Erschienen in: Methoden geschichtsdidaktischer Unterrichtsforschung / Thünemann, Holger; Zülsdorf-Kersting, Meik (Hrsg.). - Schwalbach/Ts. : Wochenschau Verlag, 2016. - (Geschichtsunterricht erforschen ; 5). - S. 63-88. - ISBN 978-3-7344-0212-8

Christiane Bertram

Entwicklung standardisierter Testinstrumente zur Erfassung der Wirksamkeit von Geschichtsunterricht[1]

1. Einführung

Die Frage, wie nachhaltige Lernprozesse bei Schülerinnen und Schülern angestoßen werden können, steht im Kern jeder (fachdidaktischen) Unterrichtsforschung. Empirisch ist vielfach belegt, dass weniger die Organisationsformen (wie z. B. die Klassengröße) einen Einfluss auf das Lernen (bzw. den Lernerfolg) und die Entwicklung der Schülerinnen und Schüler haben (Hattie 2009; Seidel/Shavelson 2009), sondern die Tiefenstrukturen des Unterrichts, das heißt die im Unterricht stattfindenden Lehr-Lernprozesse (Kunter/Trautwein 2013).[2] Allerdings bietet „guter" Unterricht, der Lernprozesse anstoßen und verstetigen könnte, noch keine Garantie für die Lernerfolge der Schülerinnen und Schüler. Im „Angebot-Nutzungsmodell" (Helmke 2012, 73) wird Unterricht als ein Angebot verstanden, das von den Lernenden genutzt werden kann (oder auch nicht). Kulturelle und institutionelle Rahmenbedingungen, Voraussetzungen der Lehrkräfte und der Lernenden wie auch der Klassenkontext haben einen Einfluss auf die Akzeptanz des Unterrichts.

Unterrichtsqualität lässt sich daran festmachen, inwieweit es Lehrkräften gelingt, bei den Lernenden Lernprozesse anzustoßen und aufrechtzuerhalten, die zu den gewünschten Lerneffekten führen. Die Bewertung des Unterrichts kann normativ erfolgen, indem Wertvorstellungen, wie Unterricht

1 Im Rahmen eines vom BMBF (Bundesministerium für Bildung und Forschung, Projekt-Nummer: 01JG0913) geförderten Dissertationsprojekts wurden Testinstrumente zur Erfassung der Wirksamkeit einer Unterrichtsintervention entwickelt.

2 Kunter und Trautwein (2013) haben in ihrer Einführung zur „Psychologie des Unterrichts" den Beitrag der Psychologie und empirischen Bildungsforschung bei der Theoriebildung des guten (= gelingenden) Unterrichts wie auch hinsichtlich der Forschungsmethoden auf Basis der aktuellen Unterrichtsforschung zusammengefasst. Der vorliegende Beitrag schließt in den einleitenden Ausführungen an diese Überlegungen an.

zu sein hat, vorgegeben werden (Berliner 2005, 207). Zum Beispiel wird im Fach Geschichte als Norm gesetzt, dass die Lernenden die Bedeutung der Geschichte für ihre eigene Gegenwart reflektieren. Daher wird in einer Lehrprobenstunde als Pluspunkt gewertet, wenn von dem im Unterricht behandelten historischen Thema eine Brücke in die Gegenwart geschlagen wird. Neben der normativen Setzung kann die Qualität des Unterrichts daran festgemacht werden, ob die gewählten Methoden oder Strategien zu den erwünschten Effekten führen (Berliner 2005; Ditton 2006). Für das obige Beispiel bedeutet dies, dass – beispielsweise in einem Test – in den Blick genommen wird, ob die Lernenden die Bedeutung eines historischen Themas tatsächlich im Hinblick auf die eigene Gegenwart bedenken. „Qualitive teaching" im Verständnis von Berliner (2005) liegt vor, wenn beide Perspektiven miteinander verknüpft werden, wenn also die Zielsetzung des Unterrichts sinnvoll und wünschenswert ist und gleichzeitig die gewünschten Wirkungen zeigt. In unserem Beispiel würde das bedeuten, dass die Lernenden aufgrund einer im Unterricht vorgenommen „Aktualisierung" das Thema der Stunde in der Bedeutung für ihre eigene Gegenwart reflektieren.

Wenn qualitätsvoller Geschichtsunterricht empirisch untersucht wird, tragen die Fachdidaktik Geschichte und die empirische Bildungsforschung spezifische Anteile bei. Basierend auf einer Theorie des historischen Denkens formulieren Fachdidaktiker/innen Lernziele, Einsichten oder Kompetenzen, die die Lernenden in der Beschäftigung mit dem Fach erreichen sollen. Zudem werden Lern- und Unterrichtsmethoden entwickelt, von denen man annimmt, dass damit die gewünschten Wirkungen bestmöglich erreicht werden können. Theoretische Überlegungen wie auch praktische Erfahrungen spielen hierbei eine wesentliche Rolle. Von der empirischen Bildungsforschung hingegen kommen Impulse zur Frage, welche Faktoren der Unterrichtsgestaltung zu gelingendem Unterricht führen können. Zudem stellt sie ein seit Jahren erprobtes Methodenarsenal zur Verfügung, um die Wirksamkeit des Unterrichts im Hinblick auf die avisierten Ziele empirisch zu überprüfen.

Will man nicht nur Aussagen über den jeweils beobachteten Unterricht treffen, sondern die Wirksamkeit bestimmter Methoden oder Unterrichtsformen grundsätzlich in den Blick nehmen, kommt die Prüfung mittels quantitativer, statistischer Verfahren ins Spiel. Die Überprüfung, ob bestimmte Beobachtungen in einer Stichprobe auch in der Gesamtpopulation gelten (z.B. ob aus den Ergebnissen zur Wirksamkeit einer speziellen Unterrichtsmethode in einer Stichprobe von neunten Klassen des Gymnasiums im Regierungspräsidium Tübingen auf die Effektivität dieser Methode bei allen Gymnasialschülerinnen und -schülern der neunten Klassenstufe in Baden-Württemberg

geschlossen werden kann), gehört in den Bereich der „Inferenzstatistik" (= "schließende Statistik"). Diese Aussagen sind allerdings immer zu einen gewissen Grade mit „Unsicherheit" behaftet. Statistische Tests können beispielsweise Auskunft darüber geben, wie wahrscheinlich es ist, dass ein bestimmter Unterschied zwischen zwei unterschiedlichen Unterrichtsmethoden „per Zufall" auftritt, obwohl beide Methoden in Wirklichkeit gleich effektiv sind. Wenn mit einem statistischen Test nachgewiesen wird, dass die Zufallswahrscheinlichkeit gering ist und unterhalb eines festgesetzten (Signifikanz)Niveaus liegt, ist das Testergebnis „statistisch signifikant".

Da bei inferenzstatistischen Analysen meist mit relativ großen Stichproben (oftmals mehrere hundert Schülerinnen und Schüler) gearbeitet werden muss, ist man auf standardisierte Instrumente angewiesen, die die Wirkung von Unterricht objektiv, reliabel, valide und (zeit)ökonomisch messen. Bevor das Vorgehen bei der Entwicklung von Messinstrumenten detailliert beschrieben wird, wird im Folgenden vorgestellt, welche Wirkfaktoren Geschichtsunterricht gelingen lassen und welche Ziele hiermit erreicht werden sollen.

2. Wirkfaktoren und Ziele gelingenden Geschichtsunterrichts

Was sind die Wirkfaktoren, die Unterricht gelingen lassen? Eine Vielzahl von Merkmalen „guten Unterrichts", die im Zusammenhang mit günstigen Schülerergebnissen stehen, wurden in den letzten Jahren empirisch bestätigt (Hattie 2009; Helmke 2012; Seidel/Shavelson 2007) und in Form von Listen zusammengestellt (z.B. Brophy/Good 1986; Helmke 2012; Lipowsky 2009; Meyer 2004). Diese Merkmale lassen sich nach den Befunden der empirischen Unterrichtsforschung drei Dimensionen des Unterrichts zuordnen: der kognitiven Aktivierung, der Klassenführung und der konstruktiven Unterstützung (Klieme u.a. 2001; Klieme 2006; Klieme/Rakoczy 2008). Diese Tiefenstrukturen gelten für jeden Unterricht, ungeachtet der Fachdisziplin.

Die Ziele des Geschichtsunterrichts hingegen werden fachdidaktisch definiert. Gemeinhin sollen im Geschichtsunterricht historische Kompetenzen und themenspezifisches Wissen vermitrelt werden. Häufig wird das Verhältnis zwischen „Kompetenzen" und „Wissen" in der Geschichtsdidaktik diskutiert. Unter „Kompetenzen" werden seit der sogenannten „Klieme-Expertise" (Klieme u.a. 2003) meist „Fähigkeiten, Fertigkeiten und Bereirschaften" (72) verstanden. Für das Fach Geschichte bedeutet dies „die Fähigkeit, Fertigkeit und Bereitschaft, historisch zu denken" (Schreiber u.a. 2007, 17). Unter „Wissen" hingegen wird eher ein Daten- und Faktenkanon als Minimalanforderung für bestimmte Themen verstanden (Kühberger 2012).

Will man in standardisierten Messinstrumenten erfassen, ob und wodurch der Geschichtsunterricht die vorab gesetzten Ziele erreicht, dann müssen zum einen die drei Tiefendimensionen des Unterrichts in den Blick genommen werden, zum anderen die Ergebnisse hinsichtlich des erworbenen Wissens und der vermittelten Kompetenzen. Um die Tiefendimensionen des Unterrichts zu bestimmen, stehen erprobte Instrumente aus einer Vielzahl groß angelegter Studien zur Verfügung, beispielsweise aus der „Pythagoras"-Studie, die den Zusammenhang von Unterrichtsqualität und Mathematikleistungen untersuchte (Rakoczy u. a. 2007), oder aus der DESI-Studie, in der deutsch-englische Schülerleistungen im internationalen Vergleich erfasst wurden (Klieme/Beck 2007). Das „Deutsche Institut für Internationale Pädagogische Forschung" (DIPF) stellt in einer allgemein zugänglichen Datenbank die bisher eingesetzten Instrumente zur Erfassung der Unterrichtsqualität zur Verfügung (http://www.dipf.de/de/Infrastrukturen/forschungsinstrumente/datenbank-zur-qualitaet-von-schule). Die Reliabilität und Validität dieser Instrumente zur Unterrichtseinschätzung aus Sicht der Schüler, der Lehrkräfte oder der externen Beobachter wurden bereits mehrfach untersucht (z.B. Clausen 2002; Gruehn 2000; Piskol 2008), sodass erprobte Instrumente zur Verfügung stehen. Bei der Messung der Lernergebnisse ist hingegen die Geschichtsdidaktik gefordert, neue Instrumente zu entwickeln.

Ob historische Kompetenzen mit standardisierten Instrumenten überhaupt messbar sind, wurde und wird oft prinzipiell in Frage gestellt (Körber u. a. 2008). Vor allem die Frage- und Orientierungskompetenzen, bei denen es um ein „Orientierungsbedürfnis" (Körber u. a. 2008, 81) gehe, aber auch die Prozesse historischen Denkens seien in geschlossenen Aufgabenformaten schwer zu fassen (Körber u. a. 2008). Die Operationalisierbarkeit von Kompetenzen historischen Denkens sei bei dem hohen Abstraktionsgrad der Kompetenzmodelle wie auch angesichts der Komplexität des Gegenstandes „Geschichte" selbst problematisch. Da die Lernenden verstehen sollen, dass es *die* Geschichte und die *eine* richtige Antwort gar nicht gibt (Borries 2007; VanSledright 2014), stoßen ökonomische quantitative Testverfahren, die auf einem Richtigkeitsstandard beruhen, an ihre Grenzen (Hartmann 2008). Ein Blick in die Tagungsbände der seit 2007 ca. zweijährig stattfindenden Tagung „geschichtsdidaktik empirisch" (Hodel/Ziegler 2008; Hodel/Ziegler 2010; Hodel u. a. 2013) zeigt, dass in der empirischen Fachdidaktik sehr viel häufiger qualitative Methoden eingesetzt werden, die eher hermeneutisch ausgerichtet sind. Allerdings können diese Methoden wegen ihres hohen Auswertungsaufwands nur in verhältnismäßig kleinen Stichproben eingesetzt werden. Interventionsstudien in Large-Scale-Dimensionen, die not-

wendig wären, um die Wirksamkeit bestimmter Methoden im Sinne der Inferenzstatistik empirisch belastbar zu überprüfen, können mit offenen Messinstrumenten kaum durchgeführt werden.

Die Konstituierung des Arbeitskreises „Empirische Geschichtsdidaktik" im Rahmen der „Konferenz für Geschichtsdidaktik" wie auch die Entstehung des vorliegenden Buches zeigen, dass sich die geschichtsdidaktische Forschung methodisch öffnet und zunehmend auch quantitative Methoden in den Blick nimmt. Im Folgenden soll ausgeführt werden, wie man bei der Entwicklung eines Messinstruments vorgehen kann. Diese Schritte werden am Beispiel der Entwicklung von Instrumenten, die im Rahmen einer groß angelegten, vom BMBF finanzierten Interventionsstudie zur Wirksamkeit von Zeitzeugenbefragungen im Geschichtsunterricht in enger Zusammenarbeit der Autorin mit Wolfgang Wagner und Ulrich Trautwein entstanden sind, veranschaulicht (vgl. zur Entwicklung der Messinstrumente: Bertram u. a. 2013; 2014). Zum Verständnis dieser Studie werden im Folgenden zunächst die Fragestellung und das Design der Zeitzeugenstudie skizziert. Nachfolgend werden einige zentrale statistische Begriffe geklärt.

Zielsetzung, Design und Fragestellung der Zeitzeugenstudie

Auf der Basis eines experimentellen Ansatzes wurden in der Studie „Chancen und Risiken von Zeitzeugenbefragungen. Eine randomisierte Interventionsstudie im Geschichtsunterricht" die Vor- und Nachteile der Arbeit mit Zeitzeugeninterviews untersucht. Die Intervention bildete eine kompetenzorientierte sechseinhalbstündige Unterrichtseinheit zum Thema „Friedliche Revolution in der DDR", in deren Zentrum in einer Doppelstunde Zeitzeugen live befragt wurden bzw. mit den Videos bzw. mit den Transkripten von Zeitzeugeninterviews gearbeitet wurde. Die Stichprobe umfasste insgesamt 38 Klassen (N = 962), von denen jeweils zehn zufällig einer der drei Interventionsgruppen zugewiesen wurden und weitere acht Klassen als Kontrollgruppe lediglich getestet wurden (vgl. zum Design und zur Ausgestaltung der Unterrichtseinheit: Bertram 2012).

Begriffsklärungen

Zum besseren Verständnis der folgenden Ausführungen sollten einige Begriffe vorab geklärt werden. Zum einen muss zwischen (Leistungs-)Tests und Fragebögen deutlich unterschieden werden. Beide arbeiten mit standardisierten Instrumenten, doch während der Fragebogen erfasst, „was jemand tut (Persönlichkeit), wie er es tut (Temperament) und warum er es tut (Motivation, Emotion, Einstellung, Interesse)", erfassen Tests „wie gut jemand es tut" (Eid/Gollwitzer/Schmitt 2011, 33). Einem Kompetenz- oder Wissenstest liegt also ein „Richtigkeitsstandard" zugrunde, während bei einem

Fragebogen, mit dem beispielsweise das Interesse am Fach oder Thema oder die Einschätzung des Unterrichts erfasst wird, nicht von richtigen oder falschen Lösungen gesprochen werden kann. Als zweite wichtige Voraussetzung sollte das Verhältnis zwischen dem empirischen Objektbereich und dem numerischen Messbereich geklärt werden. Messen bedeutet, dass der Ausprägungsgrad bestimmter Merkmale von Personen oder Objekten (empirisches Relativ) durch die Angabe von Zahlen (numerisches Relativ) repräsentiert wird, sodass bestimmte mathematische Vergleiche oder Operationen numerische Aussagen über die Verhältnisse im empirischen Merkmalsbereich ermöglichen. Dabei kommen – drittens – verschiedene Skalen (= Messsysteme) ins Spiel, mit denen die Datenwerte erfasst werden. Abhängig von der jeweiligen Skala sind verschiedene Analysen möglich. Likert-Skalen, bei denen z. B. auf einer fünfstufigen Skala von „trifft gar nicht zu" bis „trifft völlig zu" die Gültigkeit bestimmter Aussagen eingeschätzt werden sollen, werden häufig als Intervallskalen behandelt. In der Datenerfassung wird jeder der fünf Ausprägungen eine Zahl (1 bis 5) zugewiesen. Bei den Intervallskalen wird angenommen, dass die Abstände zwischen den einzelnen Antwortkategorien gleich sind, sodass von diesen Skalen Mittelwerte, Standardabweichungen etc. berechnet werden können. Nominalskalen – wie z. B. Geschlechtsangaben oder Religionszugehörigkeit – werden in der Dateneingabe zwar auch mit Zahlen erfasst, also auch hier gibt es für den empirischen Objektbereich eine numerische Entsprechung, doch bei der Auswertung kann nur eine Häufigkeit des Auftretens dieses Merkmals angegeben werden. Einen Mittelwert zu bilden, würde bei solchen Skalen keinen Sinn machen.

3. Vorgehen bei der Entwicklung eines Messinstruments[3]

Im Folgenden werden die Entwicklungsschritte hin zu einem standardisierten Messinstrument vorgestellt. Der Fokus liegt hierbei auf den Instrumenten, die fachdidaktisch relevant sind (Kenntnis- und Kompetenztest, Einschätzung der Unterrichtsmethode aus Schülersicht).

3 Das psychologische Lehrbuch „Testtheorie und Fragebogenkonstruktion" (Moosbrugger/Kelava 2012) bietet eine hervorragende Hilfestellung bei der Konzeption und Auswertung von standardisierten Messinstrumenten. Die folgenden Ausführungen nehmen Anregungen aus dem Kapitel „Planung und Entwicklung von Tests und Fragebogen" (Jonkisz u. a. 2012) auf.

1. Schritt: Festlegung der Kernkonstrukte

Was in einem Test bzw. Fragebogen erfasst werden soll, hängt eng mit der zugrunde liegenden Theorie und den im Forschungsprojekt formulierten Forschungsfragen und Thesen zusammen. Da die interessierenden „Konstrukte" nicht direkt messbar sind, müssen das Vorhandensein und die Struktur dieser Konstrukte aus messbaren Sachverhalten (den „Indikatoren") geschlossen werden. Die Zeitzeugenstudie adressierte spezifische Forschungsfragen, aus denen sich die Wahl des theoretischen Modells erklärt, wie auch die ausgewählten Konstrukte, die in Testaufgaben „übersetzt" wurden.

Die Zeitzeugenstudie geht davon aus, dass wegen der Ambiguität des Zeitzeugenberichts als Quelle *und* Darstellung und wegen der dem Bericht innewohnenden (Retro-)Perspektivität die Lernenden im reflektierten Umgang mit Zeitzeugeninterviews etwas über den grundsätzlichen Unterschied zwischen Quellen und Darstellungen wie auch über den Konstruktcharakter von Geschichte lernen können (Schreiber/Árkossy 2009). Zudem könnte es sein, dass es den Lernenden in der Live-Gruppe wegen der Präsenz des Zeitzeugen schwerer fällt, Distanz zu wahren („Aura der Authentizität", Sabrow 2012, 27) Auf der anderen Seite wäre es plausibel, dass Live-Zeitzeugenbefragungen wegen der Möglichkeit der direkten Interaktion das Interesse der Schülerinnen und Schülern an der Unterrichtseinheit fördern Darüber hinaus wird in der Studie untersucht, ob hinsichtlich des Kenntniserwerbs zur DDR-Geschichte Unterschiede zwischen den drei Interventionsgruppen zu beobachten sind.

2. Schritt: Theoretischer Rahmen und Forschungsfragen

Um die Kernkonstrukte zu definieren, werden im zweiten Schritt die theoretischen Grundlagen geklärt. Ein differenziertes theoretisches Modell bietet die Grundlage für die Formulierung der Forschungsfragen und für die im Zentrum der Untersuchung stehenden Konstrukte.

Theoretischer Rahmen

In der Zeitzeugenstudie wurde das FUER-Modell (Körber u a 2007) zu Grunde gelegt. Ausgehend vom Konzept der „disziplinären Matrix" (Rüsen 1983, 29) und dem Prozessmodell „Geschichtsbewusstsein dynamisch" (Hasberg/Körber 2003, 189) definierte die FUER-Gruppe ein Kompetenzmodell historischen Denkens (Schreiber u.a. 2007) bestehend aus Frage-, Methoden-, Orientierungs- und Sachkompetenzen. Verunsicherungen und Interessen setzen den Prozess historischen Denkens in Gang, der sich – in einer Fragestellung gebündelt (*Fragekompetenz*) – entweder in re-konstruierender Absicht an die Vergangenheit richtet oder sich in de-konstruierender Absicht mit vorliegenden historischen Narrationen auseinandersetzt (*Methodenkompetenzen = Re- und*

De-Konstruktionskompetenz). Das Ergebnis, das sich als eigene Narration bzw. als Stellungnahme gegenüber einer Darstellung präsentiert, befriedigt entweder bereits die Orientierungsbedürfnisse (*Orientierungskompetenz*) oder führt zu einer neuen historischen Frage. Durch den an verschiedenen Themen und Fragestellungen immer wieder durchlaufenen Prozess historischen Denkens bilden sich historische *Sachkompetenzen* heraus, d.h. die Schülerinnen und Schüler verfügen in zunehmendem Maße über die für den Umgang mit Geschichte relevanten Prinzipien, Konzepte und Skripts, z.B. über zentrale geschichtswissenschaftliche Begriffe und epistemologische Prinzipien. Letztere hat Baumgartner (1997) als „Prinzipien der Retroperspektivität, der Partikularität und der Konstruktivität von Geschichte" (Schreiber u.a. 2007, 32) definiert.

Zentrale Fragestellung

Von den oben skizzierten Überlegungen zur Wirksamkeit von Zeitzeugenbefragungen im Geschichtsunterricht leiten sich die Forschungsfragen ab zu den differentiellen Effekten der Arbeit mit Zeitzeugeninterviews (Live, Video, Transkription) bei den Lernenden hinsichtlich (a) Kompetenzen, (b) des themenspezfischen „Faktenwissens" und (c) der Schülereinschätzung der Unterrichtsmethode „Zeitzeugenbefragung". Etwas genauer formuliert: Es wurde untersucht, ob die Arbeit mit Zeitzeugeninterviews in den verschiedenen Interventionsgruppen hinsichtlich (a) der Einsicht in zentrale epistemologische Prinzipien und des Begriffsverständnisses von Quellen und Darstellungen (beides im FUER-Modell in den Sachkompetenzen verortet, Schreiber u.a 2007, 32), (b) des Erwerbs von Faktenwissen zur DDR und Friedlichen Revolution und (c) der Selbsteinschätzung der Lernenden zu ihren Lerneffekten in inhaltlicher, methodischer und motivationaler Hinsicht zu unterschiedlichen Effekten führt.

3. Schritt: Operationalisierung der Konstrukte

Um die theoretisch definierten Konstrukte zu operationalisieren, werden üblicherweise mehrere Aufgaben bzw. Items generiert, die für sich genommen einzelne Facetten des Konstrukts erfassen und insgesamt alle Facetten des Konstrukts repräsentieren sollen. Die „Operationalisierung" stellt eine entscheidende Etappe auf dem Weg zu einem Messinstrument (ob Fragebogen oder Test) dar. Für quantitative (und qualitative) Erhebungen müssen theoretische Konstrukte „messbar" gemacht werden, das heißt die Konstrukte werden in direkt beobachtbare Indikatoren, z.B. Testaufgaben, übersetzt (vgl. das vorab erwähnte Verhältnis zwischen dem empirischen und dem numerischen Relativ). Eine intensive Recherche nach Untersuchungen, die sich mit ähnlichen Fragestellungen befassen bzw. in denen ähnliche Konstrukte messbar gemacht werden, hilft bei der Ideengenerierung. Bewegt man sich jedoch auf wissenschaftlichem Neuland, kann es sinnvoll sein, in kleineren

qualitativen Vorstudien Anregungen für die Formulierung von geschlossenen Aufgaben zu generieren (vgl. Meyer-Hamme 2007). Bei der Ausformulierung der Aufgaben muss genau überlegt werden, mit welchen Aufgabenformaten die Konstrukte angemessen erfasst werden können und wie die „Items" – ein Item ist die kleinste Einheit einer Aufgabe, z. B. eine einzelne Aussage, zu der in einer Ratingskala Stellung genommen wird – formuliert werden.

Im Folgenden wird vorgestellt, welche vorhandenen Instrumente in der Zeitzeugenstudie genutzt bzw. adaptiert wurden, wie in einer kleinen qualitativen Vorstudie Ideen zur Itemformulierung generiert wurden und welche Instrumente neu entwickelt wurden. Daran anschließend bieten zwei längere Exkurse eine Übersicht über gängige standardisierte Aufgabenformate und Tipps für die Itemformulierung.

Entwicklung der fachdidaktisch relevanten Instrumente

Bei der Erfassung der Einsicht der Lernenden in die epistemologischen Prinzipien wurde auf Vorarbeiten in anderen Studien zurückgegriffen (u.a. von Borries u. a. 2005; Maggioni u.a. 2009). Hiervon ausgehend wurde ein Kurzinstrument mit vierzig Items entwickelt. Ideen zur Aufgabenformulierung hinsichtlich des Wissens über die DDR-Geschichte lieferten zwei prominente Studien, in denen die Faktenkenntnisse von Jugendlichen zur DDR untersucht worden sind (Arnswald u.a. 2006; Deutz-Schröder/Schröder 2008). Einige Aufgabenformate hieraus wurden übernommen bzw. für unsere Fragestellung adaptiert (z.B. Zuordnung von Politikern zur DDR oder BRD). Darüber hinaus entwarfen wir einen Lückentext zur Endphase der DDR. Um Anregungen für die Formulierung geschlossener Items zu bekommen, mit denen das Konzeptverständnis der Lernenden zu den Begriffen Quelle und Darstellung erfasst werden kann, wurde eine qualitative Vorstudie durchgeführt. Schülerinnen und Schüler der neunten Klasse beurteilten und begründeten, warum sie vier Dokumente zur Demonstration am 9 Oktober 1989 in Leipzig (Schulbuchtext, Aufruf vor der Demonstration, Stasi-Bericht nach der Demonstration, Zeitzeugenbericht) als eine Quelle oder Darstellung einschätzten. Sie wandten also ihr Begriffsverständnis von Quellen und Darstellungen auf die Texte an, womit sie ihre historischen (Sach)Kompetenzen unter Beweis stellten.

Individuelle Voraussetzungen der Lernenden

Das Rahmenmodell Helmkes (2012) impliziert, dass die Wirkung der Unterrichtseinheit im Zusammenhang mit dem generellen Interesse der Lernenden an Geschichte und am Thema wie auch mit ihren sozio-kulturellen Voraussetzungen steht. Daher wurden im Vortest der Interventionsstudie Daten zu den Voraussetzungen und dem Hintergrund der Lernenden erhoben. Für die Erfassung des Interesses an Geschichte konnte auf bewährte Instrumente aus Längsschnitt-Studien zurückgegriffen werden (z.B TRAIN, vgl. Jonkmann u.a.

2013). Die sozio-kulturellen Voraussetzungen wurden mit Instrumenten, die aus der PISA-Studie be kannt sind, erfasst (Kunter u. a. 2003). Die motivationalen und sozio-kulturellen Voraussetzungen der Lernenden wurden in die Wirksamkeitsanalysen als „Kovariaten" hineingenommen Dies bedeutet, dass ihr Einfluss auf die Wirksamkeit des Unterrichts statistisch kontrolliert wurde.

Einschätzung der Wirkung des Unterrichts und der Unterrichtsprozesse

Um zu erfassen, wie die Lernenden die Effektivität der Arbeit mit Zeitzeugen-Interviews in inhaltlicher, methodischer und motivationaler Hinsicht einschätzten, entwickelten wir einige Items selbst Darüber hinaus mussten die Prozesse während des Unterrichts, die das Lernen der Schülerinnen und Schüler bekanntermaßen beeinflussen, erfasst werden. Wenn sich die Lehrperson in den unterschiedlichen Interventionsbedingungen unterschiedlich verhalten würde (z.B. wenn die Lehrkraft die Live-Befragung bevorzugen und deswegen den Unterricht in den Live-Klassen effizienter gestalten und die Lernenden mehr unterstützen würde), dann hätte dies einen Einfluss auf die Effekte der Intervention. Daher schätzten die Lernenden nach der Unterrichtseinheit die Qualität des Unterrichts hinsichtlich der zentralen Dimensionen der kognitiven Aktivierung, der Klassenführung und der Unterstützung ein. Hierbei konnte auf die erprobten und validierten Instrumente aus der Unterrichtsforschung zurückgegriffen werden (vgl u. a Clausen 2002; Gruehn 2000; Piskol 2008). Bei der Überprüfung, wie der Unterricht in den dreißig Klassen abgelaufen ist, stellte sich heraus, dass die Lernenden auf der Klassenebene zwar aufmerksamer bzw. unaufmerksamer gewesen waren, was sicherlich Einfluss auf die Lernfortschritte der Lernenden hatte. Doch waren diese Unterschiede nicht durch die Zugehörigkeit zu einer bestimmten Interventionsgruppe bedingt. Daher konnte ausgeschlossen werden, dass diese unterschiedlich verlaufenen Unterrichtsprozesse die Ergebnisse verzerren hinsichtlich der Fragestellung, ob die Form der Arbeit mit Zeitzeugen zu differenziellen Effekten zwischen den drei Gruppen führt.

Exkurs I: Aufgabenformate

Eine Vielzahl von Aufgabentypen im freien (offene Aufgaben) und gebundenen Antwortformat (standardisierte Aufgaben) werden in der sehr informativen Einführung von Jonkisz u. a. (2012) zusammengestellt. Auf die Aufgaben mit freiem Antwortformat, bei denen keine Antwortalternativen vorgegeben werden, sondern die Antwort von der Person selbst formuliert wird (z.B. Kurzaufsatzaufgaben oder Ergänzungsaufgaben), soll an dieser Stelle nicht weiter eingegangen werden. Wer sich für die Codierung und Skalierung der offenen Aufgaben in der Zeitzeugenstudie interessiert, sei auf den Beitrag im gde13-Band verwiesen (Bertram u. a. 2015). Der Fokus dieses Beitrags liegt auf der Entwicklung und Auswertung der standardisierten

Aufgabenformate. Diese können als Ordnungsaufgaben (Zuordnung oder Umordnung), Auswahlaufgaben (z.B. dichotome Aufgaben, in denen zwischen richtig oder falsch gewählt werden soll, oder Mehrfachwahlaufgaben wie Multiple-Choice-Aufgaben) oder als Beurteilungsaufgaben (z.B. Ratingskalen wie die oben schon angesprochene Likertskala) formuliert werden. Ein besonderer Schwerpunkt wird im Folgenden auf den Beurteilungsaufgaben liegen, weil diese in Fragebögen wie auch in Tests sehr häufig eingesetzt werden.

Ordnungsaufgaben: In Zuordnungsaufgaben sollen die Testteilnehmer eine richtige Zuordnung von jeweils zwei Elementen vornehmen (z.B. Jahreszahlen bestimmten Ereignissen zuordnen). Der Vorteil besteht in der einfachen Handhabung, der platzsparenden Darbietung und der ökonomischen Auswertung. Besonders für Wissens- und Kenntnisaufgaben ist dieses Format geeignet. Das Problem der Ratewahrscheinlichkeit kann dadurch verringert werden, dass die Zahl der Antwortalternativen die der Fragen übersteigt. Allerdings ist bei diesem Format keine Reproduktionsleistung, sondern lediglich eine Wiedererkennungsleistung erforderlich. Bei Umordnungsaufgaben hingegen sollen die Probanden einzelne Teile (z.B. Worte, Satzteile, Bilder) umsortieren. Zum Beispiel könnte aus Textbausteinen ein logischer narrativer Text rekonstruiert werden. Hier müssen die Probanden eine eigenständigere Leistung abgeben und die Ratewahrscheinlichkeit wird reduziert.

Auswahlaufgaben: Bei Auswahlaufgaben muss die Disjunktheit der Antwortmöglichkeiten (keine Überlappung der Antwortalternativen) und die Exhaustivität (kein Fehlen von Antwortalternativen) berücksichtigt werden. Wird in einem Fragebogen beispielsweise danach gefragt, wie oft die Probanden historische Fernsehsendungen sehen, sind Antwortalternativen wie (a) „einmal im Jahr", (b) „ein- bis sechsmal im Jahr", (c) „sechs- bis zwölfmal im Jahr", (d) „ein- bis zweimal im Monat", und (e) „einmal wöchentlich" nicht sinnvoll, weil sich die Alternativen (b), (c) und (d) überlappen und mögliche Antwortalternativen nicht angeboten werden (z.B. gar nicht oder mehrmals wöchentlich). Für Leistungstests werden häufig Auswahlaufgaben gewählt, bei denen die Probanden aus mehreren vorgegebenen Antwortalternativen eine zutreffende Alternative (single choice) oder mehrere zutreffende Alternativen (multiple choice) auswählen sollen. Dieses Format ist ökonomisch einsetzbar und auswertbar. Allerdings funktionieren diese Aufgaben nur dann, wenn gute Distraktoren gefunden werden, wenn also die falschen Antwortmöglichkeiten den richtigen stark ähneln und damit plausibel erscheinen, sodass sie nicht leicht identifiziert werden können. Die Anzahl der (unzutreffenden) Antwortalternativen vermindert die Ratewahrscheinlichkeit.

Beurteilungsaufgaben (Ratingskalen): Beurteilungsaufgaben werden häufig in Fragebögen eingesetzt, um Einstellungen, Motive oder Persönlichkeitseigenschaften zu messen. Der Grad der Zustimmung oder Ablehnung zu einer vorgelegten Aussage wird als Indikator für die Ausprägung des Merkmals verwendet. Unterschieden wird zwischen verbalen Ratingskalen, bei denen jede Stufe beschrieben wird (z.B. „trifft gar nicht zu", „trifft eher nicht zu", „trifft manchmal zu", „trifft eher zu" und „trifft völlig zu"), und numerischen Ratingskalen, bei denen jede Stufe mit einer Zahl markiert wird. Da zu viele Stufen (mehr als sieben) in der Regel das Differenzierungsvermögen der Befragten übersteigen und zu wenige Stufen (weniger drei) zu wenig Bewertungsspielraum lassen, werden Skalen mit fünf plus bzw. minus einer Stufe meist empfohlen. Die Anzahl der Stufen sollte möglichst über alle Ratingaufgaben hinweg in einem Fragebogen oder Testinstrument identisch sein.

Mittlere Kategorie: Häufig wird darüber nachgedacht, ob es eine neutrale mittlere Kategorie geben soll. Da die Probanden einerseits die Mittelkategorie häufig nicht im Sinne einer mittleren Merkmalsausprägung, sondern als Ausweichoption nutzen, wenn der angegebene Wortlaut als unpassend empfunden, die Frage nicht verstanden oder die Antwort verweigert wird, und andererseits diese Kategorie von besonders motivierten Probanden häufig gemieden wird, führt die Nutzung einer Mittelkategorie zu einer Verzerrung der Befunde, da die verschiedenen Ursachen des Ankreuzens konfundiert sind und eine mittlere Antwort daher nicht unbedingt im Sinne des Konstrukts interpretiert werden kann. Daher wird eine neutrale Mittelkategorie meist abgelehnt (vgl. Jonkisz u.a. 2012, 54) oder eine zusätzliche „Weißnicht"-Kategorie angeboten, die das Problem der konstruktfremden Verwendung der neutralen Mittelkategorie verringern kann. Wenn man annehmen muss, dass es Probanden gibt, die zu dem Untersuchungsgegenstand keine Meinung haben oder die die Antwort nicht wissen, sollte diese Kategorie angeboten werden. Allerdings bietet sie gleichzeitig eine Ausweichoption an, die von Probanden genutzt werden kann, wenn sie über den Sachverhalt nicht nachdenken möchten. Daher sollte die Aufnahme einer zusätzlichen „Weiß nicht"-Kategorie im Vorfeld genau bedacht werden.

Umpolungen: Beim Einsatz von Beurteilungsaufgaben sollte die Tendenz, unabhängig von dem zu messenden Merkmal in einer bestimmten Art und Weise Antwortkategorien auszuwählen, berücksichtigt werden. Als Reaktion auf die sogenannte Ja-sage-Tendenz (Akquieszenz) wird empfohlen, auch negativ formulierte Items in einer Ratingskala einzusetzen, die Items also „umgepolt" zu formulieren. Zum Beispiel könnte in einem Persönlichkeitsfragebogen zur Erfassung der Gelassenheit in einer vierstufigen Skala von „trifft gar nicht zu" bis „trifft völlig zu" ein Item positiv formuliert werden

(„Ich bin jemand, der ruhig und gelassen bleibt."), und in einem anderen Item könnte das Gegenteil ausgedrückt werden (z.B. „Ich bin schnell aufgewühlt."). Bei der Aufbereitung des Datensatzes müssen die Werte des zweiten, gegenteilig formulierten Items in der Reihenfolge umgedreht werden („umkodiert") werden, um das Persönlichkeitsmerkmal „Gelassenheit" korrekt abzubilden.[4]

Vor- und Nachteile von Ratingskalen: Ratingaufgaben sind hinsichtlich der Erhebung ökonomisch handhabbar. Da sich die Testteilnehmer/innen auf einen Antwortmodus einstellen können, müssen sie nicht bei jeder Aufgabe „umdenken", sodass sich die Bearbeitungsdauer verkürzt. Hinsichtlich der Auswertung werden die Skalenpunkte in Zahlen übersetzt. Das bedeutet, dass bei einer vierstufigen Skala von „trifft gar nicht zu", „trifft eher nicht zu", „trifft eher zu" und „trifft völlig zu" die Zahlen 1 bis 4 vergeben werden. Diese Zahlen werden im Sinne einer Intervallskalierung verstanden, sodass Mittelwerte und Standardabweichungen gebildet werden können. Unter Voraussetzung dieses Skalenniveaus können eine Vielzahl von statistischen Analysen durchgeführt werden (z.B. Reliabilitätsanalysen oder Faktorenanalysen). Streng genommen stellen die Ratingskalen lediglich eine Reihenfolge her. Damit die Probanden die Abstandsgleichheit zwischen den Skalenpunkte erkennen, werden die Antwortmöglichkeiten wie im oben genannten Beispiel symmetrisch formuliert.

Exkurs II: Tipps für die Itemformulierung

Nachdem die Gestaltungsmöglichkeiten von Aufgabenformaten aufgezeigt wurden, wird nun die Itemformulierung in den Blick genommen. Entscheidend ist die eindeutige und für die Zielgruppe verständliche Formulierung. Hieraus leiten sich einige Hinweise ab, die an einigen Beispielen verdeutlicht werden.

„Ich bin angriffslustig."	Begriffe mit mehreren Bedeutungen sollten vermieden werden. Die Charakterisierung „angriffslustig" kann vom Probanden positiv oder negativ konnotiert werden und damit zu einer unterschiedlichen Interpretation und Beantwortung führen.

4 Obwohl dies häufig in der Praxis eingesetzt wird, zeigen (aktuelle) Befunde, dass eine gemischte Anbietung von invertierten und nicht-invertierten Items auch die Dimensionalität des Konstrukts beeinflussen kann. Rauch u.a. (2007) haben herausgefunden, dass bei der Nutzung eines Fragebogens mit Mischvarianten zur Erfassung des Konstrukts „Optimismus" statt eines Optimismus-Konstrukts zwei Dimensionen (Optimismus, Pessimismus) aufgetreten sind.

„Ich bin hedonistisch." in einem Fragebogen für Grundschüler	Die Wortwahl sollte zur Zielgruppe passen. Ein Grundschüler wird den Begriff „hedonistisch" vermutlich nicht kennen
„Ich fahre sehr gerne und sehr schnell Auto."	Pro Item sollte nur ein Aspekt genannt werden, da sich der Proband bei der Beantwortung nur auf einen der Itemteile oder auf beide beziehen kann. Eine eindeutige Interpretation der gegebenen Antwort ist somit nicht mehr möglich.
„Ich bin nie unlustig."	Doppelte Verneinungen vermeiden, da diese üblicherweise eines hohen kognitiven Aufwands seitens des Probanden bedürfen
„Alle Kinder machen immer nur Lärm."	Verallgemeinerungen sind in der Regel nicht günstig.
„Wie oft sind Sie in den letzten Wochen nur mühsam aus dem Bett gekommen?"	Zeitspannen sollten eindeutig definiert werden: „Wie oft sind Sie in der letzten Woche nur mühsam aus dem Bett gekommen?"
Beispiel 1: „Fallschirmspringen würde ich gerne ausprobieren": Ja – Nein	Das Antwortformat sollte zum Item passen: Beispiel 1: besser eine mehrstufige Likertskala von „würde ich gar nicht gerne" bis „würde ich sehr gerne"
Beispiel 2: „Mein Kind kann zugleich mit beiden Füßen eine Treppenstufe herunter hüpfen": Macht es a) nie; b) gelegentlich, c) häufig.	Beispiel 2: besser Ja-Nein-Antwortalternative

Zusammenfassend: Man sollte bei der Itemformulierung klare sprachliche Formulierungen wählen (einfache und klare Sätze konstruieren, unklare Begriffe vermeiden, möglichst positive Formulierungen verwenden) und die Items eindeutig formulieren (keine zweideutigen Begriffe, eindeutige zeitliche Bezüge, keine hypothetischen Fragen, keine doppelten Stimuli oder Verneinungen, keine Unterstellungen oder Suggestivfragen).

4. Schritt: Itemgenerierung und erste Erprobung

Nach der Festlegung, welche Konstrukte genau erfasst werden sollen, werden in einem offenen Brainstorming möglichst viele Items und Aufgaben generiert. Danach werden diese an der Zielpopulation erprobt, um Aspekte wie die Verständlichkeit der Items zu evaluieren. Hierbei hat sich die Thinking-aloud-Technik, in der die Befragten während der Aufgabenbearbeitung alles aussprechen, was sie gerade denken, worauf sie schauen, was sie tun und

fühlen, bewährt (Häder 2006). Eine Abwandlung der Thinking-aloud-Technik sind die sogenannten „Cognitive Labs", in denen die spontanen Äußerungen der Probanden durch situative Nachfragen, so genannte „probes" (Willis 2005), ergänzt werden. Die Nachfragen beziehen sich unter anderem auf das Verständnis der Aufgabe und Lösungsstrategien, auf das vorgängige Begriffsverständnis oder die Begründung, warum eine bestimmte Antwortalternative gewählt wurde. Bei der Entwicklung eines historischen Kompetenztests im Rahmen des vom BMBF geförderten HiTCH-Projekts („Historical Thinking – Competencies in History"; Projekt-Nummer: LSA006; vgl. Trautwein u. a. 2011) haben sich Cognitive Labs als hilfreich erwiesen, um die Denkprozesse der Probanden zu verstehen und die Aufgaben schülernah und verständlich zu formulieren (Werner/Schreiber 2015).

5. Schritt: Pilotierung

Nach der Überarbeitung auf der Grundlage der Rückmeldungen in den ersten Erprobungen wird der Testentwurf in der Pilotierung erstmals in einer ausreichend großen Stichprobe (meist > 100) überprüft. Auch hier sollten die Probanden zur Zielpopulation gehören. Neben der Frage, ob die Aufgaben und Items verständlich sind und wie lange die Bearbeitung dauert, geht es in dieser Pilotierung um die Überprüfung der psychometrischen Gütekriterien.

In der Zeitzeugenstudie wurden in einer ersten Pilotierung im Januar und Februar 2011 aus anderen Studien adaptierte wie auch neu entwickelte Testaufgaben in einer Schüler-Stichprobe (15 zehnte Klassen des Gymnasiums, $N = 311$) während einer Doppelstunde eingesetzt. Wir wählten als Stichprobe zehnte Schulklassen (und nicht neunte Klassen) aus, weil diese das Thema „DDR" im vorhergehenden Schuljahr behandelt hatten. Zum Einsatz kamen verschiedene Instrumente zur Messung der Faktenkenntnisse der Schülerinnen und Schüler, Items zur Erfassung ihrer Einsicht in epistemologische Prinzipien, offene und geschlossene Aufgaben, die sich auf vier Dokumente zur Demonstration am 9. Oktober 1989 in Leipzig bezogen, wie auch allgemeine Interesse- und Motivationsitems. Auch die Hintergrundvariablen der Schüler (z.B. Migrationshintergrund, kulturelles Kapital und Schulnoten in Deutsch, Geschichte und Mathematik) wurden erfasst. Die Items zur Einschätzung der Unterrichtseinheit mit Zeitzeugen-Interviews konnten nicht eingesetzt werden, da die Lernenden in ihrem Unterricht nicht mit Zeitzeugenbefragungen gearbeitet hatten. Die Daten wurden im Hinblick auf psychometrische Gütekriterien untersucht

Exkurs III: Psychometrische Gütekriterien

Die psychometrischen Gütekriterien lassen sich in drei Hauptkriterien (Objektivität, Reliabilität, Validität) und eine Reihe von Nebenkriterien differenzieren. Zu den Nebengütekriterien gehören Aspekte wie Testfairness oder Testökonomie (Moosbrugger/Kelava 2012). *Testfairness* bedeutet, dass Personen mit gleichen Merkmalsausprägungen (z.B. Intelligenz) unabhängig von z. B. Alter, Geschlecht, Regionen auch dieselben Testwerte (IQ-Score) erhalten. Die *Testökonomie* fragt danach, ob der Aufwand der Verfahrensanwendung im Verhältnis zum Nutzen durch das Verfahren steht. Im Folgenden werden die Hauptkriterien näher erklärt.

Objektivität bedeutet, dass die Messung und Auswertung unabhängig vom Testleiter oder der Testleiterin bzw. des Forschers oder der Forscherin vorgenommen werden können. Hierbei sollte auf drei Bereiche der Objektivität geachtet werden: Eine *Durchführungsobjektivität* kann durch eine Standardisierung der Testsituation erreicht werden, d.h. durch konkrete, verschriftlichte Anweisungen für die Probanden. Die *Auswertungsobjektivität* gilt für Fragebögen oder Testinstrumenten mit geschlossenen Antwortformaten meist als gegeben, da durch die numerische Kodierung klare Auswertungsregeln gelten, die die Berechnung eines Testwerts objektiv erlauben. Zuletzt kann die *Interpretationsobjektivität* z. B. dadurch sichergestellt werden, dass Normierungstabellen vorliegen, die eine objektive Interpretation des Testwerts einer Person in Relation zu einer interessierenden Population ermöglichen. Insgesamt ist festzustellen, dass eine genaue Dokumentation und feste Richtlinien für das Vorgehen bei der Messung, Auswertung und Interpretation die Objektivität erhöhen.

Reliabilität bezieht sich auf die Messgenauigkeit eines Fragebogens. Basierend auf der klassischen Testtheorie, kann man davon ausgehen, dass jede Messung (jedes Item) messfehlerbehaftet ist und nicht ausschließlich eine „wahre" Merkmalsausprägung widerspiegelt. Diese Messfehler können sich aus unterschiedlichsten Quellen speisen, wie Aufmerksamkeit des Probanden, Besonderheiten der Items oder der Situation, etc. Im Rahmen der Reliabilitätsschätzung wird angenommen, dass dieses Konglomerat von Einflüssen insgesamt betrachtet eine zufällige Störgröße darstellt. Für die Reliabilitätsabschätzung können verschiedene Methoden angewandt werden. Zum einen kann hinsichtlich der *Retest-Reliabilität* überprüft werden, ob die Ergebnisse in einer Stichprobe, die zu zwei Testzeitpunkten den gleichen Test bearbeitet, miteinander korrelieren, ob also die Testwerte über die Zeit hinweg stabil sind. Das Retest-Verfahren ist jedoch nur sinnvoll, wenn es sich um stabile Merkmale handelt, die nicht variabel oder kurzfristig veränderbar

sind. Die *Paralleltest-Reliabilität* überprüft, ob eine oder mehrere gleich schwere Parallelformen des Fragebogens hoch miteinander korrelieren. Da hierfür eine große Anzahl an gleich schweren („parallelen") Aufgaben vorhanden sein muss, kann dieses Verfahren sehr aufwendig sein. Die *Split-Half-* (Testhalbierungs) *Reliabilität* ist hoch, wenn zwei zufällig erzeugte Testhälften hoch miteinander korrelieren. Auch hierfür ist eine große Anzahl an Items erforderlich. Am häufigsten wird die Bestimmung der *internen Konsistenz* vorgenommen. Hierbei wird überprüft, ob die verschiedenen Items, die ein und dasselbe Merkmal erfassen, im Durchschnitt hoch miteinander korrelieren. Je höher diese Korrelationen zwischen den Items sind, desto höher ist die interne Konsistenz eines Tests. Am bekanntesten ist der Konsistenzkoeffizient Cronbachs Alpha, für den als Faustregel gilt, dass ein Alpha von > 70 für die Kollektivdiagnostik ausreichend ist. Im Weiteren wird die Itemtrennschärfe der einzelnen Items untersucht. Hier geht es um die Korrelation des Einzelitems mit dem Gesamttestwert. Eine hohe Itemtrennschärfe zeigt an, dass das Item gut zwischen den Probanden differenziert. Bezogen auf Testaufgaben bedeutet dies, dass Probanden mit einer hohen Leistungsausprägung das Item eher richtig lösen und die mit einer niedrigen Ausprägung das Item eher falsch lösen. Als Richtwert gilt eine Itemtrennschärfe von > 30 als wünschenswert (Schermelleh-Engel/Werner 2012).

Validität als drittes Gütekriterium gibt den Grad an, inwiefern ein Verfahren tatsächlich das misst, was es messen soll, also inwiefern ein Testwert auf das dahinterliegende Merkmal schließen lässt. Unter der *Inhaltsvalidität*, die z. B. aufgrund einer Expertenmeinung ermittelt wird, versteht man, dass die zentralen Inhalte des Gegenstandsbereichs oder Konstrukts durch die Items abgedeckt werden. Wenn ein Instrument einen hohen Zusammenhang mit einem Außenkriterium aufweist, spricht man von einer *Kriteriumsvalidität* (z.B. Zusammenhang von einem curricular vorgegebenen Wissenstest und der Schulnote). Von *Konstruktvalidität* spricht man, wenn ein Instrument hohe Zusammenhänge mit Instrumenten aufweist, die auf den gleichen bzw. ähnlichen Gegenstandsbereich abzielen (konvergente Validität) bzw. niedrige Zusammenhänge zu Instrumenten, die auf einen anderen Gegenstandsbereich abzielen (diskriminante Validität).

6. Schritt: Datenanalysen

Nach der Durchführung der Erhebung werden die Daten in ein Statistikprogramm (z.B. SPSS) eingegeben. Bei der Eingabe des ausgefüllten Fragebogens wird jedem Probanden für jedes Item im Datensatz ein Wert zugewie-

sen. In der Datenmatrix repräsentieren die Spalten die Items (=Variablen) und die Zeilen die Probanden.

Zunächst werden die Items deskriptiv anhand ihrer *Mittelwerte* und *Streuung* beschrieben, um das Antwortverhalten der Probanden in der Stichprobe zu erfassen. An der *Lösungshäufigkeit* zeigt sich die Schwierigkeit der Items, die weder zu leicht (Deckeneffekte) noch zu schwer (Bodeneffekte) sein sollten. Im nächsten Schritt werden *Reliabilitätsanalysen* durchgeführt, bei denen die interne Konsistenz der Skala (siehe oben: Cronbachs Alpha) wie auch die Itemtrennschärfe überprüft wird.

Auf der Grundlage von *Faktorenanalysen* kann bestimmt werden kann, ob die erhobenen Items sich auf dasselbe Merkmal beziehen, also mit derselben latenten Variable verbunden sind (d.h. auf denselben Faktor „laden"). Als Grundüberlegung steht hinter den Faktorenanalysen ähnlich wie bei den Reliabilitätsanalysen, dass das eigentlich interessierende latente Merkmal (oder „latente Variable"), z. B. die Einsicht der Lernenden in epistemologische Prinzipien, nicht direkt messbar ist, sondern anhand mehrerer Items operationalisiert werden muss. Werden verschiedene Items zu einem latenten Merkmal in ähnlicher Weise gelöst, dann zeigt dies, dass diese Items inhaltlich zusammengehören und dass es sich um eine abgrenzbare Kompetenzdimension handelt. Neben der Überprüfung der Dimensionalität geben die Faktorenanalysen eine Auskunft darüber, wie gut die einzelnen Items die zugrunde liegende Konstrukte 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 der zugrunde liegenden latenten Variablen betrachtet werden. Niedrige Ladungen hingegen bedeuten, dass ein Item sich als Indikator zur Messung der latenten Variablen weniger gut eignet. Auf Basis der deskriptiven Analysen wie auch der Reliabilitäts- und Faktorenanalysen wird eine Itemselektion vorgenommen, bei der das theoretisch definierte Konstrukt im Blick behalten sollte.

Testentwicklung und Ergebnisse in der Zeitzeugenstudie

Zusammenfassung: Vorgehen bei der Testentwicklung

Die theoretisch angenommenen Chancen und Risiken von Zeitzeugenbefragungen begründeten die Forschungsfragen, die auf der Grundlage einer größeren Stichprobe beantwortet werden sollten Das FUER-Modell mit seiner dezidierten Unterscheidung zwischen Re- und De-Konstruktionsprozessen schien für die Untersuchung der Fragestellung nach der Wirksamkeit von Zeitzeugen besonders geeignet. Bei der Suche nach schon vorhandenen Instrumenten konnten für den Kenntnisbereich Instrumente auf unsere Belange angepasst werden Darüber hinaus entwickelten wir einen Lückentext. Bei der Ent-

wicklung eines Kurzinstruments zur Erfassung der Einsicht in epistemologische Prinzipien konnte auf Vorarbeiten (z. B. Borries u. a. 2005; Maggioni u. a. 2009) zurückgegriffen werden, jedoch wurden die Items des Kurzinstruments auf einem „Richtigkeitsstandard" basierend ausformuliert. Von einer qualitativen Vorstudie ausgehend, wurden Items entwickelt, mit denen die Anwendung der Konzepte „Quelle" und „Darstellung" bezogen auf historische Dokumente erfasst wurde. Hinsichtlich der Einschätzung der inhaltlichen, methodischen und motivationalen Effekte der Unterrichtseinheit aus Schülersicht wurden ebenfalls eigene Items formuliert. Außer den letztgenannten Skalen zur Unterrichtseinschätzung wurden alle Aufgaben in einer Pilotierungsstudie mit über dreihundert Schülerinnen und Schülern eingesetzt und auf die psychometrischen Gütekriterien hin untersucht. Das Kurzinstrument zur Erfassung der Einsicht in die epistemologischen Prinzipien wurde darüber hinaus in zwei Studierenden-Erhebungen erprobt und faktorenanalytisch überprüft. In der Haupterhebung kamen darüber hinaus Instrumente zur Erfassung des sozio-kulturellen Hintergrunds der Lernenden, zu ihren motivationalen Voraussetzungen zu Geschichte und zum Thema wie auch zu ihrer Einschätzung der Unterrichtsqualität zum Einsatz. Da diese Instrumente aus anderen Studien adaptiert wurden, nicht spezifisch historisch konnotiert sind und in den Analysen lediglich als Kovariaten genutzt wurden, werden diese Instrumente im Folgenden nicht detailliert vorgestellt.

Faktenkenntnisse

Die Auswertung der Schülerstichprobe (N = 311) ergab, dass hinsichtlich der Aufgaben zur Erfassung der Faktenkenntnisse nur der selbst entwickelte Lückentext, bei dem die inhaltliche Richtigkeit (nicht die Rechtschreibung) bei der Trefferkodierung den Ausschlag gab, den psychometrischen Gütekriterien entsprach (14 Items, Cronbachs Alpha (α) = .82). Nicht funktioniert haben die aus anderen Studien adaptierten Wissenstestformate, z. B. Personenzuordnung (α = .40), Herstellung einer Chronologie (α = .51), Kenntnis von Politikern (α = .44). Diese Aufgaben, mit denen die themenspezifische Fachkenntnisse nicht reliabel erfasst werden konnten, wurden in der Interventionsstudie nicht eingesetzt.

Historische Sachkompetenzen I: Quelle oder Darstellung?

Ausgehend von den Schülerformulierungen in der qualitativen Vorstudie wurden zu vier Dokumenten zum 9. Oktober 1989 geschlossene Items formuliert. Die Schülerinnen und Schüler wurden gefragt, ob sie der Begründung für die Einordnung des jeweiligen Textes als Quelle oder Darstellung zustimmen würden, zum Beispiel, ob der Schulbuchtext eine Darstellung sei,

5 In der empirischen Bildungsforschung hat sich bei dem Berichten der Ergebnisse die anglo-amerikanische Schreibweise durchgesetzt. Das Reliabiltätsmaß „Cronbachs Alpha" wird als „α" notiert und bei Dezimalzahlen wie ‚0,82' wird die Schreibweise „.82" genutzt.

„weil der Verfasser viele Informationen recherchiert hat" Das Item wurde umgepolt formuliert. Dies wurde in der Trefferkodierung berücksichtigt, d.h. wenn dieses Item verneint wurde, wurde das Item als richtig gelöst gewertet. Der Mittelwert lag bei diesem Item bei 19 Prozent, das bedeutet, dass über 80 Prozent aller Probanden das Item richtig gelöst haben. Es war also ein relativ leichtes Item. Die Trennschärfe lag bei 30. Elf Items wurden für die Skala „Schulbuchtext: Quelle oder Darstellung" in der Validierungsstudie formuliert. Die Skala wies ein Cronbachs Alpha auf von .67 mit allen Items. Bei der Kürzung der Skala auf sechs Items verbesserte sich das Alpha auf .75. Das oben beschriebene Item wurde aufgrund der akzeptablen Trennschärfe in die später eingesetzte Skala übernommen. Alle vier Skalen haben sich in der Validierungsstudie als reliabel erwiesen. Trotzdem wurden aus Zeitgründen der Stasi-Bericht und die dazu gehörenden Items in der Haupterhebung nicht eingesetzt.

Historische Sachkompetenzen II: Einsicht in epistemologische Prinzipien

Das Kurzinstrument, mit dem die Einsicht der Schülerinnen und Schüler in epistemologische Prinzipien erfasst wurde, wurde nicht nur in der Validierungsstudie mit den Zehntklässlern, sondern auch in zwei Studierenden-Erhebungen eingesetzt. Neben der Reliabilitätsüberprüfung wurden faktorenanalytische Analysen durchgeführt. Im Folgenden werden die Ergebnisse der Faktorenanalysen zusammengefasst.

Da die Einsicht in epistemologische Prinzipien im FUER-Modell den Sachkompetenzen zugeordnet wird (Schreiber u.a 2007, 32), kann davon ausgegangen werden, dass mit diesem Kurzinstrument eine Facette der historischen Sachkompetenz erfasst wird. Die exploratorischen Faktorenanalysen legten ein Modell mit drei Faktoren nahe Auf Basis der in den Items formulierten Aussagen wurde ein Faktor „Sachkompetenz ‚Re-Konstruktion'„ genannt. Die Items, die auf diesem Faktor luden, adressierten die grundsätzliche Einsicht der Lernenden in den Konstruktcharakter von Geschichte, zum Beispiel das Verständnis, dass „Geschichte" aus dem Vergleich und der Interpretation von Quellen und Darstellungen entsteht (z.B. „Darstellungen sind das Ergebnis der Quellenanalyse und -interpretation wie auch der Auswertung anderer Darstellungen."). Die zweite Skala wurde als „Sachkompetenz ‚De-Konstruktion'" bezeichnet und enthielt Items, in deren Beantwortung die Schülerinnen und Schüler zeigten, ob es ihnen bewusst ist, dass Narrationen über die Vergangenheit de-konstruiert werden müssen (z.B. das umgepolte Item „Geschichtswissenschaftler beschreiben vergangene Ereignisse genauso, wie sie wirklich passiert sind."). Die dritte Skala „Sachkompetenz ‚Eigenart des Zeitzeugen'" umfasste Items, in denen die Besonderheit von Zeitzeugenberichten beispielsweise im Hinblick auf ihre Perspektivität und Zeitgebundenheit angesprochen wurde (z B. „Wenn man mehrere Zeitzeugen zum selben Ereignis befragt, bekommt man verschiedene Antworten.")

Bei einem Vergleich der exploratorischen Faktorenanalysen in den beiden Studierendenerhebungen zeigte sich ein sehr ähnliches Ladungsmuster. Zum Beispiel lud das Item „Darstellungen sind das Ergebnis der Quellenanalyse

und -interpretation wie auch der Auswertung anderer Darstellungen." in der ersten Studierendenerhebung mit .74 auf dem ersten Faktor, in der zweiten Erhebung mit .63. Das Item „Geschichtswissenschaftler beschreiben vergangene Ereignisse genauso, wie sie wirklich passiert sind." erreichte in der ersten Studierendenerhebung eine Ladung von .52, in der zweiten Erhebung hingegen eine Ladung von .91 auf dem jeweils dritten Faktor. Das Item „Wenn man mehrere Zeitzeugen zum selben Ereignis befragt, bekommt man verschiedene Antworten." lud in beiden Erhebungen mit .86 auf dem zweiten Faktor.

Einschätzung der Unterrichtseinheit

Die drei Testinstrumente zur Einschätzung der Unterrichtseinheit wurden in einem Trainingsdurchlauf, in dem die Unterrichtseinheit und die Testinstrumente ausprobiert wurden, erstmals in einer kleinen Stichprobe ausprobiert. Die folgenden Reliabilitätsmaße beziehen sich auf die Ergebnisse in der Interventionsstudie. Die Schülerinnen und Schüler schätzten die Wirksamkeit der Unterrichtseinheit in drei Skalen mit jeweils fünf Items ein: zu ihrem inhaltlichen Lernfortschritt (z.B. „Ich habe im Zeitzeugeninterview Neues zum Thema ‚DDR und Friedliche Revolution' erfahren."), zu ihren methodischen Erkenntnissen (z.B. „Ich habe in der Auswertung des Zeitzeugeninterviews gelernt, wie Historiker arbeiten.") und zum Motivationspotenzial durch die Arbeit mit Zeitzeugeninterviews (z.B. „Ich fand es toll, dass die Arbeit mit einem Zeitzeugeninterview im Mittelpunkt der Unterrichtseinheit stand."). Auch diese drei Skalen erwiesen sich als reliabel ($.69 < \alpha < .89$).

4. Zusammenfassung und Ausblick

Ziel des vorliegenden Beitrags war es, einen Einblick in das Vorgehen einer quantitativ ausgerichteten empirischen Untersuchung im Rahmen der Geschichtsunterrichtsforschung zu geben. Entscheidend bei empirischen Untersuchungen ist ein solides theoretisches Modell, das sowohl der Fragestellung als auch den zu messenden Konstrukten zugrunde liegt. Hinsichtlich der fachlich konnotierten Konstrukte (Kenntnisse, historische Kompetenzen) ist die Geschichtsdidaktik gefordert. Im Hinblick auf motivationale Konstrukte, auf Lernprozesse, auf die Dimensionen gelingenden Unterrichts wie auch auf die Methoden leistet die empirische Bildungsforschung wichtige Beiträge.

Ausgehend von einer theoretisch begründeten Fragestellung wird das Design einer Untersuchung entworfen (in diesem Beispiel eine Interventionsstudie). Anschließend werden die theoretisch fundierten Konstrukte in Messinstrumente „übersetzt". Bei dieser Operationalisierung können verschiedene Aufgabenformate genutzt werden, wobei die Hinweise zur Aufgaben- und

Itemformulierung beachtet werden sollten. In kleineren qualitativen Vorstudien (Thinking-aloud und/oder Cognitive Labs) kann das Messinstrument weiterentwickelt werden. Darüber hinaus sollte in einer größer angelegten Pilotierungsstudie überprüft werden, ob die Messinstrumente den psychometrischen Gütekriterien entsprechen. Nur Instrumente, die objektiv, reliabel und valide messen, sollten in einer experimentellen Wirksamkeitsstudie eingesetzt werden. In den statistischen Analysen zur Überprüfung der Wirksamkeit einer Intervention sollten weitere mögliche Wirkfaktoren des Unterrichts berücksichtigt werden (z.B. Interaktion in der Klasse, sozio-ökonomische Voraussetzungen, individuelle Motivation und Interesse).

Wenn Testinstrumente eingesetzt werden, die den psychometrischen Gütekriterien entsprechen, und wenn ein sinnvolles Design zur Untersuchung bestimmter Fragestellungen gefunden wurde, dann können evidenzbasierte Aussagen über die Wirksamkeit bestimmter Unterrichtsmethoden getroffen werden. Ähnlich wie in der Medizin, aus der der evidenzbasierte Forschungsansatz kommt, wird die Effektivität bestimmter Interventionen oder Methoden in einem kontrollierten Setting und mit funktionierenden Messinstrumenten empirisch überprüft. Es liegt auf der Hand, dass eine so verstandene empirische Forschung, in der fachdidaktische und empirische Expertise gefordert ist, in der Vorbereitung, Durchführung und Auswertung sehr aufwendig ist. Daher empfiehlt es sich, in Kooperationsprojekten die inhaltlich-fachdidaktischen wie auch die empirisch-messtheoretischen Anforderungen abzudecken. Gelingt dies, dann kann die empirische Geschichtsunterrichtsforschung wichtige Beiträge für die Gestaltung qualitätsvollen Geschichtsunterrichts liefern.

Literatur

Arnswald, Ulrich 2006: Schülerbefragung 2005 zur DDR-Geschichte. In: Arnswald, Ulrich u.a. (Hrsg.): DDR-Geschichte im Unterricht: Schulbuchanalyse – Schülerbefragung – Modellcurriculum. Berlin, S. 107-176.

Baumgartner, Hans Michael 1997: Narrativität. In: Bergmann, Klaus u.a. (Hrsg.): Handbuch der Geschichtsdidaktik, 5. Aufl. Seelze-Velber, S. 157-160.

Berliner, David C. 2005: The Near Impossibility of Testing for Teacher Quality. In: Journal of Teacher Education 56, 3, doi: 10.1177/0022487105275904, S. 205-213.

Bertram, Christiane 2012: Zeitzeugen zur Friedlichen Revolution: Live – Video – Text. Vorstellung einer kompetenzorientierten Unterrichtseinheit. In: Gerhard, Fritz/Wittneben, Eva L. (Hrsg.): Landesgeschichte in Forschung und Unterricht. Beiträge des Tages der Landesgeschichte in der Schule vom 26. Oktober 2011 in Bühl. Stuttgart, S. 63-80.

Bertram, Christiane u. a. 2013: Chancen und Risiken von Zeitzeugenbefragungen – Entwicklung eines Messinstruments für eine Interventionsstudie. In: Hodel, Jan/Ziegler, Béatrice (Hrsg.): Forschungswerkstatt Geschichtsdidaktik 12. Beiträge zur Tagung „geschichtsdidaktik empirisch 12". Bern, S. 108-119.

Bertram, Christiane u. a. 2014: Zeitzeugenbefragungen im Geschichtsunterricht: Entwicklung eines Kurzinstruments für die Wirksamkeitsmessung. In: Arand, Tobias/Seidenfuß, Manfred (Hrsg.): Neue Wege – neue Themen – neue Methoden? Ein Querschnitt aus der geschichtsdidaktischen Forschung des wissenschaftlichen Nachwuchses. Göttingen, S. 191-208.

Bertram, Christiane u. a. 2015: Historische Kompetenzen mit offenen Antwortformaten messen – Eine Studie auf Basis der „Sechser-Matrix" des FUER-Modells. In: Waldis, Monika/Ziegler, Béatrice (Hrsg.): Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13". Bern.

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

Borries, Bodo von 2007: Ergebnisse messen (Lerndiagnose im Fach Geschichte). In: Körber, Andreas u. a. (Hrsg.): Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik. Neuried, S. 651-673.

Brophy, Jere/Good, Thomas L. 1986: Teacher Behavior and Student Achievement. In: Wittrock, Merlin C. (Hrsg.): Handbook of Research on Teaching. New York, S. 340-370.

Clausen, Marten 2002: Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität. Münster.

Deutz-Schroeder, Monika/Schroeder, Klaus 2007: Abschlussbericht. Das DDR-Bild von Schülern in Berlin. Arbeitspapiere des Forschungsverbundes SED-Staat, Nr. 38/2007. Berlin.

Ditton, Hartmut 2006: Unterrichtsqualität. In: Arnold, Karl-Heinz u. a. (Hrsg.): Handbuch Unterricht. Bad Heilbrunn, S. 177-183.

Eid, Michael/Gollwitzer, Mario/Schmitt, Manfred 2010: Statistik und Forschungsmethoden. Lehrbuch. Weinheim.

Gruehn, Sabine 2000: Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung. Münster.

Häder, Michael 2006: Empirische Sozialforschung – eine Einführung. Lehrbuch. Wiesbaden.

Hartmann, Ulrike 2008: Perspektivenübernahme als eine Kompetenz historischen Verstehens (Dissertation). Göttingen 2008. Verfügbar unter https://ediss.uni-goettingen.de/handle/11858/00-1735-0000-0006-AD13-2.

Hasberg, Wolfgang/Körber, Andreas 2003: Geschichtsbewusstsein dynamisch. In: Körber, Andreas (Hrsg.): Geschichte – Leben – Lernen. Bodo von Borries zum 60. Geburtstag. Schwalbach/Ts., S. 177-200.

Hattie, John 2009: Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement. London.

Helmke, Andreas 2012: Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts. 4. Aufl. Seelze.
Hodel, Jan/Ziegler, Béatrice (Hrsg.) 2008: Forschungswerkstatt Geschichtsdidaktik 07. Beiträge zur Tagung ‚geschichtsdidaktik empirisch 07". Bern.
Hodel, Jan/Ziegler, Béatrice (Hrsg.) 2010: Forschungswerkstatt Geschichtsdidaktik 09. Beiträge zur Tagung „geschichtsdidaktik empirisch 09". Bern.
Hodel, Jan u. a. (Hrsg.) 2013: Forschungswerkstatt Geschichtsdidaktik 12. Beiträge zur Tagung „geschichtsdidaktik empirisch 12". Bern.
Jonkisz, Ewa u. a. 2012: Planung und Entwicklung von Tests und Fragebogen. In: Moosbrugger, Helfried/Kelava, Augustin (Hrsg.): Testtheorie und Fragebogenkonstruktion. 2. Aufl. Berlin, S. 27-74.
Jonkmann, K. u. a. 2013: Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen. Unveröffentlichter Projektbericht. Universität Tübingen.
Klieme, Eckhard u. a. 2001: Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung. In: Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.): TIMSS – Impulse für Schule und Unterricht. Bonn, S. 43-57.
Klieme, Eckhard u. a. 2003: Zur Entwicklung nationaler Bildungsstandards – Eine Expertise. http://www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf. (Aufruf vom 02.01.2012).
Klieme, Eckhard 2006: Empirische Unterrichtsforschung: aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. Einleitung in den Thementeil. In: Zeitschrift für Pädagogik 52, 6, S. 765-773.
Klieme, Eckhard/Beck, Bärbel (Hrsg.) 2007: Sprachliche Kompetenzen – Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim.
Klieme, Eckhard/Rakoczy, Katrin 2008: Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. In: Zeitschrift für Pädagogik 54, 2, S. 222-237.
Körber, Andreas u. a. (Hrsg.) 2007: Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik. Neuried.
Körber, Andreas u. a. 2008: Sind Kompetenzen historischen Denkens messbar? In: Frederking, Volker (Hrsg.): Schwer messbare Kompetenzen. Herausforderungen für die empirische Fachdidaktik. Baltmannsweiler, S. 65-84.
Kühberger, Christoph 2012 (Hrsg.): Historisches Wissen. Geschichtsdidaktische Erkundungen zu Art, Tiefe und Umfang für das historische Lernen. Schwalbach/Ts.
Kunter, Mareike u. a. 2003: Pisa 2000 – Dokumentation der Erhebungsinstrumente. Berlin.
Kunter, Mareike/Trautwein, Ulrich 2013: Psychologie des Unterrichts. Paderborn
Lipowsky, Frank 2009: Unterricht. In: Wild, Elke/Möller, Jens (Hrsg.): Pädagogische Psychologie. Heidelberg, S. 73-102.

Maggioni, Liliana u. a. 2009: Walking on the Borders: A Measure of Epistemic Cognition in History. In: The Journal of Experimental Education 77, 3, doi: 10.3200/JEXE.77.3.187-214, S. 187-213.

Meyer, Hans 2004: Was ist guter Unterricht? Berlin.

Meyer-Hamme, Johannes 2007: Historische Kompetenzen empirisch – Ein Versuch, das Kompetenz-Strukturmodell empirisch zu wenden. In: Körber, Andreas u. a. (Hrsg.): Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Fachdidaktik. Neuried, S. 674-693.

Moosbrugger, Helfried/Kelava, Augustin 2012: Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: dies. (Hrsg.): Testtheorie und Fragebogenkonstruktion. Berlin, S. 7-26.

Piskol, Kathleen 2008: Unterrichtsqualität aus der Schülerperspektive: Ein Verfahren zur Unterrichtsentwicklung. Diplomarbeit, Universität Mannheim. Verfügbar unter https://ub-madoc.bib.uni-mannheim.de/2664/

Rakoczy, Katrin u. a. 2007: Structure as a Quality Feature in Mathematics Instruction: Cognitive and Motivational Effects of a Structured Organisation of the Learning Environment vs. a Structured Presentation of Learning Content. In: Prenzel, Manfred (Hrsg.): Studies on the Educational Quality of Schools. The Final Report on the DFG Priority Programme. Münster, S. 101-120.

Rüsen, Jörn 1983: Historische Vernunft. Die Grundlagen der Geschichtswissenschaft. Grundzüge einer Historik I. Göttingen.

Sabrow, Martin 2012: Der Zeitzeuge als Wanderer zwischen zwei Welten. In: Sabrow, Martin/Frei, Norbert (Hrsg.): Die Geburt des Zeitzeugen nach 1945. Göttingen, S. 13-32.

Schermelleh-Engel, Karin/Werner, Christina 2012: Methoden der Reliabilitätsbestimmung. In: Moosbrugger, Helfried/Kelava, Augustin (Hrsg.): Testtheorie und Fragebogenkonstruktion. Berlin, S. 119-141.

Schreiber, Waltraud u. a. 2007: Historisches Denken. Ein Kompetenz-Strukturmodell (Basisbeitrag). In: Körber, Andreas u. a. (Hrsg.): Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik. Neuried, S. 17-53.

Schreiber, Waltraud/Árkossy, Katalin (Hrsg.) 2009: Zeitzeugengespräche führen und auswerten. Historische Kompetenzen schulen. Neuried.

Seidel, Tina/Shavelson, Rich 2007: Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. In: Review of Educational Research 77, 4, doi: 10.3102/0034654307310317, S. 458-499.

Trautwein, Ulrich u. a. 2011: Entwicklung und Validierung eines Tests historischer Kompetenzen zum Einsatz in Large-Scale-Assessments. Antragsskizze im Rahmen der BMBF-Ausschreibung zur Förderung von Forschungsvorhaben in Ankopplung an Large-Scale-Assessments vom 1. Februar 2011 (unveröffentlicht).

VanSledright, Bruce 2014: Assessing Historical Thinking & Understanding. Innovative Designs for New Standards. New York.

Werner, Michael/Schreiber, Waltraud 2015. Testfragen befragen. Pretesting und Optimierung des Large-Scale-Kompetenztests ›HiTCH‹ durch Cognitive Labs. In: Waldis, Monika/Ziegler, Béatrice (Hrsg.): Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13". Bern.
Willis, Gordon (2005). Cognitive Interviewing. A Tool for Improving Questionnaire Design. Thousand Oaks, CA.