wurde ein eindimensionales Modell spezifiziert, d.h. alle Items wurden als Indikatoren für eine gemeinsame latente Eigenschaft (historische Kompetenz) betrachtet. Mit einer Ausnahme lagen bei der Verwendung dieses eindimensionalen 2PL-Modells alle standardisierten Ladungen im positiven Bereich. Als statistisches Auswahlkriterium für den Einschluss/Ausschluss einzelner Items wurde als untere Grenze des 95%-Konfidenzintervalls der standardisierten Ladung ein Wert von .30 anvisiert. Ein 95%-Konfidenzintervall gibt dabei den Bereich an, in dem der Parameter (also der „wahre" Wert in der Population) mit einer 95%igen Sicherheit liegt. Im vorliegenden Fall kann also mit 97.5-prozentiger Sicherheit davon ausgegangen werden, dass die jeweilige standardisierte Ladung mindestens .30 beträgt, da Werte außerhalb der Grenzen des 95%-Konfidenzintervalls in beiden Richtungen (oberhalb bzw. unterhalb) jeweils nur eine 2.5-prozentige Sicherheit aufweisen. Eine stringente Umsetzung dieses Kriteriums hätte zum Ausschluss von insgesamt 34 Items geführt.[22] 19 dieser Items wurden allerdings aus inhaltlichen bzw. testökonomischen Gründen (z.B. Items, die das Kriterium knapp verfehlten, aber die Bearbeitungsdauer kaum erhöhen) beibehalten. Der finale HiTCH-Test besteht somit aus 91 Items, davon 16 Complex Multiple-Choice-Items aus drei Aufgaben (mit jeweils 3 Subitems) sowie ein Reihenfolgeaufgabe-Item basierend auf 9 Einzelitems. Eine kurze Beschreibung der Aufgabenblöcke ist in Tabelle 1 aufgeführt.

22 Eine Itemselektion auf Basis des *weighted mean-square*-Kriteriums (WMNSQ oder Infit) anhand eines Rasch-Modells in ConQuest (Wu et al., 2007) hätte zu vergleichbaren Ergebnissen geführt, da diese Statistik erwartungsgemäß (Wu & Adams, 2013) einen engen Zusammenhang mit der standardisierten Ladung aus dem 2PL-Modell aufwies (r = -.98).

Tabelle 1: Kurzbeschreibung der Aufgabenblöcke

Ordnungsnummer im Test		Kurzbeschreibung/Aufgabeninhalt	Dimension
1	**10**	Interpretation eines Holzschnittes zum Thema Hexenverfolgungen	RK
2	**11**	Interpretation von zwei Textquellen zum Thema Hexenverfolgungen	RK
3	**12**	Anordnen von Textbausteinen zu logischen Aussagen zum Thema Hexenverfolgungen	OK
4	**13**	Angemessener Umgang mit Vergangenheit (epistemologische Prinzipien)	SK
5	**14**	Anordnen von Ereignissen aus zwei Darstellungen der Geschichte Nordamerikas	RK
6	**15**	Interpretation von zwei Darstellungen zum Thema Geschichte Nordamerikas	DK
7	**16**	Beschreibung von zwei Bildquellen zum Thema Geschichte Nordamerikas	RK
8	**17**	Unterschiedliche Aussagen von zwei Bildquellen zum Thema Geschichte Nordamerikas	DK
9	**18**	Zuordnen von Staatsbegriffen zu typischen Merkmalen	SK
10	**19**	Zuordnen von Staatsbegriffen zu typischen Epochen	SK
11	**20**	Anordnen von Textbausteinen zu sinnvollen Strategien, um heutige Krisen durch historische Überlegungen zu erklären	FK
12	**21**	Einschätzung von historischen Aussagen hinsichtlich ihrer Absicht („Sinnbildungsmuster")	SK
13	**22**	Zuordnung von Fragen hinsichtlich ihres Bezugs auf Vergangenheit / Geschichte / Gegenwart und Zukunft	FK
14	**23**	Interpretation von Historikeraussagen zum Thema Scherbengericht	DK
15	**24**	Orientierungsangebote von Historikern zum Thema Scherbengericht	DK

Abkürzungen: RK = Re-Konstruktionskompetenz, OK = Orientierungskompetenz, SK = Sachkompetenz, DK = De-Konstruktionskompetenz, FK = Fragekompetenz

Vor der Itemauswahl lagen die WMNSQ-Werte (*weighted mean-square*), die die Passung der einzelnen Items im Sinne ihrer im Rasch-Modell als konstant angenommenen Diskriminationsparameter kennzeichnen und möglichst nahe am Wert von 1.0 liegen sollen, im Bereich von 0.84 bis 1.37 bei einer Varianz der latenten Variable von $\mathrm{Var}(\Theta) = 0.65$.[23] Bezogen auf die finale Itemauswahl der 91 Items aus dem HiTCH-Instrument lagen die WMNSQ-Werte im Bereich von 0.84 bis 1.17, also relativ nahe

23 Die Varianz der latenten Variable im Rasch-Modell (mit auf 1 fixierten Diskriminationsparametern) kann im Sinne der Itemdiskrimination interpretiert werden, die mit zunehmender Varianz steigt. Äquivalente Modelle (1PL-Modelle) würde man durch Fixierung der Varianz der latenten Variable (z.B. $\mathrm{Var}(\Theta) = 1$) bei freier Schätzung des (für alle Items konstanten) Diskriminationsparameters erhalten (eine hohe Varianz im „Standardmodell" würde hier also durch hohe Diskriminationsparameter repräsentiert werden). Analog dazu lässt sich die Varianz der latenten Variable auch als standardisierte Ladung der Indikatoren auf die latente Variable ausdrücken. Bei einer Varianz von $\mathrm{Var}(\Theta) = 0.65$ und einer „Residualvarianz" von $\mathrm{Var}(\varepsilon) = \pi^2 : 3$ für die Logistische Verteilung (Logit-Link) ergibt sich eine standardisierte Ladung von $\lambda_{std} = [1^2 \cdot 0.65 : (1^2 \cdot 0.65 + \pi^2 : 3)]^{0.5} = .41$.

am angestrebten Wert von 1.0; in PISA werden Items innerhalb des Bereichs $0.8 \leq$ WMNSQ ≤ 1.2 als modellkonform betrachtet (OECD, 2014). Die Varianz der latenten Variable betrug $\mathrm{Var}(\Theta) = 0.88$, was einer standardisierten Ladung der Indikatoren von $\lambda_{std} = .46$ entspricht.

In Abbildung 10 werden die Verteilungen der Itemschwierigkeiten der einzelnen Items, die mit ihrer Itemnummer aufgeführt sind, und der Schülerleistungen, gegenübergestellt. Auf der rechten Seite der Abbildung sind die Items abgetragen. Je schwieriger ein Item (je geringer also die mittlere Lösungswahrscheinlichkeit), desto weiter oben in der Abbildung ist das Item aufgeführt. Die Itemschwierigkeit bei dichotomen Items ist dabei so definiert, dass bei einer Übereinstimmung von Schülerleistung und Itemschwierigkeit die Lösungswahrscheinlichkeit 50% beträgt. Bei einer Itemschwierigkeit von $b = 1.0$ würde man also für Schülerinnen und Schüler mit einer Ausprägung von $\Theta = 1.0$ auf der latenten Variable erwarten, dass 50% dieser Schülerinnen und Schüler die entsprechende Aufgabe lösen.

Es sei jedoch darauf hingewiesen, dass es sich bei den Itemschwierigkeiten um Schätzungen handelt, die mit einer gewissen Unsicherheit verbunden sind, weshalb die Reihenfolge der Itemschwierigkeiten nicht zwingend exakt der in der Population entsprechen muss. Diese Unsicherheit kann in Form sogenannter Konfidenzintervalle dargestellt werden. Ein solches Intervall lässt sich mithilfe des Stichprobenkennwertes und des (geschätzten) Standardfehlers bestimmen. Im vorliegenden Fall beträgt der maximale Standardfehler (der mit dem größten Konfidenzintervall einhergeht) $SE = 0.04$ bei einer Itemschwierigkeit von $b = 1.53$ (Aufgabenblock 20; Item 3b). Unter der Annahme normalverteilter Stichprobenkennwerte ergibt sich für dieses Item ein 95%-Konfidenzintervall mit einer unteren Grenze von $1.53 - 1.96 \cdot 0.04 = 1.45$ und einer oberen Grenze von $1.53 + 1.96 \cdot 0.04 = 1.61$.

Auf der linken Seite der Abbildung ist die Leistungsverteilung der Schülerinnen und Schüler auf Basis von Punktschätzern – hier wurden sogenannte *weighted likelihood estimates* (WLEs; Warm, 1989) verwendet – dargestellt. Wie Abbildung 10 zeigt, ergab sich bezogen auf die Itemschwierigkeiten eine gute Abdeckung des Leistungsspektrums der Schülerinnen und Schüler in der Stichprobe. Bei den im Mittel relativ leichten Aufgabenblöcken (10, 11, 16, 17; $M(b) \leq -0.47$) wurden als Materialien meist Bilder verwendet (10, 16, 17) und häufig sollten die Materialien als Belege für Aussagen bewertet werden (10, 11, 16). Die im Durchschnitt eher schwierigen Aufgabenblöcke (20, 22, 23, 24; $M(b) \geq 0.57$) wiesen durchweg einen Gegenwartsbezug auf, d.h. es ging um die Reflexion geschichtlicher Ereignisse im Hinblick auf aktuelle Geschehnisse. Die höchste durchschnittliche Aufgabenschwierigkeit (*partial credit*-Item mit drei Ausprägungen, $b = 1.38$) fand sich bei Aufgabenblock 14 (bestehend lediglich aus einem Item basierend auf 9 Subitems). Grund hierfür ist aber vermutlich weniger der inhaltliche Anspruch der Aufgabe (es geht um die Erstellung einer zeitlichen Reihenfolge von Ereignissen auf Basis zweier Texte), sondern die eher „strenge" Kodierung der Aufgabe: Bei insgesamt neun Feldern durfte maximal eine benachbarte Kategorie vertauscht sein (1 Punkt für „Teillösung"; 2 Punkte für vollständig korrekte Reihenfolge). Die Art der Kodierung könnte allerdings auch bei den Aufga-

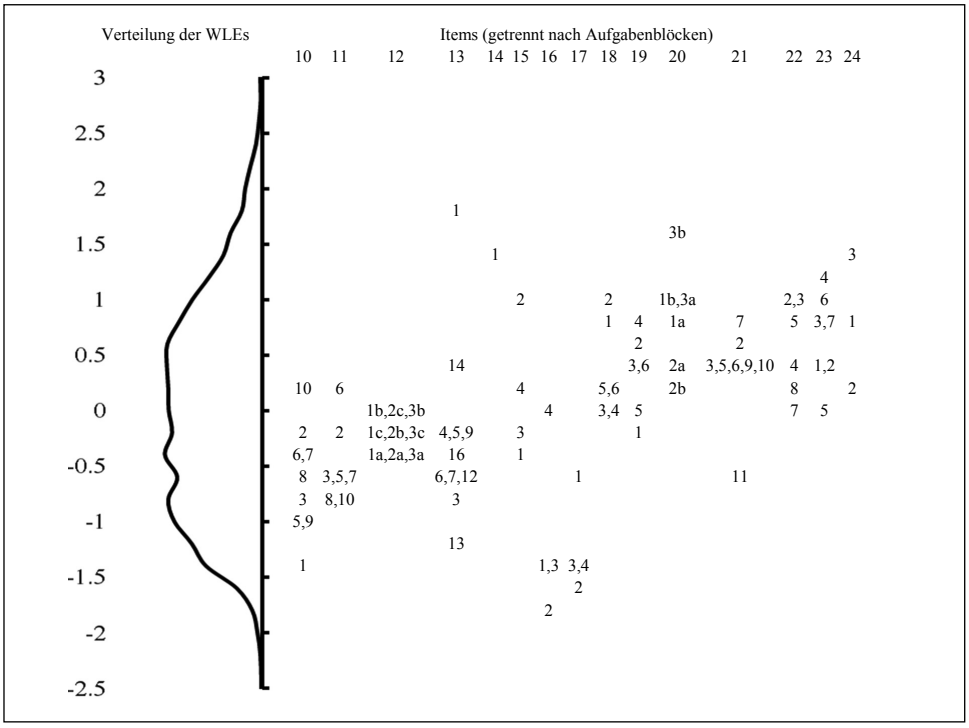


Abbildung 10: Verteilungen der Personenparameterschätzungen (WLEs) und der Itemschwierigkeiten des HiTCH-Instruments (91 Items) aus einem eindimensionalen 1PL-Modell

Anmerkung. Negative Werte stehen für unterdurchschnittliche, positive für überdurchschnittliche Personenfähigkeit. Da die Itemschwierigkeit als der Punkt auf dem Kontinuum der Personenfähigkeit definiert ist, an dem eine 50-prozentige Itemlösungswahrscheinlichkeit besteht, stehen niedrigere Werte für leichtere, höher Werte für schwierigere Items.

benblöcken 22, 23 und 24 eine Rolle gespielt haben. Es handelte sich dabei um sogenannte Complex Multiple-Choice-Aufgaben, bei denen jeweils drei dichotome Einzelitems (Antwortalternativen „ja“ und „nein“) zu einer bestimmten Aussage zu einem Indikator zusammengefasst wurden („richtig“ = alle Einzelitems korrekt, ansonsten „falsch“). Die Wahl des Antwortformats in Kombination mit der Kodierung kann u.U. einen (erheblichen) Einfluss auf die Aufgabenschwierigkeit haben (Kubinger & Gottschall, 2007; Kubinger, Holocher-Ertl, Reif, Hohensinn & Frebort, 2010), die inhaltlich begründete Schwierigkeitsunterschiede relativieren.

Die *WLE person separation*-Reliabilität (WLE PSR; Andrich, 1982) betrug Rel = .91. Die WLE PSR ist definiert als 1 minus den relativen Messfehleranteil, der als durchschnittlicher quadrierter Standardfehler geteilt durch die Gesamtvarianz der WLE-Scores definiert ist. Das Konzept der WLE PSR ist analog zur Klassischen Testtheorie zu verstehen, wobei die Messgenauigkeit in der IRT personenspezifisch ist (und nicht wie in der KTT als konstant angenommen wird; vgl. de Ayala, 2009). Für das HiTCH-Instrument gilt also, dass der durchschnittliche quadrierte Standardfehler der Testscores (der die Messgenauigkeit jedes individuellen Testscores quantifi-

ziert) nur einen Anteil von 9% an der Gesamtvarianz der Testscores hat. In anderen Worten: Die Variabilität der HiTCH-Testscores repräsentiert im Wesentlichen Kompetenzunterschiede und nur einen relativ geringen Anteil an „Messfehler"-Varianz.

Insgesamt liegt mit dem HiTCH-Instrument damit – gemessen an den üblichen Standards – ein reliables Instrument zur Erfassung historischer Kompetenzen vor, dessen Items in wünschenswerter Weise in ihrer Schwierigkeit streuen.

5.2 Überprüfung der Ein- bzw. Mehrdimensionalität

Der HiTCH-Test wies auf den im vorigen Abschnitt berichteten Ergebnissen zufrieden stellende psychometrische Kennwerte auf, was eine im Wesentlichen eindimensionale Struktur des HiTCH-Instruments nahelegt. Ein wichtiger Vorteil eines eindimensionalen Tests liegt darin, dass bei seinem Einsatz ein Gesamtwert für „historische Kompetenz" resultiert, der nicht (oder nur wenig) von der spezifischen inhaltlichen Zusammensetzung des Tests abhängt. Gleichwohl lässt sich prüfen, ob die Annahme einer mehrdimensionalen Struktur nicht zu einer noch besseren Modellgüte führen würde. Dies wurde deshalb in einem nächsten Schritt auf der Basis des eindimensionalen 1PL-Modells untersucht.

Tatsächlich lagen die Interkorrelationen auf Aufgabenebene (repräsentiert durch 15 Aufgabenblöcke mit den insgesamt 91 ausgewählten Items) deutlich unterhalb von $r = 1.0$ (vgl. Tabelle 2), welche beim Vorliegen einer perfekten Eindimensionalität zu erwarten wäre. Als mögliche Ursachen für diese Heterogenität könnten u.a. die Subdimensionen des FUER-Modells (Frage-, Orientierungs- und Sachkompetenz sowie die beiden Methodenkompetenzen Re- und De-Konstruktion, die hier separat behandelt wurden), die unterschiedlich stark in den einzelnen Items und Aufgabenblöcken repräsentiert sind, sowie – unabhängig von der mit dem Item erfassten inhaltlichen Kompetenz – die unterschiedlichen Aufgabenformate und die Themen in Frage kommen.

Das Korrelationsmuster in Tabelle 2 ergab nur sehr geringe Hinweise drauf, dass sich die theoretisch unterscheidbaren Dimensionen in den Daten als eigene Faktoren identifizieren lassen.[24]

24 Da die Aufgaben hier jeweils durch die anhand eines eindimensionalen Modells ausgewählten Items (s.o.) repräsentiert wurden, wäre es auch möglich, dass diese Art der Itemselektion der theoretischen Mehrdimensionalität entgegengewirkt hat. Die 15 ausgeschlossenen Items stammen allerdings aus vier verschiedenen Kompetenzbereichen (Re-Konstruktion: 4 Items; Sachkompetenz: 5 Items; Orientierungskompetenz: 3 Items; Fragekompetenz: 3 Items), so dass diese Annahme eher unplausibel erscheint. Ein entsprechendes Modell auf Basis der gesamten 106 Items ergab eine nahezu identische Interkorrelationsmatrix. Die maximale absolute Differenz zwischen beiden Korrelationsmatrizen betrug $|r_{xy,\text{Itemauswahl}} - r_{xy,\text{ Gesamttest}}| = .05$.

Tabelle 2: Latente Interkorrelationen auf Aufgabenblockebene sowie Varianzen der latenten Variablen (15-dimensionales Rasch-Modell geschätzt mithilfe von ConQuest)

		Re-Konstruktion				De-Konstruktion				Sachkompetenz			Orientierungskompetenz		Fragekompetenz	
	Aufgabenblock	10	11	14	16	15	17	23	24	13	18	19	12	21	20	22
Re-Konstruktionskompetenz	10															
	11	.84														
	14	.65	.69													
	16	.65	.70	.68												
De-Konstruktionskompetenz	15	.73	.75	.72	.75											
	17	.55	.59	.55	.81	.64										
	23	.60	.62	.65	.57	.70	.53									
	24	.58	.59	.58	.53	.63	.50	.87								
Sachkompetenz	13	.80	.81	.75	.74	.79	.65	.70	.63							
	18	.61	.61	.65	.59	.68	.54	.64	.57	.65						
	19	.54	.59	.69	.61	.66	.50	.58	.53	.64	.72					
Orientierungskompetenz	12	.44	.44	.43	.39	.38	.35	.37	.34	.46	.40	.36				
	21	.63	.67	.69	.63	.73	.56	.76	.68	.75	.65	.59	.39			
Fragekompetenz	20	.55	.59	.63	.56	.67	.50	.66	.60	.67	.60	.57	.35	.67		
	22	.58	.61	.68	.64	.69	.51	.77	.63	.72	.63	.58	.36	.75	.62	
	Varianz	1.37	1.79	2.12	2.72	0.96	3.85	1.59	2.84	1.31	2.18	2.50	3.74	0.88	1.82	0.96

Anmerkung. Interkorrelationen innerhalb einer Subdimension sind jeweils umrahmt.

Eine exploratorische Faktorenanalyse auf der Basis der Interkorrelationsmatrix auf Aufgabenblockebene ergab einen Eigenwert für die erste Komponente von 9.6, wohingegen bereits die zweite Komponente mit einem Eigenwert von 0.9 den Wert 1.0 (Kaiser-Kriterium) unterschritt. Dieser Befund spricht dafür, dass auf Basis dieser Daten eine Modellierung mit mehreren Faktoren wenig sinnvoll erscheint. Der erste Faktor (der Generalfaktor „historische Kompetenz") erklärte 64% der Gesamtvarianz. Einiges an Varianz verblieb auf der Ebene der einzelnen (thematischen) Aufgabenblöcke.

Um die Struktur des HiTCH-Tests noch besser zu verstehen, wurden in einem nächsten Schritt zwei sogenannte Nested-Factor-Modelle (vgl. Kapitel 4.4) berechnet. In diesen Nested-Factor-Modellen wurde, vereinfacht ausgedrückt, geprüft, ob man mit dem Generalfaktor, den wir als „historische Kompetenz" bezeichnen, das Muster der richtigen Beantwortung der Items hinreichend gut erklären kann oder ob man zusätzliche Einflussfaktoren identifizieren kann. Wenn – wiederum vereinfacht aus-

gedrückt – bestimmte Items höher miteinander korrelieren als andere, könnte dies ein Hinweis auf solche zusätzlichen Einflüsse sein. In dem ersten Nested-Factor-Modell wurde geprüft, ob die theoretisch angenommenen Subdimensionen einen Teil der Varianz, den der Generalfaktor nicht bindet, aufklären kann; im zweiten Nested-Factor-Modell wurden zusätzlich noch mögliche spezifische Effekte der einzelnen Aufgabenblöcke (aufgabenblockspezifische Varianz) geprüft.

Im ersten Nested-Factor-Modell wurden neben dem Generalfaktor die Subdimensionen des FUER-Modells (Re- und De-Konstruktionskompetenz wurden erneut separat betrachtet) spezifiziert. Die Prüfung des Modells führte zu dem Ergebnis, dass der Generalfaktor erwartungsgemäß die höchste Varianz aufwies (0.90), während die subdimensionsspezifischen Varianzen der latenten Variablen im Bereich von 0.21 bis 0.50 lagen (Re-Konstruktion: 0.29, De-Konstruktion: 0.22, Sachkompetenz: 0.21, Orientierungskompetenz: 0.50, Fragekompetenz: 0.27). Verglichen mit dem eindimensionalen Modell zeigte sich eine bessere Modellpassung für das komplexere Modell (Deviance-Differenz-Test: $\chi^2(5) = 2881.2$, $p < .001$). Dies bedeutet, dass das um Subdimensionen erweiterte Modell dem „einfachen" eindimensionalen Modell aus messtheoretischen Gründen vorzuziehen wäre. Doch aufgrund der geringen Itemanzahl und der niedrigen Varianz der spezifischen Faktoren wäre mit einer niedrigen Reliabilität der Testscores für die Subdimensionen zu rechnen – in der vorliegenden Studie betrug die maximale Reliabilität für *Expected-a-posteriori-Schätzer* (EAP; Bock & Aitkin, 1981) EAP PSR = .48 –, so dass kaum relevante zusätzliche Informationen gewonnen werden können. In anderen Worten: Messtheoretisch passt dieses komplexere Modell besser, aber der inhaltliche Mehrwert ist gering, weil die spezifischen Faktoren (hier also die Dimensionen des FUER-Kompetenzmodells) nicht noch zusätzlich in reliabler Weise gemessen werden.

Im nächsten Schritt wurden in einem zweiten Nested-Factor-Modell neben dem Generalfaktor nicht nur für die Subdimensionen, sondern darüber hinaus auch für die einzelnen Aufgabenblöcke latente Variablen spezifiziert.[25] Dieses zweite Nested-Factor-Modell prüft letzten Endes zusätzlich, wie „ähnlich" das Antwortverhalten der Schülerinnen und Schüler innerhalb eines Aufgabenblocks versus zwischen Aufgabenblöcken ist. Finden sich Belege dafür, dass die Aufgabenblöcke systematisch Varianz binden, so lässt sich dies als Hinweis darauf deuten, dass Inhalte und/oder das Format der jeweiligen Aufgabenblöcke einen relevanten Faktor für die Unterschiedlichkeit der Lösungswahrscheinlichkeit zwischen den Schülerinnen und Schülern darstellten. In diesem zweiten Nested-Factor-Modell zeigte sich (s. Tabelle 3), dass die Varianzkomponenten (quadrierte standardisierte Ladungen), die die Frage-, Re-Konstruktions-, De-Konstruktions-, Sach- und Orientierungskompetenz abbildeten, im Bereich von 0% bis 5% (M = 2%) bezogen auf die Gesamtvarianz lagen. Diese Varianzanteile waren damit weniger bedeutsam als die aufgabenblockspezifischen Va-

25 Aufgrund der hohen Komplexität wurde dieses Modell mit insgesamt 21 latenten Variablen (1 Generalfaktor „historische Kompetenz", 5 subdimensionsspezifische Faktoren und 15 aufgabenspezifische Faktoren) in Mplus mit Bayes-Estimator und dem (bisher) dafür ausschließlich verfügbaren Probit-Link geschätzt.

rianzkomponenten, die im Bereich von 1% bis 42% lagen (M = 15%). Die deutlich größten Varianzanteile ließen sich erwartungsgemäß auf den Generalfaktor „historische Kompetenz" zurückführen, wobei die relativen Anteile zwischen 30% bis 57% (M = 38%) der Gesamtvarianz betrugen. Unter Berücksichtigung aufgabenblockspezifischer Varianzanteile zeigten sich also kaum subdimensionsspezifische Anteile über die allgemeine historische Kompetenz (repräsentiert über den Generalfaktor) hinaus.

Tabelle 3: Quadrierte standardisierte Ladungen („Varianzkomponenten", Angaben in Prozent) eines Nested-Factor-Modells für die Itemauswahl des HiTCH-Instruments (91 Items) mit Generalfaktor und spezifischen Faktoren für Aufgabenblöcke und Subdimensionen (Probit-Link, Faktorladungen jeweils fixiert auf λ = 1)

		quadrierte standardisierte Ladungen (Angaben in Prozent aufgeklärter Varianz[1])			
Subdimension	Aufgabenblock Nr.	Aufgabenspez. Komponente	subdimensionsspez. Komponente	Generalfaktor	Summe
Re-Konstruktionskompetenz	10	8	5	23	35
	11	9	5	22	36
	14	1	5	24	30
	16	17	4	20	42
De-Konstruktionskompetenz	15	4	2	24	30
	17	31	1	17	50
	23	9	2	23	34
	24	20	2	20	42
Sachkompetenz	13	6	1	24	31
	18	19	1	21	41
	19	23	1	20	44
Orientierungskompetenz	12	42	0	15	57
	21	6	0	24	31
Fragekompetenz	20	18	0	21	39
	22	9	0	23	33

[1] Die aufgeklärte Varianz ist hier im Sinne eines Pseudo-R^2 nach McKelvey & Zavoina (1975) für die Itemgruppe des jeweiligen Aufgabenblocks zu betrachten.

Das jetzt vorliegende HiTCH-Instrumentarium mit 91 Items weist dementsprechend eine hinreichend deutlich ausgeprägte eindimensionale Struktur auf. Aufgrund des Vorgehens bei der Konstruktion des Tests, bei der die theoretisch als bedeutsam eingeschätzten Bestandteile von historischer Kompetenz berücksichtigt wurden, kann der Gesamttest beanspruchen, „historische Kompetenz" in wünschenswerter Weise in einem Gesamtwert zu erfassen; gleichzeitig finden sich Hinweise darauf, dass die einzelnen Aufgabenblöcke einen nicht komplett vernachlässigbaren Teil der Varianz binden, so dass vertiefende (fachdidaktische) Analysen zu den Eigenschaften der

Aufgabenblöcke angezeigt sind. Im Hinblick auf die Dimensionalität erbrachten vertiefende Prüfungen Hinweise darauf, dass sich die im FUER-Modell theoretisch definierten Kompetenzbereiche – zumindest mit den bislang entwickelten Items – empirisch nur bedingt trennen lassen, womit die Befunde der Haupterhebung in Einklang mit den Befunden der zweiten Pilotierung stehen (vgl. Kapitel 3.3.4). Geringe empirische Belege für eine konzeptionell plausible mehrdimensionale Struktur sind bei Kompetenztests nicht ungewöhnlich: Ein ähnlicher Befund zeigte sich beispielsweise auch bei dem Test zur Erfassung der Bildungsstandards im Fach Mathematik, in dem die theoretisch trennbaren Kompetenzbereiche in den empirischen Daten eng miteinander korrelierten (Blum et al., 2006).

Die empirisch gefundene eindimensionale Struktur könnte auf unterschiedliche Gründe zurückzuführen sein. Erstens wäre es möglich, dass die Aufgaben des HiTCH-Instruments die verschiedenen Kompetenzbereiche noch nicht hinreichend spezifisch adressieren. Zweitens liefern die in Kapitel 2.3.3 beschriebenen Überlappungsbereiche eine konzeptuelle Erklärung. Die Subdimensionen der Kompetenz historischen Denkens hängen sehr eng miteinander zusammen. Es ist beispielsweise zu vermuten, dass Lernende, die eine differenzierte Einsicht in die epistemologischen Prinzipien von Geschichte haben und über eine hohe Strukturierungs- und Begriffskompetenz (Sachkompetenz) verfügen, auch die prozessualen Operationen besser ausführen können. Drittens kann spekuliert werden, dass ein überzeugend durchgeführter, kompetenzorientierter Geschichtsunterricht alle Kompetenzbereiche adressiert und in ähnlich guter Weise fördert, so dass ein empirischer Nachweis einer etwaig tatsächlich vorhandenen Dimensionalität erschwert werden würde. Hier sind weitere Untersuchungen, ggf. auch unter Nutzung von weiteren Aufgaben/Items sowie homogenere Stichproben, notwendig, um ein noch besseres Verständnis der Dimensionalität zu erhalten.

5.3 Vergleichbarkeit der Itemschwierigkeiten in Subgruppen

Aus psychometrischer Perspektive sollte ein Test idealerweise in verschiedenen Adressatenkreisen (z.B. Mädchen vs. Jungen; Schülerinnen und Schüler unterschiedlicher Schulformen) identische Messeigenschaften aufweisen, damit Unterschiede zwischen diesen Gruppen (z.B. Mittelwertunterschiede nach Geschlecht oder Schulform) sinnvoll interpretiert werden können. Im Falle eines eindimensionalen, einparametrischen Modells bedeutet dies, dass sich die relativen Itemschwierigkeiten zwischen Subgruppen nicht (bzw. möglichst wenig) unterscheiden sollten. In anderen Worten, die Lösungshäufigkeiten für die einzelnen Items sollten bei Personen mit jeweils gleicher Ausprägung auf der latenten Variable (also bei gleicher Fähigkeit) in allen Subgruppen identisch sein. Analysen, die prüfen, ob dies der Fall ist, haben also nicht den Zweck, Mittelwerte zwischen einzelnen Gruppen zu vergleichen, sondern zu prüfen, ob die Beurteilung von Mittelwertunterschieden mit den jeweiligen Items überhaupt gerechtfertigt bzw. sinnvoll ist. Nur wenn sich die Itemschwierigkeiten im

Rahmen eines solchen Modells zwischen Gruppen nicht relevant unterscheiden, ist ein Mittelwertvergleich dieser Gruppen ohne weiteres interpretierbar, da in diesem Fall aus psychometrischer Sicht die Betrachtung von Gruppenunterschieden auf der Itemebene keine zusätzlichen Informationen liefern würde (und somit als redundant betrachtet werden könnte).

Die Vergleichbarkeit der Itemschwierigkeiten bei Schülerinnen und Schülern aus unterschiedlichen Gruppen wird mithilfe sogenannter Differential Item Functioning-Analysen (DIF) ermittelt (vgl. Kapitel 4.4). Dabei werden die Differenzen der Itemschwierigkeiten zwischen Gruppen betrachtet, wobei die Mittelwerte über alle Itemschwierigkeiten innerhalb jeder Gruppe auf einen bestimmten Wert (i.d.R. Null, so auch in der vorliegenden Studie) fixiert werden: Unterschiedliche Lösungshäufigkeiten zwischen den Gruppen werden dabei als Mittelwertdifferenz der latenten Variable modelliert. Unterschiede bezüglich gruppenspezifischer Itemschwierigkeiten lassen sich wie folgt interpretieren: Angenommen, für ein bestimmtes Item *i* ergibt sich eine Schwierigkeit von $b_{\text{i, Mädchen}} = 0$ für Mädchen und $b_{\text{i, Jungen}} = 1$ für Jungen, dann benötigen Jungen jeweils eine um eine Einheit höhere Ausprägung auf der latenten Variable (hier Fähigkeit in Logit-Metrik) um eine identische Lösungswahrscheinlichkeit wie Mädchen auf diesem Item zu erhalten. Liegt die Varianz der latenten Variable bei $\text{Var}(\Theta) = 1$ (für das HiTCH-Instrument mit 91 Items lag die Varianz wie oben bereits erwähnt bei $\text{Var}(\Theta) = 0.88$, d.h. $\text{SD}(\Theta) = 0.94$), dann entspricht eine Einheit einer um eine Standardabweichung höheren Leistung. Eine Differenz der Itemschwierigkeiten von einer Einheit könnte man also aus inhaltlicher Sicht als sehr groß betrachten, wohingegen Differenzen kleiner als 0.4 als wenig bedeutsam beurteilt werden können (Pohl & Carstensen, 2012).

Die DIF-Analysen für das HiTCH-Instrument ergaben vier zentrale Befunde. Erstens fanden wir hinsichtlich der Vergleichbarkeit der Itemschwierigkeiten in verschiedenen Subgruppen zwischen Jungen und Mädchen insgesamt geringfügige Unterschiede ($-0.57 \leq b_{\text{i, Jungen}} - b_{\text{i, Mädchen}} \leq 0.42$), wobei sich für 80 der insgesamt 91 Items absolute Differenzen kleiner als 0.30 ergaben. Zweitens fanden sich etwas größere Unterschiede für die Schulform: Hier lagen die Differenzen der Itemschwierigkeiten im Bereich von $-0.66 \leq b_{\text{i, nicht-gymnasiale Schulform}} - b_{\text{i, Gymnasium}} \leq 0.73$, wobei sich für 70 Items absolute Differenzen kleiner als 0.40 ergaben. Drittens ergaben sich bezüglich des häuslichen Buchbesitzes als einem Indikator für die familiäre Herkunft insgesamt eher niedrige relative Schwierigkeitsunterschiede ($-0.43 \leq b_{\text{i, geringer Buchbestand}} - b_{\text{i, hoher Buchbestand}} \leq 0.54$). Viertens zeigten sich in einem Vergleich der drei Länder, in denen der Test eingesetzt wurde (Deutschland, Österreich, Schweiz), Unterschiede in vergleichbarer Größenordnung wie bei der Schulform: $-0.41 \leq b_{\text{i, Deutschland}} - b_{\text{i, Schweiz}} \leq 0.68$ (davon bei 78 Items absolute Differenzen kleiner als 0.30), $-0.49 \leq b_{\text{i, Deutschland}} - b_{\text{i, Österreich}} \leq 0.56$ (davon bei 78 Items absolute Differenzen kleiner als 0.40); $-0.66 \leq b_{\text{i, Schweiz}} - b_{\text{i, Österreich}} \leq 0.71$ (davon bei 68 Items absolute Differenzen kleiner als 0.40).

Insgesamt weisen die DIF-Analysen (also die Analysen zur Prüfung, ob die Items bei den unterschiedlichen Gruppen von Schülerinnen und Schülern ähnlich „funktionieren") darauf hin, dass das HiTCH-Instrument in den untersuchten Subgruppen

relativ vergleichbare Messeigenschaften aufwies und damit – zumindest in Hinblick auf die vier untersuchten Gruppierungsmerkmale – eine faire Erfassung historischer Kompetenzen ermöglicht.

5.4 Vergleichbarkeit der Itemschwierigkeiten nach Testheftversion

Im Rahmen der Erhebungen wurden zwei verschiedene Testheftversionen eingesetzt, die sich lediglich darin unterschieden, dass die hier untersuchten Aufgabenblöcke in unterschiedlicher Reihenfolge präsentiert wurden (Testheft 1: Aufgabenblöcke 10 bis 24, aufsteigend geordnet; Testheft 2: 22, 23, 24, 21, 20, 18, 19, 14, 15, 16, 17, 13, 10, 11, 12). Aus verschiedenen Studien ist bekannt, dass in solchen Fällen mit sogenannten Itempositionseffekten zu rechnen ist (Hartig & Buchholz, 2012). In Tabelle 4 sind die Differenzen der Aufgabenpositionen in beiden Testheftversionen (Bsp.: Aufgabenblock 10 steht in Version A an Position 1 und in Version B an Position 13, was einer Differenz von 1 – 13 = -12 entspricht) sowie der aufgabenspezifischen Itemschwierigkeitsdifferenzen (Mittelwert der versionsbezogenen Itemschwierigkeitsdifferenzen innerhalb einer Aufgabe) aufgeführt.

Die Ergebnisse legen deskriptiv einen Zusammenhang von mittlerer Itemschwierigkeit und Position im Testheft nahe. Dabei ist zu berücksichtigen, dass zunehmende Itemschwierigkeiten gegen Ende des Testhefts hier nicht darauf zurückgeführt werden können, dass das Testheft nicht komplett bearbeitet wurde (nicht bearbeitete Items wurden nicht als „falsch", sondern im Sinne nicht vorgelegter Items behandelt). Insbesondere bei den Aufgabenblöcken mit großen Unterschieden bezüglich der Position in den beiden Testheftversionen (Aufgaben 10, 11, 12, 22, 23, 24 mit jeweils 12 Positionen Unterschied) fanden sich die größten mittleren Itemschwierigkeitsdifferenzen. Dabei ergaben sich höhere mittlere Itemschwierigkeiten bei den Aufgabenblöcken, die in Testheftversion 2 „nach hinten" verschoben wurden (Aufgaben 10, 11, 12; negative Differenzwerte der aufgabenspezifischen Itemschwierigkeiten in Version 1 minus der in Version 2, d.h. höhere Schwierigkeiten in Version 2). Umgekehrt ergeben sich (deutlich) niedrigere mittlere Itemschwierigkeiten für die in Testheftversion 2 „nach vorne" verschobenen Aufgabenblöcke (22, 23, 24). Dieses Muster findet sich bei allen Aufgabenblöcken – mit Ausnahme von Aufgabe 19, die allerdings auch nur um drei Positionen in Testheftversion 2 „nach hinten" verschoben wurde; zudem ist die Differenz der mittleren Itemschwierigkeiten hier sehr klein.

Bis auf zwei Ausnahmen (Aufgabenblöcke 11 und 23) können die Itemschwierigkeitsunterschiede als geringfügig bis moderat betrachtet werden. Bezogen auf den Gesamttestscore sind diese Unterschiede irrelevant: Für beide Versionen ergab sich jeweils ein Mittelwert für den WLE-Score von $M = 0.03$ in der Stichprobe.

Da die Schwierigkeitsunterschiede auch potentiell einen Einfluss auf die Zusammenhänge zwischen den Aufgabenblöcken haben könnten, wurden die Aufgabeninterkorrelationen zusätzlich testheftspezifisch berechnet und jeweils mit denen der Gesamtstichprobe (beide Testheftversionen) in Tabelle 2 verglichen. Dabei zeigten

sich insgesamt nur geringe Unterschiede bezüglich der geschätzten Korrelationen ($\text{Max}(|r_{xy,\text{ Gesamtstichprobe}} - r_{xy,\text{ Version 1}}|) = .08$, $\text{Max}(|r_{xy,\text{ Gesamtstichprobe}} - r_{xy,\text{ Version 2}}|) = .14$).

Tabelle 4: Itemschwierigkeiten nach Testheftversion (unterschiedliche Aufgabenreihenfolge) basierend auf der Itemauswahl (91 Items)

Aufgabenblock	Differenz der Aufgabenpositionen[a] (Position A – Position B)	mittlere aufgabenspez. Itemschwierigkeitsdifferenzen[b]
10	-12	-0.26
11	-12	-0.47
12	-12	-0.20
13	-8	-0.18
14	-3	-0.23
15	-3	-0.06
16	-3	-0.08
17	-3	-0.12
18	3	0.07
19	3	-0.01
20	6	0.11
21	8	0.36
22	12	0.32
23	12	0.64
24	12	0.32

[a] Aufgabenposition in Testversion A minus Aufgabenposition in Testversion B. Die Aufgabenreihenfolge in Version A entsprach dabei der Reihenfolge in der Tabelle (Aufgabenblock 10 an erster, Aufgabenblock 24 an letzter Position). Die Reihenfolge (Aufgabennummern) in Version B sah wie folgt aus: 22, 23, 24, 21, 20, 18, 19, 14, 15, 16, 17, 13, 10, 11, 12

[b] Für sämtliche 91 Items wurden Differenzen der Itemschwierigkeiten (Version A – Version B) bestimmt. Hier werden jedoch ausschließlich die aufgabenspezifischen Mittelwerte dieser Itemschwierigkeitsdifferenzen berichtet. Negative Werte entsprechen durchschnittlich höheren Itemschwierigkeiten in Version B.

Während die beobachteten Itempositionseffekte beim Einsatz des vollständigen HiTCH-Tests also vermutlich eine zu vernachlässigende Rolle spielen, haben sie gleichwohl eine wichtige Implikation: Sollte man in zukünftigen Studien den Schülerinnen und Schülern nur einen Teil der Aufgaben des HiTCH-Tests vorlegen, so sind deren Leistungen nicht direkt vergleichbar mit Leistungen, die im Gesamttest erzielt werden. Unter der Annahme eines linearen positiven Effekts der Testlänge auf die Itemschwierigkeit würden bei einer gemeinsamen Skalierung die Itemschwierigkeiten von gekürzten Fassungen relativ zu der Langfassung überschätzt werden. Dies hätte zur Folge, dass auch die Leistungen von Subgruppen, denen Kurzfassungen des Tests vorgelegt wurden, überschätzt werden würde.

Als potenzielle Wirkfaktoren, die Itempositionseffekten zugrunde liegen können, kommen sowohl positive (z.B. Lerneffekte durch die Aufgabenbearbeitung) als auch negative (z.B. Ermüdungseffekte bzw. motivationale Aspekte) Faktoren in Betracht (Hartig & Buchholz, 2012). Diese können inter-individuell variieren, so dass der Zu-

sammenhang zwischen Itemposition und Itemschwierigkeit inter-individuell unterschiedlich sein kann. Um dies zu untersuchen, wurde in der vorliegenden Studie – in Anlehnung an die Studie von Hartig und Buchholz (2012) – eine latente Variable für den Itempositionseffekt in Mplus[26] spezifiziert. Für diese latente Variable ergab sich ein statistisch signifikant von Null verschiedener Mittelwert von $M = 0.03$ ($p < .001$) bei einer Streuung von $SD = 0.06$. Der positive Mittelwert ist so zu interpretieren, dass Schülerinnen und Schüler im Durchschnitt – unter Berücksichtigung von „positionsunabhängigen" Itemschwierigkeitsunterschieden – am Ende des Tests mehr Fehler machen als am Anfang. Dabei ist allerdings zu berücksichtigen, dass – unter der Annahme einer normalverteilten latenten Variable bei einer Streuung von $SD = 0.06$ – für einen relativ großen Schüleranteil die Fehlerquote gegen Testende sogar (leicht) abnimmt. (Bei einer Normalverteilung liegen ca. 31 % der Fälle im Bereich unterhalb $M - ½\ SD$, bei dem im vorliegenden Fall der negative Wertebereich beginnt: $0.03 - ½ \cdot 0.06 = 0$).

In einem nächsten Schritt wurde untersucht, ob sich vorhersagen lässt, warum manche Schülerinnen und Schüler gegen Ende des Tests „besser" bzw. „schlechter" werden. Als potenzielle Prädiktoren für diesen individuellen Itempositionseffekt wurden – neben der latenten Variablen für die Ausgangsleistung im HiTCH-Instrument – das Interesse an Geschichte, das Geschlecht, die Zugehörigkeit zur Schulform Gymnasium (Dummy-Kodierung: 1 = Gymnasium, 0 = anderer Bildungsgang), die Skala zur Testmotivation (5 Items; Cronbachs $\alpha = .76$) sowie die Anzahl nicht bearbeiteter Items am Ende des Tests („not reached") und die Anzahl nicht bearbeiteter Items innerhalb des bearbeiteten Testteils (d.h. bis zu der Position, ab der nur noch nicht bearbeitete Items bis zum Ende des Tests folgen) überprüft. Ein Gesamtmodell mit sämtlichen sechs Prädiktoren ergab einen positiven Effekt für die Ausgangsleistung, d.h. die Testleistung zu Beginn der Testbearbeitung ($\beta = 0.57$, $p < .001$) – also einen für leistungsstarke verglichen mit leistungsschwachen Schülerinnen und Schülern höhere Zunahme (bzw. geringere Abnahme) der relativen Itemschwierigkeiten im Laufe der Testbearbeitung (also einen höheren Anstieg der relativen Fehlerquote). Weiterhin fanden sich negative Effekte für das Interesse an Geschichte ($\beta = -0.11$, $p < .001$), die Schulform Gymnasium ($\beta = -0.35$, $p < .001$) sowie die Testmotivation ($\beta = -0.18$, $p < .001$) auf die Itemschwierigkeitsveränderung über die Aufgabenpositionen hinweg. Bei vergleichbaren Ausprägungen auf den jeweils anderen Prädiktoren im Modell ergaben sich also für Schülerinnen und Schüler am Gymnasium, Schülerinnen und Schüler mit höherem Interesse an Geschichte sowie Schülerinnen und Schüler mit höherer Testmotivation jeweils geringere relative Fehlerquoten vom Anfang zum Ende des Tests hin. Ein negativer Effekt fand sich für die Anzahl nicht bearbeiteter Items am Ende des Tests („not reached"; $\beta = -0.07$, $p = .033$), also eine geringere Zunahme der relativen Itemschwierigkeiten im Laufe der Testbearbeitung bei zunehmender Anzahl nicht bearbeiteter Items am Testende. Dies könnte dadurch zu

26 In den Modellen wurde eigentlich der Itempositionseffekt auf die „Itemleichtigkeit" modelliert. Berichtet werden hier aber die umgepolten Effekte, d.h. Itempositionseffekte auf die Itemschwierigkeit.

erklären sein, dass ein Teil der Schülerinnen und Schüler aufgrund der zeitlichen Restriktionen bei der Testung gegen Testende hin zügiger, aber weniger gründlich arbeitete, um möglichst viele Aufgaben beantworten zu können. Für die Anzahl nicht bearbeiteter Items innerhalb des bearbeiteten Testteils sowie für das Geschlecht zeigten sich keine statistisch signifikanten Effekte.

Zusammengefasst lässt sich festhalten, dass die Testlänge jeweils mit relativen Leistungsvor- bzw. -nachteilen für bestimmte Schülergruppen einhergeht. Eine Möglichkeit, solche Effekte zu minimieren, könnte darin bestehen, Testhefte mit relativ geringem Aufgabenumfang einzusetzen, da dadurch der Einfluss von Aufgabenpositionseffekten nur noch eine untergeordnete Rolle spielen sollte.[27] Es sei darauf hingewiesen, dass Itempositionseffekte bei der Erfassung von Kompetenzen nicht spezifisch für das HiTCH-Instrument sind, sondern auch beispielsweise bei den PISA-Instrumenten (etwa Naturwissenschaft in PISA 2006; Hartig & Buchholz, 2012) oder den Standard-Tests zu den Bildungsstandards in Österreich (Mathematik, vierte Klassenstufe; Hohensinn et al., 2008) eine Rolle spielen. Da Itempositionseffekte unter bestimmten Konstellationen einen problematischen Einfluss auf die Befundlage haben können (aber nicht müssen), ist es generell (und somit auch bei der Anwendung des HiTCH-Tests) sinnvoll, die Möglichkeit von Itempositionseffekten bei der Planung des Studiendesigns in Betracht zu ziehen.

5.5 Testleistung und Anzahl nicht bearbeiteter Items

Bei der Skalierung des HiTCH-Instruments wurden Items, die nicht bzw. nicht entsprechend der Vorgaben bearbeitet wurden (z.B. Mehrfachkreuzungen bei Aufgaben, bei denen nur eine vorgegebene Antwortalternative laut Instruktion angekreuzt werden sollte), im Sinne fehlender Schülerantworten betrachtet, d.h. die entsprechenden Items wurden in diesen Fällen nicht als falsch kodiert, sondern so behandelt, als wären sie gar nicht vorgelegt worden. Der Grund hierfür liegt darin, dass es sich bei dem HiTCH-Instrument von seiner Konzeption her um einen sogenannten Power-Test und nicht um einen Speed-Test handelt (Lord & Novick, 1968). Als reine Speed-Tests werden Instrumente bezeichnet, bei denen für (nahezu) alle Schülerinnen und Schüler Itemlösungswahrscheinlichkeiten von 100% angenommen werden können, also alle Schülerinnen und Schüler prinzipiell sämtliche Items lösen können. Bei solchen Tests ist theoretisch nur die Anzahl an bearbeiteten Aufgaben relevant. Bei reinen Power-Tests hingegen sollte den Schülerinnen und Schülern zur Aufgabenbearbeitung so viel Zeit wie nötig zur Verfügung stehen. Hier ist man ausschließlich an den Itemlösungen interessiert (weil eine Bearbeitung aller Aufgaben vorausgesetzt wird). In der Praxis werden jedoch meist „partially speeded tests“ (Lord &

27 Allerdings ließe sich ggf. auch argumentieren, dass sich bestimmte Kompetenzaspekte in wünschenswerter Weise gerade mit zunehmender Länge des Tests besonders stark in der Aufgabenlösung niederschlagen könnten. Eine entsprechende Prüfung bleibt weiterer Forschung vorbehalten.

Novick, 1968, S. 132) eingesetzt, also eine Mischform aus Speed- und Power-Test, da eine zeitliche Limitierung aus ökonomischen Gründen meist unabdingbar ist. Auch bei dem Einsatz des HiTCH-Instruments gab es eine zeitliche Begrenzung der Bearbeitungsdauer. Unter solchen Bedingungen ist eine Behandlung der nicht bearbeiteten Items im Sinne nicht-administrierter Items (also als fehlender Wert anstatt einer inkorrekten Antwort) eine sinnvolle Approximation eines Power-Tests, da sich diese Strategie empirisch als akzeptabel erwiesen hat (Pohl, Gräfe & Rose, 2014; Rose, 2013) und u.a. im Rahmen der Skalierungen der Kompetenztests des Nationalen Bildungspanels eingesetzt wird (Pohl & Carstensen, 2013). Entsprechend sollte die Geschwindigkeit der Bearbeitung nicht bewertet werden. Gleichwohl sind Analysen zu dem Muster von nicht-bearbeiteten Aufgaben im Rahmen einer Testvalidierung sowie zur Klärung von interindividuellen Unterschieden im Bearbeitungsverhalten sinnvoll.

Sie wird beispielsweise oft vermutet, dass leistungsschwache (deutlich) häufiger als leistungsstarke Schülerinnen und Schüler Items „überspringen". Somit sollte sich ein (deutlich) negativer Zusammenhang zwischen der Anzahl nicht bearbeiteter Aufgaben und dem Leistungsscore finden. Für das HiTCH-Instrument (WLE-Score) zeigte sich hier eine schwach negative Korrelation (r = -.09; p = .003). Eine höhere Anzahl an nicht bearbeiteten Aufgaben (im Durchschnitt fanden sich bei M = 15.19 der 106 Items der ungekürzten Testversion, die für die Analysen zu nicht bearbeiteten Items herangezogen werden sollte, keine bzw. keine validen Angaben, SD = 19.80) ging also mit einer geringfügig geringeren Korrektheit der Aufgabenlösungen einher. Man kann sagen, dass sich Speed (gemessen als Anzahl an bearbeiteten Aufgaben im Rahmen der vorgegebenen Testzeit) und Power (über die korrekten und nicht korrekten Lösungen der bearbeiteten Aufgaben erfasste Kompetenz) deutlich unterscheiden lassen.

Interessant ist zudem die Unterscheidung zwischen der Anzahl an fehlenden Werten am Ende des Tests („not reached"; gezählt vom Ende des Tests bis zum ersten validen Wert) und der Anzahl an fehlenden Werten in dem Bereich des Tests, der bearbeitet wurde (also bis zur Position, ab der bis zum Ende des Tests nur noch fehlende Werte vorliegen).[28] Fehlende Werte am Ende des Tests könnten dafür sprechen, dass Schülerinnen und Schüler teilweise langsam, aber sehr gründlich die Aufgaben bearbeitet haben – und deshalb viele der bearbeiteten Aufgaben korrekt gelöst haben, aber den Test aus Zeitgründen nicht bis zum Ende bearbeiten konnten. Eine große Anzahl an fehlenden Werten im „bearbeiteten" Bereich des Tests hingegen dürfte eher auf eine „oberflächlichere" oder strategisch orientierte Auseinandersetzung mit den Aufgaben hinweisen, die beispielsweise in einer geringeren Motivation oder Kompetenz begründet sein könnte. In unserer Testvalidierung war die Korrelation der Anzahl an fehlenden Werten am Ende des Tests mit dem Testscore nicht statistisch signifikant (r = .03; p = .268). Für die Anzahl an fehlenden Werten im „bear-

28 Als Grundlage für die Ermittlung der Anzahl der fehlenden Werte wurde der Gesamttest mit 106 Items verwendet (wobei die Berechnung getrennt nach Testheftvarianten erfolgte).

beiteten“ Bereich des Test zeigte sich hingegen ein statistisch signifikant negativer Zusammenhang mit dem Testscore (r = -.17; p < .001).

Weiterführende Analysen galten zudem noch dem Unterschied im Muster der fehlenden Werte in den beiden Testheftversionen mit unterschiedlicher Reihenfolge der Aufgabenblöcke. Abbildung 11 zeigt den Ausfüllungsgrad der Aufgaben beider Testhefte (wobei für diese Veranschaulichung nur solche Aufgaben als „ausgefüllt“ gewertet wurden, bei denen wenigstens zwei Items bearbeitet wurden). Das linke Diagramm in Abbildung 11 bildet die Bearbeitung der Testhefte in der jeweiligen Aufgabenreihenfolge ab. In dem rechten Diagramm zeigen die übereinander liegenden Werte die gleichen Aufgabenblöcke, die jedoch an unterschiedlicher Stelle in den Testheften bearbeitet werden. Beispielsweise wurde der Aufgabenblock 10 in der ersten Testheftvariante von praktisch allen Schülerinnen und Schüler bearbeitet, in Testheftvariante 2, in denen der Aufgabenblock erst weiter hinten kam, dagegen „nur“ von knapp 90% der Schülerinnen und Schüler.

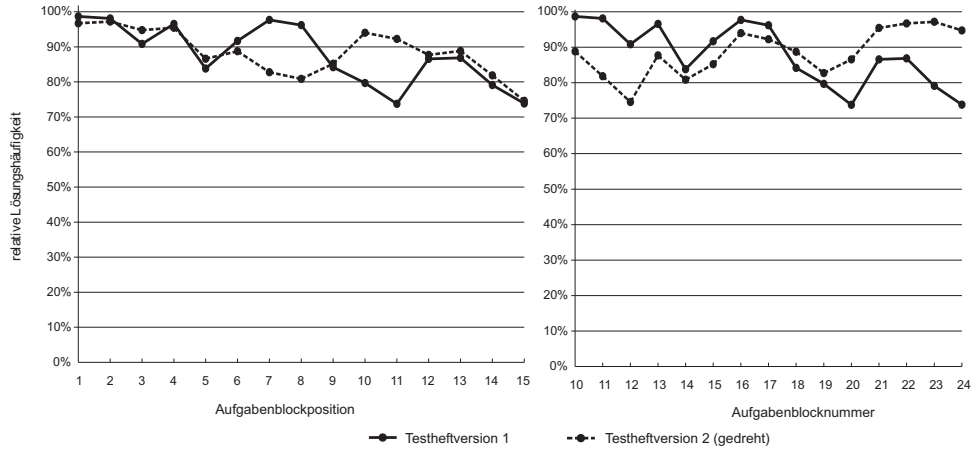


Abbildung 11: Relative Anteile an validen Antworten pro Aufgabe nach Position im Testheft (Diagramm links) bzw. nach Aufgabenblock (Diagramm rechts)

Wie die bereits ausgeführten Analysen verdeutlichen, traten deutlich unterschiedliche Bewältigungsquoten (durchschnittliche Itemschwierigkeiten pro Aufgabenblock; s. Tabelle 3) vor allem dort auf, wo Aufgaben sehr unterschiedliche Positionen in den Testheftvarianten einnahmen, und sie können mit kognitiven und motivationalen Merkmalen vorhergesagt werden. Gleichzeitig sind – und dies mag für die fachdidaktische Forschung besonders interessant sein – die Aufgabenblöcke möglicherweise unterschiedlich anfällig. Einige De-Konstruktions- und Textauswertungs-Aufgaben, aber auch eine Orientierungskompetenz-Aufgabe weisen eher hohe Unterschiede in den Bewältigungsquoten auf. Bestimmte Merkmale von Aufgaben (etwa Leseabhängigkeit, Komplexität der geforderten Operation), vielleicht auch des Themas, könnten Einfluss auf die Verarbeitungstiefe haben sowie unterschiedliches Motivations-

potenzial aufweisen. Hier sind weitere Analysen nötig – insbesondere mit Blick auf die Entwicklung weiterer Aufgaben.[29]

Zusammenfassend lässt sich resümieren, dass eine Kodierung der nicht bearbeiteten Items als inkorrekte Lösungen sowohl aus methodischer Sicht (ein deterministisches Ersetzen ist im Rahmen der IRT, die eine stochastische Beziehung zwischen latenter Variable und Itemlösung annimmt, nicht sinnvoll; Rose, Davier & Xu, 2010) als auch aus inhaltlicher Sicht (geringer Zusammenhang zwischen Anzahl fehlender Werte und Fähigkeit) unangebracht wäre. Die Approximation eines Power-Tests mithilfe der von uns gewählten Skalierungsmethode (nicht bearbeitete Items werden als fehlende Werte behandelt) kann auf Basis der vorgenommenen Kontrollanalysen insgesamt als angemessen betrachtet werden. Die Diskussion über den angemessenen Umgang mit nicht bearbeiteten Testitems kann aber noch nicht als abgeschlossen betrachtet werden, so dass auch andere Skalierungsvarianten prinzipiell in Frage kommen könnten. Da der Anteil an nicht bearbeiteten Items in der vorliegenden Studie nicht besonders hoch ist, sollte der konkrete Umgang mit fehlenden Werten allerdings auch keinen erheblichen Einfluss auf die Ergebnisse haben. Eine Skalierung auf Basis der identischen 91 Items, bei der fehlende Werte durchgängig als Null (= falsch) kodiert wurden, ergab eine Score-Reliabilität von WLE PSR = .94. Dieser WLE-Score korrelierte hoch mit dem hier verwendeten WLE-Score, bei dem fehlende Angaben nicht umkodiert wurden (r = .85).

5.6 Kriteriumsbezogene und diskriminante Validität

Durch die enge Zusammenarbeit von Fachdidaktik und Empirischer Bildungsforschung bei der Entwicklung und Überprüfung der Items konnte eine große inhaltliche Breite (Abdeckung der vier Dimensionen des FUER-Modells) der im finalen Instrument berücksichtigten Items bei insgesamt akzeptablen bis guten Messeigenschaften des HiTCH-Instruments erreicht werden. Von daher ist von einer guten inhaltlichen Validität, bezogen auf das zugrunde gelegte Modell, auszugehen.

Im Sinne der kriteriumsbezogenen Validität wurden zunächst die Zusammenhänge zwischen dem HiTCH-Instrument (WLE-Score) und der Schulform sowie den Schulnoten, insbesondere im Fach Geschichte, untersucht. Sämtliche Noten wurden so kodiert, dass höhere Werte günstigere Leistungen repräsentieren (und somit positive Zusammenhänge mit der Testleistung erwartet wurden). Fälle mit fehlenden Werten auf mindestens einer der genannten Variablen wurden ausgeschlossen, um

29 Differentielle Interpretationen der einzelnen Aufgaben hinsichtlich ihrer schwierigkeitsgenerierenden Merkmale können derzeit noch nicht in allen Fällen plausibel und befriedigend vorgelegt werden. Zwar hat die Forschung zur geschichtsdidaktischen Aufgabenkultur sich dieser Frage bereits angenommen (Waldis, 2013). Die Untersuchung erfolgte allerdings an Lern-, nicht an Leistungsaufgaben. Waldis hat herausgearbeitet, dass neben allgemeinpädagogischen auch psychologische Kriterien (wie etwa der Zahl und das Typenspektrum, gleichzeitig zu verarbeitender Wissenseinheiten) als Teilmaß für den Cognitive Load genommen werden können, aber noch kaum fachspezifische Kriterien vorliegen.

eine möglichst gute Vergleichbarkeit der einzelnen Ergebnisse miteinander zu gewährleisten. Auf Basis dieser Teilstichprobe (N = 2525) wurde eine z-Standardisierung der WLE-Scores (HiTCH-Test) durchgeführt.

Ein Vergleich der einzelnen Schulformen (Haupt- und Realschule wurden aufgrund des geringen Stichprobenumfangs zusammengefasst betrachtet) ergab statistisch signifikante Leistungsunterschiede in erwarteter Richtung im HiTCH-Test (Regressionsmodelle mit der MIXED-Prozedur in SAS unter Berücksichtigung der Schachtelung der Stichprobe; SAS Institute Inc., 2013): Gesamtschule versus Haupt- und Realschule: b = 0.40, p = .029; Gymnasium versus Gesamtschule: b = 0.79; p < .001; Gymnasium versus Haupt- und Realschule: b = 1.16, p < .001. Für die Gruppe der Gymnasialschüler ergab sich also eine um mehr als eine Standardabweichung höhere durchschnittliche Leistung im HiTCH-Test verglichen mit der Gruppe der Haupt- bzw. Realschülerinnen und -schüler. Wichtig ist hierbei nochmals zu betonen, dass keine repräsentativen Stichproben gezogen worden waren und die identifizierten Mittelwertunterschiede zwar als Validitätsbelege herangezogen werden können, aber keine exakte Abbildung von Schulformunterschieden darstellen.

Innerhalb der einzelnen Schulformen zeigten sich darüber hinaus auf Basis von Dreiebenenmodellen in SAS (SAS Institute Inc., 2013; MIXED-Prozedur) teilweise deutliche Variabilitäten auf Schulebene (also Unterschiede zwischen Schulen) sowie auf Klassenebene (Unterschiede zwischen Klassen innerhalb von Schulen). Der größte relative Varianzanteil auf Schulebene (38 %) fand sich in der Teilstichprobe der Gesamtschulen, der geringste bei den Gymnasien (5 %). Dabei ist zu berücksichtigen, dass bei den Analysen verschiedene Schulformen unter dem Oberbegriff „Gesamtschule“ zusammengefasst wurden, Unterschiede zwischen Schulen also auch teilweise Schulformunterschiede repräsentieren können. Die relative Variabilität auf der Klassenebene (innerhalb von Schulen) lag zwischen 0 % (Gesamtschulen) und 7 % (Haupt- und Realschule, Gymnasium).

Bezogen auf die Frage der Zusammenhänge von Schulnoten in den Fächern Geschichte, Deutsch und Mathematik und der mithilfe des HiTCH-Tests gemessenen historischen Kompetenz wurden zunächst einfache Korrelationen auf Basis gruppenzentrierter Daten berechnet. Bei der „Gruppenzentrierung“ wurde von den individuellen Ausprägungen auf den entsprechenden Variablen (Noten und Testleistung) jeweils der Klassenmittelwert subtrahiert, wodurch Unterschiede auf höheren Ebenen (Schulformunterschiede, Klassenunterschiede) ausgeblendet werden. Es ergaben sich moderate Zusammenhänge der Testleistung mit den (im Sinne von Leistung gepolten) Schulnoten (jeweils relativ zum Klassenmittelwert; Geschichte: r = .27; Deutsch: r = .26; Mathematik: r = .19; jeweils p < .001).

Unter Rückgriff auf Mehrebenen-Regressionsmodelle wurden die Zusammenhänge simultan und ebenenspezifisch innerhalb der verschiedenenen Schulformen untersucht. Für sämtliche Schulformen ergaben sich statistisch signifikante positive Zusammenhänge der (im Sinne von Leistung gepolten) Geschichtsnote mit der Testleistung (Haupt- und Realschule: b = 0.13, p < .001; Gesamtschule: b = 0.10, p = .003; Gymnasium: b = 0.15, p < .001). Für die Deutschnote ergaben sich lediglich für die

Gesamtschule sowie für das Gymnasium statistisch signifikante Zusammenhänge mit der Testleistung (Haupt- und Realschule: $b = 0.01$, $p = .833$; Gesamtschule: $b = 0.14$, $p < .001$; Gymnasium: $b = 0.19$, $p < .001$). Die Mathematiknote war in keinem der drei Modelle über die Geschichts- und Deutschnote hinaus prädiktiv für die Testleistung (Haupt- und Realschule: $b = 0.05$, $p = .055$; Gesamtschule: $b = 0.02$, $p = .504$; Gymnasium: $b = 0.04$, $p = .061$). Der aufgeklärte Varianzanteil in den drei Modellen lag zwischen 0 % (Haupt- und Realschule) und 11 % (Gymnasium; Gesamtschule: 8 %).

Zusammengefasst lässt sich konstatieren, dass sich ein – durchaus erwarteter – positiver Zusammenhang zwischen HiTCH-Testleistung und der Note im Fach Geschichte identifizieren ließ. Schülerinnen und Schüler, die im Schuljahr vor der Testung bessere Noten in Geschichte bzw. dem affinsten Fach (Gemeinschaftskunde, Gesellschaft) erhalten haben, erzielen somit im HiTCH-Test tendenziell höhere Leistungen. Dass dieser Zusammenhang nicht noch höher ausfiel und dass der Zusammenhang mit der Deutschnote eine ähnliche Höhe aufwies, ist vermutlich (auch) darauf zurückzuführen, dass erstens die Noten im Fach Geschichte auf nur relativ wenigen Leistungsnachweisen beruhen und deshalb selbst Reliabilitätseinschränkungen aufweisen dürften und dass zweitens typische Leistungsnachweise in Geschichte weniger konsequent kompetenzorientiert gestaltet sind als der HiTCH-Test.

Zur Ermittlung der diskriminanten Validität, d.h. der Prüfung, inwiefern der HiTCH-Test tatsächlich Fähigkeiten historischen Denkens misst und nicht generischere Fähigkeiten wie etwa Lesen oder allgemeine Denkfähigkeit (die bei historischem Denken ja auch eine Rolle spielen), wurden neben dem HiTCH-Instrument mehrere zusätzliche Instrumente eingesetzt. Bei der Überprüfung der diskriminanten Validität durch Korrelationsanalysen zeigten sich auf latenter Ebene relativ deutliche Zusammenhänge des HiTCH-Instruments mit den beiden Lesekompetenztests und dem KFT-verbal-Test ($.83 \leq r \leq .85$). Etwas niedriger fiel der Zusammenhang mit der figuralen Analogie-Facette des KFT ($r = .68$) aus. Ein niedrigerer Zusammenhang fand sich zwischen dem HiTCH-Instrument und der Leistung im Lesegeschwindigkeitstest ($r = .40$). Die Korrelation zwischen dem HiTCH-Instrument und dem Interesse an Geschichte betrug $r = .26$.

Auch auf der Ebene der einzelnen Aufgabenblöcke wurden die Zusammenhänge mit den Außenkriterien (Lesekompetenz, kognitive Fähigkeiten) ermittelt wie auch mit den jeweils verbleibenden Aufgaben des HiTCH-Instruments (Tabelle 5). Mit wenigen Ausnahmen (Aufgabenblöcke 10, 14, 15 und 24) zeigten sich erwartungsgemäß höhere Zusammenhänge der einzelnen Aufgabenblöcke mit den jeweils verbleibenden 14 HiTCH-Aufgabenblöcken der 91-Item-Version ($.47 \leq r \leq .88$) als mit den Außenkriterien (Lesekompetenz-Sachtext: $.37 \leq r \leq .89$; Lesekompetenz literarischer Text: $.43 \leq r \leq .88$; KFT-verbal: $.41 \leq r \leq .83$; KFT-figural: $.34 \leq r \leq .70$).

Tabelle 5: Latente Korrelationen der HiTCH-Aufgabenblöcke mit den jeweils verbleibenden Items (der insgesamt 91 Items) sowie mit eng assoziierten Außenkriterien (pro Aufgabenblock separates 6-dimensionales Modell in ConQuest)

Aufgabenblock	HiTCH-Instrument (verbleibende Aufgabenblöcke)	Lesekompetenz (Sachtext)	Lesekompetenz (lit. Text)	KFT-Verbal	KFT-Figural
10	.79	.79	.67	.71	.61
11	.82	.75	.69	.71	.55
12	.47	.37	.43	.41	.34
13	.88	.83	.79	.83	.66
14	.77	.80	.72	.70	.70
15	.88	.89	.88	.80	.61
16	.77	.72	.61	.70	.61
17	.68	.65	.44	.56	.44
18	.75	.73	.62	.63	.54
19	.71	.69	.64	.57	.47
20	.72	.70	.64	.59	.57
21	.81	.74	.78	.71	.58
22	.78	.75	.73	.66	.64
23	.82	.79	.79	.79	.64
24	.73	.77	.68	.61	.49

Zusammenfassend lässt sich somit konstatieren, dass die vorliegenden Validitätsbelege (Inhaltsvalidität, kriteriumsbezogene und diskriminante Validität) darauf hinweisen, dass das HiTCH-Instrument tatsächlich das misst, was es messen soll, nämlich historische Kompetenz. Bei der Überprüfung der diskriminanten Validität zeigte sich, dass sich die im HiTCH-Instrument erfasste Kompetenz historischen Denkens von der Lesekompetenz wie auch von allgemeinen kognitiven Fähigkeiten trennen ließ. Jedoch erschien der Zusammenhang mit den Tests, die verbale Fähigkeiten adressieren, mit einer Korrelation von bis zu .85 relativ hoch. Um die Bedeutung des hohen Zusammenhangs einschätzen zu können, lohnt sich erneut ein Vergleich mit anderen Schulleistungsstudien. So fanden sich in PISA für die mathematische und sprachliche „literacy" ähnlich hohe Korrelationen von rund .85, die nicht als Hinweis darauf gedeutet werden sollten, dass die PISA-Tests für die unterschiedlichen Domänen in Wirklichkeit nur eine dahinterliegende Fähigkeit messen würden (vgl. Baumert, Brunner, Lüdtke & Trautwein, 2007). Zudem ist ein relativ hoher Zusammenhang zwischen allgemeiner Lesekompetenz und der Kompetenz zu historischem Denken zu erwarten: Die aus Quellen und/oder Darstellungen lesend erschlossenen Informationen sind Bausteine für den re- und de-konstruierenden Umgang mit Vergangenheit bzw. Geschichte. Wer eine hohe Lesekompetenz aufweist, sollte auch Vorteile beim historischen Denken haben. Dabei sollte man aber nicht von einer Einbahnstraße ausgehen: Es ist zu erwarten, dass ein guter, kompetenzorientierter

Geschichtsunterricht auch substanzielle Rückwirkungen auf die allgemeine Lesekompetenz hat. Ohne Zweifel: In der Untersuchung des Zusammenhangs und Zusammenspiels von Lesekompetenz und historischer Kompetenz steckt viel Potenzial für Forschung am Schnittbereich zwischen der Fachdidaktik Geschichte und der Empirischen Bildungsforschung. Erste vertiefende Analysen werden in den beiden kommenden Abschnitten berichtet.

5.7 Allgemeine Lesekompetenz vs. historische Kompetenz: Vertiefende Analysen

In einer vertiefenden Analyse wurde die Frage untersucht, ob die Nähe zur Lesekompetenz für alle HiTCH-Items gleichermaßen gilt oder ob sich unter psychometrischer Perspektive eine Auswahl von Items treffen lässt, bei der die diskriminante Validität des HiTCH-Instruments besonders hoch ausfällt. Die Identifikation entsprechender Aufgabenblöcke bzw. Items könnte dann wiederum Ausgangspunkt weiterer theoretischer oder empirischer Analysen mit dem Ziel sein, die spezifisch „historischen" Anteile einer Aufgabe mit geringen Anforderungen an die Lesekompetenz besser zu verstehen und diese in einer Weiterentwicklung der Aufgabe noch besser zur Geltung zu bringen. Zudem könnte ein Instrument zusammengestellt werden, das nur Aufgabenblöcke enthält, bei denen die divergente Validität zum Lesen besonders hoch ausfällt. Dieser Korpus könnte zur empirischen Überprüfung theoretischer Überlegungen zum historischen Lesen genutzt werden. Allerdings sind mit einer solchen Kürzung eines Instruments häufig zwei große Probleme verbunden. Erstens kann die inhaltliche Validität eingeschränkt werden, beispielsweise indem die Breite des erfassten Konstrukts reduziert wird. Zweitens führt eine Reduktion der Itemzahl in der Regel zu Einbußen bei der Reliabilität. In dieser Hinsicht ist das Ziel eine möglichst hohe Effizienz der Messung historischer Kompetenz, also eine Reduktion des Testumfangs mit möglichst geringen Einbußen bezüglich der Reliabilität der Testscores.

Diesen Fragen wurde in explorierenden Post-hoc-Analysen nachgegangen, in denen die diskriminante Validität gegenüber der Lesekompetenz maximiert werden sollte. Durch eine psychometrisch basierte Aufgabenzusammenstellung sollten solche Aufgabenblöcke ausgewählt werden, die möglichst niedrige Korrelationen mit den – wenn auch nahe verwandten – Außenkriterien (Lesekompetenzen, kognitive Grundfähigkeiten; jeweils zwei Facetten; vgl. Tabelle 5) aufweisen. Als Obergrenze wurde hierfür eine maximale Korrelation von $r_{max} \leq .75$ sowohl mit dem Sachtext als auch mit dem literarischen Text gewählt, was zum Ausschluss von 7 der insgesamt 15 Aufgabenblöcke führte. Von den verbleibenden wurden solche Aufgaben ausgewählt, die (1) eine hohe standardisierte Ladung im Rahmen der 1-Faktor-Lösung der EFA (s. Tabelle 6) und (2) eine hohe Varianz der latenten Variable (s. Tabelle 2) aufwiesen. Konkret wurde der Varianzanteil, der auf den gemeinsamen Faktor zurückgeht, bestimmt (quadrierte standardisierte Ladung multipliziert mit der Varianz der latenten Variable; s. Tabelle 5). Auf dieser Basis wurden fünf Aufgaben des HiTCH-

Instruments für eine im Hinblick auf die diskriminante Validität rein psychometrisch optimierte Testvariante ausgewählt (Aufgaben 11, 16, 17, 18, 19; insgesamt 27 Items).

Tabelle 6: Standardisierte Ladungen der Aufgabenblöcke aus einer unidimensionalen Faktorenanalyse (basierend auf der latenten Interkorrelationsmatrix der Aufgabenblöcke) und erklärte Varianzen

Aufgabenblock	Ladung (stand.)	durch Faktor erklärte Varianz des Aufgabenblocks
17	0.71	1.93
16	0.81	1.77
19	0.74	1.35
18	0.77	1.29
11	0.84	1.28
20	0.75	1.02
12	0.48	0.88
22	0.80	0.62

Die Itemschwierigkeiten eines eindimensionalen Rasch-Modells für diese Aufgabenauswahl lagen im Bereich $-2.04 \leq b \leq .94$ bei einer Varianz der latenten Variable von 1.40 und einer Reliabilität von WLE PSR = .80, was einer akzeptablen Reliabilität entspricht, wenn man den verglichen mit der Originalversion mit 91 Items deutlich geringeren Itemumfang berücksichtigt. Ein zweidimensionales Rasch-Modell in ConQuest mit den Items dieser Aufgabenauswahl auf einer Dimension und den verbleibenden Items der 91-Item-Version des Tests ergab eine relativ hohe, aber nicht perfekte latente Korrelation von $r = .88$. Die Zusammenhänge mit den eng assoziierten Außenkriterien (s. Tabelle 7) waren jeweils (deskriptiv) etwas niedriger verglichen mit denen der Langfassung. Die Zusammenhänge mit der Lesegeschwindigkeit sowie dem Interesse an Geschichte waren nahezu identisch. Interessant und ein Ausgangspunkt vertiefender Analysen ist, dass diese Testvariante lediglich Aufgaben zur Methodenkompetenz (Re- und De-Konstruktion im Hinblick auf Bild- und Textquellen) wie auch zur Sachkompetenz (Begriffskompetenz am Beispiel von Staats- und Herrschaftsformen) umfasst.

Tabelle 7: Latente Korrelationen des HiTCH-Instruments (91 Items) sowie der rein psychometrisch orientierten Aufgabenauswahl (27 Items) mit Außenkriterien

	Gesamttest (91 Items)		Rein psychometrisch orientierte Auswahl (27 Items)	
Kriterium	r^1	*S.E.*	r^1	*S.E.*
Lesekompetenz (Sachtexte)	.85	0.03	.83	0.03
Lesekompetenz (literarische Texte)	.84	0.03	.70	0.04
KFT-V	.83	0.03	.75	0.05
KFT-F	.68	0.04	.60	0.05
Lesegeschwindigkeit	.40	0.06	.40	0.06
Interesse Geschichte	.26	0.02	.28	0.02

Anmerkung. Sämtliche Zusammenhänge sind statistisch signifikant ($p < .001$).

[1]Es handelt sich hier strenggenommen um ein standardisiertes Regressionsgewicht aus einer einfachen Regression. Dieses ist identisch mit dem jeweils analogen Korrelationskoeffizienten. Aufgrund des Matrix-Designs waren jedoch einzelne Korrelationen zwischen den Außenkriterien nicht schätzbar, da die Coverage dort 0 betrug, weshalb keine komplette Interkorrelationsmatrix geschätzt werden konnte. Deshalb wurden in einem Gesamtmodell mit sämtlichen Variablen die Regressionen des HiTCH-Instruments auf die Kriterien spezifiziert. Als Residuen wurden unkorrelierte latente Variablen spezifiziert. Bei einer Coverage größer Null (d.h. bei Variablen, die von einem Teil der Stichprobe gemeinsam bearbeitet wurden), wurden mögliche Zusammenhänge zwischen den Residuen anhand von Regressionen spezifiziert.

Auch wenn diese Aufgabenauswahl des HiTCH-Instruments hinsichtlich der Reliabilität in der vorliegenden Stichprobe nahezu gleichwertig mit dem HiTCH-Instrument selbst ist und sich Hinweise auf Vorteile in Hinblick auf die diskriminante Validität ergeben, ist die inhaltliche Validität des Instruments nur eingeschränkt gegeben, weil diejenigen Aufgaben, die die Frage- und Orientierungskompetenz adressieren, herausfielen. Da zentrale Kompetenzbereiche in dieser Aufgabenzusammenstellung nicht mehr adressiert werden, wird das Konstrukt der historischen Kompetenz aus fachdidaktischer Sicht nicht umfassend repräsentiert (construct underrepresentation, Messick, 1995). Die eingeschränkte Aufgabenauswahl könnte allerdings in solchen Studien durchaus von Interesse sein, in denen z.B. historisches von allgemeinem Lesen getrennt werden soll und historisches Lesen und historische Methodenkompetenz in Bezug gesetzt werden sollen.

5.8 Historische Kompetenz als Prädiktor für die erfolgreiche Nutzung multipler historischer Dokumente

Ein abschließender Baustein im Rahmen der Entwicklung des HiTCH-Tests bestand in der Prüfung, ob die Leistung im HiTCH-Test prädiktiv ist für die Anwendung adäquater Strategien beim Umgang mit multiplen historischen Quellen. Dieser Baustein der Validierung ergänzt die bisher berichteten Schritte in zweierlei Hinsicht. Erstens wurde bei den Schülerinnen und Schülern nicht nur die Leistung im HiTCH-Test erfasst, sondern sie mussten noch eine zweite Testaufgabe meistern; diese Aufgabe unterschied sich von dem HiTCH-Test sowohl im Darbietungsformat als auch in den verlangten Reaktionen. Zweitens erlaubt dieser Schritt der Validierung, die HiTCH-

Leistung direkt in Zusammenhang zu bringen mit dem wesentlich von Wineburg (1991, 2001) geprägten Forschungsstrang, bei dem die methodischen Kompetenzen von *sourcing*, *corroboration* und *contextualization* im Vordergrund stehen (vgl. Kapitel 2.5.3). Im Folgenden werden zentrale Befunde einer Studie von Merkt, Werner und Wagner (in press) vorgestellt.

Ein zentrales Charakteristikum der Arbeit von Historikerinnen und Historikern besteht nach Wineburg darin, dass sie sich nicht auf Informationen aus einzelnen Dokumenten verlassen, sondern ihre gewonnenen Informationen mit weiteren Dokumenten abgleichen (Rouet et al., 1996; Rouet et al., 1997; Wineburg, 1991, 2001). Hierbei sind die in Kapitel 2.5.3 bereits beschriebenen methodischen Kompetenzen corroboration, sourcing und contextualization im Umgang mit verschiedenen Informationen zentral (Wineburg, 1991). Im FUER-Modell finden sich diese drei Operationen im Re-Konstruktionsprozess. Zur Verdeutlichung kann auch die Sechs-Felder-Matrix herangezogen werden (Kapitel 2.3.1.1), die die Re- und De-Konstruktion operationalisiert: Im ersten Feld (Re-Konstruktion mit der Fokussierung auf Vergangenheit) werden die Quellenkritik und der Quellenvergleich verortet. Sourcing und corroboration können als Teiloperationen im Rahmen der Quellenkritik respektive des Quellenvergleichs verstanden werden. Im Rahmen der Fokussierung auf die Geschichte werden die – unter anderem auf der Basis der Quellenanalyse und des -vergleichs gewonnenen – „Vergangenheitspartikel" in einen synchronen und diachronen Zusammenhang gestellt (also in einen historischen Kontext gestellt, somit „kontextualisiert"). Während die Operationen des sourcing und der corroboration nicht nur von historisch Denkenden ausgeführt werden, sondern für die erfolgreiche Nutzung multipler Dokumente generell notwendig sind, stellt die contextualization eine Operation dar, die nicht ohne spezifisch historische Kompetenzen bzw. Kontextwissen auskommt. Insgesamt lässt sich annehmen, dass historische Kompetenz, wie sie durch das HiTCH-Instrument gemessen werden soll, eine wesentliche Voraussetzung für die erfolgreiche Nutzung multipler Dokumente im Fach Geschichte darstellt.

In einer Experimentalstudie mit 111 Schülerinnen und Schülern der 9. Klasse wurden daher die mithilfe des HiTCH-Instruments erfassten Kompetenzen in Beziehung gesetzt zu Indikatoren für die erfolgreiche Bearbeitung multipler Dokumente, während weitere notwendige, aber nicht als hinreichend angenommene Faktoren wie basale Lesefertigkeiten (SLS5-8; Auer, Gruber, Mayringer & Wimmer, 2005) sowie rein behaviorale Indikatoren für die Anwendung der Strategien sourcing und corroboration statistisch kontrolliert wurden. Konkret wurde den Schülerinnen und Schülern zunächst das HiTCH-Instrument vorgelegt (91 Item-Version; WLE PSR = .86), bevor sie im Rahmen einer computer-basierten Lernumgebung zwei Fragen zu den im Jahre 1968 in der Bundesrepublik Deutschland verabschiedeten Notstandsgesetzen beantworteten. Als Grundlage für diese Aufgabe wurden den Schülerinnen und Schülern unter anderem Wochenschauen, Zeitungsberichte und Zeitungsinterviews aus der Deutschen Demokratischen Republik (DDR) und der Bundesrepublik Deutschland (BRD) vorgelegt, die verschiedene Perspektiven auf das dargestellte Ereignis einnahmen. Die Lernumgebung beinhaltete neben den Materialien selbst auch

Informationen über die Materialien (z.B. Autorenschaft, Erscheinungsjahr), die für die Schülerinnen und Schüler optional abrufbar waren. Zudem konnten mittels einer Tagging-Funktion aus einer Liste von vorgegebenen Begriffen beliebige Auszüge aus den Lernmaterialien mit Schlüsselbegriffen versehen werden.

In Regressionsanalysen wurden verschiedene Indikatoren für die erfolgreiche Bearbeitung dieser Lernumgebung als abhängige Variablen mit dem HiTCH-Score, mit basalen Lesefertigkeiten (SLS5-8) und mit behavioralen Indikatoren für die Nutzung der Strategien sourcing und corroboration in Beziehung gesetzt. Die behavioralen Indikatoren für die Strategien sourcing und corroboration wurden mittels der Logdateien aus der Lernumgebung bestimmt. So wurde der Aufruf von Informationen über die Materialien als behavioraler Indikator für sourcing gewertet, während die Nutzung der Tagging-Funktion zur Vergabe eines übereinstimmenden Schlüsselbegriffs in mindestens zwei unterschiedlichen Dokumenten als behavioraler Indikator für corroboration gewertet wurde.

Bei der ersten Aufgabe ging es darum, Unterschiede und Gemeinsamkeiten hinsichtlich der in den BRD- bzw. DDR-Quellen formulierten Zustimmung bzw. Ablehnung gegenüber den Notstandsgesetzen zu nennen. Somit erforderte die Aufgabe zwar historisches Lesen und die kategorisierende Integration von Informationen aus verschiedenen historischen Dokumenten, prozedurale Kompetenzen waren jedoch nicht erforderlich. Trotzdem erwies sich der HiTCH-Score hier in linearen Regressionsanalysen sowohl für die Anzahl der genannten Gemeinsamkeiten als auch für die Anzahl der genannten Unterschiede prädiktiv (wobei das Gesamtmodell bezüglich der genannten Unterschiede nicht statistisch signifikant war, weshalb dieses Ergebnis mit Vorsicht zu interpretieren ist).

Zusätzlich wurde für diese erste Aufgabe ausgezählt, inwiefern sich die Schülerinnen und Schüler in ihren Essays in einer Art und Weise auf eine Quelle bezogen, die eine eindeutige Identifikation der verwendeten Quelle ermöglichte (spezifische Referenzierung, z.B. Material 8 oder Zeitungsbericht aus der DDR). Die Anzahl der Aufrufe von Informationen über die Materialien stand in einem positiven Zusammenhang mit der Anzahl spezifischer Referenzierungen. Der HiTCH-Score erwies sich für die Anzahl spezifischer Referenzierungen hingegen als nicht prädiktiv. Dies entsprach den Erwartungen, da eine genaue Referenzierung wie auch die Prüfung von Texten über verschiedene Domänen hinweg eine wesentliche Komponente des (wissenschaftlichen) Umgangs mit multiplen Dokumenten darstellt und daher nicht zwingend mit historischen Kompetenzen einhergehen muss. Die Beachtung von Quellinformationen und die konkrete Referenzierung von Materialien dürfte domänenübergreifend relevant zu sein und Ausdruck einer allgemeinen Lesefähigkeit.

In der zweiten Aufgabe sollten die Schülerinnen und Schüler mögliche Ursachen für die in Aufgabe 1 genannten Gemeinsamkeiten und Unterschiede zwischen den verschiedenen Dokumenten nennen. Bei der Auswertung wurde analysiert, ob die Schülerinnen und Schüler die Herkunft der Materialien aus der DDR und der BRD als Hauptursache für die in den Dokumenten vermittelte Aussage identifizierten. So erforderte die zweite Aufgabe nicht nur die Integration von Informationen

aus verschiedenen Materialien, sondern zugleich die Fähigkeit, Hintergrundinformationen zu den Materialien vor dem Hintergrund des historischen Kontextes korrekt auf die Beantwortung einer Fragestellung anzuwenden (contextualization). Ähnlich anspruchsvoll war auch die dritte Aufgabe, bei der die west- oder ostdeutsche Herkunft wörtlicher Textzitate bestimmt werden sollte. Hierfür mussten die Schülerinnen und Schüler grundlegende, von der Herkunft der Materialien abhängige Argumentationsstile extrahieren (Subtext; Wineburg, 1991) und auf die Klassifikation der Textzitate anwenden.

Aus der Beschreibung dieser beiden Aufgaben wird ersichtlich, dass hier eine tiefer gehende Auseinandersetzung mit dem Quellenmaterial gefordert war. Diese ging über die bloße Identifikation von Informationen hinaus und erforderte eine Strukturierung der gegebenen Informationen in historisch akkurate Kategorien, um darauf aufbauend Schlüsse für die Beantwortung der jeweiligen Fragestellung zu ziehen. Sowohl bei der zweiten als auch bei der dritten Aufgabe erwies sich das HiTCH-Instrument (auch nach statistischer Kontrolle weiterer Prädiktoren wie basale Lesefertigkeiten) erwartungsgemäß als prädiktiv. Der Prozess der contextualisation, der die Einordnung der präsentierten Informationen in einen sinnvollen historischen Zusammenhang (= Re-Konstruktion) verlangt, wird somit durch die Leistung im HiTCH-Test vorhergesagt. Dieser Befund lässt sich als weiteren Beleg für die Validität des HiTCH-Tests in dem Sinne anführen, dass er sich als prädiktiv ist für historisch kompetentes Handeln erwies, das in einer anderen Modalität (im Rahmen einer multimedialen Assessment-Situation) erfasst wurde.

6 Diskussion und Ausblick

6.1 Das HiTCH-Projekt: Ein kurzes Fazit

Kompetenzen historischen Denkens sind von individueller wie gesellschaftlicher Bedeutung. Ihr Potenzial bei der Orientierung in einer modernen Welt ist so bedeutsam wie das der derzeit hoch im Kurs stehenden MINT-Fächer (Mathematik, Informatik, Naturwissenschaften, Technik), wenn auch anders gelagert: Historische Kompetenzen sind die Grundlage dafür, zentrale gesellschaftliche Herausforderungen der Moderne zu meistern. Sie ermöglichen es z.B., die historischen Dimensionen in gegenwärtigen Entwicklungen (z.B. Flüchtlingsproblematik, Syrienkrieg, Krieg in der Ukraine, Re-Nationalisierung in Europa, Wachstumsabschwächung in China, ...) zu erkennen und zur Fundierung von Entscheidungen für die Zukunft zu nutzen. Sie erlauben es, triftige von untriftigen Vergangenheitsbezügen und Sinnbildungen zu unterscheiden, und schaffen so die Grundlage für einen regulierten Diskurs über Geltungsbehauptungen in den Medien, in Politik und Gesellschaft und in anderen Feldern der Lebenswelt.

Zudem erschließen historische Kompetenzen einen Zugang zur Welt und ihren Menschen, wie er prototypisch für die Kulturwissenschaften ist. Geschichte ist nicht nur ein „Fach", sondern ein methodischer Zugang zur Welt. Historisch gedacht und argumentiert wird nicht nur in der Forschungsdisziplin und im Schulfach Geschichte, sondern unter anderem auch in den Sprachen und Literaturen, in Politik, Soziologie und Wirtschaft, in Geographie, Recht und Kunst. Systematisch erklärt und eingeübt wird historisches Denken vor allem in „Geschichte", seine Bedeutung reicht aber über das Fach Geschichte in der Schule hinaus.

Die Entwicklung eines standardisierten Tests zur Erfassung historischer Kompetenzen hat in dieser Situation viel Potenzial. (1) Sie schärft die fachdidaktische Diskussion darüber, wie Kompetenzen im Fach Geschichte (und in anderen Kulturwisssenschaften) erfasst und gemessen werden können. (2) Ein standardisierter Test kann in vielfältigen Zusammenhängen genutzt werden, beispielsweise für den Nachweis, dass kompetenzorientierter Unterricht im Fach Geschichte tatsächlich lernwirksam ist. (3) Darüber hinaus kann ein entsprechender Test, der auch in Schulleistungsstudien eingesetzt wird, die notwendige Diskussion über die Ziele und die Bedeutung des Faches Geschichte vorantreiben.

Das zentrale Ziel des HiTCH-Projekts war deshalb die Entwicklung eines Instruments, mit dem historisches Denken in psychometrisch „sauber" konstruierten standardisierten Aufgaben erfasst werden kann. Das nach einer Entwicklungszeit von rund drei Jahren mit zwei Pilotstudien und einer Hauptstudie vorliegende HiTCH-Instrument stellt ein standardisiertes, reliables und valides Instrument zur Erfassung historischer Kompetenz dar, das – wie es das erklärte Ziel des HiTCH-Projekts war – in Schulleistungsstudien eingesetzt werden kann. Da die richtige Beantwortung der Testaufgaben eine vorherige Beschäftigung mit den behandelten Gegenständen im

Unterricht nicht voraussetzt und die Testaufgaben auch im unteren Leistungsbereich ausreichend differenzieren, kann in Zukunft in schulform- und (bundes-)länderübergreifenden Schulleistungsstudien auch das Fach Geschichte adressiert werden.

Der vorliegende HiTCH-Test beruht auf einem narrativistisch-konstruktivistischen Verständnis des domänenspezifischen Prozesses historischen Denkens und erfasst „generische" Aspekte historischer Kompetenzen. Damit ist dieses Instrument nicht nur bezogen auf das zu Grunde gelegte FUER-Modell relevant, sondern auch auf alle anderen national und international in der Geschichtsdidaktik diskutierten Kompetenzmodelle.

Im Folgenden werden die dargestellten Befunde und Implikationen der Arbeit an dem HiTCH-Test für die Weiterentwicklung der geschichtsdidaktischen Forschung und die schulische Praxis diskutiert.

6.2 Was misst der HiTCH-Test – und was misst er nicht?

Die psychometrischen Kennwerte weisen aus, dass das HiTCH-Instrument objektiv, reliabel und hinreichend valide ein Konstrukt erfasst, welches auf Grundlage einer dahinter stehenden Theorie als „Kompetenz historischen Denkens" bezeichnet werden kann. Die theoretischen Grundlagen (Kapitel 2) und Methodologie (Kapitel 4 und 5) sind breit dargelegt worden.

Gleichwohl wird der Test fachlich bzw. fachdidaktisch interessierten Leserinnen und Lesern möglicherweise noch immer wie eine „Black Box" vorkommen, insbesondere auch deshalb, weil die bislang entwickelten Testaufgaben der Geheimhaltung unterliegen und nur vor Ort eingesehen werden können. Die Beispiele der nachgebauten Aufgaben ermöglichen einen ersten Einblick. Dennoch soll die Aussage, es handle sich bei dem gemessenen Konstrukt um „Kompetenzen historischen Denkens", im Folgenden durch weitere Erörterungen vertieft werden, die für die Einschätzung des Tests und seiner späteren Einsetzbarkeit von Bedeutung sein könnten.[30] Fünf Aspekte sollen noch einmal betrachtet werden. Erstens (Kap. 6.2.1) wird detaillierter darauf eingegangen, inwieweit es gelungen ist, die theoretisch definierten und pragmatisch durchaus operationalisierbaren Kompetenzbereiche durch Aufgaben auch im Test zu repräsentieren, zweitens (Kap. 6.2.2) wird der Stellenwert historischen Wissens bei der Testung fachspezifischer Kompetenzen noch einmal beleuchtet, drittens (Kap. 6.2.3) wird der Frage der gemessenen Kompetenzniveaus nachgegangen; viertens (Kap. 6.2.4) wird geklärt, inwieweit der HiTCH-Test individuelle Kompetenzausprägungen misst; und fünftens (Kap. 6.2.5) wird erläutert, warum der HiTCH-Test keinen Ersatz für Klassenarbeiten darstellen kann.

30 Für eine noch umfassendere Auslotung der Charakteristika des Tests, insbesondere aus fachlich-fachdidaktischer Sicht, sei auf eine weitere Publikation verwiesen (Meyer-Hamme & Körber, in Vorb.).

6.2.1 Abdeckung des zugrundeliegenden Kompetenz-Konstrukts im HiTCH-Test

Der theoretische Bezug auf das FUER-Modell historischen Denkens ist allein nicht hinreichend für die Behauptung, es handle sich bei dem Konstrukt, das mit Hilfe des HiTCH-Tests reliabel gemessen wird, tatsächlich um historisches Denken. Entscheidend ist vielmehr, dass sowohl bei der Konstruktion der Aufgaben als auch bei der Auswahl der Items neben den Kriterien der psychometrischen Qualität immer auch die Kriterien geschichtsdidaktischer und geschichtstheoretischer Qualität beachtet wurden, dass der Test also aus beiden Sichten die „Kompetenz historischen Denkens" misst.

Der HiTCH-Test hat somit gewissermaßen eine doppelte Qualitätskontrolle durchlaufen, indem einerseits nur solche Aufgabenblöcke und Items in der finalen Testversion verblieben sind, die aus psychometrischer Sicht nennenswert und stabil zum Gesamtkonstrukt beitragen, andererseits aber der Test aus geschichtsdidaktischer Sicht daraufhin begutachtet wurde, dass Items zu allen Kompetenzbereichen vertreten sind, so dass keine Einseitigkeit entsteht. Dies ist keineswegs ein Automatismus.

Von Beginn der Testentwicklung an wurde darauf geachtet, die Aufgaben gleichmäßig auf die Kompetenzbereiche zu verteilen. Im Zuge der Arbeit am HiTCH-Test wurden von Seiten der Geschichtsdidaktik die Kriterien für die Berücksichtigung dieser Bereiche geschärft. Dazu dienten die mehrstufigen Verfahren von Cognitive Labs, kollegialen Diskussionen und Expertenratings, aber auch der interdisziplinäre Diskurs mit der Empirischen Bildungsforschung.

So galt es bei den Testoptimierungen aufgrund psychometrischer Kennwerte immer darauf zu achten, dass es sich auch aus geschichtsdidaktischer Sicht um eine Optimierung des Tests handelte. Auf diese Weise wurde vermieden, dass z.B. nur die mit bereits bekannten Itemstrukturen zu messenden Ausschnitte des Konstrukts historische Kompetenz Berücksichtigung fanden, was zu einer construct underrepresentation geführt hätte. Dazu mussten Aufgaben mehrfach optimiert, auch neu konstruiert und immer wieder durch Schülerinnen und Schülern getestet werden. Der HiTCH-Test enthält in der Folge Aufgaben zu allen vier Kompetenzbereichen, auch zu den besonders schwer operationalisierbaren Bereichen der Orientierungs- und Fragekompetenz.

Die Items aller Kompetenzbereiche haben sich schließlich als psychometrisch akzeptabel und mit einem Globalfaktor verträglich erwiesen. Dies ist als Hinweis auf eine recht hohe „innere Konsistenz" des zugrundeliegenden Konzepts wie der Operationalisierung der „Kompetenzen historischen Denkens" zu interpretieren. Nur durch die konsequente und geduldige interdisziplinäre Zusammenarbeit von Empirischer Bildungsforschung und Geschichtsdidaktik ist eine solche Qualitätsaussage möglich geworden.

Trotz aller Bemühungen sind im Test die verschiedenen Kompetenzbereiche nach FUER aber nicht ganz gleichgewichtig abbildet. Operationalisierungen historischen

Fragens bzw. des Umgangs mit vorliegenden historischen Fragen, die Verfügung über Konzepte und Kategorien, über Verfahrensschritte der historischen Re- und De-Konstruktion sind etwas breiter vertreten, als etwa solche Aufgaben, deren erfolgreiche Bearbeitung Komponenten des historischen Orientierens erfordern. Diese leichte Schieflage dürfte der Thematisierungs- und Aufgabenkultur im Geschichtsunterricht und der Geschichtsdidaktik entsprechen. Sie spiegelt aber auch die besondere Herausforderung bei der Generierung von Aufgaben wider. Die bislang etwas unterrepräsentierten Bereiche zu stärken ist somit eine Aufgabe für die weitere Testentwicklung.

Obwohl der HiTCH-Test das Gesamtmaß „Kompetenz historischen Denkens" durch die Erfassung der beteiligten Kompetenzbereiche in einiger Breite misst, kann nicht in Umkehr geschlossen werden, dass jede mögliche Denkleistung – etwa der Orientierungskompetenzen – auch tatsächlich bzw. in wünschenswerter Breite erfasst ist. Wohl unter anderem auch darauf ist zurückzuführen, dass es noch nicht gelungen ist, innerhalb des Gesamtmaßes historischer Kompetenz stabil Subdimensionen zu differenzieren, d.h. Maße für einzelne Komponenten zu entwickeln, die sowohl voneinander unterscheidbar sind als auch insgesamt so hoch untereinander zusammenhängen, dass sie als Komponenten des Gesamtkonstrukts „historisches Denken" angesehen werden können. Das Ergebnis des HiTCH-Projekts ist somit die erfolgreiche Konstruktion eines Large-Scale-fähigen Tests der Kompetenz historischen Denkens im Ganzen, nicht aber die empirische „Bestätigung" der unterschiedlichen Kompetenzbereiche, die z.B. das FUER-Modell ausweist.

Dass statistisch-psychometrische und fachdidaktische Urteile nicht immer automatisch zusammen fallen, soll abschließend verdeutlicht werden. Auf der Basis der statistischen Kennzahlen wäre es – wie in Kapitel 5.7 dargestellt – beispielsweise durchaus möglich gewesen, den Test in Hinblick auf die diskriminante Validität und die Untersuchungsökonomie zu optimieren, auch zu verkürzen. Dies wäre aber auf Kosten der Repräsentanz aller Kompetenzbereiche erfolgt.

6.2.2 HiTCH-Test – zu herausfordernd für die Schülerinnen und Schüler?

Die in Kapitel 3 vorgestellten Beispielaufgaben haben die Funktionsweise der Messung dahingehend transparent gemacht, dass die Art und Weise der Operationalisierung historischen Denkens – wenn auch nicht in der gesamten vom HiTCH-Test erfassten Breite, so aber doch exemplarisch – verdeutlicht wurde. Die Aufgabenbeispiele könnten aber auch den Eindruck erweckt haben, dass der aus solchen Aufgaben zusammengesetzte Test für Schülerinnen und Schüler der 9. Klasse, zumindest jenseits des Gymnasiums, eine zu große Zumutung darstellt hinsichtlich Ausdauer und kognitiver Belastung. In diesem Abschnitt sollen derartige Anfragen aufgegriffen werden.

Sowohl vom Umfang als auch von der Konstruktion einiger Aufgaben her ist der HiTCH-Test durchaus anspruchsvoll. Er ist dies aber nicht in einer Art, die als unfair gelten müsste oder seine Aussagekraft grundsätzlich in Frage stellen würde. Wür-

de der Test für alle Schülerinnen und Schüler gleichermaßen „leicht“ zu bearbeiten sein, könnte er nicht zwischen den Schülerinnen und Schülern, die über die Kompetenzen in einem höheren Maß verfügen und denen mit eher basalen Kompetenzen unterscheiden. Was allerdings eine Einschränkung der Aussagekraft des Tests bedeuten könnte, wären Totalausfälle in einigen „Zellen“ der Stichprobe (Schulformen, Geschlecht, Bundesland oder Staat) oder stark ungleiche Ausfallquoten zwischen den Zellen. Dies ist aber nicht der Fall.

Weder die allgemeine Höhe der Datenausfälle noch deren gesonderte Betrachtung für die beiden Testheftvarianten des HiTCH-Tests, in denen die Aufgaben jeweils blockweise unterschiedlich gereiht waren (vgl. Kapitel 5.5), ergibt Abbrüche in untragbarem Maße. Damit können sowohl die Gesamtzahl der Aufgaben und somit die Konzentrations- oder Ausdauer-Gesamtbelastung wie auch die besonderen Anforderungen einzelner Aufgaben als akzeptabel gelten.

Dazu sind auch genauere Einblicke in die Rezeption der Aufgaben und Materialien sowie der Lösungsstrategien der Schülerinnen und Schüler von Bedeutung. Zur Beurteilung liegen einerseits Interviewdaten, die in Cognitive Labs gewonnenen wurden, vor (Werner & Schreiber, 2015), anderseits Kommentare aus der Experten-Evaluierung des HiTCH-Tests sowie Schüler- und Lehrerkommentare zum Test.[31] Die Auswertung der Datenbestände von Schülerinnen und Schülern und Lehrkräften wird derzeit betrieben (Meyer-Hamme & Körber, in Vorb.).

6.2.3 Verhältnis von Wissen und Kompetenzen im HiTCH-Test

Die Rolle des Wissens ist für das Verständnis des HiTCH-Instruments wie auch für den Einsatz bedeutsam. Der Test ist als Kompetenztest so konstruiert, dass er nicht systematisch ein bestimmtes inhaltliches Wissen voraussetzt. Daher ist er nicht daran gebunden, dass die Schülerinnen und Schüler zuvor bestimmte Themen eines inhaltlichen historischen Curriculums im Unterricht behandelt haben. Das bedeutet aber nicht, dass die Kompetenzen historischen Denkens in den Aufgaben rein „abstrakt“, ohne konkrete historische „Inhalte“ und Themen erfasst würden. Die Testsituation fordert von den Schülerinnen und Schülern also kein „Stricken ohne Wolle“, wie es zuweilen der Kompetenzorientierung insgesamt vorgeworfen wird (Körber, 2010; Pallaske, 2015) – ganz im Gegenteil: Die Aufgaben sind so konstruiert, dass die Schülerinnen und Schüler sich mit konkreten und validen Problemen historischen Denkens auseinandersetzen müssen, die so formuliert sind, dass Personen mit historischen Kompetenzen gegenüber denjenigen im Vorteil sind, die den Test bzw. einzelne Auf-

31 Die Schülerinnen und Schüler z.B. waren aufgefordert, nach der Bearbeitung den Test als Ganzes zu kommentieren, regelmäßig wurden aber auch einzelne Aufgaben kommentiert, so dass Schülerkommentare in zwei Datentypen vorliegen. Die Lehrerinnen und Lehrer der Klassen, die am Test teilnahmen, durften den Test einsehen und kommentieren, aus Gründen der Testsicherheit aber nicht behalten. Zum Teil haben die Lehrkräfte von der Möglichkeit, ihre Sicht auf den Test zu dokumentieren, umfangreich Gebrauch gemacht.

gaben allein mit Hilfe evasiver, fachfremder Prozeduren und Operationen (etwa simplen oder komplexeren „sinnentnehmenden Lesens“) zu lösen versuchen.

Diese Konzeption ist dadurch möglich, dass die Aufgaben die nötigen Einzelinformationen zur Verfügung stellen. Das gilt vornehmlich für fallbezogenes Wissen, also nicht durch Transfer aus anderen historischen Denkprozessen übertragbare Begriffe und Konzepte. Solch fallbezogenes Wissen ist ein wesentliches Lehr- und Lernziel historischen Unterrichts, stellt aber keinen Bestandteil von Kompetenzen im Sinne des FUER-Modells dar. Es ist vielmehr neben diesen zu verorten (Borries, 2007b) und bildet gewissermaßen das Substrat, an welchem Kompetenzen erworben und entwickelt werden. Solch fallbezogenes Wissen kann und soll der HiTCH-Kompetenztest nicht erfassen und messen (dies muss mit anderen Instrumenten erfolgen). Er darf es auch nicht voraussetzen, insofern sonst die Verfügung über dieses Wissen, nicht aber über die Kompetenzen gemessen würde.

Im Gegensatz zum fallbezogenen Wissen wird Wissen, das auf mehrere historische Gegenstände anwendbar ist, sehr wohl in einem Kompetenztest erfasst (Borries, 2007b; Bräuer et al., 2016; Körber, 2010; Kühberger, 2012). Es geht die dabei um Sachkompetenz, also um das Verfügen über transferierbare Konzepte und Kategorien, die sich auf Inhalte, Methoden, Theorien wie historische Orientierungen beziehen können. Diese werden in den Materialien und Aufgaben (Items) nicht mitgeliefert, sondern vorausgesetzt: Die erfolgreiche Lösung der jeweiligen Aufgabe kann als Indikator für die Verfügung über konzeptuelles und kategoriales Wissens interpretiert werden. Die Korrektheit der Einzelergebnisse ist somit Indikator für die Kompetenz. Auf diese Weise misst der HiTCH-Test auch und gerade an Inhalten, die die Schülerinnen und Schüler in vorhergehendem Unterricht nicht behandelt haben müssen, deren historische Kompetenz, nicht aber in inhaltlicher Sicht den „Lernerfolg“ einer konkreten Unterweisung in bestimmte historischen Themen. Ein Einwand mangelnder „curricularer Validität“ trifft somit mit Bezug auf Inhalte und Themen des Geschichtsunterrichts nicht zu.

Zusammenfassend kann eine Kompetenztestung mit dem HiTCH-Test in Situationen angebracht sein, in denen die Kompetenzförderung zuvor explizites Lehrziel gewesen war und beispielsweise die Wirksamkeit einer Unterrichtsmethode im Hinblick auf die Kompetenzförderung untersucht wird. Darüber hinaus kann eine übergreifende, komparative Messung aber auch Einblicke darin liefern, inwiefern – unabhängig von speziellen Unterrichtskonzepten oder –methoden – von den Lernenden Kompetenzen erworben worden sind. Daher kann ein Kompetenztest auch dann zum Einsatz kommen, wenn der Unterricht nicht explizit auf Kompetenzförderung zielt.

6.2.4 Graduierungen der Kompetenz

Der HiTCH-Test zielte nicht darauf ab, das im FUER-Modell theoretisch formulierte Graduierungsmodell der Niveaus historischen Denkens vollständig zu erfassen. Höhere Testwerte sind zwar Indikatoren für die Verfügung über Kompetenzen auf höhe-

rem Niveau. Ob Schülerinnen und Schüler mit guten Ergebnissen aber bereits „elaboriert“ gemäß der FUER-Graduierung denken können, kann nicht gesagt werden, denn das elaborierte, transkonventionelle Niveau zeichnet sich gerade dadurch aus, dass über Fähigkeiten zur distanzierten Reflexion konventioneller Konzepte und Fähigkeiten verfügt werden kann. Der HiTCH-Test hingegen überprüft aber Kompetenzausprägungen bezogen auf Konventionen, die sich aus einem narrativistisch-konstruktivistischen Geschichtsverständnis ergeben. Qualitative Umschläge, z.B. hin zu einer neuen, anderen Logik historischen Denkens, kann er nicht messen.

Weil die HiTCH-Aufgaben auf Konzepte und Prozesse eines narrativistisch-konstruktivistischen historischen Denken zurückgreifen und die Verfügung über diese erfassen, stellen die Testwerte ein Maß für die Sicherheit des Verfügens über diese Konvention dar. In Bezug auf das FUER-Modell ist der Bezugspunkt also das intermediäre Niveau der Verfügung über konventionelle Konzepte und Prozesse. Das gemeinsame Niveaumerkmal ist die Verfügung über Prozeduren und Konzepte historischen Denkens, die gewissermaßen eine gesellschaftliche Konvention auf der Basis fachwissenschaftlicher Theorie darstellen. Das lässt sich an einigen wenigen Beispielen verdeutlichen: In Aufgaben, die darauf abstellen, „Perspektiven“ zu erkennen, ist die Denkfigur der Perspektivität historischer Deutungen gefragt. Aufgaben, welche Materialien als „Quellen“ und/oder „Darstellungen“ auszuwerten erfordern, verlangen eine Einsicht nicht nur in die implizite Logik, sondern auch in das zugrunde liegende Konzept.

Insofern die Testwerte im HiTCH-Test sich als Summe (bzw. Mittelwert) aus den Scores einer ganzen Reihe von Aufgaben errechnen, die alle einer gemeinsamen Verfügungslogik folgen, sind hohe Testwerte als Ausweis einer besseren (stabileren) Verfügung über dieses Niveau zu interpretieren, niedrige Testwerte als unsichere, fragile, nicht stabile Verfügung. Differenzieren lässt sich somit zwischen einem basalen Niveau (sehr niedrige Testwerte = sehr instabile, erratische Lösungen), einem ansatzweise (mittlere Testwerte) und ausgeprägt intermediärem Niveau (hohe Testwerte).

Für eine systematische Messung auch des elaborierten Niveaus historischen Denkens müsste der HiTCH-Test weiterentwickelt oder ein anderer Test erarbeitet werden. Angesichts der Definition des elaborierten Niveaus als der Fähigkeit zu je eigenständiger Reflexion erweist sich eine solche Itemkonstruktion als komplex. Eine Erfassung auch höherer Kompetenzniveaus mit Aufgaben, die ähnliche Konstruktionsmustern folgen wie die bisherigen HiTCH-Aufgaben, ergänzt um den theoretisch konstitutiven Aspekt der Reflexivität, erscheint aber nicht ausgeschlossen, wie komparative Spezialauswertungen von Schüler- und Studierendendaten nahelegen, über die an anderer Stelle zu berichten ist (Meis & Zuckowski, in Druck; Meyer-Hamme & Körber, in Vorb.). Die Entwicklung eines reliablen und validen Tests auch bezogen auf das elaborierte Niveau (interessant auch mit Blick auf die Erfassung von Kompetenz-Entwicklungsprozessen) erfordert weitergehende Konstruktionsarbeiten.

6.2.5 HiTCH-Test als Ersatz für schulische Leistungskontrollen?

Es kann aus geschichtsdidaktischer und psychometrischer Sicht festgehalten werden, dass es theoretisch wie praktisch durchaus möglich ist, Kompetenzen historischen Denkens mit Hilfe standardisierter Aufgaben in quantitativer Methodik zu messen. Dieser Befund bedarf jedoch einer Differenzierung: Als möglich erweist sich die Testung der prinzipiellen Fähigkeit, Fertigkeit und Bereitschaft historischen Denkens. Die dabei jeweils bearbeiteten Frage-, Re- und De-Konstruktions- sowie Orientierungsaufgaben wie auch diejenigen, welche die Verfügung über die kategorialen und konzeptuellen Grundlagen dieses Denkens (Sachkompetenz) erfassen sollen, beziehen sich dabei jeweils auf Gegenstände und Themen, die typischerweise nicht die ureigensten Orientierungsbedürfnisse der Schülerinnen und Schüler darstellen. Was also nicht geleistet werden kann (und was auch gar nicht beabsichtigt war), ist eine Ermittlung der Qualität der einzelnen authentischen, lebensweltlichen Orientierungsleistung. Wie es der Kompetenztheorie gemäß ist, wird vielmehr an lebensnah konstruierten beispielhaften Aufgaben die grundsätzliche Fähigkeit gemessen, nicht die jeweilige eigene Orientierung.

Aufgrund der statistischen Logik, welche die Reliabilität der Messung dadurch sicherstellt, dass Antworten auf mehrere ähnliche Fragen (Items) auf Konsistenz geprüft wird, sind Abweichungen jedes einzelnen Probanden in einzelnen Fragen ohne große Beeinflussung der psychometrischen Qualität des Tests möglich. Einzelne Schülerinnen und Schüler lassen sich – bei Gültigkeit des zugrundegelegten Modells – auf der Ebene der erreichten Kompetenzen unterscheiden, wobei bei nicht perfekter Reliabilität des Tests eine gewisse Unsicherheit über die Position der einzelnen Schülerinnen und Schüler besteht. Aussagen über die Fähigkeiten einzelner Schülerinnen und Schüler hinsichtlich einzelner Aufgaben sind mit dem vorliegenden Instrumentarium dagegen nicht reliabel zu treffen. Der HiTCH-Test kann auch deswegen nicht als Grundlage von Leistungsbeurteilungen (Benotungen) von Schülerinnen und Schülern in der Schule herangezogen werden, da der Test auf der Basis eines Kompetenzmodells entwickelt wurde und nicht auf der Basis schulischer Curricula.

Eine hohe Aussagekraft hat der HiTCH-Test dagegen mit Blick auf Vergleiche zwischen Gruppen – etwa hinsichtlich unterschiedlicher durchschnittlicher Kompetenzausprägungen („Entwickelt eine Klasse unter Bedingung x durchschnittlich bessere Kompetenz als eine unter Bedingung y?“) und hinsichtlich der Verteilungen dieser Maße („Sind die Kompetenzen von Schülerinnen und Schülern unter Bedingung x homogener als unter Bedingung y?“).

6.3 Ausblick: Die Bedeutung von HiTCH für Forschung und Schule

6.3.1 Anregungen für die Forschung

Bei den Überlegungen zum Einsatz des HiTCH-Tests müssen die Ziele der Testentwicklung beachtet werden: So ist der HiTCH-Test für Large-Scale-Assessments, aber nicht für Individualdiagnostik angelegt und zielt auf die Erfassung von Kompetenz, nicht auf die Erfassung bestimmter Curriculumsinhalte, so dass z.B. keine Aussagen zum Erwerb fallbezogenen Wissens getroffen werden können. Doch Elemente des HiTCH-Tests können auch in einem weiteren Forschungskontext genutzt werden.

Der HiTCH-Test spielt besonders dort seine Stärken aus, wo er über die (statistische) Erfassung des Maßes und der Streuungen historischer Kompetenzwerte in und (vergleichend) über Gruppen hinweg weitere Forschung unterstützen kann. Dies ist z.B. der Fall, wenn der Blick darauf gerichtet werden soll, welche Bedingungen sowie welche Gestaltungen historischen Lernens der Entwicklung historischer Kompetenzen förderlich oder hinderlich sind. Der HiTCH-Test dient dann dazu, die unter den jeweiligen Bedingungen erreichten Kompetenzausprägungen zu vergleichen und/oder auf die breitere Gruppe von Neuntklässlerinnen und -klässler zu beziehen, für die er entwickelt wurde. Genutzt werden kann er insbesondere auch für alle Interventionsstudien, die untersuchen wollen, ob die erprobten Treatments Auswirkungen auf die Entwicklung historischer Kompetenzen haben.

Ebenso ist der Test geeignet, um in der Messung anderer Kompetenzen deren Konfundierung mit oder Trennung von historischen Kompetenzen einschätzen zu können. Er kann also als Instrument zur divergenten Validierung in der Entwicklung von anderen Kompetenztests eingesetzt werden, um den Einfluss oder auch Anteil historischen Denkens zu bestimmen – so etwa für Messungen religiöser, philosophischer, literarischer politischer Kompetenz oder für die Messung historischer Lesekompetenz.

Darüber hinaus bietet der HiTCH-Test Anregungen für qualitative Studien zur Erforschung historischen Denkens. Die Ergebnisse (Kapitel 5) des HiTCH-Tests bergen in sich Anregungen für weitere Forschung. Denkbar und lohnend erscheint es uns, die Logik der Operationalisierung der Kompetenzanforderungen in den HiTCH-Aufgaben zur Grundlage für die Entwicklung *diagnostischer Instrumente* zu machen, welche einer qualitativen Methodik folgen. Dabei ist aber zu bedenken, dass damit ein entscheidendes Kriterium quantitativer Testung aufgegeben wird: Während hier nämlich über Zahl, Anteil und Verteilung richtiger Lösungen auf die Stabilität der Verfügung über die jeweils operationalisierte Kompetenz geschlossen wird und man wegen der statistischen Stabilität (Reliabilität) nicht auf die Sichtbarmachung und Erhebung der dahinter liegenden Denkprozesse angewiesen ist (unsichere Verfügung zeigt sich in Fehlerquoten nahe der Ratewahrscheinlichkeit), ist dieses Vorgehen bei individuellen Erhebungen nicht möglich. Das Ziel ist hier vielmehr, Einsichten in die Prozesse hinter der Aufgabenlösung zu gewinnen.

Für Interview- und Beobachtungssituationen, welche individuelle Prozesse historischen Denkens und Ausprägungen historischer Kompetenzen untersuchen sollen, können etwa Aufgaben des HiTCH-Typs (original oder in einiger Abwandlung) als „Grundreiz" herangezogen werden, zu denen Schülerinnen und Schülern sich, z.B. laut denkend, verhalten sollen. Unter Zugrundlegeung der für den HiTCH-Test identifizierten richtigen Lösungen oder der aus den quantitativen Untersuchungen abzuleitenden „typischen" Abweichungen lassen sich so individualdiagnostische Materialien erstellen bzw. die „Codes" zur Klassifizierung der Aussagen der Probanden gewinnen. Die dabei relevanten Daten für die Auswertungen sind die Erläuterungen und Begründungen für das Vorgehen bei der Lösung der Aufgabe.

Bei der Materialentwicklung unter Nutzung der HiTCH-Aufgaben müssen die im Entwicklungsprozess für die Operationalisierung der Aufgaben leitenden Gesichtspunkte als kategoriale Indikatoren mitgedacht und parat gehalten werden. Man wird sich also nicht nur auf die endgültigen Aufgabenformate des HiTCH-Tests stützen dürfen, sondern wird die (oben beispielhaft dargelegten) Funktionsprinzipien der Aufgaben heranziehen müssen. Zum Abgleich müssen dann die Denkprozesse mit erhoben werden (etwa in Form gleichzeitigen oder nachträglichen lauten Denkens oder schriftlicher Begründungen), die durch die Aufgaben ausgelöst wurden (vgl. auch das Vorgehen bei den zur Testentwicklung des HiTCH-Tests eingesetzten Cognitive Labs; Werner & Schreiber, 2015).

6.3.2 Anregungen für den Geschichtsunterricht

Auch wenn der HiTCH-Test nicht für den unmittelbaren Einsatz durch die Lehrkraft *in* der Schule entwickelt worden ist, bietet er unseres Erachtens auch für den Geschichtsunterricht große Potenziale.

Erstens ist da die Bedeutung für die Lehrkraft, die im Studium, im Rahmen der schulischen Arbeit oder bei Fortbildungen mit dem HiTCH-Test und seinen Konstruktionsprinzipien in Kontakt kommen mag. Nach unserer Auffassung können die im HiTCH-Test und dem ihm zugrunde liegenden Kompetenzmodell formulierten Einsichten in Fähigkeiten des historischen Denkens und ihres Potenzials für historische Orientierung Lehrkräfte dabei unterstützen, ihren Unterricht (neu) zu profilieren. Die in Kapitel 2 dieser Publikation gebündelten theoretischen Grundlagen schaffen dafür eine gesicherte Basis. Die Überlegungen zur Testentwicklung und die Aufgabenbeispiele (vgl. Kapitel 3) beinhalten Impulse sowohl für die Entwicklung von Förder- als auch Prüfungsaufgaben und können im Rahmen der fachdidaktischen Ausbildung von Diagnosekompetenz herangezogen werden.

Zweitens kann der HiTCH-Test prinzipiell zur formativen Evaluation und Weiterentwicklung eines kompetenzorientierten Unterrichts herangezogen werden. Zu Beginn und am Ende der Schuljahrs eingesetzt, kann er z.B. Kompetenzentwicklungen verdeutlichen. So kann geprüft werden, inwiefern sich kompetenzorientierter (oder nicht kompetenzorientierter) Unterricht auf die Ergebnisse ausgewirkt hat, ob die

Geschichtsnote mit den Ergebnissen des Test korreliert, inwiefern die Schülerinnen und Schüler im Test dazu in der Lage waren, beim an spezifischen Fällen Erlernten vom Inhalt zu abstrahieren und damit zusammenhängende Kompetenzanforderungen in HiTCH-Aufgaben zu bewältigen. Für diese Anwendungen wäre es allerdings wünschenswert, dass bereits „typische" Entwicklungsverläufe über ein Schuljahr im Sinne von „Normen" dokumentiert wären, was aber nicht der Fall ist.

Auch für den Vergleich zwischen Klassen kann der HiTCH-Test prinzipiell genutzt werden. Es ließe sich dann z.B. feststellen,

- inwiefern Schülerinnen und Schüler unter unterschiedlichen Bedingungen (etwa: verschiedener Bücher, Stoff- vs. Kompetenzorientierung; mehr Unterricht; etc.) unterschiedlich gut historische Kompetenzen erwerben;
- inwiefern im Zeitverlauf Kompetenzen tatsächlich kumulativ erworben werden;
- inwiefern Schülerinnen und Schüler mit gleicher Note in unterschiedlichen Klassen, Schulen, Ländern unterschiedlich gut in „Kompetenzen" sind.

Für den systematischen Einsatz des HiTCH-Tests im Umfeld des Geschichtsunterrichts wäre allerdings eine deutliche Erweiterung des Itempools notwendig, die auch einen wiederholten Einsatz des Tests mithilfe unterschiedlicher Items in der gleichen Lerngruppe ermöglichen würde.

Drittens kann nicht nur der Test selbst bzw. seine theoretische Einbettung für den Geschichtsunterricht genutzt werden, sondern auch die in ihm umgesetzte Logik der Operationalisierung entsprechender (Teil-)Kompetenzen. Dabei geht es keineswegs um ein „Test-Training". Denkbar wäre hingegen eine reflexive Thematisierung von Aufgaben und dem in ihnen intendierten historischen Denken als Ausgangspunkt von Unterrichtsprozessen zu wählen. Lehrkräfte könnten sich, wie oben bereits angedeutet, von den Aufgabenstellungen für die Entwicklung kompetenzorientierter Förder- und Prüfungsaufgaben anregen lassen. Vielfach wurde der Wunsch nach solchen Aufgabenformaten in den oben angesprochenen Lehrerkommentaren zum HiTCH-Test geäußert. Allerdings erfordert die Bereitstellung solcher unterrichtlich einsetzbarer, HiTCH-förmiger Reflexionsaufgaben einiges an Vorarbeit der Erstellung und Validierung. Hierfür Beispielaufgaben bereitzustellen, wäre somit eine Aufgabe für ein Folge- bzw. Umsetzungsprojekt.

6.3.3 Weiterarbeit im HiTCH-Projekt

Der hier vorgestellte HiTCH-Test ist ein „fertiges" Instrument, mit dessen Hilfe die Resultate kumulativen Kompetenzerwerbs im Bereich der Geschichte erfasst werden können. Gleichzeitig betrachten die Projektbeteiligten das Projekt nicht als „abgeschlossen", sondern im besten Sinne als „work in progress". Zur Fortsetzung der Forschungsarbeit wurde nach Auslaufen der Förderung des Projekts durch das BMBF das HiTCH-Konsortium gegründet (vgl. die HiTCH-Website: www.hitch-projekt.de). Die Agenda des Konsortiums bzw. seiner Mitglieder umfasst u.a. die folgenden Arbeitsschwerpunkte.

1. In Anlehnung an den bestehenden Itempool sollen auch in Zukunft fortlaufend weitere strukturgleiche oder ähnliche Aufgaben neu- und weiterentwickelt werden, um den Itempool des HiTCH-Tests zu vergrößern. Im Zuge eines solchen Entwicklungsprojekts kann dann auch die verbliebene Ungleichgewichtigkeit der Repräsentanz behoben werden.
2. Der Itempool soll möglichst auch durch Aufgaben erweitert werden, die eine kurze schriftliche Antwort (*short answer format*) erfordern. Hier stellen sich – wegen der Kodierungsnotwendigkeit – insbesondere Fragen nach der Reliabilität der entsprechenden Items.
3. Es sollen Varianten des HiTCH-Tests, z.B. für erwachsene Lerner entstehen; hierauf abzielende Analysen werden im Sammelband Meyer-Hamme & Körber (in Vorb.) publiziert (für erste Untersuchungen von Studierendendaten im HiTCH-Kontext: Borries, 2016; Meis & Zuckowski, in Druck. Bedarf gäbe es auch für ein Instrument für junge Lerner (Bräuer & Schreiber, 2016).
4. Die Entwicklung weiterer Items wird auch dazu genutzt, über eine weitere Ausdifferenzierung der Erfassung der Kernkompetenzen die Separierbarkeit einzelner Kompetenzbereiche im Detail zu prüfen.
5. Zudem soll in künftigen Arbeiten zur Validierung des HiTCH-Tests die Gemeinsamkeiten und Unterschiede zu anderen Kompetenzbereichen wie beispielsweise dem „Lesen" vertiefend nachgegangen werden.
6. Der Datensatz kann – beispielsweise im Rahmen von Qualifikationsarbeiten – für Sekundärauswertungen genutzt werden. Denkbar sind psychometrische Fragestellungen wie vertiefte Analysen zum Ausfüllverhalten (z.B. Itempositionseffekte oder Rolle der Motivation). Auch fachdidaktisch orientierte Detailanalysen zu einzelnen Aufgabenblöcken bzw. Items könnten auf der Grundlage des Datensatzes erfolgen.
7. Darüber hinaus sind Studien geplant, in denen der HiTCH-Test als abhängige oder als unabhängige Variable eingesetzt wird. Beispielsweise können die Effekte bestimmter institutioneller Strukturen (z.B. Schulformen), curricularer Orientierungen sowie Unterrichtssettings bzw. der Unterrichtsqualität auf die Ausprägung historischer Kompetenzen, wie sie mit dem HiTCH-Test erfasst wird, untersucht werden. Eine Frage könnte beispielsweise sein, wie sich ein kompetenzorientierter Unterricht auf die Schülerleistung im HiTCH-Test auswirkt. Auch weitere Merkmale des Unterrichts – beispielsweise der Einbezug von digitalen Medien bzw. die Anbindung an außerschulische Lernorte, aber auch Aspekte der Inklusion – können untersucht werden.

Nicht nur das Konsortium, sondern die wissenschaftliche Community ist eingeladen, den HiTCH-Test für weitere Forschungsarbeiten zu nutzen. Wie oben bereits beschrieben, bieten sich zahlreiche Felder an, in denen die Erhebungen von Ausprägung und Verteilung der Kompetenz(en) historischen Denkens wertvolle Ergebnisse bringen kann, bzw. für die die Erfahrungen mit dem HiTCH-Test genutzt werden können.

Die Internationalisierung der Kompetenzdebatte wie auch der Debatte zur Kompetenzmessung wird unter der Federführung von Andreas Körber und Johannes Meyer-Hamme betrieben. Dazu fanden bereits zwei internationale Workshops in Hamburg statt. Mit Blick auf die angestrebte und im deutschsprachigen Rahmen mit dem HiTCH-Test bereits eingelöste Unabhängigkeit des Tests von konkreten Curricula ist eine Weiterentwicklung auf eine international vergleichende Erfassung der Kompetenzen historischen Denkens wünschenswert. Wohl wissend, dass damit weitere Bedingungen kontrolliert werden müssen (insbesondere solche terminologischer, aber auch epistemologischer Übersetzung; Körber, 2016; Seixas, 2015, 2016) und dass die deutschen Kompetenzmodelle als bestimmte Formen von Kognitionsmodellen geschichtsdidaktischer Forschung und Entwicklung gelten müssen, denen insbesondere in den Niederlanden (van Boxtel & van Drie, 2008), den USA (Stanford History Education Group (SHEG) um Wineburg) und Kanada (Seixas, 2016; Lévesque, 2016) andere Kognitionsmodelle partiell entsprechen, aber auch Unterschiede zu ihnen aufweisen, erscheint ein solcher Zugriff sinnvoll, um die Gemeinsamkeiten, aber auch Unterschiede des Erwerbs historischer Denk- und Reflexionsfähigkeiten in der emergierenden globalen Welt vergleichend erfassen zu können.

Mit der Entwicklung und Validierung des nun vorliegenden HiTCH-Tests wurde ein wichtiger Zwischenschritt erreicht. Darüber hinaus hat sich mit dem HiTCH-Konsortium eine interdisziplinäre Forschergruppe gebildet, in der sich fachdidaktische und psychometrische Expertise hervorragend ergänzen. Damit war es möglich, die Herausforderungen zu meistern, die mit der psychometrischen Ausmessung einer komplex strukturierten kulturwissenschaftlichen Disziplin, wie es das Fach Geschichte darstellt, verbunden ist. Die Grundlage wurde damit geschaffen, um weitere Aufgaben zu entwickeln und das Projekt auf weitere Forschungsfelder (z.B. Zielgruppen) auszuweiten. Das HiTCH-Konsortium hofft auf eine Vielzahl von weiterführenden Forschungsprojekten auch außerhalb der Gruppe und auf einen intensiven Austausch mit Vertretern und Vertreterinnen der Fachdidaktik Geschichte und der Empirischen Bildungsforschung.

Literatur

Adams, R. J., Wilson, M. & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23. doi: 10.1177/0146621697211001

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research and Perspectives, 9*(1), 95–104.

Angvik, M. & Borries, B. von (Eds.). (1997). *Youth and history*: *A comparative European survey on historical consciousness and political* a*ttitudes among adolescents.* Vol. A. Hamburg: Körber-Stiftung.

Auer, M., Gruber, G., Mayringer, H. & Wimmer, H. (2005). *SLS 5–8. Salzburger Lese-Screening für die Klassenstufen 5–8.* Bern: Huber.

Baron, C. (2012). Understanding historical thinking at historic sites. *Journal of Educational Psychology*, *104*(3), 833–847.

Bäuerlein, K., Lenhard, W. & Schneider, W. (2012). *Lesetestbatterie für die Klassenstufen 8–9.* Göttingen: Hogrefe.

Barricelli, M., Gautschi, P. & Körber, A. (2012). Historische Kompetenzen und Kompetenzmodelle. In M. Barricelli & M. Lücke (Hrsg.), *Handbuch Praxis des Geschichtsunterrichts* (S. 207–235). Schwalbach/Ts.: Wochenschau Verlag.

Baumert, J., Brunner, M., Lüdtke, O. & Trautwein, U. (2007). Was messen internationale Schulleistungsstudien? – Resultate kumulativer Wissenserwerbsprozesse. Eine Antwort auf Heiner Rindermann. *Psychologische Rundschau, 58*(2), 118–128.

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 15–68). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J., Trautwein, U. & Artelt, C. (2003). Schulumwelten – institutionelle Bedingungen des Lehrens und Lernens. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–331). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumgartner, H.-M. (1975). Narrative Struktur und Objektivität. Wahrheitskriterien im historischen Wissen. In J. Rüsen & H.-M. Baumgartner (Hrsg.), *Historische Objektivität. Aufsätze zur Geschichtstheorie* (S. 45–67). Göttingen: Vandenhoeck & Ruprecht.

Baumgartner, H.-M. (1997). Narrativität. In K. Bergmann, K. Fröhlich & A. Kuhn (Hrsg.), *Handbuch der Geschichtsdidaktik* (5. Aufl.) (S. 157–160). Seelze-Velber: Kallmeyer.

Bergmann, K. (2000). *Multiperspektivität. Geschichte selber denken.* Schwalbach/Ts: Wochenschau Verlag.

Bertram, C. & Wagner, W. (im Druck). Der Weg von NAEP zu HiTCH. Erfassung historischer Kompetenzen in standardisierten Formaten am Beispiel von Aufgaben zur US Geschichte. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 15. Beiträge zur Tagung „geschichtsdidaktik empirisch 15".* Bern: hep.

Bertram, C., Wagner, W. & Schaser, E. (2015). Historische Kompetenzen mit offenen Antwortformaten messen – Eine Studie auf Basis der „Sechser-Matrix" des FUER-Modells. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13"* (S. 165–180). Bern: hep.

Bertram, C., Wagner, W. & Trautwein, U. (2013). Chancen und Risiken von Zeitzeugenbefragungen – Entwicklung eines Messinstruments für eine Interventionsstudie. In J. Hodel & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 12. Beiträge zur Tagung „geschichtsdidaktik empirisch 12"* (S. 108–119). Bern: hep.

Bertram, C., Wagner, W. & Trautwein, U. (2014). Zeitzeugenbefragungen im Geschichtsunterricht: Entwicklung eines Kurzinstruments für die Wirksamkeitsmessung. In T. Arand & M. Seidenfuß (Hrsg.), *Neue Wege – neue Themen – neue Methoden? Ein Querschnitt aus der geschichtsdidaktischen Forschung des wissenschaftlichen Nachwuchses* (S. 191–208). Göttingen: V&R Academic.

Bertram, C., Wagner, W. & Trautwein, U. (in press). Learning historical thinking with oral history interviews: A cluster randomized controlled intervention study of oral history interviews in history lessons. *American Educational Research Journal*. doi: 10.3102/0002831217694833

Blum, W., Drüke-Noe, K., Hartung, R. & Köller, O. (2006). *Bildungsstandards Mathematik konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsideen und Fortbildungsmöglichkeiten.* Berlin: Cornelsen/Scriptor.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. doi: 10.1007/bf02293801

Borries, B. von (1974). *Lernziele und Testaufgaben für den Geschichtsunterricht, dargestellt an der Behandlung der Römischen Republik in der 7. Klasse. Anmerkungen und Argumente zur historischen und politischen Bildung.* Stuttgart: Klett.

Borries, B. von (1988). *Geschichtslernen und Geschichtsbewußtsein. Empirische Erkundungen zu Erwerb und Gebrauch von Historie.* Stuttgart: Klett.

Borries, B. von (1995). *Das Geschichtsbewusstsein Jugendlicher. Eine repräsentative Untersuchung über Vergangenheitsdeutungen, Gegenwartswahrnehmungen und Zukunftserwartungen von Schülerinnen und Schülern in Ost- und Westdeutschland.* Weinheim: Juventa.

Borries, B. von (2004). Perspektivenwechsel und Sinnbildungsfiguren im Umgang mit der Geschichte. In B. von Borries (Hrsg.), *Lebendiges Geschichtslernen. Bausteine zu Theorie und Pragmatik, Empirie und Normfrage* (S. 236–287). Schwalbach/Ts.: Wochenschau Verlag.

Borries, B. von (2007a). Ergebnisse messen (Lerndiagnose im Fach Geschichte). In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 651–673). Neuried: ars una.

Borries, B. von (2007b). ‚Kompetenzmodell' und ‚Kerncurriculum'. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 334–360). Neuried: ars una.

Borries, B. von (2013). 1989 – Erinnerung für die Zukunft. In W. Beutel & P. Fauser (Hrsg.), *Demokratie erfahren. Analysen, Berichte und Anstöße aus dem „Förderprogramm Demokratisch Handeln"* (S. 163–196). Schwalbach/Ts.: Wochenschau Verlag.

Borries, B. von (2016). „Zweimal Untergang Roms?" Werkstattbericht zur (versuchten) Replikation einer britischen Studie im deutschen Kompetenztest (unter Mitarbeit von A. Zuckowski). In K. Lehmann, M. Werner & S. Zabold (Hrsg.), *Historisches Denken jetzt und in Zukunft: Wege zu einem theoretisch fundierten und evidenzbasierten Umgang mit Geschichte; Festschrift für Waltraud Schreiber zum 60. Geburtstag* (S. 235–252). Berlin: LIT.

Borries, B. von, Fischer, C., Leutner-Ramme, S. & Meyer-Hamme, J. (2005). *Schulbuchverständnis, Richtlinienbenutzung und Reflexionsprozesse im Geschichtsunterricht. Eine qualitativ-quantitative Schüler- und Lehrerbefragung im deutschsprachigen Bildungswesen 2002.* Neuried: ars una.

Borries, B. von & Meyer-Hamme, J. (Hrsg.). (2014). *Zwischen „Genuss" und „Ekel". Ästhetik und Emotionalität als konstitutive Momente historischen Lernens.* Schwalbach/Ts.: Wochenschau Verlag.

Borries, B. von, Pandel, H.-J. & Rüsen, J. (Hrsg.). (1991). *Geschichtsbewusstsein empirisch*. Pfaffenweiler: Centaurus.

Borries, B. von & Rüsen, J. (Hrsg.). (1994). *Geschichtsbewußtsein im interkulturellen Vergleich. Zwei empirische Pilotstudien.* Pfaffenweiler: Centaurus.

Bracke, S., Flaving, C., Köster, M. & Zülsdorf-Kersting, M. (2014). History education research in Germany. Empirical attempts at mapping historical thinking and learning. In M. Köster, H. Thünemann & M. Zülsdorf-Kersting (Eds.), *Researching history education. International perspectives and disciplinary traditions* (pp. 9–55). Schwalbach/Ts.: Wochenschau Verlag.

Brauch, N. (2015). *Geschichtsdidaktik.* Berlin: De Gruyter Oldenbourg.

Brauch, N. (2016). „Wer nicht fragt bleibt dumm!" Geschichtswissenschaftliche Fragekompetenz als Bestandteil professioneller Geschichtslehrer_innenkompetenz. In K. Lehmann, M. Werner & S. Zabold (Hrsg.), *Geschichtsdidaktik in Vergangenheit und Gegenwart: Bd. 10. Historisches Denken jetzt und in Zukunft. Wege zu einem theoretisch fundierten und evidenzbasierten Umgang mit Geschichte* (S. 189–199). Berlin: LIT.

Brauch, N. (in press). Bridging the gap – Comparing history curricula in history teacher education in western countries. In M. Carretero, S. Berger & M. Grever (Eds.), *Palgrave handbook of research in historical culture and education.* Basingstoke: Palgrave Macmillan.

Bräuer, B., Lehmann, K. & Werner M. (2016). Historisches Wissen und Orientierung: Eine explorative Interviewstudie am Beispiel des Umgangs Studierender mit dem Ersten Weltkrieg. In K. Lehmann, M. Werner & S. Zabold (Hrsg.), *Historisches Denken jetzt und in Zukunft: Wege zu einem theoretisch fundierten und evidenzbasierten Umgang mit Geschichte; Festschrift für Waltraud Schreiber zum 60. Geburtstag* (S. 253–267). Berlin: LIT.

Bräuer, B. & Schreiber, W. (2016). Orientierungsgelegenheiten – Theoriebildung für gemeinsames Geschichtslernen in inklusiven Klassen. In C. Kühberger & R. Schneider (Hrsg.), *Inklusion im Geschichtsunterricht: zur Bedeutung geschichtsdidaktischer und sonderpädagogischer Fragen im Kontext inklusiven Unterrichts* (S. 85–102). Bad Heilbrunn: Klinkhardt.

Breakstone, J. (2013). *History assessment of thinking. Design, interpretation and implementation*. Doctoral dissertation, Stanford University, CA. Retrieved from https://purl.stanford.edu/nt301xp3169

Breakstone, J., Smith, M. & Wineburg, S. (2013). Beyond the bubble in history/social assessments. *Phi Delta Kappan, 94*(5), 53–57.

Buchstein, H., Frech, S. & Pohl, K. (Hrsg.). (2016). *Beutelsbacher Konsens und politische Kultur. Siegfried Schiele und die politische Bildung.* Schwalbach/Ts.: Wochenschau Verlag.

Cain, T. & Chapman, A. (2014). Dysfunctional dichotomies? Deflating bipolar constructions of curriculum and pedagogy through case studies from music and history. *Curriculum Journal*, *25*(1), 111–129.

Chin, C. & Chia, L.-G. (2004). Problem-based learning: Using students' questions to drive knowledge construction. *Science Education*, *88*, 707–727.

Danto, A. C. (1965). *Analytische Philosophie der Geschichte*. Frankfurt am Main: Suhrkamp.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Deutz-Schroeder, M. & Schroeder, K. (2007). Das DDR-Bild von Schülern in Nordrhein-Westfalen. *Arbeitspapiere des Forschungsverbundes SED-Staat*, *39*.

Droysen, J. G. (1857). *Historik*. (Historisch-kritische Ausgabe von P. Leyh. Bd. I., 1977). Stuttgart: Frommann-Holzboog.

Droysen, J. G. (1881). *Historik. Vorlesungen über Enzyklopädie und Methodologie der Geschichte.* (Herausgegeben von R. Hübner, 1974). Darmstadt: Wissenschaftliche Buchgesellschaft.

Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden. Lehrbuch*. Weinheim: Beltz.

Eisele-Brauch, N. (2010). Nachhaltiges historisches Lernen als Gegenstand empirischer Lehr-/Lernforschung: Ein Rückblick auf den Tagungssommer 2009. *Zeitschrift für Geschichtsdidaktik*, *9*(1), 143–158.

Eliasson, P., Alvén, F., Axelsson Yngveús, C. & Rosenlund, D. (2015). Historical consciousness and historical thinking reflected in large-scale assessment in Sweden. In E. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 171–182). New York: Routledge.

Ercikan, K. & Seixas, P. (Eds.). (2015). *New directions in assessing historical thinking*. New York: Routledge.

Ercikan, K., Seixas, P., Lyon-Thomas, J. & Gibson, L. (2015). Cognitive validity evidence for validating assessments of historical thinking. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 206–220). New York: Routledge.

Erdmann, E. & Hasberg, W. (Eds.). (2011). *Facing mapping bridging diversity. Foundation of a European discourse on history education.* (Part 1 and 2). Schwalbach/Ts.: Wochenschau Verlag.

Feiks, D., Laubert, V. & Rothermel, G. (1975). *Objektivierte Leistungsmessung, Leistungsbeurteilung und Lerndiagnose: Testmodell am Beispiel des Geschichtsunterrichts*. Stuttgart: Klett.

Frederking, V. (Hrsg.). (2008). *Schwer messbare Kompetenzen. Herausforderungen für die empirische Fachdidaktik.* Baltmannsweiler: Schneider.

Frederking, V. & Brüggemann, J. (2012). Literarisch kodierte, intendierte bzw. evozierte Emotionen und literarästhetische Verstehenskompetenz. Theoretische Grundlagen einer empirischen Erforschung. In C. Frickel, G. Kammler & G. Rupp (Hrsg.), *Literaturdidaktik im Zeichen von Kompetenzorientierung und Empirie. Perspektiven und Probleme* (S. 15–41). Freiburg: Fillibach.

Frederking, V., Brüggemann, J. & Hirsch, M. (2016). Das fünfdimensionale ‚Literary Literacy' – Modell und seine interdisziplinären Implikationen am Beispiel der Geschichtsdidaktik. In K. Lehmann, M. Werner & S. Zabold (Hrsg.), *Historisches Denken jetzt und in Zukunft: Wege zu einem theoretisch fundierten und evidenzbasierten Umgang mit Geschichte; Festschrift für Waltraud Schreiber zum 60. Geburtstag* (S. 211–234). Berlin: LIT.

Frederking, V., Meier, C., Brüggemann, J., Gerner, V. & Friedrich, M. (2011). Literarästhetische Verstehenskompetenz – theoretische Modellierung und empirische Erforschung. *Zeitschrift für Germanistik, XXI*(1), 131–144.

Frederking, V., Meier, C., Stanat, P. & Dickhäuser, O. (2008). Ein Modell literarästhetischer Urteilskompetenz. *Didaktik Deutsch*, *25*, 11–31.

Frederking, V., Roick, T. & Steinhauer, L. (2011). Literarästhetische Urteilskompetenz – Forschungsansatz und Zwischenergebnisse. In H. Bayrhuber, U. Harms, B. Muszynski, B. Ralle, M. Rothgangel, L. Schön, H. J. Vollmer & G. Weigand (Hrsg.), *Empirische Fundierung in den Fachdidaktiken* (S. 75–94). Münster: Waxmann.

Gautschi, P. (2009). *Guter Geschichtsunterricht: Grundlagen, Erkenntnisse, Hinweise.* Schwalbach/Ts.: Wochenschau Verlag.

Glockner-Rist, A. & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(4), 544–565. doi: 10.1207/S15328007SEM1004_4

Gustafsson, J.-E. & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97–121). Washington, D.C.: American Psychological Association.

Halldén, O. (1997). Conceptual change and the learning of history. *International Journal of Educational Research, 27*(3), 201–210.

Hartig, J. & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54*(4), 418–431.

Hartig, J., Klieme, E. & Leutner, D. (2008). *Assessment of competencies in educational contexts*. Washington: Hogrefe.

Hartmann, U. (2008). *Perspektivenübernahme als eine Kompetenz historischen Verstehens.* Unveröffentlichte Dissertation, Georg-August-Universität zu Göttingen.

Hartung, O. (2013). Sprache und konzeptionelles Schreibhandeln im Fach Geschichte. Ergebnisse der empirischen Feldstudie „Geschichte–Schreiben–Lernen". In M. Becker-Mrotzek, K. Schramm, E. Thürmann & H. Vollmer (Hrsg.), *Sprache im Fach – Sprachlichkeit und fachliches Lernen* (S. 335–352). Münster: Waxmann.

Hartung, O. (2015). Generisches Geschichtslernen: Drei Aufgabentypen im Vergleich. *Zeitschrift für Geschichtsdidaktik*, *14*(1), 47–62.

Hasberg, W. & Körber, A. (2003). Geschichtsbewusstsein dynamisch. In A. Körber (Hrsg.), *Geschichte – Leben – Lernen. Bodo von Borries zum 60. Geburtstag* (S. 177–200). Schwalbach/Ts.: Wochenschau Verlag.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision.* Göttingen: Hogrefe.

Hodel, J. & Ziegler, B. (Hrsg.). (2008). *Forschungswerkstatt Geschichtsdidaktik 07. Beiträge zur Tagung „geschichtsdidaktik empirisch 07".* Bern: hep.

Hodel, J. & Ziegler, B. (Hrsg.). (2010). *Forschungswerkstatt Geschichtsdidaktik 09. Beiträge zur Tagung „geschichtsdidaktik empirisch 09".* Bern: hep.

Hodel, J., Waldis, M. & Ziegler, B. (Hrsg.). (2013). *Forschungswerkstatt Geschichtsdidaktik 12. Beiträge zur Tagung „geschichtsdidaktik empirisch 12".* Bern: hep.

Hodel, J., Waldis, M., Zülsdorf-Kersting, M. & Thünemann, H. (2013). Schülernarrationen als Ausdruck historischer Kompetenz. *Zeitschrift für Didaktik der Gesellschaftswissenschaften, 2,* 121–145.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L. & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quaterly, 50*(3), 391–402.

Ingenkamp, K. & Mielke, H. (o.J.). *Geschichtstest Neuzeit* (GTN 8-10): Teil I: 1890–1932, Teil II: 1933 – Gegenwart. Weinheim, Berlin: Beltz.

Jeismann, K.-E. (1977). Didaktik der Geschichte. Die Wissenschaft von Zustand, Funktion und Veränderung geschichtlicher Vorstellungen im Selbstverständnis der Gegenwart. In E. Kosthorst (Hrsg.), *Geschichtswissenschaft. Didaktik – Forschung – Theorie* (S. 9–33). Göttingen: Vandenhoeck & Ruprecht.

Jeismann, K.-E. (1980). ‚Geschichtsbewußtsein': Überlegungen zur zentralen Kategorie eines neuen Ansatzes der Geschichtsdidaktik. In H. Süssmuth (Hrsg.), *Geschichtsdi-*

daktische Positionen. Bestandsaufnahme und Neuorientierung (S. 179–222). Paderborn: Schöningh.
Jeismann, K.-E. (2000). *Geschichte und Bildung: Beiträge zur Geschichtsdidaktik und zur historischen Bildungsforschung.* Paderborn: Schöningh.
Jensen, B. E. (2003). *Historie – livsverden og fag* (1. udgave, 1. oplag). [Kbh.]: Gyldendal.
Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl.) (S. 27–74). Berlin: Springer.
Keirn, T. (2016). History curriculum. A transatlantic analysis. In D. Wyse, L. Hayward & J. Pandya (Eds.), *The SAGE handbook of curriculum, pedagogy and assessment* (pp. 408–423). London: SAGE.
King, P. & Kitchener, K. S. (1994). *Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults.* San Francisco: Jossey-Bass.
Klieme, E., Avenarius, H., Blum, W., Döblich, P., Gruper, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Zugriff am 25.06.2015. Verfügbar unter http://www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf
Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M. & Stanat, P. (Hrsg.). (2010). *PISA 2009: Bilanz nach einem Jahrzehnt*. Münster: Waxmann.
KMK, Ständige Konferenz der Kultusminister (1989/2005). *Einheitliche Prüfungsanforderungen in der Abiturprüfung. Geschichte.* Beschluss vom 01.12.1989 i. d. F. vom 10.02.2005. Zugriff am 12.02.2015. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1989/ 1989_12_01-EPA-Geschichte.pdf
Körber, A. (2007a). Graduierung: Die Unterscheidung von Niveaus der Kompetenzen historischen Denkens. In A. Körber (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 415–472). Neuried: ars una.
Körber, A. (2007b). Grundbegriffe und Konzepte: Bildungsstandards, Kompetenzen und Kompetenzmodelle. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 54–86). Neuried: ars una.
Körber, A. (2007c). Niveaus der Verfügung über einen Quellenbegriff. Eine Skizze der Graduierung einer Einzelkompetenz im Bereich der historischen Sachkompetenzen. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik*. (S. 546–562). Neuried: ars una.
Körber, A. (2010). Kompetenzorientierung versus Inhalte. Eine alte Debatte zu neuem Thema. *Schul-Management, 6*, 8–11.
Körber, A. (2012). Graduierung historischer Kompetenzen. In M. Barricelli & M. Lücke (Hrsg.), *Handbuch Praxis des Geschichtsunterrichts* (S. 236–254). Schwalbach/Ts.: Wochenschau Verlag.
Körber, A. (2013). *Historische Sinnbildungstypen: Weitere Differenzierung.* Zugriff am 09.08.2016. Verfügbar unter http://www.pedocs.de/volltexte/2013/7264/
Körber, A. (2016). Translation and its discontents II. A German perspective. *Journal of Curriculum Studies 48*(4), 440–456. doi: 10.1080/00220272.2016.1171401
Körber, A., Albroscheit, J., Bauer, J.-P., Borries, B. von, Baumgarten, S. & Meyer-Hamme, J. (2007). Sinnvolle Kompetenzorientierung durch Prüfungsvorgaben. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmo-*

dell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik (S. 712–768). Neuried: ars una.

Körber, A., Borries, B. von, Pflüger, C., Schreiber, W. & Ziegler, B. (2008). Sind Kompetenzen historischen Denkens messbar? In V. Frederking (Hrsg.), *Schwer messbare Kompetenzen. Herausforderungen für die empirische Fachdidaktik* (S. 65–84). Baltmannsweiler: Schneider.

Körber, A. & Meyer-Hamme, J. (2007). Ausdifferenzierung und Graduierung der ‚Gattungskompetenz. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 389–412). Neuried: ars una.

Körber, A., Meyer-Hamme, J. & Schreiber, W. (2007). Überlegungen zu Graduierungslogiken der Kernkompetenzen im Kompetenzbereich historische Orientierungskompetenzen. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 473–504). Neuried: ars una.

Körber, A., Schreiber, W. & Schöner, A. (Hrsg.). (2007). *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik.* Neuried: ars una.

Körber, A. & Meyer-Hamme, J. (2015). Historical thinking, competencies and their measurement: Theoretical challenges and testing concepts. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (p. 89–101). New York: Routledge.

Köster, M., Thünemann, H. & Zülsdorf-Kersting, M. (Eds.). (2014). *Researching history education. International perspectives and disciplinary traditions*. Schwalbach/Ts.: Wochenschau Verlag.

Kraus, A. (2013). Kategoriale Inhalts- und Strukturanalyse zur Auswertung von Schüleräußerungen zu Zeitzeugen – Wirksamkeitsforschung für kompetenzorientierten Geschichtsunterricht an Hauptschulen. In W. Schreiber, A. Schöner & F. Sochatzy (Hrsg.), *Analyse von Schulbüchern als Grundlage empirischer Geschichtsdidaktik* (S. 194–211). Stuttgart: Kohlhammer.

Kubinger, K. D. & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependant on different item response formats – An experiment in fundamental research on psychological assessment. *Psychology Science, 49*(4), 361–374.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C. & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*(1), 111–115.

Kühberger, C. (Hrsg.). (2012). *Historisches Wissen. Geschichtsdidaktische Erkundung zu Art, Tiefe und Umfang für das historische Lernen.* Schwalbach/Ts.: Wochenschau Verlag.

Kühberger, C. (2013). *Geschichte denken. Zum Umgang mit Geschichte und Vergangenheit von Schüler/innen der Sekundarstufe I am Beispiel „Spielfilm". Empirische Befunde – Diagnostische Tools – Methodische Hinweise.* Innsbruck: Studien Verlag.

Kühberger, C. (2016). Historische Fragekompetenz in der Primarstufe. In A. Becher, E. Gläser & B. Pleitner (Hrsg.), *Die historische Perspektive konkret. Begleitband 2 zum Perspektivrahmen Sachunterricht* (S. 27–39). Bad Heilbrunn: Klinkhardt.

Kühberger, C. & Mellies, D. (Hrsg.). (2009). *Inventing the EU. Zur De-Konstruktion von „fertigen Geschichten" über die EU in deutschen, polnischen und österreichischen Schulgeschichtsbüchern.* Schwalbach/Ts.: Wochenschau Verlag.

Lange, K. (2011). *Historisches Bildverstehen oder Wie lernen Schüler mit Bildquellen? Ein Beitrag zur geschichtsdidaktischen Lehr-Lern-Forschung* (Geschichtskultur und historisches Lernen, 7). Berlin: LIT.

Lee, P. J. (1983). History teaching and philosophy of history. *History and Theory*, *22*(4), 19–49. doi:10.2307/i342935

Lee, P. J. & Ashby, R. (2000). Progression in historical understanding among students ages 7–14. In P. N. Stearns, P. Seixas & S. Wineburg (Eds.), *Knowing, teaching & learning history. National and international perspectives* (pp. 199–222). New York: New York University Press.

Lee, P. J. (2004). 'Walking backwards into tomorrow': Historical consciousness and understanding history. *International Journal of Historical Learning, Teaching and Research, 4*(1). Retrieved from http://centres.exeter.ac.uk/historyresource/journal7/lee.pdf

Lee, P. J. (2005a). Historical literacy: theory and research. *International Journal of Historical Learning, Teaching, and Research, 5*(2). Retrieved from http://centres.exeter.ac.uk/historyresource/ journal9/papers/lee.pdf

Lee, P. J. (2005b). Putting principles into practice: Understanding history. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: History, mathematics, and science in the classroom* (pp. 31–77). Washington, D.C.: National Academies Press.

Lee, P. J. (2014). Fused horizons? UK research into students' second-order ideas in history: A perspective from London. In M. Köster, H. Thünemann & M. Zülsdorf-Kersting (Eds.), *Researching history education. International perspectives and disciplinary traditions* (pp. 170–194). Schwalbach Ts.: Wochenschau Verlag.

Lee, P. J. & Shemilt, D. (2003). A scaffold, not a cage: Progression and progression models in history. *Teaching History, 113,* 13–23.

Lehmann, K. (2015). Lernaufgaben zur Förderung historischer Kompetenzen mittels historischer Theaterarbeit. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13"* (S. 205–215). Bern: hep.

Lehmann, K., Werner, M. & Zabold, S. (2016). *Historisches Denken jetzt und in Zukunft: Wege zu einem theoretisch fundierten und evidenzbasierten Umgang mit Geschichte; Festschrift für Waltraud Schreiber zum 60. Geburtstag.* Berlin: LIT.

Lévesque, S. (2009). *Thinking historically: Educating students for the twenty-first century.* Toronto: University of Toronto Press.

Lévesque, S. (2016). Why should historical thinking matter to students? *Agora, 51*(2), 4–8.

Lingelbach, G. & Rudolph, H. (2005). *Geschichte studieren. Eine praxisorientierte Einführung für Historiker von der Immatrikulation bis zum Berufseinstieg.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Logtenberg, A. (2012). *Questioning the past: student questioning and historical reasoning.* Doctoral dissertation, University of Amsterdam. Retrieved from http://dare.uva.nl/document/2/105943

Lord, F. M. & Novick, M. R. (Eds.). (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maggioni, L. (2010). *Studying epistemic cognition in the history classroom: Cases of teaching and learning to think historically*. Doctoral dissertation, University of Maryland. Retrieved from http://drum.lib.umd.edu/bitstream/1903/10797/1/Maggioni_umd_0117E_11443.pdf

Maggioni, L., Alexander, P. A. & VanSledright, B. (2004). At a crossroads? The development of epistemological beliefs and historical thinking. *European Journal of School Psychology, 2*(1–2), 169–197.

Maggioni, L., VanSledright, B. & Alexander, P. A. (2009). Walking on the borders: A measure of epistemic cognition in history. *The Journal of Experimental Education, 77*(3), 187–213. doi: 10.3200/JEXE.77.3.187-214

Mandell, N. (2008). Thinking like a historian: A framework for teaching and learning. *OAH Magazine of History, 22,* 55–63.

Martens, M. (2010). *Implizites Wissen und kompetentes Handeln. Die empirische Rekonstruktion von Kompetenzen historischen Verstehens im Umgang mit Darstellungen von Geschichte.* Göttingen: V&R unipress.

Marz, F., Arnold, R. & Reischmann, J. (1978). *Lernkontrollen im politischen Unterricht.* Stuttgart: Klett.

Mayer, U. (2014). Keine Angst vor Kompetenzen: Kompetenzorientierung – eine typologische, historische und systematische Einordnung. *Geschichte für heute, 7*(3), 6–19.

McKelvey, R. D. & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology, 4*(1), 103–120. doi: 10.1080/0022250X.1975.9989847

Meis, F. & Zuckowski, A. (im Druck). Zur Kompetenz historischen Denkens angehender Geschichtslehrerinnen und -lehrer. Quantitative Befunde eines Extremgruppenvergleichs mit Schülerinnen und Schülern. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 15. Beiträge zur Tagung „geschichtsdidaktisch empirisch 15"*. Bern: hep.

Merkt, M., Werner, W. & Wagner, W. (in press). Historical thinking skills and mastery of multiple document tasks. *Learning and Individual Differences.*

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Meyer-Hamme, J. (2005). Schulbuchverständnis und Schulbuchvergleich zu Bonifatius II – Erhebung mittels Kurzessays. In B. von Borries, C. Fischer, S. Leutner-Ramme & J. Meyer-Hamme (Hrsg.), *Schulbuchverständnis, Richtlinienbenutzung und Reflexionsprozesse im Geschichtsunterricht. Eine qualitativ-quantitative Schüler- und Lehrerbefragung im deutschsprachigen Bildungswesen 2002* (S. 121–157). Neuried: ars una.

Meyer-Hamme, J. (2007). Historische Kompetenzen empirisch. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 674–693). Neuried: ars una.

Meyer-Hamme, J. (2009). *Historische Identitäten und Geschichtsunterricht. Fallstudien zum Verhältnis von kultureller Zugehörigkeit, schulischen Anforderungen und individueller Verarbeitung.* (Schriften zur Geschichtsdidaktik, 26). Idstein: Schulz-Kirchner.

Meyer-Hamme, J. (2014). Geschichtskultur: Extremfall Holocaust Comics. In B. von Borries (Hrsg.), *Zwischen „Genuss" und „Ekel". Ästhetik und Emotionalität als konstitutive Merkmale historischen Lernens* (S. 92–127). Schwalbach/Ts.: Wochenschau Verlag.

Meyer-Hamme, J. (2015). Formate geschlossener Aufgaben zur Messung von Kompetenzen historischen Denkens. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13"* (S. 139–152). Bern: hep.

Meyer-Hamme, J. (2016). Im Spannungsfeld historischer Uneindeutigkeit, notwendiger Exaktheit und sozialer Erwünschtheit. Eine Re-Analyse von Fragebogen- und Testkonstruktionen in quantitativen Studien zum Geschichtsbewusstsein und historischem Lernen. In H. Thünemann & M. Zülsdorf-Kersting (Hrsg.), *Methoden geschichtsdidaktischer Unterrichtsforschung* (S. 89–113). Schwalbach/Ts.: Wochenschau Verlag.

Meyer-Hamme, J. & Körber, A. (Hrsg.). (in Vorbereitung). Erfassung von Kompetenzen historischen Denkens: Teilergebnisse aus und Reflexionen zum HiTCH-Projekt (2012–2015) [Arbeitstitel].

Mielitz, R. (1969). Das Faktenwissen der Studienanfänger: Ergebnisse eines Tests. In R. Mielitz (Hrsg.), *Das Lehren der Geschichte. Methoden des Geschichtsunterrichts in Schule und Universität* (S. 90–102). Göttingen: Vandenhoeck & Ruprecht.

Mierwald, M. & Brauch, N. (2015). Historisches Argumentieren als Ausdruck historischen Denkens. *Zeitschrift für Geschichtsdidaktik*, *14*(1), 104–120.

Mierwald, M., Seiffert, J., Lehmann, T. & Brauch, N. (im Druck). Fragebögen auf dem Prüfstand. Ein Beitrag zur Erforschung und Weiterentwicklung eines bestehenden Instruments zur Erfassung epistemologischer Überzeugungen in der Domäne Geschichte. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 15. Beiträge zur Tagung „geschichtsdidaktik empirisch 15"*. Bern: hep.

Monte-Sano, C. & Reisman, A. (2016). Studying historical understanding. In L. Corno & E. M. Anderman (Eds.), *Handbook of educational psychology* (pp. 280–294). New York: Routledge.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion* (2. Aufl.). Berlin: Springer.

National Center for History in the Schools UCLA (1996). *National standards for history. Basic edition.* Retrieved from http://www.nchs.ucla.edu/history-standards/historical-thinking-standards

Nitsche, M. & Waldis, M. (2016). Narrative Kompetenz von Studierenden erfassen – Zur Annäherung an formative und summative Vorgehensweisen im Fach Geschichte. *Zeitschrift für Didaktik der Gesellschaftswissenschaften, 7*(1), 17–35.

Nokes, J. D. (2010a). Observing literacy practices in history classrooms. *Theory and Research in Social Education, 38*(4), 298–316.

Nokes, J. D. (2010b). (Re)Imagining literacies for history classrooms. In R. J. Draper (Ed.), *(Re)imagining content-area literacy instruction* (pp. 54–68). New York: Teachers College Press.

Nokes, J. D. (2011). *Historical literacy.* Retrieved from http://www.slcschools.org/departments/curriculum/ social-studies/documents/Historical-Literacy.pdf

OECD (2014). *PISA 2012. Technical report*. Paris: OECD.

Oehler, H. (1969). Geschichtswissen und Geschichtsbild der Abiturienten. In R. Mielitz (Hrsg.), *Das Lehren der Geschichte. Methoden des Geschichtsunterrichts in Schule und Universität* (S. 46–55). Göttingen: Vandenhoeck & Ruprecht.

Pallaske, C. (19. Dezember 2015). *Was wäre wenn: Eine Geschichtsdidaktik ohne Kompetenzbegriff?* Zugriff am 11. Oktober 2016. Verfügbar unter http://historischdenken.hypotheses.org/3099

Pandel, H.-J. (2005). *Geschichtsunterricht nach PISA. Kompetenzen, Bildungsstandards und Kerncurricula.* Schwalbach/Ts.: Wochenschau Verlag.

Pellegrino, J. W., Chudowsky, N. & Glaser, R. (Eds.). (2001). *Knowing what students know. The science and design of educational assessment*. Washington D.C.: National Academic Press.

Perfetti, C. A., Britt, M. A. & Georgi, M. C. (1995). *Text-based learning and reasoning: Studies in history*. New York: Routledge.

Pohl, S. & Carstensen, C. H. (2012). *NEPS technical report – scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität.

Pohl, S. & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online, 5*(2), 189–216.
Pohl, S., Gräfe, L. & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74*(3), 423–452.
Ricoeur, P. (1988). *Zeit und Erzählung, Bd. 1: Zeit und historische Erzählung*. München: Fink. (Erstausgabe 1983. Temps et récit. Paris: Éd. du Seuil)
Rogers, P. J. (1980). *The new history: Theory into practice*. London: Historical Association.
Rose, N. (2013). *Item nonresponses in educational and psychological measurement*. Unpublished doctoral dissertation, Friedrich-Schiller-University of Jena.
Rose, N., Davier, M. von & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Rep. no. RR-10-11), Princeton, NJ: Educational Testing Service.
Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2., überarb. u. erw. Aufl.). Bern: Huber.
Röttgers, K. (1982). *Der kommunikative Text und die Zeitstruktur von Geschichten*. München: Alber.
Rouet, J. F., Britt, M. A., Mason, R. A. & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology*, *88*(3), 478–493. doi: 10.1037/0022-0663.88.3.478
Rouet, J. F., Favart, M., Britt, M. A. & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction*, *15*(1), 85–106.
Rüsen, J. (1982). Die vier Typen des historischen Erzählens. In R. Koselleck, H. Lutz & J. Rüsen (Hrsg.), *Formen der Geschichtsschreibung* (S. 514–606). (Theorie der Geschichte. Beiträge zur Historik, 4). München: Deutscher Taschenbuch Verlag.
Rüsen, J. (1983). *Historische Vernunft. Die Grundlagen der Geschichtswissenschaft. Grundzüge einer Historik I*. Göttingen: Vandenhoeck & Ruprecht.
Rüsen, J. (1986). *Rekonstruktion der Vergangenheit. Die Prinzipien der historischen Forschung*. (Kleine Vandenhoeck-Reihe, 1515). Göttingen: Vandenhoeck & Ruprecht.
Rüsen, J. (1989). *Lebendige Geschichte. Grundzüge einer Historik III: Formen und Funktionen des historischen Wissens*. (Kleine Vandenhoeck-Reihe, 1489). Göttingen: Vandenhoeck & Ruprecht.
Rüsen, J. (1994). *Historische Orientierung. Über die Art des Geschichtsbewusstseins, sich in der Zeit zurechtzufinden*. Köln: Böhlau.
Rüsen, J. (2001). Historisches Erzählen. In J. Rüsen (Hrsg.), *Zerbrechende Zeit. Über den Sinn der Geschichte* (S. 43–106). Köln: Böhlau.
Rüsen, J. (2005). *History. Narration—Interpretation—Orientation*. New York: Berghahn Books.
Rüsen, J. (2008). *Historische Orientierung. Über die Arbeit des Geschichtsbewußtseins, sich in der Zeit zurechtzufinden* (2. Aufl.). (Klassiker der Geschichtsdidaktik). Schwalbach/Ts.: Wochenschau Verlag.
Rüsen, J. (2013). *Historik. Theorie der Geschichtswissenschaft*. Köln: Böhlau.
Sachse, M. (2005). *Fächer ohne Bildungsstandards – Fächer zweiter Güte*? München: Staatsinstitut für Schulqualität und Bildungsforschung. Zugriff am 06.07.2015. Verfügbar unter http://www.kompas.bayern.de/userfiles/Faecher_ohne_BS.doc
Sauer, M. (2006). Kompetenzen für den Geschichtsunterricht – ein pragmatisches Modell als Basis für die Bildungsstandards des Verbandes der Geschichtslehrer. *Informationen für den Geschichts- und Gemeinschaftskundelehrer, 74*, 7–20.

Sahm, F. (2015). *Eine Extremgruppenbefragung bei Studierenden zur Validierung eines Kompetenztests historischen Denkens. Eine Analyse ausgewählter Aufgaben der HiTCH-Pilotstudie 2013.* Unveröffentlichte Masterarbeit, Universität Hamburg.

Schiele, S. & Schneider, H. (Hrsg.). (1977). *Das Konsensproblem in der politischen Bildung*. Stuttgart: Klett.

Schneider, G. (2010). Die Arbeit mit schriftlichen Quellen. In H.-J. Pandel & G. Schneider (Hrsg.), *Handbuch Medien im Geschichtsunterricht* (5. Aufl.). Schwalbach/Ts.: Wochenschau Verlag.

Schönemann, B., Thünemann, H. & Zülsdorf-Kersting, M. (2010). *Was können Abiturienten? Zugleich ein Beitrag zur Debatte über Kompetenzen und Standards im Fach Geschichte.* Berlin: LIT.

Schöner, A. (2007). Kompetenzbereich historische Sachkompetenzen. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 265–314). Neuried: ars una.

Schreiber, W. (2002). Reflektiertes und (selbst-)reflexives Geschichtsbewußtsein durch Geschichtsunterricht fördern – ein vielschichtiges Forschungsfeld der Geschichtsdidaktik. *Zeitschrift für Geschichtsdidaktik, 1*, 18–43.

Schreiber, W. (2007). Kompetenzbereich historische Methodenkompetenzen. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 195–264). Neuried: ars una.

Schreiber, W. (2012). Zum Verhältnis zwischen Wissen und Kompetenzen. Ein Essay. In C. Kühberger (Hrsg.), *Historisches Wissen. Geschichtsdidaktische Erkundungen über Art, Umfang und Tiefe für das historische Lernen* (S. 119–134). Schwalbach/Ts.: Wochenschau Verlag

Schreiber, W. & Árkossy, K. (2009). *Zeitzeugengespräche führen und auswerten. Historische Kompetenzen schulen.* Neuried: ars una.

Schreiber, W. & Gruner, C. (Hrsg.). (2010). *Geschichte durchdenken. Schüler de-konstruieren internationale Schulbücher. Das Beispiel „1989/1990 – Mauerfall".* Neuried: ars una.

Schreiber, W., Körber, A., Borries, B. von, Krammer, R., Leutner-Ramme, S., Mebus, S., Schöner, A. & Ziegler, B. (2006). *Historisches Denken. Ein Kompetenz-Strukturmodell.* Neuried: ars una.

Schreiber, W., Kraus, A., Lehmann, K. & Zabold, S. (2015). Empirische Erforschung von Ausprägungen historischer Kompetenzen in museumspädagogischen Programmen – Die Fallstudie „Erinnern und Gedenken" zur Europaratsausstellung „Verführung Freiheit" am Deutschen Historischen Museum in Berlin. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13"* (S. 182–192). Bern: hep.

Schreiber, W. & Mebus, S. (2005). *Geschichte denken statt pauken.* Meißen: Sächsische Akademie für Lehrerfortbildung.

Schreiber, W., Schöner, A. & Sochatzy, F. (Hrsg.). (2012). *Analyse von Schulbüchern als Grundlage empirischer Geschichtsdidaktik.* Stuttgart: Kohlhammer.

Schreiber, W., Sochatzy, F. & Ventzke, M. (2014). *Auf dem Weg zu digital-multimedialen Lehr- und Lernmitteln für kompetenzorientiertes inklusives Unterrichten und Lernen.* Online Publikation, Medienberatung NRW. Zugriff am 11.02.2017. Verfügbar unter: http://www.medienberatung.schulministerium.nrw.de/Medienberatung-NRW/Dokumentationen/2014/140625_Symposium-Kriterien-f%C3%BCr-Lernmittel-im-Gemeinsamen-Unterricht/Schreiber_multimediale-Lernmittel-f%C3%BCr-inklusives-Lernen-und-Lehren.pdf

Schroeder, K., Deutz-Schroeder, M., Quasten, R. & Schulze Heuling, D. (2012*). Später Sieg der Diktaturen. Zeitgeschichtliche Kenntnisse und Urteile von Jugendlichen.* Frankfurt am Main: Peter Lang.
Seixas, P. (Ed.). (2004). *Theorizing historical consciousness*. Toronto: University of Toronto Press.
Seixas, P. (2008). *"Scaling Up" the benchmarks of historical thinking.* A report on the Vancouver meetings. February 14–15. Retrieved from http://historicalthinking.ca/sites/default/files/Scaling%20Up%20Meeting%20Report.pdf
Seixas, P. (Ed.). (2011). *Theorizing historical consciousness* (2nd. ed.). Toronto: University of Toronto Press.
Seixas, P. (2015). Translations and its discontents: key concepts in English and German history education. *Journal of Curriculum Studies, 48*, 427–439.
Seixas, P. (2016). A history/memory matrix for history education. *Public History Weekly*, *4,* 6.
Seixas, P. & Ercikan, K. (2015). Introduction: The new shape of history assessment. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 1–14). New York: Routledge.
Seixas, P., Gibson, L. & Ercikan, K. (2015). A design process for assessing historical thinking. The case of a one-hour test. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 102–116). New York: Routledge.
Seixas, P. & Morton, T. (2013). *The big six historical thinking concepts*. Retrieved from http://www.nelson.com/thebigsix/documents/The%20Big%20Six%20Sample%20Chapter%20with%20BLM_Aug%2030.pdf
Siebeck, C. (2013). Später Sieg des Kalten Krieges. *Gedenkstättenrundbrief, 169*, 44–54.
Smith, M. & Breakstone, J. (2015). History assessments of thinking. An investigation of cognitive validity. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 233–245). New York: Routledge.
Sochatzy, F. & Merkt, M. (2015). Kompetenzförderung durch ein multimediales Filmmethodentraining. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13"* (S. 193–204). Bern: hep.
Sochatzy, F. (2016). *Das multimediale Schulbuch (mBook) – von der Theorie in die Praxis: Konzeption, Produktion und empirische Überprüfung eines multimedialen Geschichtsschulbuchs*. Eichstätt: Institut für digitales Lernen.
Stearns, P. N. (1998). Why study history. *American Historical Association*. Retrieved from http://www.historians.org/about-aha-and-membership/aha-history-and-archives/archives/why-study-history-%281998%29
Stoel, G. L., van Drie, J. P. & van Boxtel, C. A. M. (2015). Teaching towards historical expertise. Developing a pedagogy for fostering causal reasoning in history. *Journal of Curriculum Studies*, *47*(1), 49–76.
Stradling, R. (2004). *Multiperspectivity in history teaching. A guide for teachers.* Strasbourg: Council of Europe. Retrieved from http://tandis.odihr.pl/documents/hre-compendium/CD%20SEC%202%20ENV/PARTNERS%20RESOURCES/CoE%20Multiperspectivity%20in%20history%20teaching%20ENG.pdf
Taylor, T. & Young, C. (2003). *Historical literacy. Making history: a guide for the teaching and learning of history in Australian schools.* Retrieved from http://www.hyperhistory.org/images/assets/pdf/complete.pdf
Van Drie, J. & van Boxtel, C. (2008). Historical reasoning: Towards a framework for analyzing students' reasoning about the past. *Educational Psychology Review 20*(2), 87–110. doi: 10.1007/s10648-007-9056-1

VanSledright, B. A. (2014). *Assessing historical thinking & understanding. Innovative designs for new standards*. New York: Routledge.
VanSledright, B. A. & Reddy, K. (2014). Changing epistemic beliefs? An exploratory study of cognition among prospective history teacher. *Tempo e Argumento*, *6*(11), 28–68.
Ventzke, M. (2012). Begriffliches Arbeiten und „Geschichte denken" – theoretische Voraussetzungen und unterrichtliche Vorgehensweisen. In C. Kühberger (Hrsg.), *Historisches Wissen. Geschichtsdidaktische Erkundungen über Art, Umfang und Tiefe für das historische Lernen* (S. 75–102). Schwalbach/Ts.: Wochenschau Verlag.
Ventzke, M. (2016). *Temporal turn – Grundlagen historischer Zeitanalysen im Prozess kompetenzorientierten Geschichtsdenkens.* Habilitationsschrift, Katholische Universität Eichstätt-Ingolstadt.
Ventzke, M., Sochatzy, F. & Schreiber, W. (Hrsg.). (2013). *mBook Geschichte Bd. 1 bis 5 für die Oberstufe des Gymnasiums in der Deutschsprachigen Gemeinschaft Belgiens*. Eichstätt: Institut für digitales Lernen.
Ventzke, M., Sochatzy, F. & Schreiber, W. (Hrsg.). (2014). *mBook Geschichte Bd. 1 bis 3 für die Sekundarstufe I des Gymnasiums NRW*. Düsseldorf: Medienberatung.
Verband der Geschichtslehrer Deutschlands (Hrsg.). (2007). *Bildungsstandards Geschichte. Sekundarstufe I. Rahmenmodell Gymnasium. 5.–10. Jahrgangsstufe*. Schwalbach/Ts.: Wochenschau Verlag. Zugriff am 23.06.2015. Verfügbar unter http://www.geschichtslehrerverband.org/fileadmin/images/pdf/bildungsstandards.pdf
Verband der Geschichtslehrer Deutschlands (Hrsg.). (2011). *Bildungsstandards Geschichte.* (Fassung mit erster Überarbeitung). Zugriff am 22.12.2015. Verfügbar unter http://www.geschichtslehrerverband.org/fileadmin/images/Bildungsstandards/Druckfassung/Standards_Druckformat__10.5.2011_.pdf
Wang, W.-C. & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149. doi: 10.1177/0146621604271053
Waldis, M. (2013). Fachdidaktische Analysen von Aufgaben im Fach Geschichte. In M. Kleinknecht, T. Bohl, U. Maier & K. Metz (Hrsg.), *Lern- und Leistungsaufgaben im Unterricht. Fächerübergreifende Kriterien zur Auswahl und Analyse* (S. 145–162). Heilbronn: Klinkhardt.
Waldis, M. (2016). Erzählung oder Argumentation? Zum Einfluss von Textgenre, Aufgabenprompt und Materialauswahl auf das historische Erzählen. In S. Keller & C. Reintjes (Hrsg.), *Aufgaben als Schlüssel zur Kompetenz. Didaktische Herausforderungen, wissenschaftliche Zugänge, empirische Befunde* (S. 237–260). Münster: Waxmann.
Waldis, M., Hodel, J., Thünemann, H., Zülsdorf-Kersting, M. & Ziegler, B. (2015). Material-based and open-ended writing tasks to assess narrative competence among students. In P. Seixas & K. Ercikan (Eds.), *New directions in assessing historical thinking* (pp. 119–133). New York: Routledge.
Waldis, M., Marti, P. & Nitsche, M. (2015). Angehende Geschichtslehrpersonen schreiben Geschichte(n) – Zur Kontextabhängigkeit der Erfassung narrativer Kompetenz. *Zeitschrift für Geschichtsdidaktik, 14,* 63–86.
Waldis, M. & Ziegler, B. (Hrsg.). (im Druck). *Forschungswerkstatt Geschichtsdidaktik 15. Beiträge zur Tagung „geschichtsdidaktik empirisch 15".* Bern: hep.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. doi: 10.1007/BF02294627
Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.
Werner, M. & Schreiber, W. (2015). Testfragen befragen – Pretesting und Optimierung des Large-Scale-Kompetenztests „HITCH" durch Cognitive Labs. In M. Waldis & B.

Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13“* (S. 153–164). Bern: hep.
Westhoff, L. M. (2009). Lost in translation. The use of primary sources in teaching history. In R. G. Ragland & K. A. Woestman (Eds.), *The teaching american history project: Lessons for history educators and historians* (pp. 62–76). New York: Routledge.
Willis, G. B. (2005). *Cognitive interviewing. A tool for improving questionnaire design.* Thousand Oaks, CA: SAGE.
Wineburg, S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documetary and pictorial evidence. *Journal of Educational Psychology*, *83*(1), 73–87.
Wineburg, S. (2001). *Historical thinking and other unnatural acts: Charting the future of teaching the past*. Philadelphia: Temple University Press.
Wineburg, S., Martin, D. & Monte-Sano, C. (2013). *Reading like a historian. Teaching literacy in middle and high school history classrooms*. New York: Teachers College Press.
Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.
Wu, M. L. & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*(4), 339–355.
Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest version 2.0: generalized item response modelling software.* Camberwell, AUS: ACER Press.
Ziegler, B. (2007). Die Graduierung der Re-Konstruktionskompetenz. In A. Körber, W. Schreiber & A. Schöner (Hrsg.), *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik* (S. 523–545). Neuried: ars una.
Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa: Directorate of Human Resources Research and Evaluation, National Defense Headquarters.

Appendix

Abbildung 9 zeigt exemplarisch ein Nested-Factor-Modell für einen Test bestehend aus vier Aufgaben mit jeweils zwei Items (Indikatoren), die historische Kompetenz anhand spezifischer Kompetenzfacetten (A und B) erfassen. Die unstandardisierten Faktorladungen sind im Sinne eines 1PL-Modells auf den Wert $\lambda = 1$ fixiert, so dass lediglich Varianzen der latenten Variablen geschätzt werden (gekennzeichnet durch den Doppelpfeil an den durch Kreise dargestellten latenten Variablen). Die Indikatoren ($y_{111} - y_{242}$) stehen hier für kategoriale Daten (z.B. als richtig/falsch kodierte Items) und werden entsprechend ohne Residualvarianzen dargestellt.

Bei kategorialen Variablen im IRT-Kontext wird eine diesen Daten zugrundeliegende latente kontinuierliche Variable angenommen, deren Kategorisierung (durch Unterteilung des Kontinuums an bestimmten Stellen) in der jeweiligen beobachteten kategorialen Variable resultiert. Damit lassen sich IRT-Modelle in Strukturgleichungsmodelle überführen (Glockner-Rist & Hoijtink, 2003), die komplexe Modellierungen erlauben (z.B. faktorenanalytische Verfahren). Die Varianz der zugrundeliegenden latenten Variablen ergibt sich aus den Varianzanteilen, die auf Dimensionen (Faktoren) zurückgeführt werden, und einer „Residualvarianzkomponente", die sich aus der verwendeten Link-Funktion ergibt (Probit-Link: $Var = 1$; Logit-Link: $Var = \pi^2/3$). Die durch Faktoren erklärten Varianzanteile entsprechen – analog zu Faktorenanalysen mit kontinuierlichen Daten – dem Produkt der quadrierten unstandardisierten Faktorladungen und der Faktorvarianz.

Auf Basis dieser Varianzkomponenten lassen sich standardisierte Ladungen als die Wurzel aus dem relativen Varianzanteil, der auf einen bestimmten Faktor zurückgeführt werden kann, bezogen auf die Gesamtvarianz (also sämtliche faktorbezogenen Varianzkomponenten sowie die Residualvarianz) errechnen. Für die beiden Items der ersten Aufgabe in Abbildung 9 (y_{111} und y_{112}) lässt sich beispielsweise die standardisierte Ladung ($\lambda_{hK,\,std}$; unstandardisierte Ladungen jeweils auf $\lambda = 1$ fixiert) auf den Faktor historische Kompetenz (Varianz: Φ_{hK}) bei Verwendung des Probit-Links (Residualvarianz: $\sigma_\varepsilon^2 = 1$) wie folgt berechnen (mit Φ_{FA} bzw. Φ_{A1} für die Varianzen der spezifischen historischen Kompetenz A bzw. der aufgabenspezifischen Faktoren):

$$\lambda_{hK,\,std} = \sqrt{\frac{\lambda_{hK}{}^2\phi_{hK}}{\lambda_{hK}{}^2\phi_{hK} + \lambda_{FA}{}^2\phi_{FA} + \lambda_{A1}{}^2\phi_{A1} + \sigma_\varepsilon^2}} = \sqrt{\frac{\phi_{hK}}{\phi_{hK} + \phi_{FA} + \phi_{A1} + 1}}$$

Die quadrierten standardisierten Ladungen können also im Sinne prozentualer Varianzanteile interpretiert werden, die auf den jeweiligen Faktor zurückführbar sind.[32]

32 Dies entspricht einer Interpretation analog zum Pseudo-R^2 nach McKelvey und Zavoina (1975), wobei hier latente Variablen (Faktoren) als Prädiktoren in einem Regressionsmodell (logistische Regression bzw. Probit-Regression) mit dichotomen bzw. ordinalen abhängigen Variablen betrachtet werden.