Jan M. Boelmann
(Hrsg.)

Empirische Forschung in der Deutschdidaktik

Band 1: Grundlagen


V
H

Empirische Forschung in der Deutschdidaktik

Band 1

Grundlagen

Herausgegeben von

Jan M. Boelmann

Schneider Verlag Hohengehren GmbH

Umschlagidee:
Herausgeber

Bildquelle:
Fotolia: business analysis concept on a blackboard/134401817

Die Veröffentlichung dieses Bandes als Open-Access-Veröffentlichung wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 16PGF0074 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Bibliografische Information der Deutschen Nationalbibliothek

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

ISBN: 978-3-8340-1881-6

Schneider Verlag Hohengehren GmbH
Wilhelmstrasse 13
D-73666 Baltmannsweiler
Homepage: www.paedagogik.de

Inhaltsverzeichnis

Forschungsdesigns

Praxisfragen: Organisation und Durchführung des Projekts

Vorbereitung

Verarbeitung

JAN M. BOELMANN

Vorwort

Die Schritte auf dem Weg zum eigenen Forschungsprojekt sind gespickt mit großen und kleinen Herausforderungen und es lässt sich nicht vermeiden, dass der Prozess Zeit, Nerven und viele Gedanken kosten wird. Dies mag darin begründet liegen, dass es auf der Suche nach *einer* oder auch *der einen* Wahrheit keine einfachen Herangehensweisen, keine Abkürzungen oder leicht adaptierbaren Fahrpläne gibt, deren Befolgung auch nur zu solide guten Ergebnissen führen würde. Empirische Forschung als einer der Königswege der wissenschaftlichen Auseinandersetzung mit einem Thema verlangt Forscherinnen und Forschern einiges ab.

Dementsprechend kann dieser Band keine einfachen Antworten auf die Probleme liefern, die ein Forschungsprozess stellt, aber doch die Leserinnen und Leser auf dem Weg begleiten und in die zentralen theoretischen Konzepte und praktischen Erwägungen einführen. Hierdurch wird das Dickicht, vor dem sich mancher Novize und manche Novizin zu Beginn wähnen, in seiner Struktur überschaubarer und die Anforderungen klarer. Gerade ein solcher Band, der Anfängerinnen und Anfängern zwischen Praxissemester und Promotion die ersten Schritte erleichtern soll, stellt die Autorinnen und Autoren der einzelnen Beiträge vor die Herausforderung, dass die Inhalte einerseits einsteigerfreundlich aufbereitet aber andererseits nicht simplifiziert werden sollen – gerade bei den theoretischen Beiträgen bedeutet das: ‚Von hier aus bitte weiterlesen!'. Die Beiträge vermitteln das Fundament, auf dem alle weiteren Schritte aufbauen müssen, zugleich erhalten die Leserinnen und Leser auf überschaubarem Umfang ein fundiertes Überblickswissen.

Dieser Band sieht sich in enger Verbindung mit den beiden Folgebänden dieser Reihe, weshalb ein zentraler Aspekt hier nur kursorisch Berücksichtigung findet, der in Band 2 dieser Reihe ausführlich behandelt wird: die Auswahl von zur Fragestellung passenden Erhebungs- und Auswertungsverfahren. Eine Kurzvorstellung in Form eines 20-seitigen Beitrags hätte dem Umfang und der Komplexität dieses zentralen Teilprozesses nicht im Ansatz Rechnung tragen können, sodass die Lektüre des Folgebandes dringend anempfohlen wird.

Zwar wird die Wahl des Forschungsfeldes – dies stellt den Schwerpunkt in Band 3 dieser Reihe dar – zum Zeitpunkt der Planung der eigenen Studie zumeist schon abgeschlossen sein, doch sei auch auf diesen Band verwiesen, in dem die verschiedenen empirischen Forschungsfelder der Deutschdidaktik überblicksartig

vorgestellt werden – vielleicht ergeben sich durch den Blick in die empirische Forschung anderer Forscherinnen und Forscher der eigenen Disziplin noch Anregungen für die Planung des Forschungsvorhabens.

Dank sei diesem Band vorangestellt: Ich danke den zahlreichen Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftlern für ihre Rückmeldungen zu dem bereits länger bestehenden *Erhebungs- und Auswertungsverfahren*-Band. Hierdurch wurde die Notwendigkeit dieser Reihe deutlich und ich freue mich bereits auf weitere Rückmeldungen. Der größte Dank gilt meiner Doktorandin Lisa König, die bei der Konzeption und Erstellung von Reihe und Band rat- wie tatkräftig mitgewirkt hat und den Blickwinkel des wissenschaftlichen Nachwuchses konsequent in jeder Phase der Entstehung zentral stellte.

JAN M. BOELMANN

Zur Konzeption des Bandes

Bei der Planung dieses Grundlagenbandes standen zwei zentrale Ziele im Mittelpunkt: Erstens soll der Band einsteigerfreundlich die *theoretischen Grundlagen* vermitteln, die Novizinnen und Novizen der empirischen Forschung benötigen, um fundiert erste Entscheidungen auf dem Weg zum eigenen Projekt treffen zu können. Zweitens soll dieser Band die Leserinnen und Leser *praxisnah an die Planung und Durchführung* ihres eigenen Projektes heranführen und dabei Modelle des Best-Practice aufzeigen und vor typischen und folgenschweren Fehlern warnen.

Wie ist der Band strukturiert?

Der Band gliedert sich in einen Theorie- und einen Praxisteil. Nach einer Einführung in die Entwicklung der Fragestellung, die den ganzen Forschungsprozess leitet, setzt der erste Teil *Theoriefragen: Ansätze, Paradigmen, Designs* an und liefert einen Überblick über die theoretischen Grundlagen von den ersten Überlegungen bis hin zum fertigen Design. Im ersten Abschnitt ***Forschungsansätze*** wird die grundlegende Zielrichtung des Projekts in den Blick genommen, wobei mit *Grundlagenforschung*, *Design-Research*, *Evaluationsforschung* und *Metaanalysen* vier Großformen aus theoretischer und praktischer Perspektive vorgestellt werden. Hierbei problematisieren die Autorinnen und Autoren die verschiedenen Konzepte einerseits wissenschaftstheoretisch, zeigen aber andererseits auch praktisch, wie in der Deutschdidaktik mit diesen Ansätzen gearbeitet wird. Der folgende Abschnitt ***Forschungsparadigmen*** widmet sich der hochrelevanten Frage, für welche Forschungsfragen ein *qualitatives*, *quantitatives* oder aber *gemischtes* Vorgehen hilfreich sein werden. Auch hier wird über die Aufbereitung theoretischer Grundlagen und praktischer Anwendungsbeispiele ein breites Fundament gelegt, wobei neben den grundsätzlichen Verfahrensweisen mit den Gütekriterien auch *Qualitätsstandards* für deren Einsatz vorgestellt werden. Abschließend leitet der dritte Abschnitt ***Forschungsdesigns*** in die konkrete Entwicklung der Studie über: Während grundlegende theoretische planerische Überlegungen im ersten Beitrag skizziert werden, stellen die beiden Folgebeiträge eine Gelenkstelle zwischen Theorie und Praxis dar, indem sie am Beispiel konkreter Forschungsprojekte die Entwicklung und Erprobung unterschiedlicher Designs

aufzeigen. Hierbei nutzen sie die in den anderen Beiträgen des Bandes vorgestellten Leitlinien und reflektieren die jeweiligen Entscheidungsstellen im Entwicklungs- und Durchführungsprozess.

Der zweite Teil des Bandes widmet sich praktischen Erwägungen, mit denen alle empirisch Forschenden im Verlauf ihrer Arbeit in Kontakt kommen. In der **Vorbereitung der Erhebung** müssen sowohl allgemeine planerische Aspekte als auch Vorgaben durch Datenschutz und Forschungsethik berücksichtigt werden, denen sich der erste Abschnitt widmet: Neben Aspekten des Best-Practice im Feld der Planung und Vorbereitung wird mit einem Beitrag zur Erfassung biographischer Daten ein Feld fokussiert, das viel zu häufig nebenbei erhoben wird, ohne dass die spezifischen Potenziale erkannt oder ausgeschöpft würden. Die Beiträge zu Forschungsethik und Datenschutz fokussieren hingegen das Forschungsgeschehen stärker aus der Metaebene, wobei sich hieraus dennoch normativ gültige Handlungsanweisungen für die Praxis ergeben. Während der Beitrag zur Forschungsethik vielfältige Fragen auch zu einer spezifisch-deutschdidaktischen Forschungsethik aufwirft, stellt der Beitrag zum Datenschutz deutlich heraus, was im Rahmen von Forschung erlaubt ist und welche gesetzlichen Grenzen nicht übertreten werden dürfen.

Der zweite Abschnitt vermittelt Grundlagenwissen zur **Verarbeitung der erhobenen Daten**, wobei das vermittelte Praxiswissen zu Transkription, Softwareeinsatz und statistischen Grundlagen unmittelbar in die Arbeit münden kann, wenngleich auch hier Perspektiven zur lohnenswerten Vertiefung angeboten werden.

Mit diesem Aufbau begleitet der Band die Leserinnen und Leser vom Praxissemester bis zur Dissertation beim Einstieg in die empirische Forschung und zeigt vielfältig auf, wie von hier aus die Weiterarbeit gestaltet werden kann.

JAN M. BOELMANN

Fragestellung

Wie die Fragestellung den Forschungsprozess leitet

Inspiration exists, but it has to find you working.
Picasso

1. Notwendigkeit einer Fragestellung

Forschung beginnt immer mit einer Frage. Vermutlich ist sie noch nicht präzise und ausgereift, aber sie deutet in eine erste Richtung: „Warum lesen die Kinder in meiner Klasse so schlecht?“, „Wie verstehen Menschen grammatische Regeln?“, „Wenden Lehrerinnen und Lehrer in der Praxis das an, was sie im Studium gelernt haben?“. Zwar handelt es sich bei diesen Fragen noch nicht um eine Fragestellung, aber sobald man beschließt, ihnen wissenschaftlich auf den Grund zu gehen, werden sie ein wichtiger Ausgangspunkt. Von hier aus wird der Gegenstand weiter konkretisiert, eingegrenzt und fundiert, um in der Folge eine spezifische, theoriebasierte und relevante Fragestellung zu entwickeln.

Wenn diesem Grundlagenband der Beitrag zur Fragestellung vorangestellt wird, trägt dies ihrer zentralen Rolle im Forschungsprozess Rechnung: Die Fragestellung ist Ausgangspunkt und zentraler Anker aller Entscheidungen im Forschungsprozess und dient diesem als wegweisender Kompass, an dem sich die Notwendigkeit und das Gelingen jedes einzelnen Schrittes zeigt. Während eine gute Fragestellung zur Folge hat, dass sich die darauf aufbauende Forschungsarbeit fast von selbst konzeptioniert, wird eine schlechte Fragestellung jeden einzelnen Schritt der Arbeit erschweren oder verunmöglichen, das Projekt aufblähen und Schlussfolgerungen verhindern.

Zugleich ist die Entwicklung der Fragestellung ein arbeitsreicher Prozess. Im Sinne des vorangestellten Zitats muss betont werden, dass es sich bei der Fragestellung nicht um einen Einfall handelt, der den Forschenden in einem Moment der Muße zufliegt, sondern sie vielmehr das Ergebnis eines intensiven Auseinandersetzungsprozesses mit den eigenen Forschungsinteressen und dem aktuellen Forschungsstand darstellt. Erst wenn man all dies gründlich und umfassend aufgearbeitet hat, mag sich die Inspiration zur Fragestellung einstellen.

Auch muss an dieser Stelle betont werden, dass die Entwicklung der Fragestellung zwar vor der näheren Beschäftigung mit Ansätzen, Paradigmen und Designs

beginnt, jedoch nicht vor einer Auseinandersetzung mit ihnen abgeschlossen werden kann. Somit stellt die Arbeit an der Fragestellung einen forschungsbegleitenden Prozess dar, der jedoch vor der Erhebung der ersten Forschungsdaten abgeschlossen sein muss[1].

Die Fragestellung legt fest, was der *Gegenstand* der Forschung sein wird und begrenzt somit das Vorhaben bzw. grenzt es von anderen Vorhaben ab. Hiermit erfüllt es intern wie extern eine wichtige Funktion: Für die Forschenden gibt die Fragestellung Orientierung und Halt und zeigt nach Außen, was Rezipientinnen und Rezipienten der Forschung zu erwarten haben – aber auch, was nicht. Hierbei darf eine Fragestellung durchaus vielschichtig bleiben, sofern anschließend eine Ausdifferenzierung in weitere Unterfragen erfolgt. Zugleich muss sie trotzdem den selbstformulierten Anspruch in der Folge erfüllen. Das Verhältnis von empirischer Forschung und einer Fragestellung lässt sich leicht formulieren: Nicht jede Fragestellung braucht Empirie, aber jede Empirie braucht eine Fragestellung.

Wie bereits benannt, geht der Entwicklung der Fragestellung ein Forschungsinteresse voraus, das durchaus in Form einer allgemeineren Frage formuliert sein kann. Dieses allgemeine Interesse gilt es, in einem ersten Schritt bestmöglich zu präzisieren. Ohne diese Präzisierung ufert das Thema und die sich hieraus ergebenden Aufgaben aus: Die anfänglich genannte Frage nach der Anwendung der Studieninhalte durch Lehrerinnen und Lehrer in der Praxis lässt sich etwa auf Grund ihres schierem Umfangs nicht im Rahmen einer Forschungsarbeit beantworten. Um dieses Vorhaben anzugehen, müsste man zuerst alles rekonstruieren, was die Lehrerinnen und Lehrer im Studium gelernt haben und zweitens abprüfen, inwieweit diese Kenntnisse in den aktuellen Unterricht einfließen. Die Rekonstruktion würde bestenfalls alle universitären Lehrveranstaltungen und sonstigen Lerngelegenheiten der jeweiligen Probandinnen und Probanden umfassen, wobei in der Auswertung zwischen Lernangebot – ‚Welche Lernmöglichkeiten hatten die Beforschten?' – und Lernertrag – ‚Was haben sie wirklich gelernt?' – unterschieden werden sollte. Diese Informationen lassen sich zum einen nicht rückwirkend erheben, sondern müssten parallel zum Studium der Probandinnen und Probanden beispielsweise videographisch erhoben werden, wobei somit die Datengrundlage am Ende des Studiums ca. 2.000 Stunden Videomaterial betrüge, was bedeutet, dass die einmalige Sichtung bereits 250 Tage in Anspruch nähme – all dies, ohne dass eine Sekunde ausgewertet wäre – eine empirisch saubere Auswertung würde grob überschlagen Zeit im Faktor 10 beanspruchen, also 20.000 Stunden oder 2.500 Tage oder 7 Jahre. Anschließend müsste der vom Probanden/der Probandin gehaltene Unterricht darauf hin untersucht werden, ob sich Spuren des Gelernten – gerade Kompetenzen im Sinne von Handlungswissen zeigt sich, anders als die Reproduktion von Wissen, in vielfältigen Varianten, die es aufzu-

[1] Es ist möglich, auch ohne Forschungsfrage Daten zu erheben, allerdings hat dies wenig mit empirischer Forschung zu tun, da dem Handeln eine Richtung fehlt. Ggf. können solche Daten im Rahmen von Vorstudien genutzt werden, um die Forschungsfrage zu präzisieren, aber auch dies setzt eine implizite Fragestellung voraus.

schlüsseln und zu identifizieren gälte – in den Unterrichtshandlungen des Probanden implizit. Wenn die Arbeit jedoch nicht nur eine Fallstudie darstellen soll, die beispielhaft für *eine* Lehrerbiographie steht, sondern eine etwas größere Tragweite für sich beansprucht, müssen weitere Probanden auf die gleiche Weise untersucht werden. Nach wenigen Jahrzehnten ununterbrochener Arbeit könnte der/die Forschende erste Ergebnisse präsentieren, die jedoch bereits zu diesem Zeitpunkt als veraltet angesehen werden müssten.

2. Entwicklung einer Fragestellung

Dieses pointierte Beispiel zeigt, dass die anfängliche Frage noch keine taugliche Fragestellung für die Forschungsarbeit darstellt und dringend präzisiert werden muss. Grundsätzlich gibt es hierfür verschiedene Stellschrauben, die sich in der Frage: „Wenden Lehrerinnen und Lehrer in der Unterrichtspraxis das an, was sie im Studium gelernt haben?“, bereits zeigen: Gegenstand, Fragefokus, Probandengruppe und Rahmung.

Gegenstand: Der im Beispiel gewählte Gegenstand ‚alles, was im Studium gelernt wurde‘ ist selbstverständlich zu groß gewählt und lässt sich durch Fokussierung auf einzelne Aspekte verkleinern: So könnte beispielsweise das Wissen über Lesestrategien, Textanalysefähigkeiten, Einstellung zu inklusiven Lerngruppen oder Kommunikationsstrategien im Fokus der Untersuchung stehen. Die Verkleinerung des Gegenstands wirkt sich unmittelbar auf die Tragweite der Erhebung aus, zugleich eröffnet sie Räume für eine intensive Auseinandersetzung, die bei einer breit angelegten Untersuchung nicht möglich wäre.

Fragefokus: Die Wahl des Blickwinkels bestimmt maßgeblich, was im Rahmen des Projekts erhoben und ausgewertet wird. Im Beispiel wird durch das Prädikat ‚anwenden‘ ein klarer Fokus auf die konkreten Handlungen des Probanden gelegt, die für die Forschung erhoben und ausgewertet werden müssten. Es wäre eine andere Arbeit, wollte man erheben, *über welches Studiumswissen sie noch verfügen* oder *welche Bedeutung sie den Studieninhalten beimessen*. Auch machen an dieser Stelle sprachliche Nuancen einen großen Unterschied: Die Frage, ob Schülerinnen und Schüler eine Unterrichtseinheit *lehrreich finden* oder ob diese Unterrichtseinheit für die Schülerinnen und Schüler *lehrreich war*, ziehen völlig unterschiedliche Projekte nach sich.

Probandengruppe: Im Beispiel ist das Feld der benannten Gruppe noch weit und umfasst Lehrerinnen und Lehrer aller Schulformen und Alters- bzw. Erfahrungsstufen. Auch wurde nicht präzisiert, ob sie gewisse Rahmenbedingungen oder Eigenschaften aufweisen müssen. Eine Eingrenzung könnte somit beispielhaft die Fokussierung auf Lehrerinnen und Lehrer in den *ersten drei Berufsjahren nach dem Referendariat* (‚Junglehrer‘), *Sekundarstufenlehrkräfte* oder *Quereinsteiger aus der Kreativbranche mit dem Unterrichtsfach Deutsch* nach sich ziehen.

Rahmung: Nicht jede Fragestellung benötigt eine Rahmung. Im vorliegenden Fall ist sie mit ‚in der Unterrichtspraxis‘ breit gewählt, jedoch präzisiert sie den

Rahmen der Untersuchung, indem die *Praxis* fokussiert wird und außerunterrichtliche Bereiche ausgespart bleiben. Grundsätzlich lässt sich die Fragestellung über die Rahmung beliebig eingrenzen: ‚in einer Grundschule', ‚in einer Vorbereitungsklasse', ‚vor den Sommerferien', etc.

Jede Veränderung auf einer dieser Ebenen zieht für das Forschungsprojekt teils weitreichende Konsequenzen nach sich und führt in weiteren Schritten dazu, dass eine mit pragmatisch-realistischem Aufwand beantwortbare Fragestellung entsteht. Der Prozess des Zuspitzens lässt sich in drei zirkulär ablaufende Phasen ausdifferenzieren: eine *Phase der theoretischen Fundierung*, eine *Phase der Präzisierung und Weitung* und abschließend eine *Phase der Formulierung und Festlegung*. Die Dauer dieses Prozesses kann nicht seriös benannt werden, da er von zu vielen Faktoren beeinflusst wird, wobei Erfahrung und Vorwissen zentrale Einflüsse darstellen: Erfahrenere Wissenschaftlerinnen und Wissenschaftlerinnen mögen in ihrem Spezialgebiet innerhalb weniger Stunden eine geeignete Fragestellung für ein Forschungsprojekt entwickeln, die meisten Doktorandinnen und Doktoranden benötigen mehrere Monate intensiven Arbeitens, bis ihre Fragestellung feststeht, wobei zusätzlich das Austarieren zwischen einem zu großen, zu kleinen oder angemessenen Thema Schwierigkeiten bereitet. Etwas einfacher gestaltet sich die Suche nach Fragestellungen für Studienarbeiten, da diese zumeist einen sehr kleinen Bereich abdecken, der dann wiederum sehr intensiv bearbeitet wird – dennoch müssen die folgenden Phasen bei der Entwicklung der Fragestellung eingehalten werden.

2.1 Phase der theoretischen Fundierung

Wissenschaftliches Arbeiten erfordert immer die Kenntnis des aktuellen Forschungsstandes. Forschung soll einen Beitrag dazu leisten, das Wissen der Welt zu vergrößern, was nur gelingen kann, wenn auf bestehenden Kenntnissen aufgebaut wird. Dies gilt natürlich nur, sofern nicht etwas gänzlich Neues erforscht oder erschaffen wird, was aber außerordentlich selten der Fall ist – und auch hierfür lassen sich in der Regel bestehende Verfahren und Erkenntnisse adaptieren.

Um sich einem Thema zu nähern, muss in einem ersten Schritt der aktuelle Forschungsstand aufgearbeitet werden. Insbesondere für Novizen, die sich neu in ein Forschungsfeld einarbeiten, ist diese Phase mühsam. Allerdings gibt es für die theoretische Fundierung der Forschungsfrage keine Abkürzung. Ein unzureichendes Quellenstudium führt zu Lücken im Argumentationsgang, übersieht bereits vorhandene Lösungen oder benennt Desiderate, die keine sind.

Erst nach dem Abschluss der theoretischen Fundierung erfolgt die genauere Eingrenzung des eigenen Forschungsanteils und somit die Bestimmung des Gegenstands der eigenen Arbeit (siehe hierzu den Beitrag von Boelmann zu Forschungsdesign in diesem Band). Sobald ein *Forschungsfeld* mit einem konkreten *Gegenstand* ausgewählt wurde, besteht zwar noch keine Fragestellung, doch kann das *Thema* des Projekts formuliert werden, etwa: ‚Meine Arbeit wird sich mit der Bedeutung von Leseflüssigkeit beschäftigen'.

2.2 Phase der Präzisierung und Weitung

Nachdem in der ersten Phase aus dem großen Feld aller möglichen Forschungsfelder eines ausgewählt wurde, gilt es nun, sich diesen Gegenstand umfassend zu erschließen, womit in einem ersten Schritt erneut ein Quellenstudium gemeint ist[2]. Hierbei ist es wichtig, die negativen Auswirkungen des Dunning-Kruger-Effekts zu vermeiden: Zusammengefasst besagt der Effekt, dass inkompetente Menschen ihre eigene Inkompetenz nicht erkennen können, da ihnen die Einsicht in das Feld fehlt, die notwendig wäre, um zu erkennen, was sie nicht wissen. Umgekehrt werden sich kompetente Menschen ihrer fachlichen Unzulänglichkeiten stetig bewusster, da sie mit steigendem Überblick über ein Forschungsfeld abschätzen können, was sie noch nicht wissen oder können (vgl. Kruger/Dunning 1999). In dieser Phase der Fragestellungsentwicklung sei somit vor zu schneller Zufriedenheit gewarnt, aber auch vor zu schneller Frustration: Auch ein unübersichtlich und endlos erscheinendes Forschungsfeld lässt sich mit der Zeit erschließen.

Die tiefere Durchdringung des Gegenstands geht mit einer Weitung des eigenen methodischen und fachlichen Horizonts einher, sodass in dieser Phase einerseits das eigene Interesse zunehmend präzisiert, der Blick auf das große Ganze aber geweitet wird. Insbesondere die sich öffnenden weiteren Bezugsfelder, etwa aus den Nachbardisziplinen oder angrenzenden Forschungsbereichen, können an dieser Stelle nicht umfassend erschlossen werden, doch sollten sie für die spätere Arbeit präsent bleiben, da hier wertvolle Anregungen oder Lösungsvorschläge warten könnten. Zentral steht bei der Auseinandersetzung mit dem Forschungsstand die Frage, welche *spezifischen* Probleme sich in einem Feld ergeben, welche Aspekte zu wenig untersucht wurden oder welche Fragen noch gänzlich unbeantwortet blieben. Es sollte nicht unterschätzt werden, dass die meisten Projekte und fachwissenschaftlichen Beiträge, die Antworten auf spezifische Probleme liefern, zugleich weitere Forschungsbedarfe kennzeichnen, denen sich ggf. in einer eigenen Arbeit gewidmet werden könnte.

Zum Abschluss dieser Phase kann das Thema des Projekts sehr präzise benannt werden: ‚Meine Arbeit untersucht die Auswirkungen von Lautlesetrainings mittels funktionaler Texte auf die Leseflüssigkeit von Neuntklässlern der Hauptschule‘.

2.3 Phase der Formulierung und Festlegung

Die abschließende Formulierung der Forschungsfrage erfordert mehr als die Umstellung des Themas in einen Fragesatz. Vielmehr muss hierbei der gesamte folgende Forschungsprozess durchdacht und auf die pragmatische Umsetzbarkeit überprüft werden. Einige Festlegungen, im vorliegenden Beispiel der gewählte *Forschungsansatz* und das notwendige *Forschungsdesign,* ergeben sich organisch

[2] Später kann dies auch über empirische Vorstudien geschehen, doch sei vor einem verfrühten Gang in die Forschungspraxis gewarnt, da dies erhebliche Ressourcen bindet und im Zweifelsfall Aspekte erhoben wurden, die von anderen Forscherinnen und Forschern bereits mit besseren Mitteln erforscht wurden.

aus dem formulierten Thema: Da bereits vorhandene Erkenntnisse in den Unterricht implementiert und ihre Wirkung untersucht werden sollen, entspricht dies dem Forschungsansatz *Design Research* (siehe den Beitrag von Dube in diesem Band) und da ‚Auswirkungen' untersucht werden sollen, wird ein Prä-/Post-Kontrollgruppen-Design (siehe den Beitrag von Pissarek und Wild in diesem Band) benötigt, bei dem ein Leistungsstand vor und nach der Lautlesetraining-Intervention erhoben wird. Da der Gegenstand um den Zusatz ‚mittels funktionaler Texte' erweitert wurde, können dessen spezifischen Ausprägungen nur erfasst werden, wenn eine Kontrollgruppe eine vergleichbare Intervention mit literarischen Texten durchläuft.[3]

Andere Fragen, etwa nach dem eingesetzten *Forschungsparadigma*, den konkreten *Erhebungs- und Auswertungsverfahren* oder der notwendigen Spezifizierung der *Gegenstände* sowie der konkreten Entwicklung des *Forschungsdesigns*, gehen nicht eindeutig aus dem *Thema* hervor, müssen an dieser Stelle aber ebenfalls durchdacht und später in der Entwicklung des Forschungsdesigns (siehe den entsprechenden Beitrag von Boelmann in diesem Band) festgelegt werden.

An dieser Stelle geschieht es häufig, dass durch den Blick in das konkrete Design der Studie Veränderungsbedarfe an der Fragestellung zu Tage treten, was zugleich bedeutet, dass die *Phase der Präzisierung und Weitung* erneut durchlaufen werden muss.

Sobald Design und Fragestellung passend aufeinander abgestimmt sind, kann die Forschungsfrage finalisiert und formuliert werden. Hierbei ist es möglich und bei größeren Forschungsvorhaben üblich, eine übergeordnete Fragestellung in Unterfragen auszudifferenzieren, sodass die innere Struktur des Projekts besser heraustritt.

3. Eine Frage, viele Fragen

Am Beispiel des Vorgehens aus Boelmann (2015) soll in der Folge exemplarisch gezeigt werden, wie sich eine große Fragestellung in verschiedene kleine Fragestellungen untergliedern lässt. Hierbei wird deutlich werden, dass nicht alle Teile der Fragestellung empirisch untersucht werden müssen, sondern auch heuristische Anteile bestehen, in welchen der bestehende Forschungsstand zur Theorie- und Modellbildung genutzt wird oder Grundlagen für die spätere empirische Erhebung herausgearbeitet werden.

Die im Feld der Literatur- und Mediendidaktik angesiedelte Dissertation *Literarisches Verstehen mit narrativen Computerspielen* verfolgt die übegeordnete Fragestellung (Boelmann 2015, 44):

1. Eignen sich Computerspiele als Gegenstände literarischen Verstehens?

[3] Es wäre auch möglich, die Fragestellung so anzupassen, dass verschiedene Typen von funktionalen Texten auf ihre Lernpotenziale hin untersucht werden. Dies zöge aber eine weitere Eingrenzung des Themas nach sich.

Hierbei handelt es sich um Grundlagenforschung, die darauf abzielt, die grundsätzliche Eignung der Computerspiel-Rezeption als Ausgangspunkt für literarische Verstehensprozesse zu erheben. Mit Blick auf die oben genannten Stellschrauben werden zwei *Gegenstände* (Computerspiele, literarisches Verstehen) und der *Fragefokus* (Eignung) explizit benannt, *Rahmung* und *Probandengruppe* werden nicht ausgeführt. Letzteres begründet sich daraus, dass die Tragweite der Ergebnisse nicht auf eine spezifische Population oder Altersstufe begrenzt ist: Da literarische Verstehensprozesse über den schulischen Kosmos hinaus wirken, würde sich der Nachweis der Eignung von Computerspielen als literarisches Medium auf alle Menschen übertragen lassen. Ein exemplarischer Nachweis wäre auf eine Grundgesamtheit generalisierbar.

Als Autor der Studie kann ich berichten, dass bei der Entwicklung der Unterfragestellungen die Überlegung zentral stand, wie sich die Eignung konkret nachweisen lässt. Einerseits wäre es hierzu möglich, bestehende Modelle der Computerspielanalyse und des literarischen Verstehens zu nutzen, um über einen heuristisch angelegten Vergleich eine Passung herauszuarbeiten, andererseits könnte auch empirisch nachgewiesen werden, dass es Probanden gelingt, literarische Verstehensprozesse an Computerspielen zu vollziehen. Die Entscheidung fiel, die übergeordnete Frage in drei Teilen zu beantworten: theoretisch, unterrichtspraktisch und empirisch.

Da zum Zeitpunkt der Entstehung der Arbeit noch keine Modelle vorlagen, die einen theoretischen Vergleich der Potenziale unkompliziert ermöglicht hätten, lauten die ersten Unterfragen der Arbeit:

2. Wie lässt sich literarisches Verstehen modellieren, damit es auch intermedial erfasst werden kann?
3. Nach welchen Kategorien lässt sich das Feld Computerspiele systematisieren, um grundsätzlich für das literarische Verstehen geeignete und ungeeignete Computerspiele voneinander zu separieren?

Im Verlauf des Theoriekapitels der Arbeit (Boelmann 2015, 49-148) wird entsprechend mit dem Bochumer Modell literarischen Verstehens eine Modellierung literarischen Verstehens vorgestellt (vgl. Boelmann/Klossek 2018), die eine intermediale Erfassung von literarischem Verstehen ermöglicht und in einem zweiten Schritt eine genrespezifische Systematik von Computerspielen entwickelt, die das große mediale Feld überschaubar macht und so geeignete von ungeeigneten Spielen trennt.

Die eigentliche Beantwortung der Ausgangsfrage erfolgt in einem zweischrittigen Verfahren, indem die zuerst unterrichtspraktische Dimension heuristisch hergeleitet wird (Boelmann 2015, 149-184) und anschließend eine empirische Absicherung der bis dahin generierten Ergebnisse erfolgt (Boelmann 2015, 185-272). Die unterrichtspraktische Herangehensweise weist die Eignung von sogenannten *narrativen Computerspielen* für die Erreichung curricularer Ziele nach, indem theoriegeleitet die herausgearbeiteten Potenziale und strukturellen Eigenschaften des Mediums auf die Anforderungen der Kernlehrpläne angewendet werden:

4. Wie kann der Einsatz von narrativen Computerspielen zum Erreichen curricularer Ziele konkretisiert werden?

Im Verlauf der Arbeit zeigt sich, dass alle curricularen Ziele, die für die Arbeit mit Literatur vorgesehen sind, auch mit Computerspielen erreicht werden können. Ein erster Beweis der Anschlussfähigkeit dieses neuen Mediums für schulische literarische Verstehensprozesse ist damit erbracht.

Die bis dahin geleistete Forschung erlaubt aber noch keine qualitative Aussage über die Potenziale des Computerspiels im Vergleich zum literarischen Text. Dieser Nachweis wird empirisch erbracht:

5. Welche Anforderungen an die literarische Kompetenz stellen narrative Computerspiele im Vergleich zu literarischen Texten?

Hierfür werden zwei Gegenstände, ein literarischer Text und ein narratives Computerspiel, ausgewählt, die einen „ähnlichen Rezeptions-Schwierigkeitsgrad, inhaltliche Parallelen und zudem strukturelle Ähnlichkeiten“ (Boelmann 2015, 206) aufweisen, und die Ausprägung des Verstehens durch eine Lerngruppe von 29 Schülerinnen und Schülern kontrastierend gegenübergestellt. Hierbei erhärtet sich der zuvor heuristisch hergestellte Befund, dass sich narrative Computerspiele ohne Einschränkungen für die Initiierung literarischer Verstehensprozesse eignen, auch empirisch.

Ohne an dieser Stelle näher auf das Design der Studie einzugehen (dies geschieht im Beitrag zu Forschungsdesigns mit einmaliger Erhebung in diesem Band), zeigt der Blick auf die fünf Fragen auf, wie einerseits die übergeordnete Fragestellung die Arbeit gliedert und andererseits die vier Unterfragen das Vorgehen präzisieren und die Struktur der Forschungsarbeit vorgeben.

4. Fazit

Die Fragestellung einer Arbeit nimmt bereits zu Beginn eines Projekts dessen Ende in den Blick und verhilft in einem ersten Schritt bei der Annäherung an das Forschungsfeld und in einem zweiten bei der Ausdifferenzierung des eigenen Forschungsinteresses. Sobald dies gelungen ist, gestaltet sich der weitere Fortgang des Projekts leichter, da alle relevanten Arbeitsschritte und Zwischenstationen hergeleitet werden können.

Es lohnt sich, der Entwicklung der Fragestellung die notwendige Zeit zu widmen, da mit ihr nicht nur der/die Forschende den eigenen Erkenntniszuwachs reguliert und zunehmend klarer formulieren kann, was exakt Teil der Forschung sein soll und was nicht, sondern auch die Tragweite und Relevanz des gesamten Projekts definiert werden.

Literatur

Boelmann, Jan M. (2015): Literarisches Verstehen mit narrative Computerspielen. Eine empirische Studie zu den Potenzialen der Vermittlung von literarischer Bildung und literarischer Kompetenz mit einem schüleraffinen Medium. München: kopaed.

Boelmann, Jan M./Klossek, Julia (2018, i.V.): Literarische Kompetenz erheben, literarische Bildung fördern. Das Bochumer Modell literarischen Verstehens.

Kruger, Justin/Dunning, David (1999): Unskilled and unaware of it. How difficulties in recognizing one's own incompetence lead to inflated self-assessments. In: Journal of Personality and Social Psychology, 77, 6, 1121-1134. Online unter: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.2655&rep=rep1&type=pdf (letzter Zugriff: 01.08.2018).

STEFFEN GAILBERGER

Grundlagenforschung

Wissenschaftstheoretische Reflexionen aus deutschdidaktischer Perspektive

1. Problemstellung

Der vorliegende Artikel[1] befasst sich mit einem schillernden Begriff im Rahmen sprach-literarischen Lernens. Um illustrieren zu können, worin diese Begriffsunschärfe begründet liegt und warum es schwieriger ist, in deutschdidaktischen Kontexten von Grundlagenforschung zu sprechen, als es etwa in Forschungsgebieten der Elementarteilchenphysik, der Biochemie oder der Kosmologie der Fall ist, werden zunächst drei authentische Beispiele skizziert.

Beispiel 1: Beim ersten *PISA*-Test des Lesens im Jahre 2000 erzielten die Schülerinnen und Schüler des Landes Schleswig-Holstein auf nationaler Ebene unterdurchschnittliche Ergebnisse. Im Vergleich zum Mittelwert der gesamtdeutschen Population betrug der Rückstand sechs, im Vergleich zum damaligen Spitzenreiter Bayern sogar 32 Punkte (vgl. Stanat et al. 2002, 16; ausführlich in Stanat et al. 2001). Im Lichte der ernüchternden Resultate wurden anschließend in Zusammenarbeit des Instituts für Qualitätsentwicklung an Schulen Schleswig-Holstein (IQSH) mit dem Kieler Ministerium für Bildung, Wissenschaft und Kultur verstärkt regionale und zentrale Fortbildungsveranstaltungen auf Kreis- und Landesebene ebenso wie verbindliche, wissenschaftlich wie empirisch begleitete Förderprojekte[2] initiiert. Für diese wurden spezielle Unterrichtsmaterialien entwickelt, die die Schülerinnen und Schüler beispielsweise bei der selbstständigen Erschließung von Texten mittels empirisch bewährter Lesestrategien fördern

[1] Das Entstehen des Textes wurde begleitet von inspirierenden und gewinnbringenden Gesprächen. Hierfür möchte ich Michael Krelle und Volker Frederking herzlich danken.

[2] Beispielhaft hierfür sind etwa die seit zwanzig Jahren jährlich stattfindenden Landesfachtage Deutsch an der Christian-Albrecht-Universität zu Kiel ebenso wie Projekte wie *Lesen macht stark* zu nennen. Abrufbar unter http://nzl.lernnetz.de/index.php/lesen-macht-stark.html (letzter Zugriff: 01.08.2018).

sollten.[3] Auswirkungen dieses Bemühens scheinen sich nun bemerkbar zu machen. Zumindest legen dies die vergleichenden Daten der Bildungstrends von 2009 und 2016 zum Kompetenzbereich *Lesen* nahe. Neben Ländern wie Sachsen oder Bayern, fällt auch der Anteil der Schülerinnen und Schüler aus Schleswig-Holstein, die im Lesen mindestens den Regelstandard für den Mittleren Schulabschluss (MSA) erreichen, mit 54% signifikant höher als der deutsche Gesamtwert (48%) aus (vgl. Stanat et al. 2016, 133f.). Umgekehrt verfehlen in Schleswig-Holstein ‚nur' 17,7% der jugendlichen Leserinnen und Leser den Mindeststandard des MSA, während der Wert im gesamtdeutschen Mittel 23,4% beträgt (ebd., 134). Insgesamt konnte damit der Anteil der schleswig-holsteinischen Schülerinnen und Schüler, deren Lesekompetenz mindestens dem Regelstandard für den MSA entspricht, statistisch signifikant um 6 Prozentpunkte gesteigert werden. Damit liegt das nördlichste Bundesland nun signifikant über dem deutschen Gesamtwert (ebd., 143), während es (wie soeben gezeigt) im Jahr 2000 noch unterdurchschnittliche Werte aufwies (vgl. Stanat et al. 2002, 16).

Beispiel 2: Hanno Freys Dissertation zur Förderung des Lesens (Frey 2010) wurde in der deutsch- bzw. lesedidaktischen Diskussion bislang kaum rezipiert (vgl. etwa Philipp 2015; Rosebrock/Nix 2017; Kepser/Abraham 2016), weswegen davon auszugehen ist, dass ihre Ergebnisse auch in Fortbildungsarrangements der Landesinstitute eine untergeordnete Rolle spielen dürften. Ersteres ist eigentlich verwunderlich, letzteres als misslich zu bezeichnen. Denn Frey expliziert einen methodischen Weg, wie mithilfe von systematisch eingeführten Strategien zur Bewusstmachung von Rezeptionsprozessen die Lesekompetenz von Schülerinnen und Schülern (hier von Achtklässlerinnen und Achtklässlern an Gymnasien) nach einer Förderdauer von nur vier Wochen um zwei Schuljahre gesteigert werden kann (vgl. Frey 2010, 125ff.). Sein Förderkonzept basiert auf der Leseprozesstheorie der *DESI*-Studie (vgl. Willenberg 2007a), ebenso wie auch seine empirische Evaluation mithilfe der curricular validen *DESI*-Lesetests zu zwei Zeitpunkten durchgeführt wurde (vgl. Willenberg 2007b; Nold/Willenberg 2007). Hierzu unterteilte Frey die einzelnen Schritte seines Förderprojektes entlang der von Willenberg postulierten sechs Teilfähigkeiten des Lesens (vgl. Willenberg 2007b, 105ff.), um diese durch entsprechende Bewusstmachungsstrategien zur metakognitiven Steuerung des eigenen Leseprozesses zu didaktisieren. Die systematische Förderung des Lesens erfolgt bei Frey somit u.a. über Strategien zur Förderung der *Inferenzbildung* (Leseanforderung 2 bei Willenberg), der *Textgliederung* zur Findung passender Überschriften (Leseanforderung 3 bei Willenberg) sowie zur graphischen Darstellung mithilfe von Concept Maps (im Sinne der Leseanforderung 6 bei Willenberg; vgl. Frey 2010, 82-97). Die teilnehmenden Jungen und Mädchen verzeichneten nach den von Frey vorgesehenen 14 Unterrichtsstunden einen durchschnittlichen Lesekompetenzzuwachs von zwei Schuljahren

[3] Lesetagebücher für die Eingangsphase, die Jahrgangsstufen 3 und 4 sowie schulartübergreifend für die Jahrgangsstufen 5-7; abrufbar unter http://www.lesezeit.lernnetz.de, siehe auch https://www.bildungsserver.de/onlineressource.html?onlineressourcen_id=24974 (letzter Zugriff: 01.08.2018).

[sic!], wobei gesagt werden muss, dass eine empirische Überprüfung seines Programms für Gesamt-, Ober- bzw. Stadtteilschulen noch aussteht.

Beispiel 3: In der Lehrprobe des Unterrichtsfaches Deutsch zum Abschluss seines Referendariats (in Hamburg *Unterrichtspraktische Prüfung* genannt) erzielte ein hier nicht näher genannter Kandidat im Sommer 2015 eine sehr gute Note, da ihm – trotz einiger Fehler, die ihm in den gezeigten 90 Minuten unterlaufen waren – im besonderen Maße der innovative Charakter seiner Stunde (hier v.a. im Sinne von Schul- und Unterrichtsentwicklung) zugutegehalten wurde. In der Lehrprobe didaktisierte der Hamburger Referendar das (damals noch dreidimensionale) Kompetenzstrukturmodell der *LUK*-Studie (nach Frederking 2013). Hierzu teilte er seine zwölf Schülerinnen und Schüler der elften Jahrgangsstufe für ein Gruppenpuzzle (das wiederum in das Setting einer nachgestellten Redaktionskonferenz eingegliedert war) nach den *LUK*-Dimensionen der *semantischen*, *idiolektalen* und *kontextuellen* literarästhetischen Urteilskompetenz in drei unterschiedliche Arbeitsgruppen ein, die mithilfe ausgewählter Materialien auf je unterschiedliche Weise und auf eigenen Pfaden dazu beitrugen, Franz Kafkas maximal polyvalente Parabel *Gibs auf* für sie subjektiv wie intersubjektiv handhabbar und damit verhandelbar zu machen (vgl. ebd., 134ff.). Nach der Reflexion der Unterrichtsstunde wurde deutlich, dass nicht allein der Kandidat das Setting für sich erfolgreich abschloss: Auch die Schülerinnen und Schüler gaben an, durch die Explikation der drei Ebenen von *LUK* nun gezielter (und damit auch zielführender) an literarische Rezeptionsprozesse herangehen zu können.

So dimensional unterschiedlich die drei genannten Beispiele auch gelagert sind:

- das erste fokussiert auf bildungspolitische Entscheidungen und entsprechende Konsequenzen eines Bundeslandes auf Ergebnisse eines groß angelegten Large-Scale-Tests des Lesens (und damit auf nationales Bildungsmonitoring),
- das zweite auf deutschdidaktische Forschung und Qualifikation (mit dem Ziel der Promotion),
- das dritte schließlich auf Unterrichtsplanung und -durchführung, also auf methodische Anwendung mit anschließender Reflexion (in Referendariat und Unterricht),

sie eint doch eine Voraussetzung, die auf dem Weg zu einer Begriffsexplikation hervorzuheben ist: In allen drei Beispielen (*Schleswig-Holstein*, *Frey*, *Lehrprobe*) wird sich mit *PISA*, *DESI* und *LUK* ursprünglich auf Studien bezogen, die allesamt der sprachlich-literarischen Grundlagenforschung im deutschsprachigen Raum nach 2000 zuzurechnen sind. Zugleich zeigen die gewählten Beispiele aber ein begriffliches und damit ein wissenschaftstheoretisches Dilemma auf, da sie im Anschluss ihrer Bezüge auf eben diese Studien eine wichtige Grundvoraussetzung grundlagenintendierter Forschung im engeren Sinne verletzen: die nutzungsorientierte Überwindung der *Zweckfreiheit* mit dem Ziel der *praktischen Anwendung* (vgl. etwa Mittelstraß 1992, 62; Carrierer 2011, 10; Zintzen 2000, 13f.; Wegner 2000, 36; Helmchen 2011, 27; OECD 2015, 47) – wenn auch (wie angemerkt) in unterschiedlichen Dimensionen.

Um dieses scheinbare Paradoxon aufzulösen, d.h. um klären zu können, inwiefern in einer auf lange Sicht naturgemäß anwendungsorientierten Disziplin wie der Deutschdidaktik (als einer Wissenschaft zur konzeptionellen Entwicklung, theoretischen Fundierung und empirischen Erforschung sprachlichen Lernens und seiner langfristigen Förderung innerhalb und außerhalb des Deutschunterrichts) Grundlagenforschung betrieben werden kann, obwohl sie aufgrund eben dieser fachdidaktisch bedingten Anwendungsorientierung *per definitionem* nicht zweckfrei sein bzw. bleiben kann, wird zunächst geklärt, was allgemein und wissenschaftstheoretisch unter dem Begriff Grundlagenforschung (bzw. ‚reine' Grundlagenforschung) verstanden wird. Darauf aufbauend wird die Frage beantwortet, in welchem Verhältnis Grundlagenforschung zur Anwendungsforschung steht. Diese zwei vorbereitenden Schritte sind notwendig, um schließlich Grundlagenforschungen vorstellen (und benennen) zu können, die als *genuin deutschdidaktisch* zu bezeichnen bzw. im Bereich sprachlich-literarischen Lernens und Testens zu verorten sind.

2. Begriffsklärung: Was ist Grundlagenforschung?

2.1 Kann Grundlagenforschung ‚rein' sein?

Der Begriff Grundlagenforschung wurde aus deutschdidaktischer Perspektive eingangs als schillernd und damit als definitorisch klärungsbedürftig bezeichnet. Das hat unterschiedliche Gründe:

- Ein erster Grund hierfür ist darin zu sehen, dass auf Grundlagenforschung als wissenschaftstheoretischer Begriff innerhalb der deutschdidaktischen Methodendiskussion im Grunde gänzlich verzichtet wird und somit naturgemäß auch nicht als klärungswürdig angesehen werden kann. Wie etwa bei Neumann und Mahler (2014), Kämper-van den Boogaart und Spinner (2010), Feilke und Pohl (2014), Pohl und Ulrich (2011), Bredel und Reißig (2011) oder bei Kammler und Knapp (2002) zu sehen ist, gilt dies sowohl für die theoretische wie empirische Methodendiskussion.
- Wo er doch Verwendung findet (und dies gilt ebenso für das breite Feld der empirischen Lehr-Lern-Forschung in direkter oder indirekter Nachbarschaft zur Deutschdidaktik), kann weiter die Beobachtung gemacht werden, dass der Begriff Grundlagenforschung nur beispielhaft und/oder implizit bemüht wird, anstatt ihm zunächst einer explizierenden Definition zu unterziehen, an der sich (theoretische oder empirische) Beiträge flankierend orientieren können (vgl. etwa Pflugmacher 2016; Nickel-Bacon 2006; Frickel/Kammler/Rupp 2012; Bayrhuber et al. 2012; Koch-Priewe 2004; Walter 2001; als seltene Ausnahme seien hier Köster 2016 und Frederking 2016 genannt).
- Erschwerend kommt hierbei als dritter Grund hinzu, dass Grundlagenforschung nicht als konkrete Methode oder gar als ‚Schule' theoretischen oder empirischen Arbeitens zu umreißen ist (vgl. hierzu z.B. die Beiträge von Pissarek und Schieferdecker in diesem Band), sondern zunächst einer begrifflichen und damit wissenschaftstheoretischen Klärung bedarf, ehe diese auf

> Wissenschaftsdisziplinen, konkrete Methoden oder gar ‚Schulen' bezogen werden kann.

Schlägt man zum Ziele einer ersten groben Orientierung in der von Jürgen Mittelstraß (2008) herausgegebenen *Enzyklopädie Philosophie und Wissenschaftstheorie* unter dem Lemma *Grundlagenforschung* nach, so werden sogleich die beiden Diskussionskorridore deutlich, die auf dem Weg einer deutschdidaktischen Begriffsklärung notwendigerweise nachgezeichnet und dabei aufgehellt werden müssen. Dort heißt es nämlich:

> Terminus der Wissenschaftstheorie: 1. Die wissenschaftl. Beschäftigung mit dem systemat. und method. Fundament einer wissenschaftl. Disziplin, die wissenschaftl. Bemühung also um deren method. erste Schritte, Ziele und grundlegende Verfahrensweise. […] 2. Insbes. in den Natur- und Technikwissenschaften im wesentl. die nicht auf Anwendung hin orientierte, zweckfreie Forschung […]. Daß die nicht anwendungsbezogenen Gegenstände der Forschung dabei >Grundlagen< genannt werden, verbindet sich häufig mit der unbegründeten Meinung, zweckfrei angesetzte Forschung (z.B. über Naturphänomene) schaffe quasi durch eine List der Wissenschaftsentwicklung die Grundlagen für angewandte Wissenschaft, insbes. Technologie. Ein Plädoyer für die Förderung von nicht an bestimmten Einzelzwecken orientierter Wissenschaft lässt sich jedoch offenbar dann berechtigt führen, wenn diese ein für sehr verschiedene Anwendungen wesentliches Wissen allgemein, d.h. aus dem Zusammenhang jeder bestimmten Anwendung herausgelöst, bereitstellen soll. (Kambartel 2008, 233)

Die gegebene Definition kann uns aus *zwei* Gründen nicht befriedigen.

Folgte man ihr unwidersprochen, so bedeutete Grundlagenforschung in der Deutschdidaktik (*erstens*) das theoretische wie empirische Ausloten ihres systematischen und methodischen Fundaments als Wissenschaft, mit dem Ziel, sich der eigenen Disziplin methodisch wie intentional überhaupt erst bewusst werden zu können.

Für eine Disziplin wie die Deutschdidaktik, die zwar (verglichen mit anderen Geisteswissenschaften) keine besonders lange Wissenschaftstradition, aber doch fünf Jahrzehnte intensiver, vor allem theoretischer (mehr und mehr aber auch empirischer) Forschung aufzuweisen hat, muss die in diesem ersten Abschnitt mitschwingende wissenschaftstheoretische *Exklusivität* verwundern, wenn nicht gar auf Ablehnung stoßen. Hat die Deutschdidaktik – so ließe sich schließlich einwenden – denn kein systematisches und methodisches Fundament? Hat sie keine eigenen Methoden, Ziele und Verfahrensweisen entwickelt, um mit diesen als fachdidaktische Disziplin die Grundlagen des Deutschunterrichts und des sprachlich-literarischen Lernens zu erforschen?

Hinsichtlich der ferner formulierten Einforderung nach *Zweckfreiheit* grundlagenintendierter Forschung wäre (*zweitens*) zu diskutieren, ob Grundlagenforschung generell ihren Status *als* Grundlagenforschung einbüßte, wenn sich ihre Forschungen und Erkenntnisse („durch List der Wissenschaftsentwicklung") zu Grundlagen für andere, nämlich für *angewandte* Wissenschaften (weiter)entwi-

ckelten *oder* ob es umgekehrt nicht vielmehr im besonderen Maße als förderwürdig gilt, vor allem jene Grundlagenforschungen (auch mit öffentlichen Mitteln) voranzutreiben, die aus inter- oder transdisziplinärer Sicht überhaupt erst wesentliches (allgemeines) Wissen generiert, das in anderen (inter- oder transdisziplinären) Forschungszusammenhängen erneut zur *Anwendung* kommen kann oder gar soll.

Was aber – so muss schließlich die Anschlussfrage lauten – soll man denn mit Ergebnissen machen, die einer deutschdidaktisch zu nennenden Grundlagenforschung entspringen und die dennoch (oder gerade deswegen) unterrichtliches sprachliches Lernen verbessern, das heißt also zur Anwendung kommen sollen? Etwa wegschließen?

Es bedarf also augenscheinlich einer tiefergehenden Analyse, um klären zu können, ob der Begriff Grundlagenforschung nicht vielleicht doch bereits eine deutschdidaktische Entsprechung hat, oder aber ob die Deutschdidaktik zumindest eine solche erlangen kann. Zu diesem Ziele sind im Folgenden die oben bereits angedeuteten zwei Diskussionskorridore zu durchschreiten.

a) Hierzu wird einerseits jener (dreischrittige) wissenschaftshistorische Prozess nachgezeichnet, der von einer Fachdisziplin zunächst durchlaufen werden muss, möchte sie einen Status erlangen, der es ihr (wissenschaftstheoretisch gesehen) überhaupt erst erlaubt, grundlagenintendiert Forschungsfragen, Methoden und Ergebnissen zu generieren. Dabei Bezüge zu Geschichte und Gegenwart der Deutschdidaktik, d.h. zu ihren theoretischen wie empirischen Diskussionsstadien von 1970 bis heute herzustellen, liegt auf der Hand.

b) Dieser historisch orientierte wissenschaftstheoretische Klärungsversuch wird anschließend terminologisch, und das heißt dann freilich auch inhaltlich ergänzt. Hierzu werden die Begriffe „reine Grundlagenforschung“ (vgl. etwa Mittelstraß 1992, 60) einerseits und „anwendungsorientierte“ Grundlagenforschung (vgl. etwa Brüggemann/Bromme 2006) andererseits miteinander verglichen und (wenn nötig) voneinander abgegrenzt – nicht zuletzt deswegen, da eine Fachdidaktik, wie die des sprachlich-literarischen Lernens im Deutschen, naturgemäß als eine genuin „anwendungsorientierte“ (Frederking 2014) bzw. „eingreifende“ (Kepser 2013) Wissenschaftsdisziplin gelten muss.

Zu a) Die *wissenschaftshistorische* Perspektive nimmt Wolfgang van den Daele ein. Er zeichnet in seinem mittlerweile als kanonisch geltenden Beitrag *Autonomie contra Planung: Scheingefecht um die Grundlagenforschung?* (1975) den Entwicklungsprozess einer sich im Konstituierungsvorgang begriffenen Fachdisziplin mithilfe eines *Drei-Phasen-Modells* nach, an dessen (vorläufigem) Ende überhaupt erst die Voraussetzungen für grundlagenintendiertes Forschen erfüllt seien. Dabei lag sein genuines Interesse ursprünglich gar nicht so sehr auf dem Feld der Wissenschaftsgeschichte als vielmehr auf der Frage, was eine „reife“ (ebd., 32) Wissenschaft ausmache, deren Reife schließlich daran zu erkennen sei,

dass von externer Seite (d.h. politisch, ökonomisch oder sozial) versucht werde, planerisch auf sie Einfluss zu nehmen. Wie aber erlangt eine Disziplin ‚Reife'?

1. Der mit van den Daele hierfür nachzuzeichnende Entwicklungsprozess einer wissenschaftlichen Disziplin, „sozusagen seine Lebenskurve" (ebd., 29f.), beginnt mit einer Phase, die wahlweise als „explorativ", „empiristisch" oder als „vorparadigmatisch" bezeichnet werden kann (ebd.). Diese *erste* Phase ist durch den Zustand gekennzeichnet, dass zwar auch hier bereits so etwas wie ‚Grundlagenforschung' (in einem sehr weiten, mehr oder weniger *vor*wissenschaftlichen Sinne) betrieben werden kann, dass diese aber *genau deswegen* als primitiv und als „Art of the Soluble" (d.h. als die Kunst des Lösbaren) gelten muss (ebd.). Verantwortlich hierfür ist v.a. die in diesem Stadium noch fehlende Theoriedynamik, die diese frühe Phase auf negative Weise kennzeichnet und definiert. Zwar ist, in den Worten van den Daeles, „die Theorie des Gegenstandsbereichs das Ziel der Wissenschaft, aber zunächst liegt ihre Akkumulation mehr im Bereich von Daten und Tatsachen als in der Entwicklung und Prüfung von Modellen und Theorien", sodass in dieser Phase also „Entdeckungen einen Vorrang vor Erklärungen" haben (ebd.).

   Überträgt man diese (*allgemein*wissenschaftlichen) Gedanken auf den Deutschunterricht, ließe sich die erste Phase im Sinne van den Daeles auf die Geschichte des Faches bis zum Jahr 1970 datieren, in der Vertreter wie etwa Robert Ulshöfer (1952ff.) und Hermann Helmers (1966), bzw. ein Jahrhundert zuvor Robert Heinrich Hieckes (1841) oder Philipp Wackernagel (1842) erstmals (und jeweils in ihrer Zeit verhaftet) den Versuch unternahmen, eine fachdidaktische Fundierung des Deutschunterrichts vorzunehmen (vgl. Kämper-van den Boogaart 2010). Letztere richteten ihre Didaktik dabei eindeutig an der Förderung national geprägten Denkens aus (vgl. Frederking/Abraham o.J., 2), erstere bemühten sich um eine „Erneuerung im Horizont klassisch-konservativer Bildungsvorstellungen" nach 1945 bzw. um formalästhetische Aspekte von Sprache und Literatur (ebd.), womit Deutschunterricht und Deutschdidaktik erstmals einer *explorativen Systematisierung* unterzogen worden war.

2. Die *zweite* Phase des wissenschaftlichen Entwicklungsprozesses eines Faches ist van den Daele zufolge durch die Überwindung der ersten Phase, und das heißt durch eine deutliche Zunahme an kognitiver Eigendynamik, gekennzeichnet. Er nennt sie daher auch die Phase der „Theoriedynamik" bzw. der „Paradigmatisierung." (ebd., 30) Sie beginnt, wenn das Forschungsprogramm nicht mehr durch explorativ-empiristische Ansätze abgedeckt werden kann (s.o.), sondern wenn es durch die Implikationen und Probleme von Erklärungsansätzen (d.h. durch Modelle, Hypothesen und deren Stützung oder Verwerfung etc.) bestimmt wird. Die dringlichen Aufgaben der Wissenschaft in dieser Phase liegen in der Aufbereitung von begrifflichen Ungereimtheiten der Erklärungsansätze, in der Ausdehnung dieser auf verwandte oder benachbarte Phänomene, in der Transformierung der Modelle und ihrer Integration

in andere theoretische Konzepte. Vor allem aber zeichnet sich diese *zweite* Phase durch eine erste experimentelle Realisierung durch Anwendung im Feld aus. In dieser Phase der Autonomie – so ließe sich etwas salopp formulieren – ist die Wissenschaft ‚ganz bei sich' und vor externen Planungsintentionen gefeit, so lange nämlich, wie der theoriedynamische Prozess anhält und noch zu keinen eindeutigen (d.h. hier: empirischen) Ergebnissen geführt hat.

Wissenschaftshistorisch ist diese Phase für die Deutschdidaktik auf die Jahrzehnte nach 1970 zu datieren, Jahrzehnte der (fast ausschließlich) theoretischen Auseinandersetzungen, die vor dem Hintergrund von Ideologiekritik, Strukturalismus, Rezeptionsästhetik, Poststrukturalismus, Dekonstruktion etc. geführt wurden (vgl. Kammler 2010). Aus heutiger Perspektive ist es allerdings schwerlich möglich darüber zu entscheiden, ob die *zweite* Phase mit Beginn des neuen Jahrtausends als abgeschlossen bezeichnet werden sollte (Stichwort: Leseforschung, Lesedidaktik, Leseförderung), oder ob es dafür noch zu früh ist (Stichwort: Literarisches Lernen, Kreatives Schreiben, Digitalisierung etc.).

3. Die für die Phase 2 soeben als konstitutiv angenommene *Autonomie* einer Wissenschaft unterliegt jedoch einer Dialektik, da ihr diese im Zuge ihres eigenen Erfolges wieder genommen wird. Diese Beobachtung kennzeichnet schließlich die *dritte* Phase bei van den Daele. Ihr gibt er den Namen „Finalisierungsphase" (ebd.), da sie auf der als gelungen geltenden *theoretischen* Grundierung des Faches fußend den relativen Abschluss der Theoriedynamik und ihre gleichzeitige Überwindung markiert. Dieser Abschluss zeigt sich schließlich in einer *weiter*gehenden Spezifikation der zuvor entwickelten theoretischen Grundlagen durch *weiter*führende Forschung, die an dieser Stelle erstmals (und anders als noch in Phase 1) als theoriegestützte und/oder empirische Grundlagenforschung im eigentlichen Sinne bezeichnet werden kann. Der dabei zu beobachtende (potenzielle oder reale) Verlust von Autonomie einer Wissenschaft z.B. durch politische oder soziale Interessen und Einflussnahmen (Stichwort: *KMK*, *PISA*, *IGLU*, *VERA*, *IQB*-Bildungstrend etc.) ist in der Perspektive van den Daeles aber nicht von vornherein als negativ zu bewerten (vgl. etwa Wintersteiner 2007). Er erkennt darin vielmehr einen Ausweis von „Reife" der betreffenden Wissenschaft (ebd., 32), die sich (*im* und *durch* den Prozess ihrer Reifung) in der Überwindung der explorativen wie der theoriedynamischen Phasen Bahn bricht und damit nunmehr nicht allein die *Darstellung* der theoretischen Grundlagen eines Faches ermöglicht, sondern die *tiefer*liegende und *weiter*führende Durchdringung dieser *als Grundlagenforschung* vorantreibt – eine Sicht, die aus dezidiert geisteswissenschaftlicher Perspektive auch von Hellenkemper (1996) vertreten wird.

Wie ich meine, ist damit die aktuelle Situation weiter Teile der Deutschdidaktik umrissen, verstanden als die Wissenschaft nicht nur zur konzeptionel-

len Entwicklung und theoretischen Fundierung (im Sinne der Phase 2), sondern ebenso auch zur weiteren Durchdringung und schließlich zur empirischen Erforschung des sprachlich-literarischen Lernens innerhalb und außerhalb des Deutschunterrichts: im Kompetenzbereich *Sprechen und Zuhören* (vgl. etwa Behrens 2014; Gätje et al. 2016; Krelle 2013), beim *Schreiben* (vgl. im Überblick Becker-Mrotzek/Grabowski/Steinhoff 2017), beim *Lesen und dem Umgang mit Texten und Medien* (vgl. etwa Baumert et al. 2001; Willenberg 2007b; Frederking/Brüggemann/Hirsch 2016) oder – quer dazu – im Kompetenzbereich *Sprache und Sprachgebrauch untersuchen* (vgl. Bremerich-Vos/Böhme 2009; Oomen-Welke/Bremerich-Vos 2014).

In Bezug auf ihre Wissenschafts- und Theorie*geschichte*, und das heißt dann auch in Bezug auf ihren derzeitigen Forschungsstand, kann damit als geklärt gelten, dass die Deutschdidaktik als wissenschaftliche Disziplin einen Entwicklungsstatus (sprich: ‚Reife') erlangt hat, der es ihr (selbstverständlich, möchte man sagen) erlaubt, theoretische und/oder empirische Grundlagenforschung zu betreiben (ob diese dann extern oder intern intendiert ist, kann an dieser Stelle hintangestellt bleiben und an anderen Orten diskutiert werden).

Zu b) Als ungeklärt markiert bleibt indes die Frage, ob Fachdidaktiken als auf lange Sicht anwendungsbezogene (Frederking 2014), d.h. also als eingreifende (Kepser 2013) Wissenschaften überhaupt Grundlagenforschung betreiben können, wenn Definitionen wie jene aus der *Enzyklopädie Philosophie und Wissenschaftstheorie* ins Feld geführt werden, deren begriffliche Enge den Diskussionskorridor b) markieren.

Bei dessen Sichtung sticht zunächst (und wie bereits durch die *Enzyklopädie* angedeutet) eine immer wieder vorgenommene Differenzierung ins Auge, die die sogenannte „reine" Grundlagenforschung (z.B. Brüggemann/Bromme 2006, 112ff.) auf der einen Seite von der sogenannten „anwendungsbezogenen" Forschung auf der anderen Seite unterscheidet – und damit trennt (ebd.).

Um dieses Gegensatzpaar begrifflich zu konturieren, greift Martin Carrierer etwa auf die Gegenüberstellung von „Verstehen" *vs.* „Können" zurück und stellt dabei fest, dass „reine" Grundlagenforschung mit dem Ziel der Erkenntnis und des Verstehens, „anwendungsorientierte" Forschung aber mit dem mittelfristigen Erreichen *praktischer Ziele* betrieben werde (vgl. Carrierer 2011, 10). In diesem Sinne definiert auch Jürgen Mittelstraß den Begriff der „reinen" Grundlagenforschung eben als Forschung, „deren Ergebnisse keine praktische Anwendung erwarten lassen." (Mittelstraß 1992, 62) Mittelstraß geht dabei aber noch einen Schritt weiter: Bezüglich der „reinen" Grundlagenforschung stellt er den Begriff der „Anwendung" sogar gänzlich in Frage (ebd.) und führt zur Fundierung seiner Argumentation Beispiele aus der Elementarteilchenphysik oder aus der Kosmologie an. Ihre Forschungsergebnisse zur Anwendung zu bringen, wäre Mittelstraß zufolge dann ausschließlich in Sphären des „Science Fiction" möglich – de facto also *un*möglich (ebd.).

Auf das von Mittelstraß damit gezeichnete Bild von Grundlagenforschung als Forschung *um ihrer selbst willen* greift auch Clemens Zintzen in der Einleitung des von ihm herausgegebenen Bandes zur *Zukunft der Grundlagenforschung* zurück (Zintzen 2000, 13f. zum „Selbstzweck der Grundlagenforschung"), die sich, wie auch Gerhard Wegner im selben Band an anderer Stelle erläutert, eben gerade nicht in und durch ihre Anwendbarkeit (in welchem Sinne auch immer) auszeichne, sondern durch ihre „Originalität und Qualität der Fragestellung" (Wegner 2000, 36). In diese Richtung argumentiert auch Andreas Müller (2005). Mit ihm handelt es sich bei Grundlagenforschung um rein „erkenntnisorientierte und zweckfreie Forschung", bei der „der reine Erkenntnisgewinn" im Fokus stehe und die potenzielle Anwendung keine Rolle spiele (ebd., 1; vgl. hierzu auch Helmchen 2011, 27).

Die hiermit versammelten Stimmen der deutschsprachigen Diskussion decken sich freilich mit internationalen Standards, die zuletzt im Auftrag der OECD im sogenannten Frascati-Manual zusammengetragen wurden, in dem es (in eben diesem Sinne zum Begriff *Basic Research*) heißt:

> Basic research is experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts, without any particular application or use in view. (OECD 2015, 47)

Im Sinne eines knappen Zwischenfazits aus deutschdidaktischer Perspektive müssen wir also feststellen, dass der in *a*) erzielten wissenschafts*historischen* Klärung (Stichwort „Lebenskurve" und „Reife" der Deutschdidaktik zur Grundlagenforschung) eine in *b*) deutlich gewordene *begriffliche* Unklarheit gegenübersteht, die sich in einer qualitativ motivierten Distinktion zweier scheinbar verschiedener Arten von Forschung äußert: in eine „reine" (weil *zweckfreie* und damit mögliche) Forschung, die die Bezeichnung Grundlagenforschung tragen darf, und eine „anwendungsbezogene" (weil *zweckgebundene*, mithin praktischen Zielen folgende Forschung), der aus eben diesen Gründen die Bezeichnung Grundlagenforschung augenscheinlich verweigert, also *un*möglich wird.

Aus der Perspektive einer auf lange Sicht anwendungsorientierten Wissenschaft wie die Fachdidaktik Deutsch ist diese von Exklusion begleitete Distinktion nun mindestens diskussionswürdig, wenn ihr nicht gar in Gänze zu widersprechen ist. Intuitiv kann zumindest nicht eingesehen werden, warum beispielsweise die grundlegende Erforschung literarischen Verstehens in der Sekundarstufe 1 (vgl. etwa Frederking/Brüggemann/Hirsch 2016) oder des Erwerbs orthographischer Kompetenzen in den ersten Schuljahren (vgl. im Überblick Thomé 2003) in jenem Moment ihren Status als (reine) Grundlagenforschung verlieren sollte, in dem sie unterrichtliche Relevanz durch konkrete Planung oder Methodenauswahl gewinnt (vgl. hierzu etwa Gailberger 2018a; Fay 2013).

Deswegen wird in den folgenden Abschnitten der soeben deutlich gewordenen exklusiven Distinktion durch eine kritische Reflexion der Begriffe *Anwendung* und *Zweckfreiheit* begegnet, die klären soll, inwiefern Grundlagenforschung ggfs. doch anwendungsbezogen betrieben werden ‚darf' und was Zweckfreiheit aus wissenschaftstheoretischer Perspektive dann überhaupt noch bedeuten kann.

Diese Schritte sind notwendig, um schließlich die bereits mehrfach angedeutete Frage diskutieren zu können, ob Grundlagenforschung wirklich ihren Status als Grundlagenforschung verlieren muss, wenn zuvor generierte Grundlagenerkenntnisse letztlich doch (direkt oder indirekt) zur *Anwendung* gelangen. Die hierbei gewonnenen Antworten werden schließlich die bislang lediglich als ‚intuitiv' deklarierten Zweifel an der vorgenommenen Distinktion wissenschaftstheoretisch untermauern.

2.2 Was bedeutet „theoria cum praxi" und in welchem Verhältnis stehen Grundlagenforschung und Anwendungsforschung?

Um die soeben als qualitativ motivierte wissenschaftstheoretische Unterscheidung von reiner Grundlagenforschung hier und anwendungsbezogener Forschung dort begrifflich fassen, in ihrer inhaltlichen Revisionsbedürftigkeit darstellen und in einem dritten Schritt konzeptionell überwinden zu können, liegt es (nicht zuletzt auch aus fachdidaktischer Perspektive) nahe, auf Donald E. Stokes' wissenschaftssoziologische Studie *Pasteurs Quadrant: Basic Science and Technological Innovation* (1997) zu verweisen, deren Argumentation den wissenschaftstheoretischen Blick auf das Verhältnis von Grundlagen- und Anwendungsforschung zu einem radikalen Perspektivwechsel gezwungen hat. Was ist der Kern von Stokes Argumentation?

Um seine neue Perspektive nachvollziehbar entwickeln zu können, greift Stokes zunächst auf den klassisch zu nennenden Begriff *Basic Research* im Sinne der epochemachenden Expertise *Science: The Endless Frontier*[4] von Vannevar Bush (1945) zurück. Dessen Konzept übernimmt Stokes allerdings nur, um von dort aus aufzeigen und kritisieren zu können, dass eine auf diese Weise gewählte und festgeschriebene Dichotomie von Grundlagenforschung (hier) und Anwendungsforschung (da) im Sinne zweier *entgegengesetzter* Pole aus wissenschaftstheoretischer Sicht scheitern muss. Kritik- und revisionswürdig erschien Stokes dabei nämlich die – zunächst leicht nachvollziehbare – Beobachtung, dass Forschungsvorhaben, die sich stärker dem einen der beiden Pole annähern, sich zugleich *nolens volens* vom anderen entfernen müssen (vgl. Stokes 1997, 9f.) – ein theoretisches Artefakt, das Stokes als wissenschaftshistorisch nicht haltbar ansah und schließlich widerlegte.

Zur Illustration dessen verwies er auf die Wissenschaftsbiographien von Nils Bohr (1885-1962), Thomas Edison (1847-1931) und Louis Pasteur (1822-1895), um (quasi im Sinne des Kritischen Rationalismus vorgehend) zu zeigen, a) dass Bohrs, Edisons und Pasteurs Ansätze eben *nicht* immer eindeutig als Grundlagen-

[4] In dieser plädierte der damalige *Science Adviser* der US-Regierung und Vorsitzende des *National Defense Research Committee* aus wissenschaftsstrategischen Erwägungen dafür, den durch das Kriegsende in die Diskussion geratenen Konnex von „reiner" Grundlagenforschung und Zweckfreiheit auch für die Jahre nach 1945 beizubehalten. (vgl. Bush 1945, o.S.)

forschung *hier* oder als Anwendungsforschung *dort* im Sinne der eindimensionalen Anordnung Vannevar Bushs zu deklarieren sind und b) dass Bushs eindimensionales Modell damit als falsifiziert gelten muss. Konsequenterweise überwindet Stokes im Anschluss dessen die Idee der soeben angedeuteten eindimensionalen Anordnung zugunsten eines zweidimensionalen Modells, bei dem die *Erkenntnis-* und *Anwendungs*interessen von Forschung zwei unterschiedliche Dimensionen markieren, deren hohe bzw. niedrige Ausprägungen schließlich vier Quadranten entstehen lassen.

Nils Bohrs Suche nach einem Modell des Atomaufbaus entsprang Stokes zufolge allein dessen wissenschaftlicher Neugier, ohne dass dieser bereits an eine praktische Verwendung seiner Erkenntnisse denken konnte oder wollte. Bohrs Studien teilt Stokes daher den linken oberen Quadranten mit hoher Ausprägung des Erkenntnisinteresses und geringer bzw. fehlender Orientierung am Nutzen zu (vgl. Stokes 1997, 73f.) – sie entspricht faktisch dem Ideal ‚reiner Erkenntnis', wie er auch in Vannevar Bushs Report für Franklin D. Roosevelt zum Ausdruck kommt (vgl. Bush 1945, o.S.).

Im Gegensatz zu Bohrs Forschungsintentionen kennzeichnen *Thomas Edisons* Bemühungen um profitable elektrische Beleuchtungen ein hohes Interesse am Nutzen aber ein geringes Erkenntnisinteresse. Ihn weist Stokes dem rechten unteren Quadranten zu, der das Bemühen um Anwendung ohne besonderes Interesse an allgemeineren Wissenschaftserkenntnissen symbolisiert (vgl. Stokes 1997, 74).

Die Überwindung der bis dato vorherrschenden eindimensionalen Anordnung hin zum zweidimensionalen Modell durch die Synthese aus Erkenntnis- und Nutzungszielen verdeutlicht sich schließlich im für uns wichtigsten Quadranten, der bei Stokes mit dem Namen *Louis Pasteurs* assoziiert wird (ebd., 63). In den ersten Jahren seiner Forscherlaufbahn befasste sich Pasteur zwar auch mit offenen Fragen der Grundlagenforschung, später bezog er jedoch seine Fragestellungen immer häufiger aus dem Anwendungskontext, etwa aus Problemen in den Fertigungsabläufen in Fabriken. Zum Beispiel erforschte er, wie man vermeiden kann, dass Lebensmittel bei der Herstellung verderben, und entdeckte so, dass manchen Mikroorganismen zum Leben keinen Sauerstoff benötigen (ebd., 12f.). Er verband auf diese Weise entscheidende Fortschritte in der Grundlagenforschung mit direkter Verbesserung von Technologien der Produktion und lieferte wissenschaftlich abgesicherte Informationen für praktische Entscheidungen, etwa durch seine Erkenntnisse über Krankheitsursachen, auf die sich dann Bestrebungen zur Verbesserung der Hygiene berufen konnten (ebd., 22).

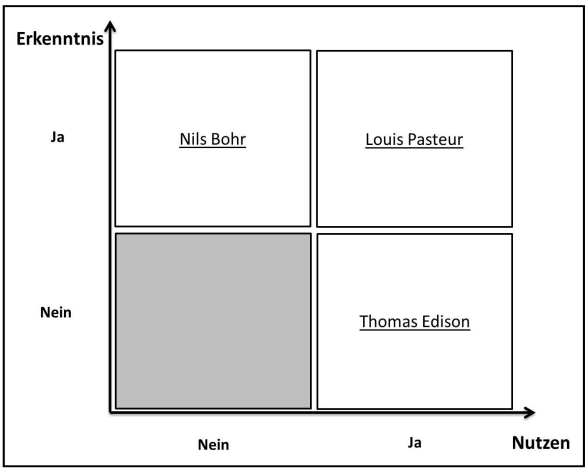

Abb. 1: Quadrantenmodell wissenschaftlicher Forschung (eigene Abbildung nach Stokes 1997, 74)

Louis Pasteurs Art zu forschen, oder genauer: seine aus Interessen, Zielen, Designs, Experimenten und Ergebnisverwertungen gebündelten Projekte bezeichnet Stokes daher als *nutzungsorientierte Grundlagenforschung* (i.O. „used-inspired basic research"; Stoke 1997, 87). Sie ist nicht höherwertig (aber auch nicht minderwertig) als reine Grundlagenforschung, ebenso ist sie nicht höherwertig (resp. minderwertig) als Anwendungsforschung. Doch spielt sie als alternativ zu betrachtende Art von Grundlagenforschung eine wichtige Rolle, da sie ausgehend von zu lösenden Problemen auf *Anwendung*sfeldern dennoch *grundlegende* Erkenntnisse generiert, die dann (sogleich oder auch indirekt und damit später) zur Anwendung gebracht werden dürfen, ohne damit eine Aberkennung des Labels Grundlagenforschung befürchten zu müssen.

Wissenschafts*historisch* betrachtet war dieser Blick auf Grundlagenforschung und der damit verbundenen Idee, „daß wir mit dem, was wir tun, irgendwie nützlich sein können, nützlich sein wollen: nützlich nicht nur uns selbst, sondern für die Zukunft aller" (Winnacker 2000, 22), indes gar nicht so neu. Friedrich Schiller war dieses Thema präsent, als er bereits in seinem Musenalmanach auf das Jahr 1797 spöttelte, dass „die Wissenschaft dem einen die hohe, die himmlische Göttin [ist], dem anderen eine tüchtige Kuh, die ihn mit Butter versorgt." (zitiert nach Winnacker 2000, 23). Vielleicht spielte Schiller damit auf Gottfried Wilhelm Leibniz an, der in seiner Denkschrift die Einrichtung einer *Societas Scientiarum et Artium in Berlin* (Leibniz 1923) betreffend bereits hundert Jahre zuvor forderte, dass es „nicht allein die Künste und die Wissenschaften, sondern auch Land und Leute, Feldbau, Manufacturen und Commercien zu verbessern gelte." (zitiert nach Winnacker 2000, 23)

Seine berühmt gewordene wissenschaftstheoretische Losung einer „theoria cum praxi" (Leibniz, IV: 8, 426), die damit zum Ausdruck kommt, führte in der deutschen Hochschulgeschichte wissenschaftssoziologisch gesehen allerdings genau

zu jener Trennung, die es nun mithilfe des Zweidimensionenmodells von Donald E. Stokes zu überwinden gilt, wenn eben dieser im Rückblick und unter besonderer Bezugnahme auf das quasi ‚deutsche Modell' feststellt:

> The Germans used very different arrangements to support their rapid technological advances in the nineteenth century. They lodged applied science and development in the *Technische Hochschulen* and industry and imparted a new prestige to these technical schools and the careers to which they led. The Germans thereby institutionalized a strong sense of the separation of technology from the pure science lodged in the universities and research institutes. […] The spectacular achievements of the Germans both in pure and in applied science made their system extraordinarily influent. They were so excellent in each that their institutional arrangements were thought to be the natural order of things by an admiring world. Thousands of Americans flocked to the German universities in the late nineteenth and early twentieth centuries. (Stokes 1997, 37f.)

Leibniz' ursprüngliche Losung einer „theoria cum praxi" wurde also aufgetrennt, und zwar unter Überbetonung der „theoria" einerseits (Stichwort: ‚reine' erkenntnisgenerierende Grundlagenforschung an Universitäten), und unter Überbewertung der „praxi" andererseits (Stichwort: pure Anwendungsforschung an Technischen Universitäten, Fachhochschulen oder Forschungsinstituten).

Wissenschaftstheoretisch ist die daraus entspringende dichotome Trennung von (reiner) Grundlagenforschung, die Erkenntnisse (scheinbar) zweckfrei generiert und sammelt, und (reiner) Anwendungsforschung, die aufgrund spezieller Zwecke überhaupt erst auf den Plan gehoben wird, allerdings ein *Artefakt*.

- Ihr artifizieller Charakter entspringt einerseits aus der einzig und allein als potenziell und spekulativ zu deklarierenden Nicht-Anwendung von zuvor gewonnenen Grundlagenerkenntnissen, deren Nicht-Verwertung aber zu keinem Zeitpunkt als garantiert gelten kann (dazu gleich mehr).
- Andererseits wird dabei eine ausschließlich praktischen Zwecken folgende Anwendungsforschung postuliert, die (so verstanden) wiederum keinen Beitrag zur Theoriediskussion bzw. der grundlegenden (empirisch gesicherten) Erforschung der jeweiligen Disziplinen beizutragen habe – eine Sicht, die beispielsweise mit Blick auf die Medizin (Gerok 1996, 128) oder auf die Psychologie resp. Psychiatrie (Helmchen 2011) ebenfalls nicht haltbar ist.

Um eben dieses *dichotome Artefakt* von ‚reiner' Grundlagenforschung und Anwendungsforschung zu belegen, deutschdidaktisch zu reflektieren und damit schließlich zu revidieren, folgt die bereits angekündigte und nun notwendige Überprüfung der Begriffe *Zweck* und Zweck*freiheit* in Bezug auf Forschung und Wissenschaft.

2.3 Kann Forschung ‚zweckfrei' sein?

Wie gezeigt wurde, erlaubt uns Donald E. Stokes' zweidimensional angelegtes Modell von Grundlagenforschung, eben diese als erkenntnis- *wie auch als* nutzenorientiert zu betrachten. Um dies schließlich deutschdidaktisch erweitern und nutzen zu können, müssen nun noch zwei Schritte vollzogen werden: eine Klärung des Lexems „Zweck" (bzw. seiner wissenschaftstheoretischen Negation im Sinne zweck*freier* Forschung), und seine anschließende Tauglichkeitsprüfung, bezogen auf die Stokes'schen Quadranten ‚Erkenntnis' (*sensu* Nils Bohr), ‚Nutzen' (*sensu* Thomas Edison) und ‚Nutzungsorientierte Erkenntnis' (*sensu* Louis Pasteur, in dessen Nähe schließlich auch die deutschdidaktische Forschung zu verorten ist).

Für den rasch getätigten Schritt 1 ziehen wir den Duden zurate. Er definiert „Zweck" als „[Beweggrund und] Ziel einer Handlung", also als „etwas, was jemand mit einer Handlung beabsichtigt, zu bewirken, zu erreichen sucht." (vgl. Duden 2018[5]).

Vor diesem Hintergrund ist es für den nachfolgenden Schritt 2 naheliegend, mit Blick auf das bisher Gesagte zunächst die Anwendungsforschung (im Sinne Edisons) zu skizzieren, schwingt in ihrer Attribuierung doch bereits die Dudendefinition implizit mit: „Nutzen" als Beweggrund, Ziel und Absicht einer Handlung, d.h.: erhoffter oder anvisierter Nutzen durch Forschung und ihre Verwertung.

In diesem Sinne wird anwendungsorientierte Forschung allgemein als Forschung definiert, die „mit Blick auf zumindest mittelfristig realisierbare konkrete praktische Ziele [betrieben wird]" (vgl. etwa Carrierer 2011, 10) und sich dabei auch „[…] mit Fragen konfrontiert [sieht], die von Nicht-Wissenschaftlern für dringlich gehalten werden […]." (ebd.; vgl. aber auch Brüggemann/Bromme 2006, 113; Mittelstraß 1992, 63f.) Dieser Nexus von Anwendungsforschung und der Orientierung an einer nützlichen Verwertung ihrer Erkenntnisse sollte allerdings nicht von vornherein als wertend missverstanden und damit a priori missbilligt werden. Bezogen auf die Deutschdidaktik können solche Nicht-Wissenschaftler beispielsweise Eltern sein, die zur sprachlichen Kompetenzförderung ihrer Kinder oder zur Reduzierung von Lernschwierigkeiten spezielle Lernstrategieprogramme einfordern, deren Effektivität im besten Fall empirisch geprüft sein sollten.

Als „Transferforschung" begriffen, die nämlich „auf praktische Probleme [reagiert]", d.h. also Erkenntnisse *transferiert*, definieren auch Brüggemann und Bromme (2006, 114) „Anwendungsforschung [als Forschung], bei der nicht ein Erkenntnisinteresse, sondern [allein] die praktische Problemlösung im Vordergrund steht." (ebd., 113) Bleiben wir bei unserem Beispiel von eben, so wären in der Terminologie von Jürgen Mittelstraß (1992) Evidenzstudien zu Lesestrategieprogrammen wie die *Textdetektive* (vgl. Gold 2007) oder *LESEN. Das Training* (Bertschi-Kaufmann et al. 2007) als „produktorientierte Anwendungsforschung"

[5] Vgl. https://www.duden.de/rechtschreibung/Zweck (letzter Zugriff: 01.08.2018).

zu bezeichnen (Mittelstraß 1992, 63), als Forschung also, „die entweder bereits mit Blick auf besondere Anwendungen stattfindet oder solche Anwendungen kurzfristig erwarten läßt." (ebd., 64)

In Feldern der *Anwendungs*forschung kann also eindeutig nicht nur nicht von Zweckfreiheit in Forschung und Wissenschaft gesprochen werden – wie zu erwarten war, verhält es sich vielmehr umgekehrt: Ohne erwartbaren Nutzen, der sofort oder zumindest mittelfristig zu erwarten ist, würden Untersuchungen zu konkreten Anwendungsfragen gar nicht erst auf die Schienen gehoben. Sie blieben damit unerforscht.

Dessen eingedenk, müsste nun eigentlich das genaue Gegenteil von der erkenntnisgenerierenden Grundlagenforschung (im Sinne Bohrs) erwartet werden: Ihr müsste jede Zweckorientierung fremd sein. Entsprechend wurden in Kapitel 2.1 jene Stimmen zitiert, die eine intentionale und zweckdienliche Nutzung von Grundlagenforschungen strikt zurückweisen. Diese Annahme erweist sich allerdings bei einer tiefer gehenden Analyse als nicht haltbar, u.a. deswegen, da sich wissenschaftliche Erkenntnis zuweilen Wege bahnt (und dadurch ‚Nutzen' bringt), die zuvor so gar nicht intendiert waren. Hierzu ist erneut zu unterscheiden zwischen einer *allgemeinen* Betrachtungsweise und einer *speziellen*.

- Erstere ist banal und verweist lediglich auf das, was bereits implizit geklärt wurde, nämlich dass Grundlagenforschung per definitionem den Zweck verfolgt, fachspezifische Erkenntnisse für ihre jeweilige Wissenschaftsdisziplin zu generieren, womit zugleich (und zwar von Anfang an) von einer angeblichen Zweckfreiheit gar nicht gesprochen werden kann (vgl. Stock 2011; Helmchen 2011; Gerok 1996).
- Mir erscheint aber (bzw. deswegen) die zweite Betrachtungsweise wichtiger und gewichtiger zu sein. Sie verdeutlicht, was (mit Mittelstraß 1992, 48ff. und Wilke 1996, 14-18 gesprochen) als die unvorhersehbare Entwicklung wissenschaftlicher Forschungsstände bezeichnet werden könnte. In ihrem Sinne ist darauf zu verweisen, dass von keiner Studie vorab gesagt werden kann, ob sie nicht zu einem späteren Zeitpunkt in gänzlich anderen Kontexten zu vollkommen neuen Erkenntnissen führen wird. Ihr würde damit im Nachhinein ein ‚Zweck' zugeschrieben werden können, der bei der Entwicklung der Forschungsfrage und bei der Planung des Designs noch gar nicht präsent war – ja vielleicht noch gar nicht präsent sein konnte. Beispiele hierfür sind Legion, ein typisches sei hier erwähnt: So verweist etwa Gerhard Wegner (2000, 41) auf ein Ende des 20. Jahrhunderts entwickeltes astrophysisches Verfahren, mit dessen Hilfe ursprünglich eine mathematische Analyse des Signalflusses aus dem Weltall möglich gemacht werden sollte. Wie sich in nachfolgenden Studien aber erwies, ist eben dieses Verfahren ebenso dazu in der Lage, in der klinischen Diagnostik (und dadurch schließlich in der Heilung) von Herzkrankheiten eingesetzt zu werden. Somit wurde aus einem Instrument der beobachtenden Astrophysik ein Instrument der medi-

zinischen Heilung (Vergleichbare weitere Beispiele finden sich in zahlreicher Form auch bei Müller 2005; Zintzen 2000; Wilke 1996 oder Mittelstraß 1992).

So wird also deutlich, dass auch erkenntnisgenerierende (‚reine') Grundlagenforschung aus zweierlei Gründen nicht als zweckfrei gelten kann: Erstens lässt sich fachintern *per definitionem* nicht frei von Zwecken forschen, da innerhalb der eigenen Disziplin stets zum Zwecke des Erkenntnisfortschritts geforscht wird. In diesem Sinne konkretisiert auch Hans-Martin Gauger (2011, 80), dass die so häufig angemahnte Zweckfreiheit ‚reiner' Grundlagenforschung immer nur außerhalb der jeweils betreffenden Wissenschaft angesiedelt sein könne („Mit »zweckfrei« meine ich hier lediglich: frei von Zwecken *außerhalb* der Wissenschaften, von Zwecken also, die man von *außen* her von ihnen verlangt, an sie – zu Recht oder zu Unrecht – heranträgt."). Innerhalb der jeweiligen Wissenschaft lasse sich die Forderung nach Zweckfreiheit der Forschung demnach nicht aufrechterhalten.

Wer diesen Umstand wahr- und annimmt, muss dann zweitens einsehen, dass sich nicht vorhersehen (geschweige denn steuern) lässt, zu *welchen* Zwecken die gewonnenen Erkenntnisse im Nachhinein (und wenn auch erst Jahrzehnte später), von wem und welcher Institution auch immer, g*enutzt* werden. Diese zuletzt gewonnene Erkenntnis ist aus deutschdidaktischer Perspektive von immenser Wichtigkeit:

> Wenn nämlich Forschungen, die eingangs noch als ‚reine' Grundlagenforschung zitiert wurden, eben *nicht* als zweckfrei angesehen werden können, wie soll dann noch die Ansicht aufrechterhalten werden, der nutzungs- und anwendungsorientierten fachdidaktischen Forschung, die neben wissenschaftlicher Erkenntnis eben *auch* (und nicht zuletzt) den (zumindest potenziellen) Zweck verfolgt, auf lange Sicht zur Anwendung zu gelangen, den Status der Grundlagenforschung abzusprechen?

Es wird (im Gegenteil) vielmehr deutlich, dass *jegliche* Forschung,

- die generalisierende Aussagen die jeweilige Fachdisziplin betreffend ermöglicht (vgl. Brüggemann/Bromme 2006, 113)
- und dabei in konzeptioneller Hinsicht theoretisch fundiert (vgl. van den Daele 1975, 30),
- bei genuin human- und sozialwissenschaftlichen (d.h. hier bei bildungswissenschaftlichen resp. fachdidaktischen) Fragestellungen aber darüber hinaus im besten Fall auch empirisch abgesichert sein sollte (vgl. Bayrhuber et al. 2011; Reinders et al. 2010; Terhart 2002; Schwippert/Bos 2002; Fischer/Wecker 2006, 32f.),

als *Grundlagenforschung* zu bezeichnen sind – und zwar unabhängig davon, ob die Erkenntnisse tatsächlich zur praktischen Anwendung kommen und damit letztlich ‚nutzen', oder nicht.

Begrifflich lässt sich dies mithilfe des bereits zitierten Frascati-Manuals der OECD (2015) und seiner terminologischen Unterscheidung von *pure basic research* und *oriented basic research* fassen und anschließend im Sinne Jürgen Mittelstraß‘ (1992) dreifach differenzieren:

> Pure basic research is carried out for the advancement of knowledge, without seeking economic or social benefits or making an active effort to apply the results to practical problems or to transfer the results to sectors responsible for their application. (OECD 2015, 50)

> Oriented basic research is carried out with the expectation that it will produce a broad base of knowledge likely to form the basis of the solution to recognized or expected current or future problems or possibilities. (OECD 2015, 51)

Vor dem Hintergrund dieser OECD-Definition nimmt sich grundlegendes *fachdidaktisches* Forschen also zumeist als *oriented basic research* aus, da es gleichzeitig auf Wissenszuwachs und Anwendung zielt, d.h. (mit Frederking)

> die Generierung von neuem, grundlegendem Wissen im Zusammenhang mit fachspezifischem Lehren und Lernen im Bewusstsein und in der Erwartung, dass damit Lösungen für aktuelle oder zukünftige Fragen bzw. Probleme fachlicher Lehr-Lernprozesse, Akteure, Inhalte etc. verbunden sind oder sein können. (Frederking 2016, 199)

Diese Unterscheidung von *pure* vs. *oriented basic research* kann nun mithilfe des von Mittelstraß vorgeschlagenen „Forschungsdreiecks“ (2003) erweitert und erneut deutschdidaktisch spezifiziert werden. Darin differenziert Mittelstraß in „Grundlagenforschung“ (s.o., Kapitel 2.1), in „anwendungsorientierte Grundlagenforschung“ (s.o., Kapitel 2.2) und in „produktorientierte Anwendungsforschung“ (vgl. Mittelstraß, 1992, 60ff.). Im Sinne der Mittelstraßschen Differenzierung ist also fach- bzw. deutschdidaktische Forschung (wenn sie denn als Grundlagenforschung angelegt ist) das, was das Frascati-Manual *oriented basic research* nennt und was im vorliegenden Beitrag bislang als anwendungs- oder nutzungsorientierte Grundlagenforschung (*bzw.* unter Rückgriff auf Stokes 1997, 87 als „used-inspired basic research“) bezeichnet wurde.

- Wie bereits an obiger Stelle referiert, intendiert eine als „rein“ zu klassifizierende Grundlagenforschung nach Mittelstraß (analog zur *pure basic research* im Frascati-Report) keinerlei Anwendungsbezug (vgl. Mittelstraß 1992, 60) – mit all den von uns bislang vorgenommenen Relativierungen dieser Begrifflichkeit.
- Diese wird nach Mittelstraß in der „anwendungsorientierten Grundlagenforschung“ (also der *oriented basic research* im Frascati-Report bzw. der *used-inspired basic research* bei Stokes) zwar auch nicht zwingend und unmittelbar angestrebt. Im Gegensatz zur ‚reinen‘ Grundlagenforschung wird sie aber als Möglichkeit mitgedacht, und mehr noch: sie wird sogar mittel- bis langfristig „intendiert“ (Mittelstraß 1992, 63).

- Bleibt im Mittelstraßschen Forschungsdreieck nur mehr auf die von ihm postulierte „produktorientierte Anwendungsforschung“ zu verweisen, die in einem Handbuchartikel zum Thema Grundlagenforschung allerdings keine größere Bedeutung erlangen kann. Sie ist (um auch hier bei Mittelstraß zu bleiben) von vornherein ganz unmittelbar ökonomisch, gesellschaftlich oder politisch motiviert, strebt in diesem Sinne rein praktischen Nutzen an und ist daher eher dem Ansatz des *Design Research* zuzuordnen (siehe den Beitrag von Dube in diesem Band).

Mit den bisher vollzogenen Analysen aus wissenschaftstheoretischer und wissenschaftshistorischer Perspektive, mit der terminologischen Klärung des Begriffs Grundlagenforschung und der vollzogenen Revidierung seiner Exklusivität als angeblich ‚reine‘ (= zweckfreie) Forschung gegenüber anwendungs- oder nutzungsintendierten Forschungen, die nicht als Grundlagenforschung zu bezeichnen seien, sind wir schließlich an einem Punkt angelangt, der uns nun auch inhaltlich mit einer grundlagenerforschenden Deutschdidaktik beschäftigen lässt, *obwohl* diese als eine auf lange Sicht anwendungsbezogene Wissenschaft angesehen werden muss.

3. Grundlagenforschung und sprachlich-literarisches Lernen aus deutschdidaktischer Perspektive

Wie für alle Fachdidaktiken zu konstatieren ist, so ist auch deutschdidaktisches Forschen immer mehr auf inter- bzw. transdisziplinäre Designs bzw. auf personale und/oder institutionelle Kooperationen angewiesen (vgl. etwa Winkler/Schmidt 2016 oder Frederking 2014), weswegen die Deutschdidaktik zuletzt auch als „Kaleidoskop-Wissenschaft“ (Gailberger 2018b) bezeichnet wurde. Je nach fachdidaktischem Forschungsinteresse und Fragestellung entstehen inter- bzw. transdisziplinäre Designs ständig neu und werden personale oder institutionelle Kooperationen ständig (neu) gegründet, so wie sich auch in einem Kaleidoskop je nach Drehwinkel und Bewegung das sich zusammensetzende Erscheinungsbild stets aufs Neue wandelt, wohingegen die farbigen Steine stets die gleichen bleiben.

In den Fachdidaktiken zeigt sich diese Inter- bzw. Transdisziplinarität (z.B. im Sinne von Mittelstraß 1992, 90; ausführlich Mittelstraß 2003; zusammengefasst Mittelstraß 2012) darin, dass diese zuvorderst auf Erkenntnisse und Forschungsmethoden der eigenen Fachwissenschaft(en) zurückgreifen, dass sie darüber hinaus aber auch von der notwendigen breiten Berücksichtigung ihrer Bezugswissenschaften gekennzeichnet sind, zu denen, deutschdidaktisch gesprochen, beispielsweise die Psychologie, die Erziehungswissenschaften, die Soziologie, die Philosophie oder die Geschichtswissenschaft zählen (vgl. Frederking 2016, 193). In der Folge ist zu beobachten, dass damit ein ebenso breites Spektrum an theoretischen wie empirischen deutschdidaktischen Forschungsmethoden und Ansätzen einhergeht, so dass das Erscheinungsbild der deutschdidaktischen Grundlagenforschung von *historischen* Analysen (vgl. etwa Ivo 1994), *theoretischen*

Grundlegungen (vgl. etwa Kreft 1977) und schließlich *empirischen* Projekten (vgl. etwa Frederking 2013) geprägt wird.

Sie an dieser Stelle flächendeckend zu berücksichtigen, ist unmöglich. Es soll aber der Versuch unternommen werden, schlaglichtartig auf jene deutschdidaktischen Forschungen zu verweisen, die den Charakter und die Vorbedingungen grundlagenintendierter Forschungen aufweisen. Dabei finden nicht-empirische wie empirische Studien und Projekte Erwähnung, die im Rahmen der oben vorgenommenen begrifflichen Herleitungen vornehmlich als „anwendungsorientierte Grundlagenforschung" (Mittelstraß 1992, 62), als „oriented basic research" (OECD 2015, 51) bzw. als „used-inspired basic research" (Stokes 1997, 87) bezeichnet wurden – als Studien also, die die Verwendung ihrer Erkenntnisse nicht zwingend und unmittelbar anstrebten, im Gegensatz zur ‚reinen' Grundlagenforschung dies aber als Möglichkeit von vornherein mitdachten, wenn nicht gar mittel- bis langfristig intendierten.

3.1 Nicht-empirische Grundlagenforschung sprachlich-literarischen Lernens

Nicht-empirische deutschdidaktische Grundlagenforschung tritt vor allem in historisierenden sowie in systematisierenden Erscheinungsformen auf. Literatur*geschichtlich* ausgerichtete Forschung in der Deutschdidaktik erfolgt in einer Definition von Frederking und Abraham (2018, o.S.) „mit dem Ziel, die historischen Prägungen der Kultur der Gegenwart für unterrichtliche Lehr-Lern-Prozesse zugänglich zu machen […]." Mit ihnen können beispielhaft folgende Studien als genuin ‚historisch' bezeichnet werden: Hubert Ivos Untersuchung zum Zusammenhang von Muttersprache, Identität und Nation und die Reflexion ihres Zusammenhangs und Einflusses auf die sprachliche Bildung von Schülerinnen und Schülern (vgl. Ivo 1994); Kilians historische Dialogforschung, im Rahmen derer er den Konnex von Sprachgeschichte und Lehrgespräch in Schule und Unterricht historisch untersucht (vgl. Kilian 2002); die umfassenden Forschungen zur ideologischen Vereinnahmung der Kinder- und Jugendliteratur im Dritten Reich (z.B. von Hopster/Josting/Neuhaus 2005); als im weitesten Sinne historisch orientiert und interessiert ist darüber hinaus ebenso die in den letzten zwei Jahrzehnten wieder stärker in das Blickfeld getretene Kanonforschung zu nennen, wie sie z.B. von Dawidowski und Korte (2009) aus sozial- und bildungsgeschichtlicher Perspektive vorangetrieben wurde.

Greifen solche als vornehmlich historisch zu bezeichnenden Projekte notwendigerweise auf Ansätze und Methoden von Bezugswissenschaften wie der Geschichtswissenschaft, der Soziologie oder der historischen Bildungsforschung zurück, sind deutschdidaktische Forschungen mit einem eher *systematisierenden* Ansatz vielmehr durch eine besondere Nähe zur Sprach- bzw. Literaturwissenschaft gekennzeichnet. Als Beispiele mit genuin systematisierender Ausrichtung nennen Frederking und Abraham (2018, o.S.) beispielsweise Michael Kämpervan den Boogaarts Auseinandersetzung mit Begründungszusammenhängen literarischer Lektüre aus kultursoziologischer Sicht (1997), Thomas Zabkas Studie

zur Pragmatik der Literaturinterpretation (2005) wie auch Forschungen zum Metaphernverstehen (Jost 2007; Lessing/Wieser 2013). Innerhalb der sprachdidaktischen Forschung verweisen Frederking und Abraham exemplarisch auf Theorie, Erwerb und didaktisch-mediale Modellierung von Schreib- und Textroutinen (z.B. von Feilke/Lehnen 2012) sowie auf Untersuchungen zur Textsortenkompetenz und ihrer Genese bei Schülerinnen und Schülern (vgl. Augst et al. 2007).

3.2 Empirische Grundlagenforschung sprachlich-literarischen Lernens

Beim empirischen Arbeiten in der Deutschdidaktik muss zwischen quantitativen und qualitativen Designs unterschieden werden. Mit Beginn des 21. Jahrhunderts und der großen medialen Aufmerksamkeit, die der ersten *PISA*-Studie (und in ihrem Fahrwasser den *DESI*- und *IGLU*-Studien) zuteil wurde, ist hierbei zunächst die quantitativ ausgerichtete Kompetenzforschung zu nennen, die es im deutschsprachigen Raum (wenn auch mit einer Verzögerung von mehreren Jahrzehnten; vgl. Schwippert 2005, 2ff.) mittlerweile zu einigem Erfolg gebracht hat.

Für den Kompetenzbereich *Schreiben* wird dies in Publikationen wie dem 2017 erschienenen Handbuch über empirische Schreibdidaktik zur Erforschung schriftsprachlicher Textproduktionskompetenzen deutlich (vgl. Becker-Mrotzek/Grabowski/Steinhoff 2017), in das der bisher zusammengetragene Forschungsstand beispielsweise zu ausdifferenzierten Modellen zur Entwicklung und Beurteilung von Schreibkompetenzen (vgl. Becker-Mrotzek/Böttcher 2006) oder das reichhaltige Wissen über Prädiktoren von Schreibkompetenz einfließen konnte (vgl. Schmitt/Knopp 2017). Der Kompetenzbereich *Sprechen und Zuhören* ist ebenfalls bereits umfangreich empirisch erforscht (vgl. etwa Becker-Mrotzek 2012) und wird (wie auch der Kompetenzbereich *Schreiben*) durch die regelmäßigen Erhebungen der Schreib- und Zuhörkompetenzen auf unterschiedlichen Alters- und Jahrgangsstufen durch das IQB kontinuierlich weiter fundiert (vgl. etwa Zingg Stamm et al. 2016 oder Böhme et al. 2017). Im Kompetenzbereich *Sprache und Sprachgebrauch* kann mit Verweis auf Bremerich-Vos und Böhme (2009) oder auf Oomen-Welke und Bremerich-Vos (2014) ebenfalls auf eine solide grundlagenintendierte Forschung verwiesen werden.

Als weniger einheitlich muss der Stand zur empirischen Grundlegung des Kompetenzbereichs *Lesen – Mit Texten und Medien umgehen* gelten, wobei eine mehr als solide aufgestellte empirische Forschung mit *lese*didaktischer Perspektive einer empirischen *literatur*didaktischen Forschung gegenübersteht, die noch am Anfang zu stehen scheint. Wie bereits mehrfach angemerkt, ist die Lesekompetenz von Schülerinnen und Schülern in Deutschland mit quantitativen Designs vor allem durch Studien wie *PISA*, *IGLU* oder *DESI* gut erforscht. Jenseits rein quantitativ ausgerichteter Designs haben aber auch andere Projekte wie das DFG-Schwerpunktprogramm *Lesesozialisation in der Mediengesellschaft* von Norbert Groeben und Bettina Hurrelmann (2002-2006) wichtige empirische Grundlagen-

forschung in diesem Bereich geleistet, die (mit wenigen Ausnahmen wie bei Richter/Plath 2005) zumeist aber qualitativen Ansätzen folgten (vgl. hierzu beispielsweise Pieper et al. 2004; Rupp/Heyer/Bonholt 2004; Pette/Charlton 2006).

Im Gegensatz zur (stark von der Kognitionspsychologie und der Psycholinguistik beeinflussten) *Lese*didaktik (vgl. etwa Christmann/Groeben 1999; Richter/Christmann 2006; Christmann 2010), gibt es in der *Literatur*didaktik bislang nur vereinzelte Bemühungen, mit quantitativen Designs und Forschungsmethoden Grundlagenforschung zu betreiben. So ist die für Literaturunterricht wie Literaturdidaktik gleichermaßen basale disziplinäre Annahme weiterhin empirisch zu erhellen, dass es einen kognitionspsychologisch wie literaturtheoretisch modellierbaren Unterschied zwischen dem Lesen und Verstehen informatorisch-pragmatischer und literar-ästhetischer Texte gebe.

Nach einer tiefer gehenden Analyse diesbezüglich passender Items und der lesebegleitenden Verarbeitung dieser durch Probandinnen und Probanden im Rahmen der ersten *PISA*-Studie (vgl. Artelt/Schlagmüller 2004), die diese Annahme in der deutschsprachigen Forschung erstmals empirisch (wenn auch auf symptomatische Weise nicht didaktisch, sondern kognitionspsychologisch und psychometrisch intendiert) stützte, ist es vor allem den Forschungen des Teams um Volker Frederking zu verdanken, das hier genuin (literaturdidaktische) Grundlagenforschung geleistet und den diesbezüglichen (empirischen) Forschungsstand auf eine neue Ebene gehoben hat (vgl. etwa Frederking 2013; Frederking/Brüggemann/Hirsch 2016; Brüggemann et al. 2017). So ist es im Rahmen der seit 2009 bis heute ständig erweiterten *LUK*[6]-Studien gelungen, auf Basis systematischer Modellierungen und Erhebungen empirische Belege dafür zu gewinnen, dass literarisches Verstehen tatsächlich einen eigenen Kompetenzbereich darstellt, der sich deutlich von informatorischen, auf Sachtexte bezogenen Leseprozessen unterscheidet. Fünf Teildimensionen literarischer Verstehenskompetenz konnten bislang als empirisch bestätigt nachgewiesen werden: Semantische (1) und idiolektale (2) literarische Verstehenskompetenz, ästhetische Aufmerksamkeit (3), die Fähigkeit zur Anwendung literarischen Fachwissens (4) und die Fähigkeit zum Erfassen textseitig intendierter Emotionen (5) (vgl. im einzelnen Frederking/Brüggemann/Hirsch 2016; Meier et al. 2017; Brüggemann et al. 2017).

Als literaturdidaktische Grundlagenforschung, die mithilfe von *qualitativen* Designs das literarische Lesen, Verstehen und Verarbeiten von Schülerinnen und Schülern untersuchen, sind abschließend noch Studien zu nennen, die sich beispielsweise der Erforschung zum Metapher-Verstehen verschrieben haben (vgl. Pieper/Wieser 2011), die die Bedeutung von Lernaufgaben (Winkler 2010) oder des Vorwissens (Freudenberg 2012) im Literaturunterricht fokussieren, die den Begriff des literarischen Lernens aus inklusiver Perspektive reflektieren und fruchtbar für ebensolche Lehr-Lern-Prozesse machen (Wiprächtiger-Geppert

[6] Das Akronym *LUK* steht für Literar-Ästhetische Urteilskompetenz. Einen guten Überblick über die Genese des Projektes aus literaturtheoretischer, psychometrischer und schließlich literaturdidaktischer Perspektive liefert Frederking 2013.

2009), die mit kultursoziologischer Perspektive die Rolle von Lesen und Literatur zwischen Abitur und Studium (vgl. Dawidowski 2009) oder in Peer-Groups (Philipp 2010) untersuchen oder die Fähigkeit zum Perspektivverstehen und zum Umgang mit Fiktionalität mithilfe nachträglicher lauter Denkprotokolle auswerten (vgl. Stark 2012).

4. Fazit

Im vorliegenden Text wurde der strittigen Frage nachgegangen, ob und inwiefern in Fachdidaktiken wie der Deutschdidaktik Grundlagenforschung betrieben werden könne und wie diese schließlich aussehen müsste. Strittig war diese Frage aus zwei Gründen:

- Erstens, da ihr innerhalb der deutschdidaktischen Methodendiskussion bislang noch nicht zufriedenstellend nachgegangen worden war und sie somit naturgemäß auch nicht beantwortet werden konnte.
- Und zweitens, da (eben aus genuin deutschdidaktischer Perspektive) der allgemeinen wissenschaftstheoretischen Annahme entgegenzutreten war, dass eine auf lange Sicht anwendungsorientierte Wissenschaftsdisziplin (wie Fachdidaktiken auf lange Sicht eben immer anwendungsorientierte Wissenschaftsdisziplinen darstellen) keine Grundlagenforschung betreiben können, wenn diese definiert wird als eine ‚reine' Forschung, deren Erkenntnisse aus *zweckfrei* intendierten Forschungsprojekten entspringen und *nicht* zur Anwendung gelangen (vgl. etwa Mittelstraß 1992, 60 oder Müller 2005, 1).

Ausgehend von drei authentischen Beispielen aus dem Umfeld der grundlagenintendierten sprachlich-literarischen Kompetenzstudien *PISA*, *DESI* und *LUK* haben wir zu diesem Zwecke klären können, dass der Begriff der ‚reinen' Grundlagenforschung (in Anführungszeichen!) zu erweitern ist und vielmehr von Grundlagenforschung immer dann gesprochen werden sollte, wenn entsprechende (z.B. deutschdidaktische) Studien in konzeptioneller Hinsicht

- *theoretisch* fundiert sind (vgl. van den Daele 1975),
- *generalisierbare* Aussagen die jeweilige Fachdisziplin betreffend ermöglichen (vgl. Brüggemann/Bromme 2006) und darüber hinaus ggf. auch
- *empirisch* abgesichert argumentieren können (vgl. Frederking 2016) – und zwar unabhängig von der Frage, ob ihre Erkenntnisse später zur (praktischen) Anwendung gelangen oder nicht.

Diese Erkenntnis ermöglichte nicht zuletzt die wirkungsmächtige wissenschaftssozilogische Untersuchung von Donald E. Stokes (1997), in der die bis dahin *ein*dimensionale Dichotomie von Erkenntnisorientierung hier und Anwendungsorientierung dort zugunsten eines *zwei*dimensionalen Modells ersetzt wurde, was eine für fachdidaktische Forschungen gewinnbringende Differenzierung im Sinne einer *nutzungsorientierten Grundlagenforschung* ermöglicht (vgl. Stokes

1997, 87) – *ohne* dass diese Attribuierung der fachdidaktischen Forschung den Anspruch streitig machte, dennoch als Grundlagenforschung zu gelten.

In diesem Sinne ist fachdidaktische Grundlagenforschung im sogenannten Forschungsdreieck (nach Mittelstraß 2003) als „anwendungsorientierte Grundlagenforschung" zu verorten, die zwischen der „*reinen* Grundlagenforschung" und der „produktorientierten *Anwendung*sforschung" oszilliert, da sie die Anwendung ihrer Erkenntnisse zwar nicht explizit und zwingend angestrebt, auf lange Sicht aber auch nicht negiert (vgl. Mittelstraß, 1992, 60ff.).

Vor diesem Hintergrund wurden schließlich deutschdidaktische Forschungsprojekte schlaglichtartig vorgestellt, die auf der Grundlage des bis dahin Dargelegten als deutschdidaktische Grundlagenforschung gelten können, wobei zwischen *nicht*-empirischer Grundlagenforschung sprachlich-literarischen Lernens (vor allem in *historisierenden* sowie in *systematisierenden* Erscheinungsformen) auf der einen Seite und *empirischer* Grundlagenforschung sprachlich-literarischen Lernens (aller vier Kompetenzbereiche des Deutschunterrichts) auf der anderen Seite unterschieden wurde. Diese Schlaglichter zeigten, dass die deutschdidaktische Grundlagenforschung zwar bereits weit vorangeschritten ist, dass es aber ebenso noch eine Reihe von Desiderata zu schließen gilt, damit auch in Zukunft Grundlagenforschungen zum sprachlich-literarischen Lernen im weiteren Sinne wie *PISA*, *DESI* oder *LUK* dazu beitragen mögen, Schule und Deutschunterricht auf System-, Qualifikations- oder Unterrichtsebene zu verbessern.

Literatur

Abraham, Ulf/Frederking, Volker (2016): Deutsch und Deutschdidaktik. In: Vorstand der Gesellschaft für Fachdidaktik (Hrsg.): Auf dem Weg zu einer Allgemeinen Fachdidaktik. Münster: Waxmann, 53-73.

Artelt, Cordula/Baumert, Jürgen/Klieme, Eckard/Neubrand, Michael/Prenzel, Manfred/Schiefele, Ulrich/Schneider, Wolfgang/Stanat, Petra/Tillmann, Klaus-Jürgen/Weiß, Manfred (Hrsg.) (2001): Zusammenfassung zentraler Befunde: PISA 2000. Die Länder der Bundesrepublik Deutschland im Vergleich. Berlin: Max-Planck-Institut für Bildungsforschung.

Artelt, Cordula/Schlagmüller, Matthias (2004): Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In: Schiefele, Ulrich/Artelt, Cordula/Schneider, Wolfgang/Stanat, Petra (Hrsg.): Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000. Wiesbaden: VS Verlag für Sozialwissenschaften, 169-196.

Augst, Gerhard/Disselhoff, Katrin/Henrich, Alexandra/Pohl, Thorsten/Völzing, Paul-Ludwig (2007): Text – Sorten – Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter. Frankfurt a.M.: Peter Lang.

Bayrhuber, Horst/Harms, Ute/Muszynski, Bernhard/Ralle, Bernd/Rothgangel, Martin/Schön, Lutz-Helmut/Vollmer, Helmut J./Weigang, Hans-Georg (Hrsg.) (2011): Empirische Fundierung in den Fachdidaktiken. (= Fachdidaktische Forschungen Bd. 1) Münster: Waxmann.

Bayrhuber, Horst/Harms, Ute/Muszynski, Bernhard/Ralle, Bernd/Rothgangel, Martin/Schön, Lutz-Helmut/Vollmer, Helmut J./Weigang, Hans-Georg (Hrsg.) (2012): Empirische Fundierung in den Fachdidaktiken. (= Fachdidaktische Forschungen Bd. 2) Münster: Waxmann.

Baumert, Jürgen/Klieme, Eckhard/Neubrand, Michael/Prenzel, Manfred/Schiefele, Ulrich/Scheider, Wolfgang/Tillmann, Klaus-Jürgen/Weiß, Manfred (2001): PISA 2000. Zusammenfassung zentraler Befunde. Berlin: MPI für Bildungsforschung.

Becker-Mrotzek, Michael (2012): Mündliche Kommunikationskompetenz. In: Ders. (Hrsg.): Mündliche Kommunikation und Gesprächsdidaktik. Baltmannsweiler: Schneider Hohengehren.

Becker-Mrotzek, Michael/Böttcher, Ingrid (2006): Schreibkompetenzen entwickeln und beurteilen. Praxishandbuch für die Sekundarstufe I und II. Berlin: Cornelsen.

Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hrsg.) (2017): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann.

Behrens, Ulrike (2014): Zuhörfähigkeit testen. In: Neumann, Astrid/Mahler, Isabelle (Hrsg.): Empirische Methoden in der Deutschdidaktik: Audio- und videografierte Unterrichtsforschung. Baltmannsweiler: Schneider Hohengehren, 33-47.

Bertschi-Kaufmann, Andrea/Hagendorf, Petra/Kruse, Gerd/Rank, Katharina/Riss, Maria/Sommer, Thomas (2007): Lesen. Das Training. Ausgaben für das 4.-6. und das 7.-9. Schuljahr. Seelze-Velber: Lernbuch Verlag bei Friedrich.

Böhme, Katrin/Schipolewski, Stefan/Canz, Thomas/Krelle, Michael/Bremerich-Vos, Albert (2017): Kompetenzstufenmodelle im Bereich Schreiben. In: Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hrsg.): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann, 55-74.

Bredel, Ursula/Reißig, Tilo (Hrsg.) (2011): Weiterführender Orthographieunterricht. Baltmannsweiler: Schneider Hohengehren.

Bremerich-Vos, Albert/Böhme, Katrin (2009): Kompetenzdiagnostik im Bereich „Sprache und Sprachgebrauch untersuchen“. In: Granzer, Dietlinde/Köller, Olaf/Bremerich-Vos, Albert (Hrsg.): Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule. Weinheim: Beltz, 376-392.

Brüggemann, Anne/Bromme Rainer (2006): Anwendungsorientierte Grundlagenforschung in der Psychologie: Sicherung von Qualität und Chancen in den Beurteilungs- und Entscheidungsprozessen der DFG. In: Psychologische Rundschau, 57, 2, 112-118.

Brüggemann, Jörn/Frederking, Volker (2017): Die Bedeutung literarisch evozierter Emotionen für die Aktivierung von Empathie und literarischer Textverstehenskompetenz im Fokus einer literaturdidaktischen Bildungsforschung. In: Rat für Kulturelle Bildung (Hrsg.): Von Mythen zu Erkenntnissen. Beiträge der 7. Netzwerktagung des Netzwerks Forschung Kulturelle Bildung. München: Kopaed.

Brüggemann, Jörn/Frederking, Volker/Henschel, Sofie/Gölitz, Dietmar (2016): Emotionale Aspekte literarischer Textverstehenskompetenz: Theoretische Annahmen und empirische Befunde. In: Mitteilungen des Deutschen Germanistenverbandes, 63, 2, 105-118

Bush, Vannevar (1945): Science: The Endless Frontier. A Report to the President on a Program for Postwar Scientific Research, http://www.nsf.gov/od/lpa/nsf50/vbush1945.htm (letzter Zugriff: 01.08.2018).

Carrier, Martin (2011): Verstehen und Können: Zum Verhältnis von Grundlagen- und Anwendungsforschung. In: Gegenworte – Hefte für den Disput über Wissen, 26, 11-13.

Christmann, Ursula (2010): Lesepsychologie. In: Kämper-van den Boogaart, Michael/Spinner, Kasper H. (Hrsg.): Lese- und Literaturunterricht. Band 1. Baltmannsweiler: Schneider Hohengehren, 148-200.

Christmann, Ursula/Groeben, Norbert (1999): Psychologie des Lesens. In: Franzmann, Bodo/Jäger, Georg (Hrsg.): Handbuch Lesen. München: Saur, 145-223.

Dawidowski, Christian (2009): Literarische Bildung in der heutigen Mediengesellschaft. Eine empirische Studie zur kultursoziologischen Leseforschung (= Siegener Schriften zur Kanonforschung Bd. 7). Frankfurt a.M.: Peter Lang.

Dawidowski, Christian/Korte, Hermann (Hrsg.) (2009): Umbrüche, Literaturkanon und Literaturunterricht in Zeiten der Modernisierung: Die 1920er und die 1960er Jahre. Vorträge des 3. Siegener Symposions zur literaturdidaktischen Forschung (=Siegener Schriften zur Kanonforschung Bd. 6). Frankfurt a.M.: Peter Lang.

Duden online Wörterbuch: „Zweck“ https://www.duden.de/rechtschreibung/Zweck (letzter Zugriff: 01.08.2018).

Fay, Johanna (2013): Orthographie in der Primarstufe. In: Gailberger, Steffen/Wietzke, Frauke (Hrsg.): Handbuch Kompetenzorientierter Deutschunterricht. Beltz: Weinheim, 172-194.

Feilke, Helmuth/Lehnen, Katrin (Hrsg.) (2012): Schreib- und Textroutinen. Theorie, Erwerb und didaktisch-mediale Modellierung (= Forum Angewandte Linguistik Bd. 52). Frankfurt a.M.: Peter Lang.

Feilke, Helmuth/Pohl, Thorsten (Hrsg.) (2014): Schriftlicher Sprachgebrauch – Texte verfassen. Baltmannsweiler: Schneider Hohengehren.

Fischer, Frank/Wecker, Christof (2006): Pasteurs Quadrant und die Diskussion in den USA um die Verbesserung des praktischen Nutzens der Bildungsforschung. In: Brüggemann, Anne/Bromme, Rainer (Hrsg.): Entwicklung und Bewertung von anwendungsorientierter Grundlagenforschung in der Psychologie – Rundgespräche und Kolloquien der Deutschen Forschungsgemeinschaft. Berlin: Akademie Verlag, 27-37.

Frederking, Volker (2013): Literarische Verstehenskompetenz erfassen und fördern. In: Gailberger, Steffen/Wietzke, Frauke (Hrsg.): Handbuch Kompetenzorientierter Deutschunterricht. Weinheim/Basel: Beltz Juventa, 117-144.

Frederking, Volker (2014): Deutschdidaktik als transdisziplinäre, anwendungs- und grundlagenorientierte empirische Wissenschaft. In: Mitteilungen des Deutschen Germanistenverbandes, 61, 2, 109-119.

Frederking, Volker (2016): Allgemeine Fachdidaktik – Metatheorie und Metawissenschaft der Fachdidaktiken. Begründungen und Konsequenzen. In: Vorstand der Gesellschaft für Fachdidaktik (Hrsg.): Auf dem Weg zu einer Allgemeinen Fachdidaktik. Münster: Waxmann, 179-204.

Frederking, Volker/Abraham, Ulf (o.J.): Deutschunterricht und Deutschdidaktik. Manuskript.

Frederking, Volker/Brüggemann, Jörn/Hirsch, Matthias (2016): Fünf Dimensionen von Literary Literacy und ihre interdisziplinären Implikationen am Beispiel der Geschichtsdidaktik. In: Lehmann, Katja/Werner, Michael/Zabold, Stefanie (Hrsg.): Historisches Denken jetzt und in Zukunft. Münster: LIT, 211-234.

Freudenberg, Ricarda (2012): Zur Rolle des Vorwissens beim Verstehen literarischer Texte. Eine qualitativ-empirische Untersuchung. Wiesbaden: Springer.

Frey, Hanno (2010): Lesekompetenz verbessern? Lesestrategien und Bewusstmachungsverfahren nutzen! Münster: Waxmann.

Frickel, Daniela/Kammler, Clemens/Rupp, Gerhard (Hrsg.) (2012): Literaturdidaktik im Zeichen von Kompetenzorientierung und Empirie. Perspektiven und Probleme. Freiburg i.B.: Fillibach.

Gailberger, Steffen (2018a): *YouTube*, *Audible* und *Co*. Literarisches Lernen in einem digitalisierten Deutschunterricht aller Kompetenzbereiche. Unterrichtsanregungen zum Zusammenhang von literarischem Hören und literarischem Lesen für die Jahrgänge 10 bis 12/13. In: Gailberger, Steffen/Wietzke, Frauke (Hrsg.): Deutschunterricht in einer digitalisierten Gesellschaft. Unterrichtsanregungen für die Sekundarstufen. Weinheim/Basel: Beltz, 194-214.

Gailberger, Steffen (2018b): Deutschdidaktik als Kaleidoskop-Wissenschaft – Das Fach Deutsch als kaleidoskopädischer Unterricht. In: Grundler, Elke (Hrsg.): Wirksamer Deutschunterricht. Baltmannsweiler: Schneider Hohengehren, 87-98.

Gauger, Hans-Martin (2011): Dürfen zwecklose Wissenschaften sein? In: Gegenworte – Hefte für den Disput über Wissen, 26, 80-83.

Gätje, Olaf/Krelle, Michael/Behrens, Ulrike/Grundler, Elke (2016): Präsentieren als literale Kompetenz? In: leseforum.ch 1/2016. https://www.leseforum.ch/myUploadData/files/2016_1_Gaetje_et_al.pdf (letzter Zugriff: 01.08.2018).

Gerok, Wolfgang (1996): Wechselwirkungen von Grundlagenforschung und angewandter Forschung: Ergebnisse und Perspektiven. In: Konferenz der deutschen Akademien der Wissenschaften; Nordrhein-Westfälische Akademie der Wissenschaften (Hrsg.): Entdeckung, Erkenntnis, Fortschritt – Wechselwirkungen von Grundlagenforschung und angewandter Forschung. Mainz: Philipp von Zabern, 135-136.

Gold, Andreas (2007): Lesen kann man lernen. Lesestrategien für das 5. und 6. Schuljahr. Göttingen: Vandenhoeck und Ruprecht.

Groeben, Norbert/Hurrelmann, Bettina (Hrsg.) (2002): Lesekompetenz. Bedingungen, Dimensionen, Funktionen. Weinheim: Juventa.

Groeben, Norbert/Hurrelmann, Bettina (Hrsg.) (2004): Lesesozialisation in der Mediengesellschaft. Ein Forschungsüberblick. Weinheim: Juventa.

Hellenkemper, Hansgerd (1996): Ende einer Verpflichtung? Angewandte Forschung in den Geisteswissenschaften. In: Konferenz der deutschen Akademien der Wissenschaften; Nordrhein-Westfälische Akademie der Wissenschaften (Hrsg.): Entdeckung, Erkenntnis, Fortschritt – Wechselwirkungen von Grundlagenforschung und angewandter Forschung. Mainz: Philipp von Zabern, 101-107.

Helmchen, Manfred (2011): Grundlagenforschung in der Psychiatrie. In: Gegenworte – Hefte für den Disput über Wissen, 26, 25-28.

Helmers, Hermann (1966): Didaktik der deutschen Sprache. Einführung in die Theorie der muttersprachlichen und literarischen Bildung. Stuttgart: Klett.

Hiecke, Robert Heinrich (1841): Der deutsche Unterricht auf deutschen Gymnasien. Leipzig: Eduard Eisenach.

Hopster, Norbert/Josting, Petra/Neuhaus, Joachim (2005): Kinder- und Jugendliteratur 1933-1945: Ein Handbuch. Stuttgart: Metzler.

Ivo, Hubert (1994): Muttersprache, Identität, Nation. Sprachliche Bildung im Spannungsfeld zwischen einheimisch und fremd. Wiesbaden: VS Verlag für Sozialwissenschaften.

Jost, Jörg (2007): Topos und Metapher. Zur Pragmatik und Rhetorik des Verständlichmachens. Heidelberg: Winter.

Kambartel, Friedrich (2008): Grundlagenforschung. In: Mittelstraß, Jürgen (Hrsg.): Enzyklopädie Philosophie und Wissenschaftstheorie. Band 3, G-Inn. Stuttgart: Metzler, 233-234.

Kammler, Clemens/Knapp, Werner (2002): Empirische Unterrichtsforschung und Deutschdidaktik. Baltmannsweiler: Schneider Hohengehren.

Kammler, Clemens (2010): Literaturtheorie und Literaturdidaktik. In: Kämper-van den Boogaart, Michael/Spinner, Kaspar H. (Hrsg.): Lese- und Literaturunterricht. Band 1. Baltmannsweiler: Schneider Hohengehren, 201-237.

Kämper-van den Boogaart, Michael (1997): Schönes schweres Lesen. Legitimität literarischer Lektüre aus kultursoziologischer Sicht. Wiesbaden: Dt. Univ.-Verl.

Kämper-van den Boogaart, Michael (2010): Geschichte des Lese- und Literaturunterrichts. In: Kämper-van den Boogaart, Michael/Spinner, Kaspar H. (Hrsg.): Lese- und Literaturunterricht. Band 1. Baltmannsweiler: Schneider Hohengehren, 3-83.

Kämper-van den Boogaart, Michael/Spinner, Kaspar H. (Hrsg.) (2010): Lese- und Literaturunterricht. 3 Teilbände. Baltmannsweiler: Schneider Hohengehren.

Kepser, Matthis (2013): Deutschdidaktik als eingreifende Kulturwissenschaft. Ein Positionierungsversuch im wissenschaftlichen Feld. In: Didaktik Deutsch, 18, 34, 52-68.

Kepser, Matthis/Abraham, Ulf (2016): Literaturdidaktik Deutsch. Eine Einführung. 4., aktual. u. erw. Aufl. Berlin: Erich Schmidt.

Kilian, Jörg (2002): Lehrgespräch und Sprachgespräche. Untersuchungen zur historischen Dialogforschung. In: Zeitschrift für Dialektologie und Linguistik, 71, 353-355.

Koch-Priewe, Barbara (2004): Professionsforschung und Didaktik der LehrerInnenbildung. In: Koch-Priewe, Barbara/Kolbe, Fritz-Ulrich, Wildt, Johannes (Hrsg.): Grundlagenforschung und mikrodidaktische Reformansätze zur Lehrerbildung. Bad Heilbrunn/Obb.: Julius Klinkhardt, 7-21.

Köster, Juliana (2016): Die dilemmatische Disziplin – Deutschdidaktik zwischen Eklektizismus und Partialisierung. In: Bräuer, Christoph (Hrsg.): Denkrahmen der Deutschdidaktik: Die Identität der Disziplin in der Diskussion. Frankfurt a.M.: Peter Lang, 59-78.

Kreft, Jürgen (1977): Grundprobleme der Literaturdidaktik. Eine Fachdidaktik im Konzept sozialer und individueller Entwicklung und Geschichte. Heidelberg: Quelle & Meyer.

Krelle, Michael (2013): Gesprächskompetenz in der Grundschule und der Sekundarstufe I – Konzepte und didaktische Erläuterungen. In: Gailberger, Steffen/Wietzke, Frauke

(Hrsg.): Handbuch kompetenzorientierter Deutschunterricht: Diagnostizieren – Binnendifferenzieren – Fördern. Weinheim: Beltz, 422-438.

Leibniz, Gottfried Wilhelm (seit 1923): Sämtliche Schriften und Briefe. Hrsg. von der Berlin-Brandenburgischen Akademie der Wissenschaften und der Akademie der Wissenschaften zu Göttingen. Berlin/Boston: de Gruyter.

Lessing, Marie/Wieser, Dorothee (Hrsg.) (2013): Zugänge zu Metaphern – Übergänge durch Metaphern. Paderborn: Wilhelm Fink.

Mittelstraß, Jürgen (1992): Leonardo-Welt. Über Wissenschaft, Forschung und Verantwortung. Frankfurt a.M.: Suhrkamp.

Mittelstraß, Jürgen (2003): Transdisziplinarität – wissenschaftliche Zukunft und institutionelle Wirklichkeit. Konstanz: Universitätsverlag.

Mittelstraß, Jürgen (Hrsg.) (2008): Enzyklopädie Philosophie und Wissenschaftstheorie. Stuttgart: Metzler

Mittelstraß, Jürgen (2012): Zwischen den Wissenschaften. Über Inter-, Multi- und Transdisziplinarität. In: Gegenworte – Hefte für den Disput über Wissen, 28, 10-13.

Müller, Andreas (2005): Brauchen wir Grundlagenforschung? https://www.spektrum.de/astrowissen/grundlagen.html (letzter Zugriff: 01.08.2018).

Neumann, Astrid/Mahler, Isabell (Hrsg.) (2014): Empirische Methoden der Deutschdidaktik: Audio- und videografierende Unterrichtsforschung. Baltmannsweiler: Schneider Hohengehren.

Nickel-Bacon, Irmgard (2006): Positionen der Literaturdidaktik – Methoden des Literaturunterrichts. Ein heuristischer Explikationsversuch für die empirische Grundlagenforschung. In: Groeben, Norbert/Hurrelmann, Bettina (Hrsg.): Empirische Unterrichtsforschung: Literatur- und Lesedidaktik. Weinheim: Juventa, 95-114.

Nold, Günter/Willenberg, Heiner (2007): Lesefähigkeit [Englisch und Deutsch]. In: Beck, Bärbel/Klieme, Eckhard (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messungen. Die DESI-Studie. Weinheim: Beltz, 23-41.

OECD (2015): Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development. The Measurement of Scientific, Technological and Innovation Activities, http://www. oecd-ilibrary.org/science-and-technology/frascati-manual-2015_9789264239012-en;jsessionid=4e5iip7a6trt2.x-oecd-live-03 (letzter Zugriff: 01.08.2018).

Oomen-Welke, Ingelore/Bremerich-Vos, Albert (2014): Sprache und Sprachgebrauch untersuchen. In: Behrens, Ulrike/Bremerich-Vos, Albert/Krelle, Michael/Böhme, Katrin/Hunger, Susanne (Hrsg.): Bildungsstandards Deutsch: konkret – Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen. Berlin: Cornelsen, 215-246.

Pette, Corinna/Charlton, Michael (2006): Empirisches Beispiel: Differenzielle Strategien des Romanlesens: Formen, Funktionen und Entstehungsbedingungen. In: Groeben, Norbert/Hurrelmann, Bettina (Hrsg.): Lesesozialisation in der Mediengesellschaft. Ein Forschungsüberblick. Weinheim: Juventa, 195-213.

Pflugmacher, Torsten (2016): Abstand durch Nahe – Nahe durch Abstand. Deutschdidaktik als reflexive Wissenschaft. In: Bräuer, Christoph (Hrsg.): Denkrahmen der Deutschdidaktik: Die Identität der Disziplin in der Diskussion. Frankfurt a.M.: Peter Lang, 79-94.

Philipp, Maik (2010): Eine Längsschnittstudie zur Bedeutung von *peer groups* für Lesemotivation und -verhalten. Wiesbaden: VS Verlag für Sozialwissenschaften.

Philipp, Maik (2015): Lesestrategien. Bedeutung, Formen und Vermittlung. Weinheim: Beltz/Juventa.

Pieper, Irene/Rosebrock, Cornelia/Wirthwein, Heike/Volz, Steffen (Hrsg.) (2004): Lesesozialisation in schriftfernen Lebenswelten: Lektüre und Mediengebrauch von HauptschülerInnen. Weinheim: Juventa.

Pieper, Irene/Wieser, Dorothee (2011): Forschungsüberblick: Empirische Studien zum Verstehen von Metaphern in literarischen Texten. In: Didaktik Deutsch, 17, 30, 74-95.

Pohl, Inge/Ulrich, Winfried (Hrsg.) (2011): Wortschatzarbeit. Baltmannsweiler: Schneider Hohengehren.

Reinders, Heinz/Ditton, Hartmut/Gräsel, Cornelia/Gniewosz, Burkhard (Hrsg.) (2010): Empirische Bildungsforschung. Strukturen und Methoden. Wiesbaden: Beltz/UTB.

Richter, Tobias/Christmann, Ursula (2006): Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In: Groeben, Norbert/Hurrelmann, Bettina (Hrsg.): Lesekompetenz: Bedingungen, Dimensionen, Funktionen. Weinheim: Juventa, 25-58.

Richter, Karin/Plath, Monika (2005): Lesemotivation in der Grundschule. Empirische Befunde und Modelle für den Unterricht. Weinheim: Juventa.

Rosebrock, Cornelia/Nix, Daniel (Hrsg.) (2017): Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung. 8., überarb. und erw. Aufl. Baltmannsweiler: Schneider Hohengehren.

Rupp, Gerhard/Heyer, Petra/Bonholt, Helge (2004): Folgefunktionen des Lesens – Von der Fantasie-Entwicklung zum Verständnis des sozialen Wandels. In: Groeben, Norbert/Hurrelmann, Bettina (Hrsg.): Lesesozialisation in der Mediengesellschaft. Ein Forschungsüberblick. Weinheim: Juventa, 95-141.

Schmitt, Markus/Knopp, Matthias (2017): Prädikatoren der Schreibkompetenz. In: Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hrsg.): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann, 239-252.

Schwippert, Kurt (2005): Vergleichende Lernstandsuntersuchungen, Bildungsstandards und die Steuerung von schulischen Bildungsprozessen. In: Berufs- und Wirtschaftspädagogik, 8, 1-14.

Schwippert, Kurt/Bos, Wilfried (2002): TIMSS, PISA, IGLU & Co. Vom Sinn und Unsinn internationaler Schulleistungsuntersuchungen. Bildung und Erziehung, 55, 1, 5-23.

Stanat, Petra/Artelt, Cordula/Baumert, Jürgen/Klieme, Eckhard/Neubrand, Michael/Prenzel, Manfred/Schiefele, Ulrich/Schneider, Wolfgang/Schümer, Gundel/Tillmann, Klaus-Jürgen/Weiß, Manfred (2001): PISA und PISA-E: Zusammenfassung der bereits vorliegenden Befunde. In: Baumert, Jürgen/Klieme, Eckard/Neubrand, Michael/Prenzel, Manfred/Schiefele, Ulrich/Schneider, Wolfgang/Stanat, Petra/Tillmann, Klaus-Jürgen/Weiß, Manfred (Hrsg.): PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske+Budrich.

Stanat, Petra/Artelt, Cordula/Baumert, Jürgen/Klieme, Eckard/Neubrand, Michael/Prenzel, Manfred (2002): PISA 2000: Die Studie im Überblick. Grundlagen, Methoden und Ergebnisse. Berlin: Max-Planck-Institut für Bildungsforschung.

Stanat, Petra/Böhme, Katrin/Schipolowski, Stefan/Haag, Nicole (Hrsg.) (2016): IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich.

Stark, Tobias (2012): Zum Perspektivenverstehen und zum Umgang mit Fiktionalität beim literarischen Verstehen. Ausgewählte Ergebnisse einer qualitativen empirischen Untersuchung. In: Pieper, Irene/Wieser, Dorothee (Hrsg.): Fachliches Wissen und literarisches Verstehen. Studien zu einer brisanten Relation. Frankfurt a.M.: Peter Lang, 153-169.

Stock, Günter (2011): »Zweckfreie Forschung« – Eine im 21. Jahrhundert taugliche Begrifflichkeit? In: Gegenworte – Hefte für den Disput über Wissen, 26, 8-9.

Stokes, Donald E. (1997): Pasteur's Quadrant. Basic Science and Technological Innovation. Washington, DC: Brookings Institution Press.

Terhart, Ewald (2002): Wie können die Ergebnisse von vergleichenden Leistungsstudien systematisch zur Qualitätsverbesserung in Schulen genutzt werden? In: Zeitschrift für Pädagogik, 48, 1, 91-110.

Thomé, Günter (2003): Entwicklung der basalen Rechtschreibkenntnisse. In: Bredel, Ursula/Günther, Hartmut/Klotz, Peter/Ossner, Jakob/Siebert-Ott, Gesa (Hrsg.): Didaktik der deutschen Sprache, Bd. 1. Paderborn: Schöningh, 369-379.

Ulshöfer, Robert (1952-57): Methodik des Deutschunterrichts. 3 Bände. Unterstufe. Stuttgart: Klett.

van den Daele, Wolfgang (1975): Autonomie contra Planung: Scheingefecht um die Grundlagenforschung? In: Wirtschaft und Wissenschaft, 23, 2, 29-32.

Wackernagel, Philipp (1842): Der Unterricht in der Muttersprache, 4. Teil des Deutschen Lesebuchs. Stuttgart: Liesching.

Walter, Jürgen (2001): Förderung bei Lese- und Rechtschreibschwäche: Grundlagenforschung, methodische Konsequenzen, Praxisbeispiele und mediendidaktische Anregungen auf der Basis empirischer Forschungsmethoden. Göttingen: Hogrefe.

Wegner, Gerhard (2000): Grundlagenforschung im deutschen Forschungssystem. In: Zintzen, Clemens (Hrsg.): Die Zukunft der Grundlagenforschung. Mainz: Philipp von Zabern, 33-42.

Wilke, Günther (1996): Vom Reagenzglas zur Produktion. Drei Beispiele zum Thema. In: Konferenz der deutschen Akademien der Wissenschaften; Nordrhein-Westfälische Akademie der Wissenschaften (Hrsg.): Entdeckung, Erkenntnis, Fortschritt – Wechselwirkungen von Grundlagenforschung und angewandter Forschung. Mainz: Philipp von Zabern, 13-18.

Willenberg, Heiner (2007a): Lesestufen – Die Leseprozesstheorie. In: Ders. (Hrsg.): Kompetenzhandbuch für den Deutschunterricht. Baltmannsweiler: Schneider Hohengehren, 11-23.

Willenberg, Heiner (2007b): Lesen. In: Beck, Bärbel/Klieme, Eckhard (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messungen. Die DESI-Studie. Weinheim: Beltz, 103-113.

Winkler, Iris (2010): Lernaufgaben im Literaturunterricht. In: Kiper, Hanna/Meints, Waltraud/Peters, Sebastian/Schlump, Stefanie/Schmit, Stefan (Hrsg.): Lernaufgaben und Lernmaterialien im kompetenzorientierten Unterricht. Stuttgart: Kohlhammer, 103-113.

Winkler, Iris/Schmidt, Frederike (Hrsg.) (2016): Interdisziplinäre Forschung in der Deutschdidaktik. „Fremde Schwestern“ im Dialog. Frankfurt a.M.: Peter Lang.

Winnacker, Ernst-Ludwig (2000): Grundlagenforschung und Forschungsförderung – Eine ständige Herausforderung. In: Zintzen, Clemens (Hrsg.): Die Zukunft der Grundlagenforschung. Mainz: Philipp von Zabern, 17-31.

Wintersteiner, Werner (2007): Die Innenwelt der Außenwelt der Innenwelt. Deutschdidaktik im Sog gesellschaftlicher Interessen. In: Didaktik Deutsch, 22, 51-70.

Wiprächtiger-Geppert, Maja (2009): Literarisches Lernen in der Förderschule. Baltmannsweiler: Schneider Hohengehren.

Zabka, Thomas (2005): Pragmatik der Literaturinterpretation. Theoretische Grundlagen – kritische Analysen. Tübingen: Niemeyer.

Zingg Stamm, Claudi/Käser-Leisibach, Ursula/Behrens, Ulrike/Krelle, Michael/Weirich, Sebastian (2016): Neue Aufgabenformate für die Messung von Zuhörkompetenzen. In: Keller, Stefan/Reintjes, Christian (Hrsg.): Aufgaben als Schlüssel zur Kompetenz. Münster: Waxmann, 129-140

Zintzen, Clemens (2000): Einleitung. In: Ders. (Hrsg.): Die Zukunft der Grundlagenforschung – Vorträge des Symposiums vom 15. Juli 1999. Stuttgart: Franz Steiner, 7-15.

JULIANE DUBE

Design Research

Anwendungsorientierte Grundlagenforschung

1. Problemstellung

Seit der Veröffentlichung internationaler und nationaler Schulleistungsstudien richtet sich der Blick didaktischer Forschung auf die Gestaltung von Unterricht. Zur Beseitigung beobachteter gegenstandsspezifischer Lernschwierigkeiten fordert die Forschungsebene auf der einen Seite eine intensivere Rezeption und Adaption neuer wissenschaftlicher Erkenntnisse durch die Praxis (vgl. u.a. Einsiedler 2010; Gräsel 2010; Tenorth 2012), während Praktikerinnen und Praktiker auf der anderen Seite die Wissenschaft in der Pflicht sehen, ihre Forschungsbemühungen stärker an den Belangen des Unterrichts auszurichten und konkrete Lernumgebungen zu entwickeln (vgl. Wilhelm/Hopf 2014, 32).

Angesichts der wachsenden Herausforderungen u.a. bei der Gestaltung eines inklusiven Deutschunterrichts gewinnen anwendungsorientierte Forschungszugänge, in denen vor allem Antworten auf die Fragen des *wie* und *warum* generiert werden sollen, wieder an Bedeutung (vgl. Prengel et al. 2008, 193). Den Ausführungen von Hargreaves und Fink folgend: „Change in education is easy to propose, hard to implement, and extraordinarily difficult to sustain“ (ebd. 2006, 1), bedarf es dafür jedoch der systematischen Verzahnung genuiner Grundlagenstudien und anwendungsbezogener Entwicklungsarbeiten.

Im vorliegenden Beitrag soll das Format des Design Research als ein Ansatz für eine anwendungsorientierte Grundlagenforschung vorgestellt werden, das bereits in seinem Termini Entwicklung und Forschung als gleichberechtigt anerkennt (vgl. Prediger/Gravemeijer/Confrey 2016, 878).[1] Ziel des jungen Forschungszugriffes, welches national in den letzten Jahren vor allem als Fachdidaktische Entwicklungsforschung bekannt wurde (vgl. Prediger/Link 2012; Hußmann et al. 2013; Dube/Prediger 2017; Dube/Hußmann 2018, i.Dr.), ist es, den konträr zueinander liegenden Erwartungen von Theorie und Praxis gerecht zu werden, indem die konkurrierende Dualität von empirischer Beforschung und theoretisch gestützter Entwicklung durch eine wechselseitige Diskursivität abgelöst wird (vgl.

[1] Vergleiche zur Begriffsvielfalt einer anwendungsorientierten Grundlagenforschung die Ausführungen in Kapitel 2.

Dube/Hußmann 2018, i.Dr.). Damit bietet sie z.B. auch der Literaturdidaktik einen Forschungsrahmen, mit dem das gestiegene Interesse an Fragen zu Voraussetzungen, Wirkungen und Funktionen beim Lesen literarischer Texte (vgl. Olsen 2011, o.S.) befriedigt werden kann.

2. Design Research

Zentrales Ziel von Design Research ist dementsprechend nicht die *koexistenzielle Anerkennung*, sondern die systematische *Verknüpfung* von Grundlagenforschung (siehe den Beitrag von Gailberger in diesem Band) und Praxisforschung unter der Zielstellung, neue theoretische Erkenntnisse zu Lehr- und Lernprozessen auf der einen Seite und empiriebasierten Unterrichtsmaterialien auf der anderen Seite (weiter) zu entwickeln (vgl. Dube/Prediger 2017, 3). Im Fokus der beschriebenen Form von Lehr-Lern-Forschung ist demnach nicht in erster Linie die Haltung des „Nachweisens, dass", sondern die des „Explorierens und Prüfens, was" (Euler 2012, 37).

International und national hat jene bipolare Zielstellung von Unterrichtsforschung in den letzten Jahren breiten Zuspruch erfahren. Wie die folgende Zusammenstellung zeigt (vgl. Dube/Hußmann 2018, i.Dr.), entwickelte sich in Konsequenz jedoch weniger ein festes Paradigma als vielmehr „a series of approaches" (Barab/Squire 2004, 2):

- Instructional design (Reigeluth 1983)
- Design Experiments (Brown 1992)
- Design Science (Wittmann 1995)
- Development(al) Research (van den Akker 1999)
- Curricular Design Research (van den Akker 2003)
- Design-Based Research (Design-Based Research Collective 2003)
- Educational Design Research (van den Akker et al. 2006)
- Design-Based Implementation Research (Penuel/Martin 2015)

Auch auf nationaler Ebene sind in den letzten Jahren unterschiedliche Ansätze einer anwendungsorientierten Grundlagenforschung entstanden:

Im Bereich der Bildungsforschung:

- Didaktische Entwicklungsforschung (Kahlert 2007; Einsiedler 2010)
- Entwicklungsorientierte Bildungsforschung (Reinmann/Sesink 2011)

Im Bereich der Fachdidaktik:

- Didaktik als Design Science (Fischer et al. 2005)
- Gestaltungsbasierte Forschung bzw. Gestaltungsforschung (Euler 2011)
- Gegenstandsorientierte Fachdidaktische Entwicklungsforschung (u.a. Hußmann et al. 2012, 2013)
- Bremer Design-Research Modell (Peters/Roviró 2016)

Unabhängig vom Laibling, hinter dem sich teilweise unterschiedliche Vorstellungen zu Zielen und Umsetzungen anwendungsbezogener Grundlagenforschung verbergen, gibt es eine Reihe gemeinsamer Kerncharakteristika (vgl. u.a. Kelly 2005, 107; Cobb 2003, 9ff.; Confrey/Maloney 2015; Dube/Hußmann 2018, i.Dr.).

3. Kerncharakteristika von Design Research

3.1 Praxis- oder theoriebezogener Handlungsdruck als Ausgangspunkt

Ausgangspunkt von Design-Forschungsprojekten sind in erster Linie komplexe Probleme bzw. Herausforderungen der Schul- und Unterrichtspraxis sowie gegenstandsübergreifende Lerntheorien (vgl. Prediger et al. 2015, 880). Gewählte Praxisprobleme sollten analog zum Ausgangspunkt der Aktionsforschung häufig beobachtbar und von allgemeinem Interesse sein. Zudem sollte in Folge einer Veränderung der unterrichtlichen Ansätze und Konzepte das Problem verringert oder behoben werden (vgl. Ralle/Di Fuccia 2014, 44). Forschungsvorhaben im Format des Design Research greifen die beschriebenen Probleme auf und generieren anwendungsbezogene Lösungsansätze. Wie der Überblick zu verschiedenen Design-Research-Projekten u.a. in Plomp/Nievee (2013) oder bei Komorek/Prediger (2013) zeigt, reichen die generierten Ergebnisse von fachunabhängigen internationalen und nationalen Bildungsprogrammen sowie fachbezogenen Curricula über Projektbausteine, teilweise mit konkret, abgrenzbaren Lernangeboten bzw. Lehr- und Lernmaterialien, bis hin zur isolierten Aufgabenreflexion.

Supra-Ebene:	Arbeit an internationalen Curricula
Macro-Ebene:	Arbeit an nationalen Curricula
Meso-Ebene:	Arbeit an fachlichen Curricula
Mikro-Ebene:	(Weiter-)Entwicklung von fachbezogenen Lehr-/Lernarrangements
Nano-Ebene:	(Weiter-)Entwicklung von fachbezogenen Aufgabenformaten

Abb. 1: Zieldimensionen Fachdidaktischer Entwicklungsforschung (Dube/Hußmann 2018, i.Dr.)

Angesichts der pluralen Zielstellungen unterscheiden Prediger, Gravemeijer und Confrey (2015) deshalb zwei zentrale „archetypes of design research“ (ebd., 877). So grenzen sie die curriculumsorientierten Forschungsbemühungen, welche Veränderungen auf Supra-, Makro- und Mesoebene fokussieren, indem sie übergreifende Design-Prinzipien formulieren und Handreichungen sowie Fortbildungen für Lehrpersonen konzipieren, von Anstrengungen ab, die sich auf Veränderungen auf Mikro- und Nanoebene richten (vgl. Abb. 1).

Durch die zuvor gewählte adjektivistische Zuschreibung der Problemsituation als komplex wird jedoch deutlich, dass Lösungsvarianten nicht allein theoretisch entwickelt werden können.

3.2 Arbeit in und mit Design-Experimenten

Anders als in klassischen Implementationsstudien, in denen meist „von externen Experten bereits ausgearbeitete Innovationen" in den Unterricht integriert werden (top down) und sowohl Ziele und Methoden, aber auch Kriterien für den Erfolg des Vorhabens ex ante festgelegt sind (vgl. Gräsel/Parchmann 2004, 198), ist die Entwicklung einer Lernumgebung im Design Research wesentlicher Bestandteil des Forschungsprozesses. Als eine erfolgreiche Variante zur Zusammenführung theorie- und empiriegestützter Überlegungen in der Unterrichtsforschung beschrieb Ann Brown (1992) die Durchführung von „design experiments". Ihr Ziel ist es: „[…] to carry out formative research to test and refine educational designs based on theoretical principles derived from prior research" (Collins et al. 2004, 17). Angelehnt an medizinische Laborstudien, in denen z.B. das Schlafverhalten untersucht wird, um herauszufinden, *wie* genau der Schlaf des Probanden bzw. der Probandin verläuft und worin mögliche Ursachen eines schlechten Schlafes liegen, steht auch in Design-Experimenten der fachdidaktischen Lehr- und Lernforschung die Beobachtung der Prozesshaftigkeit des Lernens im Vordergrund. Hierzu werden einzelne Lernende, kleinere Schülergruppen oder ganze Klassen gebeten, vorbereitete Materialien zu bearbeiten, während ihr Lern- und Arbeitsverhalten audio- oder videografisch aufgezeichnet wird (vgl. von Aufschnaiter 2014, 81f.).

Aufgrund der Schwerpunktverschiebung von messbaren Veränderungen hin zur Erkenntnisgenerierung zu Potenzialen und Hürden des Lernprozesses ist es dabei nicht unüblich, auch nur einzelne isolierte Teilschritte und Aufgaben in einem Design-Experiment zu erforschen. In den unterschiedlichen Forschungsprojekten finden sich somit Design-Experimente im Rahmen groß angelegter Schulstudien, aber auch auf Klassen- und Lernerebene.

3.3 Iterative und zyklische Prozessforschung

Dem Ziel folgend, Lehr- und Lernmaterialien eng am Lehr- und Lernprozess zu entwickeln, um das komplexe Wechselspiel verschiedener Parameter von Beginn an mit im Blick zu haben, ist Design-Forschung stark prozessorientiert ausgerichtet, d.h. im Fokus stehen vor allem die Lernenden im konkreten Umgang mit den Materialien bzw. neuen Unterrichtsmethoden. Die Auswertung der Beobachtungen fließt anschließend in die Überarbeitung des Designs, welches erneut durchgeführt und empirisch begleitet wird. Arbeiten zur Theoriebildung flankieren die Forschung in wiederholenden Zyklen. Schrittweise (iterativ) wird die Intervention folglich optimiert und eingangs aufgestellte Hypothesen reformuliert sowie Entwicklungsprozesse und -prinzipien festgehalten.

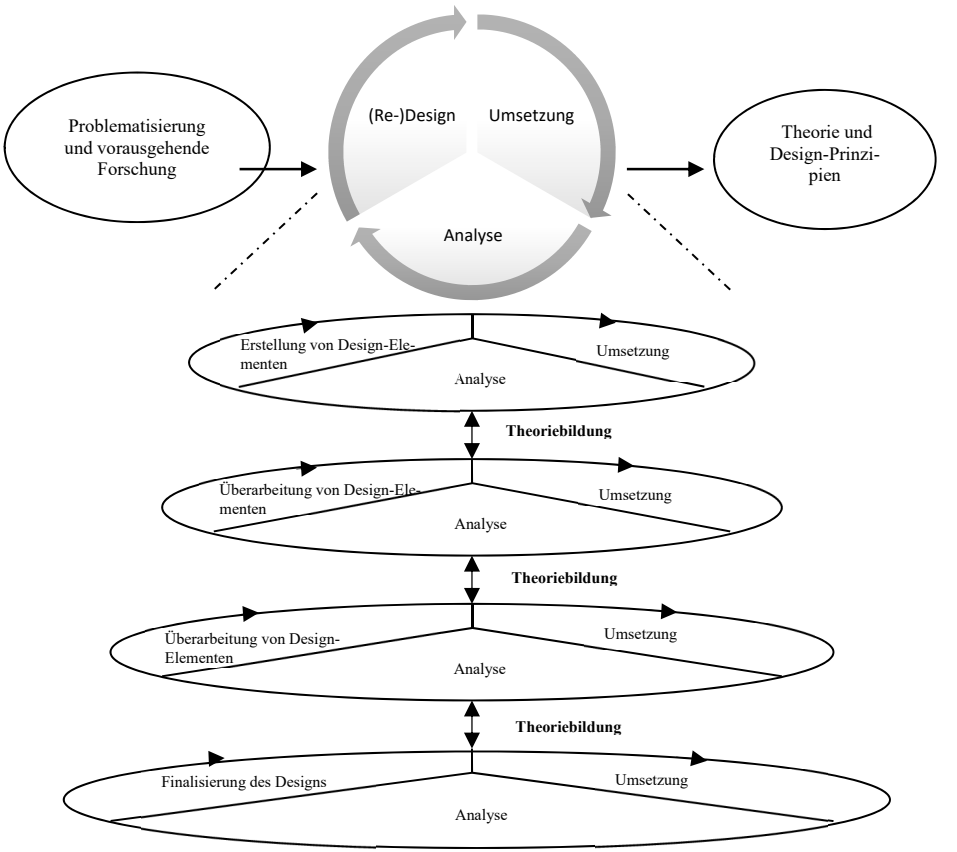

Abb. 2: Zyklisches Arbeiten im Design Research (Dube/Hußmann 2018, i.Dr.)

3.4 Theoriebasiert und theoriegenerierend

Der vorläufige Endpunkt der Forschungs- und Entwicklungsarbeit ist erreicht, wenn eine gewisse empirische Sättigung vorliegt. Die Zusammenführung aller Daten gibt nun Aufschluss über Bedingungen und Wirkungen der eingesetzten Design-Elemente, die als „Design-Prinzipien“ (van den Akker 1999; Reeves 2006) „ein Interpretationsangebot zum Vor- und Nachdenken über Praxisprobleme“ (Euler/Hahn 2007, 67) bzw. als „Erkenntnis- und Handlungsprinzipien“ (Hußmann et al. 2013, 26) die Konstruktion analoger Lernumgebungen in einem sich ständig verändernden Handlungsgefüge ermöglichen.

Andererseits bilden die einzelnen Prozessergebnisse auch die Voraussetzung für Theorien zur Strukturierung von (gegenstandsspezifischen) Lehr-/Lernprozessen (z.B. problemorientiertes Lernen, Interpretation literarischer Metaphern) oder für übergreifende didaktische Theorien. Hierzu werden meist in einem Wechselspiel von Induktion und Deduktion typische Verläufe und Hürden der beobachteten

Lernprozesse rekonstruiert. Wenngleich in ihrer Aussagekraft auf den Entstehungsprozess begrenzt, können sie durch den Vergleich mit weiteren Fällen eine globalere Theorienentwicklung vorbereiten (Prediger et al. 2012, 456).

Konsequenz jener hier vorgestellten Kerncharakteristika von Design-Forschung ist ein komplexes Forschungsdesign, in dem sowohl quantitative als auch qualitative Erhebungs- und Auswertungsinstrumente zum Einsatz kommen. Die damit gewonnenen Daten werden mit entsprechenden Methoden zur Interaktions- und Textanalyse ausgewertet und in einer Datenmatrix für die abschließende Interpretation zusammengefasst (vgl. Kapitel 4.4).

3.5 Verwandte Ansätze

Verwandte Ansätze sind neben den klassischen Evaluationsstudien, in denen Lernumgebungen ausschließlich theoriebasiert entwickelt und im Rahmen eines Pre-Post-Test-Designs empirisch überprüft werden, Forschungsbemühungen zur Didaktischen Rekonstruktion (vgl. Kattmann et al. 1997; Reinfried et al. 2009). Basierend auf den drei Bereichen: *Fachliche Klärung*, *Erfassung der Schülerperspektiven und Didaktische Strukturierung* stehen sachbezogene Strukturen und die Rekonstruktion von kognitiven Konstrukten und Theorien der Lernenden im Mittelpunkt der Forschung (vgl. Reinfried et al. 2009, 405). Wenngleich weniger auf die Gestaltung von Lernprozessen und die Entwicklung empiriebasierter Lernmaterialien ausgelegt, dient dieses „Orientierungsschema" (ebd., 404) ebenfalls dazu, Erkenntnisse des Lernenden, in diesem Fall seine „lebensweltlichen, vorunterrichtlichen Vorstellungen nicht als Hindernisse für das fachliche Lernen, sondern als Lernvoraussetzungen und potenzielle Lernhilfen" (ebd., 405) zu begreifen.

Große Ähnlichkeiten gibt es darüber hinaus auch zur Handlungsforschung und ihrem Ansatz der Aktionsforschung (vgl. Ralle/Di Fuccia 2014) bzw. partizipativen fachdidaktischen Aktionsforschung (vgl. Eilks/Ralle 2002). Der auf den Amerikaner Lewin (1948) zurückgehende Ansatz der Aktionsforschung, der in den 70er Jahren in den Erziehungs- und Sozialwissenschaften (vgl. u.a. Altrichter/Posch 1998) wiederentdeckt wurde und sich inzwischen in der Organisationsentwicklung, aber auch in den naturwissenschaftlichen Fachdidaktiken etabliert hat, basiert auf der Idee, Effekte und Veränderungen in einem realen sozialen Umfeld zu erforschen. Anders als im Design Research steht jedoch nicht der Lernende, sondern die Lehrperson als zentrale, handelnde Figur und die Weiterentwicklung ihrer Praxis im Mittelpunkt der Forschung. Eingebunden in eine professionelle Lerngemeinschaft aus weiteren Lehrpersonen, Wissenschaftlerinnen und Wissenschaftlern werden Phasen der *Zustandsdiagnose, Analyse* und *aktiven Entwicklung* gemeinsam zyklisch bearbeitet (vgl. Ralle/Di Fuccia 2014, 44). Dies bedeutet, dass Zugänge zur Problemlösung z.B. neue Unterrichtseinheiten und methodische Variationen von Lernprozessen und Medien, aber auch neue Wege und Formen der Reflexion und Evaluation von der Lehrperson in ihrem eigenen Unterricht erprobt und dokumentiert werden. Den beteiligten Wissenschaftlerinnen und Wissenschaftlern obliegt hingegen die Initiierung und Koordinierung von

Forschungsprozess und -team, das Datensammeln und -auswerten sowie die Literaturrecherche (vgl. Eilks/Ralle 2002).

Überschneidungen in den Zielsetzungen gibt es aber auch zur Praxisforschung, in der Praktikerinnen und Praktiker überwiegend ohne wissenschaftliche Unterstützung an der Optimierung ihres Unterrichts arbeiten (vgl. Prengel et al. 2008).

4. Was ist für ein Designforschungsprojekt zu bedenken?

4.1 Theoretische Vor- und Begleitarbeiten

Ausgangspunkt einer jeden guten Forschung ist die theoretische Aufarbeitung bisheriger und aktueller Erkenntnisse, die zur Beantwortung der ausgewählten Fragestellung beitragen könnten. Hierzu sollte nicht nur ein intensives Literaturstudium durchgeführt werden, sondern auch das Gespräch mit anderen Forscherinnen, Forschern und Arbeitsgruppen gesucht werden. Aufgrund der binären Zielstellung sowohl Erkenntnisse zu (gegenstandsspezifischen) Lerntheorien als auch zur Gestaltung empiriegestützter Lernmaterialien zu entwickeln, müssen im Rahmen von Design-Forschung ganz unterschiedliche Aspekte ggf. auch in ihrem Wechselspiel berücksichtigt werden. Expertise ist folglich nicht nur in der eigenen Fachdidaktik/-wissenschaft gefordert, sondern auch im Bereich der Motivations- und Lernpsychologie, der Methodik sowie der Bildungstheorie. Dementsprechend bietet es sich an, Kooperationen mit Vertreterinnen und Vertretern der verschiedenen Wissenschaften anzubahnen (vgl. Hopf et al. 2014, 35).

4.2 Schulische Feldforschung

Um Interventionsmaßnahmen im Feld durchzuführen, müssen Lehrkräfte im Vorfeld für die eigene Forschungsidee gewonnen und Einverständniserklärungen eingeholt werden. Da Lehrpersonen für Unterrichtsforschung nur in wenigen Schulen Entlastungsstunden erhalten, um sich in die Thematik einarbeiten zu können, empfehlen Wilhelm und Hopf (2014) den Lehrpersonen Medien und Schülerarbeitsmaterialien ausführlich vorzustellen. Damit verhindert wird, dass Lehrkräfte z.B. in verschiedenen Klassen das gleiche Lernmaterial unterschiedlich durchführen oder einzelnen Schülerinnen und Schüler in unsystematischer Weise Hinweise und Rückmeldungen geben (vgl. von Aufschnaiter 2014, 81), ist darauf zu achten sowohl die Grundideen der Interventionsmaßnahme, an die sich die Lehrkräfte unbedingt halten sollten, als auch jene Zugänge, die nur als unverbindliches Angebot verstanden werden, ausführlich am Material zu besprechen. Für eine gelingende Kooperation sollte dabei darauf geachtet werden, die Vorgaben für die Lehrkräfte nicht allzu eng zu halten und Rückmeldungen der Praktikerinnen und Praktiker frühzeitig aufzugreifen. Ebenso sollte kritisch zwischen *Nice to know*- und *need to know*-Fragen abgewogen werden, um sowohl die Störungen des Unterrichtsalltags zu minimieren als auch die zeitlichen Ressourcen durch Interviews, Tests und Fragebögen etc. nicht überzustrapazieren.

Praktische Empfehlungen u.a. zur Erstattung von Fahrtkosten für Lehrkräfte bei gemeinsamen Arbeitssitzungen sowie die Anmeldung der Treffen als Lehrerfortbildungsveranstaltung finden sich bei Wilhelm/Hopf (2014, 36).[2]

4.3 Sonderform Laborstudien

Vor dem Hintergrund der oben aufgeführten Herausforderungen schulischer Feldforschung haben sich in den letzten Jahren Laborstudien zur Aufzeichnung von Lernaktivitäten im Feld als Alternative zur schulischen Feldforschung etabliert (vgl. von Aufschnaiter 2014, 81). Die Lernenden werden dafür außerhalb der üblichen schulischen Lehr- und Lernsituation bestenfalls in Räumen der Hochschule eingeladen. Die Probandenzahl muss dabei nicht auf eine geringe Menge begrenzt bleiben, sondern kann in verschiedene Sitzungen aufgeteilt werden, sodass am Ende Datenmengen in Klassenstärke vorliegen können.

Durch die Arbeit außerhalb klassischer Unterrichtssettings werden negative Einflussfaktoren wie Lärm, Leistungsdruck oder soziale Spannungen auf das Lernen minimiert, sodass Lern- und Arbeitsprozesse deutlich ungestörter ablaufen. Nahezu unverstellt, ergibt sich ein detaillierter Einblick in die Nutzung sowie Bearbeitung von Instruktionen und Aufgaben. So wird z.B. deutlich, ob Verstehensschwierigkeiten im Umgang mit literarischen Metaphern auf ein unzureichendes konzeptionelles Vorwissen zur Metapher, auf die Aufgabenkonstruktion oder die Textgrundlage zurückzuführen sind (vgl. Dube/Prediger 2016). Zudem liefert der Blick auf dialogische Lernsituationen spezifische Parameter zur Rekonstruktion der Lernprozesse (vgl. von Aufschnaiter 2014, 85), die Rückschlüsse auf Denk- und Verstehensprozesse ermöglichen (vgl. ebd. 82f.).

In der Deutschdidaktik werden Laborstudien in den letzten Jahren vielfach eingesetzt, um Textbearbeitungs- und -interpretationsprozesse aufzuzeichnen. Hierzu wird neben der audio- oder videografischen Dokumentation auf die Methode des Lauten Denkens zurückgegriffen (vgl. Stark 2010; Dannecker 2016; Lessing-Sattari 2017; Dube/Prediger 2016).

Im Rahmen von fachdidaktischer Design-Forschung werden Laborstudien nicht nur eingesetzt, um die Denk- und Verstehensprozesse zu dokumentieren, sondern auch um etwas über die Bedingungen des Lernens zu erfahren. Dies ist z.B. der Fall, wenn:

- Lernmaterialien unter kontrollierten Bedingungen zunächst erprobt werden sollen, um zu explorieren, wie die Lernenden mit dem Material umgehen bzw. wie sich ihr Handeln z.B. durch die Ergänzung von mündlichen und schriftlichen Scaffolds oder Darstellungswechseln verändert.
- Hypothesen zu alltagsnahen oder fachlichen Vorstellungen und deren Veränderung durch eine Reformulierung der Lernmaterialien überprüft werden sollen (vgl. von Aufschnaiter 2014, 86).

[2] Weitere nützliche Hinweise zur Organisation und Durchführung von Erhebungen im schulischen Feld finden Sie in den Beiträgen von König und Iberer in diesem Band.

Im Mittelpunkt meist mehrerer aufeinanderfolgender Sitzungen steht folglich der durch gezielt gestaltetes Material initiierte Lernprozess, der Aufschluss geben soll zur Strukturierung und Spezifizierung des Lerngegenstands sowie zu den gegenstandsspezifischen Design-Prinzipien (vgl. Dube/Hußmann 2018, i.Dr.). Die kontrollierten Situationen bieten hier die Möglichkeit, Interaktionen zwischen Lernenden besser steuern zu können und ggf. gezielt Impulse zur Weiterarbeit zu setzen. Sind die Probandinnen und Probanden in der Lage, ihr neu erworbenes Wissen auch an ähnlichen Aufgaben anzuwenden, liegt die Annahme nah, dass die Probandinnen und Probanden etwas gelernt haben.

Neben der erleichterten Dokumentation von Verstehens- und Lernprozessen sei an dieser Stelle auch auf den pragmatischen Nutzen von Laborstudien verwiesen. Werden die Untersuchungen außerhalb der Schule durchgeführt, müssen hierfür keine schulischen oder ministeriellen Genehmigungsverfahren durchlaufen werden. Lediglich die Lernenden selbst oder bei Minderjährigkeit deren Erziehungsberechtigte müssen der Studienteilnahme zustimmen. Damit die Probandinnen und Probanden jedoch auch ggf. mehrmals zu den Sitzungen kommen, sollte über die Vergabe von Incentives in Form von Bargeld oder Gutscheinen nachgedacht werden, die allerdings erst bei der Teilnahme an allen Sitzungen ausgegeben werden sollten (vgl. von Aufschnaiter 2014, 90f.).[3]

Wenngleich Laborstudien bei Fragen zur Verstehens- und Lernprozessanalysen durch ihre kontrollierte Umgebung überzeugen, bleibt meist unklar, ob sich die ermittelten Effekte auch realen Unterricht einstellen würden. So kann es passieren, dass die Reduzierung auf einzelne Parameter den Blick auf Wechselwirkungen verstellt, die für die unterrichtliche Praxis konstitutiv sind (vgl. ebd. 85). Folglich beschreiben Laborstudien analog zu anderen Erhebungsverfahren stets nur einen kleinen Ausschnitt der Wirklichkeit, sodass zu ergänzenden Feldstudien geraten wird. Aufgrund des Erhebungs- und Auswertungsaufwands kann diese Kombination jedoch oft nicht im Rahmen eines einzelnen Projektes oder einer Dissertation erfolgen, sondern erfordert eine Folge von Forschungsprojekten, die auf ähnliche Fragestellungen abzielen und aufeinander aufbauen.

4.4 Datenerhebung und -auswertung

Mit dem Ziel sowohl gegenstandsbezogene als auch prozessbezogene Strukturen sowie ggf. markante Einflussgrößen auf das Lernen zu erheben, wird im Rahmen fachdidaktischer Entwicklungsforschung sowohl auf quantitative als auch qualitative Erhebungs- und Auswertungsmethoden zurückgegriffen. So finden sich in den Forschungsprojekten Frage- und Beobachtungsbögen sowie Einzelinterviews und Gruppendiskussionen unterschiedlichsten Strukturierungsgrades, aber auch

[3] Weiterführende Anregungen zur Vor- und Nachbereitung sowie zur Durchführung von Laborstudien finden sich zusammen mit Materialien u.a. zum Einholen von Einverständniserklärungen, Checklisten für Laborsitzungen etc. im Aufsatz von Claudia von Auschnaiter (2014).

Leistungstests sowie audio- und videobasierte Analysen von Unterrichtshandlungen.

Damit sowohl die Anzahl der Instrumente als auch die mit ihnen zu erhebende Datenmenge handhabbar bleibt, sollte die detaillierte Datenanalyse erst mit Abschluss der Konkretisierung der Forschungsfrage beginnen. Damit kann verhindert werden, dass die bereits zu Beginn investierte Arbeit in die Datenauswertung nach dem dritten Zyklus, der ggf. ganz wesentliche Erkenntnisse geliefert hat, umsonst war, da die Forschungsfrage reformuliert werden muss. Manchmal sind die Beobachtungen jedoch auch so ergiebig, dass im Anschluss an jeden neuen Zyklus eine eigene Dissertation verfasst werden könnte. Folglich sollten in einem Forschertagebuch nicht nur die konkreten Methoden und Lernsettings sowie grobe Verlaufsprotokolle, sondern nach Begutachtung der erhobenen Daten auch zentrale Beobachtungen (Schlüsselstellen) mit Datums- und Zeitangabe festgehalten werden. Sobald ausreichend Erkenntnisse durch eine erste deskriptive Analyse des Datenmaterials gewonnen sind, kann mit der kontrollierten Auswertung z.B. mithilfe von Kategoriensystemen begonnen werden. Datensätze, die nicht in diese Bearbeitungsschleife eingehen, sind jedoch nicht verloren. Vielmehr können sie als Beobachtungen in das Einleitungskapitel der Arbeit einfließen oder Anregungen für folgende Forschungsprojekte am Ende der Arbeit bieten.

4.5 Gütekriterien von Design Research

Forschungsarbeiten im Format von Design Research generieren, von einigen Ausnahmen abgesehen, insbesondere, wenn sie sich der gegenstandsspezifischen Rekonstruktion von Lernprozessen widmen (vgl. Hußmann et al. 2012), ihre Erkenntnisse anhand kleinerer Fallzahlen. Folglich sind Überlegungen zur Frage, ab wann die generierten Ergebnisse allgemeingültige Aussagen ermöglichen, stets in die Reflexion und Wertung der Ergebnisse eingeschlossen. Zur Beantwortung muss dabei auf wissenschafts- und erkenntnistheoretische Positionen zugegriffen werden, die sich entlang der gängigen Unterscheidung von quantitativer und qualitativer Sozialforschung entfalten. In den folgenden Überlegungen soll jedoch nicht verkannt werden, dass sich zwischen beiden Zugriffen oftmals nur schwer eine Trennlinie ziehen lässt. So werden Beobachtungsprotokolle z.B. nach dem Auftreten bestimmter Häufigkeiten von Äußerungen und Hinweisen quantitativ erfasst, während „quantitative Daten im Anschluss an eine standardisierte Erhebung und Auswertung interpretiert, also in Qualität übersetzt werden“ (Petri 2014, 97). Ungeachtet vielfacher Überschneidungen kann dennoch festgehalten werden, dass quantitative Designs überwiegend zur *Hypothesentestung* und qualitative Settings zur *Hypothesengenerierung* eingesetzt werden. Folglich besitzt das Forschungsdesign in Abhängigkeit der mit ihm zu beantworteten Fragestellungen sein besonderes Potenzial[4].

[4] Zur allgemeinen Unterscheidung von qualitativer und quantitativer Forschung sowie ihren jeweiligen Gütekriterien siehe die Beiträge zu *Forschungsparadigmen* in diesem Band.

Der überwiegende Einsatz von qualitativen Instrumenten in der Design-Forschung rechtfertigt sich dementsprechend vor dem Hintergrund, Beiträge zur Theoriebildung zu erbringen, indem spezifische Leistungen systematisch fokussiert und damit erste Hinweise auf Handlungsnotwendigkeiten gesammelt werden. Vor diesem Hintergrund bemisst sich die Aussagekraft jener Forschungen nicht in der Größe der Fallzahl, sondern in dem Anspruch, auf Basis von Einzelfallanalysen Ergebnisse zu formulieren, die typische Handlungsmuster und generelle Strukturen sichtbar machen (vgl. Lammek 2010, 284). Dabei unterliegen die Ausführungen jedoch auch den zentralen Kriterien der Validität, Objektivität und Reliabilität (siehe hierzu auch den Beitrag von Schmidt im vorliegenden Band). Während erstere über die Zuverlässigkeit und Vollständigkeit der erhobenen Daten sowie die Eindeutigkeit und Nachvollziehbarkeit der Interpretation beschrieben wird (vgl. Petri 2014, 98), meint Objektivität den „interpersonalen Konsens“ (Bortz/Döring 2006, 326). Sie gibt folglich an, ob unterschiedliche Forscherinnen und Forscher bei der Untersuchung desselben Sachverhalts mit denselben Methoden zu vergleichbaren Resultaten gekommen sind (vgl. ebd.).

Die Zuverlässigkeit (Reliabilität) der Ergebnisse von Design-Forschung kann jedoch auch über die methodologische Triangulation (vgl. u.a. Schründer-Lenzen 2010; Müller in diesem Band) ihrer unterschiedlichen Verfahren zur Datenerhebung beschreibbar werden. Zur Verifizierung oder Falsifizierung der Hypothesen können demnach Fragebogendaten und Feldbeobachtungen oder Testergebnisse auf Übereinstimmungen überprüft werden (vgl. Petri 2014, 99).

5. Fazit

Im vorliegenden Beitrag wurde mit dem Forschungsformat des Design Research ein noch junger Forschungszugriff in der Deutschdidaktik beschrieben. Thematisiert wurden nicht nur die begriffliche Vielfalt der teilweise synonym verwendeten Bezeichnungen sowie notwendige Überlegungen zur Vorbereitung, Durchführung und Auswertung, sondern auch dessen Kernmerkmale, wie:

- der Praxis- oder theoriebezogene Handlungsdruck als Ausgangspunkt des Forschungsprozesses.
- die iterative Abfolge von theoriebasierter Designentwicklung, Umsetzung und Analyse in mehreren Zyklen, dessen einzelne Phasen immer einen Rückbezug auf vorangegangene Phasen beinhalten.
- die komplexe Ergebnisgenerierung auf methodischer (empiriebasierter Lernmaterialien) und theoretischer Ebene (Erkenntnisse zu (gegenstandsspezifischen) Lehr- und Lernprozessen).

In der Zusammenschau der Ausführungen wird deutlich, dass Design Research nicht nur die in der Deutschdidaktik bisher bestehenden methodischen Zugangsweisen ergänzt, sondern eine für die Fachdidaktik spezifische Forschungsmethodik liefert, welche die Erforschung von gegenstandsspezifischen Lernprozessen in den Fokus rückt (vgl. Dube/Prediger 2017, 10).

Dabei wird die intensive Arbeit auf methodischer Ebene in Form von Modifizierungen des Materials nach jedem Zyklus (vgl. Abb. 2) vielmehr als Erkenntnisfortschritt, anstatt als Beitrag zur Allgemeingültigkeit gewertet (vgl. Reinmann/Sesink 2011, 18). Folglich stehen am Ende weniger abstrakt-distanzierte wissenschaftliche Aussagen „feststellenden Charakters“ als vielmehr „konkret-distanziert“ Erkenntnisse. Während erstere die nicht-verallgemeinerbaren Aspekte der empirischen Realität abstrahieren und dadurch zur Distanzierung der Theorie von der Praxis beitragen, bilden die Ergebnisse der Entwicklungsforschung die Grundlage für die Entwicklung neuer Perspektiven (vgl. ebd.).

Literatur

Altrichter, Herbert/Posch, Peter (1998): Lehrer erforschen ihren Unterricht. Bad Heilbrunn: Julius Klinkhardt.

Aufschnaiter von, Claudia (2014): Laborstudien zur Untersuchung von Lernprozessen. In: Krüger, Dirk/Parchmann, Ilka/Schecker, Horst (Hrsg.): Methoden in der naturwissenschaftsdidaktischen Forschung. Berlin: Springer, 81-95.

Barab, Sasha/Squire, Kurt (2004): Design-based research: Putting a stake in the ground. The Journal of the Learning Sciences, 13, 1, 1-14.

Bortz, Jürgen/Döring, Nicola (2006): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 4. Aufl. Berlin: Springer.

Brown, Ann (1992): Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. In: The Journal of the Learning Scienes, 2, 2, 141-178.

Cobb, Paul et al. (2003): Design Experiments in Educational Research. In: Design Experminets in Educational Research, 32, 9-13.

Collins, Alan/Joseph, Diana/Bielaczyc, Katerine (2004): Design Research. Theoretical and Methodological Issues. In: Journal of the Learning Sciences, 13, 15-42.

Confrey, Jere/Maloney, Alan (2015): A design study of a curriculum and diagnostic assessment system for a learning trajectory on equipartitioning. In: ZDM Mathematics Education, 47, 6, 919-932.

Dannecker, Wiebke (2016): Lautes Denken. Leise lesen und laut denken. Eine Erhebungsmethode zur Rekonstruktion von 'Lesespuren'. In: Boelmann, Jan M. (Hrsg.): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. Baltmannsweiler: Schneider Hohengehren, 131-146.

Design-based research collective (2003): Design-based research: an emerging paradigm for educational inquiry. In: Educational Researcher, 32, 5-8.

Dube, Juliane/Hußmann, Stephan (2018, i.Dr.): Fachdidaktische Entwicklungsforschung (Design Research). Theorie- und empiriegeleitete Gestaltung von Unterrichtspraxis. In: Sommer, Katrin/Mattiesson, Christiane/Priebe, Claudia (Hrsg.): Früher Bildungsdialog – Wissenschaftskommunikation zwischen Bildungsforschung und Schule. Bad Heilbrunn: Klinkhardt.

Dube, Juliane/Prediger, Susanne (2017): Design-Research – Ein Forschungszugang für praxisnahe Lernprozessforschung in der Deutschdidaktik. In: leseforum.ch 1/2017,

http://www.leseforum.ch/sysModules/obxLeseforum/Artikel/602/2017_1_Dube_Prediger.pdf (letzter Zugriff: 01.08.2018).

Eilks, Ingo/Ralle, Bernd (2002): Partizipative fachdidaktische Aktionsforschung – ein Modell für eine praxisnahe curriculare Entwicklungsforschung in der Chemiedidaktik. In: Chemcon, 9, 1, 13-18.

Euler, Dieter/Hahn, Angela (2007): Wirtschaftsdidaktik. Göttingen: UTB.

Euler, Dieter (2011): Wirkungs- vs. Gestaltungsforschung eine feindliche Koexistenz? In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 107, 520-542.

Einsiedler, Wolfgang (2010): Didaktische Entwicklungsforschung als Transferförderung. In: Zeitschrift für Erziehungswissenschaft, 13, 59-81.

Fischer, Frank/Waibel, Mira/Wecker, Christof (2005): Nutzenorientierte Grundlagenforschung im Bildungsbereich. Argumente einer internationalen Diskussion. In: Zeitschrift für Erziehungswissenschaften, 8, 3, 427-442.

Gräsel, Cornelia (2010): Stichwort: Transfer und Transferforschung im Bildungsbereich. In: Zeitschrift für Erziehungswissenschaft, 13, 7-20.

Gräsel, Cornelia/Parchmann, Ilka (2004): Implementationsforschung – oder: der steinige Weg, Unterricht zu verändern. In: Unterrichtswissenschaft, 32, 3, 196-214.

Hammann, Marcus/Jördens, Janina (2014): Offene Aufgaben codieren. In: Krüger, Dirk/Parchmann, Ilka/Schecker, Horst (Hrsg.): Methoden in der naturwissenschaftsdidaktischen Forschung. Berlin: Springer, 169-178.

Hußmann, Stephan/Richter, Vanessa (2012): ‚Wieso kann ein Navi so genau rechnen?' Mit linearen Funktionen modellieren. Praxis Mathematik, 54, 44, 15-19.

Hußmann, Stephan et al. (2013): Gegenstandsorientierte Unterrichtsdesigns entwickeln und erforschen – Fachdidaktische Entwicklungsforschung im Dortmunder Modell. In: Komorek, Michael/Prediger, Susanne (Hrsg.): Der lange Weg zum Unterrichtsdesign. Münster: Waxmann, 25-42.

Kahlert, Joachim (2007): Was kommt nach der Erkenntnis? Zum schwierigen Verhältnis pädagogischer Disziplinen zu der Erwartung, sich nützlich zu machen. In: Reinmann, Gabi/Kahlert, Joachim (Hrsg.): Der Nutzen wird vertagt… Bildungswissenschaften im Spannungsfeld zwischen wisseschaftlicher Profilbildung und praktischem Mehrwert. Lengerich: Pabst.

Kattmann, Ulrich/Duit, Reinders/Gropengießer, Harald/Komorek, Michael (1997): Das Modell der Didaktischen Rekonstruktion – Ein Rahmen für naturwissenschaftliche Forschung und Entwicklung. In: Zeitschrift für Didaktik der Naturwissenschaften, 3, 3, 3-18.

Komorek, Michael/Prediger, Susanne (Hrsg.) (2013): Der lange Weg zum Unterrichtsdesign: Zur Begründung und Umsetzung genuin fachdidaktischer Forschungs- und Entwicklungsprogramme. Münster: Waxmann.

Kelly, Anthony (2003): Quality criteria for design research: evidence and commitments. In: Van den Akker, Jan et al. (Hrsg.): Educational design research. New York: Routledge, 107-118.

Lamnek, Siegfried (2010): Qualitative Sozialforschung, 5. Aufl. Weinheim/Basel: Beltz.

Lessing-Sattari, Marie (2017): Didaktische Analyse der Metapher: Theoretische und empirische Rekonstruktionen von Verstehensanforderungen und Verstehenspotenzialen. In: Positionen der Deutschdidaktik, Band 5. Frankfurt a.M.: Peter Lang.

Mayer, Herbert et al. (2004): Qualitätskriterien von Assessmentinstrumenten – Cohen's Kappa als Maß der Interrater-Reliabilität. In: Die wissenschaftliche Zeitschrift für Pflegeberufe, 17, 36-46.

Olsen, Ralph (2011): Das Phänomen Empathie beim Lesen literarischer Texte. Eine didaktisch-kompetenzorientierte Annäherung. In: zeitschrift ästhetische bildung, 1, http://www.zaeb.net/index.php/zaeb/article/view/41/37 (letzter Zugriff: 01.08.2018).

Penuel, William R./Martin, Catherine (2015): Design-Based Implementation Research as a Strategy for Expanding Opportunity to Learn in School Districts. Paper presented at the Research Conference of the National Council of Teachers of Mathematics. Boston: MA.

Peters, Maria/Rovieró, Bàrbara (2017): Fachdidaktischer Forschungsverbund FaBiT: Erforschung von Wandel im Fachunterricht mit dem Bremer Modell des Design-Based Research. In: Doff, Sabine/Komoss, Regine (Hrsg.): Making Change Happen. Wandel im Fachunterricht analysieren und gestalten. Wiesbaden: Springer, 19-32.

Petri, Jürgen (2014): Fallstudien zur Analyse von Lernpfaden. In: Krüger, Dirk et al. (Hrsg.): Methoden der naturwissenschaftsdidaktischen Forschung. Berlin: Springer, 95-105.

Plomp, Tjeerd/Nieveen, Nienke (Hrsg.) (2013): Educational design research. Enschede: SLO.

Prediger, Susanne/Link, Michael (2012): Fachdidaktische Entwicklungsforschung – Ein lernprozessfokussierendes Forschungsprogramm mit Verschränkung fachdidaktischer Arbeitsbereiche. In: Bayrhuber, Horst et al. (Hrsg.): Formate Fachdidaktischer Forschung. Empirische Projekte – historische Analysen – theoretische Grundlegungen. Fachdidaktische Forschungen, Band 2. Münster: Waxmann, 29-46.

Prediger, Susanne et al. (2012): Lehr-Lernprozesse initiieren und erforschen – Fachdidaktische Entwicklungsforschung im Dortmunder Modell. In: Der Mathematische und Naturwissenschaftliche Unterricht, 65, 8, 452-457.

Prediger, Susanne/Gravemeijer, Koeno/Confrey, Jere (2015): Design research with a focus on learning processes – an overview on achievements and challenges. In: ZDM Mathematics Education, 47, 6, 877-891.

Prengel, Annedore/Heinzel, Friederike/Carle, Ursula (2008): Methoden der Handlungs-, Praxis- und Evaluationsforschung. In: Helsper, Werner (Hrsg.): Handbuch der Schulforschung. Wiesbaden: Verlag für Sozialwissenschaften, 181-197.

Ralle, Bernd/Di Fuccia, David-Samuel (2014): Aktionsforschung als Teil fachdidaktischer Entwicklungsforschung. In: Krüger, Dirk/Parchmann, Ilka/Schecker, Horst (Hrsg.): Methoden der naturwissenschaftsdidaktischen Forschung. Berlin: Springer, 43-57.

Reeves, Thomas (2000): Enhancing the worth of instructional technology research through "design experiments" and other development research strategies, http://it.coe.uga.edu/~treeves/AERA2000Reeves.pdf (letzter Zugriff: 01.08.2018).

Reeves, Thomas (2006): Design research from a technology perspective. In: Van den Akker, Jan/Gravemeijer, Koeno/McKenney, Susan/Nieveen, Nienke (Hrsg.): Educational design research. London: Routledge, 52-66.

Reigeluth, Charles (1983): Instructional design: What is it and why is it? In: Reigeluth Charles (Hrsg.): Instructional-design theories and models: An overview of their current status. Hillsdale, NJ: Lawrence Erlbaum Associates, 3-36.

Reinfried, Sibylle et al. (2009): Das Modell der Didaktischen Rekonstruktion. Eine innovative Methode zur fachdidaktischen Erforschung und Entwicklung von Unterricht. In: Beiträge zur Lehrerinnen- und Lehrerbildung, 27, 404-414.

Reinmann, Gabi/Sesink, Werner (2011): Entwicklungsorientierte Bildungsforschung (Diskussionspapier). http://gabi-reinmann.de/wp-content/uploads/2011/11/Sesink-Reinmann_Entwicklungsforschung_v05_20_11_2011.pdf (letzter Zugriff: 01.08.2018).

Schründer-Lenzen, Agi (2010): Triangulation. Ein Konzept zur Qualitätssicherung von Forschung. In: Friebertshäuser, Barbara/Langer, Antje/Prengel, Annedore (Hrsg.): Handbuch qualitative Forschungsmethoden in der Erziehungswissenschaft. Weinheim: Beltz Juventa, 149-158.

Stark, Tobias (2010): Lautes Denken in der Leseprozessforschung. Kritischer Bericht über eine Erhebungsmethode. In: Didaktik Deutsch, 29, 58-83.

Tenorth, Heinz Elmar (2012): Bildungsphilosophie – Bildungsforschung – Erziehungswissenschaft. In: Zeitschrift für Erziehungswissenschaft, 15, 403-407.

Van den Akker, Jan (1999). Principles and methods of development research. In: Van den Akker, Jan et al. (Hrsg.): Design approaches and tools in education and training. Dordrecht: Kluwer Academic Publishers, 1-14.

Van den Akker, Jan (2003): Curriculum Perspectives: An Introduction. In: Curriculum Landscapes and Trends. Dordrecht: Springer, 1-10.

Van den Akker, Jan et al. (2006): Educational design research. New York: Routledge.

Wilhelm, Thomas/Hopf, Martin (2014): Design-Forschung. In: Krüger, Dirk/Parchmann, Ilka/Schecker, Horst (Hrsg.): Methoden in der naturwissenschaftsdidaktischen Forschung. Berlin: Springer, 31-46.

Wittmann, Erich Ch. (1995): Mathematics Education as a »Design Science«. Educational Studies in Mathematics, 29, 355-374.

JASMIN BENZ

Evaluationsforschung

1. Einleitung

Evaluationsforschung ist ein anwendungsorientierter empirischer Forschungsansatz (vgl. Döring/Bortz 2016, 976), der in den letzten Jahren immer mehr an Bedeutung gewonnen hat (vgl. Stockmann 2016, 27). Es gibt zahlreiche Anwendungsbereiche für Evaluationsforschung (vgl. z.B. Gollwitzer/Jäger 2014, 32f.), wie neue Therapieansätze in der Psychologie, Weiterbildungsmodule für medizinische Fachkräfte oder die Folgen einer veränderten Rechtsprechung. Deutschdidaktisch orientierte Evaluationsforschung kann sich beispielsweise mit der Veränderung von Lesekompetenz bei Schülern durch neu konzipierte Fördermaterialien, der Wirksamkeit vom Einsatz realer Schülertexte in schreibdidaktischen Seminaren für Lehramtsstudierende oder der Akzeptanz von Unterrichtskonzepten zum Einsatz von Tablets bei Lehrkräften beschäftigen.

In diesem Kontext stellt der vorliegende Beitrag eine erste Einführung zu Evaluationsforschung dar, um vor allem Entscheidungsprozesse für oder gegen eine eigene empirische Evaluationsstudie zu unterstützen. Da Evaluationsforschung methodisch anspruchsvoll ist (vgl. Döring/Bortz 2016, 976), wird immer wieder deren Eignung für Bachelor- und Masterarbeiten sowie Promotionsvorhaben thematisiert. Zugleich kann an dieser Stelle keine umfassende Darstellung erfolgen, stattdessen werden die nachstehenden Schwerpunkte gesetzt und überblicksartig ausgeführt:

Zunächst wird der Begriff Evaluationsforschung definiert (Kapitel 2), wobei eine Abgrenzung zur Grundlagenforschung (siehe den Beitrag von Gailberger in diesem Band) stattfindet sowie mögliche Funktionen von Evaluationsforschung vorgestellt werden. Kapitel 3 enthält einen Überblick zentraler Ansätze von Evaluationsforschung und eine vertiefende Darstellung zu wirksamkeitsorientierten Evaluationsstudien, denn diese sind für Evaluationsforschung von besonderer Bedeutung (vgl. Döring/Bortz 2016, 998). Da der Erfolg einer Evaluationsstudie, wie generell beim wissenschaftlichen Arbeiten, von deren Qualität abhängt (vgl. Stockmann 2016, 50), werden im nächsten Schritt (Kapitel 4) Standards für Evaluationsforschung vorgestellt. Um eine erste Vorstellung vom wissenschaftlichen Evaluationsprozess zu ermöglichen, beschäftigt sich Kapitel 5 mit dem möglichen Ablauf einer Evaluationsstudie. Das abschließende Fazit verweist u.a. auf Checklisten für Evaluationsstudien.

2. Was ist Evaluationsforschung?

Die moderne Evaluationsforschung[1] hat ihren Ursprung in den USA und gewann gegen Ende der 60er Jahre auch in Deutschland an Bedeutung, u.a. in den Bereichen Bildung und Erziehung (vgl. Stockmann 2016, 31f.). In der Literatur zu Entwicklung und Charakteristika der Evaluationsforschung wird diese begrifflich teils synonym und teils in Abgrenzung zu Evaluation verstanden. Grundsätzlich sollte man im Rahmen einer wissenschaftlichen Studie das entsprechende Begriffsverständnis von Evaluation bzw. Evaluationsforschung thematisieren. Dies ist vor allem deshalb notwendig, da Evaluation in der heutigen Zeit auch in alltäglichen Kontexten zur Bezeichnung von Bewertungsvorgängen verwendet wird (vgl. Döring/Bortz 2016, 978). Evaluationsforschung dagegen meint Bewertungen, die wissenschaftliche Methoden nutzen und wissenschaftlichen Standards genügen – entsprechend findet sich auch der Begriff wissenschaftliche/wissenschaftsbasierte Evaluation (vgl. z.B. Stockmann/Meyer 2014, 63). Döring und Bortz betrachten verschiedene Definitionsvorschläge in der nationalen sowie internationalen Literatur und führen diese zusammen; im Folgenden findet sich ein zentraler Auszug:

> Die Evaluationsforschung (‚evaluation research') bzw. wissenschaftliche Evaluation (‚evaluation') nutzt sozialwissenschaftliche Methoden, um einen **Evaluationsgegenstand** (z.B. ein Produkt oder eine Maßnahme) unter Berücksichtigung der relevanten **Anspruchsgruppen** (z.B. Patienten, Angehörige, Produktentwickler, Evaluationsauftraggeber) anhand bestimmter **Evaluationskriterien** (z.B. Akzeptanz, Wirksamkeit, Effizienz, Nachhaltigkeit) und Maßgaben zu ihren Ausprägungen zu **bewerten**. […] (Döring/Bortz 2016, 979)

In der deutschdidaktischen Forschung sind zahlreiche Evaluationsgegenstände denkbar: Beispielsweise kann neu konzeptioniertes Unterrichtsmaterial bewertet werden. Als Evaluationskriterium ist etwa die Akzeptanz der Materialien bei den Lehrkräften oder die Wirksamkeit der Materialien in Bezug auf das Lernen der Schüler denkbar. Die Evaluationskriterien sowie die eingesetzten Forschungsmethoden hängen eng mit den jeweiligen Forschungsfragen zusammen. Evaluationsforschung benötigt außerdem festgelegte und nachvollziehbare Bewertungsmaßstäbe (vgl. Döring/Bortz 2016, 983): Bei welchen Merkmalsausprägungen kann man beispielsweise von einer hohen Akzeptanz seitens der befragten Lehrkräfte gegenüber dem neu konzipierten Unterrichtsmaterial sprechen? Insgesamt zeigt die bisherige Begriffsklärung, dass Evaluationsforschung einen hohen wissenschaftlichen Anspruch an die durchführende Person stellt. Daher kann es insbesondere für Bachelor- und Masterarbeiten sinnvoll sein, einen bereits bestehenden Evaluationsgegenstand (z.B. innovatives Unterrichtsmaterial, das von einer drit-

[1] Einen historischen Überblick zu Evaluation/Evaluationsforschung findet sich z.B. bei Gollwitzer/Jäger 2014; Stockmann 2016; Stockmann/Meyer 2014.

ten Person konzipiert wurde) zu untersuchen. So besteht die Möglichkeit, die eigenen Ressourcen zu schonen und einen vergleichsweise unabhängigen Blick von außen (Fremdevaluation[2]) auf den Evaluationsgegenstand zu erhalten.

Einen weiteren zentralen Aspekt bei der Begriffsklärung hinsichtlich Evaluationsforschung stellt die Abgrenzung zur Grundlagenforschung (siehe Gailberger in diesem Band) dar: So geht es bei Grundlagenforschung darum, Erkenntnisse zu sammeln, „[…] ohne dass dabei die Frage nach der Nützlichkeit dieses Tuns für die Gesellschaft gestellt wird […]“ (Stockmann 2016, 35). Im Vergleich dazu steht bei Evaluationsforschung deren Nutzen für die Bewältigung konkreter gesellschaftlicher Problemstellungen im Vordergrund. Vor diesem Hintergrund „[…] bewegt sich die Evaluationsforschung in einem *Spannungsverhältnis zwischen Wissenschaftlichkeit und Nützlichkeit*.“ (Stockmann/Meyer 2014, 66).

Die Nützlichkeit von Evaluationsforschung hängt mit deren möglichen Funktionen zusammen. Stockmann (2016, 37ff.) unterscheidet vier Leitfunktionen, die miteinander verbunden sind und in wissenschaftlichen Evaluationsstudien unterschiedlich gewichtet werden können:

- Erkenntnis: Das Erkenntnisinteresse bei Evaluationsstudien kann vielfältig sein und beispielsweise betrachten, inwieweit ein neues Seminarangebot die Bedürfnisse der Studierenden berücksichtigt. Im Vordergrund steht somit der Nutzen gewonnener Erkenntnisse für die beteiligten Akteure, wie in diesem Beispiel für die Studierenden.
- Kontrolle: Dabei steht im Mittelpunkt, inwieweit vorab gesetzte Ziele im Hinblick auf den Evaluationsgegenstand (z.B. ein neues Seminarkonzept) erreicht wurden. Es kann etwa den Einsatz finanzieller Ressourcen und die Eignung der eingesetzten Akteure geprüft werden.
- Entwicklung: Diese Leitfunktion zielt darauf, Anhaltspunkte für die (Weiter-)Entwicklung des Evaluationsgegenstandes zu sammeln. Dabei wird sowohl betrachtet, was (aus welchem Grund) bereits gelungen erscheint als auch Potenzial zur Überarbeitung.
- Legitimation: An dieser Stelle geht es darum, die eingesetzten Ressourcen nach außen zu legitimieren. Eine Evaluationsstudie mit positiven Ergebnissen hilft außerdem, zukünftige Mittel zu organisieren. Dennoch gebietet es der wissenschaftlich-ethische Anspruch, die Ergebnisse von Evaluationsforschung im Hinblick auf eine mögliche Legitimation nicht zu beschönigen.

In Bezug auf die dargestellten Leitfunktionen lassen sich außerdem formative und summative Evaluationsstudien unterscheiden (vgl. Döring/Bortz 2016, 990): Die formative wissenschaftliche Evaluation zielt darauf, den Evaluationsgegenstand zu optimieren. Daher ist es sinnvoll, die Ergebnisse gemeinsam mit den beteiligten Akteuren zu reflektieren. Entsprechend der Zielsetzung kommen beim forma-

[2] Weitere Ausführungen zu Fremd- und Selbstevaluation finden sich z.B. bei Döring/Bortz 2016, 898f.

tiven Evaluieren häufig qualitative Untersuchungsmethoden zum Einsatz. Summative Evaluationsstudien stellen eine Zusammenfassung dar und fokussieren auf Kontrolle sowie Legitimation; dabei werden oft quantitative Erhebungsverfahren gewählt. Auch beim summativen Evaluieren sollten die Ergebnisse den beteiligten Akteuren zugänglich gemacht werden.

Im folgenden Kapitel geht es um verschiedene Ansätze der Evaluationsforschung, deren Kenntnis zur Verortung eigener Studien notwendig ist.

3. Ansätze von Evaluationsforschung

3.1 Erster Überblick zentraler Ansätze der Evaluationsforschung

In der Theoriebildung zur Evaluationsforschung existieren verschiedene Ansätze, die sich hinsichtlich zentraler Zielvorstellungen, dem grundlegenden Verständnis wissenschaftlicher Evaluation und den eingesetzten Forschungsmethoden unterscheiden (vgl. Döring/Bortz 2016, 995). Allerdings liegt in der entsprechenden Fachliteratur keine einheitliche Systematik bestehender Ansätze vor. Für den folgenden Überblick wird auf die Darstellung von Döring und Bortz (2016, 995ff.) zurückgegriffen, da diese zentrale Ansätze gegeneinander abgrenzt:

- *Ergebnisorientiert:* Im Mittelpunkt des Forschungsinteresses stehen ausgewählte Aspekte des Evaluationsgegenstands, wie die Akzeptanz einer Lehrerfortbildung durch die teilnehmenden Lehrpersonen. Die eingesetzten Forschungsmethoden fokussieren auf die Erhebung der ausgewählten Aspekte. Häufig werden die Wirksamkeit und/oder die Effizienz einer Maßnahme untersucht.
- *Systematisch:* Hier geht es darum, den Evaluationsgegenstand umfassend einzuschätzen. Entsprechend werden nicht nur ausgewählte, ergebnisorientierte Aspekte in den Blick genommen, sondern vielmehr der Prozess insgesamt (am Beispiel einer Lehrerfortbildung: von der Konzeption [Berücksichtigung des aktuellen Forschungsstands] über die Durchführung [Beobachtung der einzelnen Akteure] bis zu den Ergebnissen [Veränderungen durch die Fortbildung]). Von Interesse ist auch der Einfluss möglicher Umgebungsfaktoren. Daher sind im systematischen Ansatz häufig mehrere Teilevaluationen nötig.
- *Theorieorientiert:* Bei theorieorientierten Ansätzen wird der Evaluationsgegenstand hinsichtlich dem Kriterium Wirksamkeit differenziert betrachtet: „**Welche** Interventionskomponenten funktionieren für **welche** Zielgruppe unter **welchen** Umständen **wie** und **warum**?“ (Döring/Bortz 2016, 1009). Dabei ist wichtig, auf welche zugrundeliegenden Theorien sich vorgefundene Evaluationsergebnisse zurückführen lassen. Für die Betrachtung von Wirkungszusammenhängen ist in der Regel der Einsatz mehrerer Erhebungsinstrumente notwendig.

- *Akteursorientiert:* Der Bezeichnung entsprechend stehen bei diesen Ansätzen die jeweils beteiligten Akteure im Mittelpunkt. Man kann klienten- von stakeholderorientierten Ansätzen unterscheiden: Im ersten Fall (klientenorientiert) wird eine Akteursgruppe besonders berücksichtigt. Bei der Evaluation eines neuen Seminarkonzepts können dies beispielsweise die teilnehmenden Studierenden sein. Die stakeholderorientierten Ansätze (Stakeholder: Anspruchsgruppen einer Evaluation) berücksichtigen dagegen alle Akteursgruppen (bei einem neuen Seminarkonzept z.B. Studierende und Dozenten) – allerdings nicht unbedingt in einem ausgewogenen Verhältnis. Grundsätzlich sollten bei jeder Evaluationsstudie die jeweiligen Akteure einbezogen werden – bei akteursorientierten Ansätzen geschieht dies in besonders ausgeprägter Form.

In der Evaluationspraxis werden die oben skizzierten Ansätze selten in Reinform umgesetzt – häufig finden sich Elemente von mehr als einem Ansatz wieder. Für die Darstellung eigener Studien ist es dabei wichtig, solche theoretischen Kombinationen zu thematisieren und in Bezug auf die Fragestellung zu begründen. Insbesondere hinsichtlich der Konzeption von Bachelor- und Masterarbeiten sollte darauf geachtet werden, die Fragestellung und den entsprechenden theoretischen Ansatz so zu wählen, dass der Einsatz von Erhebungsinstrumenten überschaubar bleibt. Insgesamt betrachtet hängt die Qualität einer Evaluationsstudie – unabhängig von dem zugrundeliegenden Ansatz – vom konkreten Forschungsdesign ab. Im nächsten Teilkapitel wird dies für Evaluationsstudien zur Wirksamkeit beispielhaft verdeutlicht.

3.2 Vertiefung: Evaluationsstudien zur Wirksamkeit

Die Wirksamkeit (auch: Effektivität) gilt als eines der zentralen Evaluationskriterien und spielt bei allen beschriebenen Ansätzen eine Rolle. Um die Wirksamkeit einer Maßnahme, z.B. eines neuen Seminar- oder Förderkonzepts, aussagekräftig aufzuzeigen, „[…] ist der wissenschaftliche Nachweis der Kausalität zentral […]“ (Döring/Bortz 2016, 998): Dazu eignet sich als Forschungsdesign das Experiment. Da es nicht immer möglich ist, Interventions- und Kontrollgruppen nach dem Zufallsprinzip zu besetzen, finden in der empirischen Bildungsforschung häufig Quasi-Experimente statt. Ein Beispiel aus der deutschdidaktischen Forschung für ein quasi-experimentelles Design findet sich im Rahmen des Projekts „Wissenschaftliche Begleitforschung Gemeinschaftsschulen in Baden Württemberg (WissGem)“ (Bohl/Wacker 2016). In einem Teilprojekt wird die Wirksamkeit einer Lehrerfortbildung zur Diagnose und Förderung von Schreibkompetenz untersucht (vgl. Grausam et al. 2016, 115). Um aufzuzeigen, dass mögliche Veränderungen nach der Lehrerfortbildung tatsächlich auf die Fortbildung (und nicht z.B. auf zunehmende Erfahrung im Unterrichtsalltag) zurückzuführen sind, wird ein Prä-Post-Design mit Interventions- und Kontrollgruppe verwendet[3]. Somit

[3] Die Verteilung der Lehrkräfte auf Interventions- und Kontrollgruppe erfolgt hier nicht zufällig, sondern aufgrund äußerer Gegebenheiten. Daher handelt es sich um ein Quasi-Experiment.

kommen die eingesetzten Erhebungsinstrumente vor und nach der Fortbildung zum Einsatz (Prä-Post-Design) und lediglich die Interventionsgruppe erhält eine Fortbildung. Neben diesem Ansatz gibt es in der Theorie noch zahlreiche weitere Umsetzungsmöglichkeiten für Experimente und Quasi-Experimente; einen zielführenden Überblick findet man in Döring/Bortz (2016).

Die folgende Tabelle stellt einen Bezug zwischen den Wirksamkeitsebenen bei Maßnahmen allgemein nach Kirkpatrick und Kirkpatrick (2012), den Ebenen nach Lipowsky (2010) für Lehrerfortbildungen und dem soeben beschriebenen Beispiel aus der deutschdidaktischen Forschung (vgl. Grausam et al. 2016) her. Insgesamt soll aufgezeigt werden, dass bei Studien zur Wirksamkeit einer untersuchten Maßnahme verschiedene Wirksamkeitsebenen berücksichtigt werden können:

Tab. 1: Mögliche Wirksamkeitsebenen einer Maßnahme

Kirkpatrick/ Kirkpatrick (2012) … bei Maßnahmen allgemein	**Lipowsky (2010) … spezifisch bei Lehrerfortbildungen**	**Grausam et al. (2016) … Beispiel aus der deutschdidaktischen Forschung**
Level 1: Reaction	Ebene 1: Einschätzungen der Lehrkräfte (z.B. Akzeptanz einer Fortbildung)	wird nicht erhoben (könnte z.B. durch einen Fragebogen am Ende der Fortbildung erhoben werden)
Level 2: Learning	Ebene 2: Veränderungen von Lehrerkognitionen (z.B. fachdidaktisches Wissen)	Bei Interventions- und Kontrollgruppe: Beurteilungen zu Schülertexten eingesammelt[4]
Level 3: Behavior	Ebene 3: Veränderungen beim Lehrerhandeln im tatsächlichen Unterricht	
Level 4: Results	Ebene 4: Veränderungen bei den Schülern (z.B. Schülerleistung)	Die eingesammelten Schülertexte werden hinsichtlich ihrer Qualität geratet.

Die vier Level nach Kirkpatrick und Kirkpatrick (2012) beziehen sich auf Maßnahmen ganz verschiedener Art (beispielsweise auch im wirtschaftlichen Bereich) und müssen für den, in der eigenen Studie untersuchten, Evaluationsgegenstand angepasst werden. Tabelle 1 zeigt eine Konkretisierung für Lehrerfortbil-

[4] Da es sich um authentische Schülertexte aus dem tatsächlichen Unterricht der jeweiligen Lehrkräfte handelt, gibt es hier eine Überschneidung der Ebenen 2 und 3. Würde man den Lehrkräften vor und nach der Fortbildung einen vom Forschenden mitgebrachten Schülertext zur Beurteilung vorlegen, hätte man Ergebnisse nur zu Ebene 2.

dungen (vgl. Lipowsky 2010) mit einem deutschdidaktischen Beispiel (vgl. Grausam et al. 2016). Anhand des Beispiels wird auch deutlich, dass man nicht alle Ebenen zwingend untersuchen muss. Wichtig ist nur, zu thematisieren, warum welche Wirksamkeitsebene untersucht bzw. nicht untersucht wird. Stellenweise können dabei auch schon die zu Verfügung stehenden Ressourcen begrenzend wirken; dies ist bei der Diskussion der Ergebnisse zu berücksichtigen.

4. Gütekriterien – Evaluationsstandards

Je nach Forschungsinteresse und Fragestellung können bei wissenschaftlichen Evaluationsstudien sowohl qualitative als auch quantitative Erhebungs- und Auswertungsverfahren zum Einsatz kommen; darüber hinaus sind Mixed-Methods-Designs (siehe den Beitrag von Müller in diesem Band) möglich. Entsprechend sollte die eigene Evaluationsstudie an den jeweils notwendigen Gütekriterien gemessen werden. Da Gütekriterien qualitativer (siehe den Beitrag von Schmidt in diesem Band) und quantitativer (siehe den Beitrag von Schmidt in diesem Band) Forschung im vorliegenden Sammelband bereits in anderen Beiträgen ausführlicher dargestellt werden, sei an dieser Stelle lediglich auf zwei weitere zielführende Informationsquellen hingewiesen: Steineke (2013) diskutiert die Notwendigkeit eigener Gütekriterien für qualitative Forschung und stellt schließlich zentrale Kernkriterien vor. Bühner (2011) stellt Objektivität, Reliabilität und Validität als Hauptgütekriterien quantitativer Forschung differenziert dar und geht auch auf weitere Nebengütekriterien ein.

Neben den klassischen Gütekriterien gibt es für Evaluationsforschung Standards, welche zur Qualität wissenschaftlicher Evaluationsstudien beitragen sollen. National liegen von der Gesellschaft für Evaluation (DeGEval 2016) Standards vor, die sich an international bereits länger etablierten Standards (JCSEE-Standards, SEVAL-Standards) orientieren. Die DeGEval-Evaluationsstandards sind nicht rechtlich bindend, dienen aber als Orientierungsrahmen für qualitativ hochwertige Evaluationsstudien. Dazu werden vier zentrale Eigenschaften benannt und in insgesamt 25 Einzelstandards ausdifferenziert sowie erläutert (vgl. DeGEval 2016):

Tab. 2: DeGEval-Evaluationsstandards (Beispiele aus DeGEval 2016)

Übergeordnete Eigenschaft	**Beispiele für ausdifferenzierte Einzelstandards**
Nützlichkeit (N)	Insgesamt 8 Einzelstandards, Beispiele: N2: Klärung der Evaluationszwecke N6: Vollständigkeit und Klarheit der Berichterstattung
Durchführbarkeit (D)	Insgesamt 3 Einzelstandards, Beispiel: D1: Angemessene Verfahren
Fairness (F)	Insgesamt 5 Einzelstandards, Beispiele: F2: Schutz individueller Rechte F5: Offenlegung von Ergebnissen und Berichten
Genauigkeit (G)	Insgesamt 9 Einzelstandards, Beispiele: G1: Beschreibung des Evaluationsgegenstandes G6: systematische Fehlerprüfung

Die DeGEval-Standards sind als Maximalstandards gedacht – nicht alle Standards können bei jeder Evaluationsstudie in gleichem Maße berücksichtigt werden. Bereits bei den übergeordneten Eigenschaften fällt auf, dass sich die Standards je nach konkretem Forschungsfeld gegenseitig begrenzen können: Beispielsweise kann die Durchführbarkeit zu Lasten der Genauigkeit gehen (vgl. DeGEval 2016, 28f.). Wichtig für eigene Evaluationsstudien ist es, den Bezug zu übergeordneten Standards darzustellen und zu diskutieren, inwieweit das Forschungsprojekt diesen gerecht wird.

Neben den Evaluationsstandards ist ein durchdachter Ablauf von Evaluationsstudien maßgeblich für deren Qualität und Nützlichkeit.

5. Ablauf von Evaluationsforschung

Wie bei allen wissenschaftlichen Studien besteht auch bei Evaluationsforschung der Forschungsprozess aus verschiedenen Phasen. Grob unterteilen lassen sich: Initiierungs-, Konzeptions-, Planungs-, Realisierungs- und Abschlussphase (vgl. Döring/Bortz 2016, 1017ff.). Die folgende Abbildung zeigt beispielhaft den siebenschrittigen Ablauf einer qualitativen Evaluationsstudie zu einem Seminarangebot an der Phillips-Universität Marburg (vgl. Kuckartz et al. 2008); dabei ging es um die Einschätzungen der Studierenden zu einer statistischen Einführungsveranstaltung. Diese Studie ist insbesondere als Beispiel für Bachelor- und Masterarbeiten – aber auch für Promotionsvorhaben – von Interesse, da hier als zentrale Zielsetzung die Umsetzung einer ertragreichen qualitativen Studie mit einem vertretbaren Arbeitspensum genannt wird.

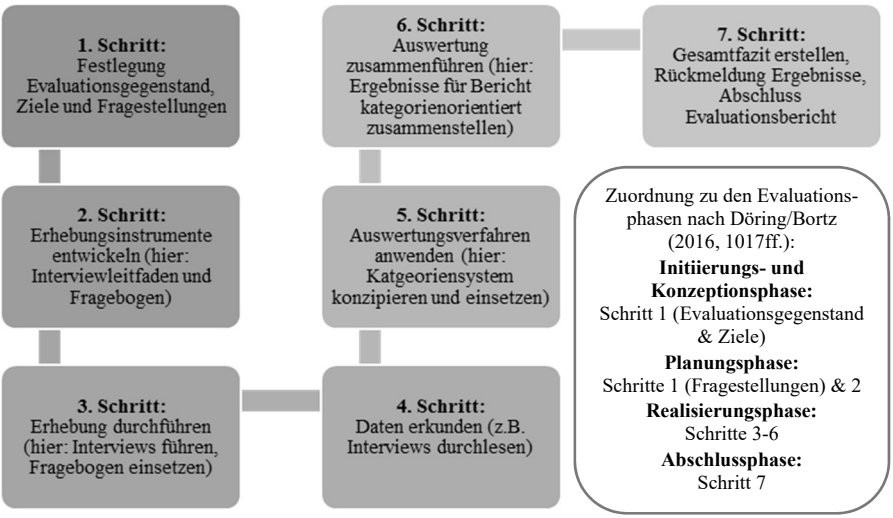

Abb 1: Beispiel für den Ablauf einer Evaluationsstudie (vgl. Kuckartz et al. 2008)

Ergänzend zu Abbildung 1 sollte noch festgehalten werden, dass eine Orientierung an Evaluationsstandards bereits in der Konzeptionsphase sinnvoll ist (vgl. Döring/Bortz 2016, 1020). Außerdem ist der Evaluationsbericht (Schritt 7, Abschlussphase) nicht mit der Umsetzung einer Abschlussarbeit (Bachelor-/ Master-/Dissertationsarbeit) gleichzusetzen, sondern richtet sich in erster Linie an die verschiedenen Akteursgruppen (hier z.B. die Studierenden) der durchgeführten Evaluation. Die „Offenlegung von Ergebnissen und Berichten“ (DeGEval 2016, 20) zählt zu den oben beschriebenen Standards hochwertiger Evaluationsstudien.

Ob eine Evaluationsstudie schließlich qualitativ, quantitativ oder mixed-methods orientiert ist, hängt vom Forschungsinteresse und den zentralen Fragestellungen ab. So raten Kuckartz et al. (2008, 11) zu einer qualitativen Evaluationsstudie, wenn „[…] eine größere Offenheit und eine Berücksichtigung der Perspektive der Beteiligten […]“ verstärkt im Forschungsinteresse liegt. Außerdem können so individuelle Sichtweisen rekonstruiert und unerwartete Einflussfaktoren eher aufgedeckt werden. Allerdings sollte gerade bei qualitativen Evaluationsstudien der Arbeitsaufwand genauestens vorab kalkuliert werden.

6. Fazit

Der vorliegende Beitrag hat u.a. gezeigt, dass der Nutzen einer Evaluationsstudie für die Gesellschaft von zentraler Bedeutung ist: Gerade bei der Verteilung von Ressourcen und der (Weiter-) Entwicklung von Konzepten, wie Lehrerfortbildungen, Hochschulseminaren, Unterrichtsmaterialien und Fördermethoden, können wissenschaftliche Evaluationsstudien einen Beitrag zu deutschdidaktischer Forschung leisten.

Für hochwertige Evaluationsforschung ist allerdings einiges zu berücksichtigen: Forschungsinteresse und zentrale Fragestellungen benötigen eine Passung zum Evaluationsgegenstand. Es sollte thematisiert werden, inwieweit die eigene Studie einzelne mögliche Leitfunktionen wissenschaftlicher Evaluation (vgl. Kapitel 2) umsetzt. Der Bezug zu theoretischen Ansätzen von Evaluationsforschung (vgl. Kapitel 3) ist begründet darzustellen. Neben den klassischen Gütekriterien (wie Objektivität, Reliabilität und Validität) spielt bei wissenschaftlichen Evaluationsstudien auch die Berücksichtigung von Evaluationsstandards eine wichtige Rolle (vgl. Kapitel 4). Da im Rahmen von Evaluationsforschung einiges bedacht werden muss, können bestehende Checklisten eine Hilfe sein (vgl. Klockgether 2015): So existieren beispielsweise Checklisten[5] für die Planungsphase von Evaluationsstudien oder abschließende Evaluationsberichte. Auch für die Anwendung der DeGEval-Evaluationsstandards steht online eine Checkliste zur Verfügung[6].

Obwohl die Bezeichnung Evaluation im Alltag inzwischen für die verschiedensten Formen an Bewertungsvorgängen verwendet wird, macht der vorliegende Beitrag deutlich, dass die Anforderungen wissenschaftlicher Evaluationsstudien hoch sind: So bedarf es Kenntnisse hinsichtlich möglicher Forschungsdesigns (z.B. quasi-experimentell) sowie verschiedener Erhebungs- und Auswertungsmethoden. Darüber hinaus muss der Forschende die beteiligten Akteure berücksichtigen und ihnen den Mehrwert aus der Evaluationsstudie zugänglich machen. Umso wichtiger ist es für die Planung eigener Evaluationsforschung, die zentrale Fragestellung nicht zu weit zu fassen und bei der Vielzahl an Evaluationsmöglichkeiten eine begründete Auswahl zu treffen.

Literatur

Bohl, Thorsten/Wacker, Albrecht (Hrsg.) (2016): Die Einführung der Gemeinschaftsschule in Baden-Württemberg. Abschlussbericht der wissenschaftlichen Begleitforschung. Münster: Waxmann.

Bühner, Markus (2011): Einführung in die Test- und Fragebogenkonstruktion. 3. Aufl. München: Pearson.

DeGEval – Gesellschaft für Evaluation e.v. (2016): Standards für Evaluation. Erste Revision 2016. Online unter: https://www.degeval.org/fileadmin/Publikationen/DeGEval-Standards_fuer_Evaluation.pdf (letzter Zugriff: 01.08.2018).

Döring, Nicole/Bortz, Jürgen (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5. Aufl. Berlin: Springer.

Gollwitzer, Mario/Jäger, Reinhold S. (2014): Evaluation kompakt. 2. Aufl. Weinheim: Beltz.

[5] Eine zusammenstellende Darstellung von Checklisten für Evaluationsstudien findet sich bei Klockgether 2015.

[6] Zu finden unter https://www.degeval.org/degeval-standards/download/

Grausam, Nina/Metz, Kerstin/Jäger, Sibylle/Maier, Uwe (2016): Diagnostik und Förderung von Schreibkompetenz in Gemeinschaftsschulen. Eine Interventionsstudie zur Prüfung von Effekten einer Lehrerfortbildung. In: Bohl, Thorsten/Wacker, Albrecht (Hrsg.): Die Einführung der Gemeinschaftsschule in Baden-Württemberg. Abschlussbericht der wissenschaftlichen Begleitforschung (WissGem). Münster: Waxmann, 115-134.

Kirkpatrick, Donald L./Kirkpatrick, James D. (2012): Evaluating training programs. The four levels. 3. Aufl. San Francisco: Berrett-Koehler.

Klockgether, Katharina (2015): Checklisten zur Planung und Steuerung von Evaluationen. In: Giel, Susanne/Klockgether, Katharina/Mäder, Susanne (Hrsg.): Evaluationspraxis. Professionalisierung – Ansätze – Methoden. Münster: Waxmann, 29-40.

Kuckartz, Udo/Dresing, Thorsten/Rädiker, Stefan/Stefer, Claus (2008): Qualitative Evaluation. Ein Einstieg in die Praxis. 2. Aufl. Wiesbaden: VS Verlag.

Lipowsky, Frank (2010): Empirische Befunde zur Wirksamkeit von Lehrerfortbildung. In: Müller, Florian H./Eichenberger, Astrid/Lüders, Manfred/Mayr, Johannes (Hrsg.): Lehrerinnen und Lehrer lernen. Konzepte und Befunde zur Lehrerfortbildung. Münster: Waxmann, 51-70.

Steineke, Ines (2013): Gütekriterien qualitativer Forschung. In: Flick, Uwe/von Kardoff, Ernst/Steineke. Ines (Hrsg.): Qualitative Forschung. Ein Handbuch. 10. Aufl. Hamburg: Rowohlt, 319-331.

Stockmann, Reinhard (2016): Entstehung und Grundlagen der Evaluation. In: Großmann, Daniel/Wolbring, Tobias: Evaluation von Studium und Lehre. Grundlagen, methodische Herausforderungen und Lösungsansätze. Wiesbaden: Springer VS, 27-56.

Stockmann, Reinhard/Meyer, Wolfgang (2014): Evaluation. Eine Einführung. 2. Aufl. Opladen: Budrich.

MAIK PHILIPP

Metaanalysen

1. Einleitung: Alles meta, oder was?

Spätestens seit der Veröffentlichung von John Hatties „Visible Learning“ im Jahr 2009 haben Metaanalysen eine erhöhte Aufmerksamkeit im Bildungssektor erfahren – die ersten metaanalytischen Ansätze bestehen freilich bereits seit Mitte der 1970er Jahre (vgl. Glass 1976). Von Anfang an ist es bei Metaanalysen erklärtes Ziel, einen kondensierten und konzentrierten quantitativen Forschungsüberblick zu geben. Dieser hohe und hehre Anspruch bei dieser Art von Zusammenfassungen des Forschungsstandes bringt es mit sich, dass Personen, die Metaanalysen durchführen, neben hohem statistischen Wissen auch eine entsprechende inhaltliche Expertise aufweisen sollten. Dies ergibt sich aus der Notwendigkeit, bei den nahezu unvermeidlichen inhaltlichen Zweifelsfällen intersubjektiv überprüfbare und vor allem klar begründete Entscheidungen im Forschungsprozess vornehmen zu können. In aller Regel sind entsprechend Metaanalysen deshalb nicht nur typisch für anspruchsvolle Qualifikationsarbeiten, sondern auch oftmals das Ergebnis von Teams aus Expertinnen und Experten.

Auch in der Deutschdidaktik gibt es inzwischen erste Versuche, fachdidaktische Fragen mittels Metaanalysen zu beantworten (vgl. exemplarisch Funke 2014 zur Effektivität des Förderansatzes *Lesen durch Schreiben* beim Schriftspracherwerb und Funke 2018 zu Effekten des Grammatikunterrichts), diese fallen aber bislang weder in- noch extensiv aus. Das ist insofern bedauerlich, als es – das beweist die Existenz dieses Bandes – deutliche Tendenzen gibt, die Deutschdidaktik hinsichtlich ihrer Methodik zu stärken (vgl. Groeben/Hurrelmann 2006), wobei bislang die qualitative Forschung erheblich mehr und intensivere Zuwendung erhalten hat (vgl. Boelmann 2016).

Dieses Kapitel möchte einen konzeptionellen Überblick über Metaanalysen geben und will in die Thematik knapp einführen. In diesem Text kann allein schon aus Platzgründen selbstverständlich nur ein selektiver Überblick gegeben werden. Für die tatsächliche Durchführung einer Metaanalyse nebst statistischem Handwerkszeug braucht es mehr Informationen, und hier sind insbesondere englische Fachbücher eine gute Wahl. Angeführt seien fünf, die sich international als einflussreich erwiesen haben und aktuell sind, nämlich:

- „Introduction to Meta-Analysis“ (Borenstein/Hedges/Higgins/Rothstein 2009),
- „Practical Meta-Analysis“ (Lipsey/Wilson 2001),
- „Research Synthesis and Meta-Analysis“ (Cooper 2017),
- „The Essential Guide to Effect Sizes“ (Ellis 2010) sowie
- „The Handbook of Research Synthesis and Meta-Analysis“ (Cooper u.a. 2009).

Diese Bände richten sich dezidiert an Forscherinnen und Forscher, die sich mit dem Gedanken tragen, eine eigene Metaanalyse durchzuführen, und versorgen sie mit diversen verständlich dargebotenen Beispielen und Informationen. Dabei setzt jedes der genannten Bücher etwas andere Akzente, sodass es sich empfiehlt, für eigene metaanalytische Studien tatsächlich alle Bände einmal konsultiert zu haben. Dies gilt insbesondere für die mathematisch-statistischen Details, ohne die eine Metaanalyse nicht auskommt. Diese Details spart dieser Beitrag aus Platzgründen gezielt aus. Für eine erste konzeptionelle Beschäftigung bietet sich die Einführung von Rost (2007, Kapitel 2.2) an, welche sich der Thematik Metaanalyse aus einer kritischen Perspektive annähert und auf konzeptionelle Kritikpunkte des Verfahrens eingeht.

Der vorliegende Beitrag ist im Weiteren in drei Kapitel unterteilt. Der erste Teil des Beitrags definiert zunächst Wesen und Anlass von Metaanalysen (Kapitel 2). Danach werden allgemeine Qualitätsindikatoren hinsichtlich der Transparenz von Metaanalysen vorgestellt und am Beispiel einer publizierten Metaanalyse zur Schreibförderung konkretisiert (Kapitel 3); dabei steht eine tabellarische Übersicht mit Qualitätskriterien im Zentrum. Zudem werden Implikationen für die Durchführung von Metaanalysen abgeleitet. Nach diesem Hauptteil des Kapitels wird im kurzen Fazit (Kapitel 4) die Essenz gebündelt dargestellt.

2. Was ist eine Metaanalyse – und wozu führt man sie durch?

Wenn *Metaanalyse* als Begrifflichkeit verwendet wird, dann ist deren Nutzung nicht immer einheitlich. Die Definition fällt leichter, wenn man „Metaanalyse“ zu einem ähnlichen, allerdings inhaltlich weiter gefassten Konzept – dem „systematischen Überblick“ – in Beziehung setzt:

- Ein *systematischer Überblick* ist ein Überblick zu einer klar formulierten Frage, der systematische und explizite Methoden dafür nutzt, relevante Forschung zu identifizieren, auszuwählen und kritisch zu beurteilen und Daten aus diesen Studien zu sammeln und zu analysieren, die im Überblick verwendet werden. Statistische Methoden (Metaanalyse) können oder können auch nicht verwendet werden, um die Ergebnisse der inkludierten Studien zu analysieren und zusammenzufassen.
- *Metaanalyse* bezieht sich auf die Verwendung statistischer Verfahren in einem systematischen Überblick, um die Ergebnisse der inkludierten Studien zusammenzufassen. (Moher u.a. 2009, 264 [Herv. und Spiegelstriche ergänzt])

Beide Konzepte eint der systematische Zugang zu Primärstudien, der im Falle der Metaanalyse mit einer zusätzlichen quantitativen Auswertung erfolgt. Insofern bildet ein systematischer Überblick einen ersten Schritt hin zu einer Metaanalyse.

Der Anlass für Metaanalysen ist von Anfang dieses Verfahrens an neben einer konkreten Fragestellung die durch die mehr oder minder rege Forschungsaktivität hervorgerufene Uneindeutigkeit in der wissenschaftlichen Literatur (vgl. Glass 1976). Beispielsweise könnte man die konkrete Forschungsfrage haben, die Effektivität verschiedener Schreibförderansätze für Schülerinnen und Schüler im Sekundarschulalter zu bestimmen. Wer dann hierfür breit und systematisch recherchiert, wird in aller Regel bald feststellen, dass es inhaltlich unterschiedliche Förderansätze bei unterschiedlichen Personengruppen (Alter, Geschlecht, etwaigen Lernschwierigkeiten etc.) mit unterschiedlich erfassten abhängigen Variablen gibt – und dass die Ergebnisse mitunter (erheblich) divergieren. Aus Sicht einer forschenden Person schildert Paul Ellis (2010, XVI) das Dilemma und einen Lösungsvorschlag:

> Wir sichten die Literatur zu einem Thema, sehen, dass es keinen Konsens gibt, und nutzen dies als Begründung, noch eine weitere Studie durchzuführen. Wir erzielen unsere eigene kleine Schlussfolgerung, und diese wird auf die Säule der Schlussfolgerungen hinzugefügt, die dann von irgendjemandem nach uns gesichtet wird. […] Ein besserer Ansatz hingegen ist, all die kleinen Schlussfolgerungen beiseite zu lassen und auf die tatsächlichen Effekte zu fokussieren, die in früheren Studien berichtet wurden. Das Zusammenbringen von unabhängigen Effektstärken wird Metaanalyse genannt. Gut gemacht können Metaanalysen präzise Schlussfolgerungen hinsichtlich der Richtung und Ausprägung eines Effekts liefern, selbst wenn die Daten aus unähnlichen Studien mit konfligierenden Daten stammen.

Damit lässt sich zusammenfassend sagen, dass Metaanalysen *quantifizierende Zusammenfassungen der verfügbaren empirischen (durchaus deutschdidaktischen bzw. deutschdidaktisch relevanten) Forschungsresultate* sind. Sie haben den Anspruch, unter der Wahrung von Qualitätsmerkmalen präzise Fragestellungen eindeutig anhand transparenter Prinzipien auf der Basis vorhandener Primärstudien durch deren Re-Analysen quantifizierend zu beantworten. Um dies zu erreichen, unterliegen Metaanalysen hohen wissenschaftlichen Qualitätsansprüchen, die als Kriterienlisten zugänglich sind, sich als Gradmesser der Qualität einzelner Metaanalysen anbieten und damit dieses Forschungsfeld bereichern. Einer solchen Kriterienliste widmet sich das folgende Kapitel, in dem die Liste auf eine konkrete Metaanalyse hin Anwendung findet.

3. Qualitätsansprüche an Metaanalysen in Form von Kriterien – ein selektiver Einblick am Beispiel der Schreibinterventionsforschung

Das Stichwort der Systematik bei Metaanalysen führt zur Frage, was denn eine gute, systematische Metaanalyse auszeichnet. Dieser wichtigen Frage ist größere Aufmerksamkeit zuteil geworden: Unter dem Akronym *PRISMA* (*P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses) hat eine große Gruppe von Medizinern – sicher ist es kein Zufall, dass ausgerechnet diese Disziplin tätig geworden ist – konsensuell Qualitätsansprüche für Metaanalysen in Form einer im Internet frei verfügbaren Checkliste mit 27, zum Teil nicht immer trennscharfen Einträgen zusammengestellt. Diese Checkliste soll einerseits einer transparenten und nachvollziehbaren Ergebnisdarstellung dienen (vgl. Moher u.a. 2009; www.prisma-statement.org). Andererseits strukturieren solche Hinweise auch den Prozess und den Inhalt einer Metaanalyse von der Entwicklung einer Fragestellung bis hin zur Präsentation der Ergebnisse. Insofern bilden diese Qualitätsindikatoren eine wichtige Ressource für die Zwecke dieses Kapitels (vgl. die auf die Belange von Metaanalysen in bildungswissenschaftlichen Zusammenhängen spezifizierte Kodier-/Checkliste bei Ahn/Ames/Myers 2012, 465-469).

Ergänzt werden die Einträge der *PRISMA*-Checkliste zur Illustration in diesem Kapitel mit Daten aus einer einflussreichen, in vielerlei Hinsicht vorbildlichen und zum Zeitpunkt ihres Erscheinens besonders umfangreichen Metaanalyse, die Steve Graham und Dolores Perin (2007) vorlegt haben. Das Forschungsteam beantwortete aus einer pädagogisch-psychologischen Perspektive die folgende *Forschungsfrage*: Welche instruktionalen Praktiken verbessern die Qualität des Schreibens von Adoleszentinnen und Adoleszenten? Grahams und Perins Metaanalyse ist nicht nur prominent erschienen und breit rezipiert, sie kam für die Belange dieses Kapitels auch deshalb infrage, weil sie auf einer für die Schreibforschung ungewöhnlich breiten Datenbasis Erkenntnisse bündelt. Zudem ist sie für die Belange der Deutschdidaktik unmittelbar relevant, da sie sich mit Fragen des Kompetenzoutputs durch konkrete Fördermaßnahmen befasst. Außerdem ist sie insofern nachahmenswert, weil in der Metaanalyse ein aus methodischer Sicht sehr wichtiger und für diesen Band erwähnenswerter Katalog von Qualitätsindikatoren für jede einzelne Interventionsstudie systematisch erfasst und kodiert wurde (vgl. dazu Graham/Perin 2017, 452).

Durch den Griff zu der kurz vor der Veröffentlichung der *PRISMA*-Checkliste publizierten Metaanalyse von Graham und Perin soll dieses Kapitel anschaulicher werden und zugleich durch die Anbindung an die *PRISMA*-Checkliste für die hohen Anforderungen an Forschende sensibilisieren – gleichwohl lohnt sich die genaue Lektüre der *PRISMA*-Checkliste samt flankierender und vertiefender Begleitdokumente (vgl. Liberati u.a. 2009). Der Überblick über 18 besonders wichtige Einträge aus der Mitte der *PRISMA*-Checkliste als auch über die Daten aus Grahams und Perins Metaanalyse sind in Tab. 1 zusammenfassend dargestellt. Eine Bemerkung erscheint noch in Hinblick auf die Informationsdichte aus der

Tabelle nötig: Die Angaben in der Tabelle 1 sind thematisch gebündelt, stark aggregiert und lassen sich für eigene metaanalytische Forschungsansinnen als Anforderungskatalog hinsichtlich nötiger inhaltlicher Punkte übersetzen:

- *Methodische Aspekte* enthalten die Listeneinträge Nr. 6-16. Dies reicht von einer transparenten Auswahl der Anforderungen an Primärstudien (Nr. 6 und 9 in der Liste) und der replizierbar beschriebenen Recherche nach solchen Studien (Nr. 7 und 8) hin zu der Datenextraktion aus den einbezogen Primärstudien (Nr. 10 und 11). Neben der Datengewinnung geht es bei der Methodik um weitere analyserelevante Punkte wie die klare Bezeichnung der abhängigen Variablen (Nr. 13 – tlw. mit expliziter Nennung der Formeln für die Datenberechnung), des Umgangs mit Daten (Nr. 14), dem Umgang mit Datenverzerrungen (Nr. 12 und 15) und etwaig durchgeführter zusätzlicher Datenanalysen (Nr.16).
- Aspekte der *Ergebnisse der Metaanalysen* werden in den Listeneinträgen 17-23 thematisiert und haben zum Teil ihre Entsprechung in den Listen zur Methodik. Es geht neben einer Darlegung der Befunde der Re-Analysen sowohl für die Einzelstudien (Nr. 19 und 20) als auch die allgemeinen studienübergreifenden Analysen (Nr. 21 und 22) und die Resultate aus Zusatzanalysen (Nr. 23).

Tab. 1: *PRISMA*-Checkliste von ausgewählten Bestandteilen für einen systematischen Überblick/eine Metaanalyse mit Beispieldaten aus einer Metaanalyse aus der Schreibinterventionsforschung[1]

Nr.	Bestandteil aus der *PRISMA*-Checkliste	Beispielangaben aus der Metaanalyse von Graham und Perin
		Methoden-Teil
6	*Auswahlkriterien:* Studienmerkmale (z.B. abhängige Variable, berücksichtigte Jahre, Sprachen, Publikationsstatus) als Auswahlkriterien berichten und begründen	• Es werden *sieben explizite Auswahlkriterien* verwendet: 1) Sample nur aus Klassenstufe 4-12); 2) Schülerschaft aus regulären Schulen (keine sonderpädagogisch betreuten Heranwachsenden); 3) Schreibqualität als abhängige Variable; 4) ausreichende Reliabilität der Schreibqualität als abhängige Variable; 5) (quasi-)experimentelles Design; 6) ausreichende Daten zur Berechnung von (gewichteten) Effektstärken und deren Homogenität; 7) Veröffentlichung der Studien auch jenseits der Fachzeitschriften (d.h. Publikationen sowohl mit als auch ohne Peer-Review) • *Ein implizites Auswahlkriterium* angewendet: Es wurden nur englischsprachige Publikationen verwendet.

[1] Auswahl von *PRISMA*-Inhalten mit einigen Anpassungen sowie Angaben aus Moher et al. 2009, 266 und Graham/Perin 2007; die Zahlen in der Spalte Nr. wurden aus Gründen der Nachvollziehbarkeit direkt übernommen aus Moher et al. 2009.

Nr.	Bestandteil aus der *PRISMA*-Checkliste	Beispielangaben aus der Metaanalyse von Graham und Perin
7	*Informationsquellen:* sämtliche Informationsquellen (z.B. Datenbanken, Suchmaschinen, Kontakt zu anderen Forschenden, um zusätzliche Studien zu lokalisieren) und Datum der letzten Suche beschreiben	• *Datenquellen:* vorherige ähnliche Metaanalysen, Fachzeitschriften (nicht spezifiziert), Abschlussarbeiten (nicht spezifiziert), Dissertationen (nicht spezifiziert), Konferenzbeiträge (nicht spezifiziert), Bücher (nicht spezifiziert), Datenbanken (nur teilweise spezifiziert) • *Datum der letzten Suche:* Mai 2005 (nur für Datenbanken berichtet)
8	*Suche:* die gesamte elektronische Suche innerhalb mindestens einer Datenbank mit allen Limitierungen so beschreiben, dass die Suche wiederholt werden kann	Suchtermini bei der elektronischen Suche in Datenbanken: Suchbegriffe „writing OR composition“ in Verbindung mit 45 weiteren, in der Publikation explizit genannten Suchbegriffen
9	*Studienauswahl:* Prozess der Studienauswahl beschreiben (z.B. Screening, Auswahl, Berücksichtigung in der Metaanalyse)	Unklar. Als Rechercheergebnis werden 582 Dokumente benannt, und die sukzessive Eliminierung der Studien durch den Einsatz von sechs Kriterien gemäß der Beschreibung in der Publikation summiert sich auf 91 Prozent Ausschlüsse auf, sodass von den ursprünglich 582 Dokumenten nur 52 übrig geblieben wären. Im Artikel werden 123 Dokumente als passend angegeben, 117 Dokumente sind im Literaturverzeichnis als Primärstudien aufgeführt.
10	*Datengewinnungsprozess:* Datenextraktion aus den Studien oder von anderen Forschenden beschreiben	• Zuweisung der einzelnen Förderansätze zu einer definierten Kategorie von Förderansätzen (iterativer Vorgang – gerade mit Blick auf die zunächst uneindeutigen Fördermaßnahmen, s.u., Nr. 12 und 20) • Kodierung durch Erstautor, 15 Prozent der Studien zufällig ausgewählt und von einem Doktoranden parallel kodiert • Bestimmung des standardisierten Mittelwertsunterschieds bei der Posttestung hinsichtlich der holistisch eingeschätzten Textqualität als abhängiger Variable • Wurde die Textqualität ausschließlich analytisch eingeschätzt, erfolgte eine Mittelwertbildung über alle eingesetzten analytischen Skalen. • Bei alternativen Treatments in den Experimentalgruppen: Bestimmung von Effektstärken für jeden einzelnen Schreibförderansatz.
11	*Datenspezifikation:* alle Daten, nach denen gesucht wurde, aufführen und defi-	• S.u., Nr. 12 und 20 • Explikation der Subkodierung bei zwei einzelnen Förderansätzen (s.u., Nr. 18)

Nr.	Bestandteil aus der *PRISMA*-Checkliste	Beispielangaben aus der Metaanalyse von Graham und Perin
	nieren und jegliche Annahmen und Simplifikationen beschreiben	
12	*Intra-Verzerrungsrisiko (Bias):* Methoden zum Umgang mit einem etwaigen Verzerrungsrisiko (sei es bei Einzelstudien, sei es bei deren Ergebnissen) und den Umgang damit in der Metaanalyse beschreiben	• Kodierung von *neun Qualitätsindikatoren pro Studie:* 1) Art der Zuteilung der Teilnehmenden; 2) Berücksichtigung von Effekten einzelner Lehrpersonen; 3) geringe Samplemortalität; 4) Vergleichbarkeit der Gruppen vor der Intervention; 5) Vorhandensein von Boden-/Deckeneffekten nach dem Training; 6) Training der instruierenden Personen; 7) Vorhandensein von alternativen Fördermaßnahmen in der Kontrollgruppe; 8) Reduktion des „Hawthorne-Effekts“; 9) Treatment-Integrität • Kodierung der *Randomisierung* von Kontroll- und Experimentalgruppe • *Doppelkodierung* von 15 Prozent aller Studien (randomisierte Auswahl, s. Nr. 10) • Korrektur der Verzerrung durch kleine Samples ($N < 20$) in drei Fällen
13	*Maße:* wichtigste Maße (z.B. Mittelwertdifferenzen) benennen	Effektstärken: gewichtete und ungewichtete Mittelwertdifferenzen bei der Textqualität (als abhängige Variable bzw. als Indikator für den Fördererfolg), geteilt durch gepoolte Standardabweichung
14	*Umgang mit Daten:* Methoden im Umgang mit Daten (z.B. Maße zur internen Konsistenz wie I^2) beschreiben	• Berechnung von insgesamt 155 Effektstärken (eine für jeden einzelnen Förderansatz) für die abhängige Variable • Metaanalyse nur für Förderansätze mit mindestens vier Effektstärken (ohne Begründung) • Berechnung von Homogenitätsmaßen (Q-Test) und Konfidenzintervallen innerhalb der Förderansätze mit mindestens vier Einzelstudien für gewichtete Effektstärken • Bildung von dichotomen Kategorien von fünf Moderatorvariablen (s. Nr. 15) für Moderatoranalysen (Q-Tests)
15	*Inter-Verzerrungsrisiko (Bias):* jegliche Bias-Maße, die den Gesamteffekt betreffen könnten (z.B. Publikationsstatus oder selektive Ergebnisdarstellung in den Studien)	• Berechnung der prozentualen Anteile von Studien, die jedem der neun einzelnen *Qualitätsindikatoren* (s. Nr. 12) genügen, sowie Berechnung von Mittelwert und Standardabweichung eines Summenwertes (alle neun Kriterien) über alle Studien innerhalb einer der 15 aufgeführten Schreibförderkategorien

Nr.	Bestandteil aus der *PRISMA*-Checkliste	Beispielangaben aus der Metaanalyse von Graham und Perin
		• Prüfung von *fünf Moderatoreffekten* bei statistisch signifikanter Homogenitätsanalyse bei jenen Förderansätzen mit mindestens 18 Effektstärken hinsichtlich a) Publikationstyp (Fachzeitschrift [mit Peer-Review] vs. andere), b) Klassenstufe (4-6 vs. 7-12), c) Genre der geschriebenen Texte (Sach- vs. narrative Texte, d) Art der Schülerschaft (schwach Schreibende vs. volle Bandbreite), e) Zuweisung zu Experimental-/Kontrollgruppe (zufällig vs. nicht zufällig)
16	*Zusatzanalysen:* Methoden etwaiger, möglichst vorab festgelegter zusätzlicher Analysen (etwa für Subgruppen oder Metaregressionen)	s. Nr. 15 sowie Nr. 18 (Subkodierungen bei Einzelstudien)
Ergebnis-Teil		
17	*Studienauswahl:* Anzahl der gescreenten, auf Kriterien analysierten und berücksichtigen Studien beschreiben und begründen, idealerweise mit einem Flussdiagramm	s. Nr. 9. Kein Flussdiagramm.
18	*Studienmerkmale:* für jede Studie die Merkmale benennen, welche aus der Studie extrahiert wurden, und bibliografische Angaben aufführen	• Kodierung für *alle Studien:* 1) Klassenstufe, 2) Art der Schülerschaft, 3) Anzahl der Probandinnen und Probanden, 4) Textsorte, 5) Publikationstyp, 6) Kurzbeschreibung von Treatment- und Kontrollgruppeninstruktion, 7) neun Qualitätskriterien (s. Nr. 12), 8) Randomisierung bei der Zuweisung zu Kontroll- bzw. Experimentalgruppe • Kodierung *einzelner Studien:* bei zwei besonders häufig untersuchten Förderansätzen dichotome Subkodierungen (1) Prozessansatz [mit oder ohne professionelle Weiterentwicklung der Lehrpersonen] und 2) Strategievermittlung [gemäß Programm „Self-Regulated Strategy Development“ oder anderes Programm])
19	*Intra-Verzerrungsrisiko (Bias):* für jede Studie und – wenn verfügbar – für jede abhängige Variable/jedes Ergebnis berichten (s. Nr. 12)	• Summenwert der neun Qualitätsindikatoren (s. Nr. 12) für jene Förderansätze angeführt, die mindestens vier Effektstärken aufwiesen • kein Summenwert für Förderansätze mit maximal drei Effektstärken angegeben
20	*Ergebnisse der Einzelstudien:* für jede abhängige Variable/jedes Ergebnis eine	• für jeden einzelnen berücksichtigen Förderansatz in zwei Tabellen – eine für die Förderansätze mit

Nr.	Bestandteil aus der *PRISMA*-Checkliste	Beispielangaben aus der Metaanalyse von Graham und Perin
	einzelne Zusammenfassung der Interventionsgruppen und Schätzeffekte mitsamt Konfidenzintervalle mit einem „Forest plot“ präsentieren	vier und mehr Effektstärken zu einem Förderansatz, eine andere für solche mit höchstens drei Effektstärken im Anhang (insg. 26 Förderansätze) – mit Angaben a) zum Kurzbeleg des Dokuments, b) der Klassenstufe(n), c) Art der Schülerschaft, d) Samplegröße, e) Genre der geschriebenen Texte, f) Art der Förderung in der Einzelstudie, g) Publikationstyp und h) Effektstärke sowie – nur für Förderansätze mit mehr als drei Effektstärken – i) Summenwert zur Studienqualität (s. Nr. 12) und j) randomisierte Zuweisung zu Experimental- und Kontrollgruppe • kein „Forest plot“, kein Konfidenzintervall angegeben
21	*Zusammenfassung der Metaanalysen:* für jede einzelne Metaanalyse Ergebnisse (inklusive Konfidenzintervallen und Maßen zur internen Konsistenz) darstellen	• tabellarische Angaben (in Tabelle 5) zu 11 Gruppen von Schreibinterventionsansätzen (und drei Moderatoranalysen) mit a) Anzahl der Effektstärken, b) Mittelwert und c) Standardabweichung für ungewichtete Effektstärken, gewichtete Effektstärken als d) Mittelwert und e) Median, Grenzen des f) oberen und g) unteren Konfidenzintervalls sowie h) Angaben zur Homogenität der gewichteten Effektstärken (Q-Test) • keine Berechnung der Effektstärken bei vier Förderansätzen (Textstrukturen, prozedurale Unterstützung, Feedback und zusätzlicher Schreibzeit) aus inhaltlichen und statistischen Gründen, wodurch 20 Effektstärken nicht genutzt werden konnten
22	*Inter-Verzerrungsrisiko (Bias):* Ergebnisse von Analysen zum Bias präsentieren (s. Nr. 15)	s. Nr. 15 (und 16) – konkret: • Korrelationsanalysen und ANOVAs bei den Qualitätsindikatoren (Summenwert) mit Blick auf die Effektstärken • Korrelationsanalysen mit Alter der Publikation und Qualitätsindikatoren (Summenwert) • ANOVA (Varianzanalyse) zum Zusammenhang zwischen Publikationstyp und Effektstärken (Summenwert und Einzelwerte)
23	*Zusatzanalysen:* Ergebnisse etwaiger Zusatzanalysen wie Subgruppenanalysen oder Metaregressionen berichten (s. Nr. 16)	s. Nr. 15 und 16 • Zusätzliche Moderatoranalysen bei zwei häufig untersuchten Förderansätzen mit Subkodierungen (s. Nr. 18)

Wie aus der Tab. 1 hervorgeht, sind die Bestandteile der *PRISMA*-Checkliste anforderungsreich und zwingen Forschende, die eine Metaanalyse durchführen, in

der Ergebnisdarstellung und auch in der vorgängigen operativen Arbeit zu sehr vielen Explikationen und hoher Transparenz. Diese Explikationen hängen mit einem zentralen Gütekriterium jedweder Forschung zusammen: der intersubjektiven Überprüfbarkeit, Nachvollziehbarkeit und Transparenz von methodischem Vorgehen und daraus resultierenden Befunden. Dabei liegt in der Reihung der Einträge der Checkliste zugleich auch eine Logik, die verhindern soll, dass „meta-meta Unsinn" (Rost 2007, 39) entsteht, indem Studien in die Metaanalyse geraten und berücksichtigt werden, die dort nicht hätten landen dürfen. Dabei ist nicht nur ein regelgeleitetes Vorgehen gemäß *PRISMA*-Liste vonnöten, sondern vor allem eine hohe fachliche Expertise, welche die Einschätzung der Eignung einzelner potenzieller Studien betrifft bzw. ein Expertenurteil überhaupt erst ermöglicht.

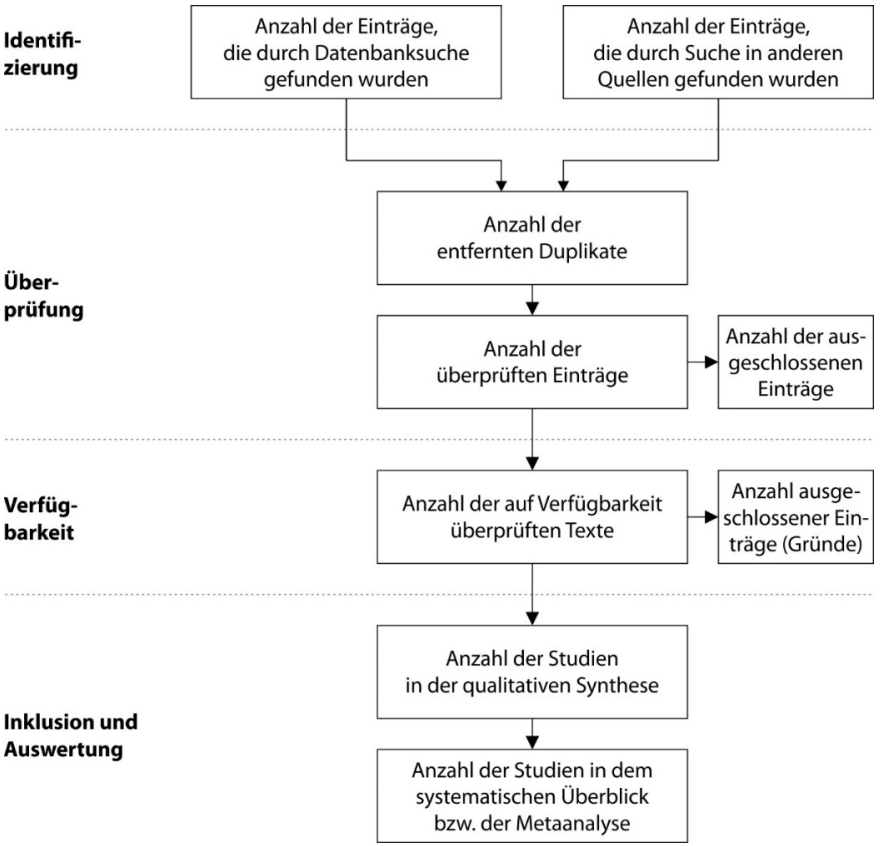

Abb. 1: Flussdiagramm zum Vorgehen bei der Auswahl von Studien für eine Metaanalyse mit verschiedenen Phasen[2]

[2] Darstellung gemäß Moher et al. 2009, 267, mit leichten Modifikationen.

Damit beim Prozess der Auswahl wirklich nur geeignete Studien metaanalysiert werden, empfiehlt das *PRISMA*-Team die Darlegung eines strukturierten Prozesses (in Schritt 17). Das dort verlangte Flussdiagramm ist in Abb. 1 dargestellt. Der darin abgebildete Reduktionsprozess sieht zu jedem Schritt vor, dass die genaue Anzahl der Studien explizit benannt wird, welche gefunden, überprüft, ausgeschlossen und zu guter Letzt tatsächlich verwendet wurden. Dieser Schritt ist essenziell, wenn es darum geht, die Güte und Vollständigkeit des Korpus an Primärstudien zu beurteilen, auf dem eine Metaanalyse basiert. Leider weist ausgerechnet bei diesem für die Transparenz und Replizierbarkeit so wichtigen Schritt aus der PRISMA-Checkliste die Metaanalyse von Graham und Perin (2007) Schwachstellen auf, die möglicherweise die Ergebnisse der gesamten Metaanalyse betreffen, weil es um die finale Studienauswahl geht.

So berichten Graham und Perin (2007, 448) nicht davon, welches Ergebnis ihre Recherchen im Schritt *Identifizierung* hatten. Wie viele Dokumente aus Datenbanken oder aus anderen Quellen stammten, bleibt ebenso ausgespart wie die Anzahl von Duplikaten, die gescreenten und im Anschluss daran ausgeschlossenen Dokumente im zweiten Schritten *Überprüfung.* Erst im nächsten, dritten Schritt schien das Forschungsteam insgesamt 582 Dokumente als Korpus *verfügbarer Studien* gehabt zu haben. In der Metaanalyse werden die ausgeschlossenen Studien in Form von Anteilen benannt:

- *Kein (quasi-)experimentelles Design in der Domäne Schreiben* (45 Prozent der Dokumente; Kriterium 5 aus PRISMA-Element 6): 15 Prozent waren Studien mit Einzelpersonen, 11 Prozent hatten keine Kontrollgruppe, in 9 Prozent waren die Probandinnen und Probanden die eigene Kontrollgruppe, 4 Prozent waren Überblicksbeiträge, 3 Prozent waren qualitativer Art, 2 Prozent nutzten ein deskriptives Design, und 1 Prozent hatte kein Bezug zum Schreiben.
- *Textqualität war nicht abhängige Variable* (22 Prozent; Kriterium 3).
- *Probandinnen und Probanden waren zu alt oder zu jung* (11 Prozent, Kriterium 1).
- *Keine Berechnung von (un)gewichteter Effektstärke möglich* (9 Prozent; Kriterium 6).
- *Keine ausreichende Reliabilität bei der Textqualität* (2 Prozent; Kriterium 4).
- *Schülerschaft stammte nicht aus regulären Schulen* (2 Prozent; Kriterium 2).

Rechnerisch ergeben die sechs Ausschlusskriterien aus dem PRISMA-Element 6 – Kriterium 7 (Veröffentlichung der Studien auch jenseits der Fachzeitschriften) ist im Kern ein *Ein*schlusskriterium – in der Summe insgesamt 91 Prozent auszuschließender Dokumente, also insgesamt 530 Dokumente. Nach dieser Berechnung wären 52 Dokumente übriggeblieben. Faktisch geben Graham und Perin (2007, 448) aber für den finalen Schritt *Inklusion und Auswertung* an, dass 123 Dokumente (im Literaturverzeichnis: 117 angegeben) „passend für die Inklusion" waren – also deutlich mehr als doppelt so viele wie nach dem berichteten Einsatz der Ausschlusskriterien. Die Anzahl der berücksichtigen Studien ist ebenso nur implizit nachvollziehbar, weil a) zum Teil Metaanalysen gar nicht durchgeführt

wurden (siehe PRISMA-Element 21), und b) aus den Studien insgesamt 155 Effektstärken extrahiert wurden. Genau genommen muss man die Metaanalyse systematisch in zwei Tabellen, im Fließtext und im Literaturverzeichnis aufwändig selbstständig durchprüfen, ehe man die genaue Zahl der Studien (nicht Dokumente und nicht Effektstärken) benennen kann. Vor diesem Hintergrund mag es wenig verwundern, dass die PRISMA-Gruppe inzwischen die Anforderungen an die Darstellung der Studienauswahl (und für andere PRISMA-Bereiche ebenfalls) nochmals verschärft hat (vgl. Moher u.a. 2015).

Dieser kurze, aber intensive Gang durch eine der populärsten Metaanalysen aus der Schreibinterventionsforschung verdeutlicht mindestens dreierlei. Erstens liegt mit der PRISMA-Checkliste ein Werkzeug vor, dass sowohl die Lektüre als auch die Durchführung und Darstellung einer Metaanalyse erleichtert. Das ist deshalb so wichtig, weil es bei Metaanalysen qua ihrem Anspruch und ihrer Funktion darum geht, zu quantifizierenden, einzelstudienübergreifenden Aussagen zu gelangen. Die empirische Bildungsforschung, zu der die deutschdidaktische Forschung zweifelsohne zählt, kann bei den Metaanalysen viel von den Qualitätsstandards lernen, die in anderen Wissenschaftsdisziplinen wie der Medizin bereits existieren.

Zweitens eröffnet die Anwendung der PRISMA-Checkliste klärende Perspektiven auf die Qualitäten und die optimierbaren Merkmale bisheriger Metaanalysen. Im Falle der Metaanalyse von Graham und Perin (2007) überwiegen im Licht der Einträge in der PRISMA-Checkliste eindeutig die handwerklichen Qualitäten. Aus Platzgründen und wegen einer Fokussierung auf die deskriptive Darstellung von PRISMA-Kriterien und ihrer Applikation auf die exemplarische Metaanalyse konnten diese in diesem Kapitel nicht umfassend gewürdigt werden. Wenn also berechtigte Kritik an der Darstellung eines zentralen Kriteriums, nämlich der Studienauswahl, erfolgt, geschieht dies zugegebenermaßen auf hohem Niveau.

Drittens illustriert der PRISMA-Katalog bzw. der in diesem Kapitel behandelte Ausschnitt daraus, dass und wie viel Arbeit eine Metaanalyse macht. Dieser Aufwand, der jede Phase der wissenschaftlichen Arbeit betrifft, unterstreicht auch den wissenschaftlichen Zugewinn von metaanalytisch gewonnenen Aussagen als retrospektive Bestandsaufnahmen aus zum Teil heterogenen Studien. Gerade wegen dieser Unübersichtlichkeit in den Primärstudien ist die intersubjektive Nachvollziehbarkeit des methodischen Vorgehens gewissermaßen die Nagelprobe für die Belastbarkeit von Metaanalysen – auch und gerade in der Bildungsforschung.

4. Fazit

In diesem Kapitel ging es um die quantitative nachträgliche Auswertung von kriterienbasiert ausgewählten Studien – ein Vorgehen, das als „Metaanalyse“ bezeichnet wird. Dieser Forschungszweig gewinnt in der internationalen empirischen Bildungsforschung an Bedeutung, und es ist inzwischen eine regelrechte ‚Industrie‘ entstanden, die Metaanalysen produziert.

Metaanalysen als sekundäranalytische Zugangsweise unterliegen wie alle anderen Forschungszugänge natürlich allgemeinen Gütekriterien, in diesem Falle jenen zur quantitativen Forschung (siehe die Beiträge von Schmidt in diesem Band). Hinzu kommen Kriterien, die spezifisch für Metaanalysen gelten. Aus dem Kontext der Medizin stammt ein PRISMA genannter Katalog von Qualitätsmerkmalen, der sich für die Lektüre und Durchführung von Metaanalysen verwenden lässt, um jenseits der konkreten spezifischen Ergebnisse einer Metaanalyse deren wissenschaftliche Qualität besser einschätzen zu können.

Handwerklich gut gemachte Metaanalysen helfen in einer stetig komplexer werdenden Forschungslandschaft gleichsam wie Richtungsweiser bei der Orientierung. Damit sie diese wichtige Funktion für einen Teil der Forschung erfüllen können, bedarf es seitens der durchführenden Personen eines hohen methodischen Wissens auch einer entsprechend profunden fachlichen Expertise – zumal wenn die Qualitätsansprüche an Metaanalysen wie auch an die Forschung allgemein steigen. Für die deutschdidaktische Forschung mit ihren genuin eigenen Fragestellungen erscheint das Potenzial von Metaanalysen insgesamt hoch – gleichwohl ist die deutschdidaktische Forschung so wie jede andere auch darauf angewiesen, dass die Primärstudien von möglichst hoher Qualität sind.

Literatur

Ahn, Soyeon/Ames, Allison J./Myers, Nicholas D. (2012): A Review of Meta-Analyses in Education: Methodological Strengths and Weaknesses. In: Review of Educational Research, 82, 4, 436-476.

Boelmann, Jan M. (Hrsg.) (2016): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. 2., durchges. Aufl. Baltmannsweiler: Schneider Hohengehren.

Borenstein, Michael/Hedges, Larry V./Higgins, Julian P. T./Rothstein, Hannah R. (2009): Introduction to Meta-Analysis. Chichester: Wiley.

Cooper, Harris M. (2017): Research Synthesis and Meta-Analysis. A Step-by-Step Approach. 5. Aufl. Los Angeles: Sage.

Cooper, Harris M./Hedges, Larry V./Valentine, Jeffrey C. (Hrsg.) (2009): The Handbook of Research Synthesis and Meta-Analysis. 2. Aufl. New York: Russell Sage Foundation.

Ellis, Paul D. (2010): The Essential Guide to Effect Sizes. Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge: Cambridge University Press.

Funke, Reinold (2014): Erstunterricht nach der Methode Lesen durch Schreiben und Ergebnisse schriftsprachlichen Lernens – eine metaanalytische Bestandsaufnahme. In: Didaktik Deutsch, 19, 36, 20-41.

Funke, Reinold (2018): Working on Grammar at School: Empirical Research from German-Speaking Regions. In: L1 – Educational Studies in Language and Literature, 18, 1-39.

Glass, Gene V. (1976): Primary, Secondary, and Meta-Analysis of Research. In: Educational Researcher, 5, 10, 3-8.

Graham, Steve/Perin, Dolores (2007): A Meta-Analysis of Writing Instruction for Adolescent Students. In: Journal of Educational Psychology, 3, 445-476

Groeben, Norbert/Hurrelmann, Bettina (Hrsg.) (2006): Empirische Unterrichtsforschung in der Literatur- und Lesedidaktik. Ein Weiterbildungsprogramm. Weinheim: Juventa.

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

Liberati, Alessandro/Altman, Douglas G./Tetzlaff, Jennifer/Mulrow, Cynthia/Gøtzsche, Peter C./Ioannidis, John P. A./Clarke, Mike/Devereaux, Philip J./Kleijnen, Jos/Moher, David (2009): The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. In: In: Annals of Internal Medicine, 151, 4, W-65-W-94.

Lipsey, Mark W./Wilson, David B. (2001): Practical Meta-Analysis. Thousand Oaks: Sage Publications.

Moher, David/Liberati, Alessandro/Tetzlaff, Jennifer/Altman, Douglas G./PRISMA Group (2009): Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. In: Annals of Internal Medicine, 151, 4, 264-269.

Moher, David/Shamseer, Larissa/Clarke, Mike/Ghersi, Davina/Liberati, Alessandro/Petticrew, Mark/Shekelle, Paul/Stewart, Lesley A. (2015): Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement. In: Systematic Reviews, 4, 1, 1-9.

Rost, Detlef H. (2007): Interpretation und Bewertung pädagogisch-psychologischer Studien. Eine Einführung. 2., überarb. und erw. Aufl. Weinheim: Beltz.

JAN M. BOELMANN

Forschungsparadigma

Es gehört zum wissenschaftlichen Basiswissen, dass Fragestellungen mit qualitativen oder quantitativen Forschungsansätzen beantwortet werden können. Landläufig ist damit die Vorstellung verknüpft, dass die Größe der Probandengruppe – qualitative Forschung arbeitet mit wenigen Probanden, quantitative mit vielen – über die Auswahl der jeweiligen Form entscheidet. Dass der eigentliche Unterschied aber nicht in der Größe der Probandengruppe liegt, sondern in den dahinterliegenden Grundannahmen, also den sogenannten Forschungsparadigmen, gehört zum Grundlagenwissen der/des Forschenden.

Unter einem Forschungsparadigma versteht man den gedanklichen Rahmen, der die verschiedenen Annahmen bündelt, die einer spezifischen Art zu forschen zu Grunde liegen.

> Ein Paradigma bezeichnet nach Kuhn (1977) das allgemein akzeptierte Vorgehen (Modus operandi) einer wissenschaftlichen Disziplin einschließlich eines gemeinsamen Verständnisses von ‚Wissenschaftlichkeit'. (Bortz/Döring 2006, 15)

Hierbei werden Paradigmen selten als solche im Vorfeld definiert, sondern bilden sich aus verschiedenen Praktiken heraus und lassen sich in der Rückschau rekonstruieren. Paradigmen lassen sich auf allen Ebenen empirischer Forschungsprozesse identifizieren, wobei schon die diesem Band zugrundeliegende Grundannahme von *Deutschdidaktik als empirischer Bildungsforschung* ebenso als Paradigma zu begreifen ist, wie die später von Ralf Schieferdecker benannten verschiedenen Ausprägungen qualitativer Forschung. Als Forschende/Forschender stellt die Auseinandersetzung mit den Paradigmen der eigenen Forschung deshalb eine wichtige Voraussetzung dar, da in ihr die zentralen Annahmen, Begrifflichkeiten, Modelle und Denkweisen zusammengefasst sind. Zugleich – und dies gestaltet sich insbesondere für Novizinnen und Novizen als große Herausforderung – werden die den verschiedenen Ansätzen zu Grunde liegenden Paradigmen nicht immer transparent und explizit ausgeführt, sondern zeigen sich erst in der Auseinandersetzung mit den jeweiligen Grundlagentexten. Entsprechend offen formuliert Asendorpf:

> Ein Wissenschaftsparadigma ist ein einigermaßen zusammenhängendes, von vielen Wissenschaftlern geteiltes Bündel aus theoretischen Leitsätzen, Fragestellungen und Methoden, das längere historische Perioden in der Entwicklung einer Wissenschaft überdauert. (Asendorpf 2007, 15)

Als Forschende und Forschender setzt man sich mit der Wahl seiner Vorgehensweise immer in die Traditionslinie einer gewissen Forschungsschule, womit es notwendig wird – möchte man nicht unwissentlich gegen zentrale Standards der eigenen Fachdisziplin verstoßen –, die Paradigmen und ihre Grundlagen zu kennen. Die Beiträge in diesem Band sollen hierfür eine erste Annäherung bieten, doch gilt auch hier, dass sie nur die ersten Schritte auf dem Weg zu einer umfassenden Auseinandersetzung begleiten.

Zum qualitativen und quantitativen Paradigma

In den folgenden Kapiteln werden das qualitative und das quantitative Paradigma von **Ralf Schieferdecker** und **Markus Pissarek** sowie ihre Gütekriterien von **Frederike Schmidt** in den Blick genommen. Hierbei stehen die grundlegenden Verfahrensweisen ebenso im Zentrum wie zentrale forschungs-philosophische Aspekte. Insbesondere der Beitrag von Ralf Schieferdecker stellt heraus, dass die Wahl eines qualitativen Forschungssettings zugleich auch die Weltsicht des/der Forschenden widerspiegelt und – sofern sie nicht angemessen reflektiert wird – das Potenzial besitzt, die Ergebnisse nachhaltig zu korrumpieren.

Bestehen die zentralen Unterschiede zwischen qualitativer und quantitativer Forschung darin, dass die qualitative Forschung sich auf die Suche nach *einer Wahrheit von vielen möglichen Wahrheiten* begibt und somit das Entdecken neuer Erkenntnisse erleichtert, spürt die quantitative Forschung *der einen Wahrheit* nach und eignet sich somit dazu, vorher definierte Aspekte zu erfassen – man könnte auch von einem *interpretativen* und einem *normativen* Paradigma sprechen. Während in der Frühphase der empirischen Orientierung der sogenannte *Paradigmenstreit* darüber entbrannte, ob – je nach Perspektive – qualitative oder quantitative Verfahren überhaupt für sich in Anspruch nehmen könnten, adäquat empirisch Forschung zu ermöglich, und sich auch heute noch Nachwirkungen hiervon zeigen, gilt dieser Disput mittlerweile als überwunden. Dementsprechend weisen zahlreiche aktuelle Forschungsarbeiten Kombinationen beider Paradigmen auf, die unter dem Schlagwort *Mixed-Methods* systematisiert wurden. Diesen querliegenden Ansatz stellt **Christian Müller** in seinem Beitrag vor.

Literatur

Asendorpf, Jens B. (2007): Psychologie der Persönlichkeit. Heidelberg: Springer.

Bortz, Jürgen/Döring, Nicola (2006): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Heidelberg: Springer.

Kuhn, Thomas S. (1977): The Essential Tension. Selected Studies in Scientific Tradition and Change. Chicago: University of Chicago Press.

RALF SCHIEFERDECKER

Qualitative Sozialforschung

Eine erkenntnistheoretische Einführung

1. Vorbemerkungen

Anspruch jeder Form empirischer Forschung ist es, dass durch die *Beobachtung der Wirklichkeit* Erkenntnisse gewonnen werden können. Darwins Apfel ist dafür das klassische Beispiel aus dem Bereich der empirischen Naturwissenschaften. Der Apfelbaum ist in diesem Fall die Wirklichkeit und die sich aus der Fallbeobachtung ergebende Formel die Erkenntnis. Die empirische Sozialforschung richtet ihren Blick nicht auf Gegenstände wie Äpfel. Sie nimmt den Menschen, bzw. die von Menschen geschaffene Gemeinschaft in den Blick. Dadurch stellt sich die Frage danach, was Wirklichkeit (Natur, Kultur, Gesellschaft usw.) ist und wie (und ob) wir zu Erkenntnissen über den Menschen kommen können. Im Gegensatz zur empirischen Naturwissenschaft bedarf die empirische Sozialwissenschaft der ständigen Reflexion des Zusammenhangs von Forschungsgegenstands und forschender Person. Um den Menschen empirisch erfassbar zu machen, haben sich zwei Herangehensweisen etabliert: die quantitativ-empirische Forschung (siehe den Beitrag von Pissarek in diesem Band) sowie die qualitativ-empirische Forschung. Mit quantitativen Methoden lässt sich *beschreiben*, was soziale Wirklichkeit ist. Beispielsweise kann präzise erfasst werden, wie viele Kinder aus bildungsfernen Milieus eine Hochschulzugangsberechtigung erhalten. Qualitative Herangehensweisen verfolgen das Ziel, diese soziale Wirklichkeit im Sinne einer Erkenntnistheorie zu *erklären*. Sie liefert etwa Antworten darauf, wie sich die Überzeugungen und Erwartungen von Eltern und Lehrpersonen an den Schulerfolg von Schülerinnen und Schülern bildungsnaher und bildungsferner Milieus unterscheiden und wie sich diese auf den Schulerfolg auswirken. Die grundlegende Herausforderung qualitativer Forschung besteht darin, soziale Wirklichkeit so zu beschreiben, dass bei dieser Beschreibung die Komplexität sozialer Wirklichkeit adäquat abgebildet wird. Hierbei spielt die grundlegende Sicht auf die Welt[1] des Forschenden eine zentrale Rolle. Diese Rolle wird in Kapitel 3 ausführlich dargestellt. Zugleich ergeben sich aus den verschiedenen Sichtweisen auf

[1] Mit *Sicht auf die Welt* wird hier ein paradigmatischer Ausgangspunkt verstanden, der alle empirisch-methodischen Überlegung grundlegend prägt.

Welt eine Vielzahl an theoretischen Ansätzen, methodischen Vorgehen und Begrifflichkeiten, die den Einstieg in das Thema qualitative Forschung erschweren können.

Der Versuch einer allgemeinen Einleitung in die Grundlagen qualitativer Forschung steht dabei vor mehreren Herausforderungen:

1. terminologische Unschärfen
2. vielfältige Strömungen und Entwicklungen
3. die Vielfalt der Methoden

Erstens werden im aktuellen Diskurs zentrale Begriffe teilweise unscharf verwendet, bzw. lassen sich deren Bedeutungshorizonte nicht immer klar voneinander trennen. Problematisch ist dies vor allem in jenen Situationen, in denen nicht klar ist, ob Bedeutungshorizont a oder b gemeint ist, bzw. ob nicht möglicherweise (bewusst oder unbewusst) die terminologische Bedeutung offen gehalten wird. Viele begriffliche Unklarheiten hängen oftmals mit unterschiedlichen Strömungen innerhalb qualitativer Methoden zusammen. Eben diese Unterschiede lassen sich zweitens auf verschiedene paradigmatische Grundlagen zurückführen. Hier fällt es schwer, die Orientierung zu behalten, welche Methode wie verortet ist, bzw. wie sich die jeweilige Verortung der verschiedenen Methoden zueinander verhält. Drittens lässt sich die Qualität qualitativer Forschung nicht einfach mit Blick auf einige wenige Gütekriterien einordnen. Die Einschätzung, ob eine Entscheidung im Forschungsprozess angemessen ist, beinhaltet eine gewisse Dynamik und ist abhängig vom Forschungsinteresse und dem Forschungskontext. Damit die Qualität qualitativer Forschung (auch der eigenen) bewertet werden kann, ist es daher notwendig, einen Überblick über verschiedenen Herangehensweisen und methodischer Möglichkeiten zu behalten und sie für die eigene Forschungsfrage angemessen auszuwählen.

Mit den Mitteln der qualitativ-empirischen Forschung lassen sich pädagogische und didaktische Fragen beantworten, die darauf abzielen zu erklären, warum etwas so ist, wie es ist. Gefragt wird z.B. nach den Einstellungen von Lehrpersonen, nach den Orientierungen von Peergruppen, der Sinnhaftigkeit von Unterrichtsroutinen u.v.m.

Der vorliegende Beitrag versteht sich als Kompass für einen systematischen Überblick zu den Grundlagen qualitativer Forschung unter Berücksichtigung ihrer jeweiligen erkenntnistheoretischen Zugänge. Zuerst werden in Kapitel 2 die Begriffe Methodologie und Methode beschrieben und voneinander abgegrenzt. Ausgehend von einer grundlegenden Systematik werden erkenntnistheoretische Grundpositionen in Kapitel 3 vorgestellt und entsprechende Methoden qualitativer Forschung zugeordnet. Anschließend folgt eine Sammlung hilfreicher Informationen und Hinweise in Kapitel 4, ehe abschließend in Kapitel 5 weiterführende Literaturhinweise exemplarisch vorgeschlagen werden.

2. Terminologische Abgrenzung: Methodologie und Methode

Die Begriffe Methodologie und Methode werden in Diskursen (sowohl sprachlich, z.B. auf Tagungen, wie auch schriftsprachlich, z.B. in Fachliteratur) nicht immer trennscharf verwendet[2]. Dies birgt zweierlei Problemlagen: Einerseits führt eine unklare Begriffsverwendung mit ziemlicher Sicherheit zu einer Deprofessionalisierung des Diskurses. Andererseits erschwert eine solche Unklarheit insbesondere für den wissenschaftlichen Nachwuchs den Einstieg in das Fachvokabular der Theorie(n) Qualitativer Forschung. Aus diesen Gründen ist es wichtig, bereits zu Beginn der eigenen akademischen Entwicklung diese beiden Begriffe klar voneinander zu unterscheiden.

Methoden sind Hilfsmittel, die es ermöglichen, seinem erkenntnisleitenden Ziel näher zu kommen. In der empirischen Forschung dienen sie dazu, dem Forschungsinteresse nachzugehen und die empirische Wirklichkeit zu beschreiben. In der qualitativen empirischen Sozialforschung ist es eine wesentliche Aufgabe der Methoden, sich kontrolliert vom Untersuchungsgegenstand (in der Regel Menschen) zu distanzieren. Dadurch wird ermöglicht, den Gegenstand der Forschung zu einem (beobachtbaren) Objekt zu machen, bzw. den Forschenden in die Lage zu versetzen, einen objektiven oder intersubjektiven Blick auf diesen Gegenstand einnehmen zu können. Methoden regulieren den Zugang zum empirischen Material und legen fest, wie und wann die einzelnen Schritte der Datenerhebung und Datenauswertung durchgeführt werden.

Von Methodologie wird immer dann gesprochen, wenn die grundlegenden Theorien, Konzepte und Termini einer Methode diskutiert, ausgehandelt oder verglichen werden. Die Methodologie ist demnach die theoretische Verortung einer (oder mehrere) Methoden. Eine solche methodologische Verortung legt basale Ausgangspunkte für die jeweilige Methode fest (z.B.: Was ist soziale Wirklichkeit? Wie kommen wir zu Erkenntnis?) und prägt damit grundlegend den empirischen Forschungsprozess. Eine Methodologie verankert das methodische Vorgehen, dahinterliegende Modelle und Theorien innerhalb erkenntnistheoretischer Grundpositionen (vgl. Flick et al. 2009, 106-109).

Die Begriffe Methodologie und Methode beziehen sich auf zwei unterschiedliche Ebenen und können so klar voneinander unterschieden werden. Damit verbunden ist die Notwendigkeit, beim Gebrauch der Begriffe diese beiden Ebenen zu unterscheiden. Es muss klar zu erkennen sein, ob es um die Theorie über Methoden geht oder um die Anwendung von Methoden.)

---

[2] Zu beachten ist, dass diese Unterscheidung jedoch typisch für den deutschsprachigen Raum ist. Der englische Begriff *methodology* umfasst im internationalen Kontext beide Bedeutungen.

Abhängig von der gewählten methodologischen Grundlage leiten sich verschiedene Möglichkeiten des Erkenntnisgewinns ab. Aus der philosophischen Disziplin der Logik sind drei Möglichkeiten bekannt, um Schlüsse zu ziehen und so zu Erkenntnissen zu kommen:[3]

- *Deduktion* bezieht sich auf Theorien, leitet daraus Kategorien ab und überprüft diese an der (empirischen) Praxis. Dabei wird davon ausgegangen, dass vom Allgemeinen (der Theorie) Rückschlüsse auf den Einzelfall möglich sind (vgl. Lamnek/Krell 2016, 235-237; Peckhaus 1999, 833-841). Dies geschieht typischerweise in Fragebögen oder in kleinschrittig geführten Interviews. In beiden Fällen besteht bereits eine theoretische Grundannahme (Hypothese) und um diese zu *überprüfen*, werden konkrete Forschungsfragen gestellt.
- *Induktion* zieht aus der Beobachtung von (empirischer) Praxis Schlüsse und generiert daraus Theorien. Dabei wird davon ausgegangen, dass vom Einzelfall Rückschlüsse auf das Allgemeine (die Theorie) möglich sind (vgl. Lamnek/Krell a.a.O.; Meyer 2009, 302-320). Vor allem zu den Themen, zu denen es noch keine umfangreiche theoretische Fundierung gibt, bietet es sich an, z.B. aus dem Material von offen geführten Interviews oder aus Beobachtungen neue Ansätze für mögliche Theorien zu *generieren*. Derart ausgerichtete Forschungsdesigns lassen sich ebenfalls mit konkreten Forschungsfragen verfolgen. Jedoch beinhaltet die Forschungsfrage in diesem Fall eine gewisse Offenheit im Hinblick auf die zu generierende Thesen.
- *Abduktion* ermöglicht, durch den Vergleich Zusammenhänge aufzudecken und daraus Theorien zu entwickeln. Dabei wird davon ausgegangen, dass durch den Moment des Vergleichs allgemeine Aussagen (Theorie) entstehen können (vgl. Reichertz 2003). Immer dann, wenn das Forschungsinteresse auf eine implizite Ebene abzielt, d.h. wenn es nicht darum geht, was jemand sagt oder tut, sondern welche impliziten Muster dieser Praxis zugrunde liegen, bietet es sich an, z.B. mit Beobachtungen, narrativen Interviews oder Gruppendiskussionen empirisches Material zu erheben und mittels des *Vergleichs* (zwischen den Fällen und innerhalb des Falls) Gemeinsamkeiten und Unterschiede zu identifizieren und so zu Theorien zu gelangen. Verfolgt man eine abduktive Forschungsstrategie, muss die Forschungsfrage offen nach möglichen Mustern fragen.

Je nachdem für welchen Zugang man sich entscheidet, ergeben sich daraus Folgen für das weitere methodische Vorgehen. Zu beachten gilt, dass die verschiedenen Ansätze auch im Forschungsprozess kombiniert werden können (z.B. induktive und deduktive Schlüsse).

Mit einem kritischen Blick auf die Abgrenzung von Deduktion, Induktion und Abduktion fällt auf, dass eine solche Trennung im konkreten Fall kaum aufrecht zu halten ist. Streng genommen lassen sich die drei Verfahren des Erkenntnisgewinns nur idealtypisch derart trennen. In der *Alltagspraxis* spielen meist alle drei

[3] Für Details: vgl. Reichertz 2014, 75-78.

Aspekte in unterschiedlicher Intensität eine Rolle. In der Praxis empirischer Forschung wird der Fokus meist auf ein (oder zwei) Formen des logischen Schließens gelegt. Dies bedeutet jedoch nicht, dass implizit (z.B. bei der Entwicklung eines Forschungsinteresses) nicht auch die anderen Formen des logischen Schließens zum Einsatz kommen – auch wenn diese dann zumeist nicht explizit reflektiert werden.

3. Erkenntnistheoretische Grundlagen qualitativer Forschung

3.1 Erkenntnistheoretische Verortung

Empirische Sozialforschung steht immer vor der Herausforderung, das Subjekt (Menschen) zum Objekt seiner Analyse zu machen. Bei all den Versuchen, sich methodisch vom Untersuchungsgegenstand (Mensch) zu distanzieren und somit eine Position einzunehmen, die eine objektive Beschreibung dessen ermöglicht, bleibt zumindest ein subjektives Moment bestehen, hinter dem die forschende Person – bei aller methodischer Kontrolliertheit – nicht zurück kann. Dieses subjektive Moment spiegelt die erkenntnistheoretische Grundposition der forschenden Person wider und kann (und soll) nicht ausgeblendet werden.

Nicht immer ist dem Forschenden seine eigene paradigmatische Sicht auf die Welt im Vorfeld bewusst, vielmehr kann es passieren (und das ist meiner Erfahrung nach bei den ersten Forschungsprojekten häufiger der Fall), dass sich erst durch eine ausführliche Beschäftigung mit dem eigenen Forschungsinteresse (bei der Suche nach der geeigneten Methode) immer klarer die dahinterliegende erkenntnistheoretische Grundposition herausstellt. So lässt sich die Welt beispielhaft als (I) Aushandlungsprozess der Bedeutung von Wirklichkeit, (II) stetig sozial (re-)produziert oder (III) System von Symbolen und Regeln verstehen. Abhängig von der Wahl eines bestimmten erkenntnistheoretischen Zugangs ergeben sich hierzu passende empirische Methoden der Datenauswertung (Flick et al. 2009, 19).

Egal auf welchem Weg man sich dem Thema nähert, wichtig ist, dass im Laufe des Entwurfs des Forschungsprojekts eine klare Positionierung gefunden wird und sich diese konsequent auf das Forschungsdesign auswirkt.

Erkenntnistheorie ist die Wissenschaft darüber, wie der Mensch sich Wissen aneignen kann. Dabei stellt sich auf einer philosophischen Ebene die Frage, *ob* Erkenntnis überhaupt möglich ist (Brugger 1976, 93f.); beantwortet man diese Frage positiv, stellt sich forschungspraktisch die Frage, *wie* Einsichten in eine mögliche Realität gewonnen werden können[4]. Die Erkenntnistheorien entstammen einer langen philosophischen Tradition. Im 20. Jahrhundert entwickelten sich beson-

[4] Aus einer wissenschaftstheoretischen Perspektive ist eine solche Entscheidung immer kriteriengeleitet und lässt sich bestimmten erkenntnistheoretischen Grundpositionen zuschreiben. D.h. es gibt innerhalb der Wissenschaftstheorie (Kron 1999; Poser 2001) theoretische Modelle, die zu erklären versuchen, wie Erkenntnisse entstehen.

ders viele Richtungen der Erkenntnistheorie (z.B. Empirismus, Idealismus, Positivismus, Pragmatismus usw., vgl. Schischkoff 1969, 147), wobei sich bis heute Modetrends bei der Verwendung erkenntnistheoretischer Positionen beobachten lassen.[5]

Entscheidet man sich für ein Forschungsprojekt (egal ob theoretisch oder empirisch, ob quantitativ oder qualitativ), verortet man dieses (bewusst oder unbewusst) in einer erkenntnistheoretischen Grundposition. Eine solche Verortung entscheidet darüber, ob ein Forschungsinteresse empirischer oder theoretischer Art ist, ob es eher mit einem qualitativen oder eher quantitativen Forschungsdesign bearbeitet werden kann und welche Auswertungsmethoden in Frage kommen.

Aus diesem Grund ist es ratsam, sich der eigenen Position und der dahinterliegenden Weltsicht bewusst zu werden und darauf das empirische Forschungsdesign stringent aufzubauen.

3.2 Erkenntnistheoretische Positionen

Für alle im Folgenden dargestellten Forschungspositionen gilt, dass es nicht die eine einheitliche Form erkenntnistheoretischer Positionen gibt. Vielmehr handelt es sich um vage Sammelbegriffe (Albrecht 2007, 1). Wenn mit dem Symbolischen Interaktionismus, der Ethnomethodologie und dem Strukturalismus exemplarisch drei zentrale erkenntnistheoretische Grundpositionen vorgestellt werden, muss den Lesenden bewusst sein, dass weder von *dem* Symbolischen Interaktionismus, *der* Ethnomethodologie oder *dem* Strukturalismus gesprochen werden kann. Zu jeder Position existieren verschiedene Strömungen. Mit den hier skizzierten Positionen wird jedoch versucht, grundlegende Gemeinsamkeiten darzustellen. Dabei können weder die jeweiligen Positionen im Detail geklärt, noch umfassend auf die Überschneidungen und Trennungsschärfen ausführlich eingegangen werden. Insbesondere bei Ethnomethodologie und Strukturalismus lassen sich je nach methodologischem Verständnis der jeweiligen Methode unterschiedliche Positionierungen vornehmen.

[5] Welche Position gerade ‚state oft the art' ist, hängt von vielen Faktoren ab. Diese können sein: die jeweilige wissenschaftliche Disziplin, das jeweilige hochschulische Umfeld oder auch das unmittelbare Betreuungsumfeld.

		Forschungsperspektiven		
		I	**II**	**III**
A	Erkenntnis-theoretische Position	Symbolischer Interaktionismus	Ethnomethodo-logie	Strukturalismus
B	Verständnis von Welt als…	…Aushandlungspro-zess der Bedeutung von Wirklichkeit	…stetig sozial (re-)produziert	…Systeme von Sym-bolen und Regeln
C	Beschrei-bungsebene des For-schungsinte-resses	Zugänge zu subjekti-ven Sichtweisen	Prozesse der Her-stellung sozialer Situationen	Hermeneutische Analyse tieferliegen-der Strukturen
D	Anwen-dungsfelder (Bsp.)	Biografieforschung, Analyse von Alltags-wissen	Analyse von Le-benswelten und Organisationen, Evaluations-forschung, Cultural Studies	Familienforschung, Biografieforschung, Generationen-forschung, Genderforschung
E	Ausgangs-punkt emp. Vorgehens	**analysierende Verfahren**	**rekonstruierende Verfahren**	
F	Methoden der Datener-hebung	Leitfadeninterview, Experteninterview, Videographie, Narratives Interview, Gruppeninterview	Gruppen-diskussion, Teilnehmende Beobachtung, Narratives Interview, Aufzeichnung von Interaktionen	Aufzeichnung von Interaktionen (z.B. Videographie), Sammlung von Dokumenten, z.B.: Fotografie, Film, Chatprotokolle usw.
G	Methoden der Interpre-tation/Daten-auswertung	Theoretisches Codieren, Qualitative Inhaltsanalyse, Narrative Analyse, Hermeneutische Verfahren	Konversations-analyse, Diskursanalyse, Dokumentarische Methode, Gattungsanalyse, Dokumenten-analyse	Objektive Hermeneutik, Hermeneutische Wissenssoziologie, Grounded Theory, Tiefenhermeneutik

Abb. 1: Übersicht zentraler Forschungsperspektiven in der qualitativen Forschung (Quelle: eigene Darstellung, in Anlehnung an: Flick et al. 2009, 19)

Aus den drei Positionen (A) folgen jeweils die für sie typischen Formen der Formulierung des Forschungsinteresses (B-D), sowie der Datenerhebungs- und Auswertungsmethoden (E-G). Die Übersicht (Abb. 1) kann als Orientierung hilfreich sein, so lange berücksichtigt wird, dass die Unterscheidung idealtypisch getroffen wurde und die Verortung für jedes Forschungsprojekt individuell geklärt werden muss. So ergeben sich aus der Systematik der drei grundsätzlich voneinander zu

unterscheidenden Perspektiven qualitativer Forschung Schlussfolgerungen für eine mögliche Kombination von Erhebungs- und Auswertungsmethoden, wobei im Einzelfall individuelle Gründe und Argumente dafür sprechen können, die Methoden auch anderweitig zuzuordnen (alternative Darstellungen vgl. Mruck/Mey 2005, 8; Heiser 2018, 273).

Grundsätzlich kann davon ausgegangen werden, dass innerhalb einer Position (die jeweiligen Spalten I, II, III) die Kombination verschiedener Methoden der Datenerhebung zu Ergebnissen führt, die sich aufeinander beziehen lassen. Im Gegensatz dazu ist eine Kombination der Auswertungsmethoden zwischen den jeweiligen Positionen kritisch zu prüfen. Eine solche Kombination läuft Gefahr, dass die so gewonnenen Ergebnisse sich auf unterschiedlichen erkenntnistheoretischen Ebenen verortet lassen können. Ist dies der Fall, so können die Ergebnisse nicht unmittelbar aufeinander bezogen werden. Ganz konkret bedeutet dies, dass z.B. der Versuch, gleichzeitig interaktionistisch und strukturalistisch zu arbeiten, scheitern muss, bzw. die Ergebnisse sich nicht aufeinander beziehen lassen, solange die jeweiligen zu Grunde liegende erkenntnistheoretische Position ernst genommen wird.

Um die jeweilige erkenntnistheoretische Position im Folgenden ausführlicher darzustellen, wird eine vierschrittige Systematik verwendet: Erstens werden die erkenntnistheoretischen Grundannahmen vorgestellt. Zweitens werden hieraus jeweilige Folgen für das empirische Vorgehen abgeleitet. Dies geschieht in knappen Aufzählungen, um den Überblick zu erleichtern. Drittens werden klassische Vertreter der entsprechenden Position exemplarisch erwähnt und beispielhaft Werke benannt, die sich auf den Bereich empirischer Sozialforschung nachhaltig ausgewirkt haben. Abschließend werden typische Anwendungsfelder empirischer Forschung auf Grundlage dieser Position beschrieben.

3.2.1 Symbolischer Interaktionismus: Die Welt als Aushandlungsprozess der Bedeutung von Wirklichkeit

Nehmen wir ein exemplarisches Forschungsinteresse an, das nach der „Aktualität der biographischen Dimension des Lesens im Medienzeitalter“ (Graf 1995, 98) fragt. Hierzu wertete Werner Graf 300 Lektüreautobiographien seiner Studierenden aus, um zu ergründen, welche Bedeutung Literatur für die Entwicklung von jungen Erwachsenen hat, die in zunehmend medialisierten Umfeldern aufwachsen.

Ein solches Forschungsbestreben geht davon aus, dass die Welt, in der wir leben, Teil einer objektiven Wirklichkeit ist und die Dinge innerhalb dieser Welt ihre Bedeutung dadurch erhalten, dass sie Bedeutungen für Subjekte haben. Deshalb ist es ratsam, die Position des *Symbolischen Interaktionismus* einzunehmen, um das eigene Forschungsinteresse zu formulieren und sein methodisches Vorgehen (Forschungsdesign) zu planen. Ziel der Interpretation ist die Beschreibung der Beziehung zwischen Subjekt (in Grafs Fall junge Erwachsene, die in einer Mediengesellschaft aufwachsen) und Objekt (der Literatur). Die Beziehung stellt dabei

den Untersuchungsgegenstand dar und es gilt, diesen empirisch zu operationalisieren.

Symbolischer Interaktionismus: Beim Symbolischen Interaktionismus wird von einer empirisch gegebenen Welt ausgegangen, die in „symbolischen Beziehungs- und Inhaltsformen organisiert ist" (Kron 1999, 188). Der *Symbolische Interaktionismus* beruht im Wesentlichen auf drei Prämissen (vgl. Blumer 1980):

1. Menschen verhalten sich zu Dingen (Dinge = alles was Menschen wahrnehmen, d.h. z.B. Gegenstände, Personen, Institutionen, Werte, Handlungen anderer usw.) abhängig davon, welche Bedeutung diese Dinge für den Menschen haben.
2. Die Bedeutung der Dinge ergibt sich aus der sozialen Interaktion mit anderen Menschen.
3. Die Bedeutung wird erst durch den Prozess der Auseinandersetzung und Interpretation mit diesen Dingen deutlich.

Ein bekannter Vertreter ist unter anderem George H. Mead, der mit seinem Werk *Geist, Identität und Gesellschaft* (1973) vor allem auch die im pädagogischen Kontext relevante Rollentheorie etabliert hat.

Typische empirische Anwendungsfelder sind: Datenerhebungen durch Leitfrageninterviews wie sie z.B. bei einigen Formen von Experteninterviews (Bogner/Littig/Menz 2014) vorzufinden sind. Charakteristisch für die Datenerhebung ist, dass hierbei bestimmte thematische Inhalte gezielt eingebracht werden (z.B. durch detaillierte Fragen). So können z.B. Deutschlehrerinnen und Deutschlehrer im Primarbereich als Expertinnen und Experten[6] für die Vermittlung der Schriftsprache interviewt werden. Das aktive Einbringen von Inhalten (z.B. durch Fragen) ist im Rahmen der erkenntnistheoretischen Grundposition angemessen, da nur auf diese Weise die Beziehungen zum Untersuchungsgegenstand empirisch beobachtbar werden. Zur Auswertung solcher empirischer Beobachtungen werden oft analytische Verfahren wie die Qualitative Inhaltsanalyse (Kuckartz 2017; Mayring 2015) eingesetzt. Typisch für solche Verfahren ist, dass sie darauf abzielen, Aussagen auf einer expliziten Ebene zu strukturieren und zu Kategorien zu verdichten.

Anders verhält es sich im Folgenden mit rekonstruktiven empirischen Verfahren (E II-III). Sie beziehen sich auf erkenntnistheoretische Positionen, die nicht den Anspruch erheben, etwas über Wirklichkeit aussagen zu können. Der soziale Konstruktions- und Reproduktionsprozess steht dabei im Fokus des Forschungsinteresses, was zu einem grundlegend anderen methodischen Vorgehen führt.

---

[6] Die Frage wer Experte ist und wer nicht, ist keine normative Bewertung, sondern eine theoretische Setzung. D.h. für die Studie X gehen wir davon aus, Y ist Experte zum Thema Z.

3.2.2 Ethnomethodologie: Soziale Welt als Produkt stetiger gesellschaftlicher Veränderungsprozesse

Daniel Scherf untersucht in seiner Dissertationsschrift, auf welcher Wissensgrundlage Lehrende Leseförderung betreiben (vgl. Scherf 2013). Hierzu ließ er Lehrer und Lehrerinnen in Gruppendiskussionen (vgl. Lamnek 2005; Scherf 2018) entsprechende Fragen diskutieren, wertete diese später mittels dokumentarischer Methode (vgl. Schieferdecker 2018) aus und arbeitet mit ihr die kollektiven Orientierungen der Lehrendengruppe heraus.

Eine solche Forschungsanlage setzt voraus, dass Welt als soziale Wirklichkeit beschrieben wird und diese von Menschen (bzw. durch deren Handlungen) ständig produziert und reproduziert wird. Der Untersuchungsgegenstand des empirisch Forschenden verliert dadurch seine konkrete Verankerung in einer physikalischen Welt (nicht weil es diese nicht gibt, sondern weil diese aus der eingenommenen Perspektive nicht zu beobachten ist). Soziale Wirklichkeit befindet sich in einem ständigen Prozess und lässt sich aus dieser Perspektive nicht fixieren und erst recht nicht generalisieren. Möglich sind ausschließlich Fallbeschreibungen, die es jedoch (so die Grundannahme) ermöglichen, mit einem Blick auf die Details innerhalb kleiner Ausschnitte sozialer Wirklichkeit die Muster und Strukturen größerer Zusammenhänge aufzudecken. So kann z.B. aus der Beobachtung von Begrüßungsritualen in Schulen auf eine dahinterliegende Schulkultur (als soziale Wirklichkeit) zurückgeschlossen werden.

Ethnomethodologie: Die Ethnomethodologie bezeichnet einen soziologischen Untersuchungsansatz, der davon ausgeht, dass sich große Zusammenhänge sozialer Ordnung bis auf einzelne Handlungsmuster ausprägen und wiederfinden lassen. So kann durch die Analyse von Details sozialer Mikro-Ausschnitte auf Makro-Strukturen sozial-konstruierter Wirklichkeit zurückgeschlossen werden. Konkret wird dieser Ansatz deutlich, wenn z.B. der Mikro-Ausschnitt des Unterrichtseinstiegs empirisch analysiert wird und so auf allgemeine Strukturen von Macht- und Hierarchie (Wernet 2009, 47-51) geschlossen werden kann. Wirklichkeit wird demnach verstanden als „fortwährende Hervorbringung und Leistung der gemeinsamen Tätigkeit“ (Bergmann 2009, 121).

Die Ethnomethodologie verbindet vier grundlegende programmatische Positionen, die im Rahmen dieser Abhandlung nur kurz skizziert werden (ausführlich hierzu und im Folgenden: Bergmann 2009, 125-128):

1. Der *Ort,* an dem die Konstruktion von Wirklichkeit stattfindet, ist die soziale Interaktion. Durch den Vollzug von Handlung wird der Konstruktionsprozess erst sichtbar und damit empirisch beobachtbar.
2. Die *Art und Weise* der Handlung (z.B. die Kommunikation zwischen zwei Menschen) sind jeweils für den Kontext (z.B. Schule) spezifisch. Der Zusammenhang zwischen Kontext und Kommunikation zeigt sich durch die Art und Weise, wie die Akteure diesen Kontext verstehen.

3. Der *Kontext* bestimmt grundlegend den Ablauf von Handlungen (oder Äußerungen). Ohne dass es den Akteuren bewusst ist, verweist ihre Handlung auf den Kontext, in dem sie entstehen.
4. *Grenzen eines reflexiven Kontextbezugs* bestehen darin, dass den Akteuren die Kontextbedingungen, unter denen sie handeln, immer unklar sind. Der Sinn sprachlicher Handlung ist demnach immer ungewiss und konstituiert sich aus dem Handlungs- bzw. Interaktionsprozess heraus.

Als Grundlagenwerk für ethnomethodologische Positionen gilt u.a. Garfinkels *Studis in ethnomethodology* (1967). Neuere Publikationen mit einem empirischen-methodischen Schwerpunkt beziehen sich oft auch auf die Konversationsanalyse (Bergmann 2010, 258-274; Deppermann 2000, 96-124).

Typisch für diese Forschungsperspektive ist u.a. eine relative Offenheit bei der Datenerhebung z.B. durch Formen der Beobachtung, narrativer Interviews oder Gruppendiskussionen. Ebenso können natürlich Situationen (z.B. eine Unterrichtssequenz) aufgezeichnet und als Grundlage für die Interpretation verwendet werden. In Fällen, in denen aktiv durch den Forschenden mittels eines Impulses (Frage, Bild, Problemstellung usw.) erhoben wird, kommt es darauf an, dass dieser Impuls möglichst wenig inhaltlich vorgibt. Anstatt explizit nach den Gründen für die Bildungskarriere von Lehrerkräften mit Migrationshintergrund zu fragen, bietet es sich z.B. an die Probanden erzählen zu lassen, wie es dazu kam, dass sie heute dort sind, wo sie sind.

Für die Datenauswertung wird das empirische Material in Sequenzen eingeteilt. Aus den jeweiligen Methoden werden Kriterien vorgegeben, um Anfang und Ende einer Sequenz zu bestimmen. Z.B. strukturieren *turns* bei der Konversationsanalyse die Sequenzeinheit. Bei der Dokumentarischen Methode werden die Sequenzen durch *Proposition-Konklusions-Verläufe* strukturiert. Durch ein sequenzanalytisches Vorgehen und u.a. methodenspezifische Techniken wird die Distanz zum empirischen Material vergrößert. Auf diese Weise wird ein Perspektivwechsel – weg von der Analyse des Expliziten, hin zur Rekonstruktion des Impliziten – ermöglicht.

3.2.3 Strukturalismus: Soziale Welt als System von Symbolen und Regeln

Gerhard Rupp, Petra Heyer und Helge Bonholt (2004) untersuchten, wie sich Verarbeitungsweisen von Wirklichkeit und mediale Wahrnehmungen in Eigenproduktionen von Schülerinnen und Schülern niederschlagen. Hierzu ließen sie die Probanden Fortsetzungstexte zu einer Kurzgeschichte, freie lyrische Texte oder Videokurzfilme produzieren und unterzogen die Ergebnisse einer mehrphasigen Dokumentenanalyse, zu denen Verfahren der Grounded Theory (vgl. Kubik 2018) und der hermeneutischen Textanalyse gehörten.

Strukturalismus: Die forschenden Personen richteten ihren Blick auf Strukturen, Systeme, Mechanismen und Muster, die das soziale Leben prägen, und nahmen somit die erkenntnistheoretische Perspektive des Strukturalismus ein. Strukturen

werden hier als Regeln verstanden, „die sich aus ihrer gegenseitigen Relation bestimmen“ (Schmidt 1969, 594). Im Gegensatz zum Interaktionismus (der ebenfalls Regelhaftigkeiten von Beziehungen betrachtet) werden im Strukturalismus die Muster von Regeln als subjektunabhängig verstanden. Eben diese Regeln gilt es durch qualitative Sozialforschung empirisch zu rekonstruieren. „Ziel […] ist das Erfassen des jeweiligen Netzes von Beziehungen zwischen Elementen, die als solche durch dieses Netz […] bestimmt sind.“ (Brugger 1976, 382)

Für die empirische Forschungspraxis ergeben sich daraus folgende Konsequenzen:

- Methoden zielen darauf ab, implizite Muster/Strukturen hinter expliziten Handlungen/Kommunikation herauszuarbeiten
- soziale Strukturen werden als objektiv verstanden
- aus der Beschreibung eines sozialen Ausschnitts kann auf größere, dahinter liegende Strukturen zurück geschlossen werden.

Einen klassischen Ansatz des soziologischen Strukturalismus findet man bei Lévi-Strauss‘ (1971) Theorie über Beziehungen zwischen der Struktur der Sprache und Kultur einer Gesellschaft (Hillmann 1994, 847), der sich u.a. auf den Linguistischen Strukturalismus von de Saussure (1967) bezieht.

Aus der Perspektive einer strukturalistischen Position heraus leiten sich unter anderem die folgenden Schlussfolgerungen für das empirische Vorgehen ab.

1. In authentischen Momenten zeigen sich Strukturen sozialer Wirklichkeit

Empirische Daten werden idealerweise aus authentischen Momenten gewonnen. Dies können Aufnahmen von Situationen, narrative Passagen in Interviews, natürliche Gruppen in Gruppendiskussionen oder auch Dokumente, die ursprünglich nicht zur Datenerhebung verfasst wurden, sein. Allerdings ist es ebenfalls durchaus üblich zur Forschungsfrage passende Daten zu erheben und einen Verfremdungsprozess durch die Datenerhebung in Kauf zu nehmen.

2. Unterscheidung zwischen impliziten und expliziten Strukturen

Eine wissenssoziologische Unterscheidung zwischen implizitem und explizitem Wissen bildet die Grundlage für die Interpretation der empirischen Daten. Explizit kann es in einem Interview z.B. um Unterrichtsstörungen und Disziplin gehen. Implizit wird vielleicht deutlich, dass es auch um Fragen der gegenseitigen Wahrnehmung und Achtung geht. Die Strukturen, auf die eine empirische Untersuchung abzielt, werden in der Regel als implizite Strukturen verstanden.

3. Vermeidung, eigene Muster und Strukturen zu beschreiben

Bei der Beschreibung von Mustern besteht die Gefahr, die eigenen Muster *ins Material hinein* zu interpretieren. Daher ist es hier besonders wichtig, mit Interpretationsgruppen die eigene Analyse intersubjektiv zu überprüfen. Um dies zu vermeiden, ist es wichtig, seine eigenen subjektiven Theorien (z.B. ‚Lehrerinnen sind empathischer als Lehrer‘) vor der Datenauswertung explizit festzuhalten, um dann kritisch die Datenauswertung überprüfen zu können, ob diese subjektiven Theorien nicht zu schnell ins Material hineingelegt werden.

4. Offenheit gegenüber dem empirischen Material

Um die Chance zu erhalten, aus dem empirischen Material tatsächlich Strukturen und Muster sozialer Wirklichkeit zu rekonstruieren, die einem möglicherweise selbst fremd sind, ist eine Offenheit gegenüber dem Material notwendig. Das Forschungsinteresse muss zwar klar umrissen werden, gleichwohl bleibt die Forschungsperspektive (im Vergleich zu anderen Forschungsansätzen) relativ offen, um dem empirischen Material möglichst nichts vorweg zu nehmen. Der Forschungsprozess bleibt so zu einem gewissen Teil unverfügbar. Beispielsweise ist die beobachtete Unterrichtssituation einem in der Regel so oder so ähnlich vertraut (durch eigene primäre oder sekundäre Lehrerfahrungen). Es geht nun darum, sich den Untersuchungsgegenstand mittels empirischer Methoden soweit fremd zu machen, dass dieser einen wieder mit neuen Erkenntnissen überraschen kann. Möglich sind auf diese Weise Schulkulturen und -milieus oder auch Schüler- und Lehrerhabitus zu rekonstruieren.

Welche erkenntnistheoretische Position letztlich für das eigene Forschungsvorhaben gewählt wird, hängt von verschiedenen Faktoren ab. Neben persönlichen Überzeugungen (der eigenen Sicht auf Welt), können ebenso bewusst gewählte theoretische Grundlagen (z.B. ein bestimmtes Konzept wie: was sind *Überzeugungen*, was ist *Lernen*?) sich prägend auf den weiteren Forschungsprozess auswirken. Schließlich lassen sich in den verschiedenen Disziplinen auch Trendbewegungen hin zu einer bestimmten erkenntnistheoretischen Position beobachten. In der aktuellen empirischen Sozialforschung im Kontext Bildung finden sich zwei Schwerpunkte: empirische Forschung auf Grundlage einer interaktionistischen Position (hier werden vorwiegend analysierende Verfahren, wie die Inhaltsanalyse verwendet), sowie eine strukturalistische Position (hier werden vorwiegende rekonstruierende Verfahren, wie die Objektive Hermeneutik, verwendet). Gerade für den Bereich der rekonstruktiven Verfahren (Abb. 1: E II + III) lässt sich jedoch eine erkenntnistheoretische Position nicht immer eindeutig zuschreiben[7].

4. Auf einen Blick

Es folgen kurze Aufzählungen von Informationen und Hinweisen, die für die konkrete Forschungspraxis von zentraler Bedeutung sind. Mit dieser Skizzierung soll ein möglichst rascher Einstieg in das Thema ermöglicht werden. Ebenso wird versucht, konkrete Fragen vorweg zu nehmen und hierfür auf weiterführende Informationsquellen zu verweisen.

[7] Zum Beispiel bezieht sich die Dokumentarische Methode auf sehr unterschiedliche Vertreter strukturalistischer Positionen (Pierce, Bourdieu, Luhmann), zugleich orientiert sich das methodische Vorgehen an ethnomethodologischen Ansätzen (Schütze).

4.1 Forschungsdesign

Es gibt unzählige Darstellungen von Forschungsdesigns für qualitativ-empirische Projekte. Diese hier aufzuführen, ist nur begrenzt sinnvoll, da sich die Designs stark unterscheiden können (u.a. wegen der vielfältigen Forschungsmethoden). Ein anderer, konstruktiver Weg ist der Folgende, in dem anstelle von gelisteten Forschungsdesigns zentrale Fragen als Inspirationsquelle stehen (vgl. Przyborski/Wohlrab-Sahr 2014, 118-131):

1. Was ist mein eigenes Erkenntnisinteresse und wie lässt sich dies in eine Forschungsfrage übersetzen?
2. Wo positioniere ich mich erkenntnistheoretisch und methodologisch?
3. Was ist das passende Forschungsfeld für mein Erkenntnisinteresse?
4. Für welche Erhebungs- und Auswertungsmethoden entscheide ich mich und was muss ich bei den jeweiligen Methoden für mein weiteres Handeln beachten?
5. Wie wähle ich mein Sample aus und was bedeutet dies für die Aussagekraft meiner Ergebnisse, bzw. die anschließende Theoriebildung?
6. Wie ist mein Forschungsvorhaben grundlagentheoretisch eingebettet?

Berücksichtigt man diese Struktur bei der Planung seines Forschungsprojekts, können das Vorgehen, die Forschungsfrage sowie die Ergebnisse grundlagentheoretisch eingeordnet werden.

4.2 Bedingungsfaktoren ‚guter' qualitativer Forschung

Insgesamt ändern sich die Methoden qualitativer Sozialforschung beständig. Entscheidend für das eigene Forschungsdesign ist jedoch immer und zuallererst die eigene Forschungsfrage, die sich aus der Beschäftigung mit dem Forschungsfeld und -stand ergibt. Diese kann sich gerade durch die Auseinandersetzung mit neuen Methoden verändern. Die Forschungsfrage kann ebenso dazu führen, dass eine Methode gewählt und dann an das eigene Projekt angepasst werden muss. So bieten die Methoden qualitativer Sozialforschung viele Möglichkeiten und Freiheit. Es gilt dabei, auf der einen Seite die Souveränität zu entwickeln diese Freiheiten zu nutzen und auf der anderen Seite für die damit einhergehende Verantwortung für gute und anspruchsvolle empirische Analysearbeit einzustehen.

Neben klassischen Gütekriterien qualitativer Forschung (siehe den Beitrag von Schmidt zu *Gütekriterien qualitativer Forschung* in diesem Band), müssen weitere Aspekte beachtet werden: Forschungsethische Kriterien (siehe den Beitrag von Bräuer/Vaupel in diesem Band; Miethe 2013, 927-937) finden sich in verschiedenen Positionspapieren, z.B. im Ethik-Kodex der Deutschen Gesellschaft für Soziologie (www.soziologie.de; von Unger 2014, 15-24), der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE 2010, 179-184), sowie den ethischen Prinzipien psychologischer Forschung (www.dgps.de) u.v.m.

Die Frage nach methodenübergreifenden Qualitätskriterien ist für die qualitative Sozialforschung nicht einheitlich zu beantworten. Dies stellen Strübing et al.

(2018, 84) dar. Die Autoren schlagen aus diesem Grund fünf Kriterien zur Bewertung qualitative Forschung vor (vgl. hierzu und im Folgenden: Strübing et al. 2018, 85-96):

- Gegenstandsangemessenheit
- Empirische Sättigung
- Theoretische Durchdringung
- Textuelle Performanz
- Originalität

Dagegen berücksichtigt die ebenfalls aktuelle Auflistung von Qualitätskriterien nach Heiser (2018, 44ff.) neben forschungsmethodischen auch forschungsethische Fragen:

- Gegenstandsangemessenheit/Flexibilität – Vereinbarkeit von Forschung und Methodik
- Offenheit – Keine theoretischen Vorfestlegungen
- Intersubjektive Nachvollziehbarkeit – Zustandekommen der Ergebnisse ist transparent
- Reflektierte Subjektivität – Einfluss des/der Forschenden wird berücksichtigt
- Empirische Verankerung – Theorie gründet in empirischen Daten
- Kohärenz – Innere Stimmigkeit von Daten und Interpretation
- Limitation – Generalisierungsniveau ist bestimmt
- Relevanz – Bedeutung der Ergebnisse für Scientific Community

Die Qualität empirischer Forschung hängt darüber hinaus mit der Passung von Forschungsdesign und den zur Verfügung stehenden Mitteln zusammen. Für bestimmte Formen der Datenerhebung und Datenauswertung werden mehr oder weniger Ressourcen (Arbeitszeit, Geld, Geräte usw.) benötigt. Abhängig von den zur Verfügung stehenden Ressourcen (individuelle Qualifikationsarbeit vs. finanziertes Forschungsprojekt im Verbund) gilt es kritisch zu prüfen, welches methodische Vorgehen machbar und sinnvoll ist. Dabei ist die entscheidende Ressource, die es zu beachten gilt, an erster Stelle *Zeit* (eigene Arbeitszeit, sowie Zeit von Mitarbeitenden und Hilfskräften, aber auch die Zeit der/des Betreuerin/Betreuers). Danach kommen an zweiter Stelle die Ressourcen, die mit dem Forschungs*kontext* (Zugang zu Interpretationsgruppen usw.) verbunden sind. Im Idealfall kann hier an dem Forschungsnetzwerk des Betreuungskontextes oder der jeweiligen Hochschule partizipiert werden. Erst an dritter Stelle steht dann die offensichtlichste Ressource *Geld*. Mit finanziellen Möglichkeiten lassen sich manche Ressourcen kompensieren (z.B. Zeit durch die Finanzierung von Hilfskräften zur Transkription der Daten oder Kontext durch Fahrten zu anderen Interpretationsgruppen). Jedoch bringen finanzielle Mittel nur wenig, wenn man selbst kaum Zeit für sein Projekt hat, bzw. die betreuende Person so eingebunden ist, dass keine wirkliche Betreuung stattfinden kann.

4.3 Scientific Community der Qualitativen Sozialforschung

Die Community der qualitativen Forschungsmethoden ist geprägt durch vielerlei unterschiedliche ‚Schulen', Interpretationsgruppen und Kooperationen. Dies angemessen zu erfassen oder im Rahmen dieses Beitrags darzustellen, ist nicht möglich. Daher konzentriert sich die folgende Aufzählung auf zentrale Anlaufstellen für qualitative Forschende und interessierte Personen.

Um mit der Community in Kontakt zu treten, bieten sich zwei forschungsmethodische *Tagungen* von zentraler Bedeutung an, die in Magdeburg und Berlin stattfinden. In beiden Fällen richtet sich das Angebot sowohl an Einsteiger, die sich einen Überblick verschaffen wollen, wie auch an Forschende, die ganz konkret mit ihren empirischen Daten vor Ort in Interpretationsgruppen arbeiten wollen.

- Zentrum für Sozialweltforschung und Methodenentwicklung Magdeburg (www.zsm.ovgu.de/)
- Berlin Methodentreffen (www.qualitative-forschung.de)

Auf der Suche nach forschungsmethodischen *Fachzeitschriften* lohnt sich der Blick über die Grenzen der eigenen Disziplin hinaus. So bietet u.a. die Soziologie, die Erziehungswissenschaft, die Politikwissenschaft usw. interessante Periodika an. Im Folgenden werden drei Medien vorgestellt, die sich durch eine hohe Relevanz für bildungswissenschaftliche Sozialforschung auszeichnen:

- ZQF – Zeitschrift für Qualitative Forschung (www.budrich-journals.de/index.php/zqf)
- ZISU – Zeitschrift für interpretative Schul- und Unterrichtsforschung (www.budrich-journals.de/index.php/zisu)
- FQS – Forum Qualitative Sozialforschung (www.qualitative-forschung.de)

Bei der ZQF und der ZISU handelt es sich um Fachzeitschriften (peer-review), die in regelmäßigen Abständen hochkarätige Beiträge von bekannten Autoren zu aktuellen Themen vorstellen. Beim FQS handelt es sich um ein umfassendes und freiverfügbares Onlinearchiv mit Methodenaufsätzen zu sehr vielen Themen. Auch hier findet man hochkarätige Beiträge. Gleichzeitig besteht die Möglichkeit, sich in den FQS-E-Mail-Verteiler einzutragen. In diesem finden eine rege Kommunikation und ein kollegialer Austausch statt. Regelmäßig werden auch Informationen über Tagungen, Publikationen, Stellenangebote und Projekte geteilt.

5. Kommentierte Literaturempfehlungen

Handbücher

Es gibt nahezu unbegrenzt viele Handbücher für die Methoden qualitativer Sozialforschung. Als Klassiker gilt u.a.:

Flick/v. Kardorff/Steinke (Hrsg.) (2009): Qualitative Forschung. Ein Handbuch.

Nachvollziehbar, systematisch und sehr umfassend werden hier alle wesentlichen Zusammenhänge und Fachbegriffe dargestellt. So eignet sich dieses Handbuch zum gezielten Nachschlagen und bietet umfangreiche Literaturempfehlungen zum Weiterlesen. Dabei werden auch methodologische Entwicklungsströme dargestellt.

Ebenfalls umfangreich, mit einem deutlich stärkeren Praxisbezug, bietet z.B. Frieberthäuser et al. ein Nachschlagewerk an, das den Bogen zwischen Grundlagenwissen und praktischer Anwendung spannt:

Frieberthäuser/Langer/Prengel (Hrsg.) (2013): Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft.

Wissenschaftstheoretische Grundlagen

Vielleicht ist es ein Symptom für aktuelle Trends innerhalb empirischer Sozialforschung, wenn viele gute (und immer noch gültige) Gedanken bereits vor einigen Jahren (oder Jahrzehnten) formuliert wurden. Im Rahmen dieses Beitrags wurde – vor allem für die erste Annäherung aus einer pädagogischen Perspektive und für den ersten Überblick – das Werk von Kron herangezogen:

Kron (1999): Wissenschaftstheorie für Pädagogen.

Interpretation von Bildern

In den letzten zehn Jahren hat die qualitative Sozialforschung mittels visueller und audio-visueller Datenerhebung enorm an Bedeutung gewonnen. Zugänge und Beispiele aus verschiedenen methodischen Disziplinen hierzu finden sich unter anderem in:

Frieberthäuser/v. Felden/Schäffer (Hrsg.) (2007): Bild und Text. Methoden und Methodologien visueller Sozialforschung in der Erziehungswissenschaft.

Einen systematischen Einstieg ins Thema *Bild* bietet:

Burkart/Meyer (Hrsg.) (2016): Die Welt anhalten. Von Bildern, Fotografie und Wissenschaft.

Dort wird Grundlagenwissen zum Bild als Medium, zur Herstellung und Funktion von Bildern und verschiedenen methodischen Zugängen geboten.

Typenbildung

Insbesondere bei rekonstruktiven Methoden qualitativer Sozialforschung stellt sich die Frage, wie man vom empirischen Fall zu Typen kommt. Hierzu bietet jede Forschungsmethode ihren eigenen Weg an. Ein Buch, das sich über die einzelnen Methoden hinweg mit dieser Frage auseinandersetzt, ist:

Ecarius/Schäffer (Hrsg.) (2010): Typenbildung und Theoriegenerierung. Methoden und Methodologien qualitativer Bildungs- und Biographieforschung.

Literatur

Albrecht, Jörn (2007): Europäischer Strukturalismus: ein Forschungsüberblick. Tübingen: Narr.

Berger, Peter/Luckmann, Thomas (1977): Die gesellschaftliche Konstruktion der Wirklichkeit. Eine Theorie der Wissenssoziologie. Frankfurt a.M.: Fischer.

Bergmann, Jörg G. (2009): Ethnomethodologie. In: Flick, Uwe/Kardorff, Ernst von/Steinke, Ines (Hrsg.): Qualitative Forschung. Ein Handbuch. Reinbeck: Rowohlt, 118-135.

Bergmann, Jörg G. (2010): Ethnomethodologische Konversationsanalyse. In: Hoffmann, Ludger (Hrsg.): Sprachwissenschaft. Ein Reader. 3. Aufl. Berlin/New York: De Gruyter, 258-274.

Blumer, Herbert (1980): Der methodologische Standort des Symbolischen Interaktionismus. In: Arbeitsgruppe Bielefelder Soziologen (Hrsg.): Alltagswissen und Interaktion und gesellschaftliche Wirklichkeit 1 – Symbolischer Interaktionismus und Ethnomethodologie, Wiesbaden: Springer VS.

Bogner, Alexander/Littig, Beate/Menz, Wolfgang (2014): Interviews mit Experten: Eine praxisorientierte Einführung. Wiesbaden: Springer VS.

Bohnsack, Ralf (1989): Generation, Milieu und Geschlecht. Opladen: Lesek + Budrich.

Bohnsack, Ralf (1993): Kollektivität als konjunktiver Erfahrungsraum (unveröffentlichtes Manuskript).

Bohnsack, Ralf (2008): Rekonstruktive Sozialforschung: Einführung in Qualitative Methoden. Stuttgart: UTB.

Bourdieu, Pierre (1976): Entwurf einer Theorie der Praxis. Frankfurt a.M.: Suhrkamp.

Breidenstein, Georg/Hirschauer, Stefan/Kalthoff, Herbert/Nieswand, Boris (2015): Ethnografie. Die Praxis der Feldforschung. Konstanz: UVK.

Brugger, Walter (1976): Philosophisches Wörterbuch. Freiburg: Herder.

Burkart, Günter/Meyer, Nikolaus (Hrsg.) (2016): Die Welt anhalten. Von Bildern, Fotografie und Wissenschaft. Weinheim/Basel: Beltz.

Deppermann, Arnulf (2000): Ethnographische Gesprächsanalyse: Zu Nutzung und Notwendigkeit von Ethnographie für die Konversationsanalyse. In: Gesprächsforschung, 1, 96-124, http://www.gespraechsforschung-ozs.de/heft2000/ga-deppermann.pdf (letzter Zugriff: 01.08.2018).

DGfE (Hrsg.) (2010): Erziehungswissenschaft. Mitteilungen der Deutschen Gesellschaft für Erziehungswissenschaft, 21, 41.

Dubs, Rolf (1995): Konstruktivismus: Einige Überlegungen aus der Sicht der Unterrichtsgestaltung. In: Zeitschrift für Pädagogik, 41, 6, 889-903.

Ecarius, Jutta/Schäffer, Burkhard (Hrsg.) (2010): Typenbildung und Theoriegenerierung. Methoden und Methodologien qualitativer Bildungs- und Biographieforschung. Opladen/Farmington Hills: Budrich.

Flick, Uwe/Kardorff, Ernst von/Steinke, Ines (Hrsg.) (2009): Qualitative Forschung. Ein Handbuch. Reinbeck: Rowohlt.

Frieberthäuser, Barbara/Felden, Heide v./Schäffer, Burkhard (Hrsg.) (2007): Bild und Text. Methoden und Methodologien visueller Sozialforschung in der Erziehungswissenschaft. Opladen/Farmington Hills: Budrich.

Frieberthäuser, Barbara/Langer, Antje/Prengel, Annedore (Hrsg.) (2013): Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft. Weinheim/München: Beltz Juventa.

Friedrichs, Jürgen (2014): Forschungsethik. In: Bauer, Nina/Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS, 81-91.

Fuß, Susanne/Karbach, Ute (2014): Grundlagen der Transkription. Eine praktische Einführung. Opladen/Toronto: UTB.

Garfinkel, Harold (1967): Studis in Ethnomethodology. Englewood Cliffs, N.J: Prentice-Hall Inc.

Glaser, Barney, G/Strauss, Anselm L. (2008): Grounded Theory: Strategien qualitativer Forschung. Bern: Huber.

Glasersfeld, Ernst von (1987): Wissen, Sprache und Wirklichkeit. Arbeiten zum radikalen Konstruktivismus. Wiesbaden: Vieweg + Teubner.

Graf, Werner (1995): Fiktionales Lesen und Lebensgeschichte. Lektürebiographien der Fernsehgeneration. In: Rosebrock, Cornelia (Hrsg.): Lesen im Medienzeitalter. Biographische und historische Aspekte literarischer Sozialisation. Weinheim: Juventa, 97-126.

Heiser, Patrick (2018): Meilensteine der qualitativen Sozialforschung. Eine Einführung entlang klassischer Studien. Wiesbaden: Springer VS.

Hillmann, Karl-Heinz (1994): Wörterbuch der Soziologie. 4. Aufl. Stuttgart: Kröner.

Kleining, Gerhard (1982): Umriss zu einer Methodologie qualitativer Sozialforschung. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 34, 2, 224-253, http://www.ssoar.info/ssoar/handle/document/861 (letzter Zugriff: 01.08.2018).

Kron, Friedrich W. (1999): Wissenschaftstheorie für Pädagogen. München/Basel: Reinhardt.

Kubik, Silke (2018): Grounded Theory. In: Boelmann, Jan M. (Hrsg.): Empirische Forschung in der Deutschdidaktik. Band 2: Erhebungs- und Auswertungsverfahren. Baltmannsweiler: Schneider Hohengehren, 245-264.

Kuckartz, Udo (2017): Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. 3. Aufl. Weinheim/München: Beltz Juventa.

Lamnek, Siegfried (2005): Gruppendiskussion. Theorie und Praxis. Weinheim: Beltz.

Lamnek, Siegfried/Krell, Claudia (2016): Qualitative Sozialforschung. Weinheim: Beltz.

Lévi-Strauss, Claud (1971): Strukturale Anthropologie. Frankfurt a.M.: Suhrkamp.

Luhmann, Niklas (1984): Soziale Systeme. Frankfurt a.M.: Suhrkamp.

Maturana, Humberto (1996): Was ist erkennen? Zürich: Piper.

Mayring, Philipp (2015): Qualitative Inhaltsanalyse: Grundlagen und Techniken. 12. Aufl. München/Basel: Beltz.

Mead, Georg H. (1973): Geist. Identität und Gesellschaft aus der Sicht des Sozialbehaviorismus. Frankfurt a.M.: Suhrkamp.

Meyer, Michael (2009): Abduktion, Induktion – Konfusion. Bemerkungen zur Logik der interpretativen Sozialforschung. In: Zeitschrift für Erziehungswissenschaft, 12, 2, 302-320.

Miethe, Ingried (2013): Institutionalisierung forschungsethischer Standards – Welchen Weg geht die Erziehungswissenschaft? Erziehungswissenschaft 24, 47, 13-21.

Mruck, Katja/Mey, Günter (2005): Qualitative Forschung: zur Einführung in einen prosperierenden Wissenschaftszeig. In: Historical Social Research, 30, 1, 5-27, http://nbn-resolving.de/urn:nbn:de:0168-ssoar-50230 (letzter Zugriff: 01.08.2018).

Muckel, Petra (2007): Die Entwicklung von Kategorien mit der Methode der Grounded Theory. In: Historical Social Research, 19, 211-231, http://nbn-resolving.de/urn:nbn:de:0168-ssoar-288620 (letzter Zugriff: 01.08.2018).

Oevermann, Ulrich/Allert, Tilman/Kronau, Elisabeth/Krambeck, Jürgen (1979): Die Methodologie einer „objektiven Hermeneutik“ und ihre allgemeine forschungslogische Bedeutung in den Sozialwissenschaften. In: Soeffner, Hans-Georg (Hrsg.): Interpretative Verfahren in den Sozial- und Textwissenschaften. Stuttgart: Metzler, 352-434.

Peckhaus, Volker (1999): Abduktion und Heuristik. In: Nida-Rümelin, Julia et al. (Hrsg.): Rationalität, Realismus, Revision. Vorträge des 3. Internationalen Kongresses der Gesellschaft für Analytische Philosophie. Berlin/New York: De Gruyter, 833-841.

Peirce, Charles, S. (1967): Schriften zum Pragmatismus und Pragmatizismus. Frankfurt a.M.: Suhrkamp.

Poser, Hans (2001): Wissenschaftstheorie. Eine Einführung. Stuttgart: Reclam.

Przyborski, Aglaja/Wohlrab-Sahr, Monika (2014): Forschungsdesigns für die qualitative Sozialforschung. In: Bauer, Nina/Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS, 116-133.

Reichertz, Jo (2003): Die Abduktion in der qualitativen Sozialforschung. Opladen: Leske+Budrich.

Reichertz, Jo (2004): Objektive Hermeneutik und hermeneutische Wissenssoziologie. In: Flick, Uwe/Kardorff, Ernst von/Steinke, Ines (Hrsg.): Qualitative Forschung. Ein Handbuch. Reinbeck: Rowohlt, 514-524.

Reichertz, Jo (2014): Empirische Sozialforschung und soziologische Theorie. In: Bauer, Nina/Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS, 75-78.

Rennie, David L. (2005): Die Methodologie der Grounded Theory als methodische Hermeneutik: Zur Versöhnung von Realismus und Relativismus. In: ZBBBS, 1, 85-104.

Rupp, Gerhard/Heyer, Petra/Bonholt, Helge (2004): Lesen und Medienkonsum. Wie Jugendliche den Deutschunterricht verarbeiten. Weinheim: Juventa.

Saussure, Ferdinand de (1967): Grundfragen der allgemeinen Sprachwissenschaft. Berlin: De Gruyter.

Scherf, Daniel (2013): Leseförderung aus Lehrersicht: Eine qualitativ-empirische Untersuchung professionellen Wissens. Wiesbaden: Springer VS.

Scherf, Daniel (2018): Gruppendiskussionen. In: Boelmann, Jan M. (Hrsg.): Empirische Forschung in der Deutschdidaktik. Band 2: Erhebungs- und Auswertungsverfahren. Baltmannsweiler: Schneider Hohengehren, 81-98.

Schieferdecker, Ralf (2018): Dokumentarische Methode. In: Boelmann, Jan M. (Hrsg.): Empirische Forschung in der Deutschdidaktik. Band 2: Erhebungs- und Auswertungsverfahren. Baltmannsweiler: Schneider Hohengehren, 265-282.

Schischkoff, Georgi (1969): Philosophisches Wörterbuch. Stuttgart: Kröner.

Schmidt, Heinrich (1969): Philosophisches Wörterbuch. Stuttgart: Kröner.

Schülein, Johann August/Reitze, Simon (2005): Wissenschaftstheorie für Einsteiger. Wien: UTB.

Schütze, Alfred (1974): Der Sinnhafte Aufbau der sozialen Welt. Frankfurt a.M.: Suhrkamp.

Seiffert, Helmut (1983): Einführung in Wissenschaftstheorie. Band 1/2. München: C.H. Beck.

Siebert, Horst (2005): Pädagogischer Konstruktivismus. Lernzentrierte Pädagogik in Schule und Erwachsenenbildung. Weinheim/Basel: Beltz.

Strauss, Anselm/Corbin, Juliet (1994): Grounded Theory Methodology. An Overview. In: Denzin, Norman/Lincoln, Yvonna (Hrsg.): Handbook of qualitative research 17, http://www.depts.ttu.edu/education/our-people/Faculty/additional_pages/duemer/epsy_5382_class_materials/Grounded-theory-methodology.pdf (letzter Zugriff: 01.08.2018).

Strauss, Anselm/Corbin, Juliet (2010): Grounded Theory. Grundlagen Qualitativer Sozialforschung. Weinheim: Beltz.

Strübing, Jörg/Hirschauer, Stefan/Ayaß, Ruth/Krähnke, Uwe/Scheffer, Thomas (2018): Gütekriterien qualitativer Sozialforschung. Ein Diskussionsanstoß. In: Zeitschrift für Soziologie 47, 2, 83-100, https://doi.org/10.1515/zfsoz-2018-1006 (letzter Zugriff: 01.08.2018).

Tschamler, Herbert (1996): Wissenschaftstheorie. Eine Einführung für Pädagogen. Bad Heilbrunn: Klinkhardt.

Unger, Hella von (2014): Forschungsethik in der qualitativen Forschung: Grundsätze, Debatten und offene Fragen. In: Unger, Hella von/Narimani, Petra/M'Bayo, Rosaline (Hrsg.): Forschungsethik in der qualitativen Forschung. Reflexivität, Perspektiven, Positionen. Wiesbaden: Springer VS.

Wernet, Andreas (2009): Einführung in die Interpretationstechnik der Objektiven Hermeneutik. Wiesbaden: Springer VS.

Onlinequellen

http://www.budrich-journals.de/index.php/zisu (letzter Zugriff: 01.08.2018)
http://www.budrich-journals.de/index.php/zqf (letzter Zugriff: 01.08.2018)
http://www.dgps.de (letzter Zugriff: 01.08.2018)
http://www.soziologie.de (letzter Zugriff: 01.08.2018)
http://www.qualitative-forschung.de (letzter Zugriff: 01.08.2018)
http://www.zsm.ovgu.de (letzter Zugriff: 01.08.2018)

FREDERIKE SCHMIDT

Gütekriterien für qualitative Forschungsansätze

1. Systematisierung des Gegenstandsfeldes[1]

Eine empirisch forschende Deutschdidaktik hat den Anspruch, Erkenntnisse über Lehr- und Lernprozesse im Deutschunterricht – sowie deren Voraussetzungen und Folgen – gegenstandsbezogen, methodisch kontrolliert und systematisch zu gewinnen. Diese Zielsetzung hat Hermann Helmers, der Begründer der wissenschaftlichen Deutschdidaktik, bereits in den siebziger Jahren betont:

> Das Forschen einer Didaktik der deutschen Sprache wird möglichst von der Unterrichtswirklichkeit ausgehen. In diesem Sinn ist das wissenschaftliche Forschen der Didaktik empirisch zu nennen. […] Die einzelnen Ergebnisse sind jeweils in eine systematische Sicht einzubringen. Dies ist der konstruktive Faktor, den die Didaktik als Forschungselement ebenfalls nötig hat. (Helmers 1971, 29 [Herv. i.O.])

Den Ausführungen von Helmers folgend, kann man die Deutschdidaktik als eine „discipline[ ] of inquiry in education“ (Shulman 1997) bezeichnen. Verbunden mit diesem Anspruch an Erkenntnisgewinnung ist die Frage, wie ‚gute empirische Forschung‘ innerhalb der Disziplin bestimmt werden kann. Dazu haben sich wissenschaftliche Standards herausgebildet, anhand derer die Qualität und Geltung von empirischen Erkenntnissen gemessen und abgesichert wird – die sog. *Gütekriterien*. Die Beschäftigung mit dem Thema Gütekriterien schließt aber nicht nur die Betrachtung forschungsbezogener Maßstäbe, sondern auch den Blick auf Standards für die Publikation von Forschungsergebnissen ein: Empirische Befunde (sowie das diesen zugrunde liegende Datenmaterial) liegen zunächst nur den Forschenden selbst vor. Diesen exklusiven Zugang gilt es aufzuheben, wenn man anstrebt, Erkenntnisse in die wissenschaftliche Community einzubringen. Denn nur wenn Forschungsbefunde für die Diskussion zur Verfügung stehen, können Sie auch zur angesprochenen Erkenntnisgewinnung in der Deutsch-

[1] Kapitel 1 dieses Handbuchbeitrags ist in weiten Teilen bewusst zum entsprechenden Kapitel im Beitrag *Gütekriterien für quantitative Forschungsansätze* in diesem Band parallelisiert, da hier allgemein gültige bzw. bedeutsame Aspekte für das Gegenstandsfeld *Gütekriterien in der empirischen Forschung* besprochen werden.

didaktik beitragen.[2] Anders gewendet: Wer in der Deutschdidaktik empirisch forschen möchte, muss sich notwendigerweise auch mit Fragen der Güte und Geltung in der Forschungs- *und* Publikationspraxis beschäftigen.

Diese Überlegungen bilden den Ausgangspunkt für den vorliegenden Beitrag. Ausgehend von einer Darstellung allgemeingültiger Maßstäbe für empirische Forschung ist es Anliegen dieses Artikels, Gütekriterien zu diskutieren, die spezifisch zur wissenschaftlichen Bewertung und Publikation qualitativ-empirischer Forschungsprojekte angelegt werden (müssen).

1.1 Definition und Funktion von Gütekriterien

Empirische Forschung sollte sich an Standards orientieren, um Güte und Geltungsanspruch von empirischen Befunden zu gewährleisten. In der einschlägigen Literatur werden diese explizit formulierten Maßstäbe zumeist unter dem Begriff *Gütekriterien* diskutiert. Ganz allgemein sind Gütekriterien „normative theoretische Konstrukte" (Schmelter 2014, 33), die festlegen, wie und welche Wege im Forschungsprozess beschritten werden sollen, um zu empirischen Erkenntnissen zu gelangen. In Artikeln oder Handbüchern werden Gütekriterien wiederum bewusst sehr allgemein gehalten. Insofern müssen die Standards „in einem Forschungsvorhaben individuell angepasst werden" (Schmelter 2014, 33). Die Beurteilung, ob die etablierten Maßstäbe in einer Studie eingehalten wurden, ist wiederum abhängig vom Untersuchungsbereich und dem gewählten Gegenstandsfeld für das Forschungsvorhaben (Dörnyei 2007, 48f.). Folglich stehen empirisch Forschende vor einer doppelten Aufgabe: Sie müssen erstens entscheiden, an welchen Gütekriterien sie sich orientieren und zweitens eigenständig bestimmen, wie sie diese jeweils untersuchungsspezifisch konkretisieren.[3]

Im Kontext der Diskussion von Forschungsstandards wird häufig auch der Begriff *Qualitätsstandard* verwendet (siehe auch Kapitel 1.3). Diese legen fest, „welche Ausprägung die Qualitätsindikatoren [in einer empirischen Studie] jeweils mindestens habe müssen" (Döring/Bortz 2016, 83). Es geht also im Kern um die Frage, worin sich eine gute von einer weniger guten Anwendung von Gütekriterien unterscheiden lässt. In vorliegenden Publikationen zeigen sich allerdings begriffliche Unschärfen: Die Differenz zwischen *Gütekriterien* und *Qualitätsstandards* im Rahmen empirischer Forschung wird nicht immer klar markiert, teilweise werden die Termini sogar synonym verwendet.

Grundlegend lassen sich verschiedene Funktionen von Gütekriterien bestimmen. Sie sind

- *Prinzipien,* die als Orientierungsrahmen zur Planung und Durchführung von Forschungsprojekten dienen. Für Forschende sind Gütekriterien folglich

[2] Flick (2010, 403) nennt neben „Forschungspraxis", „Forschungsbewertung" und „Publikationspraxis" noch „Antragstellung" und „Lehre" als Eckpfeiler für die Diskussion von Gütekriterien. Die beiden letztgenannten Aspekte werden hier – mit Blick auf den Schwerpunkt dieses Grundlagenbandes – nicht vertiefend erörtert.

[3] Dies gilt insbesondere für Gütekriterien in der qualitativen Forschung (Kapitel 2).

Zielvorgaben bzw. Prüfsteine, die zeigen, wie Wege in der empirischen Forschung beschritten werden müssen, um zu wissenschaftlichen Aussagen – in der Deutschdidaktik also über das Lehren und Lernen im Deutschunterricht – zu gelangen;

- *Bewertungsmaßstäbe* für die (nachträgliche) Bestimmung von Güte und Geltung empirisch gewonnener Befunde; meist beziehen sich die Standards auf den Forschungsprozess. Geprüft wird somit „die Qualität des Weges zur wissenschaftlichen Erkenntnisgewinnung durch bestimmte Methoden [mit Blick auf fachliche Ansprüche und Zielsetzungen]“ (Lamnek 2010, 127);
- *Strukturierungselemente* für die Darstellung der gewonnenen Forschungsergebnisse in Publikationen, da sie als Bausteine bzw. zu berücksichtigende Textelemente im Schreibprozess Anwendung finden (sollten);
- *Kommunikationsmittel* für den (inter-)disziplinären Austausch. Gütekriterien sind notwendig, um abzusichern, dass empirische Forschung „so kommuniziert wird, dass alle Beteiligten und Angesprochenen von den Ergebnissen profitieren können“ (Riemer 2011, 199, zitiert nach Schmelter 2014, 35); dies gilt insbesondere auch für die Kommunikation bzw. Vermittlung zwischen Vertretern von unterschiedlichen Forschungsparadigmen;
- *Beitrag zur gesellschaftlichen Akzeptanz* einer Disziplin, da durch disziplinübergreifende Maßstäbe einer Beliebigkeit in empirischen Forschungsprojekten entgegengewirkt wird. Auf diese Weise erhalten Ergebnisse eine stärkere Bedeutsamkeit und können auch außerhalb der eigenen wissenschaftlichen Community Anerkennung finden (dazu u.a. Ludwig 2012, 82f.).

1.2 Übergreifende Gütekriterien für die Forschungs- und Publikationspraxis

Quantitative und qualitative Forschung sowie Mixed-Methods-Designs (siehe die Beiträge von Schieferdecker, Pissarek und Müller in diesem Band) sind Vorgehensweisen innerhalb der empirischen Forschung, die auf verschiedenen Grundannahmen basieren und daher zu unterschiedlichen Erkenntnissen gelangen. Dennoch lassen sich einige Gütekriterien formulieren, die übergreifend für empirische Forschung gelten und somit unabhängig von den gewählten Ansatz zu reflektieren sind. Hinter diesen Maßstäben steht die Haltung, dass empirische Forschungsergebnisse für alle Forschenden in einer Disziplin, unabhängig von ihrem jeweiligen methodologischen Standort, transparent sein sollen und man sich darauf einigt, welcher wissenschaftliche Anspruch an empirische Forschung allgemein als akzeptabel betrachtet wird. Auf dieser Grundlage können folgende Gütekriterien formuliert werden (leicht verändert nach Schmelter 2014, 35):

- *Nachvollziehbarkeit*: Nicht nur die zentralen Erkenntnisse einer empirischen Studie, sondern der Forschungsprozess als Ganzes – d.h. die Datenerhebung, -aufbereitung und -analyse – muss für Dritte einsichtig sein. Dieses Kriterium wird in der qualitativen Forschung meist unter dem Terminus *intersubjektive Nachvollziehbarkeit* diskutiert, in quantitativen Forschungsansätzen findet größtenteils der Terminus *Objektivität* Anwendung.
- *Gegenstandsverständnis*: Forschende müssen das Gegenstandsverständnis in einer Untersuchung offenlegen und begründen; gerade und insbesondere um das deutschdidaktische Erkenntnisinteresse in einer Studie zu verdeutlichen. Das Gegenstandsverständnis bildet die Grundlage für die Bewertung des gewählten Ansatzes und der eingesetzten Erhebungs- und Auswertungsverfahren zur Bearbeitung der Forschungsfrage(n). Zugleich ist dieser Standard eine Basis für die Bewertung des Praxisbezuges und die Anschlussfähigkeit einer Studie (s.u.).
- *Anschlussfähigkeit an den Forschungsdiskurs*: Es muss erkennbar sein, an welche Theorien und Forschungsergebnisse eine Studie anknüpft und welchen innovativen Charakter sie für den bestehenden Forschungsdiskurs – hier: die Deutschdidaktik – besitzt.
- *Praxisbezug*[4]*:* Gemeinhin sollen Forschungsergebnisse gesellschaftliche Relevanz aufweisen. Dieses übergreifend formulierte Kriterium ist anschlussfähig an die Zielperspektive der Deutschdidaktik, zur Verbesserung und Weiterentwicklung des Lehrens und Lernens im Deutschunterricht beizutragen. Daher ist ein weiterer Ausweis von Güte, dass empirische Erkenntnisse nicht nur für die Forschung, sondern auch für den professionellen Alltag des Deutschunterrichts bedeutsam sind.
- *Einhaltung forschungsethischer Standards*: Empirische Forschung sollte ethische Standards genügen[5].

In der Zusammenschau wird deutlich, dass sich Gütekriterien auf zwei verschiedene Ebenen beziehen: Einerseits fokussieren die formulierten Maßstäbe die eingesetzten Methoden in einer Studie, andererseits nehmen sie das Untersuchungsdesign als Ganzes in den Blick. Bereits einleitend wurde betont, dass die Darstellung empirischer Ergebnisse eine zentrale Dimension innerhalb des Gegenstandsfeldes *Gütekriterien* ist. Eine gute Orientierung für (Nachwuchs-)Forschende bieten hier der Kriterienkatalog von Elliot/Fischer/Rennie (1999, 220), in dem Publikationsstandards[6] für quantitative und qualitative Forschung bestimmt werden.

---

4 In der einschlägigen Literatur wird teilweise auch der Terminus *Relevanz* genutzt.

5 Vertiefend dazu den Beitrag von Bräuer/Vaupel in diesem Band.

6 Elliot und Kollegen (1999) diskutieren weiterhin spezifische Standards für die Publikationspraxis bei qualitativer Forschung (s.u., Kapitel 2.2). Zur Überprüfung der Standards von Elliot/Fischer/Rennie (1999) siehe die Studie von Ilg und Boothe (2010).

Die nachfolgend angeführten Maßstäbe wurden von mir jeweils um damit verbundene Reflexionsfragen[7] zur Konkretisierung ergänzt:

- *Explicit scientific context and purpose*
  - Werden die Zielsetzungen der Studie klar benannt?
  - Wird Literatur aus dem Fachdiskurs zur Herleitung der Fragestellung(en) herangezogen?
- *Appropriate methods*
  - Werden die eingesetzten Erhebungs- und Auswertungsverfahren reflektiert?
- *Respect for participants*
  - Wird expliziert, dass ethische Standards in der Studie eingehalten wurden (z.B. durch Anonymisierung der Daten)?
- *Specification of methods*
  - Werden die gewählten Erhebungs- und Auswertungsverfahren angemessen dargestellt?
- *Appropriate discussion*
  - Werden in der Publikation die zentralen Elemente einer Studie diskutiert (z.B. theoretische Setzungen, inhaltlicher Ertrag für die Deutschdidaktik, Methodik, Schlussfolgerungen, Grenzen der Studie, …)?
- *Clarity of presentation*
  - Ist die Studie leserseitig nachvollziehbar dargestellt?
  - Werden die theoretischen Ansätze, die Methodik und die empirischen Ergebnisse angemessen dargestellt?
- *Contribution to knowledge*
  - Wird in der Studie neues empirisches Wissen für die Deutschdidaktik generiert?

1.3 Wissenschaftlichkeit und wissenschaftliche Qualität im Rahmen empirischer Forschung

In Kapitel 1.1 wurde bereits angerissen, dass im Rahmen empirischer Forschung nicht zuletzt zu klären ist, wie sich eine gute von einer weniger guten Anwendung von Gütekriterien unterscheiden lässt. Dies berührt die Frage nach dem wissenschaftlichen Anspruch an Forschungsprojekte, welche in der Debatte zumeist unter dem Begriff *Qualitätsstandards* diskutiert wird. Konkret geht es darum zu klären, welche Mindest-, Regel- und Maximalstandards für empirische Projekte angelegt werden (Döring/Bortz 2016, 82f.). Allgemein und insbesondere in der Deutschdidaktik ist dabei augenfällig, dass Qualitätsstandards bzw. Indikatoren zur Bewertung der Qualität bislang kaum diskutiert werden. Ein Erklärungsansatz

[7] Die von mir formulierten Reflexionsfragen erheben nicht den Anspruch auf Vollständigkeit.

mag sein, dass sich Pauschalaussagen hier kaum formulieren lassen, da die Bewertung und Einschätzung der Qualität einer empirischen Studie immer abhängig von Untersuchungsbereich und -gegenstand ist (s.o., Kapitel 1.1). Zumindest einen ersten Zugang zur Thematik bietet eine Übersicht von Döring/Bortz (2016, 91), die aus Perspektive der empirischen Sozialforschung vier allgemein anerkannte Standards der Wissenschaftlichkeit und die zugehörigen Kriterien wissenschaftlicher Qualität zusammenführen:

Die vier Standards der Wissenschaftlichkeit und die zugehörigen vier Kriterien der wissenschaftlichen Qualität im Überblick

Standards der Wissenschaftlichkeit	Kommentar	Kriterien der wissenschaftlichen Qualität	Kommentar
Sie müssen von jeder wissenschaftlichen Studie prinzipiell eingehalten werden.	*Diese Fragen müssen bei einer wissenschaftlichen Studie prinzipiell bejaht werden können.*	*Sie sind bei wissenschaftlichen Studien graduell sehr unterschiedlich ausgeprägt und differenzieren herausragende, gute, durchschnittliche und schwache Studien.*	*Bei diesen Fragen ist anhand von Vergleichsstudien, Referenzwerten aus der Methodenlehre und inhaltlichen Argumenten der Grad der Ausprägung in der jeweiligen wissenschaftlichen Studie abzuschätzen, um ihre Qualität einzustufen.*
1. Wissenschaftliches Forschungsproblem	Bearbeitet die Studie ein Forschungsproblem, das sich in einen anerkannten wissenschaftlichen Forschungs- und Publikationskontext einordnet?	**1. Inhaltliche Relevanz**	In welchem Maße trägt die Studie mit ihren Ergebnissen a) zum wissenschaftlichen Erkenntnisfortschritt (wissenschaftliche/theoretische Relevanz) und/oder b) zur Lösung praktischer Probleme (praktische Relevanz) bei?
2. Wissenschaftlicher Forschungsprozess	Orientiert sich die Studie an etablierten wissenschaftlichen Methodologien und Methoden, die zum Forschungsproblem passen?	**2. Methodische Strenge**	Wie anspruchsvoll sind die gewählten Methodologien und Methoden, wie gut sind sie zur Bearbeitung des Forschungsproblems geeignet, und wie regelkonform werden sie umgesetzt?
3. Wissenschafts- und Forschungsethik	Folgt die Studie den Prinzipien der Wissenschafts- und Forschungsethik?	**3. Ethische Strenge**	Wie konsequent und umfassend werden einzelne Standards der Wissenschafts- und Forschungsethik erfüllt?
4. Dokumentation des Forschungsprojektes	Sind Vorgehen und Ergebnisse der wissenschaftlichen Studie im Detail nachvollziehbar dokumentiert?	**4. Präsentationsqualität**	Wie vollständig, wohlstrukturiert und gut lesbar wird die Studie in ihrem Ablauf und mit ihren Befunden präsentiert und wie umfassend werden die Standards der Berichterstattung des jeweiligen Faches und Publikationsorgans eingehalten?

Abb. 1: Vier Standards der Wissenschaftlichkeit und dazugehörige Kriterien der wissenschaftlichen Qualität (Döring/Bortz 2016, 90)

1.4 Herausforderungen und Desiderate

Eine Systematisierung dazu, *wie* in der Deutschdidaktik empirisch geforscht werden sollte, muss auch in den Blick nehmen, wo sich aktuell noch Blindstellen ausmachen lassen und welche Schwierigkeiten bei der Einhaltung von Gütekriterien zu berücksichtigen sind. In einer ersten Zwischenbilanz sind mindestens drei Herausforderungen bzw. Desiderate erkennbar, die es zu reflektieren gilt:

1. *Fehlende fachdidaktische Profilierung*: Die im Diskurs etablierten Gütekriterien entstammen gemeinhin der Diskussion in den empirischen Sozialwissenschaften. Das bedeutet: Die skizzierten Kriterien sind an Fragestellungen der Psychologie, der Erziehungswissenschaften und der Soziologie orientiert. Für unsere Disziplin ist wiederum wichtig, dass deutschdidaktische Erkenntnisinteressen angemessen berücksichtigt werden, wenn es um die Sicherung von Güte, Geltung und Wissenschaftlichkeit geht. Entgegen der allgemeinen Bedeutsamkeit empirischer Forschung fehlt dazu aber eine qualifizierte fachspezifische Debatte.[8] Es ist daher eine zukünftige wichtige Entwicklungsaufgabe zu klären, inwiefern und wo eigenständige Standards für empirische Forschungsansätze vonseiten der Deutschdidaktik formuliert werden müssen. Für Forschende gilt vor diesem Hintergrund, dass sie Gütekriterien auf ihr Forschungsprojekt abstimmen und die Anwendung der gewählten forschungsmethodologischen Standards für ihre Zwecke offenlegen müssen.[9]
2. *„Benchmarkproblem“*[10]: Flick (2010, 405) charakterisiert mit diesem Begriff den Aspekt, dass innerhalb der empirischen Forschung noch geklärt werden muss, welcher Qualitätsanspruch an empirische Projekte angelegt wird. So ist auch ein in der Deutschdidaktik bislang noch ungeklärter Punkt, was Mindest-, Regel- und Maximalstandards sind, etwa wenn es um Stichprobengrößen im Rahmen qualitativer oder quantitativer Forschung in Promotionsvorhaben geht.[11]
3. *Bedingungen guter Forschung*: Daneben fällt auf, dass in der Debatte häufig ökonomische Gesichtspunkte ausgespart werden. Empirisch Forschende

---

[8] Wenngleich hier Entwicklungsarbeit zu leisten ist, so ist in den letzten Jahren immerhin erkennbar, dass zunehmend eine Diskussion um fachspezifische Forschungsmethodik und Untersuchungsdesigns innerhalb der Deutschdidaktik entsteht, wie einige neuere Publikationen dokumentieren (z.B. Boelmann 2016; Becker-Mrotzek 2017).

[9] Ein gutes Beispiel für Herausforderungen und den Umgang mit dem Spannungsverhältnis zwischen deutschdidaktischem Erkenntnisinteresse und normativen Setzungen seitens der empirischen Sozialforschung bietet der Beitrag von Winkler/Steinmetz (2016).

[10] Flick diskutiert diesen Aspekt für qualitative Forschungsansätze, aus meiner Sicht ist dieser Anspruch aber allgemein auf empirische Forschungsprojekte übertragbar.

[11] Die genannten Aspekte sind vermutlich auch nicht pauschal zu klären, eine Spezifizierung von Qualitätsstandards stellt aber eine notwendige Aufgabe für die Deutschdidaktik – gerade auch mit Blick auf Anforderungen an Qualifizierungsarbeiten – dar (allgemein dazu Flick 2010, 404f.).

sollten diese jedoch als weitere Gütekriterien zur Sicherung guter Forschung, der eigenen Ressourcen und nicht zuletzt der eigenen Gesundheit reflektieren. Gerade die Faktoren *Zeit* und *Machbarkeit* (Oswald 2013, 183ff.) sind hier als zentrale Eckpfeiler aufzufassen. Bezogen auf den Faktor *Zeit* ist ein häufiger Fehler, dass der Aufwand für das Auswerten der Daten und das Schreiben nicht bzw. nicht genügend in den Blick genommen wird und diese beiden Prozesse unter erhöhtem Zeitdruck (und oftmals auch in verkürzter Zeit) im letzten Abschnitt eines Forschungsvorhabens vollzogen werden. Eng damit zusammenhängend ist der Faktor *Machbarkeit*. Trotz oder gerade in der Begeisterung für ein Projekt müssen die realistischen und realisierbaren Anforderungen im Blick bleiben, d.h. wie viele Ressourcen überhaupt für die Arbeit an einem Projekt bestehen.

Die genannten Problemfelder und Herausforderungen verdeutlichen, dass zukünftig die „Qualität der Forschung in ihren Prozessen und Ergebnissen umfassender in den Blick zu nehmen [ist]" (Schmelter 2014, 43). Eine solche Debatte stellt nicht zuletzt einen Professionalisierungsanspruch für die Deutschdidaktik dar, der einen Beitrag zu (1) einer Profilierung eines geteilten empirischen Orientierungsrahmens, (2) einer systematischen Evaluation von Studienergebnissen und (3) einem innerdisziplinären Austausch leistet. Gegenwärtig ist es Aufgabe für (Nachwuchs-)Forschende, hier eigene Wege zu gehen. Positiv gewendet stellen diese (notwendigen) Adaptionen einen eigenen Beitrag zum Diskurs dar, der in der Publikation herausgestellt werden kann und sollte. Auf pragmatischer Ebene sollten die vorhergehenden Ausführungen verdeutlicht haben, dass die Einhaltung von Gütekriterien eine gute Planung (Stichwort: Zeit und Machbarkeit, s.o.) erfordert, die Standards aber zugleich einen guten Orientierungsrahmen und Prüfstein für den Forschungsprozess bieten.

2. Welche Gütekriterien gelten für qualitative Forschungsansätze?

2.1 Perspektiven auf Gütekriterien qualitativer Forschung

Im Gegensatz zur quantitativen Forschung (siehe *Gütekriterien für quantitative Forschungsansätze* in diesem Band) ist ein breit akzeptierter Konsens hinsichtlich der Festlegung von Gütekriterien für qualitative Forschungsansätze (noch) nicht festzustellen (u.a. Flick 2014, 422); vor diesem Hintergrund variieren darüber hinaus die Begriffe für (mögliche) Gütekriterien in den verschiedenen Publikationen. Die kontroverse Debatte ist vor allem auf die unterschiedlichen Zielsetzungen und methodischen Grundlagen innerhalb qualitativer Forschungsansätze zurückzuführen (siehe für einen Überblick: Schieferdecker in diesem Band); insgesamt lassen sich in der Diskussion drei Positionen unterscheiden (Steinke 2012, 319ff.; für die englischsprachige Diskussion: Madill/Jordan/Shirley 2000):

- *Quantitative Kriterien für qualitative Forschung:* Charakteristisch für diese Position ist, dass die etablierten Forschungsstandards für quantitative Forschung (siehe *Gütekriterien für quantitative Forschungsansätze* in diesem Band) auf qualitative Forschungsansätze übertragen werden können. Grundgedanke ist ein generalistisches Konzept: Es sollen einheitliche Standards formuliert werden, die für jedwede empirische Forschung gelten. Daher werden die etablierten Gütekriterien quantitativer Forschung adaptiert, operationalisiert und teilweise reformuliert, um sie an die Ausrichtung qualitativer Forschungsansätze anzupassen (Vertreter u.a.: AERA 2006; Mayring 2012; Przyborski/Wohlrab-Sahr 2009, 35ff.).
- *Eigene Kriterien qualitativer Forschung*: Vertreter dieser zweiten Grundposition betonen, dass qualitative Forschung eigenen Forschungslogiken unterliegt und man diesem Umstand Rechnung tragen müsse. Daher wird argumentiert, dass die gängigen Gütekriterien für quantitative Forschung (siehe *Gütekriterien quantitativer Forschungsansätze* in diesem Band) nicht unmittelbar auf qualitative Forschungsansätze übertragen werden können. Ausgangspunkt für diese Position ist daher, dass eigene, spezifische Gütekriterien zur Bewertung qualitativer Forschungsansätze formuliert werden müssen. Häufig werden „Glaubwürdigkeit, Authentizität und intersubjektive Nachvollziehbarkeit" als alternative Kriterien zu den (quantitativen) Standards „Objektivität, Reliabilität und Validität" genannt (Vertreter: u.a. Flick 2014; Lincoln/Guba 1985; Steinke 2012).[12]
- *Ablehnung von Gütekriterien*: Kennzeichnend für diese Position ist, dass die Formulierung und Anwendung von Bewertungskriterien für qualitative Forschung grundsätzlich abgelehnt wird. In der Argumentation wird hervorgehoben, dass Realität sozial konstruiert wird und sich daher jeglicher Bewertung durch Dritte entzieht (Vertreter: u.a. Feyerabend 1983; Smith/Hodkinson 2005)

Zusammengefasst: Die Vorstellungen von Gütekriterien variieren und zeigen dabei eine gewisse Spannbreite, wodurch die Bewertung von qualitativen Forschungsprojekten nach allgemeingültigen Standards erschwert wird. In der aktuellen Diskussion scheint sich zunehmend die Position durchzusetzen, dass Gütekriterien in der qualitativen Forschung notwendig sind, um einer Beliebigkeit im Forschungsprozess entgegenzuwirken und um Möglichkeiten zu einem interdisziplinären Dialog zu schaffen. Zugleich wird dafür plädiert, dass eigene Kriterien zu formulieren sind, um Zielsetzungen und Ansprüchen qualitativer Forschung gerecht zu werden (u.a. Steinke 2012; Schmelter 2014).[13]

---

[12] Zum Teil ist allerdings zu diskutieren, inwiefern hier wirklich ‚alternative' Standards formuliert werden (ausführlich dazu Ludwig 2012, 83).

[13] Interessant ist in diesem Zusammenhang das Textkorpus aus der seit dem Jahr 2000 geführten Debatte zu *Qualitätsstandards qualitativer Forschung* in der Online-Zeitschrift *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*.

2.2 Kernkriterien

Mit Blick auf die beschriebene Gemengelage dürfte deutlich geworden sein, dass es ‚die' Gütekriterien für qualitative Forschungsansätze nicht gibt. In der Methodenliteratur werden vielmehr verschiedene Kriterien summiert, die jeweils mit unterschiedlichen inhaltlichen Forschungsinteressen und -kontexten verknüpft sind (u.a. Dörnyei 2007, 54; Steinke 2012, 323). Dennoch lassen sich einige gemeinsam anerkannte Maßstäbe formulieren, die sich an übergreifenden Gütekriterien (Kapitel 1.1) und Grundprinzipien qualitativer Forschungsansätze[14] orientieren (siehe u.a. Flick 2014, 422; Schmelter 2014, 42f.). Zentral sind hier:

- *Intersubjektive Nachvollziehbarkeit*
  Mit diesem Kriterium wird fokussiert, dass das Zustandekommen der Ergebnisse transparent ist, sodass Außenstehende den Forschungsprozess nachvollziehen und bewerten können. Dieses Kriterium bezieht sich somit auf (1) die Dokumentation des Forschungsprozesses, (2) die Interpretationen von Forschungsdaten in (Arbeits-)Gruppen und (3) die Anwendung kodifizierter Verfahren (z.B. Tests). Grundlegend ist dabei, dass Schwierigkeiten in der Darstellung transparent gemacht werden (siehe auch das Kriterium *Offenheit*), zumal dies zum Nachvollziehen der Erkenntnisgewinnung (im Sinne eines negativen Wissens) im Forschungsprozess und einer Fehlervermeidung in möglichen (eigens durchgeführten oder durch Dritte initiierten) Anschlussstudien beiträgt (dazu auch Schmelter 2014, 36). Beide Punkte sind als Ausweis von Reflexion anzusehen.

- *Gegenstandsangemessenheit*
  Gegenstandsangemessenheit ist nicht nur für qualitative Forschung, sondern allgemein als das zentrale Kriterium im Rahmen empirischer Forschung in der Deutschdidaktik anzusehen (s.o., Kapitel 1.2). Das Kriterium bezieht sich einerseits auf die Vereinbarkeit von Forschungsfragen und der gewählten Methodik(en) sowie andererseits auf den Forschungsprozess als Gesamtkonstrukt. Häufig wird deshalb auch der Begriff *Indikation* in der einschlägigen Literatur gewählt (z.B. Steinke 2012, 326).

- *Offenheit*
  Untersuchungsgegenstand, -methode und -situation werden weitgehend offen behandelt. Theoretische Vorüberlegungen sind daher als vorläufig zu betrachten und der/die Forschende muss bereit sein, im Forschungsprozess ggf. entwickelte Theorien, Entscheidungen und geplante Analyseschritte zu verwerfen oder zu modifizieren.

- *Reflektierte Subjektivität*
  Die Standortgebundenheit und der Einfluss des/der Forschenden (eigene Biographie/wiss. Sozialisation, Forschungsinteressen, Vorannahmen usw.) muss im Untersuchungsprozess berücksichtigt und reflektiert werden.

[14] Siehe den Beitrag von Schieferdecker in diesem Band.

- *Empirische Verankerung*
  In der qualitativen Forschung gründet die Theorie in den empirischen Daten. Die Theoriebildung muss rekursiv erfolgen, d.h. es muss die Möglichkeit bestehen, Neues zu entdecken und eigene theoretische Vorannahmen zu revidieren bzw. zu modifizieren.
- *Kohärenz*
  Wichtig ist, dass die entwickelte Theorie in einer Studie in sich stimmig ist, weshalb die Kohärenz der entwickelten Theorie und mögliche Widersprüche zwischen Daten und Interpretation in den Blick genommen werden müssen.
- *Limitation/Reichweite der Daten*
  Der Geltungsbereich einer empirischen Untersuchung muss reflektiert werden. Daher müssen die Grenzen der Verallgemeinerbarkeit der empirischen Erkenntnisse bestimmt werden. Ein wichtiger Prüfstein ist hier beispielsweise die Fallkontrastierung.
- *Relevanz*
  Die empirischen Erkenntnisse aus einer Untersuchung sollen für die Disziplin bedeutsam sein. Damit verbunden sind etwa Fragen zum Praxisbezug der gewählten Fragestellung und dem Innovationspotenzial der entwickelten Theorie in einer Studie.

Die Formulierung von Indikatoren zur konkreten Umsetzung der einzelnen Kriterien würde den Rahmen dieses Beitrags sprengen. Eine zielführende Lektüre bietet hier die Übersicht von Döring/Bortz (2016, 112-113). In Bezug auf die Publikationspraxis lassen sich im Anschluss an den bereits erwähnten Kriterienkatalog von Elliot/Rennie/Fischer (1999) spezifische Standards für qualitative Forschungsansätze formulieren. Analog zu Kapitel 1.2 (s.o.) werden die einzelnen Kriterien hier jeweils um eigens ergänzte Reflexionsfragen angeführt (ebd., 220):

- *Owning one's perspective*
  - Wird die eigene Standortgebundenheit reflektiert (z.B. eigene Biographie, Vorannahmen als Forschende(r), eigenes Gegenstandsverständnis)?
- *Situating the sample*
  - Wird die Stichprobe der Studie angemessen beschrieben und reflektiert (biographischer Hintergrund, Vorwissen, Lebensumstände der Probanden usw.)?
- *Grounding in examples*
  - Werden in der Darstellung Thesen, Zusammenführungen etc. durch (Daten-)Beispiele begründet?
  - Sind die verwendeten Erhebungs- und Auswertungsdaten nachvollziehbar dargestellt (z.B. fachspezifische Modifikationen von einzelnen Verfahren)?

- *Providing credibility checks*
  - Besteht eine Glaubwürdigkeit der Interpretation der Daten (etwa durch kommunikative Validierung, Dateninterpretation in Forschungsgruppen oder Datentriangulation)?
- *Coherence*
  - Werden die konkreten Vorgehensweisen bei der Dateninterpretation expliziert und stringent dargestellt?
- *Accompishing general vs. special research tasks*
  - Basieren Generalisierungen in der Darstellung auf einer fundierten Fallrekonstruktion (z.B. durch Fallkontrastierungen, Suche von abweichenden Fällen)?
- *Resonating with readers*
  - Wird die Leserperspektive in der Darstellung hinreichend berücksichtigt?
  - Kann sich der/die Lesende ein eigenes Bild über das Forschungsvorhaben als Ganzes sowie die einzelnen Forschungsschritte machen? (z.B. Dokumentation von Entscheidungsprozessen, Darlegung von angelegten Kriterien oder Benennung von Schwierigkeiten im Forschungsprozess)

3. Bilanz

Gegenwärtig besteht nach wie vor eine rege Diskussion zu Vorschlägen für Gütekriterien der qualitativen Forschung. Ein Konsens ist in der Debatte aber noch nicht erreicht. Allgemein als auch spezifisch in der Deutschdidaktik ist aber rekonstruierbar, dass intersubjektive Nachvollziehbarkeit als zentrales Kriterium für Güte und Geltung empirischer Erkenntnisse gelten kann. Insbesondere muss in diesem Zusammenhang berücksichtigt werden, dass im qualitativen Paradigma ein weiteres Spannungsfeld entsteht: Wie verhalten sich die normativen Setzungen vonseiten der empirischen Sozialforschung und fachspezifischen Erkenntnisansprüchen innerhalb der Deutschdidaktik? Deshalb ist es zwingend notwendig, die Gütekriterien für das eigene qualitative Forschungsprojekt eigenständig zu operationalisieren. Kurz: Qualitativ Forschende in der Deutschdidaktik müssen sich (1) zu den bisher prägenden Setzungen aus der empirischen Sozialforschung positionieren und (2) ihre Haltung in der Darstellung ihrer wissenschaftlichen Befunde offenlegen. Nicht zuletzt wird so ein wichtiger Beitrag zur Professionalisierung einer empirisch forschenden Deutschdidaktik geleistet.

4. Kommentierte Literaturempfehlungen

Dörnyei (2007): Research Methods in Applied Linguistics. Quantitative, Qualitative and Mixed Methodologies.

Der Artikel bietet einen guten Überblick über die englischsprachige Diskussion zum Thema *Gütekriterien*.

Elliott/Fischer/Rennie (1999): Evolving guidelines for publication of qualitative research studies in psychology and related fields.

Elliott et al. stellen in ihrem Beitrag einen Kriterienkatalog für die Publikation von qualitativer Forschung vor. Der Artikel bietet daher Impulse für die Strukturierung von Ergebnissen und ist zugleich eine gute Reflexionsfolie für die eigene Publikationspraxis.

Schmelter (2014): Gütekriterien.

Im Artikel werden Gütekriterien empirischer Forschung aus Perspektive der DaZ-/DaF-Forschung diskutiert (mit spezifisch sprachwissenschaftlichem Fokus). Aufgrund des damit verbundenen didaktischen Blickwinkels bietet der Beitrag eine sinnvolle Ergänzung zu den Standardwerken der empirischen Sozialforschung, zumal die – hier ebenfalls noch offene – Diskussion einer fachspezifischen Profilierung von Gütekriterien mit in den Blick genommen wird.

Literatur

American Educational Research Association (AERA) (2006): Standards for reporting on empirical social science research in AERA publications. Educational Researcher, 35, 6, 33-40.

Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hrsg.) (2017): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann.

Boelmann, Jan M. (Hrsg.) (2016): Empirische Erhebungs- und Auswertungsverfahren in der Deutschdidaktik. Baltmannsweiler: Schneider Hohengehren.

Döring, Nicola/Bortz, Jürgen (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5., vollst. überarb., aktual. und erw. Aufl. Berlin: Springer.

Dörnyei, Zoltan (2007): Research Methods in Applied Linguistics. Quantitative, Qualitative and Mixed Methodologies. Oxford: Oxford University Press, 48-72.

Elliott, Robert/Fischer, Constance T./Rennie, David L. (1999): Evolving guidelines for publication of qualitative research studies in psychology and related fields. In: British Journal of Clinical Psychology, 38, 3, 215-229.

Flick, Uwe (2010): Gütekriterien qualitativer Forschung In: Mey, Günter/Mruck, Katja (Hrsg.): Handbuch Qualitative Forschung in der Psychologie. Wiesbaden: VS Verlag für Sozialwissenschaften, 395-407.

Flick, Uwe (2014): Gütekriterien qualitativer Sozialforschung In: Baur, Nina/Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS, 411-423.

Feyerabend, Paul (1983): Wider den Methodenzwang. Skizze einer anarchistischen Erkenntnistheorie. Frankfurt a.M.: Suhrkamp.

Helmers, Hermann (1971): Didaktik der deutschen Sprache. Einführung in die Theorie der muttersprachlichen und literarischen Bildung. 6., erneut bearb. u. erw. Aufl. Stuttgart: Klett.

Ilg, Stefan/Boothe, Brigitte (2010): Qualitative Forschung im psychologischen Feld: Was ist eine gute Publikation? In: Forum Qualitative Sozialforschung, 11, 2, 25.

Lamnek, Siegfried (2010): Qualitative Sozialforschung. Lehrbuch. 5., überarb. Aufl. Unter Mitarbeit von Claudia Krell. Weinheim/Basel: Beltz.

Lincoln, Yvonna S./Guba, Egon G. (1985): Naturalisitik Inquiry. Newbury Park, CA: Sage Publications.

Ludwig, Peter H. (2012): Thesen zur Debatte um Gütestandards in der qualitativen Bildungsforschung – eine integrative Position. In: Gläser-Zikuda, Michaela et al. (Hrsg.): Mixed Methods in der empirischen Bildungsforschung. Münster: Waxmann, 79-89.

Madill, Anna/Jordan, Abbie/Shirley, Caroline (2000): Objectivity and reliability in qualitative analysis. In: British Journal of Psychology, 91, 1, 1-20.

Mayring, Philipp (2012): Mixed Methods – ein Plädoyer für gemeinsame Forschungsstandards qualitativer und quantitativer Methoden. In: Gläser-Zikuda, Michaela et al. (Hrsg.): Mixed Methods in der empirischen Bildungsforschung. Münster: Waxmann, 287-300.

Oswald, Hans (2013): Was heißt qualitativ forschen? Warnungen, Fehlerquellen, Möglichkeiten In: Friebertshäuser, Barbara/Langer, Antje/Prengel, Annedore (Hrsg.): Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft. 4., durchg. Aufl. Weinheim: Beltz, 183-201.

Przyborski, Aglaja/Wohlrab-Sahr, Monika (2014): Qualitative Sozialforschung. Ein Arbeitsbuch. 4., erw. Aufl. München: Oldenbourg.

Schmelter, Lars (2014): Gütekriterien. In: Settinieri, Julia et al. (Hrsg.): Empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache. Eine Einführung. Paderborn: Schöningh, 33-46.

Shulman, Lee S. (1997): Disciplines of inquiry in education. In: Jaeger, Richard M. (Hrsg.): Complementary methods for researchers in education. Washington, D.C.: American Education Research Association, 3-19.

Smith, John K./Hodkinson, Phil (2005): Relativism, criteria, and politics. In: Denzin, Norman/Lincoln, Yvonna S. (Hrsg.): The sage handbook of qualitative research. 3. Aufl. Thousand Oaks, CA: Sage, 915-932.

Steinke, Ines (2012): Gütekriterien qualitativer Forschung In: Flick, Uwe/von Kardoff, Ernst/Steinke, Ines (Hrsg.): Qualitative Forschung. Ein Handbuch. 9. Aufl. Rowohlt: Reinbek, 319-331.

Winkler, Iris/Steinmetz, Michael (2016): Zum Spannungsverhältnis von deutschdidaktischen Fragestellungen und empirischen Erkenntnismöglichkeiten am Beispiel des Projekts KoALa. In: Krelle, Michael/Senn, Werner (Hrsg.): Qualitäten von Deutschunterricht. Stuttgart: Fillibach bei Klett, 37-56.

MARKUS PISSAREK

Quantitative Forschung

Die Fachdidaktik Deutsch weist in disziplinärer Hinsicht ein Paradoxon auf: Beinhalten die universitären Lehramts-Curricula in der Regel keine Vermittlung methodischer Ansätze aus der empirischen Sozialforschung, so zeigt sich doch in den letzten beiden Dekaden in der fachdidaktischen Forschung eine klare Tendenz zur Ausdifferenzierung und Verfeinerung der Forschungsmethodik – dies betrifft besonders quantitative Forschungsmethoden. Allein auf dem Symposium Deutschdidaktik an der Universität Hamburg im September 2018 bezogen sich 58 Vorträge und Poster bereits im Ankündigungstext explizit auf *quantitative* empirische Verfahren. In diesem einführenden Artikel sollen einige Grundprinzipien und Ansätze des quantitativen Paradigmas dargelegt werden. Dabei ist zu bemerken, dass es in der Deutschdidaktik, wie in der allgemeinen Unterrichtsforschung, schon seit längerer Zeit eine wissenschaftliche Kontroverse gibt, ob qualitative oder quantitative Verfahren besser geeignet seien, die spezifischen deutschdidaktischen Fragestellungen zu erforschen. Dieses Zwei-Lager-Denken ist in seinem Kern überholt und zudem unfruchtbar und wird inzwischen auch von manchen Autoren ironisch als „Paradigmenkrieg“ bezeichnet (für eine Ordnung der Diskursfelder und eine historische Darstellung vgl. Kelle 2008). Generell kann man Helmkes zusammenfassendem Motto „Vom Methodenpurismus zur Methodenvielfalt“ (2009, 32) folgen und konstatieren, dass immer von der zugrundeliegenden Fragestellung und dem Forschungsinteresse her entschieden werden sollte, ob man rein qualitativ, rein quantitativ oder mit einer Kombination von Forschungsansätzen beider Paradigmen, wie bei den Mixed-Methods-Verfahren (vgl. den Beitrag von Müller in diesem Band), vorgehen will. Dieser Artikel versteht sich dabei als ein selektiv-exemplarischer Einführungstext, der stark akzentuiert und vereinfacht einige Grundüberlegungen des quantitativen Paradigmas darstellt.

Dabei werden sowohl einige wenige Grundlagen der Psychometrie als auch wesentliche Schritte eines quantitativ-empirischen Forschungsprozesses erläutert (Hypothesenbildung, Konzeptualisierung, Operationalisierung, kritischer Rationalismus). Es wird einerseits versucht, insbesondere solche Aspekte zu behandeln, die vor allem Einsteiger häufig vor Probleme stellen (z.B. für welche Fragestellungen sich überhaupt quantitative Methoden empfehlen). Andererseits soll auch anhand von Beispielen ein kurzer Überblick über Elemente des gesamten empirischen Forschungsprozesses gegeben werden – von der Formulierung der

Fragestellung, der theoretischen Grundlagen und dem Deduzieren von Hypothesen bis hin zur Diskussion der Ergebnisse. Die in diesem Kapitel referierten Konzepte und Methoden sollen vor allem Nachwuchswissenschaftlerinnen und -wissenschaftlern erste Anhaltspunkte zur Planung, Durchführung und Auswertung einer empirischen Studie geben, darüber hinaus aber auch die Rezeption von Berichten aus der empirischen Forschung erleichtern. Deshalb erfolgt der Einstieg über die Erläuterung des Falsifikationsprinzips und des kritischen Rationalismus als wissenschaftstheoretische Basis der quantitativen Sozialforschung.

1. Grundprinzipien des quantitativen Paradigmas

Das quantitative Paradigma folgt grundsätzlich anderen wissenschaftstheoretischen Prämissen als das qualitative (vgl. zu erkenntnistheoretischen Überlegungen in Bezug auf das qualitative Paradigma auch den Beitrag von Schieferdecker in diesem Band), wenn sich auch beide Ansätze grundsätzlich als *empirische* Zugänge verstehen. Stark vereinfacht gesagt, folgt das quantitative Paradigma eher einem naturwissenschaftlichen Ansatz, bei dem die *Messung* von *Ausprägungen* (z.B. der Grad der Lesekompetenz) und das Herstellen von *Zusammenhängen* (z.B. die Auswirkung eines spezifischen Orthographietrainings auf die Rechtschreibkompetenz) im Vordergrund stehen und mit größeren Stichproben gearbeitet wird, um Repräsentativität zu erreichen. Repräsentativität meint in diesem Fall, dass in der ausgewählten Stichprobe (z.B. 4200 Mittelschülerinnen und -schüler) die Grundgesamtheit (alle Mittelschülerinnen und -schüler eines Landes) möglichst unverzerrt abgebildet wird. Qualitative und quantitative Forschungsansätze unterscheiden sich dabei auf Ebene der Epistemologie (z.B.: „Wie kommt man zu Erkenntnissen?“), der Ontologie (z.B.: „Welcher Realitätsbegriff liegt zugrunde?“) und ihren zugrundeliegenden Axiomen (z.B. „Von welchen beweislosen Voraussetzungen geht man aus?“). Allein zu diesen drei Feldern ließen sich ganze Bücher schreiben; zur einführenden Vertiefung sei hier auf Döring/Bortz (2016) sowie Bühner/Ziegler (2009) und für eine umfassendere Darlegung erkenntnistheoretischer und methodologischer Fragen auf die Monographie von Meinefeld (1995) verwiesen.

1.1 Kritischer Rationalismus als wissenschaftstheoretische Basis der quantitativen Sozialforschung

Ein grundlegendes Spezifikum „der“[1] quantitativ orientierten Forschung liegt darin, dass sie in wissenschaftstheoretischer Hinsicht auf dem Kritischen Rationalismus in der Prägung Karl Poppers (1989) aufsetzt, der die Auffassung vertrat, dass sich die Sozialwissenschaften epistemologisch nicht grundsätzlich von den

[1] Im Folgenden wird eine – zugegebenermaßen vereinfachende – Homogenität des quantitativen Paradigmas in forschungstheoretischer Hinsicht angenommen, um dem Einführungscharakter dieses Beitrags gerecht zu werden. Differenzierungen und anregende Kritik am Wahrheitskonzept des kritischen Rationalismus findet man unter anderem bei Kelle (2008).

Naturwissenschaften unterscheiden. Die quantitativ orientierte Sozialforschung bedient sich dabei in einem relativ streng formalisierten bzw. nahezu normierten Forschungsprozess (vgl. 1.3) spezifischer Methoden der Datenerhebung, die darauf abzielen, empirische Strukturen und Daten in numerische Strukturen und Daten zu überführen, eben zu quantifizieren. Dabei kommen verschieden elaborierte Methoden zur Analyse der Daten zur Anwendung (für eine basale Einführung vgl. Hauser/Humpert 2009; für vertiefende Einführungen vgl. Döring/Bortz 2016 und Bühner/Ziegler 2009).

Eine Kernidee des kritischen Rationalismus ist, nicht mit Induktionsschlüssen zu arbeiten und das Konzept der *Verifikation* abzulehnen. War die frühere erkenntnistheoretische Ausrichtung des Positivismus (bzw. Empirismus) noch darauf aus, durch die Sammlung empirischer Daten und Induktionsschlüsse zu allgemeingültigen Theorien zu gelangen, setzt der kritische Rationalismus hingegen auf das *Falsifikationsprinzip* bzw. den *methodologischen Falsifikationismus* (vgl. Döring/Bortz 2016, 37). Dabei geht man zunächst von einer bereits gebildeten/publizierten/anerkannten Theorie aus, die in der Regel den Startpunkt einer Studie bildet – und induziert sie nicht erst Post hoc aus den Daten. Das ist ein wesentlicher Unterschied der beiden Paradigmen und führt so weit, dass bei manchen qualitativen Ansätzen dezidiert nicht vorgesehen ist, zu Beginn des Forschungsprozesses von Hypothesen auszugehen, wohingegen die initiale Darlegung der Forschungshypothesen ein unverzichtbarer Kernbestandteil quantitativ ausgerichteter Forschungsprojekte ist.

Aus den zu Beginn vorliegenden theoretischen Grundannahmen können nun wissenschaftliche Hypothesen abgeleitet (deduziert) werden (zur eingängigen Abgrenzung von Alltagsaussagen und wissenschaftlichen Hypothesen vgl. Hauser/Humpert 2009, 13-20). Dabei stützen sich diese Hypothesen entweder auf Ergebnisse aus früheren Studien (z.B. die Ergebnisse aus den PISA-Testungen) oder sie leiten sich aus der fachdidaktischen Theorie her (z.B. einem spezifischen fachdidaktischen Modell zur Lesekompetenz – für eine synoptische Darstellung infrage kommender Aufteilungen vgl. Pissarek 2018). Entscheidend ist dabei, dass die Hypothesen so formuliert werden, dass sie statistisch überprüft werden können. Eine solche Hypothese könnte z.B. lauten: „H1: Lautleseverfahren (= Methode A) eignen sich besser zur Förderung der Leseflüssigkeit (= Merkmal/Konstrukt) als Vielleseverfahren (= Methode B)."

Der kritische Rationalismus geht davon aus, dass man diese Hypothese H1 nicht *verifizieren* kann, da man – den Gesetzen der Logik folgend – nie von einer begrenzten Datenlage zu All-Aussagen kommen kann. Wenn man beispielsweise in einer Studie mit 20 Mittelschulklassen eine signifikant höhere Wirksamkeit von Lautleseverfahren als von Vielleseverfahren für die Entwicklung von Leseflüssigkeit ermittelt, so ist nicht gesagt, dass man bei einer weiteren Stichprobenziehung mit weiteren 20 Mittelschulklassen nicht gegenteilige Ergebnisse finden könnte. Jedoch wäre es umgekehrt durch Falsifikation möglich, nicht gültige Theorien bzw. Annahmen zurückzuweisen – indem man z.B. eine Stichprobe findet, bei der sich die Annahme nicht bestätigt. Dann wäre darüber nachzudenken, ob

die Theorie Fehler aufweist (z.B. dass die beiden Verfahren eben doch gleich gut für die Entwicklung von Leseflüssigkeit funktionieren) oder ob die eingesetzten Messinstrumente (z.B. ein Leseflüssigkeitstest) das zu messende Konstrukt (Leseflüssigkeit) passend operationalisiert haben, oder ob Messfehler vorliegen könnten (z.B. die Art und Weise, wie Dekodierfehler als Subkomponente der Leseflüssigkeit operationalisiert und erfasst wurden).

Aus den genannten Gründen würde man hier mit der Nullhypothese (H0: Lautleseverfahren und Vielleseverfahren unterscheiden sich bezüglich ihrer Wirksamkeit für die Entwicklung von Leseflüssigkeit nicht) arbeiten. Wird sie zurückgewiesen, kann man davon sprechen, dass die zugrundeliegende Theorie bzw. Annahme *vorläufig bestätigt* wurde. Übersteht die Theorie (Lautlesetrainings fördern Leseflüssigkeit effizienter als Vielleseverfahren) mehrere Falsifikationsversuche (also mehrere Replikationsstudien), kann die Theorie als *bewährt* bezeichnet werden (vgl. Döring/Bortz 2016, 37). Der kritische Rationalismus beschreibt also

> Erkenntnisfortschritt als Aussondern nicht-bestätigter Theorien durch Falsifikation bzw. umgekehrt als Zurückbehalten von nicht-falsifizierten – d. h. vorläufig bestätigten bzw. bewährten – Theorien. Er wird deswegen auch als Falsifikationismus (‚falsificationism') sowie als Kritizismus (‚criticism') bezeichnet und stellt ein ausdrückliches Gegenmodell zu dem auf Verifikation basierenden Empirismus bzw. Positivismus dar. (ebd., 38)

Im Umgang mit den Hypothesen, die bei quantitativen Ansätzen *vor* der Untersuchung vorliegen (müssten), und dem Falsifikationsprinzip, das bei quantitativen Ansätzen der Hypothesenprüfung zugrunde liegt, liegen fundamentale Unterschiede zwischen dem quantitativen und dem qualitativen Paradigma. Inzwischen gibt es im Kontext der Replikationskrise der Psychologie sogar Forderungen danach, im Rahmen von Transparenzvereinbarungen im Sinne von Open Data die Hypothesen schon früh im Forschungsprozess präregistrieren zu lassen, so dass sie im Zuge der Datenauswertung nicht mehr der Datenlage angepasst werden können (vgl. 3. Kritikpunkte am quantitativen Paradigma).

1.2 Psychometrie – oder was bedeutet quantifizieren und messen?

Die Operationalisierung der theoretischen Konstrukte (z.B. der Lesekompetenz) bildet die Grundlage einer Messung. Messen bezeichnet dabei „die Abbildung empirischer Strukturen auf numerische Strukturen, so dass die numerischen Strukturen die Relationen zwischen den empirischen Strukturen widerspiegeln" (Krauss/Bruckmeier et al. 2015, 616). Bei der Operationalisierung, also der Erstellung der Messinstrumente, sind die psychometrischen Gütekriterien (vgl. das Kapitel zu den *Gütekriterien für quantitative Forschung* von Frederike Schmidt in diesem Band – auf diese wird in diesem Beitrag nicht eingegangen) zu beachten und einzuhalten. Darüber hinaus gilt es, Konstrukte, für die noch keine allgemeingültige Definition vorliegt (z.B. Professionswissen oder Sprachbewusstsein), zu „konzeptualisieren", was bedeutet, Merkmale, Fähigkeiten, Eigenschaften usf. zu

definieren, die dem zu messenden Konstrukt zugeordnet werden und die es inhaltlich füllen. So ist beispielsweise eine häufig zu findende Auffassung von Lehrenden-Professionswissen, dass es aus FW (Fachwissen), FDW (fachdidaktischem Wissen) und PW (pädagogischem Wissen) besteht, wobei diese Bereiche dann noch näher spezifiziert werden, z.B. die Fähigkeit mit Schülerkognitionen umzugehen und Anforderungen von Aufgaben an Schülerkognitionen einzuschätzen (vgl. ein ausführliches Beispiel zur Konstruktvalidierung in Pissarek/Schilcher 2017). Mit Hilfe der Operationalisierung der Konstrukte werden Theorie und Empirie erst verknüpfbar (vgl. dazu Steyer/Eid 1993, 2-4; Bühner 2011, 83-133). Die Operationalisierung eines theoretischen Konzepts bzw. einer latenten Variable legt dabei fest, „anhand welcher beobachtbaren Variablen (Indikatoren) die Ausprägung des theoretischen Konzepts bei den Untersuchungsobjekten festgestellt werden soll." (Döring/Bortz 2016, 228)

Natürlich kann man dafür auch auf bereits existierende, standardisierende Testverfahren zurückgreifen, sodass nicht alle Operationalisierung in Eigenarbeit erfolgen muss – das ist sogar ratsam. Gerade im Bereich der Leseforschung oder der Rechtschreibdiagnostik steht hier beispielsweise eine breite Auswahl an standardisierten Tests mit Normierungstabellen zur Verfügung, die einen Vergleich der Stichprobenergebnisse erleichtern (vgl. dazu Pissarek/Pronold-Günthner 2018). Tabelle 1 zeigt zur Veranschaulichung einige Beispiele für mögliche Quantifizierungen empirischer Größen.

Tab. 1: Beispiele für (deutschdidaktische) Messungen bzw. Quantifizierungen

Empirische Struktur		**Numerische Struktur**	
Merkmal/ Konstrukt	Messung	**Messwerte/ Messgrößen**	**Skalenniveau**
Lesegeschwindigkeit	→	Rohwert aus Lesegeschwindigkeitstest, z.B. 160 Wörter/Minute	Intervallskala
Leseverständnis	→	Testrohwert (z.B. aus LGVT 6-12), z.B. Werte von -6 und weniger bis 22 und mehr; gesicherte Normen	Intervallskala
Geschlecht	→	z.B. „0" für männlich und „1" für weiblich	Nominalskala
Lesemotivation	→	Skalenwert auf Zustimmungsskalen verschiedener Indikatoren (wie Lesefreude und Lesevielfalt), z.B. von „sehr gering" (1) bis (6) „sehr hoch"	Ordinalskala
Leistung im Fach Deutsch	→	Schulnoten, z.B. Ziffernoten 1 bis 6 im deutschen System (Ordinalskala) oder Testergebnisse, z.B. der Rohwert aus dem bayerischen Jahrgangsstufentest 0 bis 88 BE (Intervallskala)	Ordinalskala

Das Skalenniveau ist dabei mitzuberücksichtigen, denn nicht mit allen Messwerten können alle mathematischen Operationen sinnvoll verknüpft werden. So ist bei *Nominalskalen* lediglich die Angabe von absoluten und prozentualen Häufigkeiten sinnvoll, bestenfalls noch der Modalwert (= häufigste Kategorie). Dahingegen bieten Ordinalskalen mehr Möglichkeiten: Die Werte können kumuliert werden, man kann Prozentränge darstellen, der Median kann angegeben werden und auch die Spannweite. Noch mehr Möglichkeiten bietet die Intervallskala, die z.B. bei einem klassischen IQ-Test, oder bei der PISA-Messung vorliegt. Hiermit werden arithmetisches Mittel, Standardabweichung, Schiefe (ist die Verteilung symmetrisch?) und die Breite der Verteilung berechenbar. Verknüpft mit der Frage, ob die Operationalisierung eines Konstrukts (z.B. Lesekompetenz oder literarisches Verstehen) gelungen ist, ist auch zu klären, ob die Messung *reliabel, objektiv und valide* ist (vgl. ausführlich dazu Schmidt in diesem Band).

Ob ein Konstrukt von einigen wenigen oder besser von vielen Items gemessen wird, ist nicht immer leicht zu entscheiden und erfordert auch etwas Erfahrung bei der Entwicklung neuer Skalen. Die Item-Skala-Korrelation (die Trennschärfe) ist dabei ein wichtiger Indikator. Die Items können mit Hilfe von Summenwerten zu Skalen gruppiert werden oder auch für komplexere Verfahren wie das Erstellen von Modellen mit Hilfe der Item-Response-Theorie (probabilistische Testtheorie) verwendet werden (vgl. Döring/Bortz 2016, 482ff.). Letztere Möglichkeit stellt dabei die im forschungsmethodischen Sinne anspruchsvollere Variante dar, bietet jedoch den Vorteil, dass die Skalierung und die Dimensionalität der Skala geprüft werden können. So wird es auch möglich, nicht nur das beobachtbare Antwortverhalten (durch die manifesten Variablen, z.B. das Lösen von Textverstehensaufgaben), sondern auch die nicht direkt beobachtbaren Konstrukte (z.B. die Lesekompetenz) zu untersuchen. Für ein ausführlicheres Beispiel vgl. dazu auch die konfirmatorische Faktorenanalyse in Pissarek/Schilcher (2017, 98f.) als Beispiel für die Modellierung des fachdidaktischen Wissens und des Fachwissens von Lehrpersonen als latente Konstrukte. Grabowski weist im Kontext der empirischen Schreibdidaktik darauf hin, dass die deutschdidaktisch relevanten Konzepte häufig latente Merkmale betreffen, und häufig neue Indikatorvariablen entwickelt werden müssen; dabei wird in der Forschungspraxis oft ein entscheidender Schritt übersehen: die ausführliche Vorprüfung und Pilotierung neuer Erhebungsinstrumente (Grabowski 2017, 317f.). Werden neue Instrumente – was Novizinnen und Novizen häufig in der zeitlichen Enge von Qualifikationsphasen als „typischer Anfängerfehler“ (Döring/Bortz 2016, 98) passieren kann – zu früh bzw. ohne Konstruktvalidierung eingesetzt, kann es häufig passieren, dass ganze Messreihen unbrauchbar und aufwendige Erhebungen umsonst erfolgt sind: „Intuitiv konstruierte und nicht weiter überprüfte Erhebungsinstrumente sind in der Regel wissenschaftlich unbrauchbar“ (ebd., 318). Da dies häufig bedeutet, dass aufgrund der organisatorisch aufwendigen Erhebungsroutinen erhebliche Verzögerungen der Forschungsprojekte bis zu einer Sprengung des Zeitplans einer Qualifikationsarbeit eintreten können, ist es tatsächlich gerade im Feld „schwer messbarer Kompetenzen“ (Frederking 2008) wichtig, hier nach einem prototypischen, entlastenden Ablaufplan zu verfahren, der vor Misserfolgen schützt (vgl. z.B. den groben

Zeitplan für eine sechsmonatige Bachelor- und Masterarbeit in Döring/Bortz 2016, 154.) Insofern ist hier aus fachdidaktischer Perspektive auf jeden Fall die Ansicht Grabowskis zu unterstreichen, dass auch die Entwicklung und Erprobung eines neuen (fachdidaktischen) Erhebungsverfahrens selbst „ein anspruchsvolles Forschungsziel darstellen“ (2017, 318) kann, und nicht nur als Teilschritt eines umfassenderen Forschungsprozesses gesehen werden muss.

1.3 Prototypischer Aufbau und Elemente einer quantitativen Studie – der Forschungsprozess

Die unter 1.1 geschilderten wissenschaftstheoretischen Prämissen und die sich daraus ableitende Notwendigkeit, schon *vor* der Operationaliserung, der Datenerhebung und -auswertung die Forschungshypothesen und die Forschungsfragen detailliert zu spezifizieren sowie die zeitliche Planung der Erhebung relativ großer und repräsentativer Stichproben bringt es mit sich, dass sich bei quantitativ empirischen Studien ein relativ striktes „Neun-Phasen-Modell“ des Forschungsprozesses etabliert hat (vgl. Döring/Bortz 2016, 24f.), wenngleich auch andere Modelle existieren.

Tab. 2: (Sequentieller) Ablauf eines quantitativ-empirischen Forschungsprojekts

Hauser/Humpert 2009 **Fünf Elemente**	Krauss/Bruckmaier et al. 2015 **Schritte des Forschungsprozesses**	Döring/Bortz 2016 **Neun-Phasen-Modell**
1. Literaturrecherche (Aufarbeitung theoretischer und empirischer Hintergrund)	1. Formulierung von Fragestellung und Hypothesen	1. Forschungsthema und Forschungsproblem
2. Fragestellung und Hypothesen	[Forschungsbericht: Theorie]	2. Forschungsstand und theoretischer Hintergrund
3. Methode der Datenerhebung und Bestimmung der statistischen Verfahren	2. Entwurf des Designs (Studienlage, Gesamtplan)	3. Untersuchungsdesign
	3. Konstruktion der Untersuchungsinstrumente	4. Operationalisierung
		5. Stichprobenziehung
4. Darstellung der Ergebnisse	4. Wahl der Stichprobe	6. Datenerhebung
		7. Datenaufbereitung
5. Interpretation (Hypothesen, Gültigkeitsbereich, Alternativerklärungen, selbstkritische Hinweise, Weiterentwicklung der Theorie und weitere Forschung)	5. Statistische Analysen (Datenaufbereitung, deskriptive Statistik, Reliabilitätsanalyse, Unterschiedshypothesen, Zusammenhangshypothesen)	8. Datenanalyse
	[Im Forschungsbericht: Diskussion]	9. Ergebnispräsentation

Tabelle 2 stellt zur besseren Orientierung und Anregung eigener Anpassungen eine Synopse dreier Darstellungen mit steigendem Anforderungsgrad dar: a) eine an Lehrkräfte adressierte einführende Darstellung (Hauser/Humpert 2009), eine

an Studierende der Mathematikdidaktik gerichtete Spezifizierung (Krauss/Bruckmaier/Schmeisser/Brunner 2015) und ein Auszug aus einem umfassenden Standardwerk für Psychologiestudierende (Döring/Bortz 2016).

Alle drei Pläne unterscheiden sich geringfügig bezüglich der Segmentierung in einzelne Phasen, jedoch nicht wesentlich in Bezug auf die Sequentialität der einzelnen Schritte. Lediglich Hauser/Humpert beginnen mit der Literaturrecherche vor der Spezifizierung der Fragestellung und der Hypothesenformulierung, doch dürfte auch in den anderen Modellen sicherlich Phase 1 und 2 rekursiv ineinander greifen, denn eine Formulierung von Forschungshypothesen ohne Konsultation der Forschungsliteratur erscheint nicht sinnvoll. Letztlich gilt im Forschungsprozess Folgendes als unumstößliche Norm (vgl. Döring/Bortz 2016, 23f.):

1. Die standardisierten Testinstrumente (z.B. der Lesetest, der Fragebogen zur Lesemotivation etc.) werden während der Datenerhebung nicht mehr verändert.
2. Der vorhandene Datensatz bzw. die gezogene Stichprobe wird während der Datenanalyse nicht mehr ergänzt bzw. angepasst. (Gerade hier begründet sich die Notwendigkeit zur gründlichen theoretischen Vorarbeit und Pilotierung der Operationalisierung in der quantitativen Forschung).
3. Eine Ergänzung, die sich aus der sogenannten *Replikationskrise* der Psychologie (vgl. Abschnitt 3 zur Kritik) ergibt, ist das strikte Unterlassen des sogenannten *HARKing – hypothesizing after results are known* (vgl. Kerr 1998). Die „vermutlich weite Verbreitung“ (Scharlau 2018, 113) dieser schädlichen Praktiken (inklusive p-hacking) wird im Meta-Diskurs sehr kritisch und als disziplinschädigend gesehen (vgl. Abschnitt 3).

Mit Blick auf Tabelle 3 ist nun festzustellen, dass sich das quantitative Paradigma insofern streng vom qualitativen unterscheidet, als dort die Hypothesen- und Theoriebildung in vielen Paradigmen erst als vorletzter Schritt vor der Ergebnispräsentation und somit nach dem iterativen Zyklus Stichprobenziehung, Datenerhebung, Datenaufbereitung, Datenanalyse erfolgt (vgl. Döring/Bortz 2016, 27).

2. Forschungsfelder quantitativer fachdidaktischer Forschung und einige Begriffsklärungen

Die Fachdidaktik Deutsch hat – wie alle Fachdidaktiken – in Bezug auf die empirisch ausgerichtete Forschung mit einer Reihe an (potentiellen) Bezugsdisziplinen und Forschungsfeldern zu tun. Waren Fachwissenschaft, Pädagogik, Psychologie und das Praxisfeld schon immer Partner der Fachdidaktik, so hat sich in den letzten Jahren eindeutig eine stärkere Einbindung (eines Teils der) Fachdidaktik in die empirische Bildungsforschung mit ihren Subdisziplinen (wie der Lehr-Lern-Forschung) ergeben. Dies erkennt man unter anderem an der steigenden Zahl fachdidaktischer Beiträge auf den Tagungen der Gesellschaft für Empirische

Bildungsforschung (GEBF). Insgesamt hat sich das Feld der empirischen Bildungsforschung seit der Gründung der noch jungen GEBF im Jahr 2012 dynamisch entwickelt.

Die quantitative empirische Bildungsforschung wiederum, an der die Fachdidaktik partizipiert, hat durch die bildungspolitische Steuerung der letzten Jahre eine starke finanzielle und ideelle Förderung durch eine Vielzahl drittmittelstarker Förderprogramme erfahren, die auf das „Bedürfnis der Politik nach wissenschaftlicher Evidenz für Bildungsentscheidungen" (Leuders 2015, 223) zurückgeführt werden kann (vgl. auch Baumert/Tillmann 2016; Klieme/Rakoczy 2008; Köller 2014; Prediger et al. 2016; Reinders et al. 2015; Senn/Krelle 2016).

2.1 Quantitativ ausgerichtete deutschdidaktische Forschung

Ein Schwerpunkt der quantitativen empirischen Bildungsforschung liegt derzeit sicherlich auf dem Bildungsmonitoring (**large-scale assessment** wie PISA), in das die Deutschdidaktik zunehmend stärker eingebunden wird – an der ersten Konzeption der PISA-Instrumente und an den einschlägigen Veröffentlichungen zu den Bildungsstandards war sie noch nicht beteiligt.

Neben den Schulleistungsstudien ist das große Feld der **Lehr-Lern-Forschung** sicherlich dasjenige, das aus Sicht der Deutschdidaktik eine intensive Beschäftigung mit quantitativer Methodik erfordert. Hier zeigt sich die Stärke der fachdidaktischen Forschung besonders in *domänenspezifischen* Akzentuierungen wie sie beispielsweise die empirische Schreibforschung der dies-Gruppe (Becker-Mrotzek/Tillmann 2016) oder anderer Domänen (Lesekompetenzforschung) aufweist. Vor dem Hintergrund der in Abschnitt 1 skizzierten Voraussetzung der Theoriebildung kommt der Fachdidaktik dabei auch eine besondere Rolle als Lieferant von substanziellen (inhaltsgesättigten) Theorien oder sogenannten „lokalen Theorien" zu, ein Begriff,

> der ausdrückt, dass fachdidaktische Theorien einen größeren Lokalisierungsgrad besitzen und verschiedene Formen der Kontextspezifität explizit berücksichtigen [...]. Die besondere Stärke fachdidaktischer Forschung ist dabei die substanzielle Nutzung der fachlichen Strukturen und der psychischen Strukturen der Lernenden. (Leuders 2015, 225)

In diesem Kontext kann man auch in den letzten Jahren eine Mehrung der Evidenz durch eine Vielzahl an **Wirksamkeitsstudien** erkennen, die zum Teil wiederum in neuen Interventionen münden (zur Methodik einer Wirksamkeitsstudie vgl. Pissarek/Wild in diesem Band).

Die **Entwicklung** neuer **Testinstrumente** steht zum einen in Zusammenhang mit der Lehr-Lern-Forschung und den Wirksamkeitsstudien, wird aber durch den Bedarf nach besserer und passgenauerer Diagnostik zum Teil auch durch nicht-universitäre Institute (wie dem BIFIE in Salzburg) vorangetrieben. Hier kommen die zuvor skizzierten Verfahren der Konstruktvalidierung und der Fragebogen- und Testkonstruktion zum Tragen.

Ein viertes, derzeit sehr agiles Forschungsfeld ist jenes der deutschdidaktischen und **domänenspezifischen Professionalitätsforschung**, dem das SDD 2018 in Hamburg gewidmet war. Hier war TEDS-LT (Blömeke et al. 2011; 2013) das erste deutschdidaktische Projekt, das in der Professionalitätsforschung mit elaborierten quantitativen Forschungsmethoden eine repräsentative Stichprobe erhoben, und dabei auch mit der probabilistischen Testtheorie gearbeitet hat.

2.2 Designs, Stichproben und Auswertungsverfahren

Es versteht sich von selbst, dass hier nicht in gängige Designs und Auswertungsverfahren der quantitativen Forschung eingeführt werden kann, jedoch sollen ein paar Grundbegriffe genannt werden.

Neben der Beobachtung (z.B. Beobachtungsraster und Analyse einer Unterrichtsstunde in Bezug auf verschiedene Aktivitätsmuster) und dem Fragebogen (z.B. dem Persönlichkeitstest) unterscheidet man bei den gängigen Designs der quantitativ-empirischen Forschung zwischen *experimentellen* und *nicht-experimentellen*. Bei einem *idealen* Experiment wird die unabhängige Variable sehr genau kontrolliert und manipuliert, was es von den nicht-experimentellen Studien unterscheidet (zu einer anschaulichen Differenzierung vgl. Krauss et al. 2015, 620-622).

Die Datenanalyse kann mit Hilfe rein *deskriptiver* oder *schließender* Statistik (siehe hierzu auch den Beitrag von Schmitz in diesem Band) erfolgen. Bei der deskriptiven statistischen Analyse werden ‚lediglich' die erfassten Daten beschrieben bzw. visualisiert, z.B. in Form von Balken- oder Kreisdiagrammen (z.B. die Verteilung der Mädchen und Jungen in der Stichprobe oder die Zahl der Zweitspracheschülerinnen und -schüler im Verhältnis zur Stichprobe). Hinzu kommen die Varianz und die Standardabweichung als Streuungsmaße bzw. die Maße der zentralen Tendenz (wie Medien, Modalwert und arithmetisches Mittel).

Wie stark zwei Variablen zusammenhängen (z.B. als **Zusammenhangshypothese**: *Lesekompetenz* und *Lesemotivation* weisen eine mittlere Korrelation auf) kann durch den *Korrelationskoeffizienten r* berechnet werden, in diesem Fall als bivariates lineares Zusammenhangsmaß. Auf diese Art kann man die Zusammenhangshypothesen testen. r kann dabei einen Wert zwischen -1 und +1 einnehmen, wobei bei positivem Vorzeichen ein proportionaler Zusammenhang besteht, und je höher der Wert ausfällt (bis 1), desto stärker ist der Zusammenhang zweier Variablen.

Unterschiedshypothesen können hingegen mit dem t-Test (jedoch nur bei intervallskalierten Daten, s.o.) getestet werden. Er gibt Auskunft über den Unterschied zwischen einer Stichprobe und der Population, oder aber auch darüber, ob sich die Mittelwerte zweier Stichproben (z.B. Lesekompetenz der Erst- und Zweitsprache-Schülerinnen und Schüler) unterscheiden. Hier gibt es den Einstichproben-t-Test und den Zweistichproben-t-Test, deren Beschreibung hier den Rahmen

sprengen würde (vgl. Hauser/Humpert 2009, 138-151 für eine knappe und anschauliche Darstellung, für die Hypothesentestung auf statistischem Niveau vgl. Bühner/Ziegler 2009, 141-176).

Bei der **Stichprobenziehung** gibt es im Wesentlichen zwei qualitativ unterscheidbare Gruppen: die zufallsgesteuerten Stichproben und die nicht zufallsgesteuerten Stichproben. Von den zufallsgesteuerten Stichproben wird die *einfache Zufallsstichprobe* oder echte Randomisierung eher selten in schulbezogenen Projekten erreicht werden können (es sei denn bei zentral gesteuerten large-scale assessments mit entsprechender administrativer Unterstützung), häufiger wird die *geschichtete Stichprobe* (z.B. die Aufteilung der Grundgesamtheit nach den Schularten Mittelschule, Realschule, Gymnasium) oder die *Klumpenstichprobe* (z.B. Schulen oder Klassen als Einheiten) vorkommen. Letztere wird vermutlich in der deutschdidaktischen Forschung den häufigsten Fall einer Stichprobenziehung darstellen, denn gerade in Interventionsstudien ist man doch häufig auf kooperierende Schulleiterinnen und Schulleiter oder Lehrerinnen und Lehrer angewiesen, die einem organisatorisch zur Seite stehen und Zugang zu den Probandinnen und Probanden verschaffen (Grabowski geht auf die für Forschungszwecke in der Regel nicht manipulierbare personelle Zusammensetzung von Schulklassen ein und empfiehlt in diesem Zusammenhang das Inbetrachtziehen von quasi-experimentellen Designs, vgl. 2016, 323).

2.3 Was bedeutet nochmal „signifikant“?

Am 6. Dezember 2016 trug ein Beitrag der Journalistin Marlene Weiß zum gerade frischen PISA-Ranking diesen Titel. Es ist höchst selten, dass der populärwissenschaftliche Diskurs sich darum kümmert, wissenschaftliche Konstrukte zu klären. In diesem Fall wird das kleine Wort „**signifikant**“ – auch im fachdidaktischen Diskurs – häufig im Sinne von „bedeutungsvoll“ rezipiert und verstanden und ihm dadurch eigentlich zu viel Bedeutung beigemessen. Relevanter wäre es, das System der Hypothesentestung zu erklären, doch selbst Einführungsbände wie Bühner/Ziegler brauchen alleine dafür ganze 36 Seiten (2009, 141-176). Der SZ-Artikel bietet als Erklärung an:

> Die Definition dafür ist etwas umständlich: Eine Differenz zwischen zwei Werten ist dann signifikant, wenn sie in höchstens fünf Prozent der Fälle auch ohne einen realen Unterschied zu erwarten wäre, aufgrund statistischer Ungenauigkeit. (Weiß 2016)

Man kann hier also von einem **wahrscheinlichkeitstheoretischen** bzw. **probabilistischen** Erklärungsansatz sprechen (im Gegensatz zu einem deterministischen, vgl. Döring/Bortz 2016, 50).

Wenn wir es dabei *be-* und die Variante des Neyman-Pearson-Signifikanztests der privaten Lektüre *über*lassen, so wäre doch in Medienberichten die Frage nach den gefundenen **Effektstärken** die für die Interpretation der Ergebnisse wesentlich interessantere. Die Effektstärke d bildet als ein intuitiv interpretierbares Maß den Mittelwertunterschied zweier Gruppen ab (z.B. Mädchen und Jungen), wobei der

Gruppenunterschied an der Standartabweichung SD der Gesamtstichprobe relativiert wird. Die Effektstärke d ist somit unabhängig von absoluten Mittelwerten und Streuungen und ermöglicht eine direkte Beurteilung der Größe des Unterschieds zweier Gruppen (vgl. Krauss et al. 2015). Sie ist sehr leicht zu interpretieren (d=0,20 ist ein kleiner, d=0,50 ist ein mittlerer und d=0,80 ist ein großer Effekt) und ermöglicht es aufgrund des Rechenverfahrens auch die Ergebnisse verschiedener Studien zu vergleichen (z.B. „Wieviel lernen Schülerinnen und Schüler einer 5. Jahrgangsstufe beim Lesen hinzu?“ vs. „Wie wirksam ist ein Lesetraining von 3 Wochen?“). Für Laien wäre dieser Wert jedenfalls weit aussagekräftiger als der Signifikanztest, der bei großen Stichproben ohnehin schlicht durch die Gruppengröße relativ schnell signifikante Werte liefert. Er wird deshalb auch in den empirischen Bildungs- und Sozialwissenschaften „zunehmend kritisch beurteilt“ (vgl. Krauss et al. 2015, 632).

> Der Rückschluss vom Ergebnis eines statistischen Hypothesentests (signifikantes versus nicht-signifikantes Ergebnis) auf die zu prüfende Theorie ist definitionsgemäß mit einem statistischen Fehler behaftet. Er ist darüber hinaus auf theoretischer Ebene unsicher und muss kritisch diskutiert werden, da die Aussagekraft der Daten von der Gültigkeit diverser Hilfs- und Instrumententheorien abhängt, auf denen die Methodenentscheidungen im Forschungsprozess (z.B. Operationalisierung, Stichprobenauswahl) basieren. (Döring/Bortz 2016, 50)

3. Kritikpunkte am und Einschränkungen des quantitativen Paradigmas

Auch wenn der Paradigmenstreit im Fachdiskurs bisweilen überwunden scheint, gibt es dennoch Kritik und Vorbehalte gegenüber dem quantitativen Paradigma.

Stärken, die für das quantitative Paradigma immer wieder ins Feld geführt werden, sind die Generalisierbarkeit der mit quantitativer Methodik gefundenen Ergebnisse, die Identifizierung kausal wirkender Variablen, die gute Auswertbarkeit und Übertragbarkeit der Ergebnisse. Aufgrund der testtheoretischen Gestaltung können z.B. die Effektstärken verschiedener Interventionen verglichen und Erkenntnisse für wirksame Lernszenarien gewonnen werden – auch Metastudien (siehe den Beitrag von Philipp in diesem Band)sind möglich, deren Aufklärungsraum groß ist. Befürworterinnen und Befürworter sehen zudem die Standardisierung und die konsequente Verwendung statistischer Verfahren als großen Vorteil, da so vergleichbare Daten generiert werden können, die wiederum Voraussetzung für das Schaffen von Bildungsgerechtigkeit wären. So ist es z.B. mit Hilfe der OECD-Studien möglich geworden, die Benachteiligung sozial weniger privilegierter Schichten im Bildungssystem deutlich sichtbar zu machen und eine überzeugende Evidenz aufzubauen, die wiederum Grundlage für Bildungsreformen sein könnte, die dringend nötig wären. Zudem ermöglicht diese Messmethodik (wie PISA 2000 eindrucksvoll gezeigt hat), die Annahmen von der Qualität unserer Schulen bzw. unseres Unterrichts objektiv zu prüfen – dass in Deutschland und Österreich so viele ‚Risikoschüler‘ als 15-Jährige die Schullaufbahn verlas-

sen, hätte vor PISA 2000 kaum jemand angenommen. Für Monitoring und Trendstudien bleibt fast keine andere reliable Möglichkeit als quantitativ-empirische Erhebungen. Döring und Bortz nennen darüber hinaus noch die praktische Anwendbarkeit, die große Einfachheit bzw. Sparsamkeit und den hohen empirischen Bewährungsgrad als Vorteile dieses Ansatzes (vgl. Döring/Bortz 2016, 56f.).

Und doch gibt es berechtigte Kritik an einigen Studien im quantitativen Paradigma, die zunächst eine wahrgenommene Verselbstständigung der Methoden und einen mangelnden Reflexionsgrad der Passung an die Forschungsfrage anführen. Und selbst Döring und Bortz merken an, dass in der Forschungspraxis häufig mit Hypothesen bzw. Theoremen gearbeitet wird, „die wissenschaftstheoretisch nicht den Status einer vollwertigen Theorie haben" (Döring/Bortz 2016, 57). Um übliche Kritiklinien hier summativ wiederzugeben, sei neben dem unreflektiertem Einsatz existierender Instrumente der hohe Grad an Abstraktion und Kategorisierung, der die Probandinnen und Probanden in formulierte Sachverhalte zwänge (vgl. Lamnek 2005, 337) genannt. Auch begegnen einem bisweilen in der „Kongresspraxis" Vorbehalte dahingehend, dass Messinstrumente hinterfragt werden, weil sie nicht das gesamte zu messende Konstrukt abbilden – was ja wiederum in der Messtheorie auch nicht begründet wäre.

Eine veritable Krise, die bis heute andauert, entstand durch die sogenannte ***Replikationskrise*** in den Sozialwissenschaften, die bis in die empirische Bildungsforschung ausstrahlt. Im Jahr 2014 veröffentlichte die Zeitschrift *Lancet* – für die Biomedizin – eine Serie zur Replizierbarkeit von Forschungsergebnissen, mit dem besorgniserregenden Ergebnis, dass sich tatsächlich nur ca. ein Viertel der präklinischen Studien, die in der renommierten Zeitschrift erschienen waren, replizieren ließ. Die Replikation gilt dabei als ein sehr wichtiges Verfahren, um quantitativ-empirisch ausgerichtete Studien und ihre experimentalwissenschaftlich und quantitativ begründeten empirischen Wissensansprüche zu bestätigen. Die DFG reagierte bald auf diese Debatte mit einer Stellungnahme zur Replizierbarkeit von Forschungsergebnissen, die sich darum bemüht, den Wirkungsgrad von Replikation einzuschränken: „Replizierbarkeit ist kein generelles Kriterium wissenschaftlicher Erkenntnis. [...] Nicht-Replizierbarkeit ist kein genereller Falsifikationsbeweis." (DFG 2017, 2). Eine Gruppe um Felix Schönbrodt (LMU München) unterbreitete daraufhin Vorschläge zur nachhaltigen Qualitätssicherung quantitativ orientierter Studien, die eine umfassende Konkretisierung der DFG-Leitlinien darstellen. Sie sehen vor allem die Zugänglichkeit, Persistenz und Identifizierbarkeit der Daten vor, die auch die Nachnutzung der Daten durch andere Forscherinnen und Forscher gewährleisten soll (Schönbrodt/Gollwitzer/Abele-Brehm 2017).

Dieser Diskurs lässt sich an die eingangs skizzierte wissenschaftstheoretische Basis zurückkoppeln:

> Wenn Zeitschriften die Veröffentlichung von Replikationsstudien – d.h. von Wiederholungsstudien – ablehnen, weil diese nicht ‚originell' genug seien, dann spricht daraus eine erkenntnistheoretische Position, der gemäß die strenge Prü-

> fung vorhandener Theorien weniger bedeutsam ist als das Generieren neuer Theorien. Wenn bei der Beurteilung von Forschungsanträgen oder Doktorarbeiten methodische Strenge und die Orientierung an etablierten wissenschaftlichen Methoden und Methodologien gefordert wird, steht dahinter offenbar die wissenschaftstheoretische Position, dass Erkenntnisgewinn durch die Nutzung eines etablieren Methodeninstrumentariums gefördert und nicht behindert wird. (Döring/Bortz 2016, 36)

Es gehört also zur konsequenten Umsetzung des quantitativen Paradigmas, dass die Strenge des Verfahrens, das in Abschnitt 2 beschrieben wurde, eingehalten wird.

Die oben skizzierte Überbewertung bzw. Fehleinschätzung der „p-values", also der Signifikanzwerte, hat auch zu spezifischem Fehlverhalten in der Forschungspraxis geführt, das als *HARKing* (hypothesizing after results are known) und *p-Hacking* bezeichnet wird. Beides stellt elementare Verstöße gegenüber den in Abschnitt 1 vorgestellten Prämissen dar. HARKing meint dabei den Umstand, dass *nach*dem die Daten ausgewertet sind, manche Forscherinnen und Forscher noch einmal zu Schritt 1/2 im Forschungsprozess (vgl. Abschnitt 1.3) zurückkehren, um neue Hypothesen zu generieren, um solche zu finden, die „signifikant" und damit berichtenswert werden. In die gleiche Richtung zielt das p-Hacking, das verschiedene Manipulationsstrategien am Datensatz umfasst (Löschen einzelner Daten, Umcodierung, Umskalierung etc.), um doch noch signifikante Ergebnisse zu bekommen. Beides stellt einen fundamentalen Verstoß gegen die oben skizzierten wissenschaftstheoretischen Prämissen des quantitativ-empirischen Ansatzes dar. Die *American Statistical Association* (ASA 2016) hat inzwischen *Principles to Improve the Conduct and Interpretation of Quantitative Science* veröffentlicht, welche die folgenden sechs Prinzipien umfassen:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Gerade Prinzip 5 und 6 lassen sich gut auf die Erklärung zum Begriff *signifikant* oben in Abschnitt 2.3 beziehen. Als eine Möglichkeit, aus der Krise der quantitativ orientierten Forschung herauszufinden, erscheint die standortübergreifende Kooperation. Der offene Zugang zu Daten (inkl. der transparenten Veröffentlichung von Testbatterien und Datensätzen) für die Forschungsgemeinschaft (sogenannte Empiriedatenbanken) und die kooperative, sequentielle Testung bzw. Replikation von Studien unter nahezu identischen Bedingungen an anderen Standorten (sogenannte *daisy chain replications*) wären ein erster Schritt. Auch wäre über den Umgang mit nicht-signifikanten Forschungsergebnissen und die

Praxis, diese nicht zu publizieren, nachzudenken. Dies würde auch den Druck von vielen Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftlern nehmen, signifikante Ergebnisse produzieren zu müssen (und beispielsweise noch einmal nachzuerheben). Noch wirksamer wäre es, für hochrangige Forschungsvorhaben eine *pre-registration* der Forschungshypothesen nach Phase 1/2 des Prozesses vorzusehen (eine Forderung von Daniel Kahnemann), so dass diese nicht mehr post hoc geändert werden können.

Literatur

American Statistical Association (ASA) (2016): American Statistical Association releases statement on statistical significance and p-values. Provides principles to improve the conduct and interpretation of quantitative science. Online unter: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf (letzter Zugriff: 1.8.2018).

Baumert, Jürgen/Brunner, Martin/Lüdkte, Oliver/Trautwein, Ulrich (2007): Was messen internationale Schulleistungsstudien? – Resultate kumulativer Wissenserwerbsprozesse. In: Psychologische Rundschau, 58, 2, 118-145.

Baumert, Jürgen/Tillmann, Klaus-Jürgen (2016) (Hrsg.): Empirische Bildungsforschung. Der kritische Blick und die Antwort auf Kritiker. Zeitschrift für Erziehungswissenschaft 31.

Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hrsg.): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann.

Boelmann, Jan M. (2016) (Hrsg.): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. 2. durchges. Aufl. Baltmannsweiler: Schneider Hohengehren.

Blömeke, Sigrid/Bremerich-Vos, Albert/Haudeck, Helga/Kaiser, Gabriele/Nold, Günter/Schwippert, Kurt/Willenberg, Heiner (2011) (Hrsg.): Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen, Erste Ergebnisse aus TEDS-LT. Münster: Waxmann.

Blömeke, Sigrid/Bremerich-Vos, Albert/Kaiser, Gabriele/Nold, Günter/Haudeck, Helga/Keßler, Jörg-U./Schwippert, Kurt (2013) (Hrsg.): Professionelle Kompetenzen im Studienverlauf. Weitere Ergebnisse zur Deutsch-, Englisch-, und Mathematiklehrerausbildung aus TEDS-LT. Münster: Waxmann.

Bühner, Markus (2011): Einführung in die Test- und Fragebogenkonstruktion. 3. aktual. u. erw. Aufl. München: Pearson.

Bühner, Markus/Ziegler, Matthias (2009): Statistik für Psychologen und Sozialwissenschaftler. München: Pearson.

Deutsche Forschungsgemeinschaft (DFG) (2017): Replizierbarkeit von Forschungsergebnissen. Eine Stellungnahme der Deutschen Forschungsgemeinschaft. Online unter: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_de.pdf (letzter Zugriff: 1.8.2018).

Döring, Nicola/Bortz, Jürgen (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5. volls. überarb., aktual. u. erw. Aufl. Berlin: Springer.

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

Frickel, Daniela A./Kammler, Clemens/Rupp, Gerhard (2012) (Hrsg.): Literaturdidaktik im Zeichen von Kompetenzorientierung und Empirie. Perspektiven und Probleme. Freiburg im Breisgau: Fillibach.

Grabowski, Joachim (2017): Anforderungen an Untersuchungsdesigns. In: Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hrsg.): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann, 315-334.

Hauser, Bernhard/Humpert, Winfried (2009): signifikant? Einführung in statistische Methoden für Lehrkräfte. Zug: Klett Balmer.

Helmke, Andreas (2009): Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts. Wiesbaden: Klett Kallmeyer.

Kelle, Udo (2008): Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte. 2. Aufl. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kerr, Norbert L. (1998): HARKing: Hypothesizing after the results are known. In: Personality and Social Psychology Review, 2, 196-217.

Klieme, Eckhard (2006): Empirische Unterrichtsforschung: aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. In: Zeitschrift für Pädagogik 52, 2006, 6, 765-773.

Klieme, Eckhard/Leutner, Detlev (2006): Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. In: Zeitschrift für Pädagogik 52, 6, 876-903.

Klieme, Eckhard/Rakoczy, Katrin (2008): Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. In: Zeitschrift für Pädagogik 54, 2, 222-237.

Köller, Olaf (2014): Entwicklung und Erträge der jüngeren empirischen Bildungsforschung. In: Zeitschrift für Pädagogik, 60. Beiheft, 102–122.

Krauss, Stefan/Bruckmaier, Georg/Schmeisser, Christine/Brunner, Martin (2015): Quantitative Forschungsmethoden in der Mathematikdidaktik. In: Bruder, Regina/Hefendehl-Hebeker, Lisa/Schmidt-Thieme, Barbara/Weigand, Hans-Georg (Hrsg.): Handbuch der Mathematikdidaktik. Wiesbaden: Springer VS, 613-641.

Lamnek, Siegfried (2005). Qualitative Sozialforschung. 4., vollst. überarb. Aufl. Weinheim: Beltz.

Leuders, Timo (2015): Empirische Forschung in der Fachdidaktik. Eine Herausforderung für die Professionalisierung und die Nachwuchsqualifizierung. In: Beiträge zur Lehrerinnen- und Lehrerausbildung 33, 215-234.

Leutner, Detlev/Fleischer, Jens/Grünkorn, Juliane/Klieme, Eckhard (2017): Competence Assessment in Education. An Introduction. In: Leutner, Detlev/Fleischer, Jens/Grünkorn, Juliane/Klieme, Eckhard (Hrsg.): Competence Assessment in Education. Research Models and Instruments. Cham (CH): Springer International Publishing, 1-8.

Meinefeld, Werner (1995): Realität und Konstruktion: erkenntnistheoretische Grundlagen einer Methodologie der empirischen Sozialforschung. Opladen: Leske Budrich.

Popper, Karl Raimund (1989): Logik der Forschung. 9. Aufl. Tübingen: Mohr Siebeck.

Pissarek, Markus/Schilcher, Anita (2017): FALKO-D: Die Untersuchung des Professionswissens von Deutschlehrenden. Entwicklung eines Messinstruments zur fachspezifischen Lehrerkompetenz und Ergebnisse zu dessen Validierung. In: Krauss, Stefan/Lindl, Alfred/Schilcher, Anita/Fricke, Michael/Göhring, Anja /Hofmann, Bernhard/Kirchhoff, Petra/Mulder, Regina H. (Hrsg.): FALKO: Fachspezifische Lehrerkompetenzen. Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik. Münster: Waxmann, 67-112.

Pissarek, Markus/Pronold-Günthner, Friederike (2018): Lernvoraussetzungen ermitteln am Beispiel der Lese- und Rechtschreibkompetenz. In: Schilcher, Anita/Finkenzeller, Kurt/Knott, Christina/Pronold-Günthner, Friederike/Wild, Johannes (Hrsg.): Schritt für Schritt zum guten Deutschunterricht. Praxisbuch für Studium und Referendariat: Strategien und Methoden für professionelle Deutschlehrkräfte. Seelze: Klett Kallmeyer, 119-132.

Pissarek, Markus (2018): Zum Begriff der Lesekompetenz – Förderung von Lesekompetenz bei jungen Erwachsenen. In: Resinger, Paul/Volgger, Angela (Hrsg.): Förderung der Lesekompetenz von Lehrlingen. Innsbruck: Studien Verlag, 5-14.

Prediger, Susanne/Parchman, Ilka/Hammann, Marcus/Frederking, Volker (2016): Unterrichtsqualität braucht Fachlichkeit – Zur Bedeutung fachdidaktischer Grundlagen- und Anwendungsforschung als Bindeglied zwischen Forschung und Praxis. In: BMBF (Hrsg.): Bildungsforschung 2020. Zwischen wissenschaftlicher Exzellenz und gesellschaftlicher Verantwortung. Berlin: Bundesministerium für Bildung und Forschung, 405-435.

Reinders, Heinz/Ditton, Hartmut/Gräsel, Cornelia/Gniewosz, Burkhard (2015): Vorwort. In: Reinders, Heinz/Ditton, Hartmut/Gräsel, Cornelia/Gniewosz, Burkhard (Hrsg.): Empirische Bildungsforschung. Gegenstandsbereiche. 2. Aufl. Wiesbaden: Springer VS, 11-18.

Scharlau, Ingrid (2018): Sich verständigen. Überlegungen zur Frage der Evidenzbasierung. In: Jenert, Tobias, Reinmann, Gabi, Schmohl, Tobias (Hrsg.): Hochschulbildungsforschung. Theoretische, methodologische und methodische Denkanstöße für die Hochschuldidaktik. Wiesbaden: Springer VS, 105-123.

Schönbrodt, Felix/Gollwitzer, Mario/Abele-Brehm (2017): Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien. In: Psychologische Rundschau, 20-35.

Senn, Werner/Krelle, Michael (2016): Qualitäten von Unterricht: Empirische Unterrichtsforschung im Fach Deutsch. Stuttgart: Filibach.

Steyer, Rolf/Eid, Michael (1993): Messen und Testen. Berlin: Springer.

Weiß, Marlene (2016): Was bedeutet nochmal „signifikant“? In: Süddeutsche Zeitung vom 6.12.2016. Online unter: http://www.sueddeutsche.de/bildung/pisa-was-bedeutet-nochmal-signifikant-1.3281954 (letzter Zugriff: 01.08.2018).

FREDERIKE SCHMIDT

Gütekriterien für quantitative Forschungsansätze

1. Systematisierung des Gegenstandsfeldes[1]

Eine empirisch forschende Deutschdidaktik hat den Anspruch, Erkenntnisse über Lehr- und Lernprozesse im Deutschunterricht – sowie deren Voraussetzungen und Folgen – gegenstandsbezogen, methodisch kontrolliert und systematisch zu gewinnen. Diese Zielsetzung hat Hermann Helmers, der Begründer der wissenschaftlichen Deutschdidaktik, bereits in den siebziger Jahren betont:

> Das Forschen einer Didaktik der deutschen Sprache wird möglichst von der Unterrichtswirklichkeit ausgehen. In diesem Sinn ist das wissenschaftliche Forschen der Didaktik empirisch zu nennen. […] Die einzelnen Ergebnisse sind jeweils in eine systematische Sicht einzubringen. Dies ist der konstruktive Faktor, den die Didaktik als Forschungselement ebenfalls nötig hat. (Helmers 1971, 29 [Herv. ebd.])

Den Ausführungen von Helmers folgend, kann man die Deutschdidaktik als eine „discipline[ ] of inquiry in education“ (Shulman 1997) bezeichnen. Verbunden mit diesem Anspruch an Erkenntnisgewinnung ist die Frage, wie ‚gute empirische Forschung‘ innerhalb der Disziplin bestimmt werden kann. Dazu haben sich wissenschaftliche Standards herausgebildet, anhand derer die Qualität und Geltung von empirischen Erkenntnissen gemessen und abgesichert wird – die sog. *Gütekriterien*. Die Beschäftigung mit dem Thema Gütekriterien schließt aber nicht nur die Betrachtung forschungsbezogener Maßstäbe, sondern auch den Blick auf Standards für die Publikation von Forschungsergebnissen ein: Empirische Befunde (sowie das diesen zugrunde liegende Datenmaterial) liegen zunächst nur den Forschenden selbst vor. Diesen exklusiven Zugang gilt es aufzuheben, wenn man anstrebt, Erkenntnisse in die wissenschaftliche Community einzubringen. Denn nur wenn Forschungsbefunde für die Diskussion zur Verfügung stehen, können Sie auch zur angesprochenen Erkenntnisgewinnung in der Deutsch-

[1] Kapitel 1 dieses Handbuchbeitrags ist in weiten Teilen bewusst zum entsprechenden Kapitel im Beitrag *Gütekriterien für qualitative Forschungsansätze* in diesem Band parallelisiert, da hier allgemein gültige bzw. bedeutsame Aspekte für das Gegenstandsfeld *Gütekriterien in der empirischen Forschung* besprochen werden.

didaktik beitragen.[2] Anders gewendet: Wer in der Deutschdidaktik empirisch forschen möchte, muss sich notwendigerweise auch mit Fragen der Güte und Geltung in der Forschungs- *und* Publikationspraxis beschäftigen.

Diese Überlegungen bilden den Ausgangspunkt für den vorliegenden Beitrag. Anliegen ist ein Überblick zu Gütekriterien, deren Beachtung zentral für die Durchführung und Publikation von quantitativen Forschungsvorhaben ist. Dazu erfolgt zunächst eine Darstellung grundlegender Maßstäbe für empirische Forschung, um darauf aufbauend den Schwerpunkt auf die Auseinandersetzung mit Standards für quantitative Forschungsansätze zu legen.

1.1 Definition und Funktion von Gütekriterien

Empirische Forschung sollte sich an Standards orientieren, um Güte und Geltungsanspruch von empirischen Befunden zu gewährleisten. In der einschlägigen Literatur werden diese explizit formulierten Maßstäbe zumeist unter dem Begriff *Gütekriterien* diskutiert. Ganz allgemein sind Gütekriterien „normative theoretische Konstrukte“ (Schmelter 2014, 33), die festlegen, wie und welche Wege im Forschungsprozess beschritten werden sollen, um zu empirischen Erkenntnissen zu gelangen. In Artikeln oder Handbüchern werden Gütekriterien wiederum bewusst sehr allgemein gehalten. Insofern müssen die Standards „in einem Forschungsvorhaben individuell angepasst werden“ (Schmelter 2014, 33). Die Beurteilung, ob die etablierten Maßstäbe in einer Studie eingehalten wurden, ist wiederum abhängig vom Untersuchungsbereich und dem gewählten Gegenstandsfeld für das Forschungsvorhaben (Dörnyei 2007, 48f.). Folglich stehen empirisch Forschende vor einer doppelten Aufgabe: Sie müssen erstens entscheiden, an welchen Gütekriterien sie sich orientieren und zweitens eigenständig bestimmen, wie sie diese jeweils untersuchungsspezifisch konkretisieren.

Im Kontext der Diskussion von Forschungsstandards wird häufig auch der Begriff *Qualitätsstandard* verwendet (siehe auch Kapitel 1.3). Diese legen fest, „welche Ausprägung die Qualitätsindikatoren [in einer empirischen Studie] jeweils mindestens habe müssen“ (Döring/Bortz 2016, 83). Es geht also im Kern um die Frage, worin sich eine gute von einer weniger guten Anwendung von Gütekriterien unterscheiden lässt. In vorliegenden Publikationen zeigen sich allerdings begriffliche Unschärfen: Die Differenz zwischen *Gütekriterien* und *Qualitätsstandards* im Rahmen empirischer Forschung wird nicht immer klar markiert, teilweise werden die Termini sogar synonym verwendet.

Grundlegend lassen sich verschiedene Funktionen von Gütekriterien bestimmen. Sie sind

- *Prinzipien,* die als Orientierungsrahmen zur Planung und Durchführung von Forschungsprojekten dienen. Für Forschende sind Gütekriterien folglich

[2] Flick (2010, 403) nennt neben „Forschungspraxis“, „Forschungsbewertung“ und „Publikationspraxis“ noch „Antragstellung“ und „Lehre“ als Eckpfeiler für die Diskussion von Gütekriterien. Die beiden letztgenannten Aspekte werden hier – mit Blick auf den Schwerpunkt dieses Grundlagenbandes – nicht vertiefend erörtert.

Zielvorgaben bzw. Prüfsteine, die zeigen, wie Wege in der empirischen Forschung beschritten werden müssen, um zu wissenschaftlichen Aussagen – in der Deutschdidaktik also über das Lehren und Lernen im Deutschunterricht – zu gelangen;

- *Bewertungsmaßstäbe* für die (nachträgliche) Bestimmung von Güte und Geltung empirisch gewonnener Befunde; meist beziehen sich die Standards auf den Forschungsprozess. Geprüft wird somit „die Qualität des Weges zur wissenschaftlichen Erkenntnisgewinnung durch bestimmte Methoden [mit Blick auf fachliche Ansprüche und Zielsetzungen]“ (Lamnek 2010, 127);
- *Strukturierungselemente* für die Darstellung der gewonnenen Forschungsergebnisse in Publikationen, da sie als Bausteine bzw. zu berücksichtigende Textelemente im Schreibprozess Anwendung finden (sollten);
- *Kommunikationsmittel* für den (inter-)disziplinären Austausch. Gütekriterien sind notwendig, um abzusichern, dass empirische Forschung „so kommuniziert wird, dass alle Beteiligten und Angesprochenen von den Ergebnissen profitieren können“ (Riemer 2011, 199, zitiert nach Schmelter 2014, 35); dies gilt insbesondere auch für die Kommunikation bzw. Vermittlung zwischen Vertretern von unterschiedlichen Forschungsparadigmen;
- *Beitrag zur gesellschaftlichen Akzeptanz* einer Disziplin, da durch disziplinübergreifende Maßstäbe einer Beliebigkeit in empirischen Forschungsprojekten entgegengewirkt wird. Auf diese Weise erhalten Ergebnisse eine stärkere Bedeutsamkeit und können auch außerhalb der eigenen wissenschaftlichen Community Anerkennung finden (dazu u.a. Ludwig 2012, 82f.).

1.2 Übergreifende Gütekriterien für die Forschungs- und Publikationspraxis

Quantitative und qualitative Forschung sowie Mixed-Methods-Designs (siehe die Beiträge von Schieferdecker, Pissarek und Müller in diesem Band) sind Vorgehensweisen innerhalb der empirischen Forschung, die auf verschiedenen Grundannahmen basieren und daher zu unterschiedlichen Erkenntnissen gelangen. Dennoch lassen sich einige Gütekriterien formulieren, die übergreifend für empirische Forschung gelten und somit unabhängig von den gewählten Ansatz zu reflektieren sind. Hinter diesen Maßstäben steht die Haltung, dass empirische Forschungsergebnisse für alle Forschenden in einer Disziplin, unabhängig von ihrem jeweiligen methodologischen Standort, transparent sein sollen und man sich darauf einigt, welcher wissenschaftliche Anspruch an empirische Forschung allgemein als akzeptabel betrachtet wird. Auf dieser Grundlage können folgende Gütekriterien formuliert werden (leicht verändert nach Schmelter 2014, 35):

- *Nachvollziehbarkeit*: Nicht nur die zentralen Erkenntnisse einer empirischen Studie, sondern der Forschungsprozess als Ganzes – d.h. die Datenerhebung, -aufbereitung und -analyse – muss für Dritte einsichtig sein. Dieses Krite-

rium wird in der qualitativen Forschung meist unter dem Terminus *intersubjektive Nachvollziehbarkeit* diskutiert, in quantitativen Forschungsansätzen findet größtenteils der Terminus *Objektivität* Anwendung.

- *Gegenstandsverständnis*: Forschende müssen das Gegenstandsverständnis in einer Untersuchung offenlegen und begründen; gerade und insbesondere um das deutschdidaktische Erkenntnisinteresse in einer Studie zu verdeutlichen. Das Gegenstandsverständnis bildet die Grundlage für die Bewertung des gewählten Ansatzes und der eingesetzten Erhebungs- und Auswertungsverfahren zur Bearbeitung der Forschungsfrage(n). Zugleich ist dieser Standard eine Basis für die Bewertung des Praxisbezuges und die Anschlussfähigkeit einer Studie (s.u.).
- *Anschlussfähigkeit an den Forschungsdiskurs*: Es muss erkennbar sein, an welche Theorien und Forschungsergebnisse eine Studie anknüpft und welchen innovativen Charakter sie für den bestehenden Forschungsdiskurs – hier: die Deutschdidaktik – besitzt.
- *Praxisbezug*[3]*:* Gemeinhin sollen Forschungsergebnisse gesellschaftliche Relevanz aufweisen. Dieses übergreifend formulierte Kriterium ist anschlussfähig an die Zielperspektive der Deutschdidaktik, zur Verbesserung und Weiterentwicklung des Lehrens und Lernens im Deutschunterricht beizutragen. Daher ist ein weiterer Ausweis von Güte, dass empirische Erkenntnisse nicht nur für die Forschung, sondern auch für den professionellen Alltag des Deutschunterrichts bedeutsam sind.
- *Einhaltung forschungsethischer Standards*: Empirische Forschung sollte ethische Standards genügen[4].

In der Zusammenschau wird deutlich, dass sich Gütekriterien auf zwei verschiedene Ebenen beziehen: Einerseits fokussieren die formulierten Maßstäbe die eingesetzten Methoden in einer Studie, andererseits nehmen sie das Untersuchungsdesign als Ganzes in den Blick. Bereits einleitend wurde betont, dass die Darstellung empirischer Ergebnisse eine zentrale Dimension innerhalb des Gegenstandsfeldes *Gütekriterien* ist. Eine gute Orientierung für (Nachwuchs-)Forschende bietet hier der Kriterienkatalog von Elliot/Fischer/Rennie (1999, 220), in dem Publikationsstandards[5] für quantitative und qualitative Forschung bestimmt werden. Die nachfolgend angeführten Maßstäbe wurden von mir jeweils um damit verbundene Reflexionsfragen[6] zur Konkretisierung ergänzt:

---

[3] In der einschlägigen Literatur wird teilweise auch der Terminus *Relevanz* genutzt.

[4] Vertiefend dazu der Beitrag von Bräuer/Vaupel in diesem Band.

[5] Zur Überprüfung der Standards von Elliot/Fischer/Rennie (1999) siehe die Studie von Ilg und Boothe (2010).

[6] Die von mir formulierten Reflexionsfragen erheben nicht den Anspruch auf Vollständigkeit.

- *Explicit scientific context and purpose*
  - Werden die Zielsetzungen der Studie klar benannt?
  - Wird Literatur aus dem Fachdiskurs zur Herleitung der Fragestellung(en) herangezogen?
- *Appropriate methods*
  - Werden die eingesetzten Erhebungs- und Auswertungsverfahren reflektiert?
- *Respect for participants*
  - Wird expliziert, dass ethische Standards in der Studie eingehalten wurden (z.B. durch Anonymisierung der Daten)?
- *Specification of methods*
  - Werden die gewählten Erhebungs- und Auswertungsverfahren angemessen dargestellt?
- *Appropriate discussion*
  - Werden in der Publikation die zentralen Elemente einer Studie diskutiert (z.B. theoretische Setzungen, inhaltlicher Ertrag für die Deutschdidaktik, Methodik, Schlussfolgerungen, Grenzen der Studie, …)?
- *Clarity of presentation*
  - Ist die Studie leserseitig nachvollziehbar dargestellt?
  - Werden die theoretischen Ansätze, die Methodik und die empirischen Ergebnisse angemessen dargestellt?
- *Contribution to knowledge*
  - Wird in der Studie neues empirisches Wissen für die Deutschdidaktik generiert?

1.3 Wissenschaftlichkeit und wissenschaftliche Qualität im Rahmen empirischer Forschung

In Kapitel 1.1 wurde bereits angerissen, dass im Rahmen empirischer Forschung nicht zuletzt zu klären ist, wie sich eine gute von einer weniger guten Anwendung von Gütekriterien unterscheiden lässt. Dies berührt die Frage nach dem wissenschaftlichen Anspruch an Forschungsprojekte, welche in der Debatte zumeist unter dem Begriff *Qualitätsstandards* diskutiert wird. Konkret geht es darum zu klären, welche Mindest-, Regel- und Maximalstandards für empirische Projekte angelegt werden (Döring/Bortz 2016, 82f.). Allgemein und insbesondere in der Deutschdidaktik ist dabei augenfällig, dass Qualitätsstandards bzw. Indikatoren zur Bewertung der Qualität bislang kaum diskutiert werden. Ein Erklärungsansatz mag sein, dass sich Pauschalaussagen hier kaum formulieren lassen, da die Bewertung und Einschätzung der Qualität einer empirischen Studie immer abhängig von Untersuchungsbereich und -gegenstand ist (s.o., Kapitel 1.1). Zumindest einen ersten Zugang zur Thematik bietet eine Übersicht von Döring/Bortz (2016,

91), die aus Perspektive der empirischen Sozialforschung vier allgemein anerkannte Standards der Wissenschaftlichkeit und die zugehörigen Kriterien wissenschaftlicher Qualität zusammenführen:

Die vier Standards der Wissenschaftlichkeit und die zugehörigen vier Kriterien der wissenschaftlichen Qualität im Überblick

Standards der Wissenschaftlichkeit	Kommentar	Kriterien der wissenschaftlichen Qualität	Kommentar
Sie müssen von jeder wissenschaftlichen Studie prinzipiell eingehalten werden.	*Diese Fragen müssen bei einer wissenschaftlichen Studie prinzipiell bejaht werden können.*	*Sie sind bei wissenschaftlichen Studien graduell sehr unterschiedlich ausgeprägt und differenzieren herausragende, gute, durchschnittliche und schwache Studien.*	*Bei diesen Fragen ist anhand von Vergleichsstudien, Referenzwerten aus der Methodenlehre und inhaltlichen Argumenten der Grad der Ausprägung in der jeweiligen wissenschaftlichen Studie abzuschätzen, um ihre Qualität einzustufen.*
1. Wissenschaftliches Forschungsproblem	Bearbeitet die Studie ein Forschungsproblem, das sich in einen anerkannten wissenschaftlichen Forschungs- und Publikationskontext einordnet?	**1. Inhaltliche Relevanz**	In welchem Maße trägt die Studie mit ihren Ergebnissen a) zum wissenschaftlichen Erkenntnisfortschritt (wissenschaftliche/theoretische Relevanz) und/oder b) zur Lösung praktischer Probleme (praktische Relevanz) bei?
2. Wissenschaftlicher Forschungsprozess	Orientiert sich die Studie an etablierten wissenschaftlichen Methodologien und Methoden, die zum Forschungsproblem passen?	**2. Methodische Strenge**	Wie anspruchsvoll sind die gewählten Methodologien und Methoden, wie gut sind sie zur Bearbeitung des Forschungsproblems geeignet, und wie regelkonform werden sie umgesetzt?
3. Wissenschafts- und Forschungsethik	Folgt die Studie den Prinzipien der Wissenschafts- und Forschungsethik?	**3. Ethische Strenge**	Wie konsequent und umfassend werden einzelne Standards der Wissenschafts- und Forschungsethik erfüllt?
4. Dokumentation des Forschungsprojektes	Sind Vorgehen und Ergebnisse der wissenschaftlichen Studie im Detail nachvollziehbar dokumentiert?	**4. Präsentationsqualität**	Wie vollständig, wohlstrukturiert und gut lesbar wird die Studie in ihrem Ablauf und mit ihren Befunden präsentiert und wie umfassend werden die Standards der Berichterstattung des jeweiligen Faches und Publikationsorgans eingehalten?

Abb. 1: Vier Standards der Wissenschaftlichkeit und dazugehörige Kriterien der wissenschaftlichen Qualität (Döring/Bortz 2016, 90)

1.4 Herausforderungen und Desiderate

Eine Systematisierung dazu, *wie* in der Deutschdidaktik empirisch geforscht werden sollte, muss auch in den Blick nehmen, wo sich aktuell noch Blindstellen ausmachen lassen und welche Schwierigkeiten bei der Einhaltung von Gütekriterien zu berücksichtigen sind. In einer ersten Zwischenbilanz sind mindestens drei Herausforderungen bzw. Desiderate erkennbar, die es zu reflektieren gilt:

1. *Fehlende fachdidaktische Profilierung*: Die im Diskurs etablierten Gütekriterien entstammen gemeinhin der Diskussion in den empirischen Sozialwissenschaften. Das bedeutet: Die skizzierten Kriterien sind an Fragestellungen der Psychologie, der Erziehungswissenschaften und der Soziologie orientiert. Für unsere Disziplin ist wiederum wichtig, dass deutschdidaktische Erkenntnisinteressen angemessen berücksichtigt werden, wenn es um die Sicherung von Güte, Geltung und Wissenschaftlichkeit geht. Entgegen der allgemeinen Bedeutsamkeit empirischer Forschung fehlt dazu aber eine qualifizierte fachspezifische Debatte.[7] Es ist daher eine zukünftige wichtige Entwicklungsaufgabe zu klären, inwiefern und wo eigenständige Standards für empirische Forschungsansätze vonseiten der Deutschdidaktik formuliert werden müssen. Für Forschende gilt vor diesem Hintergrund, dass sie Gütekriterien auf ihr Forschungsprojekt abstimmen und die Anwendung der gewählten forschungsmethodologischen Standards für ihre Zwecke offenlegen müssen.[8]
2. *„Benchmarkproblem"*[9]: Flick (2010, 405) charakterisiert mit diesem Begriff den Aspekt, dass innerhalb der empirischen Forschung noch geklärt werden muss, welcher Qualitätsanspruch an empirische Projekte angelegt wird. So ist auch ein in der Deutschdidaktik bislang noch ungeklärter Punkt, was Mindest-, Regel- und Maximalstandards sind, etwa wenn es um Stichprobengrößen im Rahmen qualitativer oder quantitativer Forschung in Promotionsvorhaben geht.[10]
3. *Bedingungen guter Forschung*: Daneben fällt auf, dass in der Debatte häufig ökonomische Gesichtspunkte ausgespart werden. Empirisch Forschende

---

7 Wenngleich hier Entwicklungsarbeit zu leisten ist, so ist in den letzten Jahren immerhin erkennbar, dass zunehmend eine Diskussion um fachspezifische Forschungsmethodik und Untersuchungsdesigns innerhalb der Deutschdidaktik entsteht, wie einige neuere Publikationen dokumentieren (z.B. Boelmann 2016; Becker-Mrotzek 2017).

8 Ein gutes Beispiel für Herausforderungen und den Umgang mit dem Spannungsverhältnis zwischen deutschdidaktischem Erkenntnisinteresse und normativen Setzungen seitens der empirischen Sozialforschung bietet der Beitrag von Winkler/Steinmetz (2016).

9 Flick diskutiert diesen Aspekt für qualitative Forschungsansätze, aus meiner Sicht ist dieser Anspruch aber allgemein auf empirische Forschungsprojekte übertragbar.

10 Die genannten Aspekte sind vermutlich auch nicht pauschal zu klären, eine Spezifizierung von Qualitätsstandards stellt aber eine notwendige Aufgabe für die Deutschdidaktik – gerade auch mit Blick auf Anforderungen an Qualifizierungsarbeiten – dar (allgemein dazu Flick 2010, 404f.).

sollten diese jedoch als weitere Gütekriterien zur Sicherung guter Forschung, der eigenen Ressourcen und nicht zuletzt der eigenen Gesundheit reflektieren. Gerade die Faktoren *Zeit* und *Machbarkeit* (Oswald 2013, 183ff.) sind hier als zentrale Eckpfeiler aufzufassen. Bezogen auf den Faktor *Zeit* ist ein häufiger Fehler, dass der Aufwand für das Auswerten der Daten und das Schreiben nicht bzw. nicht genügend in den Blick genommen wird und diese beiden Prozesse unter erhöhtem Zeitdruck (und oftmals auch in verkürzter Zeit) im letzten Abschnitt eines Forschungsvorhabens vollzogen werden. Eng damit zusammenhängend ist der Faktor *Machbarkeit*. Trotz oder gerade in der Begeisterung für ein Projekt müssen die realistischen und realisierbaren Anforderungen im Blick bleiben, d.h. wie viele Ressourcen überhaupt für die Arbeit an einem Projekt bestehen.

Die genannten Problemfelder und Herausforderungen verdeutlichen, dass zukünftig die „Qualität der Forschung in ihren Prozessen und Ergebnissen umfassender in den Blick zu nehmen [ist]" (Schmelter 2014, 43). Eine solche Debatte stellt nicht zuletzt einen Professionalisierungsanspruch für die Deutschdidaktik dar, der einen Beitrag zu (1) einer Profilierung eines geteilten empirischen Orientierungsrahmens, (2) einer systematischen Evaluation von Studienergebnissen und (3) einem innerdisziplinären Austausch leistet. Gegenwärtig ist es Aufgabe für (Nachwuchs-)Forschende, hier eigene Wege zu gehen. Positiv gewendet stellen diese (notwendigen) Adaptionen einen eigenen Beitrag zum Diskurs dar, der in der Publikation herausgestellt werden kann und sollte. Auf pragmatischer Ebene sollten die vorhergehenden Ausführungen verdeutlicht haben, dass die Einhaltung von Gütekriterien eine gute Planung (Stichwort: Zeit und Machbarkeit, s.o.) erfordert, die Standards aber zugleich einen guten Orientierungsrahmen und Prüfstein für den Forschungsprozess bieten.

2. Welche Gütekriterien gelten für quantitative Forschungsansätze?

Die etablierten Standards zur Bewertung quantitativer Forschungsansätze lassen sich hinsichtlich ihrer Verwendungsweise unterscheiden (Krebs/Menold 2014, 42). Man kann zwischen *Standards für Methoden* (im Hinblick auf deren Zuverlässigkeit und Gültigkeit, Kapitel 2.1) und *Standards für das gesamte Forschungsdesign* (bezogen auf Generalisierbarkeit und Eindeutigkeit der Befunde, Kapitel 2.2) differenzieren. Wichtig zu bedenken ist dabei, dass sich die Gütekriterien übergreifend auf Aussagen beziehen, die auf *Basis* einer Methode bzw. des Forschungsdesigns abgeleitet werden können.

Von diesen Hauptgütekriterien ausgehend bildet eine Reihe von Nebengütekriterien (sog. *weiche Kriterien*) zusätzliche Anforderungen an quantitative Forschung: Praktikabilität, Durchführung, Normierung, Vergleichbarkeit, Ökonomie und Nützlichkeit (vertiefend dazu z.B. Moosbrugger/Kelava 2012, 18-24). Die Nebengütekriterien stellen weitere Bewertungsmaßstäbe zur praktischen Durch-

führung einer quantitativen Studie dar, im Forschungsprozess muss daher der Fokus (zunächst) auf die nachfolgend dargestellten Hauptgütekriterien gerichtet werden. In Bezug auf die Dokumentation und Publikation quantitativer Ergebnisse bieten die in Kapitel 1.2 dargestellten Maßstäbe eine gute Orientierung, weiterführende Überlegungen finden sich bei Albert/Marx (2016, 169-177).

2.1 Gütekriterien für Methoden

Im Hinblick auf Messinstrumente sind die aus der Testtheorie stammenden Gütekriterien *Objektivität, Reliabilität* und *Validität* in der quantitativen Forschung fest etabliert (u.a. Döring/Bortz 26, 93-106; Krebs/Menold 2014, 426-433):

(1) *Objektivität* bezieht sich auf die Unabhängigkeit von Forschungsergebnissen, d.h. dass die Ergebnisse einer Studie unbeeinflusst und hinreichend unabhängig von der Person bzw. den Personen sind, die die Untersuchung durchführt, auswertet und interpretiert. Hinsichtlich der Objektivität einer Studie unterscheidet man daher – orientiert an den einzelnen Forschungsschritten – zwischen folgenden Aspekten:
   - *Durchführungsobjektivität* – gleiches Vorgehen (Standardisierung) bei der Durchführung
   - *Auswertungsobjektivität* – gleiche Ergebnisse verschiedener Auswerter
   - *Interpretationsobjektivität* – gleiche Schlussfolgerungen verschiedener Interpreten

(2) *Reliabilität* fokussiert die Zuverlässigkeit der Aussagen in einer quantitativen Studie und bezieht sich auf die verwendeten Verfahrensweisen und die Bewertung der empirischen Ergebnisse (ungeachtet, welcher Inhalt zu messen ist, siehe Abb. 2). Es geht darum sicherzustellen, dass die Ergebnisse einer Studie keinen Einzelfall darstellen, sondern bei einer Wiederholung (unter gleichen Bedingungen) die gleichen Ergebnisse erzielt werden würden.[11] Im deutschdidaktischen Kontext sind hier etwa gleiche Lernausgangslagen bei Lernenden oder die Anwendung des gleichen Lehrmaterials denkbar, um fundierte Aussagen über die Wirksamkeit einer Innovation machen zu können. Vor diesem Hintergrund wird bzgl. der Reliabilität einer Studie zwischen drei Aspekte differenziert:
   - *Stabilität* – Übereinstimmung der Messergebnisse zu verschiedenen Zeitpunkten
   - *Konsistenz* – Maß, mit dem die zu einem Merkmal gehörenden Items das selbe Merkmal erfassen
   - *Äquivalenz* – Gleichwertigkeit von Messungen

(3) *Validität* wird im Diskurs zumeist als das wichtigste messtheoretische Kriterium bezeichnet. Validität meint die Gültigkeit von Aussagen in einer Untersuchung, d.h. ob man in einer Studie tatsächlich das misst, was gemessen werden soll. Verschiedene Aspekte sind hier zentral:
   - *Inhaltsvalidität* – Items erfassen das Konstrukt in allen wichtigen Aspekten

[11] Dies gilt unabhängig davon, ob das Kriterium *Validität* erfüllt ist, siehe Abb. 2.

- *Kriteriumsvalidität* – Übereinstimmung der gemessenen Befunde mit einem ‚außen stehenden Kriterium' (z.B. Ergebnisse, die anhand eines anderen Verfahrens gewonnen wurden)
- *Konstruktvalidität*[12] – angemessene und nachvollziehbare Operationalisierung und somit Messung des (gesamten) Konstrukts (z.B. Leseflüssigkeit von Lernenden, orthographisches Wissen von Lehrpersonen etc.)

Die drei skizzierten Gütekriterien entstammen der Testtheorie und beziehen sich auf die Methode(n) in einem Forschungsvorhaben. Die Gütekriterien stehen in wechselseitiger Abhängigkeit: so kann es ohne Objektivität keine Reliabilität geben und die Validität basiert wiederum auf den anderen beiden Gütekriterien (Schmelter 2014, 38). Was heißt das konkret für die Durchführung einer empirischen Studie? Ein geeignetes Beispiel, um die angesprochenen Abhängigkeiten zu verdeutlichen, findet sich bei Nerdel (2017, 75)[13]. Nerdel überträgt die Gütekriterien (inhaltliche) Validität und Reliabilität auf eine Zielscheibe, in welcher der Mittelpunkt das angestrebte Konstrukt in einer Untersuchung darstellt. Die Treffer auf der Scheibe bilden verschiedene Varianten ab, wie im Rahmen einer empirischen Untersuchung ein Konstrukt erfasst wird (bzw. eben auch nicht):

Abb. 2: Reliabilität und Validität (Nerdel 2017, 75)

Was dieses Beispiel zeigt: Reliabilität ist eine notwendige, aber noch *keine hinreichende* Bedingung für ein valides Messinstrument in einer Untersuchung.

2.2 Gütekriterien von Forschungsdesigns

Wie bereits einleitend zu diesem Kapitel erläutert, haben sich zwei Verwendungsweisen der Gütekriterien für quantitative Forschungsansätze entwickelt. Neben den etablierten Maßstäben *Objektivität*, *Reliabilität* und *Validität* gibt es noch Gütekriterien, die das Untersuchungsdesign fokussieren (u.a. Döring/Bortz 2016, 99-106; Dörnyei 2007, 52-53; Krebs/Menold 2014, 435f.):

[12] Gerade bei Anfängern wird die Konstruktvalidität häufig unterschätzt und zu schnell bereits die Datenerhebung durchgeführt (Döring/Bortz 2016, 98). Zu den Anforderungen an Konstruktvalidität siehe Pissarek zu quantitativer Forschung in diesem Band.

[13] Nerdel (2017) bezieht sich in den Ausführungen auf die Anwendung von Gütekriterien in Leistungsmessungen (ebd., 75f.).

- *interne Validität* liegt dann vor, wenn die empirischen Ergebnisse einer Studie mit den entwickelten Hypothesen erklären lassen. Die Ergebnisse einer Untersuchung sind eindeutig interpretierbar.
- *externe Validität* bezieht sich auf die Übertragbarkeit der Ergebnisse über die Stichprobe, Zeiträume und Situationen über die konkrete Studie hinaus. Angesprochen ist damit die Verallgemeinerbarkeit von Aussagen bzw. die Repräsentativität.

3. Bilanz: Herausforderungen und Schlussfolgerungen in der Anwendung von Gütekriterien für quantitative Forschung

Ein Vorteil quantitativer Forschung ist zweifelsohne die Exaktheit der Ergebnisse (siehe den Beitrag von Pissarek in diesem Band). Aus fachdidaktischer Perspektive zeigen sich allerdings auch Herausforderungen bzw. Grenzen bei der Einhaltung der dargestellten Gütekriterien für quantitativer Forschung, die es in Forschungsvorhaben zu reflektieren gilt, um zu fundierten Erkenntnissen zu gelangen.

Gerade die Einhaltung einer vollkommenen Interpretationsobjektivität ist in fachdidaktischen Studien kaum einlösbar, da spätestens bei der Interpretation der Daten die eigene Perspektive mit einfließt (u.a. Krebs/Menold 2014, 426; Schmelter 2014, 39). Eine ‚neutrale' Durchführung dieses Auswertungsschrittes ist als utopisch anzusehen. Zielführender erscheint es daher, hier das Kriterium *intersubjektive Nachvollziehbarkeit* (s.o., Kapitel 1.2) anzulegen, um diesem Umstand Rechnung zu tragen. Weiterhin ergibt sich ein Spannungsfeld hinsichtlich des Anspruchs, der mit der Konstruktvalidität verbunden ist: In fachdidaktischen Studien werden zumeist latente Konstrukte erfasst, die also nicht direkt beobachtbar sind, sondern erst operationalisiert werden müssen (u.a. Grabowski 2017, 317). Gerade die dazu herangezogenen Variablen sind aufgrund der hohen Komplexität fachlicher Gegenstände in der Deutschdidaktik häufig Gegenstand der Diskussion[14]. Um eine Einhaltung der Gütekriterien zu gewährleisten, müssen fachliche Gegenstände mitunter so stark zerlegt werden, dass die Frage der praktischen Bedeutsamkeit und somit die Einhaltung des allgemeinen Gütekriteriums *Praxisbezug* (Kapitel 1.2) zu diskutieren ist.[15] Gerade wenn sich die eigene Studie nicht auf gut eingeführte theoretische Konstrukte (wie z.B. Leseverständnis bei pragmatischen Alltagstexten) bezieht, stehen Forschende daher vor der Herausforderung, sich zwischen Einhaltung von Güte (und dem damit verbundenen Gewinn von

[14] Als Beispiele kann man hier die Diskussion zu Variablen für die Bewertung von Textqualität (u.a. Grabowski 2017, 318) oder die Debatte über Variablen für die Erfassung literarischer Kompetenz (z.B. Köster 2016, 67) nennen.

[15] Es soll damit nicht in Abrede gestellt werden, dass es auch gelungene Beispiele für quantitative deutschdidaktische Forschung unter Einhaltung der Gütekriterien gibt (z.B. Fay 2010).

Exaktheit, s.o.) und normativen fachlichen Ansprüchen an einen Gegenstandsbereich zu positionieren, wenn sie fachliche Konstrukte in einer quantitativen Untersuchung operationalisieren. Gerade hier ist zugleich eine Überprüfung von entwickelten Indikatorvariablen notwendig (Stichwort: Pilotierung). Im Kontext dieser Anforderungen an Forschende ist der Argumentation von Grabowski (2017, 318) zuzustimmen, dass die „die Entwicklung und Erprobung eines guten oder neuartigen Erhebungsverfahrens selbst […] ein anspruchsvolles Forschungsziel" sein kann (siehe beispielhaft dazu Winkler 2017). Nicht zuletzt wird so aber ein wichtiger Beitrag zur Professionalisierung einer empirisch forschenden Deutschdidaktik geleistet.

4. Kommentierte Literaturempfehlungen

Albert/Marx (2016): Empirisches Arbeiten in Linguistik und Sprachlehrforschung. Quantitative Studien von der Planungsphase bis zum Forschungsbericht. (hier: Kapitel 2.4)

Gut lesbare und kompakte Einführung in Gütekriterien für quantitative Forschungsansätze mit sprachwissenschaftlichem Fokus. Darüber hinaus bietet die Monographie einen sehr guten Überblick von der Planung einer quantitativen Studie bis hin zur Publikation der Ergebnisse.

Dörnyei (2007): Research Methods in Applied Linguistics. Quantitative, Qualitative and Mixed Methodologies.

Der Artikel bietet einen guten Überblick über die englischsprachige Diskussion zum Thema Gütekriterien.

Grabowski (2017): Anforderungen an Untersuchungsdesigns.

Ausgangspunkt dieses Beitrags ist, dass die Gütekriterien für quantitative Forschung bereits bei der Planung von Forschungsdesigns berücksichtigt werden müssen. Mit Blick auf die Frage, was geeignete Settings in der unterrichtsbezogenen Schreibforschung sein können, werden Anforderungen und Grenzen diskutiert und Ausführungen abschließend anhand von zwei schreibdidaktischen Forschungsprojekten konkretisiert.

Krebs/Menold (2014): Gütekriterien quantitativer Sozialforschung.

Gut lesbarer Handbuchbeitrag aus der empirischen Sozialforschung. Neben der Darstellung von Grundbegriffen ist der Beitrag auch interessant, da mögliche Fehlerarten (Stichprobenfehler, Messfehler …) in der quantitativen Forschung erörtert werden.

Schmelter (2014): Gütekriterien.

Im Artikel werden Gütekriterien quantitativer und qualitative Forschungsansätze aus Perspektive der DaZ-/DaF-Forschung erörtert. Aufgrund dieser didaktischen Perspektive auf das Thema bietet der Beitrag eine sinnvolle Ergänzung zu den Standardwerken der empirischen Sozialforschung; zumal im Beitrag auch (die

hier noch) offene Diskussion einer fachspezifischen Profilierung von Gütekriterien in den Blick genommen wird.

Literatur

Albert, Ruth/Marx, Nicole (2016): Empirisches Arbeiten in Linguistik und Sprachlehrforschung. Quantitative Studien von der Planungsphase bis zum Forschungsbericht. 3., überarb. Aufl. Tübingen: Narr.

Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hrsg.) (2017): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann.

Boelmann, Jan M. (Hrsg.) (2016): Empirische Erhebungs- und Auswertungsverfahren in der Deutschdidaktik. Baltmannsweiler: Schneider Hohengehren.

Döring, Nicola/Bortz, Jürgen (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5., vollst. überarb., aktual. u. erw. Aufl. Berlin: Springer.

Dörnyei, Zoltan (2007): Research Methods in Applied Linguistics. Quantitative, Qualitative and Mixed Methodologies. Oxford: Oxford University Press, 48-72.

Elliott, Robert/Fischer, Constance T./Rennie, David L. (1999): Evolving guidelines for publication of qualitative research studies in psychology and related fields. In: British Journal of Clinical Psychology, 38, 3, 215-229.

Fay, Johanna (2010): Kompetenzfacetten in der Rechtschreibdiagnostik. Rechtschreibleistung im Test und im freien Text. In: Didaktik Deutsch, 29, 15-36.

Flick, Uwe (2010): Gütekriterien qualitativer Forschung In: Mey, Günter/Mruck, Katja (Hrsg.): Handbuch Qualitative Forschung in der Psychologie. Wiesbaden: VS Verlag für Sozialwissenschaften, 395-407.

Grabowski, Joachim (2017): Anforderungen an Untersuchungsdesigns. In: Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hrsg.): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann, 315-334.

Helmers, Hermann (1971): Didaktik der deutschen Sprache. Einführung in die Theorie der muttersprachlichen und literarischen Bildung. 6., ern. bearb. u. erw. Aufl. Stuttgart: Klett.

Ilg, Stefan/Boothe, Brigitte (2010): Qualitative Forschung im psychologischen Feld: Was ist eine gute Publikation? In: Forum Qualitative Sozialforschung, 11, 2, 25.

Köster, Juliane (2016): Die dilemmatische Disziplin – Deutschdidaktik zwischen Eklektizismus und Partialisierung. In: Bräuer, Christoph (Hrsg.): Denkrahmen der Deutschdidaktik. Die Identität der Disziplin in der Diskussion. Frankfurt a. M: Peter Lang, 59-77.

Krebs, Dagmar/Menold, Natalja (2014): Gütekriterien quantitativer Sozialforschung. In: Baur, Nina/Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS, 425-438.

Lamnek, Siegfried (2010): Qualitative Sozialforschung. Lehrbuch. 5., überarb. Aufl. Unter Mitarbeit von Claudia Krell. Weinheim/Basel: Beltz.

Ludwig, Peter H. (2012): Thesen zur Debatte um Gütestandards in der qualitativen Bildungsforschung – eine integrative Position. In: Gläser-Zikuda, Michaela et al.

(Hrsg.): Mixed Methods in der empirischen Bildungsforschung. Münster: Waxmann, 79-89.

Moosbrugger, Helfried/Kelava, Augustin (2012): Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: Dies. (Hrsg.): Testtheorie und Fragebogenkonstruktion. 2., aktual. u. überarb. Aufl. Berlin: Springer, 7-26.

Nerdel, Claudia (2017): Grundlagen der Naturwissenschaftsdidaktik. Kompetenzbasiert und aufgabenorientiert in Schule und Hochschule. Heidelberg: Springer Spektrum.

Oswald, Hans (2013): Was heißt qualitativ forschen? Warnungen, Fehlerquellen, Möglichkeiten In: Friebertshäuser, Barbara/Langer, Antje/Prengel, Annedore (Hrsg.): Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft. 4., durchg. Aufl. Weinheim: Beltz, 183-201.

Schmelter, Lars (2014): Gütekriterien. In: Settinieri, Julia et al. (Hrsg.): Empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache. Eine Einführung. Paderborn: Schöningh, 33-46.

Shulman, Lee S. (1997): Disciplines of inquiry in education. In: Jaeger, Richard M. (Ed.): Complementary methods for researchers in education. Washington, D.C.: American Education Research Association, 3-19.

Winkler, Iris/Steinmetz, Michael (2016): Zum Spannungsverhältnis von deutschdidaktischen Fragestellungen und empirischen Erkenntnismöglichkeiten am Beispiel des Projekts KoALa. In: Krelle, Michael/Senn, Werner (Hrsg.): Qualitäten von Deutschunterricht. Stuttgart: Fillibach bei Klett, 37-56.

Winkler, Iris (2017): Potenzial zu kognitiver Aktivierung im Literaturunterricht. Fachspezifische Profilierung eines prominenten Konstrukts der Unterrichtsforschung. In: Didaktik Deutsch, 43, 78-97.

CHRISTIAN MÜLLER

Mixed Methods

Theorie und Praxis methodenpluraler Forschung

Forschende stehen bei der Planung eines Projektes stets vor der Herausforderung der Wahl eines passenden Forschungsansatzes. Wenn sich bestimmte Forschungsfragen einer exakten Beantwortung mit einer rein quantitativen oder rein qualitativen Vorgehensweise verschließen, erweist sich der Mixed-Methods-Ansatz als gebotene theoretische Grundlage für die Praxis. Dieser verknüpft beide Forschungsstränge, deren Gemeinsamkeiten in den empirischen und systematischen Vorgehensweisen liegen. Die Unterschiede der Perspektiven auf denselben Forschungsgegenstand werden zu Vorteilen im Forschungsprozess: Der Erschließung eines neuen Forschungsfeldes mit qualitativen Methoden folgt beispielsweise eine Hypothesenbildung, die mit quantitativen Methoden überprüft wird, um auf der Basis numerischer Ergebnisse eine Generalisierung abzuleiten.

Der folgende Artikel versucht, einen Beitrag zur Theorie und Praxis methodenpluraler Forschung für deutschdidaktische Forschungsdesigns zu leisten, indem im ersten Kapitel zunächst in die Grundlagen der Mixed-Methods-Forschung eingeführt und in der Folge auf verschiedene Designs (Kapitel 2) sowie die Datenanalyse (Kapitel 3) eingegangen wird. Im vierten Kapitel werden theoretische und praktische Problemfelder der Mixed-Methods-Forschung dargelegt, bevor mit einem Fazit geschlossen wird (Kapitel 5).

1. Grundlagen

Mixed Methods sind ein bisher noch selten beleuchteter Bereich empirischer Forschung (vgl. Kuckartz 2014a, 7), dessen Bearbeitung auch für die Deutschdidaktik von Interesse ist. Im Zuge weiterer Implementierungsbestrebungen empirischer Methoden in die deutschdidaktische Forschung sind u.a. solche zu berücksichtigen, die in Bezugsdisziplinen wie der Bildungsforschung bereits diskutiert und angewendet werden (vgl. Gläser-Zikuda et al. 2012). Sollen sich deutschdidaktische Forschungsdesigns nicht eindimensional auf ausschließlich quantitative *oder* qualitative Methoden beschränken, ist eine Erweiterung des empirischen Methodenrepertoires anzustreben und etablierten Forschungstraditionen „ein neues zeitgemäßes Methodenverständnis“ (Kuckartz 2014b, 18) entgegenzuset-

zen. Bei der schwierigen Entscheidung für einen Forschungsansatz bezüglich individueller deutschdidaktischer Forschungsvorhaben steht zu den zwei genannten eine weitere Option zur Auswahl, die einer „dritten methodologischen Bewegung“ (Tashakkori/Teddlie 2003, IX) entspringt. Es handelt sich dabei um die Methodenkombination, die sich nach Kuckartz wie folgt definieren lässt:

> Methodenkombination bedeutet allgemein, dass im Rahmen eines Forschungsprojektes beide Methoden und Datenarten, qualitative und quantitative, in sinnvoller Weise miteinander verbunden werden. Dies kann sowohl methodologisch begründet geschehen, [sic!] als auch in der inhaltlichen Logik eines Forschungsprojektes begründet sein. (Kuckartz 2014a, 28-29)

Methodenkombinationen folgen dem pragmatisch ausgerichteten Ansatz der Mixed-Methods-Forschung (vgl. Creswell/Plano Clark 2011, 22-36), der in den vergangenen Jahren an Popularität gewann und „keineswegs eine Randerscheinung“ (Kuckartz 2014a, 7) ist, was auf Grund der zahlreich erschienenen Publikationen zu methodenpluralistischer Grundlagenforschung offenbar wird (vgl. hierzu u.a. Creswell 2003; Tashakkori/Teddlie 2003, 2010; Hesse-Biber/Johnson 2015). Werden Mixed Methods in deutschsprachigen – und damit auch in deutschdidaktischen – Forschungsprojekten bislang tendenziell zurückhaltend verwendet (vgl. Kuckartz 2014a, 9), verdeutlicht eine Vielzahl an aktuellen Forschungsarbeiten verschiedener Wissenschaftsdisziplinen, dass der Ansatz international bereits fest verankert ist. All diese Arbeiten können in ihren Spezifika auch für die deutschdidaktische Forschung von Nutzen sein. So legt beispielsweise Mertens (2018) eine umfassende Publikation zu praxisorientierten Designs von Mixed Methods in der Evaluationsforschung vor. Das Arbeitsbuch von Bazeley (2018) und das Einführungswerk von Creamer (2018) fokussieren die bedeutsamsten Herausforderungen beim Einsatz von Mixed Methods: die systematische Integration und Analyse methodisch unterschiedlich erfasster Datensammlungen, die anhand zahlreicher praktischer Designbeispiele erklärt werden.

Da jeder methodische Ansatz für sich – also auch der quantitative und qualitative – einer gewissen Limitation bei der Beantwortung von Forschungsfragen unterliegt, ist es nur logisch, eine Verbindung von zwei Ansätzen anzustreben (vgl. Kuckartz 2014b, 16), die zwischen den rein quantitativen und qualitativen Polen unterschiedlicher Ausprägung sein kann, wie folgende Grafik[1] verdeutlicht:

quantitativ dominant		**paritätisch**	**qualitativ dominant**	
mono-methodisch rein quantitativ	Mixed Methods mit stärker quantitativer Ausprägung	Mixed Methods mit gleicher Ausprägung	Mixed Methods mit stärker qualitativer Ausprägung	mono-methodisch rein qualitativ

Abb. 1: Ausprägungen von quantitativen und qualitativen Forschungsmethoden

[1] Die Grafik lehnt sich an das von Johnson/Onwuegbuzie/Turner (2007, 124) entworfene Kontinuum verschiedener Forschungsansätze an.

So kann durch „die Integration von Methoden, Verfahren und Techniken, die zwei verschiedenen Ansätzen bzw. Methodenbereichen entstammen“ (Kuckartz 2014a, 30) eine Überwindung von monomethodischen Vorgehensweisen mit dem Ziel der Perspektiverweiterung vorgenommen werden. Ein entscheidendes Kriterium ist dabei „die *Kompatibilitätsannahme*, dass also die beiden Methoden tatsächlich miteinander vereinbar sind, sich ergänzen und unterschiedliche Perspektiven liefern“ (Kuckartz 2014a, 35 [Herv. i.O.]). Die Entscheidung, dem Mixed-Methods-Ansatz zu folgen, liegt in der Zielsetzung ein tiefergehendes Verständnis einer Forschungsproblematik (vgl. Creswell 2015, 2) und vielfältigere Forschungsergebnisse aus den zielgerichtet kombinierten qualitativen und quantitativen Daten erhalten zu können. Ein weiterer – auch forschungspragmatischer – Vorteil von Mixed Methods liegt darin, sich nicht für oder gegen quantitative und qualitative Methoden entscheiden zu müssen, sondern der in verschiedenen Wissenschaftsbereichen vorhandenen Diskussion um die vermeintlich einzig korrekte Methodologie galant begegnen zu können. In Frage gestellt wird heute beispielsweise nicht mehr die Anerkennung und der Erkenntnisgewinn durch das Konzept der Triangulation (vgl. Denzin 2009 [1970]; Flick 2011; Kuckartz 2014a, 46-47), das mit dem Mixed-Methods-Ansatz eng verzahnt ist. Unter Triangulation wird die Kombination bzw. Überlagerung verschiedener Elemente mit dem Ziel verstanden, mögliche oder spezifische Schwächen der Einzelelemente auszubessern und so ein komplexeres sowie genaueres Gesamtbild zu entwerfen. Hierbei gestaltet sich das Spektrum verschiedener Formen von Triangulation breit:

1. Theorietriangulation: Bei der Theorietriangulation werden Forschungsergebnisse vor der Folie verschiedener Theorien interpretiert, so dass einzelne, im Vorhinein theoretisch möglicherweise stark beeinflusste Hypothesen eine neue Perspektive hinzugewinnen.
2. Forschertriangulation: Untersuchen und interpretieren unterschiedliche Forscher einen Gegenstand, liegt eine Forschertriangulation vor. Der Vergleich der Einflüsse der Forscher soll in der Folge zu validen Ergebnissen bezüglich des Gegenstands führen.
3. Datentriangulation: Bei der Datentriangulation wird ein Phänomen mit einer Erhebungsmethode an unterschiedlichen Messzeitpunkten, Orten und Probanden erforscht.

Diese drei ersten Formen von Triangulation zeigen, dass der Begriff sich nicht ausschließlich auf die Verwendung mehrerer Methoden beziehen muss, was für den Mixed-Methods-Ansatz jedoch grundlegend ist. Dennoch können Mixed Methods und (Methoden-)Triangulation nicht gänzlich losgelöst voneinander betrachtet werden (vgl. Flick 2017). Fällt die Unterscheidung von Mixed Methods zur Theorie-, Forscher- und Datentriangulation noch leicht, ist eine Abgrenzung speziell zur Methodentriangulation komplizierter. Das Verständnis des im Zusammenhang mit Mixed Methods oft verwendeten Begriffs der Methodentriangulation (vgl. Flick 2018, 444-461) lässt sich am ehesten in seiner Abhängigkeit von den Phasen des Forschungsprozesses erklären:

4. Methodentriangulation: Methodentriangulation liegt grundsätzlich vor, wenn ein Phänomen mit verschiedenen Methoden untersucht wird.
   a) In Abgrenzung zu Mixed Methods ist Methodentriangulation zu attestieren, wenn anhand von zwei quantitativen *oder* zwei qualitativen Methoden geforscht wird.
   b) Kommen qualitative *und* quantitative Forschungsmethoden systematisch in den Phasen der Datenerhebung und Datenanalyse zum Einsatz, liegt eine Methodentriangulation nach dem Mixed-Methods-Ansatz vor.
   c) Werden Methoden beider Ansätze nur in der Datenauswertungsphase genutzt, ist ausschließlich von Mixed Methods zu sprechen (vgl. Creswell/Plano Clark 2011). Für Letzteres beispielhaft angeführt werden kann die quantitative Auswertung der Häufigkeit der verwendeten Codes einer zuvor erfolgten qualitativen Inhaltsanalyse. Zum Tragen kommt somit möglicherweise das bessere Verstehen einer Forschungsproblematik durch die qualitativen Maßnahmen zur Eruierung der Sinnhaftigkeit und im quantitativen Vorgehen des Zählens (vgl. Kuckartz 2014a, 53).

Somit folgen Mixed Methods einem „sehr konkreten, praktisch ausgerichteten Forschungsansatz, verbunden mit ganz eigenen Strategien des Designs, der Datenerhebung und der Datenanalyse“ (Kuckartz 2014a, 11), während der Triangulationsansatz in seinen Ursprüngen die Validierung in den Mittelpunkt stellt, so dass „mehr als eine Perspektive zur Untersuchung einer Forschungsfrage eingesetzt wird, um so das Vertrauen in die Validität der Resultate zu erhöhen“ (ebd., 47). Die Planung und Durchführung einer Studie, die auf Verfahren von Mixed Methods basiert, bedarf der passenden Wahl aus verschiedenen Mixed-Methods-Forschungsdesigns, die im nächsten Kapitel vorgestellt werden.

2. Mixed-Methods-Designs

Das Forschungsdesign einer Studie beschreibt das forschungsmethodische Vorgehen, muss bestimmten Qualitätsstandards genügen und wird in der Planungsphase des Forschungsprozesses entworfen. Das Design einer Mixed-Methods-Studie ist grundsätzlich kriteriengeleitet und anhand einer Forschungsfrage begründet ausgerichtet. So lässt sich konstatieren, dass es eine große Menge an verschiedenen Mixed-Methods-Designs gibt und neue stets entworfen werden müssen. Daher kann zur Anlage einer Mixed-Methods-Studie nur eine Auswahl an Forschungsdesigns vorgestellt werden. Nach Creswell (vgl. 2003, 211) sind bei der Entscheidung für ein Design die Abfolge der qualitativen und quantitativen Erhebung, die Priorisierung und Schwerpunktsetzung einer der Teilstudien, der Zeitpunkt der Integration beider Datenarten und die theoretische Einbettung leitend.

Eine Mixed-Methods-Studie folgt zudem unter Beachtung möglicher methodischer Schwierigkeiten einem spezifischen Designtyp. Der Konzeption von Mixed-Methods-Forschungsdesigns liegt eine Kombination aus quantitativen und

qualitativen Methoden zu Grunde, die parallel oder sequenziell der Erreichung der Forschungsziele dienlich ist.[2]

2.1 Parallele Forschungsdesigns

In parallelen (*concurrent*) Forschungsdesigns werden qualitative und quantitative Daten simultan erhoben und ausgewertet (vgl. Creswell 2003). Es können drei Arten paralleler Designs zur gleichzeitigen Integration beider Methoden unterschieden werden.

1. Bei der Anwendung der parallelen Triangulationsstrategie (*concurrent triangulation*) werden die Ergebnisse von zwei separaten Erhebungen miteinander in Beziehung gesetzt. Z.B. wird eine Forschungsfrage parallel mit einem leitfadengestützten Interview qualitativ und mit einem standardisierten Fragebogen quantitativ untersucht. Die Resultate beider Methoden werden im Anschluss kombiniert und ausgewertet.
2. Das parallele verschachtelte Design (*concurrent nested*) räumt einem Ansatz hingegen Vorrang ein: Hier findet schon in der Datenerhebungsphase eine Integration der einen in die andere Methode statt. So kann z.B. die qualitative Erhebung von Daten eines Teilaspekts mittels Interviewverfahren durch quantifizierbare Daten einer Befragung tiefergehend betrachtet werden.
3. Beim transformativen parallelen Design (*concurrent transformative*) beherrscht eine theoretische Perspektive die Erhebungen, ohne dass ein Ansatz dabei dominant sein muss (vgl. Kuckartz 2014a, 67-68). Leitend für eine solche Mixed-Methods-Studie ist dabei der theoretische Rahmen, der sich z.B. an fachdidaktischen, fachwissenschaftlichen, pädagogischen, gesellschaftlichen oder kulturellen Konzepten ausrichtet. Diese theoretische Fokussierung gibt vor, mit welcher Methodenkombination in diesem Bereich am ehesten valide Ergebnisse erzielt werden können.

[2] Die Kombinationsvarianten der Methoden folgen einem bestimmten Notationssystem: Bedeutet ‚QUAL/qual' qualitative, entspricht ‚QUAN/quan' quantitativer Forschung. Werden die Notationen großgeschrieben, bilden sie den Schwerpunkt einer Studie; bei Kleinschreibung kommt der Methode eine geringere Bedeutung zu. Finden Datenerhebungen gleichzeitig statt, wird dies mit einem ‚+' markiert. Mit einem Pfeil ‚→' wird die sequenzielle Reihenfolge der Erhebungen symbolisiert (vgl. Morse 2008 [1991], 149-158).

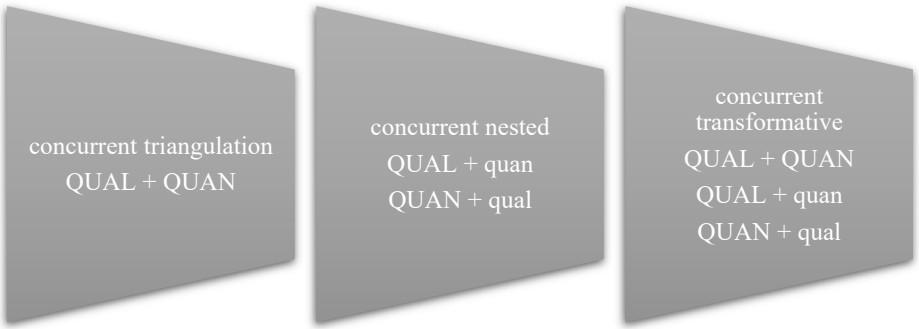

Abb. 2: Parallele Forschungsdesigns mit Notationen

2.2 Sequenzielle Forschungsdesigns

Bei den Datenerhebungen sequenzieller (*sequential*) Forschungsanlagen steht die Reihenfolge der Methoden im Vordergrund. Es wird erklärend und vertiefend vorgegangen, indem eine quantitative Methode von einer qualitativen ergänzt wird (*explanatory design*). Zum Beispiel erfolgt zunächst eine statistische Analyse von Daten einer Befragung. Eine anschließende Gruppendiskussion und die Auswertung der hieraus erlangten qualitativen Daten ermöglichen ein genaueres Verständnis der quantitativen; vice versa wird in einem explorativen Vorgehen einer qualitativen eine quantitative nachgestellt (*exploratory design*). Bei der dritten sequenziellen Form ist die Abfolge der Datenerhebung weniger bedeutsam als die theoretische Ausrichtung (*transformative design*) (vgl. Teddlie/Tashakkori 2006, 12-28; vgl. Creswell 2003, 211-215).

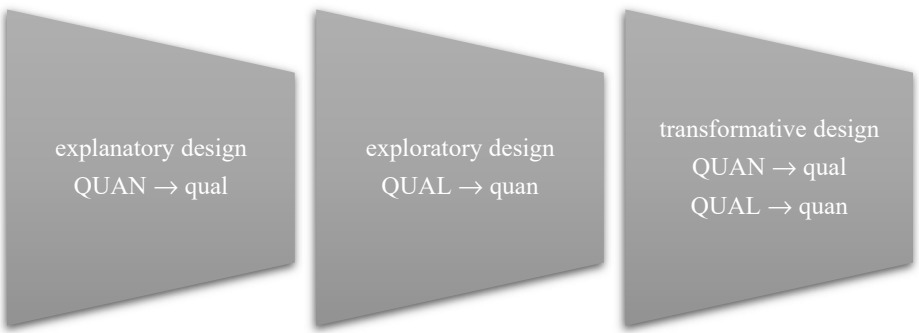

Abb. 3: Sequenzielle Forschungsdesigns mit Notationen

3. Datenanalyse

Eine Mixed-Methods-Datenanalyse richtet sich bei der Auswertung der Daten an quantitativen (Kapitel 3.1) und qualitativen (Kapitel 3.2) Vorgehensweisen aus,

indem u.a. sowohl deskriptiv mit Statistiken und Hypothesenüberprüfungen gearbeitet wird als auch beispielsweise kategorisierte und codierte Datensammlungen einer Interpretation unterliegen. Infolgedessen müssen entsprechende – auch computergestützte – Verfahren in einer integrativen Datenanalyse (Kapitel 3.3) angewendet werden.

3.1 Analyse quantitativer Daten

Der statistische Analyseprozess quantitativer Daten bedarf in der Planungsphase zunächst einer digitalen Aufbereitung durch Eingabe in Computerprogramme wie z.B. SPSS. Die Daten gilt es anschließend zu überprüfen, eventuelle Fehler zu beseitigen und auf ihre Plausibilität hin zu kontrollieren. Dies ist besonders bedeutsam, wenn für einen Untersuchungsgegenstand auch qualitative Daten erhoben werden. Eindeutig zuweisbare Kategorisierungen und Codierungen unterstützen die Kombination der Datentypen einer Analyse. Nach der Grundauszählung und der univariaten Verfahrensweise der explorativen Auswertungsphase sowie weiteren Berechnungen lassen sich die Statistiken entsprechend visuell aufbereiten. In einem nächsten Schritt erfolgt eine komplexere Analyse möglicher Zusammenhänge zwischen mehreren Variablen (bivariate oder multivariate Analyse). Zuletzt werden die Ergebnisse in einem abschließenden Forschungsbericht visualisiert, diskutiert und bewertet (vgl. Kuckartz 2014a, 105-108).

3.2 Analyse qualitativer Daten

Zur Analyse qualitativer Daten sind verschiedene Optionen zur Beschreibung, zum Vergleich und zur Herstellung von Zusammenhängen möglich. Die spezifischen Herausforderungen liegen in der Kenntnis und Anwendung von z.B. der qualitativen Inhaltsanalyse, Grounded Theory, phänomenologischen Analyse, Diskursanalyse etc. Ähnlich wie bei der quantitativen Datenanalyse werden die Daten zunächst aufbereitet, was bei der Vielzahl an möglichen Datenformen (z.B. Textdokumente, Audiodateien, Bild- und Videodaten) umfangreichere Tätigkeiten der Forschenden erfordert. Im Anschluss werden die Daten explorativ anhand einer Kategorienbildung bearbeitet und genauer mit einer QDA-Software codiert, bevor die Ergebnisse dargestellt und interpretiert werden (vgl. Kuckartz 2014a, 109-113).

3.3 Integrative Datenanalyse

Der Prozess der integrativen Datenanalyse besteht aus der Reduktion und der Darstellung der Daten sowie der Bildung von Rückschlüssen (s. Abb. 3). In einem Mixed-Methods-Design, in dem ein Mixing in jeder Phase stattfinden kann, werden diese drei Schritte sowohl quantitativ als auch qualitativ vollzogen (s. Abb. 3) mit dem Ziel, die „Frage nach der Integration und Relation der Ergebnisse“ (Kuckartz 2014a, 100) zu beantworten. Eine integrative Datenanalyse kann sich nach Creswell/Plano Clark (2011) in vier unterschiedlichen Vorgehensweisen vollziehen:

- Die Ergebnisse der Datenanalyse beider Methodenstränge können in z.B. einem parallelen Design am Ende des Forschungsprozesses in einem Forschungsbericht zusammengefasst und integrativ visualisiert werden.
- Die Daten und Ergebnisse des einen Strangs sind grundlegend für die in einem sequenziellen Design angelegte Erhebung und Auswertung der Daten des anderen Strangs. Eine Kombination beider Daten und Ergebnisse findet folglich bereits vor einem abschließenden Forschungsbericht z.B. in der Darstellung in einer Concept-Map statt.
- Die Datenformen werden in einer Qualifizierung und/oder Quantifizierung von der einen in die andere Form transformiert.
- Die Erhebung der Daten des in einer Studie nachrangigen Datentyps erfolgt in einer dezidiert hierfür geschaffenen Phase.

Unabhängig von diesen Vorgehensweisen können die Ergebnisse der beiden Datenformen mit einer entsprechenden Software in Joint Displays (integrative tabellarische und grafische Darstellungsmöglichkeiten) mit dem Ziel erstellt werden, die Zusammenhänge in übersichtlichen Visualisierungen zu verdeutlichen (vgl. Creswell/Plano Clark 2011, 224-230) und somit ein besseres Verstehen zu gewährleisten; sie sind jedoch nicht zu verwechseln mit der integrativen Datenanalyse an sich (vgl. Burzan 2016, 73-75). Auf der Grundlage der Ergebnisdarstellungen können (Meta-)Inferenzen herausgebildet werden (s. Abb. 3): „Inferences in mixed methods research are conclusions or interpretations drawn from the separate quantitative an qualitative strands of a study as well as across the quantitative and qualitative strands, called ‘meta-inferences‘“ (Creswell/Plano Clark 2011, 212-213). Meta-Inferenzen können folglich als „vergleichende Betrachtung der Schlussfolgerungen (Inferenzen) beider Stränge [so genannte ‚strands‘, C.M.]“ (Kuckartz 2014a, 101) verstanden werden.

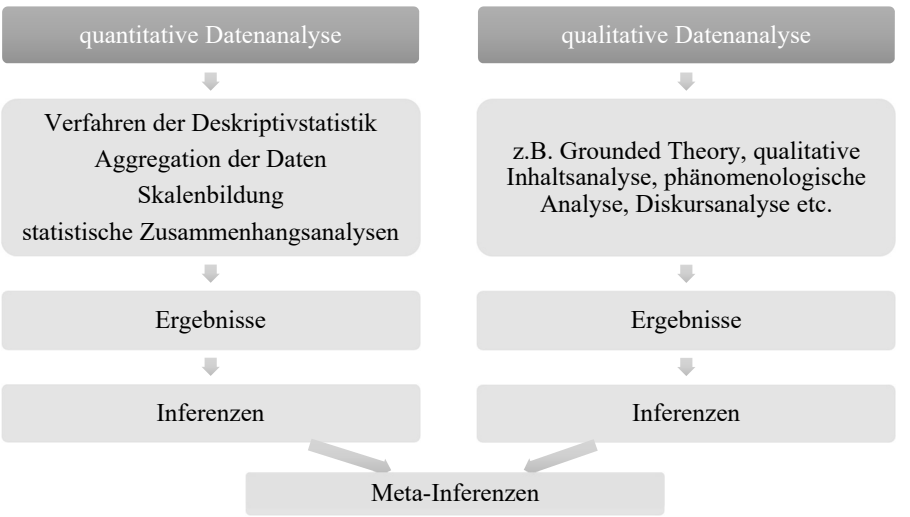

Abb. 4: Prozess einer Mixed-Methods-Datenanalyse

Ausgerichtet am jeweiligen Designtyp ist zuletzt ein entsprechender Forschungsbericht zu verfassen (vgl. Creswell/Zhang 2009, 612-621). Anschließend werden die in den Berichten dargelegten Ergebnisse der Teilstudien – z.B. in einem parallelen Design – auf Basis von Meta-Inferenzen in einem integrativen Bericht miteinander verbunden (vgl. Kuckartz 2014a, 74).

4. Theoretische und forschungspraktische Problemfelder

Eine theoriebasierte Intention und Begründung für die Verknüpfung von Methoden sind für das zu planende und durchzuführende Forschungsvorhaben von besonderer Bedeutung. Forschende bewegen sich dabei in verschiedenen theoretischen und praktischen Problemfeldern von Mixed Methods.

Aus theoretischer Perspektive würde von Anhängern monomethodischer Forschung an Mixed Methods kritisiert, „dass die wissenschaftstheoretische und epistemologische Fundierung unklar sei“ (Kuckartz 2014a, 156). Multiperspektivität im Einsatz zweier Forschungsmethoden eines Mixed-Methods-Designs führt weder automatisch zu einer ganzheitlichen noch einer objektivierbaren Forschung (vgl. Bergman 2011, 274-275). Ziel der Mixed-Methods-Forschung kann es daher nur sein, einen Gegenstand mit qualitativen und quantitativen Methoden umfassender zu erforschen, um zu validen Ergebnissen zu gelangen (vgl. Kelle 2007, 261).

Aus forschungspraktischer Sicht problematisch ist die Verbindung von (oftmals divergierenden) Fragestellungen und Ergebnissen beider Methodenstränge; gefolgt von dem aufzuwendenden zeitlichen Umfang einer Mixed-Methods-Studie, der Abhängigkeit bei der Arbeit in Forscherteams – die v.a. jedoch in interdisziplinärer Zusammenarbeit erhebliche Vorteile haben (vgl. Bartholomew/Brown 2012) – und der eventuell fehlenden Akzeptanz und Wertschätzung von Mixed Methods aus Sicht der Gutachter (vgl. Kuckartz 2014a, 157).

5. Fazit

Der Mixed-Methods-Ansatz stellt eine besondere Herausforderung für Forschende dar, da dieser der Expertise von gleich zwei theoriebasierten Methoden sowie ihrer forschungspraktischen Integration bedarf. Da zudem in der empirischen Bildungsforschung nach wie vor gemeinsame Standards für quantitative und qualitative Methoden fehlen (vgl. Mayring 2012, 287-300), wirkt sich dies auch auf das empirische Arbeiten mit Mixed Methods in der Deutschdidaktik aus: Forschende können sich bislang nur auf vage formulierte Denk- und Orientierungsrahmen sowie auf vielfältige Modelle und Vorgehensweisen für die Praxis stützen. Eine zu starke Normierung und Formalisierung – oder gar Dogmatisierung – wirkte jedoch wiederum zu Lasten der offenen, weit gefassten Überlegungen zur Theorie und Praxis von Mixed Methods. Diese theoretisch-praxisorientierten Ambivalenzen verdeutlichen die künftigen Arbeitsfelder und Aufgaben

der Mixed-Methods-Forschung, die sich noch immer in einer Phase der Etablierung befindet.

Letztlich kann hervorgehoben werden, dass Mixed Methods zwar kein a priori einlösbares Versprechen für multiperspektivisch abgesicherte Forschungsergebnisse sind; in der Nutzung der Stärken und im Ausgleich möglicher Nachteile beider Methodenstränge (vgl. Wiggins 2011) eignen sie sich indes vortrefflich zur Beantwortung von komplexen Forschungsfragen, wenn eine entsprechende Kombination das probateste Mittel zur Erreichung valider Resultate darstellt.

Literatur

Bartholomew, Theodore/Brown, Jill (2012): Mixed Methods, Culture, and Psychology: A Review of Mixed Methods in Culture-specific Psychological Research. In: International Perspectives in Psychology: Research, Practice, Consultation, 1, 3, 177-190.

Bazeley, Pat (2018): Integrating Analyses in Mixed Methods Research. London et al.: Sage.

Bergman, Manfred (2011): The Good, the Bad, and the Ugly in Mixed Methods Research and Design. In: Journal of Mixed Methods Research, 5, 4, 271-275.

Burzan, Nicole (2016): Methodenplurale Forschung. Chancen und Probleme von Mixed Methods. Weinheim/Basel: Beltz Juventa.

Creamer, Elizabeth (2018): An Introduction to Fully Integrated Mixed Methods Research. Thousand Oaks et al.: Sage.

Creswell, John (2003): Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Thousand Oaks et al.: Sage.

Creswell, John (2015): A Concise Introduction to Mixed Methods Research. Thousand Oaks et al.: Sage.

Creswell, John/Plano Clark, Vicki (2011): Designing and Conducting Mixed Methods Research. Thousand Oaks et al.: Sage.

Creswell, John/Zhang, Wanqing (2009): The Application of Mixed Methods Designs to Trauma Research. In: Journal of Traumatic Stress, 22, 6, 612-621.

Denzin, Norman (2009/1970): The Research Act: A Theoretical Introduction to Sociological Methods. New Brunswick/London: Aldine Transaction.

Flick, Uwe (2011): Triangulation. Eine Einführung. Wiesbaden: Springer.

Flick, Uwe (2017): Doing Triangulation and Mixed Methods. London et al.: Sage.

Flick, Uwe (2018): Triangulation. In: Denzin, Norman K./Lincoln, Yvonna (Hrsg.): Handbook of Qualitative Research. Thousand Oaks et al.: Sage, 444-461.

Hesse-Biber, Sharlene/Johnson, Robert (2015): The Oxford Handbook of Multimethod and Mixed Methods Research Inquiry. New York: Oxford University Press.

Johnson, Robert/Onwuegbuzie, Anthony/Turner, Lisa (2007): Toward a Definition of Mixed Methods Research. In: Journal of Mixed Methods Research, 1, 2, 112-133.

Kelle, Udo (2007): Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte. Wiesbaden: Springer.

Kuckartz, Udo (2014a): Mixed Methods. Methodologie, Forschungsdesign und Analyseverfahren. Wiesbaden: Springer.

Kuckartz, Udo (2014b): Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung. Weinheim/Basel: Beltz Juventa.

Gläser-Zikuda, Michaela/Seidel, Tina/Rohlfs, Carsten/Gröschner, Alexander/Ziegelbauer, Sascha (Hrsg.) (2012): Mixed Methods in der empirischen Bildungsforschung. Münster: Waxmann.

Mayring, Philipp (2012): Mixed Methods – ein Plädoyer für gemeinsame Forschungsstandards qualitativer und quantitativer Methoden. In: Gläser-Zikuda, Michaela/Seidel, Tina/Rohlfs, Carsten/Gröschner, Alexander/Ziegelbauer, Sascha (Hrsg.): Mixed Methods in der empirischen Bildungsforschung. Münster: Waxmann, 287-300.

Mertens, Donna (2018): Mixed Methods Design in Evaluation. Thousand Oaks et al.: Sage.

Morse, Janice (2008/1991): Approaches to qualitative-quantitative methodological triangulation. In: Creswell, John/Plano Clark, Vicki (Hrsg.): The Mixed Methods Reader. Thousand Oaks et al.: Sage, 149-158.

Tashakkori, Abbas/Teddlie, Charles (2003/2010): Handbook of Mixed Methods in Social & Behavioral Research. Thousand Oaks et al.: Sage.

Teddlie, Charles/Tashakkori, Abbas (2006): A General Typology of Research Designs Featuring Mixed Methods. In: Research the Schools, 13, 1, 12-28.

Wiggins, Bradford (2011): Confronting the Dilemma of Mixed Methods. In: Journal of Theoretical and Philosophical Psychology, 31, 1, 44-60.

JAN M. BOELMANN

Forschungsdesign

Bei der Wahl des Forschungsdesigns[1] kommen die zuvor getätigten Entscheidungen zusammen und werden mit der Zusammenstellung der Erhebungs- und Auswertungsverfahren konkretisiert. Mit Blick auf die Fragestellung der Studie wurde bereits entschieden, welcher Forschungsansatz gewählt wurde und ob eine qualitative, quantitative oder gemischte Herangehensweise benötigt wird. Nun gilt es, in die konkrete Planung des Vorhabens einzusteigen. Optimalerweise fügen sich die Überlegungen zur Forschungsfrage, -ansätzen und -typen an diesem Punkt zu einem Ganzen, ohne dass erneut recherchiert oder theoretisch fundiert werden muss, da dies bereits vorher geschah. Zugleich deckt die Festlegung des Forschungsdesigns vergleichsweise schonungslos Lücken und Fehler in der Vorbereitung auf.

Folgende Schritte stehen bei der Entwicklung des Forschungsdesigns zentral:

- Festlegung des Gegenstands
- Festlegung der Probandengruppe
- Festlegung der Erhebungs- und Auswertungsverfahren
- Festlegung des Ablaufs

All diese Schritte laufen nicht chronologisch ab, sondern finden in einem zirkulären Verfahren statt: Alle Aspekte sind mit den anderen verbunden und wirken sich aufeinander aus, sodass eine Veränderung der *Probandengruppe* ggf. einen anderen *Gegenstand* erfordert oder eine Anpassung der *Erhebungs- und Auswertungsverfahren* erfolgen muss. Dieses Vorgehen gleicht einem Balanceakt, in dem die verschiedenen Design-Elemente reflektiert aufeinander abgestimmt werden. Um die Prozesse vorzustellen, werden im Folgenden die einzelnen Festlegungen jedoch voneinander separiert skizziert.

1. Festlegung des Gegenstands

Jedes empirische Projekt hat einen Gegenstand, an dem oder zu dem geforscht wird. Das kann das *literarische Verstehen* sein, *Gattungswissen*, die Fähigkeit

[1] Dieser Beitrag deckt die theoretischen Fragen der Forschungsdesign-Festlegung ab, während der Beitrag von König in diesem Band praktische Erwägungen in den Fokus rückt. Beide Beiträge betrachten somit ähnliche Arbeitsschritte aus unterschiedlichen Blickwinkeln, weshalb die ergänzende Lektüre empfohlen wird.

Junktoren angemessen einzusetzen, die *Nutzung Neuer Medien* oder die *Überzeugung zu Lehr-/Lernprozessen im inklusiven Unterricht* – der Vielfalt ist keine Grenze gesetzt. Auf einer Makroebene lässt sich vom *Thema der Forschung* sprechen. Für die Planung der empirischen Erhebung muss dieses Thema dann differenzierter betrachtet und in kleinere Einheiten zerlegt werden, die dem Rahmen des jeweiligen Forschungsprojekts entsprechen. In vielen Fällen geschieht diese Präzisierung bereits bei der Formulierung der Fragestellung (siehe den Beitrag von Boelmann in diesem Band), in einigen Fällen – etwa wenn die Beantwortung der Frage einen theoretisch-heuristischen und einen empirischen Teil erfordert – muss der Gegenstand des empirischen Settings zur Festlegung des Forschungsdesigns neu präzisiert werden.

1.1 Die Makro-Ebene des Projekts

Die Makro-Ebene des Projekts nimmt den *übergeordneten Gegenstand* – also die Festlegung in welchem *Forschungsfeld*[2] zu welchem *Thema* geforscht werden soll – in den Blick und präzisiert ihn in seiner konkret behandelten Ausprägung. Hierbei lohnt die Unterscheidung zwischen *Breite* und *Tiefe* der Untersuchung des Gegenstands[3]. Die *Breite* zeigt sich darin, wie umfangreich bzw. vielfältig der gewählte Gegenstand ist: Wird etwa bei der Untersuchung des Gattungswissens a) nur eine Gattung oder b) die fünf in der Schule verbreitetsten Gattungen oder c) gar alle vorhandenen Gattungen berücksichtigt? Ist man bei der Suche nach Überzeugungen zum inklusiven Unterricht offen für alle Möglichkeiten oder werden ausgewählte Aspekte – etwa das Störverhalten und die Individualisierung von Aufgaben – fokussiert? Einher geht diese Festlegung mit der Wahl der *Tiefe*, die Aufschluss darüber gibt, wie intensiv mit bzw. an dem Gegenstand gearbeitet wird: Prüft der Test zum Gattungsverstehen nur ab, ob eine Gattung *erkannt* wurde oder werden die *mentalen Tiefenstrukturen* des Gattungsverstehens erforscht.

Breite und *Tiefe* der Wahl des Gegenstands sind keine normativ messbaren Aspekte, doch helfen sie bei der Orientierung, wenn sie als zwei Achsen eines Koordinatensystems verstanden werden, auf dem man sein Projekt verortet.

Verfügt der gewählte Gegenstand über zu wenig Breite und zu wenig Tiefe, muss das Thema wohl oder übel als banal gekennzeichnet werden – es eignet sich nicht für eine empirische Untersuchung. Kleinere Studienarbeiten zeichnen sich dadurch aus, dass sie ein Thema in einem überschaubaren Rahmen abdecken, hier bedarf es einer gewissen Breite und/oder Tiefe, ohne dass der Umfang ausufern

---

[2] Für einen Überblick über die Forschungsfelder der empirischen Deutschdidaktik, siehe Band 3 dieser Reihe: Boelmann 2018b.

[3] Die Begriffe Breite und Tiefe stehen ausdrücklich nicht synonym zu qualitativer und quantitativer Forschung. Sowohl qualitative Forschung bedarf neben der ihr eigenen Tiefe auch einer gewissen Breite der Gegenstände, etwa des theoretischen Konzepts und der behandelten Aspekte, wie auch die quantitative Forschung tiefenstrukturelle Aspekte abzuprüfen vermag.

würde. Abschlussarbeiten sollen ein umfassendes Verständnis eines Themengebiets dokumentieren, sodass hier sowohl Breite als auch Tiefe bei der Untersuchung des Gegenstands erwartet wird. Dissertationen weisen nach, dass der Autor/die Autorin der/die deutschlandweit führende Instanz für das untersuchte Thema ist. Dies muss sich auch in der Durchdringung der Breite und Tiefe und damit der Wahl des Themas widerspiegeln.

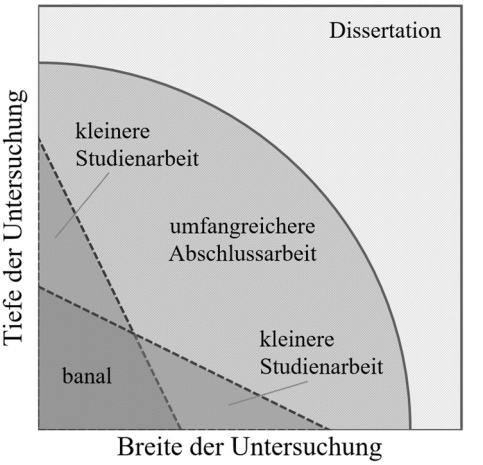

Abb. 1: Breite und Tiefe der Untersuchung

Zugleich bedeutet die Festlegung des Gegenstands auch, dass an dieser Stelle definiert werden muss, was das spezifische Arbeitsverständnis des Gegenstands auszeichnet. Insbesondere in Forschungsfeldern, in denen verschiedene Definitionen vorherrschen oder eine zentrale Begrifflichkeit mehrdeutig interpretiert werden kann, muss der Gegenstand präzise bestimmt und eingegrenzt werden (vgl. Boelmann 2018b).

Wie ein übergeordnetes Thema mit einem sehr vielfältigen Gegenstand in verschiedene Unterthemen und Untergegenstände konkretisiert werden kann, soll am Beispiel des von Ylva Schwinghammer[4] durchgeführten Projekts *Arbeitskoffer zu den Steirischen Literaturpfaden des Mittelalters* nachgezeichnet werden: Die zentrale Fragestellung zielt darauf ab, wie Schülerinnen und Schüler mit mittel- und frühneuhochdeutschen Texten umgehen. Diese Frage dient der Evaluation eines Praxisprojekts und hat vor allem fragen- und hypothesengenerierende Funktion: Schwinghammer arbeitet Elemente des Best-Practice heraus (Was gelingt gut? Was muss für folgende Projekte verbessert werden?) und legt somit ein wissenschaftliches Fundament für kommende Praxisprojekte, ohne alle Bedingungsfaktoren im Detail zu untersuchen. Vor diesem Hintergrund erklärt sich, dass der Vielfältigkeit der untersuchten Aspekte, also der Breite, Vorrang vor der Tiefe der Untersuchung eingeräumt wird. In diesem Sinne präzisiert sie die vergleichsweise offene Fragestellung, indem sie den Gegenstand des Projekts *Umgang mit mittel- und frühneuhochdeutschen Texten* in sieben Unterfragen präzisiert und so spezifische Untersuchungsgegenstände schafft, an denen sich die Erhebungen im Rahmen des Projekts orientieren:

1. Welche **Textstellen** sind für die Zielgruppe besonders schwierig oder auch spannend?
2. Welche **Strategien** wenden Schülerinnen und Schüler an, um Texte der älteren Sprachstufen zu entschlüsseln?

[4] Die folgenden Ausführungen beziehen sich auf die methodischen Betrachtungen in Schwinghammer 2018. Für detailliertere Ausführungen zum Projekt sei auf Schwinghammer 2015 sowie Hofmann/Schwinghammer 2015 verwiesen.

3. Welche **Hilfestellungen** erleichtern den Zugang zum Text (Hinweise sprachlicher und inhaltlicher Natur, Vorlesen, Hörverstehen …)?
4. Wie kommen Schülerinnen und Schüler mit dem **mittelhochdeutschen Wörterbuch** zurecht?
5. Wie verläuft die **Anschlusskommunikation** und welche Schlüsse lassen sich daraus ziehen?
6. Welche **Texte und Teilaspekte** eignen sich für welche Einsatzzwecke im Unterricht?
7. Was kann besonders gut mit welchen **Methoden** unterrichtet werden? (Schwinghammer 2018, 175 [Num. u. Herv. JMB])

Durch die Fragen wird die Blickrichtung auf gegenstandsbezogene (Frage 1, Frage 4, Frage 6), (lese-)strategische (Frage 2, Frage 3) und kommunikative (Frage 1, Frage 2, Frage 5) Aspekte gelegt, die in Frage 7 zusammengeführt werden. Schwinghammer zergliedert folglich eine große Frage und somit einen großen Gegenstand in verschiedene kleine Fragen/Gegenstände, deren jeweilige Beantwortung in der Folge einen Beitrag für die Beantwortung der großen Frage liefern.

Die Festlegung der Gegenstände auf der Makroebene wirken sich in der Folge auf die anderen Arbeitsschritte der Design-Festlegung aus, da nun *Erhebungs- und Auswertungsverfahren* (vgl. Boelmann 2018a) gewählt werden müssen, mit denen die ausgewählten Gegenstände auch valide zu untersuchen sind. In der Regel gelingt dies nicht mit einem einzigen Instrument, sondern es müssen unterschiedliche Verfahren eingesetzt werden, um allen Einzelaspekten in der Tiefe gerecht zu werden. So würde eine *Erhebung* durch Beobachtung, etwa Videographie, die spätere *Auswertung* mit verschiedenen spezialisierten Verfahren wie *Inhaltsanalysen* für gegenstandsbezogene und (lese-)strategische Aspekte und *Gesprächsanalysen* für kommunikative Aspekte ermöglichen. Schwinghammer wählt einen pragmatischen Weg, der einen Blick in die Breite mit einer angemessenen, aber nicht umfassenden Tiefe ermöglicht: Sie verwendete für die Erhebungsphase ihres Praxisprojekts die Methode der *Teilnehmenden Beobachtung* mit unterschiedlichen Fokuspunkten, um all diese Aspekte zu erheben, was aber – wie sie selber kritisch reflektiert – in der Tiefe an Grenzen stößt (ebd., 177). Dennoch ist diese Entscheidung für ein Primat der Breite vor der Tiefe für die Zielrichtung des Projekts passend und pragmatisch zugleich.

1.2 Die Mikro-Ebene des Projekts

Die Festlegung der Gegenstände auf der Mikro-Ebene erfolgt erst dann, wenn alle anderen Aspekte des Forschungsdesigns festgelegt sind und die Ziele der Erhebung feststehen. Hierbei handelt es sich um die Auswahl der konkreten Texte, Tests und/oder Materialien, die im Projekt genutzt werden: Mit welchem standardisierten Lesetest wird die Leseleistung erhoben? An welchem literarischen Text werden die literarischen Kompetenzen zum Einsatz gebracht? Welche Fragen werden im Leitfadeninterview gestellt? In jedem Fall gibt es mehr als eine Auswahlmöglichkeit, sodass die fundierte Auswahl Zeit und kognitive Energie erfordert. Insbesondere in studentischen Projekten wird dieser Schritt zu oft beiläufig

vollzogen, was zur Folge hat, dass die Ergebnisse der Studie enttäuschend oder wenig aussagekräftig ausfallen. Zugleich vermag die sorgfältige und zielgerichtete Auswahl der Gegenstände die Tragweite des Projekts erheblich zu erweitern.

Die Voraussetzung für die reflektierte Auswahl der Mikro-Gegenstände stellt eine fachwissenschaftliche Sachanalyse dar, die Potenziale und Grenzen verschiedener Materialen herausarbeitet. Auf Grundlage von Fachwissen zu Lesen lassen sich beispielsweise die Schwerpunkte von verschiedenen Lesetests herausarbeiten und auf die Ziele der Untersuchung hin abgleichen, die literaturwissenschaftliche Textanalyse verhilft zu einer Einschätzung über den Grundlagentext und die Kenntnis über den zu erhebenden Gegenstand ermöglicht die Entwicklung von zielgerichteten Leitfragen. Insofern die Projektleiter und Projektleiterinnen über die fachlichen Kompetenzen verfügen – dies sei vorausgesetzt –, handelt es sich bei den Sachanalysen um reine Fleißarbeiten, sodass Versäumnisse in diesem Feld nicht zu entschuldigen sind.

Gute Arbeiten zeichnen sich dadurch aus, dass die Leitenden nur dann zufrieden mit den Gegenständen der Mikro-Ebene sind, wenn diese bestmöglich das abdecken, was für das Projekt benötigt wird. Auf der Suche nach einem Text, mit dem das Metaphern-Verstehen untersucht wird, könnte *irgendein* literarischer Text verwendet werden, da nahezu jeder literarische Text *irgendwelche* Metaphern enthält – das Ergebnis wäre aber vermutlich auch nur *Irgendwas*. Wenn es aber gelingt, einen Text zu identifizieren, der verschiedene Ausprägungen von metaphorischen und symbolischen Ausdrucksweisen in direkter und indirekter Form enthält, der bedeutungsoffen und zugleich altersangemessen gestaltet ist, lassen sich hiermit in beachtlicher Breite tiefgreifende Erkenntnisse über das Verstehen von Metaphern erarbeiten.

Die Auswahl der Mikro-Gegenstände muss entsprechend äußerst sorgfältig vollzogen werden, da sie am Ende darüber entscheiden, ob die zuvor herausgearbeiteten Ziele des Projekts auch erreicht werden können.

2. Festlegung der Probandengruppe

Die Festlegung der Probandengruppe geht mit der Entscheidung über den Gegenstand Hand in Hand und findet in den meisten Projekten bereits mit der Festlegung der Fragestellung statt. Abhängig vom Ziel der Erhebung muss die Probandengruppe sorgfältig anhand spezifischer für die Forschungsfrage notwendiger Faktoren ausgewählt werden: Typischerweise zählt hierzu bei Schülerinnen und Schülern die Klassenstufe, bzw. das Alter der Probandinnen und Probanden, bei Lehrerinnen und Lehrern das Unterrichtsfach sowie spezifische Fähigkeiten oder Erfahrungen. Darüber hinaus ist es möglich – und zumeist sinnvoll – weitere ausgewählte Faktoren als Bedingung für die Aufnahme in den Probandenpool zu definieren, die sich nach der jeweiligen Forschungsfrage richten, etwa das Vorhandensein eines Migrationshintergrunds, Erfahrungen im inklusiven Unterricht, diagnostizierte Lese-Rechtschreib-Schwäche oder andere für die Untersuchung relevante Faktoren. Grundsätzlich sind hier der Kreativität der Forschenden keine

Grenzen gesetzt. Oberste Priorität hat bei der Probandenauswahl nicht die leichte Verfügbarkeit der Probandengruppe – etwa eine Schülergruppe/ein Kollegium, die/das man bereits kennt und die entsprechend leicht verfügbar sind –, sondern die Sinnhaftigkeit für das Forschungsprojekt – auch wenn dies bedeutet, dass ggf. zeitaufwändig neue Kontakte aufgebaut werden müssen.

Für die benötigte Anzahl der Probandinnen und Probanden oder den Zuschnitt der Stichprobe gibt es keine universellen Richtlinien, da sie sich nach den Anforderungen der individuellen Vorgehensweise richtet. Grundsätzlich benötigt man für ein quantitatives Vorgehen eine gewisse Quantität an Teilnehmenden, damit die Aussagen der Erhebung über die Breite der Datengrundlage an Qualität gewinnen, hingegen können qualitativ vorgehende Projekte auch mit wenigen, dafür aber sorgfältig ausgewählten Probandinnen und Probanden auskommen.

Grundsätzlich ist die Auswahl der Stichprobe eine Wissenschaft für sich – und diese Aussage ist mit Blick auf die Existenz des Forschungszweigs *Stichprobentheorie* weniger metaphorisch, als man annehmen könnte –, jedoch gilt es, sich mit den eigenen Untersuchungsobjekten näher auseinanderzusetzen[5], wobei insbesondere die Tragweite, die spätere Ergebnisse für sich beanspruchen können, von der Auswahl und Größe der Stichprobe abhängt[6]. Hierbei gehört auch die Auseinandersetzung mit der Frage, ob die Repräsentativität der Daten angestrebt wird, die bei den in der fachdidaktischen Forschung weit verbreiteten *Klumpenstichproben* nicht gegeben ist. Hierbei ist es notwendig, sich der Metaphorik des Begriffs bewusst zu sein, da wie Diekmann betont, eine Stichprobe „niemals sämtliche Merkmalsverteilungen der Population“ (Diekmann 2010, 430) repräsentiert und eine echte Repräsentation nicht gegeben sein kann. Dennoch besteht die Möglichkeit, durch eine Quotenstichprobe eine hohe Vergleichbarkeit mit einer speziell ausgewählten Grundgesamtheit zu erreichen, wobei dies hohe Anforderungen an die Transparenz der Auswahlkriterien und die systematische Kontrolle der Probanden stellt.

Besondere Formen stellen Querschnitt- oder Längsschnittstudien dar: In Querschnittstudien werden die Daten an einem Messzeitpunkt erhoben und sollen einen möglichst breiten Blick auf den aktuellen Stand eines Untersuchungsaspekts ermöglichen (Beispielfrage: „Wie stark ist die Leseflüssigkeit aller Schülerinnen und Schüler einer Schule ausgeprägt?“). In Längsschnittstudien stehen soziale Wandlungsprozesse im Fokus, die sich über einen Vergleich von periodisch

---

[5] Empfohlen sei an dieser Stelle die gut lesbare Einführung von Diekmann (2010, 373ff.), die umfangreichere Einführung in die Stichprobentheorie im fünften Kapitel von Hartung (2009, 269ff.) oder die vielfältigen Anmerkungen von Bortz/Döring (2006).

[6] Für fortgeschrittene und in Statistik bewanderte Forschende lohnt der Einsatz von Softwarelösungen zur Fallzahlenberechnung, bzw. zur Bestimmung der Tragweite der eigenen Ergebnisse, etwa mit der kostenlosen Software G*Power (http://www.gpower.hhu.de/) oder der im medizinischen Kontext verbreiteten nQuery (https://www.statsols.com/nquery).

durchgeführten gleichen Tests nachweisen lassen (Beispielfragen: „Wie entwickelt sich die Leseflüssigkeit der Schülerinnen und Schüler im Verlauf der Sekundarstufe I?“, „Wie verändert sich die Leseflüssigkeit der Schülerinnen und Schüler zu Beginn der fünften Klasse von 1995 bis 2010?“)[7].

Neben diesen grundsätzlichen Überlegungen zur Bedeutung der reflektierten Probandenauswahl können weitere pragmatische (siehe den Beitrag von König in diesem Band) Aspekte berücksichtigt werden.

3. Festlegung der Erhebungs- und Auswertungsverfahren

In der Geschichte der empirischen Bildungsforschung wurden zahlreiche Erhebungs- und Auswertungsverfahren entwickelt, die darauf abzielen, *spezifische* Probleme zu lösen. Sie helfen dabei, in fundierter und transparenter Weise Daten zu erheben oder auszuwerten, sodass unter Wahrung wissenschaftlicher Qualitätsstandards gearbeitet werden kann. Um diese Standards einzuhalten, bedarf es einer Auseinandersetzung mit den theoretischen Grundlagen der jeweiligen Verfahren, sodass diese Festlegung vermutlich den zeitlich größten Raum einnehmen und die größte Tragweite aufweisen wird.

Diese Verfahren verstehen sich in der Regel – standardisierte Testverfahren seien an dieser Stelle ausgeklammert – als Vorgehensweisen, die in einer spezifischen Art und Weise für die eigene Fragestellung adaptiert werden müssen und keine fertigen Bausteine bereithalten. So kann es als die Regel und nicht die Ausnahme beschrieben werden, dass für die empirische Erhebung eigene Instrumente entwickelt werden müssen, wobei die jeweiligen Verfahren eigene Gesetzmäßigkeiten aufweisen, die bei der Entwicklung zwingend beachtet werden müssen. Gerade bei den bekannteren Erhebungsformen *Interview* und *Fragebogen* geschieht es in studentischen Arbeiten häufig, dass auf einem vorwissenschaftlichen Level, also ohne Kenntnis der wissenschaftlichen Grundlagen, gearbeitet wird: Fragen und Items werden in einem solchen Fall wahllos aneinandergereiht, ohne ihre Gestaltung, Reihung und Auswertungsdimensionen zu reflektieren, was in der Folge zu wenig bis gar nicht aussagekräftigen Ergebnissen führt.

Während im folgenden Kapitel die Bedeutung einer guten Zusammenstellung von Erhebungs- und Auswertungsverfahren in spezifischen Designs näher ausgeführt wird, fällt die Vorstellung ausgewählter Verfahren aufgrund ihrer Vielfältig- und Vielschichtigkeit dem pragmatischen Rahmen dieses Beitrags zum Opfer. Dies begründet sich auch in der Existenz des zweiten Bandes dieser Reihe, der sich

[7] Anschaulich lässt sich dies an der JIM-Studie zeigen: Die Querschnittsstudie erhebt und veröffentlicht seit 1998 jährlich das Medienverhalten von Jugendlichen zwischen 12 und 19 Jahren (Online unter: https://www.mpfs.de). In der Meta-Studie *15 Jahre JIM-Studie* wurden im Sinne einer Längsschnittuntersuchung die Ergebnisse der einzelnen Querschnittsuntersuchungen über den gesamten Erhebungszeitraum nachgezeichnet und so ein Längsschnitt erstellt: https://www.mpfs.de/fileadmin/files/Studien/JIM/2013/15JahreJIMStudie.pdf.

ausschließlich diesem Thema (Boelmann 2018a) widmet und für diese Phase des Forschungsprojekts empfohlen sei. Dennoch sei erneut betont, dass insbesondere mehr als alle anderen Entscheidungen zum Forschungsdesign die Wahl zur Forschungsfrage passender Verfahren darüber entscheidet, ob das Ziel der Erhebung erreicht werden kann oder ob das Projekt scheitert.

3.1 Zusammenstellung der Erhebungs- und Auswertungsinstrumente

Ohne auf konkrete Erhebungsverfahren einzugehen, soll im Folgenden grundsätzlich auf die Bedeutung der Zusammenstellung der Forschungsinstrumente eingegangen werden.

3.1.1 Informationen erheben

Um herauszufinden, wie viele Stunden ein Kind täglich mit dem Lesen von Büchern verbringt, kann man es danach fragen. Man erhält eine Antwort und man weiß nun, wie lange es seiner eigenen Einschätzung nach am Tag liest.

Und jetzt? Was nutzt diese Information?

Einerseits steht nicht fest, ob diese Information korrekt ist – dieses Problem sei vorerst hintangestellt und wird später beim Thema *Triangulation* aufgegriffen –, andererseits lässt sich aus einer Einzelinformation wenig ableiten. Während in schulischen Diagnostiksituationen durch das tägliche Miteinander eine Fülle von Informationen vorliegen, muss im Rahmen empirischer Forschung im Vorfeld der Erhebung sehr genau überlegt werden, welche Daten zur Beantwortung der Forschungsfrage benötigt werden, da neben zeitökonomischen Fragen auch – hier sei auf den Beitrag von Iberer in diesem Band verwiesen – das Gebot der Datensparsamkeit beachtet werden muss. Verschärft wird dies dadurch, dass durch die Anonymisierung der Probandendaten keine Rückschlüsse auf konkrete Personen gezogen werden dürfen.

Welche Forschungsfrage ließe sich mit der eingangs skizzierten Frage zur Erhebung der Lesezeiten beantworten? Einzig die Frage, wie viel die untersuchten Schülerinnen und Schüler täglich nach eigener Aussage lesen. Diese Information könnte mittels statistischer Verfahren ein wenig verfeinert werden, sodass etwa der Mittelwert Auskunft darüber geben könnte, wie viele Probanden mehr oder weniger lesen als der Durchschnitt (siehe zur statistischen Basis den Beitrag von Schmitz in diesem Band). Jedoch würde hiermit nur ein spezifischer Status erhoben, der noch keinen Aufschluss über Zusammenhänge liefert.

Möglich wäre auch eine Verbindung mit repräsentativ erhobenen Vergleichsdaten, etwa den Daten der KIM- (2016) oder JIM-Studie (2017), sodass die Lesezeiten der Probandengruppe mit dem Bundesdurchschnitt abgeglichen werden könnten – vorausgesetzt, die Messskalen wurden so gewählt, dass sie mit den repräsentativen Studien übereinstimmen. Allerdings können diese Daten weder auf einzelne Schüler bezogen werden (Anonymisierungsgebot), noch detaillierter

ausgewertet werden, da weitere Informationen fehlen, die als Differenzierungsfolie dienen.

Alternativ könnte die Erhebung auch zu einem späteren Zeitpunkt wiederholt werden, sodass sich dann über einen Vergleich der beiden Erhebungszeitpunkte eine Entwicklung darstellen ließe – dass sich dies in der Praxis jedoch ein wenig komplexer gestaltet, wird in Kapitel 4 und im Beitrag von Pissarek und Wild deutlich.

3.1.2 Zusammenhänge bestimmen

Sobald mehr als eine Information erhoben wird, ergeben sich Zusammenhänge, die näher bestimmt werden können. Wenn etwa neben den Lesezeiten auch das Geschlecht und das Alter des Probanden abgefragt wird, lassen sich die Daten genauer untersuchen: Unterscheiden sich die Lesezeiten geschlechtsspezifisch? Bestehen Unterschiede in den Lesezeiten zwischen verschiedenen Altersstufen? Auch lassen sich alle drei Informationen zusammenbringen: Gibt es geschlechtsspezifische Ausprägungen der Lesezeiten in verschiedenen Altersstufen? Hieraus ergibt sich für das Design der Studie, dass die Auswahl der Untersuchungsaspekte sorgfältig betrieben werden muss: Nur was erhoben wurde, kann anschließend in die Beantwortung der Fragestellung einfließen.

Es wäre an dieser Stelle insbesondere für Novizen im Bereich der empirischen Forschung wünschenswert, einen beispielhaften Zusammenstellungsfahrplan für die Auswahl der verschiedenen Untersuchungsaspekte vorzulegen. Jedoch ist dies nicht möglich, da sich die für die Untersuchung relevanten Aspekte aus der *Fragestellung* und der *Zielrichtung* des Forschungsprojekts ergeben und somit für jede Erhebung individuell erarbeitet werden müssen.

Grundsätzlich lohnt es sich jedoch, ein Set von zentralen Aspekten und ein Set von präzisierenden Aspekten zusammenzustellen. Als zentral gelten alle Aspekte, die von der Fragestellung direkt fokussiert werden, präzisierend können alle anderen Faktoren sein, die wahrscheinlich oder möglicherweise Einfluss auf die (Ausprägung der) zentralen Aspekte nehmen und somit näher bestimmen können, wieso der zentrale Aspekt bei dem jeweiligen Probanden zur gemessenen Ausprägung gelangt. Im vorliegenden Beispiel könnten neben den Lesezeiten (zentraler Aspekt) auch präzisierende Aspekte wie Alter, Geschlecht, andere Hobbys, Einstellung zu Lesen, Intelligenzquotient, Aufmerksamkeit oder Kreativität abgeprüft werden. In der Folge würde bestimmt, ob und wie sich das Vorhandensein und ggf. die spezifische Ausprägung jedes einzelnen Aspekts auf die Lesezeit auswirkt. Ist die Datenbasis breit genug (siehe hierzu den Beitrag zu quantitativer Forschung von Pissarek in diesem Band), lassen sich aus den erhobenen Daten Korrelationen errechnen, also der Grad des Zusammenhangs beider Werte mit mathematischer Präzision bestimmen (siehe hierzu den Beitrag von Schmitz in diesem Band).

Es lohnt sich, frühzeitig darüber nachzudenken, in welchem Feinheitsgrad bestimmte Aspekte erhoben werden sollen. Es macht beispielsweise einen großen Unterschied, ob das biologische Geschlecht eines Probanden erhoben wird oder

das geschlechtliche Selbstbild (siehe den Beitrag von Maak in diesem Band): Während das biologische Geschlecht mittels eines standardisierten Ankreuz-Items (männlich/weiblich) ausgesprochen zeitökonomisch erhebbar ist – jedoch auch wenig aussagekräftig bleibt –, müsste man für eine gründliche Auseinandersetzung mit dem geschlechtlichen Selbstbild des Probanden ein längeres Interview führen. Einerseits unterscheidet sich der Aufwand der Erhebungen deutlich, andererseits lassen sich aus der differenzierteren Erhebung auch differenziertere Aussagen ableiten. Gleiches gilt für alle oben im Beispiel genannten Aspekte und auch an dieser Stelle ist wieder die Rückbindung an die Fragestellung und das Ziel des Projekts zu beachten. In der Regel gilt: Je feiner die jeweiligen Aspekte durchdrungen werden sollen, desto aufwändiger ist das dazugehörige Testverfahren, weshalb man wiederum pragmatisch abwägen muss, wie man die begrenzte Zeit der Erhebung (entweder durch externe Faktoren wie schulische Vorgaben oder Probanden-interne wie die vorhandene Aufmerksamkeitsspanne) bestmöglich nutzt.

Wenngleich Christian Müller dies in seinem Beitrag in diesem Band ausführlich darlegt, sei an dieser Stelle auf die Methoden-Triangulation und Mixed-Methods-Verfahren verwiesen, die insbesondere bei Einmalerhebungen wertvolle Dienste leisten können: Bei einer Methoden-Triangulation wird ein Aspekt über verschiedene Testverfahren abgeprüft, um die spezifischen Unschärfen einer Methode abzufedern. So könnte die Projektleitung des obigen Beispiels die Lesezeiten nicht nur durch einen *Fragebogen* erheben, sondern die dort getätigten Aussagen im Rahmen eines *problemzentrierten Interviews* hinterfragen und so die Ergebnisse besser absichern – hierdurch könnte sich z.B. zeigen, ob die von einem Probanden angegebene Lesezeit pro Tag realistisch abgeschätzt wurden oder ob eine Verzerrung etwa durch *soziale Erwünschtheit* vorliegt – und in der Folge eine breitere Datenbasis erzeugen, über die auch Erklärungsmuster generiert werden könnten. Im Rahmen einer Mixed-Methods-Erhebung würden qualitative und quantitative Verfahren aufeinander aufbauend miteinander kombiniert werden: So wäre es möglich, in einem ersten Schritt quantitativ einzelne Aspekte zu erheben (z.B. mit einem Fragebogen zum Leseverhalten) und auf Grundlage dieser Daten Typen zu bilden (z.B. *wenig lesende Kinder mit vielen Hobbys* und *viel lesende Kinder mit wenigen Hobbys*), von denen einzelne Repräsentanten im Rahmen einer qualitativen Untersuchung, etwa mittels *narrativer Interviews*, ausführlicher befragt würden.

Über die zielgerichtete Zusammenstellung der einzelnen Testinstrumente lässt sich somit gestalten, welche Ziele die Erhebung verfolgt und welche Tragweite den Ergebnissen zugeschrieben werden kann.

4. Festlegung des Ablaufs

Sobald festgelegt wurde, an was (Gegenstand), an wem (Probandengruppe) und womit (Erhebungs- und Auswertungsverfahren) die empirische Erhebung durchgeführt wird, lässt sich auch der Ablauf des Projekts präzisieren. Grundsätzlich

gestaltet sich der Ablauf eines Forschungsprojekts mit seinen Erhebungen und Auswertungen ebenso individuell wie alle anderen Schritte der Design-Entwicklung, sodass es keine vorgefertigten Ablaufschemata geben kann. Wie zuvor auch entscheidet die Fragestellung des Projekts darüber, wie vorgegangen werden muss. Während König in diesem Band auf die Praxisfragen eingeht, werden hier die theoretischen Aspekte fokussiert, wobei vor allem zwischen einmaliger und mehrmaliger Erhebung unterschieden wird.

4.1 Einmalige und mehrmalige Erhebungen

Um begrifflichen Unschärfen direkt zu Beginn entgegenzuwirken: Einmalige oder mehrmalige Erhebung im Sinne dieses Bandes meint nicht die Frage, wie häufig die Forscherinnen und Forscher an eine Schule oder anderen Erhebungsort reisen müssen, um alle Daten zu generieren – es ist durchaus üblich, dass man auch für eine einmalige Erhebung mehrmals an den Erhebungsort reist (hierzu siehe Kapitel 4.3) –, sondern ob es eines oder mehrerer Messzeitpunkte bedarf, um valide Ergebnisse zu gewinnen. In der Psychologie spricht man auch von Status- und Prozessdiagnostik: Die Statusdiagnostik untersucht den aktuellen Stand zu einem Zeitpunkt, während die Prozessdiagnostik den Verlauf einer Entwicklung in den Blick nimmt.

4.1.1 Einmalige Erhebungen

Zahlreiche aktuelle Forschungsprojekte stellen Fragen, die sich mit einer einmaligen Erhebung von Daten beantworten lassen; etwa dann, wenn es um *Überzeugungen, Wissen, Können, Praktiken* oder andere Formen von *Bestandsaufnahmen* geht. Hierbei ist es unerheblich, zu welchem *Forschungsfeld* die Arbeit gehört – wenngleich gewisse fachspezifische Präferenzen existieren (vgl. hierzu Boelmann 2018b) – und welche *Forschungsmethodik* eingesetzt werden soll, was ein Blick in eine kleine Auswahl von abgeschlossenen Forschungsprojekten zeigt: So fragte etwa Eva Maus (2014), welchen Einfluss Geschlechtermuster in jugendliterarischen Werken auf die Rezeption durch Jugendliche haben. Daniel Scherf (2013) untersuchte, auf welchen Wissensgrundlagen Lehrende Leseförderung betreiben und Jochen Heins (2017) ging der Frage nach, inwieweit ein erhöhtes Ausmaß an instruktionaler Unterstützung in Aufgaben zu komplexen Problemstellungen lernförderlich ist.[8] Reinold Funke (2005) untersucht am Beispiel von Nomen und Verben, über welche Wortartenkenntnisse Schülerinnen und Schüler verfügen, und Josef Payrhuber (1991) erhebt, welche Bedeutung das Drama im regulären Deutschunterricht hat.

[8] Die bis hierhin gestellten Fragen werden in Band 2 dieser Reihe *Erhebungs- und Auswertungsverfahren* (Boelmann 2018a) genutzt, um den konkreten Einsatz verschiedener empirischer Verfahren vorzustellen. Während diese Fragen hier zur Erklärung spezifischer Forschungsdesigns genutzt werden, kann dort nachgelesen werden, wie die Forschenden konkret vorgegangen sind und welche Ergebnisse sie erzielten.

Die Themenfelder dieser Forschungsprojekte sind denkbar unterschiedlich, doch ließen sich all diese Fragen über eine Statusdiagnostik beantworten: Die Forschenden nahmen einen Aspekt in den Fokus ihrer Forschung, bei dem nicht eine lang- oder kurzfristige Entwicklung im Zentrum steht, sondern bei dem über eine Momentaufnahme Antworten gefunden werden konnten.

Diese Momentaufnahmen wurden mit unterschiedlichen Erhebungs- und Auswertungsverfahren realisiert: Maus führte *problemzentrierte Interviews*, die sie mittels *Grounded Theory* auswertete, Scherf ließ Lehrerinnen und Lehrer in *Gruppendiskussionen* miteinander sprechen und nutzte anschließend die *Dokumentarische Methode*, um zu erfahren, welche kollektiven Überzeugungen vorherrschten. Heins ließ Schülerinnen und Schüler *Aufgaben* schriftlich bearbeiten und untersuchte die Ergebnisse mittels *qualitativer Inhaltsanalyse*, wohingegen Funke *standardisierte Tests* für Schülerinnen und Schüler entwarf und Payrhuber *Fragebögen* nutzte, um Informationen von Lehrkräften zu erhalten. Diese Beispiele werfen ein Schlaglicht auf die Vielfältigkeit einmaliger Erhebungen und zeigen doch, dass sich die Wahl des Designs mit seinen Erhebungs- und Auswertungsverfahren alleinig aus der Fragestellung des Projekts ableiten lässt.

Dennoch gibt es neben diesen inhaltlichen auch zeitlich-pragmatische Gründe, die für die Wahl eines einmaligen Erhebungszeitpunkts sprechen und ggf. eine Anpassung der zuvor als sinnvoll erachteten Fragestellung nach sich ziehen müssen:

- Die Arbeit muss innerhalb einer gewissen Frist abgeschlossen werden und der zeitliche Rahmen erlaubt nur einen Erhebungszeitpunkt. Von diesem Problem sind insbesondere studentische Arbeiten zum Studienabschluss oder im Praxissemester betroffen, da ein Prä-/Postdesign in der Regel eine Entwicklungs- oder Interventionsphase benötigt, womit gewisse Zeitspannen zwischen den beiden Messzeitpunkten liegen müssen.
- Der Zugriff auf die Probandengruppe ist beschränkt, z.B. wenn die Schulleitung nicht mehrere Erhebungen erlaubt, um Unterrichtsausfallzeiten zu begrenzen[9].

Diese pragmatischen Erwägungen müssen bei der Planung eines Forschungsprojekts mitbedacht werden, da deren (vorsätzliche) Missachtung schwerwiegende Folgen bei der Durchführung nach sich ziehen kann. Ohne weiter auf pragmatische Fragen einzugehen (verwiesen sei auf König in diesem Band), gibt es jedoch auch schlechte Gründe, um ein Design mit einmaliger Erhebung zu wählen:

- Die einmalige Erhebung erscheint einfacher zu sein als mehrfache Erhebungen.

[9] Dieser Grund kann nur dann gelten, wenn die Probandengruppe nicht ausgetauscht werden kann, da sie Spezifika aufweist, über die andere Gruppen nicht verfügen. Hier gilt es kritisch abzuwägen, ob eine tiefgreifende Veränderung im Forschungsprozess zu rechtfertigen ist.

- Die Menge an erhobenen Daten soll mit einer einmaligen Erhebung begrenzt werden, um den Arbeitsaufwand für Erhebung und Auswertung überschaubar zu halten.

Beide Vermutungen müssen zurückgewiesen werden: Weder muss man für die Planung und Durchführung einer einmaligen Erhebung weniger leisten – weder geistig, noch zumeist logistisch –, noch sagt die Erhebungsform etwas darüber aus, wie viel Arbeitsaufwand in die Durchführung des Projekts investiert werden muss. Hierüber entscheidet insbesondere die Zusammenstellung der einzelnen Erhebungs- und Auswertungsinstrumente, die für die Beantwortung der Fragestellung eingesetzt werden.

Exemplarisch stellt der folgende Beitrag von Boelmann den Einsatz einer einmaligen Erhebung vor, sodass die verschiedenen Auswahl- und Abstimmungsprozesse auch in einem konkreten Beispiel nachvollzogen werden können.

4.1.2 Mehrmalige Erhebungen

Weist die Fragestellung Begriffe wie „verändert“ oder „entwickelt“ auf, legt dies eindeutig fest, dass ein Erhebungszeitpunkt für die Beantwortung nicht genügt. Gleiches gilt bei der Überprüfung von Wirksamkeiten spezifischer Interventionen, etwa eines Lesetrainings. Es bedarf mehrerer Erhebungszeitpunkte, um die dazugehörige Frage zu beantworten. Die Forschungsfrage: „Wie hat sich die Leseflüssigkeit der Schülerinnen und Schüler der Lichtenbergschule im Verlauf der dritten Klasse verändert?“, lässt sich naheliegenderweise nicht über nur eine Erhebung beantworten, es bedarf eines Vergleichs zweier Messungen: eine Messung zu Beginn des für die Erhebung interessanten Zeitraums (im Beispiel: Schuljahresbeginn) und eine zu deren Abschluss (im Beispiel: Schuljahresende). Die Forschungsfrage: „Wie ist der Entwicklungsstand der Klasse 3a der Lichtenbergschule im Bereich Leseflüssigkeit?“, bedarf hingegen nur einer Erhebung – und bestenfalls eines normierten Vergleichswerts, der die Ergebnisse in einen größeren Kontext setzt (vgl. hierzu ausführlich Dube 2018).

Da die Gestaltung einer Interventionsstudie, also ein Design mit mehreren Messzeitpunkten, eigene komplexe Entscheidungen verlangt, wird die Forschung mit einem Prä-/Post-/FollowUp-Kontrollgruppendesign im folgenden Beitrag von Pissarek und Wild ausführlich thematisiert.

4.2 Zeitpunkt oder Zeitspanne der Erhebung

Die Festlegung des Erhebungszeitpunkts markiert den letzten Schritt der Gestaltung des Forschungsdesigns. In manchen Fällen gestaltet es sich einfach, etwa wenn von einer überschaubaren Gruppe (z.B. eine Klasse) eine ebenso überschaubare Menge an Daten (etwa ein Fragebogen zum Leseverhalten) erhoben werden soll. Hier stellt das Einholen der notwendigen Genehmigungen zumeist eine größere Herausforderung dar als die Terminierung und Durchführung der eigentlichen Erhebung, die vermutlich nach einem 15-minütigen Besuch in einer Klasse abgeschlossen wäre.

In anderen Fällen werden die Projektleiterinnen und Projektleiter vor logistische Herausforderungen gestellt, wenn etwa bei einer größeren Gruppe von mehreren Klassen aufwändigere Datenerhebungsverfahren eingesetzt werden sollen, die genutzten Verfahren keine Gruppenerhebung erlauben oder die Leistungsfähigkeit der Probanden erhoben werden soll. In einem solchen Fall muss penibel darauf geachtet werden, dass die Daten nicht durch Planungsversäumnisse korrumpiert werden. Dies gilt insbesondere, wenn ein Vorgehen mit Prä-/Post-/FollowUp-Design angestrebt wird, da hier mehrere Messzeitpunkte obligatorisch vorgesehen sind und die Vergleichbarkeit der Daten existenziellen Charakter erhält.

Auch bei einmaligen Erhebungen ist es grundsätzlich üblich, an mehreren Zeitpunkten Messungen durchzuführen und so eine große Erhebung in mehrere kleinere Erhebungen aufzuteilen, sodass von einer Erhebungszeitspanne gesprochen werden kann. Die verschiedenen Erhebungszeitpunkte werden dann allerdings zu einem Faktor, der reflektiert und bestmöglich kontrolliert werden muss, sollen die erhobenen Daten miteinander vergleichbar bleiben. Da sich Lernzuwächse, körperliche oder geistige Entwicklungen oder Einstellungen nur langsam verändern, wirkt sich ein Abstand von mehreren Tagen zwischen verschiedenen Erhebungszeitpunkten nur wenig auf die Ergebnisse aus. Die individuelle Leistungsfähigkeit variiert hingegen im Verlauf eines Tages erheblich und muss somit stärker reflektiert werden: Wenn eine Untersuchung zur Lernmotivation einer Lerngruppe am Dienstag in der zweiten Stunde deutlich positivere Ergebnisse generiert als ein Folgetest an einem Freitag zum Ende der siebten Stunde, liegt dies vermutlich stärker in einem Uhrzeit- und Wochentagseffekt begründet als in einer grundlegenden Einstellungsveränderung innerhalb der verstrichenen drei Tage.

4.3 Rahmung der Erhebung

Neben zeitlichen Aspekten können sich auch andere Einflüsse als Störfaktoren erweisen: Möglich erweisen sich etwa wechselnde Personen, die die Datenerhebung durchführen, unterschiedliche Erklärungen der Aufgabenstellung oder auch Lerneffekte bei Tests, die sich auf Folgetests auswirken, als Problem. Hierbei ist es notwendig, diese Aspekte zu reflektieren und ihren Einfluss individuell abzuschätzen bzw. häufig auch gegeneinander abzuwägen. Zentral steht hierbei das Verhindern der Verunreinigung der Ergebnisse, was etwa durch die Pilotierung der Instrumente, der Einheitlichkeit der Materialien – von Tests bis hin zu Instruktionen – und auch der Schulung der an der Erhebung und Auswertung beteiligten Personen erreicht werden kann (siehe den Beitrag von König in diesem Band). Diese Maßnahmen müssen bei der Entwicklung des Forschungsdesigns mit eingeplant und ihr Erfolg hinreichend dokumentiert werden.

Eine Abwägung zwischen verschiedenen Einflussfaktoren ergibt sich beispielsweise, wenn Schülerinnen und Schüler im Rahmen eines Projekts interviewt werden: Da die Projektleitung nicht mehrere Interviews gleichzeitig führen kann, muss zwischen einer gleichzeitigen bzw. zeitnahen Erhebung durch mehrere Interviewende oder der zeitlich zerdehnten Interviewerhebungen mit einer Interviewerin/einem Interviewer abgewägt werden. Der Vorteil einer Durchführung aller

Interviews durch eine Person läge in der Einheitlichkeit der personenbezogenen Aspekte der Erhebung: Dies reicht vom äußeren Erscheinungsbild über die Klangfarbe der Stimme bis hin zu nonverbalen Reaktionen auf spezifische Probandenäußerungen. Als Nachteil erwiese sich allerdings unter anderem, dass die Schülerinnen und Schüler bei dieser zerdehnten Form die Möglichkeit erhalten, das Gespräch mit bereits interviewten Mitschülerinnen und Mitschülern über die Interviews zu suchen, was zu Lerneffekten oder anderen Verfälschungen der später getätigten Äußerungen führen kann. Auch ist nicht sichergestellt, wie sich die variierende Erhebungs-(Uhr-)Zeit auf die Interviews auswirkt. Bei einer gleichzeitigen Erhebung mit verschiedenen Erhebenden können diese Zeit-Aspekte aufgefangen werden, jedoch wirken die interpersonellen Aspekte stärker, da hier verschiedene Individuen mit eigenen Verhaltensweisen unterschiedliche Reaktionen seitens der Probandinnen und Probanden auslösen.

Erst nach einem Abwägeprozess kann die Projektleitung über das Vorgehen entscheiden, wobei in beiden Fällen Maßnahmen zur Qualitätssicherung (Standardisierung der Instruktion, Interviewerschulung, Begleitfragen zur Leistungsfähigkeit und Motivation etc.) getroffen werden können und müssen, um die Güte der Daten sicher zu stellen.

5. Fazit

Die Entwicklung des Forschungsdesigns stellt den letzten Schritt vor dem Gang in die Praxis dar. In ihr kommen alle zuvor getätigten Entscheidungen zusammen und decken Fehler in der bisher gelaufenen Planung auf. An diesem Punkt ist es aber noch vergleichsweise konsequenzenfrei möglich, einen Schritt zurückzugehen und Fehlentwicklungen zu korrigieren. Deshalb ist es wichtig, sich ehrlich und kritisch mit der Planung auseinanderzusetzen und Lücken im Fundament zu schließen oder auch zeitintensive grundlegende und richtungsweisende Korrekturen vorzunehmen.

Vor der Hoffnung darauf, dass sich erkannte Schwächen schon *nicht zu sehr* niederschlagen werden, sei gewarnt: Wenn Fehler bei der Entwicklung des Designs nicht wahrgenommen oder ignoriert werden, treten sie bei der Auswertung der Daten noch deutlicher zu Tage – hier ist der entstandene Schaden jedoch deutlich größer, da bis dorthin erhebliche Anstrengungen in das Projekt geflossen und die Fehler zumeist irreversibel sind.

Literatur

Boelmann, Jan M. (Hrsg.) (2018a): Empirische Forschung in der Deutschdidaktik. Band 2: Erhebungs- und Auswertungsverfahren. Baltmannsweiler: Schneider Hohengehren.

Boelmann, Jan M. (Hrsg.) (2018b): Empirische Forschung in der Deutschdidaktik. Band 3: Forschungsfelder. Baltmannsweiler: Schneider Hohengehren.

Diekmann, Andreas (2010): Empirische Sozialforschung – Grundlagen – Methoden – Anwendungen. Reinbeck bei Hamburg: Rowohlt.

Döring, Nicole/Bortz, Jürgen (2006): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 4. Aufl. Berlin: Springer.

Dube, Juliane (2018): Standardisierte Testverfahren. In: Boelmann, Jan M. (Hrsg.): Empirische Forschung in der Deutschdidaktik. Band 2: Erhebungs- und Auswertungsverfahren. Baltmannsweiler: Schneider Hohengehren, 115-130.

Funke, Reinold (2005): Sprachliches im Blickfeld des Wissens. Grammatische Kenntnisse von Schülerinnen und Schülern. Tübingen: Niemeyer.

Hartung, Joachim (2009): Statistik. Lehr- und Handbuch der angewandten Statistik. München: Oldenbourg.

Heins, Jochen (2017): Lenkungsgrade im Literaturunterricht. Zum Einfluss stark und gering lenkender Aufgabensets auf das Textverstehen. Wiesbaden: VS Verlag für Sozialwissenschaften.

Hofmeister, Wernfried/Schwinghammer, Ylva (2015): Das Sparkling Science Projekt Arbeitskoffer zu den Steirischen Literaturpfaden des Mittelalters: Die Welt des Mittelalters als Ausgangspunkt für regionale und digitale Literaturerlebnisse. In: Hofmeister, Wernfried/Schwinghammer, Ylva (Hrsg.): Literatur-Erlebnisse zwischen Mittelalter und Gegenwart. Aktuelle didaktische Konzepte und Reflexionen zur Vermittlung deutschsprachiger Texte. Frankfurt a.M.: Peter Lang, 9-72.

Maus, Eva (2014): Wer (ver)führt zum Lesen? Der Einfluss von Geschlechtermustern auf die Lesemotivation von Jungen und Mädchen. Baltmannsweiler: Schneider Hohengehren.

Medienpädagogischer Forschungsverband Südwest (Hrsg.) (2016): KIM-Studie 2016. Kindheit, Internet, Medien. https://www.mpfs.de/fileadmin/files/Studien/KIM/2016/KIM_2016_Web-PDF.pdf (letzter Zugriff: 01.08.2018).

Medienpädagogischer Forschungsverband Südwest (Hrsg.) (2017): JIM-Studie 2017. Jugend, Information, (Multi-) Media. https://www.mpfs.de/fileadmin/files/Studien/JIM/2017/JIM_2017.pdf (letzter Zugriff: 01.08.2018).

Payrhuber, Franz-Josef (1991): Das Drama im Unterricht. Aspekte einer Didaktik des Dramas. Analysen und empirische Befunde – Begründungen – Unterrichtsmodelle. Rheinbreitbach: Dürr und Kessler.

Scherf, Daniel (2013): Leseförderung aus Lehrersicht. Wiesbaden: VS Verlag für Sozialwissenschaften.

Schwinghammer, Ylva (2015): Alte Sprache – schwere Sprache? Empirische Erhebungen und praktische Erfahrungen zum Einsatz mittel- und frühneuhochdeutscher Texte im Unterricht. In: Hofmeister, Wernfried/Schwinghammer, Ylva (Hrsg.): Literatur-Erlebnisse zwischen Mittelalter und Gegenwart. Aktuelle didaktische Konzepte und Reflexionen zur Vermittlung deutschsprachiger Texte. Frankfurt a.M.: Peter Lang, 147-173.

Schwinghammer, Ylva (2018): Teilnehmende Beobachtung. In: Boelmann, Jan M. (Hrsg.): Empirische Forschung in der Deutschdidaktik. Band 2: Erhebungs- und Auswertungsverfahren. Baltmannsweiler: Schneider Hohengehren, 165-178.

JAN M. BOELMANN

Forschungsdesigns mit einmaliger Erhebung

Zur Überprüfung von aktuellem Stand und Zusammenhängen[1]

1. Zum Rahmen: Statusdiagnostik und Korrelationsstudien

Im vorangegangen Beitrag finden sich zahlreiche Beispiele für Forschungsprojekte, in denen mit einer einmaligen Erhebung ein aktueller Stand (Beispiele Scherf 2013; Funke 2005; Payrhuber 1991) oder Zusammenhänge (Beispiele: Maus 2014; Heins 2017) erhoben werden. Um den Leserinnen und Lesern einen Eindruck von den Entscheidungsprozessen bei der Entwicklung eines Forschungsprojektes zu vermitteln, soll in diesem und dem folgenden Beitrag (siehe den Beitrag zu Prä-/Post-/Follow Up-Designs von Pissarek/Wild in diesem Band) die konkrete Entwicklung und Durchführung eines entsprechenden Projekts skizziert werden. Dies geschieht hier nach einer allgemeinen Vorstellung verschiedener Erhebungsdesigns mit einmaliger erhebung unter Rückgriff auf meine Dissertation ‚Literarisches Verstehen mit narrativen Computerspielen' (Boelmann 2015), die bereits im Beitrag zur Entwicklung der Fragestellung (siehe den Beitrag von Boelmann in diesem Band) exemplarisch herangezogen wurde.

1.1 Arten des Forschungsdesigns mit einmaliger Erhebung

1.1.1 Statusdiagnostik

Grundsätzlich lassen sich verschiedene Herangehensweisen an ein Design mit einmaliger Erhebung wählen: Zur Erhebung eines aktuellen Standes wird in der Regel ein Test zu einem Faktor ausgewählt bzw. entwickelt und dieser dann durchgeführt. Was in diesem Satz sehr einfach klingt, ist in der Praxis deutlich komplexer, da nur wenige Informationen objektiv und in Reinform verfügbar sind. Hierbei hängt die Verfügbarkeit stark von den Untersuchungsobjekten ab: Während die Arbeit an Materialien, etwa an einem Textkorpus, auf Grund ihres

[1] Dieser Beitrag stellt exemplarisch den gesamten Prozess einer empirischen Erhebung vor und konkretisiert somit Aspekte, die in den weiteren Beiträgen dieses Bandes vorgestellt werden. Insbesondere seien hier die Beiträge von Boelmann zu Fragestellung und Forschungsdesign, Pissarek zu quantitativer Forschung, König zu Planung und Vorbereitung empirischer Erhebungen und Schmitz zu statistischen Grundkenntnissen genannt, auf die im Folgenden nicht erneut explizit verwiesen wird.

statischen Charakters vergleichsweise unkompliziert verlaufen kann – ein vorliegender Text verändert seinen Inhalt und seine Form nicht im Angesicht seiner Auswertung –, können Forscherinnen und Forscher nicht auf alle anvisierten innerpersonellen Faktoren der Probandinnen und Probanden direkt zugreifen, was menschliche wie forschungstheoretische Gründe hat[2]. Beides sei am Beispiel der *Lesekompetenz* kurz veranschaulicht: Forschungstheoretisch handelt es sich beim Faktor *Lesekompetenz* nicht um eine normative Größe, sondern um ein Konstrukt. Das heißt, sie lässt sich nicht eindeutig messen, wie etwa Pulsrate oder Temperatur, sondern es wird anhand spezifischer Merkmale in einem Konstrukts definiert, was im jeweiligen Projekt unter Lesekompetenz verstanden wird. Aus diesem Grund liegen verschiedene Definitionen hierfür vor (z.B. Baumert et al. 2001; Hurrelmann/Groeben 2006; Rosebrock/Nix 2017), die unterschiedliche Zielaspekte fokussieren und mit unterschiedlichen Verfahren erhoben werden können. Standardisierte und insbesondere auch selbstentwickelte Tests müssen deshalb die ihnen zu Grunde liegenden Konstrukte offenlegen. Zudem reagieren menschliche Probanden – anders als Materialien – feinfühlig auf unterschiedliche Stimuli, seien sie der Fragestellung, der Erhebungsinstrumente oder dem Erhebungskontext geschuldet, sodass die Güte der erhobenen Informationen nicht zwingend vorausgesetzt werden kann.

Grundsätzlich lassen sich drei Formen von Informationen unterscheiden:

1. Allgemeine Informationen, die unzweifelhaft vorliegen (z.B.: Alter, biologisches Geschlecht, ggf. Medienbesitz[3]).
2. Möglicherweise sensible Informationen, die ggf. vom Probanden eingeschätzt werden müssen (z.B. sexuelle Orientierung, Mediennutzungszeiten, spezifische Fähigkeiten und Einstellungen etc.).
3. Informationen, derer sich der Proband nicht bewusst ist (z.B. Leistungsfähigkeit: Grad der Leseflüssigkeit, orthographische Kenntnisse; Vorwissen etc.).

Während Aspekte der *ersten Kategorie* direkt erfragt werden können, bedarf es bei der Erhebung von Informationen der zweiten oder dritten Kategorie einer gewissen Finesse, um einen menschlichen Probanden zur Offenlegung der gewünschten Informationen zu veranlassen. Insbesondere bei der *zweiten Kategorie* spielt der Störfaktor *Soziale Erwünschtheit* eine große Rolle, da Probanden dazu neigen, im Sinne des Gemocht-Werdens oder Gut-Seins, ihre Antworten an die von ihnen vermuteten Erwartungen des Testleitenden anzugleichen. Beispielhaft zeigt sich das anhand einer Längsschnittauswertung zur Vorlesestudie der Stiftung Lesen, DIE ZEIT und der Deutschen Bahn Stiftung (seit 2007):

---

[2] Schnell, Hill und Esser (2011) widmen sich ausführlich möglichen Störfaktoren 207ff.

[3] Grundsätzlich ist der Medienbesitz einzelner Personen eindeutig bestimmbar, jedoch kann die Offenlegung von Besitzverhältnissen in sozialen Kontexten schambehaftet und somit doch sensibel sein, etwa wenn die Person deutlich mehr oder weniger besitzt als der Durchschnitt der sozialen Referenzgruppe. Dies muss im Einzelfall kritisch hinterfragt werden.

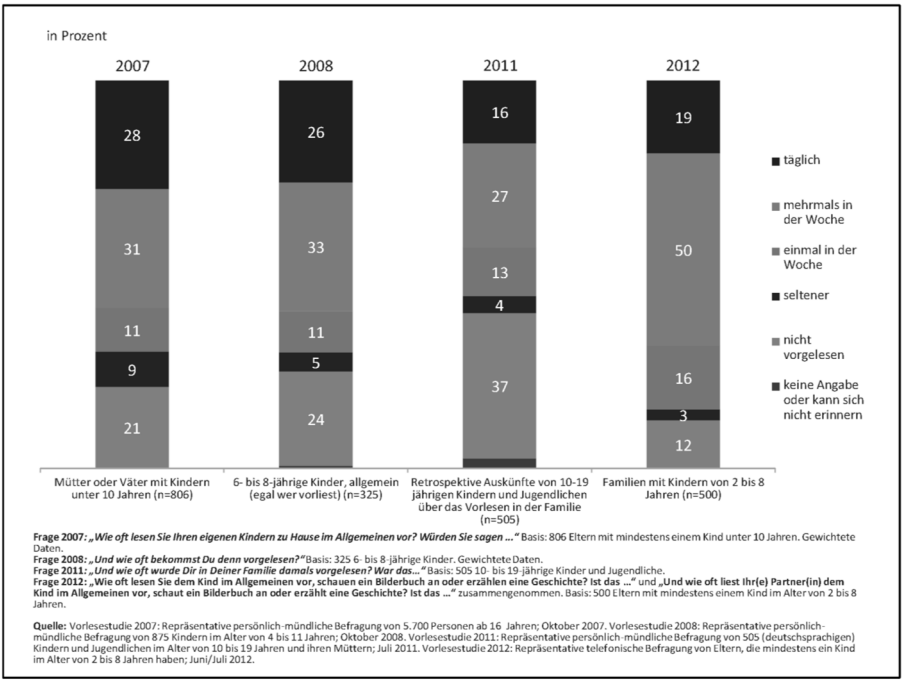

Abb. 1: Problematische Parallelstellung unterschiedlicher Items am Beispiel von ‚Vorlesehäufigkeit in der Familie nach Angaben von Eltern und Kindern' (Ehmig/Reuter 2013, 29)

Vergleicht man die Entwicklung zwischen den Jahren 2011 und 2012, läge die Vermutung nahe, dass hier eine außergewöhnliche Veränderung in der Vorlesehäufigkeit stattgefunden hat. Ein Blick auf das Kleingedruckte offenbart jedoch, dass 2011 Kinder und Jugendliche zu ihren *Erfahrungen*, 2012 Eltern zu ihrem *Vorleseverhalten* befragt wurden. Davon abgesehen, dass die empirische Güte der Schlussfolgerung durch die Parallelstellung zweier unterschiedlicher Faktoren in Zweifel gezogen werden muss, fällt auf, dass die Eltern höhere Lesezeiten angeben als die befragten Kinder – ein Einfluss der sozialen Erwünschtheit (Glaubenssätze: ‚Vorlesen ist gut. Als gutes Elternteil sollte ich vorlesen.') liegt nahe. Auch zeigen sich ähnliche Effekte, wenn sich das Befragungsformat (telefonisch vs. persönlich) ändert. Vermutlich möchten die Probandinnen und Probanden im direkten Kontakt mit den Testleitenden ungern zugeben, ein wenig sozial gewünschtes Verhalten zu zeigen, während über das Telefon eine gewisse Distanz besteht, die die Offenlegung als sensibel wahrgenommener Sachverhalte erleichtert. Dies soll nicht als Plädoyer für distanzierte oder nicht-persönliche Befragungen verstanden werden, aber auf die Notwendigkeit der bewussten Item-Konstruktion hinweisen. Hier gilt es für Forscherinnen und Forscher, Indikatoren sozialer Erwünschtheit abzumildern oder alternative Erfassungsmethoden in Betracht zu ziehen: Anstatt Zielaspekte direkt abzufragen, lohnt es sich, Umwege

für bessere Ergebnisse in Kauf zu nehmen. Dies sei an einem Beispiel verdeutlicht: Nur die allerwenigsten Schülerinnen und Schüler würden die Frage: „Hegst Du Sympathien für ein nationalsozialistisches Weltbild?“, zustimmend beantworten, da ein offenes Bekenntnis hierzu einerseits einem gesellschaftlichen Tabu gleich kommt und andererseits das Selbstbild der meisten Schülerinnen und Schüler eine solche Eigencharakterisierung nicht zulässt; verlässlicher wäre hier ein indirekter Zugang, etwa über Items der Bauart: „Wie sehr stimmst Du folgenden Aussagen zu?“, in deren Folge kontroverse Positionen präsentiert werden, über die sich das Weltbild des Probanden erschießen lässt[4].

Bei diesem Vorgehen ist die Grenze zur *dritten Kategorie* fließend, da Selbsteinschätzungen immer fehleranfälliger sind als fundierte Diagnostik, die für die Erhebung nicht bewusster Informationen notwendig ist. Eine Voraussetzung stellt das Vorliegen von Testverfahren dar, die im Rahmen der Entwicklung des Forschungsdesigns auch selbst entwickelt werden können, bzw. in den meisten Fällen sogar entwickelt werden müssen.

1.1.2 Korrelationsstudien

Zur Ermittlung eines Zusammenhangs zwischen zwei Faktoren bedarf es verschiedener Variablen, die in der empirischen Forschung als *unabhängige* und *abhängige Variablen* benannt werden. Verkürzt gesagt, stellt die *unabhängige Variable* die vermutete *Ursache* dar, während die *abhängige Variable* die *Wirkung* repräsentiert. Möchte man etwa der Frage nachgehen, wie sich der sozioökonomische Status einer Familie (unabhängige Variable) auf die Lesekompetenz der Kinder (abhängige Variable) auswirkt, müsste man beide Variablen bei einer großen Gruppe erheben und mit statistischen Verfahren (siehe den Beitrag von Schmitz in diesem Band) eine Korrelation berechnen. Dass es einen solchen Zusammenhang gibt, wurde in verschiedenen Studien seit *PISA 2000* mehrfach repliziert, doch zeigen sich an diesem Beispiel zwei wichtige Folgeaspekte, die es bei der Verbindung zweier Variablen zu beachten gilt: Erstens ergibt eine Korrelation noch keine Kausalität. Dies bedeutet, dass zwar ein *Zusammenhang* zwischen sozioökonomischem Status und schlechter Lesekompetenz besteht, der schwache sozioökonomische Status aber nicht *den Grund* für die schlechte Lesekompetenz darstellt. Wäre dem so, würde eine finanzielle Besserstellung von Familien leseförderlich wirken – ein offenkundiger Trugschluss. Auch zeigt sich besagter Zusammenhang in anderen Teilnehmerstaaten der PISA-Studie nicht oder weniger stark ausgeprägt als in Deutschland, sodass – und hierbei handelt es sich um den zweiten beachtenswerten Folgeaspekt – andere Variablen berücksichtigt werden müssen, um dieses Problem angemessen zu erklären. Auf Grund ihrer Komplexität lassen sich die wenigsten Probleme in einfachen Ursache-/Wir-

[4] Zugleich und erneut zeigt sich an diesem Beispiel die Bedeutung des Konstrukts, das der Erhebung zu Grunde liegt: Ebenso wie bei der oben genannten Lesekompetenz müsste im Rahmen dieses Projekts deutlich herausgestellt werden, was unter dem Konstrukt ‚nationalsozialistisches Weltbild‘ verstanden wird.

kungszusammenhängen beschreiben, sodass die Arbeit mit verschiedenen unabhängigen Variablen in vielen Fällen zu empfehlen ist. Nur so lassen sich die Fragen klären, *warum* ein so ausgeprägter Zusammenhang von sozioökonomischer Stellung der Familie in Deutschland und der Lesekompetenz existiert oder *welche anderen Faktoren* die Lesekompetenz noch bedingen und – weitergehend – welche Wechselwirkung diese mit dem sozioökonomischen Status aufweisen.

Die meisten Qualifikationsarbeiten widmen sich entsprechend weniger komplexen Themenstellungen als der Bestimmung aller Einflussgrößen auf die Lesekompetenz, sondern isolieren einzelne Variablen und beobachten die Wirkung, die durch ihre Veränderung hervorgerufen wird. Diese gilt es in der Planung herauszuarbeiten und auch spezifische andere Einflussvariablen zu identifizieren, sodass diese ebenfalls erhoben oder eliminiert oder zumindest kontrolliert werden können.

1.2 Zur exemplarischen Studie

Die Entwicklung eines Forschungsdesigns mit einmaliger Erhebung wird im Folgenden an der 2015 veröffentlichten Dissertationsstudien *Literarisches Verstehen mit narrativen Computerspielen*[5] vorgestellt. Diese begegnet dem Desiderat, dass Computerspiele, obwohl sie in der Lebenswelt von Schülerinnen und Schülern eine zentrale Rolle einnehmen (vgl. KIM 2016), bis dahin nicht als Unterrichtsgegenstände wahrgenommen wurden. Als die zentralen Gründe hierfür galten:

1. die mangelnde Medienausstattung der Schulen,
2. die Computerspielferne der aktuellen Lehrergeneration und
3. fehlende oder nicht ausreichend vorhandene Konzepte der Computerspielnutzung im didaktischen Kontext. (Boelmann 2015, 15)

Während die ersten beiden Punkte einem quasi-evolutionären Prozess unterliegen, deren Problemlösung nicht im Rahmen einer wissenschaftlichen Studie erfolgen kann, stellt die grundlegende Fundierung des Computerspiels als Lehr- und Lernmedium eine Aufgabe der universitären Mediendidaktik dar. Als Ausgangspunkt hierfür stand die Frage (nicht Fragestellung – siehe zur Unterscheidung den Beitrag zur Fragestellung in diesem Band), wie Computerspiele für einen curriculumsadäquaten Einsatz im Literaturunterricht aufbereitet werden müssen, damit alle Schülerinnen und Schüler gleichermaßen von ihrem Einsatz profitieren könnten. Schnell wurde klar, dass das Projekt sowohl einen theoretischen, einen schulpraktischen und auch einen empirischen Teil benötigen würde.

[5] Selbstverständlich müssen in einem solchen Beitrag zahlreiche Aspekte und Entscheidungen verkürzt dargestellt werden. Umfassende Informationen können der Monographie (Boelmann 2015) entnommen werden.

2. Vorarbeiten

Wie bei allen Forschungsarbeiten stellt die Auseinandersetzung mit dem Forschungsdiskurs die Grundvoraussetzung für die Entwicklung eines Forschungsdesigns dar. Entsprechend wurde der Forschungsstand zu den beiden Referenzgebieten *Computerspielforschung* und *literarisches Verstehen* umfassend aufgearbeitet und hierauf aufbauend Anknüpfungspunkte und Desiderate identifiziert. Als zentrale Fragestellung wurde festgelegt:

1. Eignen sich Computerspiele als Gegenstände literarischen Verstehens?

Anschließend wurde sie in weitere Unterfragen ausdifferenziert, die im Folgenden mit den sich daraus ergebenden Aufgaben vorgestellt werden:

2. Wie lässt sich literarisches Verstehen modellieren, damit es auch intermedial erfasst werden kann? (Aufgabe: Konstrukt klären)
3. Nach welchen Kategorien lässt sich das Feld Computerspiele systematisieren, um grundsätzlich für das literarische Verstehen geeignete und ungeeignete Computerspiele voneinander zu separieren? (Aufgabe: Gegenstand bestimmen)
4. Wie kann der Einsatz von narrativen Computerspielen zum Erreichen curricularer Ziele konkretisiert werden? (Aufgabe: heuristische didaktische Modellierung)
5. Welche Anforderungen an die literarische Kompetenz stellen narrative Computerspiele im Vergleich zu literarischen Texten? (Aufgabe: empirische Überprüfung)

Insbesondere bei den Unterfragen 2 und 3 traten bestehende Desiderate deutlich zu Tage: Im Bereich des literarischen Verstehens (Frage 2) ergab sich das Problem, dass die bis dahin entwickelten Konzepte und Modellierungen literarischen Verstehens weder die valide Messung literarischer Kompetenz erlaubten, noch die Einbeziehung nicht-printbasierter Medien vorsah. Im Bereich der Computerspielforschung (Frage 3) wurde das Fehlen einer Systematisierung des Gegenstandsbereichs in narrative (und somit der Hypothese nach für den Deutschunterricht geeignete) und wenig-narrative Spiele herausgearbeitet, ebenso existierte bis Studienbeginn kein dem Gegenstand angemessenes und erzähltheoretisch fundiertes Beschreibungsvokabular für die narrative Analyse von Computerspielen.

Im theoretischen Teil der Arbeit wurden diese drei zentralen Desiderate erfüllt, indem erstens eine narrative Genreeinteilung erarbeitet, zweitens ein narrativ-orientiertes Strukturebenenanalysemodell für Computerspiele konzipiert und erprobt sowie drittens mit dem *Bochumer Modell literarischen Verstehens* eine Systematisierung und Modellierung literarischer Verstehensprozesse entwickelt wurde, das sowohl literarische Kompetenzen valide messen, als auch für Gegenstände verschiedener medialer Basierung offen stehen sollte.

Die vierte Fragestellung mündete im schulpraktischen Teil der Arbeit, in dem für verschiedene Altersstufen Einsatzszenarien von Computerspielen im Literaturunterricht skizziert wurden. Die fünfte Fragestellung sollte empirisch angegangen werden und steht im Weiteren im Zentrum der Betrachtung.

2.1 Konkrete Forschungsfragen festlegen

Durch die konkrete Formulierung der Forschungsfrage werden spezifische Aspekte in den Fokus der Betrachtung gerückt und andere ausgespart. Entsprechend steht die bewusste Entscheidung für eine Formulierung zentral für das spätere Design der Studie. Betrachtet man die Fragestellung der empirischen Erhebung „Welche Anforderungen an die literarische Kompetenz stellen narrative Computerspiele im Vergleich zu literarischen Texten?“, treten verschiedene Aspekte hervor. Zum einen stützt sich die Untersuchung auf die Hypothese, dass literarische Kompetenz die zentrale Voraussetzung für das Verstehen von literarischen Texten darstellt und sich folglich in den Verstehensleistungen des Rezipienten abbildet. Zum anderen wurde die Frage so formuliert, dass sie durch den Vergleich zweier Gegenstände in einer einmaligen Erhebung beantwortet werden kann: Der Einsatz *literarischer Kompetenz* beim Verstehen der beiden Gegenstände *literarischer Text*‘ und *narratives Computerspiel* wird erhoben und das Ergebnis verglichen. Wäre die Frage alternativ so formuliert worden: „Lässt sich mit narrativen Computerspielen literarische Kompetenz fördern?“, hätte der empirische Nachweis ein Prä-/Post-Design erfordert, da erst die Messung eines nachgewiesenen Lernzuwachses die Fragestellung valide beantwortet hätte. Zugleich wäre die Frage nach der qualitativen Eignung – also ob literarische Verstehensprozesse mit Computerspielen besser oder schlechter angeregt werden können als im Literaturunterricht mit printbasierten Medien – ausgespart geblieben.

Die in der Frage zentral stehenden *Anforderungen* an die literarische Kompetenz nimmt wiederum das Konstrukt *Literarische Kompetenz* in den Blick, da hier die Hypothese zu Grunde liegt, dass sich diese beim Verstehen von Narrationen unabhängig von der medialen Basierung der Lerngegenstände zeigt. Es wird in der Erhebung somit über die Leitfrage hinaus geklärt werden müssen, ob die These der Universalität literarischer Kompetenzen im Sinne des Bochumer Modells haltbar ist, oder ob sich gegenstandsspezifische Ausprägungen literarischer Kompetenz nachweisen lassen, die nicht über unterschiedliche Anforderungen der Gegenstände bedingt sind.

Für die Beantwortung der Leitfrage wurde somit in der Folge ein Vergleich zwischen einem literarischen Text und einem narrativen Computerspiel angestrebt, der Unterschiede in den Kompetenzpotenzialen verdeutlichen sollte.

2.2 Konstrukt klären

Für die Messung literarischer Kompetenz stand das Bochumer Modell Pate, das davon ausgeht, dass literarische Kompetenz nicht in Reinform messbar ist, sondern nur über die literarische Performanz – also eine Äußerung zu einem literarischen Verstehensprozess (etwa durch einen geschriebenen Text oder eine mündliche Äußerung) – erschlossen werden kann. Als Bedingungsfaktoren gelten die *Gegenstandsschwierigkeit* – zu einem anspruchsvollen Text ist es schwieriger, seine Kompetenz anzubringen als bei einem leichten –, die *Aufgabenschwierigkeit* – eine komplex formulierte Aufgabe ist schwieriger zu beantworten als eine einfach formulierte – und die *individuelle körperliche und geistige Verfassung.*

Literarische Kompetenz =	Literarische Performanz – (Gegenstandsschwierigkeit + Aufgabenschwierigkeit + individuelle Bedingungsfaktoren)

Theoretisch ließe sich die literarische Kompetenz durch die Erhebung der literarischen Performanz, der Text- und der Aufgabenschwierigkeit sowie aller individuellen Bedingungsfaktoren unter dem Einsatz dieser Formel bestimmen. In der Praxis stößt dieses Vorgehen allerdings an Grenzen, da bislang weder die Schwierigkeit eines Textes objektiv und empirisch valide ermittelt kann[6], noch die Vielzahl individueller Bedingungsfaktoren bestimmt und erhoben werden können. Ähnliches gilt für die Aufgabenschwierigkeit und die literarische Performanz, sodass ein Vorgehen, wie das Geplante, die Entwicklung von Testverfahren für alle vier Faktoren nach sich ziehen müsste.

Um das Verfahren zu vereinfachen und es dennoch weniger anfällig für Störfaktoren zu gestalten, wurden einige der benannten Faktoren nicht gemessen, sondern lediglich kontrolliert, bzw. konstant gehalten. Hierfür wurde ein direkter Vergleich zwischen einem literarischen Text und einem narrativen Computerspiel, deren Textschwierigkeit sich entsprechen muss, angestrebt. Alle weiteren Faktoren wurden in beiden Testdurchläufen so weit identisch gehalten, dass sich ihr Einfluss auf beide Testergebnisse in gleichen Maßen auswirkt und somit als kontrolliert gelten kann.

Diese Entscheidung zog nach sich, dass zwar von der erhobenen literarischen Performanz nicht auf einen genauen Wert der literarischen Kompetenz geschlossen werden konnte, aber ein Vergleich beider Performanz-Werte es ermöglichte, mithilfe ihrer Differenz die Unterschiede in den Kompetenzpotenzialen beider Gegenstände aufzuzeigen. Über den Vergleich der Ergebnisse miteinander konnten Aussagen über das Verhältnis der Verstehensfähigkeit der unterschiedlichen Medien geklärt und Erkenntnisse über das Vorhandensein einer unabhängigen, übergreifenden Größe, namentlich der literarischen Kompetenz und ihrer Teilkompetenzen, getroffen werden.

Um die Fragestellung zu beantworten, mussten neben den Verstehensleistungen weitere mögliche Einflussfaktoren hinterfragt werden, die sich in den Messergebnissen der unterschiedlichen Medien widerspiegeln könnten. Hierbei fiel die Wahl auf die *Lesefähigkeit* als Grundvoraussetzung für die Rezeption von literarischen Printtexten, die *Konzentrationsfähigkeit* als Bedingungsfaktor für den Lese- und Spielprozess sowie das *Mediennutzungsverhalten*, um Erfahrungen der

6 „Entgegen der Alltagsintuition lässt sich die Verständlichkeit eines Textes nicht allein durch bestimmte objektiv feststellbare Textmerkmale wie etwa Wortschwierigkeit, Wortlänge, Satzlänge oder Satzkomplexität bestimmen, sondern sie erfordert neben der Berücksichtigung der semantischen Struktur und der Organisation von Textinhalten immer auch den Rückgriff auf das konkrete Verstehen eines Textes durch einen Leser bzw. eine Leserin.“ (Christmann 2004, 33). Siehe zum aktuellen Forschungsstand den Beitrag von Frickel in Band 3 dieser Reihe (Frickel 2018).

Probanden mit und Einstellungen zu den einzelnen Medien einschätzen und kontextualisieren zu können. Gleichzeitig bergen diese Messungen Erkenntnispotenziale, die über die zentrale Leitfrage hinausgehen und eine Validierung der Grundannahmen des Bochumer Modells ermöglichen.

2.3 Testinstrumente festlegen

Um die verschiedenen Ziele der Erhebung zu erreichen, wurde eine Batterie an Testinstrumenten benötigt, die die verschiedenen Aspekte erheben und zugleich wirksam, aber auch zeitsparend eingesetzt werden konnten. Während für die *Lesefähigkeit* (gemessen mit dem *SLS 5-8*) und die *Konzentrationsfähigkeit* (gemessen mit dem *d2-Aufmerksamkeits- und Belastungstest*) standardisierte Testverfahren eingesetzt werden konnten, deren Wirksamkeit erwiesen und deren Vorgehen standardisiert ist, mussten die Erhebungsinstrumente für die Ermittlung personenbezogener Daten, zur Mediennutzung und zur Messung der literarischen Kompetenz neu entwickelt werden.

Da für das exemplarische Forschungsdesign die Vorstellung der Ergebnisse nicht im Detail benötigt werden und in den Folgekapiteln insbesondere der selbstentwickelte Kompetenztest im Zentrum steht, wird an dieser Stelle neben der Testvorstellung auch die Aufbereitung der später erhobenen Daten skizziert, um ihren Nutzen herauszustellen.

2.3.1 Das Salzburger Lesescreening (SLS 5-8)

Für die Erhebung der Lesefähigkeit fiel die Wahl aus mehreren Gründen auf das *SLS 5-8*: Zum einen misst das *SLS 5-8* mit der basalen Lesefertigkeit exakt das, was für die Kontrolle der Rezeptionsfähigkeit benötigt wird. Zudem haben die Ergebnisse eine Voraussagekraft in Bezug auf die Lesetest-Ergebnisse der *PISA-Studie*, was die Generierung weiterer Hypothesen im Rahmen der Auswertung ermöglicht.

Darüber hinaus ist die Durchführung im hohen Maße ökonomisch, da die Durchführungsdauer mit allen Rahmenhandlungen, wie dem Instruieren der Schülerinnen und Schüler sowie dem Austeilen und Einsammeln der Testbögen, nur zehn Minuten beträgt und von allen Schülerinnen und Schülern der Klasse zeitgleich durchgeführt werden kann (vgl. Auer/Gruber/Mayringer/Wimmer 2005, 4). Zudem ist das *SLS 5-8* standardisiert und es liegen Normtabellen zu Vergleichsstichproben für verschiedene Schularten und Jahrgangsstufen vor; seine Aussagekraft ist demnach empirisch gesichert.

Die Methode des Tests ist vergleichsweise einfach: Die Schülerinnen und Schüler lesen in begrenzter Zeit möglichst viele inhaltlich sehr einfache Sätze, wie beispielsweise „Markus ist ein bekannter Mädchenname“, und beurteilen deren Wahrheitsgehalt als Nachweis des Verstehens mit „Richtig“ oder „Falsch“. Die Anzahl innerhalb von drei Minuten richtig beantworteter Aussagen ergibt den Datenrohwert. Anhand dieses Rohwertes können die Leistungen mit den Normtabellen abgeglichen und ein Lesequotient erstellt werden.

Die Güte des Testverfahrens ist ausgesprochen hoch. Die Paralleltestreliabilität, also die Verlässlichkeit des Testverfahrens, beträgt .89. Die Validität wird mit .78 beziffert. Die Konsistenz des Tests ist ebenfalls sehr hoch und er weist keine Boden- oder Deckeneffekte auf (vgl. Auer/Gruber/Mayringer/Wimmer 2005, 5f.).

Was leistet das *Salzburger Lesescreening 5-8*?
Was leistet das *Salzburger Lesescreening 5-8* nicht?

Das *SLS 5-8* misst die basalen Lesefertigkeiten der Schülerinnen und Schüler. Durch den Fokus auf die technischen Voraussetzungen des Leseprozesses kann für den literarischen Verstehenstest abgeklärt werden, ob eine Probandin oder ein Proband grundsätzlich in der Lage ist, den Text zu lesen und somit die Eingangsvoraussetzung für die Vergleichbarkeit erfüllt. Zudem können durch die Voraussagekraft des *SLS 5-8* auf die Ergebnisse der *PISA-Studie* weitere Hypothesen auf das Wesen der literarischen Kompetenz abgeleitet werden.

Der *SLS 5-8* bietet keine diagnostische Hilfe und liefert keine Aussagen über die Ursachen schlechter Leseleistungen. Zudem können keine Aussagen über das Textverstehen direkt, sondern nur über die technischen Aspekte des Lesens getroffen werden.

Nach der zeitökonomischen Auswertung von wenigen Minuten pro Test liegen die Ergebnisse in Form von Lesequotientwerten vor, wobei der Wert 100 den durchschnittlichen Wert aller Schülerinnen und Schüler der gleichen Schulform und Jahrgangsstufe beschreibt – man erhält über die Normtabellen somit ein Bild, das über die soziale Bezugsnorm einen weiteren Horizont verleiht. Ein Proband mit einem Wert über 100 liest entsprechend besser als seine Vergleichspopulation, ein Wert unter 100 deutet auf eine schlechtere Leseleistung hin.

Die Probandinnen und Probanden der Studie erreichten folgende Werte:

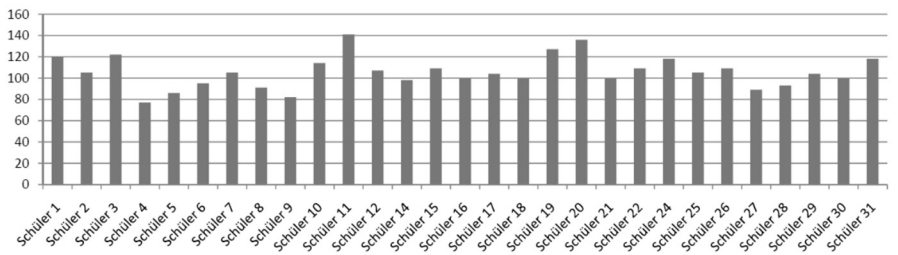

Abb. 2: Ergebnisse des SLS 5-8, aufbereitet

Einerseits können die so gewonnenen Daten mit den Ergebnissen der anderen Tests mit statistischen Mitteln verglichen werden (siehe hierzu den Beitrag von Schmitz in diesem Band), mit Hilfe dieser Darstellung lassen sich andererseits aber auch besonders leistungsstarke Schülerinnen (Proband(in) 11, 19, 20) ebenso identifizieren wie leistungsschwächere (Proband(in) 4, 5, 9), um eine Referenz für eine genauere qualitative Untersuchung aller Ergebnisse zu erhalten. Hierbei könnte der Frage nachgegangen werden, ob diese herausragenden Fälle besondere Strategien beim Verstehen literarischer Gegenstände aufweisen oder ob es neben den statistischen Kennwerten auffällige Einzelfallbeobachtungen gäbe, die in anschließenden Studien vertieft werden müssten.

2.3.2 Der d2-Aufmerksamkeits- und Belastungstest

Der d2-Aufmerksamkeits-Belastungstest ist erprobt und seine Aussagekraft belegt. Er misst allgemeine Voraussetzungen für Lernprozesse, zu denen Konzentration, Aufmerksamkeit, Anstrengung, Beachtung, innere Anspannung, Willensanspannung und Aktivierung gehören. Hierbei liegt der Fokus auf der Messung von Konzentration, also

> eine leistungsbezogene, kontinuierliche und fokussierende Reizselektion, die Fähigkeit eines Individuums, sich bestimmten (aufgaben-)relevanten internen oder externen Reizen selektiv, das heißt, unter Abschirmung gegenüber irrelevanten Stimuli, ununterbrochen zuzuwenden und diese schnell und korrekt zu analysieren. (Brickenkamp/Karl 1986, 195)

Hierfür wird das Zusammenspiel der Instanzen Tempo bzw. Quantität, Qualität und zeitlicher Verlauf der gezeigten Leistung untersucht und ausgewertet.

Die Ergebnisse des d2-Tests haben eine Voraussagekraft in Bezug auf die Aufmerksamkeits- und Konzentrationsleistungen (vgl. Brickenkamp 2002, 8) der Probandinnen und Probanden, zudem können die Ergebnisse im Vergleich zur deutschen Eichstichprobe verortet und somit die Leistungsfähigkeit der getesteten Schülerinnen und Schüler kontextualisiert werden. Für die Beantwortung der gestellten Ausgangsfrage war vor allem der Blick auf die leistungstechnischen Voraussetzungen der Untersuchungsgruppe relevant, die eine weitere Verbindung zu den Ergebnissen der selbstentwickelten Tests zum literarischen Verstehen ermöglichten. Zudem können die Schülerinnen und Schüler einem Raster verortet werden, was in der Phase der Hypothesenbildung eine Explikation bezüglich der Zielgruppe individueller Förderung mit Computerspielen ermöglicht.

Was leistet der *d2-Aufmerksamkeits-Belastungstest*?
Was leistet der *d2-Aufmerksamkeits-Belastungstest* nicht?

Der *d2-Aufmerksamkeits- und Belastungstest* misst die Konzentrationsfähigkeit der Schülerinnen und Schüler, wobei auch das Arbeitstempo und die Sorgfalt berücksichtigt werden. Er hilft, die Schülerleistungen zu kontextualisieren.

Der *d2-Test* misst weder die Intelligenz, noch besitzt er eine valide Voraussagekraft für das Verstehen von Computerspielen bzw. für die Fähigkeit, Computerspiele zu spielen. Auch wird nicht die Aufmerksamkeitsspanne gemessen, sodass hieraus resultierende Leistungsschwankungen nicht beleuchtet werden können.

Die ausgewerteten Daten lassen sich in einer Sorgfalt/Tempo-Matrix darstellen, wobei das Spektrum in seinen Extremen von ‚erheblich konzentrationsgestört' (wenig Sorgfalt, wenig Tempo), pedantisch (große Sorgfalt, wenig Tempo), Ü-Symptom (wenig Sorgfalt, hohes Tempo) und ‚hoch konzentriert' (große Sorgfalt, hohes Tempo) reicht. Der in der Grafik dunkelgrau eingefärbte Bereich stellt den Normbereich dar, innerhalb dessen sich die durchschnittlichen ‚Normalwerte' befinden. Die Schülerinnen und Schüler wurden wie folgt auf der Matrix verortet:

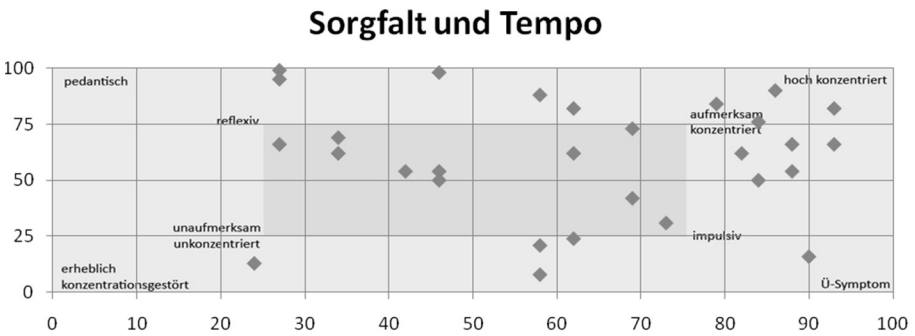

Abb. 3: Darstellung der Sorgfalt und des Bearbeitungstempos der Probanden

Während sich ein Großteil der Gruppe nahe oder oberhalb des Normbereichs ansiedelt, sind für eine intensivere Einzelfallbetrachtung auch hier insbesondere die Extremfälle interessant, da über die qualitative Analyse ihrer Leistungen möglicherweise Hinweise auf individuelle Verstehenswege gewonnen werden können.

2.3.3 Personenbezogene Daten und Fragebogen zur Mediennutzung

Die bei der Erhebung personenbezogener Daten abgefragten Aspekte können als allgemeine Informationen, die unzweifelhaft vorliegen, gekennzeichnet werden, sodass zeitökonomisch eine direkte Abfrage im Fragebogenformat gewählt wurde. Einige der Informationen zu Medienbesitz und Mediennutzung fielen auch unter diese Kategorie, jedoch mussten bei einigen Einschätzungsaspekten – etwa Lese- und Computerspielzeiten – Kontrollfragen, etwa nach konkret gespielten Spielen oder zuletzt gelesenen Romanen, in den Test integriert werden, um Störfaktoren wie die soziale Erwünschtheit zu kontrollieren.

Der selbstentwickelte Mediennutzungsbogen diente primär der Kontextualisierung der Schülerinnen- und Schülerleistungen in den folgenden Tests. Er erfasste die Medienausstattung, das Nutzungsverhalten bezogen auf Literatur und Computerspiele sowie die subjektive Bedeutung einzelner Medien. Hierbei wurden aus zeitökonomischen Gesichtspunkten auf eine ausführliche Erhebung der einzelnen Selbstkonzepte verzichtet und die Aussagen schlaglichtartig aufgenommen.

In einem ersten Teil wurde die Verbreitung spezieller Medienarten erhoben. Durch diese Informationen ließen sich erste Vermutungen über die Vorerfahrung der Probanden äußern. Gleichzeitig konnten die Angaben einzelner Schülerinnen und Schüler im Vergleich zur repräsentativen Mediennutzungsstudie JIM (vgl. JIM 2015) gesehen und somit eine Verortung der Lerngruppe zum Bundesdurchschnitt durchgeführt werden.

Im zweiten Teil des Bogens wurden die Mediennutzungszeiten für Computerspiele und Bücher erhoben, wobei zwischen Zeiten unter der Woche und am Wo-

chenende unterschieden wurde. Zudem waren die Schülerinnen und Schüler aufgefordert, die vier Computerspiele anzugeben, die sie am häufigsten spielen, und die beiden zuletzt gelesenen Bücher namentlich zu nennen.

In einem dritten Block sollte zu sieben Aussagen Stellung genommen werden. Der Fragenkomplex erfasste jeweils zwei Fragen zum Leser- und Computerspielerselbstkonzept, eine Frage zur Leistung im Deutschunterricht und eine Frage zur aktuellen Befindlichkeit.

Aufgrund der Störanfälligkeit von Selbstaussagen wurden die durch den Fragebogen erfassten Informationen nur eingeschränkt in der quantitativen Auswertung genutzt und vorwiegend in die qualitative Analyse miteinbezogen. Dennoch boten sie einen Orientierungsrahmen, da aus der Summe der Antworten ein Lese- und ein Computerspiel-Affinitätsindex errechnet werden konnte, der Klassifizierungen in verschiedene Gruppen und somit eine Binnendifferenzierung zwischen verschiedenen Mediennutzungstypen ermöglichte.

Was leistet der *Mediennutzungsbogen*?
Was leistet der *Mediennutzungsbogen* nicht?

Der *Mediennutzungsbogen* erhebt Daten zur Medienausstattung, den Mediennutzungsgewohnheiten und zu medienbezogenen Selbstkonzepten. Diese Informationen dienen dem Vergleich der Klasse mit dem in der JIM-Studie erhobenen Bundesdurchschnitt und der Verortung der Ergebnisse der Verstehenstests in Bezug auf Medienpräferenzen.

Die Validität der im Mediennutzungsbogen erhobenen Daten ist nicht belegt. Sie unterliegen möglicherweise Störfaktoren und müssen dementsprechend reflektiert in die Auswertung einbezogen werden.

Neben den genannten Aspekten ließen sich die aus den Daten gewonnenen Indizes in Leseaffinitäts- und Computerspielaffinitätstabellen umrechnen, aus denen sich ergibt, dass die Probandinnen und Probanden Literatur stärker zugetan sind als Computerspielen:

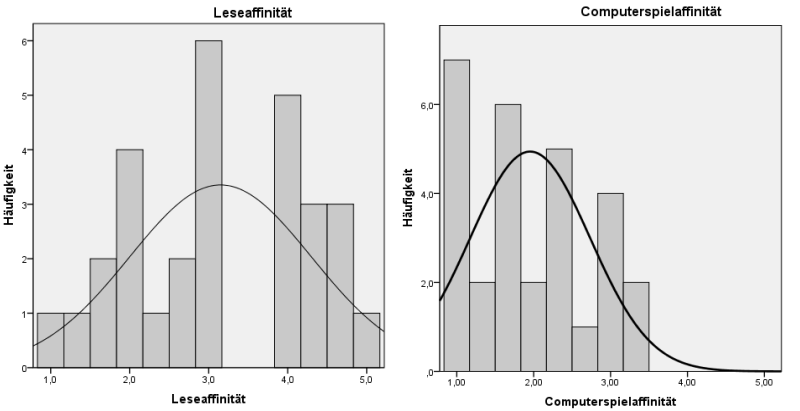

Abb. 4: Darstellung der Lese- und Spielaffinität der Probanden

2.3.4 Selbstentwickelter Test zur Messung literarischer Kompetenz

Die Entwicklung des Tests zur Messung literarischer Kompetenz als Kernelement der Erhebung muss an dieser Stelle ausgespart werden, da dies den Rahmen dieses Beitrags sprengen würde, es sei jedoch bei Interesse auf die Monographie verwiesen. Im Folgenden werden knapp zentrale Gedanken vorgestellt, die die Entwicklung des Tests leiteten.

1. Literarische Kompetenz kann nicht mit standardisierten Tests erhoben werden.

Als das Ziel literarischen Lernens und Verstehens wird im Projekt die subjektive Generierung von Sinn aus einem Text verstanden – diese Annahme entspricht dem Konstrukt *Literarischer Kompetenz* nach dem Bochumer Modell und wurde in der Testkonstruktion umgesetzt. Dieser vom Leser geschaffene Sinn kann aufgrund der *Plausibilität*, die der mentalen Operation zu Grunde liegt, beurteilt werden. Dies gelingt nur, indem nicht mit normativen Kategorien das Verstehen oder Nichtverstehen eines Textes vordefiniert wird. Polyvalenz ist eine maßgebliche Eigenheit von literarischen Texten, die bei der Entwicklung von Testverfahren Beachtung finden muss. Als für die Messung mit empirischen Mitteln problematisch erweist sich somit grundlegend die literarische Form des Gegenstands: Während Sachtexte funktional angelegt sind und sich ihr Verstehen durch den Leser durch das Erreichen einer vorher definierbaren Absicht des Textes abprüfen lässt, bieten literarische Texte dem Leser vielfältige Interpretationsspielräume. Der Text eröffnet seinen Lesern somit zahlreiche Möglichkeiten, ihn zu verstehen und Deutungshorizonte zu eröffnen. Anders als bei der Auseinandersetzung mit Sachtexten können also Aussagen über einen literarischen Text nicht zweifelsfrei in ‚richtig'-/‚falsch'-Kategorien eingeordnet werden. Viel mehr dokumentiert die Konsistenz einer Aussage über den literarischen Text das Verständnis des Textes. Um den Faktor *Plausibilität* zu messen, muss allerdings auf Verfahren der qualitativen Sozialforschung zurückgegriffen werden.

Hierzu wurde ein Test mit offenen Fragen entwickelt, der auf die verschiedenen literarischen Teilkompetenzen des Bochumer Modells abzielte, und anschließend anhand von sogenannten Durchdringungsstufen auf ihre Plausibilität hin untersucht werden konnte.

2. Zu beiden Gegenständen müssen möglichst identische Aufgaben entwickelt werden.

Um die Ergebnisse der Erhebung nicht über unterschiedliche Aufgabenschwierigkeiten zu korrumpieren, wurden die Items in einem ersten Schritt unabhängig von den Gegenständen entwickelt. Hierbei bestand eine Orientierung an den Vorgaben des theoretischen Konstrukts und erst in einem zweiten Schritt wurden die Items an die Gegenstände angepasst. Dies gewährleistet die Vergleichbarkeit der Antworten auf dieser Ebene.

3. Die Gegenstände müssen eine einheitliche Anforderungsstruktur aufweisen.

Die Auswahl der Gegenstände war für die in der Anlage der Erhebung zentrale Vergleichbarkeit der Ergebnisse aus beiden Tests elementar. Daraus folgt, dass der Text und das Computerspiel einen ähnlichen Rezeptions-Schwierigkeitsgrad, inhaltliche Parallelen und zudem strukturelle narrative Ähnlichkeiten vorweisen mussten. Dennoch durften die beiden Gegenstände inhaltlich nicht zu ähnlich sein, da Lerneffekte verhindert werden sollten, die durch die Spiel-/ Textrezeption angestoßen werden und sich auf das Verstehen des anderen Gegenstandes auswirken. Um ebenfalls bereits am Gegenstand vollzogene Lerneffekte auszuschließen, sollten beide Gegenstände den Probandinnen und Probanden unbekannt und angemessen kurz sein, sodass sie im Rahmen der Erhebung rezipiert werden konnten. Die Wahl fiel auf zwei Gegenstände, die diesen Ansprüchen gerecht werden: der Text *Kariuki und der weiße Junge*, ein Textausschnitt aus dem Roman *Kariuki und sein weißer Freund* von Meja Mwangi (2008), und die erste Mission der Menschenkampagne des Spiels *Warcraft III* (Blizzard Entertainment 2002).

4. Neben der Rezeptions- muss auch die Produktionsfähigkeit der Probanden berücksichtigt werden.

Neben Gegenständen und Items zeigten sich auch die Rezeptions- und Produktionsfähigkeiten der Rezipienten als mögliche Störfaktoren. Während die Rezeptionsfähigkeit mittels des *SLS 5-8* (für Lesefähigkeit) und Begleitung des Computerspielprozesses durch Hilfskräfte kontrolliert wurde[7], fiel die Entscheidung bei der Wahl des Produktionsmittels auf mündliche Erhebungen mittels *standardisierter Interviews*, da hiermit die Hoffnung verknüpft wurde, ausführlichere Antworten zu erhalten. Auch Probleme bei der schriftlichen Textproduktion sollten so aufgefangen werden.

2.4 Pilotierung

Vor der eigentlichen Testung wurden die neu entwickelten Tests einer Pilotierung unterzogen, um

> das vorläufige Instrument auf seine Anwendbarkeit, Vollständigkeit, Verstehbarkeit und Qualität (Einhaltung der Gütekriterien), die Erhebungssituation und eventuell die Interviewer zu prüfen. (Raithel 2006, 63).

Hierbei fiel die Wahl auf Probanden mit unterschiedlichen Erfahrungen und verschiedenen Medienpräferenzen, sodass die selbstentwickelten Tests mit drei Schülern der achten Jahrgangsstufe getestet wurden. Zwar ist die Probandengruppe der Pilotierung zu klein, um statistische Analysen vorzunehmen oder valide Vorerwartungen zu den Ergebnissen zu formulieren, dennoch konnten grundlegende Verstehensprobleme im Bereich der Instruktion und Itemformulierung

[7] Die Konzeption der zur Kontrolle von *Spielefähigkeit* und *Nervosität* eingesetzten Interviewer-Fragebögen wird in diesem Beitrag ausgespart.

erkannt und behoben werden, was eine leichte Modifikation der Frageformulierungen und der Instruktionstexte zur Folge hatte.

Die Pilotierung ermöglichte zudem, die Zeiten, die während der Erhebung für die Rezeption des Textes und des Spiels zur Verfügung stehen mussten, so zu bemessen, dass starke und schwache Leserinnen und Leser bzw. erfahrene und unerfahrene Spielerinnen und Spieler genügend Zeit für die Rezeption bekamen.

3. Planung der Erhebung

3.1 Stichprobe planen

Die Wahl der Probandengruppe fiel auf eine achte Gymnasialklasse einer Mittelstadt, die im Grenzgebiet zwischen Sauerland und Ruhrgebiet angesiedelt ist. Mit der Wahl einer Schule in einer Mittelstadt wurde die Hoffnung verbunden, dass die Ergebnisse sowohl für Schulen in Ballungsgebieten wie auch in ländlichen Regionen anschlussfähig sind. Diese Mittelstellung der Schule wurde zudem durch den *Sozialindex* des Schulministeriums NRW, der die soziodemographischen Merkmale Arbeitslosenquote, Sozialhilfequote, Migrantenquote und den Anteil der Wohnungen in Einfamilienhäusern im Schulgebiet umfasst, bestätigt (vgl. Ministerium für Schule und Weiterbildung des Landes NRW o.J.a). Der Sozialindex weist der Region einen Index zwischen 40 und 50 zu. Die Städte des Ruhrgebiets liegen im Schnitt über, die anderen Kreise des ländlichen Sauerlands teils deutlich unter diesem Indexwert. Auch die Einordnung in einen *Standorttyp* durch das Schulministerium belegte dies: Im Rahmen der Lernstanderhebung wurde das Gymnasium dem Standorttyp 3 von 5 zugeordnet, was bedeutet, dass es gemessen an den Sozialfaktoren seiner Schülerinnen und Schüler einen durchschnittlichen Wert aufweist. Zwischen 15 und 25% der Schülerinnen und Schüler haben einen Migrationshintergrund, ebenso beziehen 10 bis 15% der Familien Sozialgeld oder Sozialhilfe. Der Wohnwert der von einem Großteil der Schülerinnen und Schüler des Gymnasiums bewohnten Immobilien ist durchschnittlich (vgl. Ministerium für Schule und Weiterbildung des Landes NRW o.J.b).

Die Wahl der *Altersgruppe* achte Klasse hatte zudem curriculare wie entwicklungspsychologische Gründe: Einerseits findet die Ausbildung der literarischen Kompetenzen nach den *Kernlehrplänen NRW* schwerpunktmäßig zwischen der siebten und neunten Jahrgangsstufe statt, sodass in der achten Jahrgangsstufe ein heterogenes Bild der Kompetenzausprägung erwartet werden kann. Zudem kann durch eine Probandengruppe mit stärkeren als auch schwächeren Schülerinnen und Schüler das Potenzial aufweisen, eine nähere Bestimmung möglicher Zielgruppen für den Einsatz von narrativen Computerspielen zur individuellen Förderung zu leisten. Darüber hinaus ist die Altersgruppe der 13- bis 15-jährigen auch aus *lesesozialisatorischer Sicht* interessant und soll mit der Testung im besonderen Maße fokussiert werden: Mit der einsetzenden Pubertät und den erweiterten medialen Möglichkeiten findet in dieser Altersgruppe eine wichtige Weichenstellung in Bezug auf das spätere Leseverhalten statt.

Die Tragweite der Erhebung sollte die in der Arbeit zuvor geleisteten heuristischen Überlegungen stützen und auf ein intersubjektiv nachvollziehbares und belastbares Fundament stellen. Zugleich wurde keine Repräsentativität der Ergebnisse angestrebt, womit die Wahl auf eine kleine Klumpenstichprobe fiel: An der empirischen Erprobung nahmen 29 Schülerinnen und Schüler der 31 Jugendliche umfassenden Klasse teil, zwei fehlten krankheitsbedingt. Die Probandengruppe bestand aus 17 Mädchen und zwölf Jungen, drei Schülerinnen und Schüler wiesen einen Migrationshintergrund auf, bei einem Schüler wurde ADHS diagnostiziert. Nach Aussage des Klassenlehrers galt die Klasse als disziplinarisch unauffällig und stand leistungsbedingt prototypisch für die gesamte Jahrgangsstufe. Im Bereich der Rezeptionsfähigkeit waren zudem keine Probleme zu erwarten, was die Kontrolle dieses Faktors in der Auswertung bestätigte.

3.2 Zeitliche Organisation

Bereits früh im Prozess der Designentwicklung wurde klar, dass die Erhebung von 62 standardisierten Interviews nicht von einer einzelnen Person in einem überschaubaren Rahmen geleistet werden kann. Anstelle einer zerdehnten Erhebung über mehrere Wochen hinweg, fiel – mit dem Ziel die Umgebungsfaktoren ähnlich zu gestalten und die Kommunikation der Probandinnen und Probanden über die Gegenstände und hiermit einhergehende Lerneffekte zu verhindern – die Entscheidung, die Interviews in einem Durchlauf zu erheben. Hierzu sollten zwei große Befragungen nacheinander durchgeführt werden, in denen die eine Hälfte der Klasse erst die Begleittests und anschließend die Verstehenstests durchläuft, während die andere Gruppe erst die Verstehenstests absolviert und anschließend die Begleittests durchführt. Aus diesem Grund wurde eine vergleichsweise große Gruppe von 20 Interviewern für die Erhebung ausgebildet.

Die Erhebung wurde an einem Vormittag Mitte Januar 2011, wenige Tage vor der Vergabe der Halbjahreszeugnisse durchgeführt, was verschiedene positive Effekte aufwies: Aus wissenschaftlicher Sicht wurde hiermit die zeitliche Mitte der drei Referenzschuljahre, in denen literarische Verstehensprozesse elaboriert werden, abgepasst, was der Heterogenität der Gruppe vermeintlich zu Gute kommt. Die Klasse hatte vor der Zeugnisvergabe keine Klassenarbeiten mehr zu schreiben, sodass die Schulleitung dem Termin ohne Vorbehalte zustimmte. Einen weiteren Türöffner stellte die Weitergabe der Ergebnisse der standardisierten Tests an den Klassenlehrer dar, da dieser die Daten für die individuelle Förderung im zweiten Halbjahr nutzen wollte.

Diese Informationen wurden im Rahmen der datenschutzrechtlich notwendigen Mitteilung frühzeitig an die Erziehungsberechtigten weitergeleitet und im Rahmen eines Elternabends besprochen. Auf dieser Grundlage konnten die Eltern der Probandinnen und Probanden eine informierte Entscheidung treffen, wobei alle Eltern der Teilnahme ihrer Kinder einwilligten und sich in der Folge auch alle Schülerinnen und Schüler dazu entschieden, an der Untersuchung mitzuwirken.

3.3 Schulung für Testung und Intervention

Anders als ein Fragebogen oder eine schriftliche Arbeit mit offenen Antworten birgt das Erhebungsinstrument *Interview* die Gefahr, durch unterschiedliches Verhalten der Interviewenden eine mögliche Quelle für Störfaktoren zu werden (vgl. Sedlmeier/Renkewitz 2008, 102). Zum Ziel der Vergleichbarkeit der Interviews sollten verschiedene Störfaktoren durch eine *Standardisierung des Vorgehens* und durch die intensive Schulung der Interviewer bestmöglich ausgeschlossen werden. Während der Interviewerschulung wurden die 20 Interviewer nicht nur in der Art der Interviewführung ausgebildet, sondern zudem an den technischen Gerätschaften (Computer für die Spielerprobung und Aufnahmegeräten) unterwiesen. Diese Aspekte flossen in einen den Interviewenden ausgehändigten Interviewer-Leitfaden ein.

Thematisch umfasste die Schulung alle Aspekte der Interviewsituation, wobei von der Begrüßung über Antworten auf zu erwartende Fragen, das Rückmeldeverhalten in der Interviewsituation und die Form der Verabschiedung alle Aspekte möglichst genau vorgegeben wurden, um auch die Rahmensituation weitestgehend zu standardisieren. Neben der Standardisierung der Verhaltensweisen sollten ebenfalls die Fragesituation (Vorlesen von Instruktionstext und Items) durch eine ähnliche Betonung aller Interviewenden kontrolliert werden.

Im Anschluss an die theoretische Unterweisung erprobten die Interviewenden in Zweiergruppen das Interview, wobei die Rolle Interviewer(in)/Proband(in) nach einem Testdurchlauf wechselte. So erfuhren die Interviewenden die spezifischen Probleme der Probandensituation und erprobten ihre Rolle bereits vor der Durchführung. Auftretende Probleme („Wie gehe ich mit einer spezifischen Reaktion um?“) konnten während der Interviewer-Schulung für alle Teilnehmer verbindlich behandelt und gelöst werden.[8]

4. Durchführung der Erhebungen

Für die Durchführung standen vier Räume zur Verfügung: der Klassenraum, in dem nach der gemeinsamen Einführung die Begleittests durchgeführt wurden, sowie drei große Seminarräume, in denen die Interviews stattfanden. Auch wenn nicht für jedes Interview wie gewünscht ein eigener Raum zur Verfügung stand, wurde sichergestellt, dass die großen Seminarräume, die im Schulalltag für Abiturklausuren oder Theateraufführungen genutzt werden, eine Weitläufigkeit aufwiesen, die es den Interviewenden ermöglichte, die Probandinnen und Probanden zu interviewen, ohne dass sie von Nebengeräuschen der anderen Interviews gestört, abgelenkt oder beeinflusst wurden. Alle für die Befragung notwendigen

[8] Nach der Schulung wurden zwei Interviewerinnen von der Durchführung ausgeschlossen, da sie im größeren Maße lautliche Rückmeldungen („Ja.“, „Gut.“) gaben und es ihnen zudem nicht gelang, auf gestaltende Nachfragen, die über den standardisierten Wortlaut hinausgingen, zu verzichten.

technischen Gerätschaften wurden von der Forschergruppe mit in die Schule gebracht, zu technischen Schwierigkeiten während der Durchführung kam es nicht.

Nach einem gemeinsamen Einstieg im Klassenraum mit Vorstellung der Erhebenden, des Ablaufs und der Ziele der Erhebung, erhielten die Probandinnen und Probanden mittels eines hierfür angefertigten Videos eine kurze Einführung in das Gameplay des Strategiespiels. Auch wurden die Schülerinnen und Schüler darauf hingewiesen, dass die Verstehenstest-Interviews standardisiert ablaufen werden, was bedeutet, dass die Interviewenden nicht auf Fragen bzw. Bitten um Hilfe oder sonstige Interaktionen eingehen dürfen. Ebenso wurde auf den straffen Zeitplan und das Verhalten während der Raumwechsel hingewiesen. Dies verhinderte leistungsbeeinflussende Irritationen in späteren Phasen, die als Störfaktoren somit ausgeschlossen werden konnten. Die Atmosphäre während der Einführung war wie im gesamten Durchführungszeitraum kooperativ und die Schülerinnen und Schüler stellten Verständnisfragen, sodass in der Folge keine weiteren Fragen aufkamen, die den Ablauf verzögert oder behindert hätten.

Um die Umgebungsfaktoren möglichst vergleichbar zu halten, sollten die Interviews zeitnah zueinander stattfinden. Es war mit dem Klassenlehrer im Vorfeld abgesprochen worden, dass die Gruppe zweigeteilt wird. Während die erste Gruppe mit den standardisierten Tests begann, verließen die Probanden der Gruppe 2 den Klassenraum und teilten sich auf die drei Interview-Räume auf, in denen die Interviewer bereits auf sie warteten. Anschließend wurden die Räume getauscht und die zweite Gruppe absolvierte die standardisierten Tests, während Gruppe 1 interviewt wurde.

Der Zeitplan des Tages, der im Rahmen der Durchführung auch eingehalten werden konnte, gliedert sich wie folgt:

Zeitplanung Schüler

Beginn	Ende				
9:45 Uhr	10:00 Uhr	*Gemeinsamer Start im Klassenraum (ca. 15 Min.)*			
		Gruppe 1		**Gruppe 2**	
10:00 Uhr	10:10 Uhr	Mediennutzungsbogen	10 Min.	Computerspielen	30 Min.
10:10 Uhr	10:20 Uhr	D2-Test	10 Min.		
10:20 Uhr	10:30 Uhr	Lesescreening	10 Min.		
		ggf. Pause (Aufsicht: SG)			
10:30 Uhr	10:45 Uhr	Leseverstehenstext lesen	15 Min.	Interview	15 Min.
10:45 Uhr	10:50 Uhr	*Raumwechsel*			
10:50 Uhr	11:00 Uhr	Interview	15 Min.	Mediennutzungsbogen	10 Min.
11:00 Uhr	11:10 Uhr	Computerspielen	30 Min.	D2-Test	10 Min.
11:10 Uhr	11:20 Uhr			Lesescreening	10 Min.
11:20 Uhr	11:35 Uhr			15 Min. Pause	15 Min.
11:35 Uhr	11:50 Uhr	Interview	15 Min.	Leseverstehenstext lesen	15 Min.
11:50 Uhr	11:55 Uhr	*Raumwechsel*			
11:55 Uhr	12:10 Uhr	15 Min. Pause	15 Min.	Interview	15 Min.
12:10 Uhr	12:15 Uhr	*gemeinsamer Schlusspunkt im Klassenraum (ca. 5 Min.)*			

Abb. 5: Zeitplan des Erhebungstages

Bei beiden Testdurchläufen wurde ebenfalls auf eine größtmögliche Vergleichbarkeit Wert gelegt. So wurde neben den bereits beschriebenen Maßnahmen zur Standardisierung der Interviews auch der Ablauf der Testdurchläufe der Begleittests identisch gestaltet. Alle Instruktionstexte, wie auch Moderationen wurden im Vorfeld niedergeschrieben und entsprechend vorgetragen.

Vor den Raumwechseln wurden die Probanden darauf hingewiesen, dass sie nicht miteinander oder mit Probanden der anderen Gruppe kommunizieren dürften, was vom Projektleiter und dem Klassenlehrer überwacht und von den Schülerinnen und Schülern eingehalten wurde. Während der Durchführung kam es zu keinen besonderen Ereignissen, sodass der Zeitplan eingehalten wurde und alle Interviews in den Zeitvorgaben stattfanden. Auch funktionierten die technischen Gerätschaften fehlerfrei. Aufgrund dessen konnten alle Probandinnen und Probanden ohne Komplikationen die Passage spielen und alle Interviews aufgezeichnet werden.

5. Datenmanagement

5.1 Vorbereitende Maßnahmen und Planung

Im Vorfeld wurde mit den Eltern und der Schulleitung abgesprochen, dass der Klassenlehrer Einblick in die Ergebnisse seiner Klasse erhalten soll, während die Forschenden nur anonymisiert mit den erhobenen Daten arbeiteten. Aus diesem Grund war es notwendig, den Schülerinnen und Schülern einen eindeutigen Code zuzuordnen, mit dem ihre Dokumente versehen wurden, der es aber dem Klassenlehrer ermöglichte, diese in der Folge einzelnen Schülerinnen und Schülern zuzuordnen.

Den Schülerinnen und Schülern wurden deshalb zu Beginn der Stunde sogenannte Code-Karten ausgeteilt, die die Felder „Nr." mit einer eingetragenen Zahl zwischen 01 und 31 und „Name" enthielten, in das die Probanden ihren Namen eintrugen. Sie beschrifteten ihre Dokumente mit ihrer Nummer und der Klassenlehrer sammelte nach Abschluss der Erhebung die Code-Karten ein, sodass er die Ergebnisse zuordnen konnte. Gleichzeitig konnte mit Hilfe der Code-Nummern die Schülergruppe zweigeteilt werden: Die Schülerinnen und Schüler mit einer geraden Zahl wurden der Gruppe 1 zugeordnet, Probandinnen und Probanden mit einer ungeraden Zahl gehörten der Gruppe 2 an. Diese Zufallsverteilung ergab ein geschlechtsspezifisch ausgewogenes Verhältnis von sechs Jungen und neun Mädchen in Gruppe 1 sowie sechs Jungen und acht Mädchen in Gruppe 2. Im Verlauf der Erhebung gaben die Probandinnen und Probanden bei jedem Test ihren Code an, die Code-Karten verblieben beim Klassenlehrer.

5.2 Dateneingabe, Speicherung und Archivierung

Nach der Erhebung wurden die in Papierform vorliegenden Testbögen eingesammelt und die Audiodaten direkt vor Ort auf dem Projektleitungslaptop kopiert, jede Datei wurde auf ihre Intaktheit geprüft und die Dateien auf einem zweiten

Datenträger gesichert. Damit auf die Daten nur von autorisierten Personen zugegriffen werden konnte – insbesondere im Falle eines Verlusts oder Diebstahls der Datenträger –, wurden die Daten nicht frei zugänglich, sondern mit Veracrypt verschlüsselt abgelegt. Die Daten wurden hiermit vor unberechtigtem Zugriff geschützt. Die analogen Daten wurden anschließend digitalisiert und ebenfalls in die verschlüsselten Container geladen. Einerseits wurde so einem potenziellen Datenverlusten Vorschub geleistet, andererseits der Zugriff auf die Forschungsdaten beschränkt. Erst anschließend wurde mit der Datenauswertung begonnen.

6. Datenauswertung

Da die Angaben der Probanden in den Erhebungsbögen ohne eine Veränderung in die Analysesoftware SPSS übertragen wurden (siehe hierzu Kapitel 6 des Beitrags von Pissarek und Wild in diesem Band), handelt es sich strenggenommen nicht um eine Datenaufbereitung, allerdings wurden die angegebenen Werte in Variablen übertragen (siehe den Beitrag von Schmitz in diesem Band), sodass sie anschließend mittels statistischer Methoden bearbeitet werden konnten.

6.1 Daten aufbereiten

Die standardisierten Tests mussten nach Vorgabe ihrer Auswertungsmanuals ausgewertet und die Daten zur Weiterverarbeitung in SPSS eingespeist werden. Dazu wurden die Ergebnisse – wie in Kapitel 2 bereits vorgestellt – einzeln zusammengestellt. Das aufwändigere Verfahren stellte die Auswertung der beiden Tests zur literarischen Kompetenz dar, da hier durch eine *qualitative Inhaltsanalyse* Plausibilitätskategorien bestimmt werden mussten. Hierzu wurde mit der Qualitativen Inhaltsanalyse ein quantifizierendes Verfahren genutzt, das die Äußerungen der Schülerinnen und Schüler in einen Durchdringungswert zwischen 1 (niedrigster Wert) und 6 (höchster Wert) überführt. Hierzu wurden im Anschluss an die Datenerhebung die Tonspuren der Interviews nummeriert und einem inhaltlichen Integritätstest unterzogen[9].

Alle Interviews, die den formalen Anforderungen entsprachen, wurden drei Auswertern nach einer spezifischen Schulung zur Verfügung gestellt[10]. Das Auswertungsverfahren nutzte für jedes Item eine sechsstufige inhaltlich ausdifferenzierte Skala, auf der die Leistung verortet werden musste. Die Auswerter beurteilten die einzelnen Items und Interviews unabhängig voneinander und übergaben die Ergebnisse ohne Kenntnis der anderen Auswertungsergebnisse dem Projektleiter, der die Daten zusammenführte. Die *Interrater-Korrelation* (nach Krippendorffs-Alpha) des literarischen Text-Verstehenstest beträgt im Durchschnitt 0,933; im

[9] Ein Interview konnte nicht ausgewertet werden, da der Interviewer während des Interviews Rückfragen stellte und Anmerkungen zu den Aussagen der Schülerin machte. Das Interview erfüllt somit nicht den Anspruch der Standardisierung und hat eine Leitfaden-Interview-Charakteristik, die nicht mit den anderen Leistungen vergleichbar ist.

[10] Eine alternative Herangehensweise stellt die Arbeit an Transkriptionen der Interviewdaten dar, nach einem Abwägeprozess wurde aber beschlossen, dass die Auswerter mit den Originaldaten arbeiten sollten.

Computerspiel-Verstehenstest liegt sie bei 0,973 noch höher. Dies bedeutet, dass die Leistungseinschätzung der drei Auswerter in nur sehr wenigen Fällen differierte und in einem Großteil der Einschätzungen identisch war. Im Fall einer Anderseinschätzung lag die Differenz bei maximal einem Punkt auf der Skala. Methodenkritisch lässt sich anmerken, dass das gewählte Verfahren sehr aufwendig und mit Blick auf größere Stichproben noch optimierbar erscheint, dass aber die erhobenen Ergebnisse in ihrer Aussagekraft dennoch vor allem in der Phase der Theoriebildung denen von Multiple-Choice-Tests vorzuziehen sind, da sie sowohl qualitativ wie auch quantitativ ausgewertet werden können und dem polyvalenten Gegenstand gerechter werden.

Für die anschließende Auswertung wurden die Daten ebenfalls in die Statistiksoftware SPSS eingespeist, sodass die Auswertung der Studie mit dem vollständigen Datensatz von 115 Variablen[11] begann.

6.2 Auswertung nach Forschungsfragen und Interpretation

Die Auswertung der Daten erfolgte in mehreren aufeinander aufbauenden Schritten, wobei zuerst die einzelnen Tests ausgewertet und anschließend Zusammenhänge herausgearbeitet wurden. So konnten die beiden Tests zur literarischen Kompetenz nachweisen, dass sich der in der theoretischen Modellierung des Bochumer Modells angelegte hierarchische Aufbau der einzelnen Teilkompetenzen auch in den Antworten der Probandinnen und Probanden zeigte: Fragen, die auf die Handlungsebene des Modells abzielten, konkret zu Handlungslogik und Figurenverstehen, wurden besser beantwortet als solche, die die Meta-Ebene fokussierten und nach symbolischen und metaphorischen Ausdrucksweisen sowie zur (bild-)sprachlichen Gestaltung fragten. Nach dieser grundsätzlichen Validierung des Konstrukts wurden dann die Einzeltests miteinander in Beziehung gesetzt: Die Auswertung der printbasierten und spielbasierten literarischen Verstehenstests zeigte, dass es den Schülerinnen und Schülern gelang, ihre literarische Kompetenz nicht nur an einem Text, sondern auch an einem narrativen Computerspiel anzuwenden. Eine Gegenüberstellung der Mittelwerte beider Tests verdeutlicht, dass die Ergebnisse beider Tests große Ähnlichkeiten zueinander aufweisen.

Mittelwerte der Items organisiert nach Medienorientierung[12]

	Item 3	Item 2	Item 5	Item 6	Item 7	Item 8	Item 4
printbasiert	4,580	4,069	3,560	3,357	3,160	2,645	2,520
spielbasiert	4,951	4,250	3,512	3,274	3,148	2,859	2,654

Abb. 6: Darstellung der Itemmittelwerte nach Medienorientierung

[11] Hierbei sind die Ergebnisse der drei Auswerter bereits in gemeinsame Werte zusammengeführt worden.

[12] Item 3 und 2 fokussieren die Handlungslogik, Item 5, 6 und 7 das Figurenverstehen, Item 8 (bild-)sprachliche Mittel und Item 4 das Symbol- und Metaphernverstehen. Ein hoher Wert weist eine hohe Verstehensleistung aus, ein niedriger Wert eine niedrige.

Der Vergleich der Durchschnittswerte nach Items zeigt, dass die Leistungen der Schülerinnen und Schüler zum literarischen Verstehen in den beiden Medien sehr nah beieinander liegen. Obwohl sich die Klasse im Schnitt als deutlich literaturaffiner erwies, lassen sich keine signifikanten Leistungsunterschiede beim literarischen Verstehen zwischen Computerspiel und Text erkennen. Diese Aufstellung weist zudem darauf hin, dass der Schwierigkeitsgrad der einzelnen Anforderungen literarischer Kompetenz medienunabhängig ansteigt. Diese erste Auswertung mit der vergleichsweise niedrigschwelligen Mittelwertbestimmung macht deutlich, dass die grundsätzliche Vergleichbarkeit zwischen beiden Medienformen gegeben ist und keine grundlegenden Unterschiede zwischen dem Kompetenzeinsatz bei verschiedenen medialen Basierungen besteht. Zugleich ergibt sich hieraus noch kein Zusammenhang. Dieser wurde erst in der Folge über eine Korrelationsanalyse hergestellt. Hierbei wird deutlich, dass die vermuteten Effekte auch auf der Ebene der Korrelationen der verschiedenen Items nach Medienorientierung zueinander nachweisbar sind und Aufschluss über die Stärke der Verbindung von literarischem Verstehen über Mediengrenzen hinweg geben.

Korrelation von spielbasiertem und textbasiertem literarischem Verstehen nach Items

	Item 3	Item 2	Item 5	Item 6	Item 7	Item 8	Item 4
Korrelation	.656**	.699**	.506*	.696**	.341*	.419*	.482*

Abb. 7: Darstellung der Korrelation von spielbasierten und textbasierten literarischen Verstehen

Diese Ergebnisse zeigen, dass auf allen Teilkompetenzen literarischer Kompetenz eine signifikante Verbindung zwischen printbasiertem und computerspielbasiertem literarischen Verstehen besteht. Die durch die Mittelwerte aufgezeigte Verbindung ist somit nicht zufällig, sondern auch durch den korrelativen Vergleich der Einzelleistungen nachweisbar. Dies belegt die These, dass Schülerinnen und Schüler ihre literarische Kompetenz an verschiedenen Medien vergleichbar gut einsetzen können, wenngleich mit Blick auf die kleine Probandengruppe nicht von einer Beweisführung im engeren Sinne gesprochen werden kann, sondern nur stichhaltige Hinweise vorliegen.

Darüber hinaus ergaben sich – im Folgenden nur schlaglichtartig betrachtet – auch im Wechselspiel der anderen Tests mit dem selbstentwickelten Test zum literarischen Verstehen interessante Befunde:

- In beiden Tests zum literarischen Verstehen erzielten die *weiblichen Teilnehmenden* im Schnitt bessere Ergebnisse als die männlichen – eine geschlechtsspezifisch-medienbasierte Ausprägung lässt sich nicht nachweisen.
- Die Höhe der *basalen Lesekompetenz* hatte bei der Probandengruppe keine Auswirkung auf die Leistungen im printbasierten literarischen Verstehenstest. Dies kann mit der für die Rezeption ausreichenden Ausprägung der basalen Lesekompetenz begründet werden.

- Die Höhe der *Leseaffinität* spiegelte sich in den Items zur Handlungslogik wider, während sie in den analytisch-geprägten Items zu Figuren, sprachlichen Bildern und sprachlicher Gestaltung nicht zum Tragen kommt.

Abschließend wurden die Ergebnisse ausgewählter Probandinnen und Probanden noch vertiefenden qualitativen Untersuchungen unterzogen, wodurch sich weitere Vermutungen zu Zielgruppen, Rezeptionsweisen und Verstehensprozessen entwickeln ließen, die in folgenden Forschungsprojekten untersucht werden müssen.

7. Fazit

Grundsätzlich lassen sich durch *Forschungsdesigns mit einem Erhebungszeitpunkt* Überzeugungen, Wissen, Können und Praktiken sowie ihre Ausprägungen erheben und über die zielgerichtete Kombination verschiedener *Erhebungs- und Auswertungsverfahren* Zusammenhänge zwischen ihnen herausarbeiten. Im vorliegenden Beispiel konnte gezeigt werden, dass insbesondere der Testentwicklung auf Grundlage eines umfassenden Quellenstudiums und eines klar umrissenen Konstrukts hierbei grundlegende Relevanz zukommt, aber auch die Auswahl der bedingenden Faktoren wie die fokussierte Bestimmung der Probandengruppe erheblichen Einfluss auf die Tragweite der Erkenntnisse nimmt.

Literatur

Primärliteratur und -spiele:

Blizzard Entertainment (2002): Warcraft II.

Mwangi, Meja (2008): Kariuki und sein weißer Freund: Eine Erzählung aus Kenia. Göttingen: Lamuv Verlag.

Sekundärliteratur:

Auer, Michaela/Gruber, Gabriele/Mayringer, Heinz/Wimmer, Heinz (2005): Salzburger Lesescreening 5-8 – Handbuch. Göttingen: Hogrefe.

Baumert, Jürgen/Klieme, Eckhard/Neubrand, Michael/Prenzel, Manfred/Schiefele, Ulrich/Schneider, Wolfgang/Stanat, Petra/Tilman, Klaus-Jürgen/Weiß, Manfred (2001) (Hrsg.): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Budrich.

Boelmann, Jan M. (2015): Literarisches Verstehen mit narrativen Computerspielen. Eine empirische Studie zu den Potenzialen der Vermittlung von literarischer Bildung und literarischer Kompetenz mit einem schüleraffinen Medium. München: kopaed.

Brickenkamp, Rolf (2002): Test d2-Aufmerksamkeits-Belastungs-Test – Manual. Göttingen: Hogrefe.

Brickenkamp, Rolf/Karl, Gerhard (1986): Geräte zur Messung von Aufmerksamkeit, Konzentration und Vigilanz. In: Brickenkamp, Rolf (Hrsg.): Handbuch apparativer Verfahren in der Psychologie. Göttingen: Hogrefe, 195-211.

Christmann, Ursula (2004): Verstehens- und Verständlichkeitsmessung. Methodische Ansätze in der Anwendungsforschung. In: Lerch, Kent D. (Hrsg.): Recht verstehen. Verständlichkeit, Missverständlichkeit und Unverständlichkeit von Recht. Berlin/New York: de Gruyter, 33-62.

Ehmig, Simone C./Reuter, Timo (2013): Vorlesen im Kinderalltag. Bedeutung des Vorlesens für die Entwicklung von Kindern und Jugendlichen und Vorlesepraxis in den Familien. Zusammenfassung und Einordnung zentraler Befunde der Vorlesestudien von Stiftung Lesen, DIE ZEIT und Deutsche Bahn 2007-2012. Mainz: Stiftung Lesen. Online unter: https://www.stiftunglesen.de/download.php?type=document-pdf&id=951 (letzter Zugriff: 01.10.2018).

Funke, Reinold (2005): Sprachliches im Blickfeld des Wissens. Grammatische Kenntnisse von Schülerinnen und Schülern. Tübingen: Niemeyer.

Heins, Jochen (2017): Lenkungsgrade im Literaturunterricht. Zum Einfluss stark und gering lenkender Aufgabensets auf das Textverstehen. Wiesbaden: VS Verlag für Sozialwissenschaften.

Hurrelmann, Bettina/Groeben, Norbert (2006): Lesekompetenz: Bedingungen, Dimensionen, Funktionen. Weinheim: Juventa.

Maus, Eva (2014): Wer (ver)führt zum Lesen? Der Einfluss von Geschlechtermustern auf die Lesemotivation von Jungen und Mädchen. Baltmannsweiler: Schneider Hohengehren.

Medienpädagogischer Forschungsverband Südwest (Hrsg.) (2015): JIM-Studie 2015. Jugend, Information, (Multi-) Media. https://www.mpfs.de/fileadmin/files/Studien/JIM/2015/JIM_Studie_2015.pdf (letzter Zugriff: 01.10.2018).

Medienpädagogischer Forschungsverband Südwest (Hrsg.) (2016): KIM-Studie 2016. Kindheit, Internet, Medien. https://www.mpfs.de/fileadmin/files/Studien/KIM/2016/KIM_2016_Web-PDF.pdf (letzter Zugriff: 01.10.2018).

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (Hrsg.) (o.J.a): Schulstatistik. https://www.schulministerium.nrw.de/docs/bp/Ministerium/Service/Schulstatistik/index.html (letzter Zugriff: 01.10.2018).

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (Hrsg.) (o.J.b): Beschreibung der Standorttypen. https://www.schulentwicklung.nrw.de/e/upload/lernstand8/download/mat_2017/2017-02-08_Beschreibung_Standorttypen__weiterfhrende_Schulen_NEU_RUB_ang.pdf (letzter Zugriff: 01.10.2018).

Payrhuber, Franz-Josef (1991): Das Drama im Unterricht. Aspekte einer Didaktik des Dramas. Analysen und empirische Befunde – Begründungen – Unterrichtsmodelle. Rheinbreitbach: Dürr und Kessler.

Raithel, Jürgen (2006): Quantitative Forschung: Ein Praxiskurs. Wiesbaden: VS Verlag für Sozialwissenschaften.

Rosebrock, Cornelia/Nix, Daniel (2017): Grundlagen der Lesedidaktik: und der systematischen schulischen Leseförderung. 8. korr. Aufl. Baltmannsweiler: Schneider Hohengehren.

Scherf, Daniel (2013): Leseförderung aus Lehrersicht. Wiesbaden: VS Verlag für Sozialwissenschaften.

Schnell, Rainer/Hill, Paul B./Esser, Elke (2011): Methoden der empirischen Sozialforschung. München: Oldenbourg Wissenschaftsverlag.

Sendlmeier, Peter/Renkewitz, Peter (2008): Forschungsmethoden und Statistik in der Psychologie. München: Pearson Studium.

Stiftung Lesen/DIE ZEIT/Deutsche Bahn Stiftung (Hrsg.) (o.J.): Vorlesestudie im Rahmen des bundesweiten Vorlesetages. https://www.stiftunglesen.de/forschung/forschungsprojekte/vorlesestudie (letzter Zugriff: 01.10.2018).

MARKUS PISSAREK & JOHANNES WILD

Prä-/Post-/Follow-Up-Kontrollgruppendesign

Zur Überprüfung der Wirksamkeit von Interventionen[1]

1. Zum Rahmen: Warum Interventionsstudien?

Wenn man wissen will, ob ein Rechtschreib-, Wortschatz-, Grammatik- oder Lesetraining wirksam ist, d.h. die Kompetenzen der Schülerinnen und Schüler dadurch steigen, so liegt es zwar nahe, sie zu beobachten oder sie zu befragen, jedoch ist dies nicht zielführend. Doch woher weiß man, ob der Kompetenzzuwachs genau auf dieses Training zurückzuführen ist und nicht z.B. auf eine alterstypische Entwicklung? Es könnte ja auch sein, dass die Schülerinnen und Schüler mit einer anderen Methode oder im regulären Unterricht noch weit größere Fortschritte erzielt hätten. Die Auseinandersetzung mit der Qualität schulischer Lehr-Lernprozesse und der Lernwirksamkeit des Lehrerhandelns hat in Deutschland lange Tradition (vgl. Helmke 2009, 47-57; Klieme 2006, 765). Während zunächst aber vor allem pädagogische und methodische Debatten den Diskurs prägten, wird aktuell die empirische Überprüfung des Lernerfolgs von Schülerinnen und Schülern, d.h. der Wirksamkeit und Nachhaltigkeit von Unterricht bzw. Unterrichtskonzepten, gefordert (vgl. Leutner et al. 2017, 1; Senn/Krelle 2016, 7; Weinert 2001, 30). Erstmals ist „also statt kontrollierter Laborsettings das natürliche Umfeld von Kindern, Jugendlichen und Erwachsenen zum dominanten Erhebungskontext geworden“ (Reinders et al. 2015, 12). Will man überprüfen, wie *gut* eine Maßnahme (*Intervention*/*Treatment*) funktioniert hat, so ist das sogenannte *Prä-/Post-/Follow-Up-Kontrollgruppendesign* ein Standardverfahren, bei dem die Schülerleistungen mehrfach gemessen sowie eine Trainingsgruppe (oder mehrere) und eine Kontrollgruppe verglichen werden. Im Folgenden wird das Verfahren anhand der Ergebnisse aus einer Lesestudie illustriert.

---

[1] Dieser Beitrag stellt exemplarisch den gesamten Prozess einer empirischen Erhebung vor und konkretisiert somit Aspekte, die in den weiteren Beiträgen dieses Bandes vorgestellt werden. Insbesondere seien hier die Beiträge von Boelmann zu Fragestellung und Forschungsdesign, Pissarek zu quantitativer Forschung, König zu Planung und Vorbereitung empirischer Erhebungen und Schmitz zu statistischen Grundkenntnissen genannt, auf die im Folgenden nicht erneut explizit verwiesen wird.

1.1 Design einer Wirksamkeitsstudie

Die zugrunde liegende Annahme des Prä-/Post-/Follow-Up-Designs ist, dass Kompetenzzuwächse durch ein aktives Nutzen geeigneter Lerngelegenheiten entstehen (vgl. Zeitler/Köller/Tesch 2010, 27), was sich nach Gerrig/Zimbardo (2008, 192) in einer „relativ konsistenten Änderung des Verhaltens oder des Verhaltenspotenzials" niederschlägt. Dieser Leistungszuwachs wird meistens in einer spezifischen Domäne ermittelt (vgl. Helmke 2014, 807). Will man also die Effektivität von Unterrichtsmaßnahmen überprüfen, muss das Niveau der Fähigkeiten und Fertigkeiten vor und nach einer Intervention durch entsprechende Testinstrumente längsschnittlich erfasst werden (vgl. Klieme 2006, 768; Kormann 1984, 117; vgl. *Abbildung 3*). Es handelt sich in der Regel dabei um eine *Evaluation mit drei Messzeitpunkten* (MZP 1-3), die summativen oder formativen Charakter annehmen kann (vgl. Maier 2014; Smit 2009; Roos 2001). Dazu muss man sich zunächst im Klaren sein, welche konkreten Fähigkeiten und Fertigkeiten auf welcher Abstraktionsebene untersucht werden sollen. Erst im Anschluss daran können die konkrete *Fragestellung der empirischen Erhebung*, das *Forschungsdesign* und *quantitative bzw. qualitative Testinstrumente* festgelegt werden.

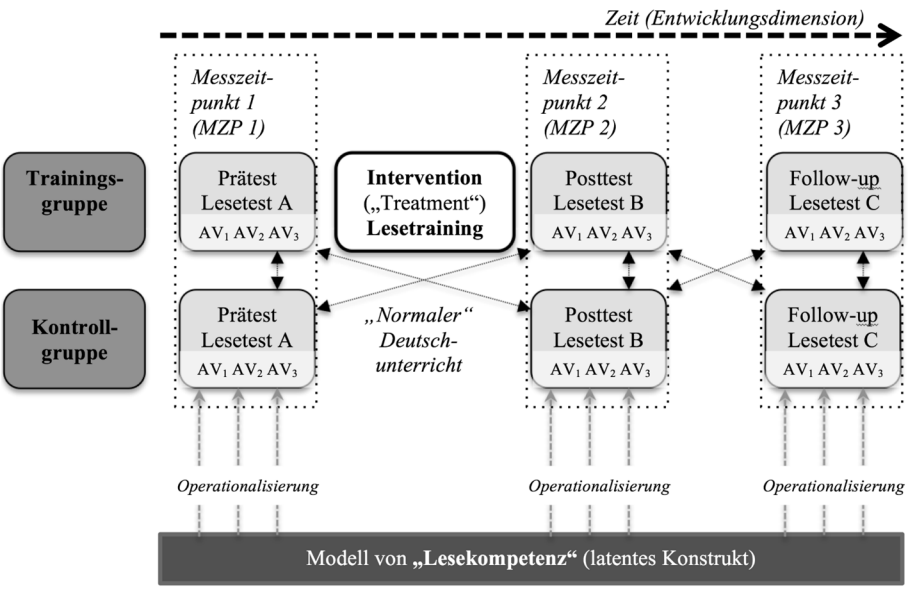

Abb. 1: Schematische Darstellung des Vorgehens bei einer Interventionsstudie

In Abbildung 1 sieht man die zeitliche Situierung der Messzeitpunkte illustriert: Messzeitpunkt 1 stellt das Eingangsniveau der Trainings- und Kontrollgruppe fest, während der Posttest (MZP 2) feststellt, ob sich die Leistungsindikatoren nach der Interventionsdauer verändert haben. Die Trainingsgruppe durchläuft das Treatment, das zugleich die UV (unabhängige Variable) bildet. Eine Kontrollgruppe nimmt nicht an der Intervention (z.B. einem Lesetraining) teil, sondern durchläuft in der Regel regulären schulischen Unterricht. Sie wird benötigt, um

den Erfolg der Intervention von einer normalen Entwicklung zu unterscheiden.[2] Wenn ein Lesetraining sieben Wochen dauert, so ist erwartbar, dass sich auch die Kontrollgruppe in der Zeit altersgemäß typisch entwickelt (vgl. dazu weiter unten das ausführliche Beispiel). Abbildung 1 zeigt außerdem eine Unterscheidung in eine *manifeste* Ebene, auf der gemessen wird (z.B. mit Hilfe eines Leseverständnistestes) und eine *theoretische* Ebene, auf der sich das latente Konstrukt befindet, das man annimmt (hier: Lesekompetenz). Um die Entwicklung sichtbar machen zu können, muss das theoretische Modell *operationalisiert* werden, d.h. in Form von Items abgebildet und überprüfbar gemacht werden, die z.B. das *Leseverständnis*, die *Leseflüssigkeit*, die *Interpretationsfähigkeit* etc. messen sollen. Sie bilden als abhängige Variablen (AV) die Indikatoren für die latenten Konstrukte, die wiederum zu Skalen zusammengefasst werden können.

1.2 Einsatzmöglichkeiten von Wirksamkeitsstudien

Eine Erhebung zu mehreren Messzeitpunkten erlaubt es im Sinne einer Lernverlaufsdiagnostik (vgl. Maier 2014, 19-22), anhand dieser Indikatoren die Entwicklung der Fähigkeiten und Fertigkeiten von Schülerinnen und Schülern über einen festgelegten Zeitraum hinweg zu untersuchen. Solche „Value-added-Modelle gelten heute als state of the art, wenn es um die quantitative Abschätzung der Wirksamkeit des Unterrichts geht“ (Helmke 2014, 812). Eingesetzt werden Studien dieser Art deshalb häufig, um die Wirksamkeit von Lehrmethoden oder Fördermaßnahmen zu evaluieren, beispielsweise eines neu konzipierten Lesetrainings. So kann man herausfinden, ob sich z.B. die Lesekompetenz (als abhängige Variable ‚AV‘) durch dieses Training im Vergleich zu regulärem (Lese-)Unterricht (unabhängige Variable ‚UV‘) effektiv steigern lässt und ob diese Steigerung nachhaltig ist. Das heißt, ob der Effekt auch längere Zeit nach dem Training noch anhält oder ob lediglich ein kurzfristiger Effekt (sogenannte Coaching-Effekt) vorliegt. Kurz gesagt: Ob das „didaktisch geplante Handeln einer Lehrperson [...], das auf den Wissens- und Kompetenzerwerb von Lernenden abzielt“ (Gräsel/Gniewosz 2015, 21), auch tatsächlich dazu beiträgt (vgl. Schrader 2011, 684; Ingenkamp/Lissmann 2008, 21). Dazu sind umfangreiche Vorarbeiten, v.a. theoretisch-konzeptioneller Art, notwendig, die im Folgenden am Beispiel der Evaluation eines selbstregulierten Lesestrategietrainings illustriert werden (*Burg Adlerstein. Lesetraining:* Schilcher/Stöger/Pissarek et al. 2013; für die Konzeption vgl. Pissarek/Schilcher/Pronold-Günthner 2012). Das Training wurde für die vierte bzw. fünfte Jahrgangsstufe konzipiert.

Eine quasi-experimentelle Studie wie diese greift zwar zur Prüfung ihrer Hypothesen

> auf Gruppen zurück, die nicht zufällig zusammengestellt, sondern vorgefunden oder anderweitig gebildet wurden (keine Randomisierung), behandelt diese je-

[2] Um die Effekte des regulären Unterrichts einschätzen zu können, ist es sinnvoll, diesen mittels eines Lerntagebuchs o. ä. dokumentieren zu lassen.

> doch ebenso wie im echten Experiment systematisch unterschiedlich (experimentelle Variation der unabhängigen Variable/n) und misst die in den Experimental- und Kontrollgruppen resultierenden Effekte (Döring/Bortz 2016, 193).

In der fachdidaktischen Forschung wird es der Regelfall sein, dass man auf ganze Klassen zurückgreift und somit eine *Klumpenstichprobe* (vgl. Döring/Bortz 2016, 314f.) vorliegt. Aus organisatorischen Gründen wird eine echte Zufallsstichprobenziehung in der fachdidaktischen Interventionsforschung eher die Ausnahme bleiben. Der Vorteil der Klumpenstichprobe liegt auf der Hand: Der organisatorische Aufwand ist deutlich geringer.

2. Vorarbeiten

Zu den Grundsätzen wissenschaftlichen Arbeitens gehört die Auseinandersetzung mit dem Diskurs; d.h. die Studie sollte ein relevantes Forschungsproblem untersuchen bzw. von praktischer Bedeutung sein. Bei der Durchführung sollte methodische (wissenschaftliche Methoden und Techniken) und ethische Strenge (Umgang mit Daten und Ergebnissen) angelegt werden (siehe hierzu die Beiträge von Bräuer und Vaupel sowie Iberer in diesem Band) sowie dies im Sinne einer vollständigen schriftlichen Dokumentation entsprechend festgehalten werden (vgl. Döring/Bortz 2016, 85f.).

2.1 Konstrukt klären

Die Qualität der Literaturrecherche, der Erarbeitung des Forschungsstandes sowie die Schlüssigkeit der Theoriebildung sind maßgeblich für die Validität eines Konstruktes (vgl. Döring/Bortz 2016, 94f.). Dadurch wird definiert, welche Fähigkeiten und Fertigkeiten durch die Intervention gefördert werden sollen. Zu welchem theoretischen Konstrukt die genannten Fähigkeiten und Fertigkeiten zu zählen sind, wie diese im aktuellen Forschungsdiskurs beschrieben werden und welche Theorien des Lernens ihrem Erwerb zugrunde gelegt werden können (vgl. Gräsel/Gniewosz 2015, 21), ist nicht trivial. Beispielsweise kann das laute Vorlesen eines Textes sowohl Teil der Lesekompetenz sein (vgl. Holle 2009) als auch der Sprech-, Gesprächs- und Zuhörkompetenz (vgl. Becker-Mrotzek 2004), je nachdem, wie bzw. in welchem Kontext gemessen wird.

Zwar sind theoretische Konstrukte wie z.B. Kompetenzmodelle keine ontologischen Größen, sie können jedoch – wie z.B. in den Bildungsstandards (vgl. Zeitler/Köller/Tesch 2010, 27) – *operationalisiert* (siehe Abbildung 1) werden und sind somit einer Messung zugänglich (vgl. zum Theorie-Empirie-Überbrückungsproblem: Beller 2008, 29). Durch die Operationalisierung wird bestimmt, mit welchen Indikatoren das theoretische Modell abgebildet werden soll (vgl. Döring/Bortz 2016, 228) – gewissermaßen eine ‚Übersetzung' der empirischen Strukturen in numerische Strukturen (vgl. Krauss/Bruckmeier et al. 2015, 616). Man erhält für jeden Probanden und jede AV jeweils genau einen Messwert, der anschließend im Hinblick auf das Konstrukt interpretiert werden kann (vgl. Ab-

bildung 2). Je nachdem, welcher Aspekt wie im jeweiligen Messinstrument operationalisiert wurde, erhält man nominal-, ordinal- oder metrisch skalierte Daten (vgl. Langfeldt 1984, 67). Sind im Diskurs bereits personen- oder untersuchungsbedingte Stör- bzw. Moderatorvariablen bekannt (s.u.), sollten diese bei einer Erhebung nach Möglichkeit ausgeschaltet oder kontrolliert werden (vgl. Döring/Bortz 2016, 200).

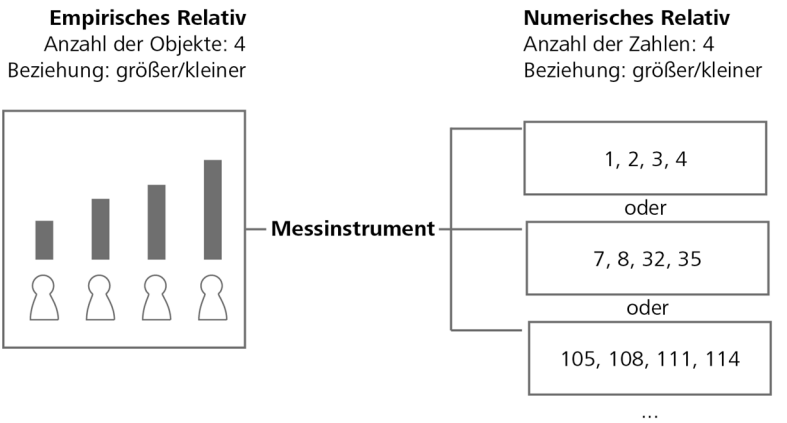

Abb. 2: Messen nach Langfeldt 1984, 66

Auf unser Interventionsbeispiel des *Regensburger selbstregulierten Lesetrainings* (*RESL*) bezogen bedeutet das, dass sich zunächst die Frage stellt, auf welchen Ebenen des Lesekompetenzmodells die Intervention wirksam sein könnte. Diese sind in einem theoretischen Modell der Lesekompetenz zu identifizieren (vgl. Rosebrock/Nix 2017; Lenhard 2013; Garbe/Holle/Jesch 2010; für eine zusammenfassende Darstellung vgl. Pissarek 2018). Der Aufbau des Lesetrainings ist an anderer Stelle ausführlich dargestellt (vgl. Pissarek/Schilcher/Pronold-Günthner 2012), sei aber hier noch einmal skizziert: Das Training umfasst zwei Induktionswochen, in denen Schülerinnen und Schüler in die Selbstregulation und drei Lesestrategien eingeführt werden. Danach folgt ein fünfwöchiges Training mit insgesamt 30 Lesetexten (mit ca. 4.500 Zeichen Länge). Um zu überprüfen, ob die Schülerinnen und Schüler ein adäquates mentales Modell der Texte bilden, befinden sich in Anschluss an jeden Text zehn Fragen, die auf den verschiedenen Ebenen der Leseanforderung (*Reproduktion*/*Zusammenhänge herstellen*/*Reflektieren und beurteilen*) angesiedelt sind. Die drei vermittelten Lesestrategien (*Nachdenken zu Fragen*/*Überfliegendes Lesen*/*Überschriften finden*) unterstützen die Prozesse auf den hierarchieniedrigen bis hierarchiehöheren Anforderungsebenen und können in der Phase des selbstregulierten Lesens von den Schülerinnen und Schülern frei gewählt und angepasst werden, um an der individuellen Schwelle zu arbeiten. Es lässt sich also erwarten, dass Effekte auf der Prozess- und Subjektebene des Lesens auftreten und auf der Prozessebene wiederum auf allen Hierarchieebenen gefördert wird. Daraus ergeben sich die Forschungsfragen.

2.2 Konkrete Forschungsfragen festlegen

Zwar besteht in der Regel schon vor Beginn einer Erhebung eine vage Forschungsidee, natürlich bereits bei der Konzeption von Interventionen, jedoch sollte diese nach der Diskussion des Konstrukts in Form von Forschungsfragen (bzw. Hypothesen[3]) konkretisiert werden. Dazu müssen konkrete, auf Basis vorliegender Theorien und Befunde „empirisch untersuchbare und auf dem aktuellen wissenschaftlichen Erkenntnisstand theoretisch erklärbare Sachverhalte adressiert werden." (Döring/Bortz 2016, 85) Werden Untersuchungsfragen zu allgemein formuliert, sind sie nicht sinnvoll untersuchbar.

Durch die Evaluation des *Burg Adlerstein*-Lesestrategietrainings sollen folgende Forschungsfragen (hier eine Auswahl) quantitativ beantwortet werden:

- Ist das Lesetraining bezüglich der Förderung von *Leseverstehen* effektiver als regulärer Leseunterricht?
- Gibt es dabei *geschlechtsspezifische* Unterschiede?
- Gibt es diesbezüglich Unterschiede bezüglich *Deutsch als Erst- und Zweitsprache*?

Aus den Forschungsfragen ergibt sich die Überlegung, was erhoben werden soll. Im vorliegenden Fall sind dies die kursiv gesetzten Konstrukte: *Leseverstehen*, *Geschlecht* und *Sprachstatus*. Da man aus der Theorie und Empirie zudem eine ausreichende *Leseflüssigkeit* allgemein als Voraussetzung für Leseverstehen annehmen kann, soll diese ebenfalls erhoben werden.

Offensichtlich geht mit der Konkretisierung der Forschungsfragen bereits eine erste Entscheidung hinsichtlich Untersuchungsdesign und Testinstrumenten einher (vgl. Döring/Bortz 2016, 149).

2.3 Testinstrumente festlegen

Wie bereits zuvor erläutert wurde, haben standardisierte Messinstrumente (vgl. Dube 2018) in der quantitativen Forschung die Aufgabe, „Merkmalsausprägungen in Form von sinnvoll interpretierbaren numerischen Messwerten" (Döring/Bortz 2016, 223) zu erfassen. Um die aufgestellten Forschungsfragen beantworten zu können, müssen in einem zweiten Schritt daher Testinstrumente gesucht werden, welche die zuvor festgelegten Teilaspekte eines Konstrukts objektiv und zuverlässig messen. Idealerweise erfolgt die Auswahl dieser *Testbatterie* mit dem Ziel, die zu untersuchenden Merkmale möglichst präzise und ökonomisch zu erfassen, aber dabei nicht mehr Daten zu produzieren, als zur Testung der Hypothesen benötigt werden (vgl. Moosbrugger/Kelava 2008, 21). In der Regel erfüllen formelle (standardisierte) Tests dieses Nebengütekriterium der

[3] Zur Unterscheidung von (Null-)Hypothesen und Forschungsfragen: Hypothesen können als wenn-dann-Formulierung konzeptualisiert werden, wobei die unabhängige Variable den ‚wenn'-Teil und die abhängige Variable den ‚dann'-Teil bildet (vgl. Bortz/Döring 2016, 43; Beller 2008, 16). Generell sind Unterschieds- und Veränderungshypothesen zu unterscheiden (vgl. Döring/Bortz 2016, 146f.).

Testökonomie sehr gut (zu den Gütekriterien siehe die Beiträge von Schmidt in diesem Band). Trotz aller Bedenken von Lehrkräften gegenüber standardisierten Tests (vgl. zusammenfassend Lukesch 1998, 519f.) ist ihr Einsatz gerade bei Evaluationsstudien nicht zuletzt deshalb sinnvoll, weil sie den Vergleich mit einer curriculumsunabhängigen Normstichprobe erlauben (vgl. Gage/Berliner 1996, 603).

Zur Auswahl geeigneter Testinstrumente bietet Langfeldt (1984, 97f.) eine ausführliche Checkliste an:

- Überprüft der Test das, was er zu messen vorgibt?
- Ist der Test reliabel (zuverlässig) genug?
- Wie präzise ist ein individueller Testpunktwert? Ist er feinkörnig genug, um Entwicklungsfortschritte bei den Zielvariablen zu erfassen?
- Wie wird eine objektive Testdurchführung gesichert?
- Wie wird die Auswertungsobjektivität gewährleistet? Wie aufwändig ist die Auswertung?
- Wie und wann ist der Test normiert? Gibt es eine Version bzw. Normierung für die angedachte Zielgruppe?
- Wie sind Testergebnisse inhaltlich zu interpretieren?
- Wie lange dauert der Test?
- Gibt es Paralleltests?

Nicht jeder Test verfügt beispielsweise über verschiedene Varianten. Während dies bei einmaligen Querschnittserhebungen (z.B. ‚Status quo') der Ausprägungen von Schülermerkmalen unbedenklich ist, entsteht für längsschnittliche Untersuchungen (sogenannte ‚Panel-Designs') wie im vorliegenden *Prä-/Post-/Follow-up-Design* ein Problem (vgl. Häder 2015, 116). Messwiederholungen können dann schlimmstenfalls

> die Ergebnisse beeinflussen und somit die interne Validität einer Studie gefährden […], sei es durch untersuchungsbedingte Testübung oder Testmüdigkeit (z.B. wiederholter Einsatz derselben oder ähnlicher Messinstrumente verändert das Ergebnis etwa im Sinne weniger sorgfältig ausgefüllter Depressionsfragebögen) oder durch Faktoren jenseits der Studie wie z.B. Reifung (z.B. im Laufe der Zeit bessert sich Depressivität teilweise von allein) oder äußerer historischer Ereignisse (Döring/Bortz 2016, 210).

Als Faustregel gilt daher: Die Anzahl der Messzeitpunkte in einem Untersuchungsdesign legt die Zahl der benötigten *Parallelversionen* eines Testinstruments fest. Liegt nur eine Parallelversion vor (z.B. beim *Salzburger Lesesceening*: SLS A und SLS B), können die Testvarianten behelfsweise im Wechsel auf die Messzeitpunkte verteilt werden (z.B. *A-B-A*), um *Übungs- oder Ermüdungseffekte* zu minimieren. Der zeitliche Abstand zwischen den Messungen sollte dann ausreichend groß sein. In unserem Fall (*RESL*) liegen zwischen MZP1 (prä) und MZP2 (post) sieben Wochen und zwischen MZP2 (post) und MZP3 (follow-up) drei Monate – man könnte also argumentieren, dass die großen zeitlichen Abstände etwaige Übungseffekte minimieren. Jedoch wäre ein Nachteil dieses Vorgehens, dass die Eichung des ursprünglichen Tests verfälscht werden könnte bzw. die Werte nicht ohne Probleme interpretiert werden können.

Zur Beantwortung der Forschungsfragen im Rahmen des vorliegenden Lesetrainings werden Instrumente benötigt, die hierarchieniedrige (*Leseflüssigkeit*) sowie hierarchiehohe Prozesse (*Leseverstehen*) des Lesens messen. Zum Erfassen von *Lesekompetenz* stehen inzwischen eine Vielzahl standardisierter und nicht-standardisierter diagnostischer Verfahren zur Verfügung (vgl. z.B. die systematische Aufstellung für das Lesen und Rechtschreiben in Pissarek/Pronold-Günthner 2018), die sich zur Lernverlaufsdiagnostik einsetzen lassen und über mehrere Parallelversionen verfügen, sodass von einer in diesem Sinne gut strukturierten Domäne gesprochen werden kann. Um eine fundierte Auswahl aus dem Angebot der in Frage kommenden Testinstrumente zu treffen, empfiehlt es sich, die Zielvariablen aufzulisten und die verfügbaren Tests nach den o.g. Merkmalen und ggf. nach Jahrgangsstufe sortiert zuzuordnen (vgl. Tab. 1). Hierbei ist zwischen ein- (z.B. *SLS*, Wimmer/Mayringer 2014; Auer/Gruber/Wimmer/Mayringer 2005) und mehrdimensionalen Test zu unterscheiden (z.B. *ELFE II*, Lenhard/Lenhard/Schneider 2018). Eindimensional bedeutet, dass in dem Test alle Testaufgaben zu einem einzigen Testwert zusammengefasst werden (bei *SLS* lediglich der *Rohwert* der korrekt gelesenen Sätze), der die Ausprägung des Konstrukts (*Leseflüssigkeit*, *Satzverstehen*) erfasst. Mehrdimensional ist hingegen ein Test, der auch Unterdimensionen des Konstruktes erfasst (z.B. *Wort-/Satz- und Textverständnis* bei ELFE II), die die Bestimmung mehrerer Subtest-Werte zulässt (zur Dimensionalität vgl. Döring/Bortz 2016, 430). Die Konstruktion eines eigenen Instruments ist nur in wenigen Fällen sinnvoll, da die Gewährleistung der Testgüte neuer Tests aufwändig und ein langwieriger Prozess ist.

Tab. 1: Exemplarische Liste der in Frage kommenden Testinstrumente

Zielvariable (AV)	**Name des Tests**	**Parallel-versionen**	**Jgst.**	**Durchführ-ungsdauer**	**Normierung (Jahr)**	**Auswertungs-aufwand**
Leseflüssig-keit	*Salzburger Le-sescreening 5-8 (SLS)*	2	5-8	3 Min.	2005 (*SLS2-9*: 2014)	gering (Schab-lone)
	...					
Lese-verste-hen	*Hamburger Le-setest 3-4 (HAMLET)*	2	3-4	90 Min.	2006	gering (Schab-lone)
	Burg Adler-stein-Fragen zum Text (trainings-nah)	3	4-5	ca. 30 Min.	-	gering (Schab-lone)

Hintergrunddaten wie *Geschlecht*, *Alter* oder *Deutschnote* (als Leistungsindikator) können relativ unkompliziert einmalig mit einem Fragebogen erhoben werden. Sie werden für die Überprüfung von Unterschiedshypothesen benötigt, um z.B. zu überprüfen, ob Mädchen und Jungen gleichermaßen oder unterschiedlich gut vom Training profitieren.

Zur Überprüfung des *RESL*-Lesetrainings wurde neben den trainingsfernen standardisierten Tests *SLS* (Leseflüssigkeit) und *HAMLET* (Leseverstehen; Lehmann/Peek/Poerschke 2006) auch ein nicht standardisiertes Testinstrument eingesetzt, das trainingsnahe Effekte beim Leseverstehen erfassen sollte. Bei Letzterem handelte es sich um Erzähltexte mit Multiple-Choice-Fragen, die analog zu den Trainingstexten konstruiert waren und zuvor pilotiert wurden. Allerdings muss man sich in diesem Fall bewusst sein, dass Schülerinnen und Schüler der Kontrollgruppe insofern benachteiligt sind, da sie im Gegensatz zur Trainingsgruppe nicht mit dem Procedere bzw. den Fragetypen vertraut sind (sogenannte ‚teaching to the test'). Standardisierte Tests haben diesen Nachteil nicht.

Wichtig zu bedenken ist auch, ob das Instrument vom Schwierigkeitsgrad her passend für die Stichprobe ist, also keine *Decken-* (zu leicht) bzw. *Bodeneffekte* (zu schwer) zu erwarten sind und ob es valide Ergebnisse für die Stichprobe liefert (z.B. so hat beispielsweise für die Gruppe der Zweitspracheschülerinnen und -schüler einen nicht unerheblichen Einfluss, in welchem Umfang der Test Wortschatz voraussetzt und es kann zu Messfehlern kommen; d.h. manche Gruppen benötigen u.U. eigene Messinstrumente).

Dies sei an einem Beispiel erläutert: Bei *RESL* war die Annahme, dass der Schwierigkeitsgrad von *HAMLET* für die Stichprobe angemessen ist, da das Training für sehr leseschwache Schülerinnen und Schüler der fünften Klasse einer Mittelschule konzipiert war und somit nicht erwartet wurde, dass dieser Leseverständnistest zu leicht ausfällt. Dennoch hat *HAMLET* bei dieser Stichprobe sogenannte Deckeneffekte bei den guten Lesern produziert, was ein Problem für die Abbildung der empirischen Daten (auch die guten Leser haben sich entwickelt) in den numerischen Daten (das Maximalergebnis aus MZP1 kann sich in MZP2 und MZP3 nicht mehr steigern) darstellt. D.h. man kann annehmen, dass auch die guten Leserinnen und Leser sich beim Leseverständnis entwickelt haben, doch in den Daten wird dies bei dieser Gruppe durch den Deckeneffekt nicht messbar.

Wichtig ist, wenn an Schulen erhoben wird, bei der Erfassung personenbezogener Daten (siehe den Beitrag von Maak in diesem Band) wie z.B. *Geschlecht*, *sozioökonomischer Status* etc. unbedingt eine Einwilligung der Erziehungsberechtigten bzw. die Zustimmung der Bildungsadministration einzuholen (vgl. Schilcher/Stöger/Wild et al. 2017). Grundsätzlich müssen die Probanden aber immer mit der Testung einverstanden sein (siehe ausführlich hierzu den Beitrag von Iberer in diesem Band).

3. Planung der Erhebung

3.1 Stichprobe planen

In der Regel wird bei einer Erhebung nicht die gesamte Population untersucht. Welcher Ausschnitt der Gesamtpopulation gewählt wird, ergibt sich aus dem Forschungsinteresse. Die für die Erhebung geplante Stichprobe sollte daher repräsen-

tativ für „genau die Population sein, für die Schlussfolgerungen (Inferenzen) getroffen werden sollen." (Krauss/Bruckmeier et al. 2015, 626) Das ist nach Beller (2008, 87) dann der Fall, wenn das Verfahren der Probandenauswahl „keine Elemente der Population in Bezug auf die interessierenden Merkmale bevorzugt" (z.B. Anteil der Kinder mit Deutsch als Zweitsprache, Verhältnis Jungen/Mädchen etc.), d.h. eine ausreichend große Zufallsauswahl ist. Wie groß die Stichprobe mindestens sein sollte, um ggf. auch kleine Effekte zu entdecken, kann mit dem Programm G*Power (http://www.gpower.hhu.de/) vorab geschätzt werden.

Denkbar sind verschiedene, jedoch unterschiedlich stark randomisierende Verfahren der Stichprobenziehung. Im schulischen Kontext und im Standardfall deutschdidaktischer Forschung dürfte die quasi-experimentelle Klumpenstichprobe mit Wartekontrollgruppe (vgl. Aeppli et al. 2011, 112ff; Kuper 2011, 139) die häufigste Form sein. In quasi-experimentellen Designs wird trotz der Untersuchung ‚im Feld' versucht, die Durchführungsbedingungen weitgehend zu kontrollieren und Störvariablen auszuschließen. Da es in der Schule meist nicht möglich ist, einzelne Schülerinnen und Schüler unterschiedlichen Gruppen (z.B. in Interventions- und Kontrollgruppe) aufzuteilen, erfolgt die Randomisierung nur auf Schul- bzw. Klassenebene (vgl. Döring/Bortz 2016, 315). Der Vergleich der *Trainingsgruppe* mit einer *Kontrollgruppe* ist schon allein deshalb sinnvoll, um die Ergebnisse der Intervention gegen Störvariablen sowie allgemeine Trends (vgl. Wottowa 1998, 118) abzusichern. *Kontrollgruppen*, in denen in der Regel regulärer Deutschunterricht stattfindet, werden häufig als *Wartekontrollgruppen* angelegt: Die beteiligten Lehrkräfte erhalten im Anschluss die gleiche Fortbildung, das gleiche Material etc. wie die Trainingsgruppe, um motivationale Einflüsse auszuschließen, die die Kontrollgruppenschülerinnen und -schüler benachteiligen würden (vgl. Kuper 2011, 139).

Bei der Replikationsstudie *RESL-Tirol* lag eine sogenannte *Konvenienz- oder Gelegenheitsstichprobe* vor, die sich aus der Kooperation mit der Bildungsadministration (dem Landesschulart Tirol und der PH Tirol, Raimund Senn) ergab: Die Klassen (Klumpen) ergaben sich automatisch durch die Lehrerpersonen, die sich eigeninitiativ und freiwillig für das Projekt interessierten und vom Landesschulrat unterstützt wurden (Freistellung für die Fortbildung, Unterstützung bei den Materialkosten). Die hohen Standards einer Randomisierung wären hier nicht möglich gewesen. Wir halten das für eine pragmatische und vertretbare Realität bei fachdidaktischen Wirksamkeitsstudien im Feld. Bei *RESL-Tirol* liegen die Daten von 447 Schülerinnen vor, die sich auf Testgruppe (N=276) und Kontrollgruppe (N=171) verteilen, wobei in der Testgruppe sieben Schulen mit 14 Klassen beteiligt waren und in der Kontrollgruppe sieben Schulen mit neun Klassen. Der Anteil der Erstspracheschülerinnen und -schüler lag bei 81,2% (Testgruppe) bzw. 80,1% (Kontrollgruppe) für beide Gruppen vergleichbar hoch. Der Anteil der Jungen fiel mit 56,9% etwas höher aus als in der Gesamtpopulation, was an der Schulart (Neue Mittelschule) liegen könnte.

3.2 Zeitliche Organisation von Messzeitpunkten und Intervention

Lernprozesse an sich können durch die oben genannten Testinstrumente nicht erfasst werden, lediglich die „zu verschiedenen Zeitpunkten beobachtbaren Zustände" (Kormann 1984, 117) von Fähigkeiten und Fertigkeiten können gemessen werden. In einem Prä-Post-Follow-up-Design werden üblicherweise drei Messzeitpunkte beobachtet. *Kontrollgruppe* und *Treatmentgruppe* werden dazu mit den jeweils gleichen Messinstrumenten gleichzeitig (*SLS*, *HAMLET*, *Burg Adlerstein*)[4] getestet. Kleine Verschiebungen des konkreten Testtages sind möglich, sollten aber nicht mehr als zwei bis drei Tage betragen (vgl. Tab. 2). Neben der Intervention muss also auch ausreichend Zeit für die Messungen eingeplant werden.

Tab. 2: Einsatz der Instrumente

	Prä (*Baseline*)		**Post**		**Follow-Up**
Treatmentgruppe	SLS A Hamlet 1 Lesetest 1 Fragebogen (Sozialdaten, Schulnoten, Geschlecht etc.)	Intervention [von 22. Oktober 2012 bis 7. Dezember 2012 (7 Wochen)]	SLS B Hamlet 2 Lesetest 2	7. März 2013 (3 Monate)	SLS A Hamlet 1 Lesetest 3
Kontrollgruppe	SLS A Hamlet 1 Lesetest 1 Fragebogen (Sozialdaten, Schulnoten, Geschlecht etc.)	Normaler Unterricht	SLS B Hamlet 2 Lesetest 2	7. März 2013 (3 Monate)	SLS A Hamlet 1 Lesetest 3

Bevor für jeden Messzeitpunkt ein konkretes Datum festgelegt wird, sollte man sich zunächst einen Überblick über das zur Verfügung stehende Zeitkontingent verschaffen. Insbesondere gilt es, schulorganisatorische Besonderheiten wie Schulferien, Feiertage, Zeugnisphase etc. bei der Planung der Untersuchung zu berücksichtigen. Besteht bereits Kontakt zu den Schulen, können auch Besonderheiten auf Schulebene wie Klassenarbeiten, Ferienlager, Exkursionen berücksichtigt werden. Hierbei kann ein Schulferienkalender hilfreich sein, in dem die in Frage kommenden Zeitabschnitte eingetragen sowie organisatorische Besonderheiten ergänzt werden. Erst dann sollte ein konkreter Zeitplan erstellt werden.

[4] Auf weitere in der Studie verwendete Instrumente wird hier aus Raumgründen nicht eingegangen.

Bereits an dieser Stelle sollte darüber nachgedacht werden, wer die Intervention bzw. Testungen durchführt und wie die Ergebnisse an die wissenschaftliche Leitung zurückgemeldet werden: z.B. Versuchsleiter oder von Lehrkräften. Eine intensive Schulung zur Durchführung (s.u.) ist allerdings in beiden Fällen notwendig.

Über *Datenmanagement* und die wissenschaftlichen Gepflogenheiten zur *Dokumentation* der Studie (vgl. Allianz der dt. Wissenschaftsorganisationen 2010, 2) sollte man sich ebenfalls schon jetzt Gedanken machen, z.B. Organisation eines Archivraums, verschließbarer Schrank, externe Speichermedien etc.

4. Durchführung der Intervention und Erhebungen

4.1 Schulung für Testung und Intervention

Sowohl die Testungen als auch die Intervention selbst müssen sorgfältig und sachgerecht durchgeführt werden. In der quantitativen Forschung betrifft dies neben der „Herstellung vergleichbarer und anonymer Untersuchungsbedingungen" (Döring/Bortz 2016, 96) auch die Schulung aller Beteiligten. Beispielsweise ist ungeschulten Versuchsleitern häufig nicht bewusst, dass sich selbst kleine Abweichungen von der Durchführungsanweisung eines Tests, wie z.B. Zeitzugaben, in mangelhafter Datenqualität, fehlender Vergleichbarkeit oder reduzierter Konstruktvalidität niederschlagen (vgl. Döring/Bortz 2016, 96). Tritt ein solcher Fall auf, müssen die Daten von der Untersuchung ausgeschlossen werden. Die Versuchsleiter sollten daher im Rahmen einer Fortbildung bereits vorab die Möglichkeit haben, sich mit den Tests auseinanderzusetzen und diese ggf. selbst auszuprobieren.

Gleiches gilt für die Intervention selbst: Da die Effekte des Trainings gemessen werden sollen, ist es wichtig, dass dieses wie vorgesehen durchgeführt wird. Andernfalls misst man lediglich Effekte des regulären Unterrichts.

Hinzu kommt das Datenmanagement und evtl. die Pseudonymisierung der Probanden (vgl. 6.1 und den Beitrag von Iberer in diesem Band). Bei *RESL-Tirol* wurde die Pseudonymisierung durch die Lehrkräfte selbst vor Ort vorgenommen. Nur Sie konnten die Schüler-IDs (Codes) mit den Klassenlisten abgleichen. Bei der Übermittlung der Testhefte gab es in der rechten oberen Ecke ein Feld (mit dem Namen), das vor der Übersendung der Testunterlagen von den Lehrkräften abgeschnitten wurde. So war einerseits eine eineindeutige Zuordnung der Schülercodes zu den Kindern gewährleistet (Vermeiden von Vertauschungen und Flüchtigkeitsfehlern), zugleich aber auch die geforderte Pseudonymisierung.

5. Testdurchführung und Intervention

Das Beachten der in der Schulung festgelegten Gütemaßstäbe ist auch bei der Durchführung sicherzustellen. Gegebenenfalls werden dazu entsprechende Terminpläne, Tagebücher oder ‚Spickzettel' etc. für die Versuchsleiter vorbereitet. Darüber hinaus betrifft dies die pseudonymisierte Übermittlung (der bearbeiteten Tests) an die wissenschaftliche Leitung sowie evtl. die Auswertung der Testbögen (s.u. Datenmanagement). Testbeschreibungen und exemplarische Aufgaben müssen in der Regel während der Studie zur Einsichtnahme für alle Beteiligten bzw. Erziehungsberechtigten vorliegen (vgl. von der Gathen 2011, 67f.).

Eine zeitnahe Aufbereitung von Ergebnissen auf Schul- bzw. Klassenebene ist nicht nur für die Forscherinnen und Forscher interessant, auch für beteiligte Lehrkräfte sind diese aufschlussreich und nützlich für die Verbesserung des eigenen Unterrichts. Bewährt hat sich hier die Arbeit mit Excel-Tabellen, in die die Ergebnisse eingetragen und im Anschluss auf Klassenebene mit einer kurzen Zusammenfassung rückgemeldet werden. Erste einfache Berechnungen können in der Datei hinterlegt werden: z.B. Wie viele Schülerinnen und Schüler lernen erfolgreich? In welchen Bereichen lassen sich Stärken bzw. Schwächen feststellen? Wie schlagen sich die Schülerinnen und Schüler im Vergleich zur Gleichaltrigengruppe? Nach dem Posttest kann zudem auch zurückgemeldet werden, wo die Intervention (nicht) wirkt (vgl. Stellungnahme der DFG zur Replikationsdebatte 2017). Ein weiterer Vorteil einer zeitnahen Rückmeldung liegt darin, dass Inkonsistenzen in den Daten gleich auffallen und in der Regel mit den Versuchsleitern geklärt werden können. Zudem können diese Tabellen ohne großen Aufwand in Statistikprogramme wie SPSS oder R importiert werden. Bei *RESL-Tirol* erhielten die Lehrkräfte unmittelbar nach der Intervention eine persönliche Rückmeldung (Weihnachten 2012), die die Ergebnisse aus MZP2 enthielt. Die Präsentation der Gesamtanalyse erfolgte im Mai 2012 im Rahmen eines weiteren Nachbereitungsseminars. Der unmittelbare Austausch und die Diskussion der Umsetzung waren für beide Seiten (Schule und Universität) höchst gewinnbringend. Wenn möglich, ist eine persönliche und zeitnahe Rückmeldung anzustreben, erhöht sie doch auch die Akzeptanz und Bereitschaft auch künftig an weiteren Studien teilzunehmen.

6. Datenmanagement

6.1 Vorbereitende Maßnahmen und Planung

Bereits vor der ersten Erhebung sollte eine Dokumentation aller Variablen (*Codebook*) und ggf. ein Kodiermanual angelegt werden: Hierin werden alle Variablen kurz, evtl. mit Literaturverweis, beschrieben und soweit nötig ihre Zusammensetzung oder Wertebereich erklärt (vgl. Tab. 3). Ein einheitliches Benennungsschema erleichtert später die Orientierung im Datensatz.

Tab. 3: Beispielhafte Dokumentation eines pseudonymisierten numerischen Schülercodes „1207“ auf Nominalskalenniveau[5]

Schülercode (Skalenniveau: nominal, 4-stellig)			
s_code	1	2	07
	Schule	*Klasse (hier Klasse Nr. 2 an Schule Nr. 1)*	*Schülernummer*

Die Variable „s_code“ bildet im vorliegenden Beispiel die Schlüsselvariable des Datensatzes. Anhand dieser können die Testwerte der Schülerinnen und Schüler (AV) eindeutig zugeordnet werden. Jeder Wert kommt in dieser Spalte nur ein einziges Mal vor und darf sich für eine Schülerin bzw. einen Schüler während der Erhebung nicht ändern.

Die unabhängigen (personenbezogenen) Variablen sind im vorliegenden Fall: Geschlecht (*s_sex*), Sprache (*s_lang*) sowie die Deutschnote im Jahreszeugnis 2012 (*s_dnote*).

Die abhängigen Variablen sind die Daten der Testinstrumente zu den drei Messzeitpunkten (x_SLS, y_SLS, z_SLS; x_HAMLET, y_HAMLET usw.). Die Kennzeichnung der MZPs durch führend „x_“, „y_“, „z_“ erlaubt ein sicheres ‚Suchen und Ersetzen‘ in der R-/SPSS-Syntax.

Variablen, die erst später berechnet werden, werden nicht manuell angelegt. Es ist allerdings empfehlenswert, sich, bereits bevor die ersten Daten eintreffen, einen leeren Datensatz mit der entsprechenden Syntax anzulegen und zu kommentieren. Variablen-, Wertelabels sowie Definitionen für Missings und evtl. Berechnungen, etwa der Klassencodes, sollten ebenfalls hinterlegt werden.

Tab. 4: Beispielhafte SPSS-Syntax zur Berechnung der Schule auf Basis des Schülercodes

```
* Nach Schulen aufteilen.                          Kommentar mit Erklärung der Berechnung
COMPUTE s_school=TRUNC(s_code/1000).               Berechnung der neuen Variable s_school
VARIABLE LABELS s_school 's_school Schul-ID'.      Beschriftung der neuen Variable
EXECUTE.                                           Ausführen des Befehls
```

Unabhängig davon muss auch die Erhebung dokumentiert werden. Das jeweilige Datum sollte auch in den Datensatz eingetragen werden. Als hilfreich haben sich zur Übersicht Checklisten erwiesen, in denen erfasst wird, von welchem Standort bereits Testinstrumente und/oder Ergebnisse eingetroffen sind und wo sie noch fehlen.

[5] Ein ausführliches Beispiel kann bei den Autoren angefordert werden.

Tab. 4: Anriss der Checklisten-Tabelle

			digital			print					
Schule	**Kl.-Code**	**Klasse**	**MZP 1**	**MZP2**	**MZP3**	**MZP 1**	**MZP2**	**MZP3**	**N**	**Ansprech-partner**	**E-Mail**
Muster-schule Muster-hausen	11xx	1a	SLS HAM-LET BA	SLS HAM-LET BA	SLS HAM-LET BA	SLS HAM-LET BA	SLS HAM-LET BA	SLS HAM-LET BA	25	1: Herr Max Muster	m.m@gmx.de
Beispiel-schule Beispiel-hausen	21xx	1c	SLS HAM-LET BA	SLS HAM-LET BA	SLS HAM-LET BA	SLS HAM-LET BA	SLS HAM-LET BA	SLS HAM-LET BA	26	1: Frau Berta Beispiel 2: Frau Brunhilde Beispiel	b_b@gmx.de
...	...	...	...	...	...	...	...	...	...	...	...

6.2 Dateneingabe, Speicherung und Archivierung (digital und analog)

Sobald die Datenerhebung zu einem Messzeitpunkt abgeschlossen ist, kann nach dem zuvor festgelegten Schema mit der Überführung der analogen Daten in eine digitale Datenmatrix (z.B. in der Software R oder SPSS) begonnen werden. Dazu

> muss zunächst sorgfältig sortiert, kommentiert, formatiert, anonymisiert, bereinigt und oft auch transformiert werden, um eine systematische Datenanalyse überhaupt zu ermöglichen (Döring/Bortz 2016, 580).

Merkmalsausprägungen werden anschließend gemäß des empirisch-numerischen Relativs kodiert (z.B. nominal für das Geschlecht *s_sex*: weiblich = 0; männlich = 1). Wird aus bestimmten Gründen davon abgewichen, sollte dies in der Variablendokumentation mit Datum vermerkt werden.

Gleiches gilt für Änderungen am Datensatz: Die Dateien sollten auf verschiedenen Medien mit Datum versehen gesichert werden. Empfehlenswert ist folgendes Schema, da die Dateien dadurch auf dem Datenträger auch nach Datum sortiert werden: [Name]-[Jahr]-[Monat]-[Tag], z.B. *RESL-2012-06-03*. Auf jeden Fall sollte nach Abschluss der Erhebungen ein *Rohdatensatz* aufbewahrt werden, der eine unbearbeitete Version der Daten enthält.

Zusätzlich zu den digitalen Daten müssen die Paper-and-Pencil-Version der Testinstrumente aufbewahrt und auf geeignete Art und Weise archiviert werden, sodass ggf. Zweifelsfälle überprüft werden können oder andere Forscherinnen und Forscher die Daten einsehen können (vgl. Allianz der dt. Wissenschaftsorganisationen 2010):

> Primärdaten sollen gesichert und mindestens zehn Jahre lang aufbewahrt werden (typischerweise an der Forschungseinrichtung, an der die Studie durchgeführt wurde), um Nachvollziehbarkeit zu gewährleisten und z.B. Wissenschaftsfälschungen aufdecken zu können. (Döring/Bortz 2016, 133)

Es ist also wichtig, die Frage der Archivierung (Raumressourcen) mitzubedenken, da hier einige Meter Regalfläche verbraucht werden.

Tab. 5: Ausschnitt der Datenmatrix der Evaluation des Burg Adlerstein-Lesetrainings. Die Variablen ‚s_school' und ‚s_class' wurde automatisch berechnet (s.o.)

s_code	s_school	s_class	s_sex	s_lang	s_dnote	x_SLS	x_HAMLET
1101	1	11	1	2	2	27	7
1102	1	11	1	1	3	19	9
1103	1	11	1	1	1	27	6
...	...	...	...	...	...	...	...
2104	2	21	1	1	3	23	10
2105	2	21	1	2	3	26	8
...	...	...	...	...	...	...	...

7. Datenauswertung

7.1 Datensatz prüfen

> Zur Datenaufbereitung („data preparation") gehören all jene begründeten und dokumentierten Bearbeitungen bzw. Veränderungen des Rohdatenmaterials, welche die Aussagekraft und (Wieder-)Verwendbarkeit der Daten steigern und die inhaltliche Datenanalyse vorbereiten. (Döring/Bortz 2016, 580)

Bevor Berechnungen durchgeführt werden können, muss der Datensatz also auf Fehler geprüft werden, die die Ergebnisse verfälschen könnten. Dies gelingt bereits mit einfachen Mitteln sehr gut. Über die Sortierfunktion und deskriptive Statistiken wie Häufigkeiten können leicht falsche Wertelabels, Fehleingaben, unplausible oder fehlende Werte identifiziert werden. Je nach Polung der Variablen müssen diese ggf. (in eine neue Variable) transformiert werden.

Tab. 6: Beispiel für eine Fehleingabe; Die Variable ‚s_lang' kann laut Dokumentation nur die Werte 0 oder 1 annehmen

s_lang

		Häufigkeit	%	Gültige %	Kumulierte %
Gültig	Deutsch als Erstsprache	360	80,4	80,7	80,7
	Deutsch als Zweitsprache	85	19,0	19,1	99,8
	11	1	0,2	0,2	100,0
	Gesamt	446	99,6	100,0	
Fehlend	System	2	0,4		
Gesamt		448	100,0		

Werden Fehler gefunden, muss ein Abgleich mit der Papierversion stattfinden. Häufig hilft auch die Rücksprache mit dem Testdurchführenden weiter: Z.B. könnte bei *SLS* Zeit zugegeben worden sein, was sich in zu hohen Testwerten niederschlägt. Können die Zweifel nicht beseitigt werden, müssen die betroffenen Daten von den Berechnungen ausgeschlossen werden. Grundsätzlich sind deshalb

stichprobenartige Vergleiche zwischen digitaler und print-Version sinnvoll, denn eine

> Studie kann sich nur dann auf hohe Datenqualität berufen, wenn das sorgfältig erhobene Rohdatenmaterial [...] einer systematischen und dokumentierten Datenbereinigung unterzogen wurde (z.B. Ausschluss unplausibler Werte und Fälle, sachgerechte Behandlung fehlender Werte; [...]). Einbußen der Datenqualität, die durch Fehler bei der Datenerhebung (z.B. mangelnde Interviewerschulung) oder Fehler bei der Operationalisierung (z.B. Nutzung eines Messinstrumentes mit zu geringer Messgenauigkeit) zustande gekommen sind, lassen sich im Zuge der Datenaufbereitung jedoch nicht mehr kompensieren. Mangelnde Datenqualität kann sich in reduzierter Konstruktvalidität [...] niederschlagen. (Döring/Bortz 2016, 96)

Bei Fragebogenerhebungen sollte zusätzlich auch auf Ankreuzmuster geachtet werden, z.B. alle Items wurden mit ‚stimme nicht zu' bewertet. Erst wenn die Überprüfung der Daten abgeschlossen ist, kann mit einer Auswertung begonnen werden.

Jegliche Änderungen an den Daten, die aufgrund dessen vorgenommen wurden, sind später im Ergebnisbericht zu dokumentieren und zu begründen (vgl. Döring/Bortz 2016, 581).

7.2 Auswertung nach Forschungsfragen und Interpretation

Für erste Auswertungen – etwa in SPSS – sind die ausführlichen und erprobten Anleitungen von Andy Field (2017) sehr empfehlenswert. Üblich ist es, zunächst deskriptive und Reliabilitätsanalysen durchzuführen.[6] Im Folgenden wird eine exemplarische Auswertung der Ergebnisse zu den in Abschnitt 2.2 aufgeworfenen Forschungsfragen erfolgen. Die Stichprobe der Replikationsstudie umfasste insgesamt N=448 Schülerinnen und Schüler, wobei in der Trainingsgruppe (N=277) und der Kontrollgruppe (N=171) der Anteil von Kindern mit Deutsch als Zweitsprache (je ca. 20%) vergleichbar war und Mädchen (insgesamt 48,9%) und Jungen (insgesamt 51,1%) ähnlich häufig vertreten waren.

Um zu beantworten, ob das *Burg Adlerstein*-Lesetraining bezüglich der Förderung des *Leseverstehens* effektiver ist als regulärer Leseunterricht, können Entwicklungsunterschiede der beiden Gruppen berechnet werden (*Treatmentgruppe* vs. *Kontrollgruppe*). Die zugrundeliegende *Unterschiedshypothese 1* wäre z.B.:

> H1: Die Intervention (TG) zeigt einen größeren Effekt auf die Lesekompetenz als regulärer Unterricht (KG).

Um Entwicklungsunterschiede standardisiert zu berichten, wird in Wirksamkeitsstudien häufig auf das Gruppendifferenz-Effektstärkenmaß *Cohens d* zurückge-

---

[6] Die üblichen statistischen Analysen zu den Messinstrumenten (Reliabilität, Trennschärfe, Schwierigkeit etc.) werden hier aus Raumgründen nicht dargestellt, aber vorausgesetzt. Deren Prinzip und Funktion wird eingehend in Schmitz in diesem Band erläutert.

griffen. Cohens *d* gibt die Wirksamkeit eines Treatments gemessen in Standardabweichungen an. Je höher der Wert ausfällt, desto wirksamer war eine Intervention. Cohens *d* kann berechnet werden, wenn ein t-Test o.ä. signifikant wird, d.h. die Unterschiede zwischen zwei Gruppen nicht zufällig sind (zu den Gütekriterien siehe die Beiträge von Schmidt in diesem Band). Als standardisiertes Effektgrößenmaß ist Cohens *d* von Messeinheiten und Stichprobenumfang unabhängig (vgl. Bortz/Döring 2016, 816) und erlaubt es so, verschiedene fachdidaktische Interventionen miteinander zu vergleichen. Zur Berechnung von Effektstärken stellt Lenhard (2018) online hilfreiche und gut dokumentiere Tools zur Verfügung, mit denen Effektstärken in *Experimental-* und *Interventionsstudien* mit Prä-/Post-/Kontrollgruppen-Design berechnet werden können. Tabelle 8 zeigt die Effektstärken-Berechnungen für die in Abschnitt 2 angeführten Unterschiedshypothesen.

Tab. 7: Untersuchung der Gruppenentwicklungen gemessen in Effektstärken

	BA prä → post	**BA prä → f.-up**	**SLS prä → post**	**SLS prä → f.-up**	**HAMLET prä → post**	**HAMLET prä → f.-up**
Trainingsgruppe vs. Kontrollgruppe	.89**	.74**	-.048 n.s.	.39**	.15 n.s.	.18 n.s.
Erstsprache Trainingsgruppe vs. Kontrollgruppe	.83**	.75**	.063 n.s.	.38*	.10 n.s.	.19 n.s.
Erstsprache Trainingsgruppe vs. Zweitsprache Trainingsgruppe	-.64**	-.51**	.12 n.s.	.05 n.s.	.35 n.s.	.23 n.s.

**: Unterschied signifikant mit p<.001; *: Unterschied signifikant mit p<.005; n.s.: nicht signifikant; BA: Burg Adlerstein (trainingsnaher Test); SLS: Salzburger Lesescreening (Leseflüssigkeit); Hamlet: Hamburger Lesetest (Textverstehen). Für die Interpretation der Effektstärken-Werte bei Gruppendifferenzmaßen gelten die Werte d=.20 als kleiner, d=.50 als mittlerer und d=.80 als großer Effekt, was „inzwischen als Faustregel weitgehend akzeptiert ist" (Bortz/Döring 2016, 820).

Aus Tabelle 8 geht hervor, dass beim trainingsnahen Testinstrument *Burg Adlerstein* jeweils große bis mittlere Effekte zugunsten der Trainingsgruppe auftreten. D.h. beim trainingsnah konzeptualisierten Textverstehen unterscheiden sich die Trainingsgruppe und die Kontrollgruppe unmittelbar nach dem Training um d=.89**. Zugleich ist dieser Unterschied nachhaltig, denn mit d=.74** bleibt er auch noch drei Monate nach Beendigung der Intervention bestehen.

Hinsichtlich der Leseflüssigkeit findet sich unmittelbar nach dem Training noch kein Unterschied zwischen Trainingsgruppe und Kontrollgruppe, jedoch liegt nach weiteren drei Monaten (Follow-up) ein mittelgroßer signifikanter Effekt (d=.39**) vor. Das heißt, die Trainingsgruppe zeigt erst verspätet einen Automatisierungseffekt gegenüber der Kontrollgruppe.

HAMLET als trainingsferner Test weist zwar zu beiden Zeitpunkten kleine Effekte auf, sie werden allerdings nicht signifikant. Hier hat vermutlich der oben geschilderte Deckeneffekt des Messinstruments verhindert, dass die Entwicklungsunterschiede voll erfasst werden.

DaZ-Schülerinnen und -schüler profitieren im trainingsnahen Test von dem *Burg Adlerstein*-Lesetraining stärker als Kinder mit deutscher Erstsprache (d=-.64** und d=-.51**) und das, obwohl die gesamte Trainingsgruppe ohnehin schon stark von dem Training profitiert. Kritisch anzumerken ist, dass diese Effekte nur bei den trainingsnahen Aufgaben sichtbar werden. Insofern ist im Sinne der gebotenen „(selbst)kritische[n] Reflexion von Forschungsergebnissen" (Bortz/Döring 2016, 132) hier auch die Frage zu stellen, ob die Ergebnisse aus dem Testinstrument *Burg Adlerstein* generalisiert werden dürfen, auch wenn das Niveau der Verständnisfragen stark an die Konzeptualisierung der Fragen in *IGLU* und *PISA* angelehnt ist. Als nächster Schritt wäre nun die Diskussion der Ergebnisse vor dem Hintergrund des aktuellen Forschungsdiskurses (Leseforschung und DaZ-Forschung) zu leisten und eine Präsentation der Ergebnisse vor Fachpublikum.

8. Fazit

Das Prä-/Post-/Follow-Up-Kontrollgruppendesign stellt ein typisches experimentelles Design fachdidaktischer Wirksamkeits-Forschung dar. Es eignet sich vor allem in Evaluationsstudien, um die Wirksamkeit von fachdidaktischen Konzepten und Materialien in der Praxis empirisch zu überprüfen. Dabei garantiert das Vorliegen einer Kontrollgruppe, dass die analysierten Ursache-Wirkung-Beziehungen auf das *Treatment* bzw. die *Intervention* zurückgeführt werden können – wobei durch die geschilderte Vorgehensweise der Klumpenstichprobenziehung die schulische Realität mitberücksichtigt wird und der hohe Anspruch einer völlig randomisierten Stichprobenziehung meist nicht erfüllt wird. Auch wenn die Schwierigkeit, die Erhebungen zu drei verschiedenen Messzeitpunkten synchron an mehreren Standorten durchzuführen, nicht unerheblich ist, so stellt sie doch eine der besten Möglichkeiten dar, die verändernde Wirkung eines Treatments auf eine abhängige Variable zu überprüfen. Dabei ist im Rahmen des Möglichen auf vergleichbare Ausgangsbedingungen in den Klassen/Schulen der Treatment- und Kontrollgruppen zu achten. Das standardisierte Effektstärkemaß *Cohens d* ermöglicht es dabei, die Wirksamkeitsüberprüfung mit den Ergebnissen aus anderen Interventionen vergleichbaren Umfangs zu vergleichen.

Literatur

Aeppli, Jürg/Gasser, Luciano/Gutzwiller, Eveline/Tettenborn, Annette (2011): Empirisches wissenschaftliches Arbeiten. Ein Studienbuch für die Bildungswissenschaften. Bad Heilbrunn: Klinkhardt.

Allianz der deutschen Wissenschaftsorganisationen: Grundsätze zum Umgang mit Forschungsdaten: http://www.allianzinitiative.de/fileadmin/user_upload/www.allianzinitiative.de/Grundsaetze_Forschungsdaten_2010.pdf (letzter Zugriff: 01.08.2018).

Auer, Michaela/Gruber, Gabriele/Wimmer, Heinz/Mayringer, Heinz (2005): Salzburger Lese-Screening für die Klassenstufen 5-8. Bern: Huber.

Becker-Mrotzek, Michael (2004): Kernkompetenzen im Bereich von Mündlichkeit und Schriftlichkeit. In: Kämper-van den Boogaart, Michael (Hrsg.): Deutschunterricht nach der PISA-Studie. Reaktionen der Deutschdidaktik. Frankfurt a.M.: Peter Lang, 143-152.

Beller, Sieghard (2008): Empirisch forschen lernen. Konzepte, Methoden, Fallbeispiele, Tipps. 2. Aufl. Bern: Huber.

Deutsche Forschungsgemeinschaft (DFG): Replizierbarkeit von Forschungsergebnissen. Eine Stellungnahme der Deutschen Forschungsgemeinschaft, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_de.pdf (letzter Zugriff: 01.08.2018).

Döring, Nicola/Bortz, Jürgen (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 5. Aufl. Berlin/Heidelberg: Springer.

Dube, Juliane (2018): Standardisierte Testverfahren. In: Boelmann, Jan M. (Hrsg.): Empirische Forschung in der Deutschdidaktik. Band 2: Erhebungs- und Auswertungsverfahren. Baltmannsweiler: Schneider Hohengehren, 115-130.

Field, Andy (2017): Discovering Statistics Using IBM SPSS. 5. Aufl. London: Sage.

Gage, Nathaniel/Berliner, David (1996): Pädagogische Psychologie. 5. Aufl. Weinheim: Beltz.

Garbe, Christine/Holle, Karl/Jesch, Tatjana (2010): Texte lesen: Lesekompetenz – Textverstehen – Lesedidaktik – Lesesozialisation. 2. Aufl. Stuttgart: Schöningh.

Gerrig, Richard/Zimbardo, Philip (2008): Psychologie. 18. Aufl. München u.a.: Pearson.

Gräsel, Cornelia/Gniewosz, Burkhard (2015): Überblick Lehr-Lernforschung. In: Reinders, Heinz/Ditton, Hartmut/Gräsel, Cornelia/Gniewosz, Burkhard (Hrsg.): Empirische Bildungsforschung. Gegenstandsbereiche. 2. Aufl. Wiesbaden: Springer VS, 19-24.

Häder, Michael (2015): Empirische Sozialforschung. Eine Einführung. 3. Aufl. Wiesbaden: Springer VS.

Helmke, Andreas (2009): Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts. Wiesbaden: Klett Kallmeyer.

Helmke, Andreas (2014): Forschung zur Lernwirksamkeit des Lehrerhandelns. In: Terhart, Ewald/Bennewitz, Hedda/Rothland, Martin (Hrsg.): Handbuch der Forschung zum Lehrerberuf. 2. Aufl. Münster: Waxmann, 807-821.

Holle, Karl (2009): Psychologische Lesemodelle und ihre lesedidaktischen Implikationen. In: Garbe, Christine/Holle, Karl/Jesch, Tatjana: Texte lesen. Lesekompetenz – Textverstehen – Lesedidaktik – Lesesozialisation. Paderborn: Schöningh, 103-166.

Ingenkamp, Karl-Heinz/Lissmann, Urban (2008): Lehrbuch der Pädagogischen Diagnostik. Weinheim: Beltz.

Klieme, Eckhard (2006): Empirische Unterrichtsforschung: aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. In: Zeitschrift für Pädagogik 52, 6, 765-773.

Kormann, Adam (1984): Konzepte der Veränderungs- und Lernmessung. In: Heller, Kurt (Hrsg.): Leistungsdiagnostik in der Schule. Bern: Huber, 117-124.

Krauss, Stefan/Bruckmaier, Georg/Schmeisser, Christine/Brunner, Martin (2016): Quantitative Forschungsmethoden in der Mathematikdidaktik. In: Bruder, Regine/Hefendehl-Hebeker, Lis/Schmidt-Thieme, Barbara/Weigand, Hans-Georg (Hrsg.): Handbuch der Mathematikdidaktik. Berlin: Springer Spektrum, 613-641.

Kuper, Harm (2011): Evaluation. In: Reinders, Heinz/Ditton, Hartmut/Gräsel, Cornelia/Gniewosz, Burkhard (Hrsg.): Empirische Bildungsforschung. Strukturen und Methoden. Wiesbaden: Springer VS, 131-144.

Langfeldt, Hans-Peter (1984): Die klassische Testtheorie als Grundlage normorientierter (standardisierter) Schulleistungstests. In: Hellert, Kurt (Hrsg.): Leistungsdiagnostik in der Schule. Bern: Huber, 65-98.

Lehmann, Rainer H./Peek, Rainer/Poerschke, Jan (2006): Hamburger Lesetest für 3. und 4. Klassen (HAMLET 3-4). Göttingen: Hogrefe.

Lenhard, Alexandra (2018): Psychometrica. Berechnung von Effektstärken. Online unter: https://www.psychometrica.de/effektstaerke.html (letzter Zugriff: 01.08.2018).

Lenhard, Wolfgang (2013): Leseverständnis und Lesekompetenz: Grundlagen – Diagnostik – Förderung. Stuttgart: Kohlhammer.

Lenhard, Wolfgang/Lenhard, Alexandra/Schneider, Wolfgang (2018): ELFE II: ein Leseverständnistest für Erst- bis Siebtklässler – Version II. 2., korr Aufl. Göttingen: Hogrefe.

Leutner, Detlev/Fleischer, Jens/Grünkorn, Juliane/Klieme, Eckhard (2017): Competence Assessment in Education. An Introduction. In: Leutner, Detlev/Fleischer, Jens/Grünkorn, Juliane/Klieme, Eckhard (Hrsg.): Competence Assessment in Education. Research Models and Instruments. Cham (CH). Springer International Publishing, 1-8.

Lukesch, Helmut (1998): Einführung in die pädagogisch-psychologische Diagnostik. 2. Aufl. Regensburg: Roderer.

Maier, Uwe (2014): Formative Leistungsdiagnostik in der Sekundarstufe – Grundlegende Fragen, domänenspezifische Verfahren und empirische Befunde. In Hasselhorn, Marcus/Schneider, Wolfgang/Trautwein, Ulrich (Hrsg.): Lernverlaufsdiagnostik. Göttingen: Hogrefe, 19-39.

Moosbrugger, Helfried/Kelava, Augustin (2008): Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: Moosbrugger, Helfried/Kelava, Augustin (Hrsg.): Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer, 7-26.

Pissarek, Markus (2018): Zum Begriff der Lesekompetenz – Förderung von Lesekompetenz bei jungen Erwachsenen. In: Resinger, Paul/Volgger, Angela: Förderung der Lesekompetenz von Lehrlingen. Innsbruck: Studien Verlag, 5-14.

Pissarek, Markus/Pronold-Günthner, Friederike (2018): Lernvoraussetzungen ermitteln am Beispiel der Lese- und Rechtschreibkompetenz. In: Schilcher, Anita/Finkenzeller, Kurt/Knott, Christina/Pronold-Günthner, Friederike/Wild, Johannes (Hrsg.): Schritt für Schritt zum guten Deutschunterricht. Praxisbuch für Studium und Referendariat: Strategien und Methoden für professionelle Deutschlehrkräfte. Seelze: Klett Kallmeyer, 119-132.

Pissarek, Markus/Schilcher, Anita/Pronold-Günthner, Friederike (2012): Strategietraining auf Burg Adlerstein – das Regensburger selbstregulierte Lesetraining (RESL). In: Philipp, Maik/Schilcher, Anita (Hrsg.): Selbstreguliertes Lesen. Ein Überblick über wirksame Leseförderansätze. Seelze: Friedrich, 158-173.

Reinders, Heinz/Ditton, Hartmut/Gräsel, Cornelia/Gniewosz, Burkhard (2015): Vorwort. In: Reinders, Heinz/Ditton, Hartmut/Gräsel, Cornelia/Gniewosz, Burkhard (Hrsg.): Empirische Bildungsforschung. Gegenstandsbereiche. 2. Aufl. Wiesbaden: Springer VS, 11-18.

Roos, Markus (2001): Ganzheitliches Beurteilen und Fördern in der Primarschule. Chur: Rüegger.

Rosebrock, Cornelia/Nix, Daniel (2017): Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung 8., korr. Aufl. Baltmannsweiler: Schneider Hohengehren.

Schilcher, Anita/Stöger, Heidrun/Pissarek, Markus/Sontag, Christine/Pronold-Günthner, Friederike/Steinbach, Julia (2013): Burg Adlerstein – Lesetraining. Braunschweig. Westermann.

Schilcher, Anita/Stöger, Heidrun/Wild, Johannes (2017): Herausforderungen und Chancen der Zusammenarbeit von Forschungs- und Bildungsinstitutionen. Über die Schwierigkeit, Daten an Schulen zu erheben. In: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache (Hrsg.): „Blick zurück nach vorn“. Perspektiven für sprachliche Bildung in Lehrerbildung und Forschung. Köln: Mercator-Institut, 21-25.

Schrader, Friedrich-Wilhelm (2011): Lehrer als Diagnostiker. In: Terhart, Ewald/Bennewitz, Hedda/Rothland, Martin (Hrsg.): Handbuch der Forschung zum Lehrerberuf. Münster: Waxmann, 683-698.

Senn, Werner/Krelle, Michael (2016): Einleitung – Qualitäten von Deutschunterricht. In: Senn, Werner/Krelle, Michael (Hrsg.): Qualitäten von Deutschunterricht: Empirische Unterrichtsforschung im Fach Deutsch. Stuttgart: Fillibach.

Smit, Robbert (2009): Die formative Beurteilung und ihr Nutzen für die Entwicklung der Lernkompetenz. Baltmannsweiler: Schneider Hohengehren.

von der Gathen, Jan (2011): Leistungsrückmeldungen bei Large-Scale-Assessments und Vollerhebungen. Rezeption und Nutzung am Beispiel von DESI und lernstand. Münster: Waxmann.

Weinert, Franz (2001): Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In: Weinert, Franz (Hrsg.): Leistungsmessungen in Schulen. 2. Aufl. Weinheim: Beltz, 17-32.

Wimmer, Heinz/Mayringer, Heinz (2014): Salzburger Lese-Screening für die Schulstufen 2-9. Bern: Huber.

Wottawa, Heinrich (1998): Evaluation. In: Rost, Detlef H. (Hrsg.): Handwörterbuch Pädagogische Psychologie. Weinheim: Beltz, 116-121.

Zeitler, Sigrid/Köller, Olaf/Tesch, Bernd (2010): Bildungsstandards und ihre Implikationen für Qualitätssicherung und Qualitätsentwicklung. In: Gehrmann, Axel/Hericks, Uwe/Lüders, Manfred (Hrsg.): Bildungsstandards und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht. Bad Heilbrunn: Klinkhardt, 23-36.

LISA KÖNIG

Planung und Vorbereitung empirischer Erhebungen

Oder: Der lange Weg zur eigentlichen Arbeit

Große Visionen und die Hoffnung nach weitreichenden Ergebnissen des eigenen empirischen Vorhabens prägen den Beginn eines jeden Forschungsprozesses. Doch egal ob aufstrebender Nachwuchswissenschaftler, arrivierter Professor oder Studierender während der Bachelor- oder Masterarbeit – jeder Forschende steht alsbald vor den Herausforderungen der konkreten Übertragung der eigenen Ideen in ein empirisches Setting.

Die Umsetzung des Forschungsvorhabens mittels der Wahl des Forschungsansatzes, Forschungsparadigma, Forschungsdesigns sowie der empirischen Erhebungs- und Auswertungsverfahren zeigt sich jedoch keineswegs als problemlos. Zumeist stammen auftretende Komplikationen aus einem der folgenden zwei Problemfelder: einer unspezifischen Zielsetzung oder Fragestellung, welche eine konkrete Planung des empirischen Settings verhindert, oder der vorschnellen Festlegung auf eine spezifische Forschungsmethode, die ohne eine reflektierte Auseinandersetzung hinsichtlich der Passung des Vorhabens ausgewählt wird. Um diesen Problematiken vorzubeugen und sowohl dem Forschungsvorhaben entsprechende als auch aussagekräftige Daten zu generieren, bedarf es einer gründlichen Planung und Vorbereitung der eigentlichen Erhebung, bevor diese anschließend durchgeführt wird.[1]

[1] Die tatsächliche Erhebung evidenzbasierter Daten wird zu Recht als das Herzstück des Forschungsvorhabens angesehen, stellt jedoch lediglich einen kleinen Abschnitt innerhalb der gesamten Projektumsetzung dar. Während dieser Beitrag die spezifischen Planungs- und Vorbereitungsphasen empirischer Erhebungen thematisiert, erfolgt innerhalb des Diskurses der empirischen Bildungsforschung die Strukturierung des ganzheitlichen Forschungsprozesses in weiterführende, makrostrukturelle Ablaufmodelle. Hierbei werden die einzelnen Phasen des gesamten Forschungsprozesses kategorisiert und anhand deren Spezifik näher erläutert. Ein erster Überblick über die grundlegende Makrostruktur empirischer Forschungsvorhaben findet sich unter anderem in Atteslander 2003; Schnell/Hill/Essner 2013; Poscheschnik 2015 oder Aeppli et al. 2016 sowie mit konkreten deutschdidaktischen Bezug in Boelmann 2016.

Die folgenden Ausführungen fokussieren daher die verschiedenen Phasen und Bereiche der Planung empirischer Erhebungen und legen Leitkriterien dar, welche in der Vorbereitungsphase beachtet werden müssen.

1. Grundlegende Konzeption der Planung und Vorbereitung empirischer Erhebungen

Die Vorbereitung und Planung empirischer Erhebungen gliedert sich in zwei Phasen, in welchen verschiedene Konzeptionen zur Spezifikation des Vorbereitungsprozesses erstellt werden.

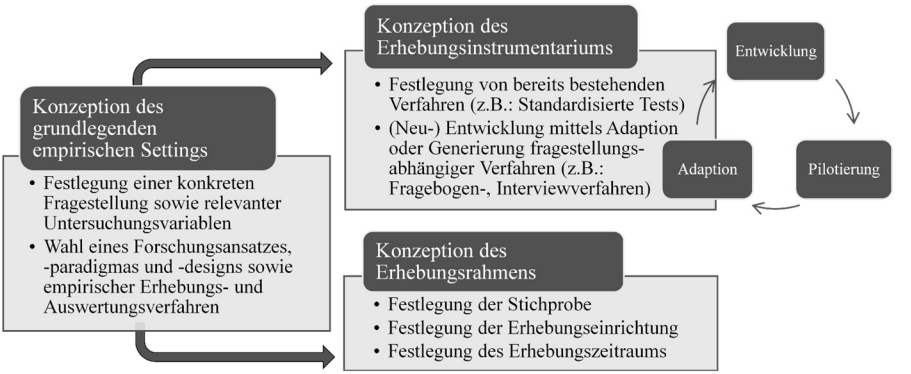

Abb. 1: Phasen der Planung und Vorbereitung empirischer Erhebungen

Die erste Phase der ‚Konzeption des grundlegenden empirischen Settings' beinhaltet Aspekte, die das empirische Forschungsdesign betreffen. Diese müssen im Vorfeld der konkreten durchführungspraktischen Planung festgelegt werden. Neben der Formulierung einer eindeutigen Fragestellung erfolgt hierbei sowohl die Bestimmung relevanter Untersuchungsvariablen als auch die Entscheidung hinsichtlich der anzuwendenden Forschungsmethodik. Die zweite Phase umfasst hingegen zwei Konzeptionsbereiche, welche beide erst im Anschluss an die ‚Konzeption des grundlegenden empirischen Settings' erfolgen können, sich jedoch mit fortschreitenden Forschungsprozess gegenseitig bedingen: Die ‚Konzeption des Erhebungsinstrumentariums' und die ‚Konzeption des Erhebungsrahmens', die dabei in enger Wechselbeziehung stehen.

2. Konkrete Ausgestaltung der Phasen des Vorbereitungsprozesses

Im Folgenden werden die einzelnen Konzeptionsbereiche der beiden Vorbereitungs- und Planungsphasen näher erläutert. Die Übertragung der beispielhaft vorgestellten Überlegungen auf das eigene Forschungsprojekt wird durch im folgenden explizierter Leitkriterien erleichtert, die das empirische Setting, das Erhebungsinstrumentarium sowie den Erhebungsrahmen auszeichnen.

2.1 Phase 1 – Konzeption des grundlegenden empirischen Settings

Bevor mit der konkreten Planung der empirischen Erhebung begonnen werden kann, müssen sich Forschende zunächst darüber im Klaren sein, wie das empirische Setting des Forschungsvorhabens aufgebaut sein soll. Das erste Ziel besteht daher in der Findung und Formulierung einer konkreten Fragestellung, welche im Rahmen des Forschungsvorhabens beantwortet werden soll.

Obgleich diese Festlegung eine der größten Herausforderungen eines jeden Forschungsprojektes darstellt, erweist sich dieser Schritt bereits innerhalb der ersten Planungsphase als unverzichtbar. Forschungsprojekte ohne eine festgelegte Fragestellung laufen sowohl innerhalb der Planungs- als auch der Durchführungsphase des Projekts Gefahr, weder ein konkretes Erhebungs- oder Auswertungsverfahren auswählen noch die generierten Daten aufgrund der mangelnden Perspektive aufbereiten und analysieren geschweige den interpretieren zu können.

Zu Beginn der Planung und Vorbereitung der empirischen Erhebung wird daher eine vorläufige Formulierung vorgenommen, welche als Arbeitsgrundlage für das weitere Vorgehen fungiert. Insbesondere die Planungs- und Vorbereitungsphase dient dazu, die Fragestellung zu konkretisieren, zu evaluieren oder auch zu revidieren und daraufhin das Forschungsvorgehen nochmals zu verändern. Die zunächst vorgenommene Formulierung der Fragestellung muss demnach innerhalb des Forschungsprozesses hinsichtlich deren Passung reflektiert und ggf. verändert werden.

Da die Fragestellung die Ausrichtung des Forschungsvorhabens jedoch erheblich beeinflusst, kann die Fortsetzung der Vorbereitung erst dann stattfinden, wenn eine vorläufige Fragestellung vorliegt (vgl. hierzu auch Häder 2010, 81ff.).

Im Anschluss an die Forschungsfrage erfolgt die Festlegung relevanter Untersuchungsvariablen, welche zur Beantwortung der Fragestellung beitragen können oder potenzielle Synergieeffekte erzeugen. Diese können dabei entweder biographische Daten (das Geschlecht, Alter, Herkunft etc.), sozialisatorische Aspekte (die Familienkonstellation, Lektüregewohnheiten etc.) oder auch andere Fähigkeiten der Probanden (Lesefähigkeit, Schreibfähigkeit, Literarische Kompetenz etc.) betreffen. Je nach Wahl der Untersuchungsvariablen kann der Einsatz von Begleittests innerhalb der empirischen Erhebung notwendig sein, welche gemeinsam mit dem Erhebungsinstrumentarium ausgewählt oder entwickelt werden müssen.

Beispiel: Verfolgt ein Forschungsprojekt beispielsweise die Fragestellung „Inwieweit beeinflusst das Leseförderkonzept der Anne-Frank-Gesamtschule die Lesefähigkeit von Schülerinnen und Schülern einer achten Klasse?“, könnte es interessant sein, das Geschlecht der betreffenden Kinder oder auch deren literarische Sozialisationsumgebung mit zu erheben. Hieraus ließen sich potenzielle Einflussfaktoren ableiten, welche Aussagen über die spezifischen Erfolge des Förderkonzepts bei bestimmten Schülergruppen ermöglichen. Andere Faktoren, die nicht auf den schulischen Förderunterricht zurückzuführen sind, etwa ein verstärktes Lese-Engagements zuhause, lassen sich durch gezielte Erhebung ihres Einflusses berücksichtigen.

Erst wenn die Fragestellung und die Untersuchungsvariablen feststehen, kann die Konkretisierung des Forschungsdesigns sowie des methodischen Vorgehens stattfinden. Hierbei müssen sowohl das Forschungsparadigma (qualitativ, quantitativ oder Mixed-Methods), der Forschungsansatz (Grundlagenforschung, Design Research etc.) sowie das Untersuchungsdesign (Einmalerhebung, Pre-Post etc.) festgelegt werden, bevor anschließend die Auswahl der Erhebungs- und Auswertungsmethoden stattfindet.[2] Sowohl die Wahl des empirischen Forschungsdesigns als auch die Festlegung des forschungsmethodischen Vorgehens orientiert sich dabei an zwei übergeordneten Leitkriterien, welche ebenfalls die ‚Konzeption des Erhebungsinstrumentariums' und den konkreten Erhebungsrahmen bestimmen:

1. Leitkriterium „Ziel- und Frageorientierung": Die Ausrichtung des empirischen Settings muss zur Beantwortung der Fragestellung beitragen!

Auch wenn das erste Leitkriterium zunächst banal klingen mag, stellt dieses eine der zentralen Gelingensbedingungen empirischer Erhebungen dar. Die Wahl des Forschungsparadigmas, Forschungsansatzes, Untersuchungsdesigns sowie der Erhebungs- und Auswertungsmethoden muss sich an der zuvor festgelegten Fragestellung orientieren, sodass auch tatsächlich ein empirisches Setting konzipiert wird, welches zu ihrer Beantwortung beiträgt: Soll die Fragstellung des Forschungsvorhabens beispielsweise durch die Entwicklung eines induktiv generierten Kategoriensystems beantwortet werden, erfordert dies die Entwicklung eines Settings, in welchem die Erhebung qualitativer Daten ermöglicht wird. Zielsetzungen, welche hingegen die Wirksamkeit spezifischer Förderinstrumente untersuchen, verlangen in der Regel Konzeptionen von Prä-Post-Formaten, da ein Untersuchungsdesign, das lediglich eine Einmalerhebung vorsieht, Veränderungsprozesse bei Probanden nicht erfassen kann. Erfordert die Beantwortung der Forschungsfrage allerdings eine große Fallzahl, wie beispielsweise bei einem Vergleich von länderweiten Schülerleistungen, muss dem auch die Erhebungsmethode Rechnung tragen. Obgleich der Einsatz von qualitativen Leitfadeninterviews womöglich interessante Ergebnisse hervorbringen würde, können diese unter Berücksichtigung der Fragestellung nicht ohne ein großes Forscherteam oder einen sehr hohen zeitlichen Aufwand umgesetzt werden. Ohne diese Ressourcen bedarf es einer alternativen Methode, welche der Zielsetzung des Vorhabens adäquater entspricht.

Aufgrund dessen muss bei der ‚Konzeption des grundlegenden empirischen Settings' die Wahl des Forschungsdesigns stets vor dem Hintergrund der Zielsetzung des Forschungsprojekts reflektiert werden.

---

[2] Einen einführenden Überblick über verschiedene empirische Erhebungs- und Auswertungsmethoden im deutschdidaktischen Kontext bietet der zweite Band *Erhebungs- und Auswertungsverfahren* dieser Reihe.

2. Leitkriterium „Intentionalität“: Innerhalb des empirischen Settings werden nur Daten erhoben, welche relevant für das Forschungsvorhaben sind!

Insbesondere bei der konkreten Planung der empirischen Erhebungsverfahren stellt das zweite Leitkriterium eine unverzichtbare Größe dar. Innerhalb des empirischen Setting dürfen nur Daten erhoben werden, welche für das Forschungsvorhaben einen bestimmten Zweck erfüllen (siehe den Beitrag von Iberer in diesem Band). Berücksichtigt man dieses Kriterium, erhöht sich die Qualität der empirischen Erhebung in zwei Bereichen: Zum einen dient es als Fokussierungshilfe, da so unnötige und nicht mehr zu bewältigenden Datenberge vermieden werden. Zum anderen hilft dieses Kriterium dabei, die Zeit der empirischen Erhebung für die Gewinnung der relevanten Daten zu nutzen. Bei zu langen Erhebungsphasen kann es passieren, dass insbesondere bei jungen Probanden im Verlauf der Erhebung die Konzentration nachlässt, was wiederum die Güte der später erhobenen Daten beeinträchtigen kann. Dies lässt sich mit einer fokussierten Erhebungsauswahl vermeiden. Wenngleich die Lust am Forschen groß sein kann, muss die Notwendigkeit jedes einzelnen Erhebungsinstruments vor der Durchführung hinterfragt und irrelevante Testverfahren gestrichen werden.

Im Sinne der Leitkriterien bestimmen folgende Fragen die ‚Konzeption des grundlegenden empirischen Settings‘:

- Hilft das gewählte empirische Setting bei der Beantwortung der Fragestellung?
- Verfolgen die Verfahren des empirischen Settings innerhalb des Forschungsvorhabens einen *konkreten* Zweck?

Im Anschluss an die Festlegung des Forschungsdesigns sowie des methodischen Vorgehens, wird das erarbeitete empirische Setting unter Berücksichtigung möglicher Störvariablen reflektiert. Hierunter versteht man alle Einflussgrößen, die auf die Untersuchungsvariablen – insbesondere die abhängige Variable – einwirken und eindeutige Schlussfolgerungen verhindern (vgl. u.a. Döring/Bortz 2016). Mithilfe einer frühzeitigen Reflektion potenzieller Störvariablen lassen sich Konstruktionsfehler des empirischen Settings bereits in der ersten Phase der Planung und Vorbereitung aufdecken und beseitigen, weshalb eine Einbettung dieses Schritts im Anschluss an eine erste Konzeption des empirischen Settings erfolgen muss. In der anschließenden Konzeption des Erhebungsinstrumentariums und des Erhebungsrahmens sowie der Durchführung der Erhebung selbst können dadurch vorhersehbare Fehler vermieden werden. Diese können sich dabei sowohl auf das gesamte empirische Konzept, einzelne Erhebungsmethoden oder zu berücksichtigende Rahmenbedingungen beziehen und lassen sich nach Versuchspersonenmerkmalen, Situationsmerkmalen und Versuchsleitermerkmalen klassifizieren (vgl. u.a. Hussy/Schreier/Echterhoff 2010, 115f.): Die Versuchspersonenmerkmale umfassen zum einen probandenspezifische Eigenschaften wie das Alter, Geschlecht, Migrationshintergrund oder Förderbedarf, aber auch zum anderen subjektbezogene Fähigkeiten, welche sich erst in Zusammenhang mit der Untersuchungsvariable als potenziell beeinflussend zeigen können. Beispiele hierfür sind unter anderem die Konzentrations- oder Problemlösefähigkeit sowie fachliche

Fertigkeiten, wie die Lese-, Schreib- oder literarische Kompetenz. Situationsmerkmale beziehen sich hingegen auf den Untersuchungskontext, sodass sich die unterschiedliche Konzeption des Erhebungsrahmens beispielsweise durch differente Tageszeiten, die Beleuchtung im Erhebungsraum aber auch das Untersuchungsmaterial störend auf die Ergebnisse auswirkt. Weiterhin kann der Versuchsleiter eine Störvariable für die Ergebnisse darstellen. Werden beispielsweise unterschiedliche Versuchsleiter bei zwei Erhebungsgruppen eingesetzt, kann deren Alter, Geschlecht, aber auch deren Verhaltensweisen störende Auswirkungen nach sich ziehen (vgl. ebd.). Anhand der gezielten Auseinandersetzung mit potenziell störenden Einflussfaktoren auf die Konzeption des empirischen Settings oder der konkreten empirische Erhebung können mögliche Konstruktionsfehler vor der eigentlichen Durchführung bereits frühzeitig erkannt oder spezifische Herausforderungen der Erhebung verdeutlich werden. Die meisten Störvariablen lassen sich durch die Konzeption des Erhebungsinstrumentariums und der Anpassung des Erhebungsrahmens kontrollieren.[3] Hierbei gelten entweder die Konstanthaltung der Erhebungsbedingungen über mehrere Erhebungen, die Eliminierung der Störvariable oder auch deren gezielte Variation durch den Einsatz von Kontrollgruppen oder einer gezielten Probandenauswahl als potenzielle Kontrolltechniken (vgl. ebd.).

Beispiel: Das Forschungsdesign sieht eine Erhebung von zwei verschiedenen Probandengruppen an zwei unterschiedlichen Tagen vor. Eine potenzielle Störvariable könnte daher der unterschiedliche Erhebungszeitpunkt sein, welcher sich auf die gezeigte Leistung der Probanden auswirkt. Diese Störvariable lässt sich jedoch leicht mithilfe der ‚Konstanthaltung' kontrollieren, indem beide Erhebungen zur selben Uhrzeit durchgeführt werden. Der Einfluss der Tageszeit wäre damit bei beiden Gruppen gleich.

2.2 Phase 2 – Erhebungsinstrumentarium und Erhebungsrahmen

Im Anschluss an die ‚Konzeption des grundlegenden empirischen Settings' finden in der zweiten Phase sowohl die Festlegung des zu verwendenden Instrumentariums als auch die Aufstellung der Rahmenbedingungen statt. Diese orientieren sich ebenfalls an den übergeordneten Leitkriterien ‚Ziel- und Frageorientierung' sowie ‚Intentionalität'. Da sich beide Konzeptionen innerhalb der Erhebungsplanung gegenseitig bedingen, wirken sich Änderungen an der einen Konzeption in der Regel auf die andere aus, was bei der Planung zu berücksichtigt werden muss.

[3] An dieser Stelle muss darauf hingewiesen werden, dass sich nicht alle potenziellen Störvariablen im Vorfeld als kontrollierbar herausstellen. Hierzu zählen insbesondere die Versuchspersonenmerkmale wie beispielsweise das individuelle Verhalten der befragten Probanden, welche auf die Erhebungssituation in andere Art und Weise reagieren, als dies zu vermuten war. Die erhobenen Ergebnisse müssen daher unter Berücksichtigung potenzieller Störvariablen analysiert hinsichtlich der Aussagekraft der Erhebung reflektiert werden.

2.2.1 Konzeption des Erhebungsinstrumentariums

Für die ‚Konzeption des Erhebungsinstrumentariums' existieren zwei mögliche Vorgehensweisen zur Verfügung, welche je nach den Spezifikationen des Forschungsvorhabens auch kombiniert verwendet werden können. Zum einen besteht die Möglichkeit, auf bereits erprobte, standardisierte Verfahren zurückzugreifen, oder zum anderen für spezifischere Fragestellungen die Verfahren selbst zu entwickeln und diese anschließend für die Datengenerierung zu verwenden. Hierbei wird zwischen der Anwendung des Erhebungsinstruments als Haupt- oder als Begleittest unterschieden. Haupttests dienen der Erhebung von Merkmalen oder Fähigkeiten der Probanden, welche explizit durch das Forschungsprojekt analysiert werden sollen. Diese Untersuchungsvariablen werden als ‚Abhängige Variablen' bezeichnet. Im Rahmen von Begleittests erfolgt hingegen die Erhebung nebenstehender Aspekte, die potenzielle Einflussfaktoren für die Ausprägung der abhängigen Variable darstellen. Sie werden als ‚Unabhängige Variable' benannt. Jede empirische Erhebung umfasst zumeist einen Korpus verschiedener Erhebungsinstrumentarien, welche entweder als Haupt- oder als Begleittest eingesetzt und sowohl erprobte aber auch selbstentwickelte Testverfahren umfassen können.

Verwendung erprobter Testverfahren

Die Verwendung bereits erprobter Verfahren bietet sich insbesondere bei Forschungsfragen an, welche die Auswirkungen einer bestimmten Fähigkeit auf anderweitige Leistungen des Probanden oder Entwicklungstendenzen untersuchen. Sie lassen sich auch in Kombination mit selbstentwickelten, auf die Fragestellung des Forschungsprojekts explizit zugeschnittenen Erhebungsinstrumenten kombinieren. Die Vorteile bereits erprobter oder standardisiert festgelegter Verfahren zeigen sich dabei in der Einhaltung der Gütekriterien forschungsmethodischen Vorgehens, sodass diese meist an großen Stichproben normiert wurden und dabei valide, reliabel und objektiv den zu erhebenden Faktor messen. Aus diesem Grund können diese Verfahren auch als Begleittests genutzt werden, die einzelne Untersuchungsvariablen erheben oder potenzielle Störvariablen kontrollieren (vgl. Döring/Bortz 2016). Steht für die Erhebung einer Untersuchungsvariable ein erprobtes oder standardisiertes Verfahren zur Verfügung, sollte das erprobte Testverfahren dem selbstentwickelten Instrument vorgezogen werden. Wenngleich ein neuentwickelter Lesekompetenztest beispielsweise auf die spezifischen Anforderungen der Probandengruppe oder des Projekts zugeschnitten werden kann, steht dieser in den meisten Fällen hinter den normierten Tests zurück: Da die Einhaltung der Gütekriterien empirischer Forschung bei der Testkonstruktion neuentwickelter Erhebungsinstrumente (siehe die Beiträge von Schmidt in diesem Band) einen erheblichen Aufwand darstellt, kann sie in der Regel nicht im selben Umfang wie bei etablierten standardisierten Verfahren gewährleistet werden. Folglich verfügten die Ergebnisse bei der Erhebung mittels selbstentwickelter Tests über eine geringere Aussagekraft.

Beispiel: Untersucht das Forschungsvorhaben den Erfolg einer Interventionsstudie, welche die Verbesserung der Lesegeschwindigkeit von Schülerinnen und Schülern der Mittelstufe fokussiert, muss das hierfür eingesetzte Instrumentarium zur Messung der Lesegeschwindigkeit nicht neu entwickelt werden, da eine Reihe von standardisierten Tests für diverse Altersklassen in normierter Qualität vorliegen, welche zudem zeitökonomisch eingesetzt werden können.

Einen Überblick über vielfältige, standardisierte Erhebungsinstrumente, welche sich für den Einsatz in empirischen Erhebungen eignen, bietet beispielsweise die Testzentrale von Hogrefe (www.testzentrale.de). Darüber hinaus bieten angrenzende Wissenschaftsdisziplinen wie die Pädagogik, Entwicklungs- oder auch Lernpsychologie forschungsmethodische Verfahren, welche insbesondere für den Einsatz in Begleittests geeignet sind (beispielsweise Erhebung der Konzentration durch den standardisierten Test ‚d2' oder Feststellung der Kreativitätsleistung durch die „Mind-Wandering Methode" (vgl. Mooneyham/Schooler 2013)).

Neuentwicklung von Erhebungsinstrumenten

Die zweite Möglichkeit der ‚Konzeption des Erhebungsinstrumentariums' in Form einer (Neu-)Entwicklung fragestellungsspezifischer Verfahren wird hingegen dann notwendig, wenn bislang keine zur Fragestellung passenden Verfahren für die Erhebung vorliegen. Selbstentwickelte Verfahren lassen sich in Folge exakt auf die leitende Fragestellung des Forschungsvorhabens, die Untersuchungsvariablen oder auch die Spezifika der Stichprobe zuschneiden. Hierzu können entweder das Verfahren der konkreten Durchführung vollständig neu entwickelt oder auf bereits veröffentlichte Studien als Orientierungspunkte für die eigene Adaption zurückgegriffen werden. Auch für die Entwicklung von Begleittests, wie zum Beispiel bei der Konzeption eines Fragebogens zur Erfassung biographischer Daten (siehe hierzu den Beitrag von Maak in diesem Band), stellt sich eine an die Erhebungssituation angepasste Erstellung des Instruments als sinnvoll heraus.

Beispiel: Verfolgt das Forschungsvorhaben beispielsweise das Ziel, ein Projekt des städtischen Jugendhauses mithilfe einer Fragebogenerhebung zu evaluieren, ist es unwahrscheinlich, dass zu explizit diesem Angebot in derselben Gemeinde mit der identischen Probandengruppe bereits eine empirische Erhebung durchgeführt wurde, auch wenn bereits ähnliche Erhebungen zu anderweitigen Angeboten oder auch in Nachbargemeinden durchgeführt worden sein könnten. Demnach müssen die Items der Befragung selbst entwickelt und hinsichtlich des konkreten Angebots spezifiziert werden. Nur durch die Selbstentwicklung der Items werden innerhalb der Erhebung Aussagen generiert, welche sich auch auf die explizite Situation beziehen. Für das Sammeln erster Ideen kann es daher sinnvoll sein, sich die Spezifika des empirischen Settings vor Augen zu führen oder für Denkanstöße ähnliche Studien für die Konzeption des Erhebungsinstrumentariums zu Rate zu ziehen.

Unabhängig von der Wahl des Vorgehens zur Festlegung des Erhebungsinstrumentariums prägen zwei Kriterien die Konzeptionsphase des verwendeten Korpus: die Notwendigkeit der Pilotierung und der Einsatz personeller Ressourcen.

Kriterium 1: Pilotierung des Instrumentariums

Die Pilotierung des Erhebungsinstrumentariums stellt eine unverzichtbare Phase des Vorbereitungsprozesses dar. Innerhalb dieses Arbeitsschritts wird das bis zu diesem Punkt geplante empirische Setting und insbesondere das Erhebungsverfahren an einer kleinen Stichprobe, welche nach Möglichkeit 10% der Gesamtstichprobe der folgenden Haupterhebung umfasst, ausprobiert (vgl. u.a. Poscheschnik 2015; Döring/Bortz 2016). Pilotierungen verfolgen dabei das Ziel, die Funktionalität des empirischen Settings sowie der verwendeten Instrumentarien zu überprüfen, potenzielle Fehler in der bisherigen Konzeption zu entdecken und noch vor der eigentlichen Haupterhebung zu korrigieren. Außerdem zeigen sich zumeist weiterführende Faktoren, welche einen Einfluss auf die Untersuchungsvariablen ausüben, jedoch bislang innerhalb des Forschungsvorhabens unberücksichtigt blieben. Daher gilt, dass jede Festlegung und Entwicklung des Erhebungsinstrumentariums nach einer Pilotierungsphase verlangt und deren Ergebnisse zu einer Veränderung oder Adaption des bisher bestehenden Vorgehens führen können (vgl. Hussy/Schreier/Echterhoff 2010). Die Potenziale, welche sich durch die Durchführung und Auswertung einer Pilotierung ergeben, lassen sich dabei in drei Reflexionskategorien systematisieren:

Die erste Kategorie „Personenspezifische Ergebnisse“ umfasst Aspekte, welche den Versuchsleiter betreffen. So ermöglichen Pilotierungen, das eigene Frageverhalten zu reflektieren und auf potenzielle Störfaktoren bezüglich der Versuchsleitermerkmale hin zu untersuchen. Beispielsweise könnte es sein, dass der Versuchsleiter durch sein Verhalten die Probanden in ihren Äußerungen beeinflusst, indem er zu viele oder zu wenige Informationen preisgibt, selbst durch seine Unsicherheit oder Nervosität die Aussagen der Befragten verfälscht oder aufgrund der Stimmlage oder Wortwahl die Versuchspersonen verunsichert. Die Pilotierung der empirischen Erhebung bietet damit einen Rahmen, sich selbst als Versuchsleiter zu erproben, ein möglichst erhebungsförderliches Verhalten zu entwickeln sowie erste Erfahrungen in der Handhabung des Erhebungs- und Auswertungsverfahrens zu sammeln.

Die zweite Kategorie „Instrumentariumsspezifische Ergebnisse“ ermöglicht hingegen Aussagen bezüglich der Eignung des Erhebungsinstrumentariums für das konkrete empirische Setting. Im Rahmen der Pilotierung kann sich zum Beispiel erweisen, ob ein selbstentwickeltes Item tatsächlich Aussagen zu einer bestimmten Untersuchungsvariable erzeugt, oder ob dieses nochmals verändert oder ausgetauscht werden muss. Darüber hinaus lassen sich Erkenntnisse über den Schwierigkeitsgrad des verwendeten Verfahrens ableiten und potenzielle Verstehensprobleme beheben. Hierbei können die notwendigen Veränderungen einzelne Formulierungen, ganze Items oder das gesamte gewählte Verfahren betreffen. Wird zum Beispiel ein Wort innerhalb eines Items verwendet, welches einige der Probanden der Pilotierungsstichprobe nicht kennen oder das einer weiteren Erklärung bedarf, sollte dieses ausgetauscht oder eine Erklärungsphase für die entspre-

chende Fragestellung eingeplant werden. Auch zeigt sich in der Pilotierungsphase, ob das für das Instrumentarium eingeschätzte Zeitlimit eingehalten oder überschritten wird.

Die dritte und letzte Reflexionskategorie beinhaltet „Fragestellungsorientierte Ergebnisse“. Die Pilotierung des Erhebungsinstrumentariums ermöglicht erste Rückschlüsse über die Fragestellung und Hypothesen des Forschungsvorhabens oder führt zu einer Veränderung sowie Weiterentwicklung der Zielsetzung. Obgleich sich aufgrund der kleinen Stichprobe der Pilotierung noch keine bestätigenden oder widerlegenden Aussagen der Ergebnisse ableiten lassen, zeigt die Auswertung erste Tendenzen auf, welche bei der finalen Vorbereitung der empirischen Erhebung miteinbezogen werden müssen. Des Weiteren können sich Variablen als relevant erweisen, welche in der bisherigen Konzeption des empirischen Settings unberücksichtigt blieben. Im Anschluss an die Pilotierung kann es daher nötig sein, weitere Untersuchungsvariablen zu definieren und in das Erhebungsinstrumentarium zu integrieren, wobei stets abgewogen werden muss, ob eine Integration eine qualitative Erweiterung der Beantwortung der leitenden Fragestellung darstellt. Ergeben sich weitreichende Veränderungen bezüglich der Fragestellung oder der Zielsetzung des Forschungsprojekts anhand der Pilotierungsergebnisse, müssen die vorgenommenen Anpassungen nochmals pilotiert und ausgewertet werden.

Kriterium 2: Aufwand personeller Ressourcen

Nicht alle Erhebungen können durch den Projektleiter alleine durchgeführt werden. Eine weitere zu berücksichtigende Größe der ‚Konzeption des Erhebungsinstrumentariums‘ stellt somit der Einsatz der personellen Ressourcen dar. Hierbei muss die Reflektion des festgelegten Erhebungsverfahrens hinsichtlich der für die Durchführung notwendigen beteiligten Personen erfolgen. Erfordert das empirische Setting beispielsweise die Befragung durch mehrere Versuchsleiter muss gewährleistet sein, dass diese über dieselben Informationen und Fähigkeiten bezüglich des Erhebungsverhaltens verfügen. Hierzu erweist es sich als notwendig, die betreffenden Personen im Vorfeld in einem Umfang zu schulen, dass sie sowohl die Intentionen des Forschungsprojekts kennen, das Erhebungsverfahren mehrfach üben konnten, auf potenzielle Schwierigkeiten innerhalb der Erhebungssituation vorbereitet sind und ein einheitliches Erhebungsverhalten festgelegt und erprobt werden konnte. Nur durch eine ausgeprägte Schulung der Versuchsleiter wird anschließend eine Vergleichbarkeit der Befragungsergebnisse erreicht (vgl. Settinieri 2014).

Bereits in dieser frühen Planungsphase sollte der Blick auch auf die Auswertung gerichtet werden: Selbst wenn die Erhebung von einem Versuchsleiter alleine erfolgen kann, sollte – abhängig von den gewählten Erhebungs- und Auswertungsverfahren – dieser nach Möglichkeit nicht die Auswertung der Ergebnisse übernehmen, da anderenfalls die Objektivität des Forschungsprozesses beeinträchtigt sein könnte.

2.2.2 Konzeption des Erhebungsrahmens

Parallel zu der Entwicklung des Erhebungsinstrumentariums findet in der zweiten Phase der Vorbereitung der empirischen Erhebung die ‚Konzeption des Erhebungsrahmens' statt. Die Rahmenbedingungen empirischer Erhebung setzen sich dabei aus drei Faktoren zusammen, welche die konkrete Durchführung unter Berücksichtigung der leitenden Fragestellung bestimmen und spezifizieren.

Festlegung der Stichprobe

Die Auswahl der Stichprobe stellt eine der zentralen Vorbereitungsaspekte empirischer Erhebungen dar. Eine jede Stichprobe besteht aus einer bestimmten Anzahl von Probanden, welche nach spezifischen Kriterien ausgewählt wurden und nach Möglichkeit repräsentativ für die Grundgesamtheit oder Gesamtpopulation steht. Die Festlegung der Probanden orientiert sich hierbei an der konkreten Fragestellung sowie dem Design des Forschungsvorhabens, sodass je nach empirischen Setting die Zusammensetzen der Stichprobe stark variieren kann. Das Untersuchungsdesign bestimmt dabei, ob für die empirische Erhebung die Befragung eine Probandengruppe genügt, ob ein Vergleich zwischen zwei Gruppen oder sogar die Integration einer Kontrollgruppe in das empirische Setting notwendig erscheint.

Nicht nur die Entscheidung über die Größe der Stichprobe, auch deren konkrete Zusammensetzung erfordert reflektierte Entscheidungen, sodass bei der Auswahl von Stichproben zwei mögliche Vorgehensweisen vorliegen: die ‚Zufallsauswahl' und die ‚bewusste Auswahl'. Die ‚zufällige Auswahl' einer Probandengruppe wird zumeist in quantitativen Forschungsvorhaben eingesetzt. Hierbei wird die Stichprobe nicht nach den abhängigen oder unabhängigen Variablen des Forschungsprojekts ausgewählt, sondern orientiert sich an den Kriterien der Repräsentativität und Wahrscheinlichkeit. Im Rahmen einer Zufallsauswahl besteht für jede Variable eine gleich große Wahrscheinlichkeit in die Stichprobe aufgenommen zu werden und damit repräsentativ für einen Teil der Gesamtpopulation zu stehen. Oftmals erfolgt die Zusammensetzung anhand von ‚Einfachen Zufallsstichproben', in welchen einzelne Probanden oder Probandengruppen unabhängig voneinander ausgewählt werden. Das Ziel der ‚zufälligen Auswahl' besteht dabei in der Übertragbarkeit der Forschungsergebnisse auf die Grundgesamtheit der Population, sodass sich nicht nur Schlussfolgerungen bezüglich einer einzelnen Probandengruppe ergeben, sondern die gewonnenen Erkenntnisse bei anderen Gruppen angewandt werden können. Da Schlussfolgerungen bezüglich der Tragweite von Forschungsergebnissen jedoch ebenso von der Größe der jeweiligen Stichproben abhängen, zeigen sich die Erkenntnisse innerhalb zufällig ausgewählter Stichproben aufgrund einer ausgewogenen Merkmalsadäquanz zwar als potenziell übertragbar, vor dem Hintergrund der in studentischen Arbeiten zumeist geringen Probandenanzahl allerdings als ebenso ausschnitthaft, wie es eine gezielte

Auswahl der Stichprobe anhand von Probandenmerkmalen erlaubt (vgl. Kaya/Himme 2007, 78).[4]

Die ‚bewusste Auswahl' einer Stichprobe folgt hingegen einer zuvor ausgewählten Variable – wie beispielsweise dem Geschlecht, dem Alter, der sozialen Herkunft oder der Teilnahme an gezielten Fördermaßnahmen –, welche entweder einen potenziellen Einflussfaktor auf die zu untersuchende Leistung, Fähigkeit oder Fertigkeit der Probanden darstellt oder selbst die fragestellungsbestimmenden Faktoren abbildet. Aufgrund der Auswahl einer spezifischen Probandengruppe bieten die Erkenntnisse des Forschungsprojekts einen ausschnitthaften Einblick, welche Hinweise auf Ergebnisse der Grundgesamtheit ermöglichen, jedoch nicht abschließend bestätigen (ebd.).

Beispiel: Sieht das Forschungsdesign qualitative Einzelinterviews von Lehrkräften bezüglich ihres unterrichtlichen Einsatzes digitaler Medien vor, kann es eine Möglichkeit sein, sowohl Lehrerinnen und Lehrer zu befragen, welche aufgrund der institutionellen Rahmenbedingungen auf ein umfangreiches digitales Angebot zurückgreifen können, als auch diejenigen, welche ausschließlich auf Medienzentren und deren eigene Ausstattung angewiesen sind. In diesem Fall liegen zwei Probandengruppen vor, welche gemeinsam die Stichprobe bilden.

Je nach Spezifikation des Forschungsvorhabens spielt bei der Zusammensetzung der Stichprobe die ‚Ähnlichkeit der Probanden(gruppen)' oder die ‚Varianz der Probanden(gruppen)' eine ausgeprägtere Rolle.

Bei der Auswahl nach Ähnlichkeitskriterien steht die Ausprägung der zu untersuchenden abhängigen Variable im Zentrum. Untersucht ein Forschungsprojekt beispielsweise die Aufgabenschwierigkeit von textgebundenen Items zu einer Schulbuchlektüre, erhalten zwei in ihrer Zusammensetzung ähnliche Probandengruppen unterschiedliche Items, sodass sich die Unterschiede der Messergebnisse zwischen beiden Gruppen vor allem auf die abhängige Variable, die Aufgabenschwierigkeit, zurückführen lassen.

Hierzu werden ähnliche Probanden(gruppen) ausgewählt, welche bei vielen Kriterien (beispielsweise bei der Altersstruktur, institutioneller Rahmen etc.) Übereinstimmungen aufweisen. Die Verwendung sich ähnelnder Probanden(gruppen) bietet dabei das Potenzial, die erhobenen Leistungen hinsichtlich auftretender Unterschiede zu analysieren und die Ausprägung der vorhabensbestimmenden Faktoren eingehender zu untersuchen. Die Ergebnisse könnten beispielsweise zeigen, dass die befragten Kinder, trotz der ähnlichen Zusammensetzung von zwei Probandengruppen, unterschiedliche Ergebnisse erzielen. Die Gründe für diese Unterschiede müssen anschließend reflektiert werden. Weiterhin wird die Auswahl nach ‚Ähnlichkeit' zumeist bei der Konzeption von Kontroll- und Interventionsgruppen eingesetzt.

[4] Eine Übersicht spezifischer Auswahlverfahren zur Generierung einer Zufallsstichprobe, wie beispielsweise von einfachen, geschichteten oder mehrstufigen (Klumpen-)Stichproben, findet sich unter anderem in Hussy/Schreier/Echterhoff 2010 oder auch Döring/Bortz 2016.

Beispiel: Soll innerhalb des Forschungsvorhabens der erleichterte Textzugang anhand visueller Unterstützungssysteme untersucht werden, durchläuft eine der Probandengruppen (Interventionsgruppe) ein empirisches Setting, in welchem neben dem textuellen Gegenstand zusätzlich Bildimpulse zur Verfügung gestellt werden. Eine zweite Probandengruppe (Kontrollgruppe) erhält ausschließlich den Text. Um auftretende Leistungsunterschiede anschließend auf die Darbietung des Unterstützungssystems zurückführen zu können, müssen die Probanden bei entscheidenden Merkmalen identische Eigenschaften aufweisen, sodass diese als Einflussfaktor ausgeschlossen werden können.

Bei der Berücksichtigung der ‚Varianz' werden hingegen die Einflussfaktoren der jeweiligen Forschungsergebnisse fokussiert. Die Zusammensetzung der Probandengruppe(n) wird so gewählt, dass sie sich hinsichtlich bestimmter Untersuchungsvariablen stark voneinander unterscheiden. Die Varianz innerhalb einer oder auch zwischen verschiedenen Probandengruppen ermöglicht anschließend die differenten Ergebnisse auf bestimmte Ausprägungsmerkmale zurückzuführen und Aussagen über den Einfluss spezifischer Faktoren zu ermöglichen. Auch kann die Stichprobe auch gezielt variiert werden, um mögliche Störvariablen auszuschließen.

Beispiel: Fokussiert das Forschungsprojekt den Einfluss familiärer literarischer Sozialisation auf die literarische Kompetenz von Schülerinnen und Schülern einer ersten Klasse, erfordert die Beantwortung der Fragestellung möglichst unterschiedliche Probandengruppen. Diese sollten sich bezüglich ihrer familiären literarischen Sozialisation stark voneinander unterscheiden, um Aussagen über den angenommenen Einflussfaktor zu ermöglichen.

Abhängig von der Spezifikation des Forschungsprojekts kann es sinnvoll sein, bezüglich eines bestimmten Merkmals ähnliche oder unterschiedliche Probandengruppen in die Erhebung aufzunehmen. In den meisten Forschungsprojekten ähneln sich die Probanden(gruppen) bei einigen Ausprägungsmerkmalen, während sie sich hinsichtlich einer oder mehrerer konkreten Untersuchungsvariablen explizit voneinander unterscheiden. Sowohl die ‚Vergleichbarkeit' als auch die ‚Varianz' stellt damit eines der zentralen Auswahlkriterien bei der Zusammensetzung von Stichproben dar.

Festlegung der Erhebungseinrichtung

Die Festlegung der Stichprobe steht in enger Verbindung mit der Auswahl der Erhebungseinrichtung. Hierbei bestimmt die Fragestellung des Forschungsvorhabens, ob zuerst die Stichprobe oder die Erhebungseinrichtung ausgewählt werden muss: Sofern Probanden und deren Leistungen im Zentrum der Erhebung stehen, orientiert sich die Auswahl der Erhebungseinrichtung anhand der für die Fragestellung notwendigen Stichprobe. Soll innerhalb eines Forschungsvorhabens beispielsweise die Ausprägung der Zuhörkompetenz von Schülerinnen und Schüler einer fünften Klasse in Folge einer explizit durchgeführten Unterrichtseinheit untersucht werden, erfolgt angesichts der Forschungsintention zunächst die Wahl

einer Probandengruppe, welche das konkrete Thema bereits bearbeitet hat. Aufgrund dieser Festlegung ergibt sich im Anschluss die Wahl der passenden Einrichtung, in welcher ein solches Programm durchgeführt wurde.

Ausschließlich in Evaluationsstudien kann die Planung auch von der Festlegung der spezifischen Erhebungseinrichtung aus vorgenommen werden, wobei im Anschluss an die Einrichtungsauswahl eine spezifische Stichprobe bestimmt werden muss. Beispiele hierfür finden sich insbesondere bei der Evaluation schulischer Entwicklungsprojekte, in welchen die Einrichtung oder spezifische Merkmale dieser selbst den Forschungsgegenstand darstellen. Bezüglich der Auswahl der Erhebungseinrichtung liegen jedoch vielfältige Kriterien vor, die ein Spektrum dessen aufzeigen, welche Einrichtungen für empirische Erhebungen in Frage kommen, oder auch Einflussfaktoren darstellen, welche sowohl die Erhebung selbst als auch die Kontextualisierung der Ergebnisse stark beeinflussen:

Kriterium 1: Institutionalität

Abhängig von der Fragestellung des Forschungsvorhabens erfolgt zu Beginn die Wahl der Erhebungseinrichtung anhand des Kriteriums der ‚Institutionalität'. Die Durchführung empirischer Erhebungen erfolgt entweder innerhalb eines institutionellen oder non-institutionellen Rahmens, wobei die verschiedenen Orte spezifischen Anforderungsstrukturen und kontextbezogenen Rahmenbedingungen unterliegen. Empirische Erhebungen können hierbei entweder in institutionellen Bildungseinrichtungen – wie Schulen, Kindergärten oder Kitas – oder in außerschulischen Angeboten – beispielsweise Jugendhäusern, Seniorenheimen, Vereinen, Freizeiteinrichtungen, aber auch Kinderspielgruppen – durchgeführt werden. Unabhängig von dem institutionellen Charakter der Erhebungseinrichtung müssen explizite Rahmenbedingungen, verbindliche Kommunikationswege und konkrete Datenschutzkonzepte eingehalten und entwickelt werden (siehe hierzu den Beitrag von Iberer in diesem Band). Insbesondere in institutionellen Bildungseinrichtungen gelten Vorgaben, welche durch Regelungen höherer Instanzen – wie durch Vorgaben des Schulbezirks, des Landkreises, der Regierungspräsidien oder auch der Bildungsministerien der verschiedenen Bundesländer – bestehen und demnach unbedingt in die Planung und Vorbereitung der empirischen Erhebung miteinbezogen werden müssen. Aufgrund dessen ist es wichtig, sich vorzeitig über bestehende Regelungen des ausgewählten Erhebungsortes zu informieren und die Realisierbarkeit der Datenerhebung sowohl rechtlich als auch pragmatisch zu überprüfen.

Kriterium 2: Merkmalsspezifika

Weiterhin kann die Wahl der Erhebungseinrichtung anhand spezifischer Merkmale erfolgen, welche die relevanten Untersuchungsvariablen des Forschungsvorhabens berücksichtigen. Neben der von den Untersuchungsvariablen abhängigen Probandenzusammenstellung (beispielsweise monoedukative Sekundarschulen, Förderzentren etc.) kann die Wahl des Ortes beispielsweise angesichts des vorliegenden Profils der jeweiligen Einrichtung vorgenommen werden, welches

sich als relevant für die Beantwortung der leitenden Fragestellung erweist. Verfügt zum Beispiel ein Kindergarten über ein explizites Programm zur Förderung früher Literalitätserfahrungen von Kleinkindern, kann dieses Merkmal ein wichtiger Faktor für die Festlegung des Erhebungsortes sein und einen Ort entweder für die Durchführung der Erhebung als unverzichtbar oder als bereits auszuschließen kennzeichnen.[5]

Weiterhin stellen Standortfaktoren einen mit einzubeziehenden Faktor dar. Handelt es sich zum Beispiel um eine Schule, welche innerhalb eines sozial schwach ausgeprägten Milieus der Stadt liegt, wird dieses Spezifikum sowohl die Zusammensetzung der Stichprobe als auch die Auswertung der Ergebnisse beeinflussen und bildet damit einen der Einflussfaktoren.

Kriterium 3: Pragmatismus

Erfordert die Fragestellung des Forschungsprojekts keine spezifischen Rahmenbedingungen, wie das Vorhandensein eines expliziten Einrichtungsprofils, kann die Wahl des Ortes entweder auf bereits durch Kooperationen zur Verfügung stehende oder von den Untersuchungsvariablen unabhängige Einrichtungen entfallen. Hierbei müssen die jeweiligen Einrichtungen dennoch hinsichtlich ihrer spezifischen Merkmale reflektiert und die potenziellen Auswirkungen bestimmter Eigenschaften auf die empirische Erhebung analysiert werden. Fokussiert das Forschungsvorhaben beispielsweise einen Vergleich zwischen der Leistung von Jungen und Mädchen, kann – abhängig von der Zielsetzung – die Merkmalsstruktur des Erhebungsortes zwar beeinflussend für die Ergebnisse wirken, jedoch irrelevant für die Auswahl des Erhebungsortes sein. Weiterhin sollte eine möglichst pragmatische Lösung für das Erreichen des Erhebungsortes nicht außer Acht gelassen werden, sodass sowohl die für die Erhebung notwendigen Materialien an den Erhebungsort gebracht und anschließend wieder mitgenommen als auch die beteiligten Personen den betreffenden Ort zeitökonomisch erreichen können.

Beispiel: Die Projektverantwortlichen befinden sich an der Universität in Greifswald. Zwei Schulen weisen eine für die Fragestellung ideale Merkmalsspezifik hinsichtlich des Schulprofils auf, sodass sie als unverzichtbar für das Forschungsvorhaben eingeschätzt werden. Da sich die eine Institution jedoch in Essen, die andere allerdings in Rostock befindet, fällt aus pragmatischen Gründen die Wahl des Erhebungsortes auf den näher gelegenen Standort in Rostock.

Unabhängig von der Auswahl der Erhebungseinrichtung anhand institutionsbedingter, merkmalsspezifischer oder pragmatischer Faktoren müssen jedoch insbesondere die kommunikativen Wege des Forschungsvorhabens geplant und eingehalten werden, um das Gelingen der empirischen Erhebung zu gewährleisten. Hierfür zeigen sich folgende Schritte als unverzichtbar:

[5] Für die Suche nach bundesweiten Bildungseinrichtungen, welche spezifische Schulprofile aufweisen und einen engen Austausch bezüglich deren schulischer Entwicklungsprozesse pflegen, sei u.a. die Website der *Deutschen Schulakademie* empfohlen (https://www.deutsche-schulakademie.de/).

1. Frühzeitige Anfrage der Erhebungseinrichtung

   Die Anfrage der Erhebungseinrichtung muss frühzeitig erfolgen, da Zeit für gemeinsame Absprachen, (Eltern-)Informationen sowie Vorbereitungstreffen gewährleistet sein müssen. Auch innerhalb der Institution benötigt die Entscheidungsfindung über die Genehmigung der Durchführung einige Zeit. Anderenfalls kann es passieren, dass die Erhebung erst verspätet oder überhaupt nicht an dem vorgesehen Ort oder Zeitpunkt stattfinden kann. Außerdem empfiehlt sich die Einplanung eines zeitlichen Puffers, um potenzielle Alternativen zu sichten, anzufragen und notfalls auszuweichen.

2. Verbindlichkeit von Absprachen

   Jede Absprache, die zwischen den Projektverantwortlichen und den Verantwortlichen am Erhebungsort besteht, muss eingehalten werden. Für spezifische Planungsaspekte oder auch zur konkreten Vorstellung des Forschungsvorhabens werden persönliche Treffen vereinbart, in welchen die Besprechung des Erhebungsrahmens erfolgen kann. Neben projektspezifischen Absprachen müssen auch beispielsweise Informationen hinsichtlich des Leistungsstandes oder der Spezifika der Probandengruppe zur Verfügung stehen, um den Erhebungsrahmen bestenfalls anzupassen.

3. Berücksichtigung und Information aller beteiligter Personen

   Während der Planung müssen alle beteiligten Personen des Erhebungsortes rechtzeitig umfassende Informationen erhalten. Im Vorfeld sollte daher festgehalten werden, welche Personen sich in welchem Umfang an der Erhebung beteiligen und in welcher Art und Weise die konkrete Durchführung diese betrifft.

Festlegung des Erhebungszeitraumes

Die Festlegung des Erhebungszeitraums stellt den dritten Faktor der Planung der Rahmenbedingungen dar. Obgleich dieser bereits im Vorfeld in die Planung und Vorbereitung des Erhebungsprozesses miteinbezogen werden muss, offenbart sich dieser Faktor zumeist als der am wenigsten planbare, da die Wünsche und terminlichen Verpflichtungen der Erhebungseinrichtung oder der Stichprobe in der Regel nicht im Vorfeld abzusehen sind. Aufgrund dessen erweist es sich als vorteilhaft, einen Spielraum hinsichtlich des konkreten Erhebungszeitraumes einzuplanen. Die Planung des Erhebungszeitraumes muss sich an dem durch den Versuchsleiter bestimmten ‚Erhebungszeitpunkt' und den durch externe Systeme bestimmten ‚Temporalen Begrenzungen' orientieren:

1. Erhebungszeitpunkt

   Der Erhebungszeitpunkt wird aktiv durch den Versuchsleiter festgelegt. Zu Beginn des Vorbereitungsprozesses muss daher zunächst die Entscheidung erfolgen, zu welchem spezifischen Zeitpunkt die empirische Erhebung stattfinden soll. Hierbei kann es beispielsweise relevant sein, dass andere Projekte

bereits abgeschlossen sind, bestimmte Themen erst in folgenden Unterrichtseinheiten erarbeitet wurden oder sich Schülerinnen und Schüler zu Beginn, in der Mitte oder auch am Ende eines Schuljahres befinden.

Außerdem muss festgelegt werden, inwieweit die Tageszeit oder auch der Wochentag entscheidend für das Gelingen der empirischen Erhebung sein kann.

2. Temporale Begrenzung

Die temporalen Begrenzungen des Erhebungsrahmens erfolgen durch äußere Einflüsse und wirken sich auf die Konzeption der Rahmenbedingungen aus. Neben den Ferienzeiten des jeweiligen Bundeslandes können dabei anderweitige Verpflichtungen der Erhebungseinrichtung oder der Probandengruppe sowie regionsspezifische Termine (beispielsweise Einschulungszeitpunkte, das jährliche Bürgerfest etc.) den Erhebungszeitraum beeinflussen.

Außerdem können an manchen Erhebungsorten lediglich kleine Zeitfenster für die empirische Erhebung zur Verfügung gestellt oder ganze Zeiträume blockiert werden, sodass die Erhebung ausschließlich in wenigen (Schul-) Stunden erfolgen muss. Die Einschränkung beeinflusst den Erhebungsrahmen erheblich.

3. Fazit

Vor dem Hintergrund der vielfältigen Herausforderungen empirischen Forschens erweist sich die konkrete Planung und Vorbereitung der Erhebung selbst als zentrale Phase des ganzheitlichen Forschungsprozesses. Hierbei ermöglicht die spezifische Vorbereitung des Forschungsprojekts einem jeden Forschenden die frühzeitige Vermeidung oder Bewältigung verschiedener Schwierigkeiten, welche hinsichtlich der Wahl des Untersuchungsdesigns, der Forschungsmethodik, der zu verwendenden Instrumentarien, der Zusammensetzung der Stichprobe oder der zu gestaltenden Rahmenbedingungen während des Forschungsvorhabens auftreten können.[6]

Die Analyse des eigenen Forschungsvorhabens anhand der planungsbestimmenden Konzeptionen des grundlegenden empirischen Settings, des Erhebungsinstrumentariums und des Erhebungsrahmens unterstützt dabei die konkrete Durchführung empirischer Erhebungen und bietet die Möglichkeit, das eigene Forschungsvorhaben zu reflektieren, durchführungsrelevante Fallstricke zu beseitigen und durch eine gezielte und fragestellungsorientierte Vorbereitung zum Gelingen des gesamten Forschungsprozesses beizutragen.

---

[6] Für eine vertiefende Auseinandersetzung mit der Planung und Vorbereitung empirischer Erhebungen seien unter anderem die Werke von Aeppli et al. 2016; Döring/Bortz 2016 sowie Hug/Poscheschnik 2015 empfohlen.

Literatur

Aeppli, Jürg et al. (2016): Empirisches wissenschaftliches Arbeiten. Ein Studienbuch für die Bildungswissenschaft. Bad Heilbrunn: Klinkhardt.

Atteslander, Peter (2003): Methoden der empirischen Sozialforschung. Berlin: Walter de Gruyter.

Boelmann, Jan M. (2016): Von der Idee zum Forschungsprojekt. In: Boelmann, Jan M. (Hrsg.): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. Baltmannsweiler: Schneider Hohengehren, 11-21.

Döring, Nicola/Bortz, Jürgen (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Heidelberg: Springer.

Häder, Michael (2010): Empirische Sozialforschung. Wiesbaden: VS Sozialwissenschaften.

Hussy, Walter/Schreier, Margit/Echterhoff, Gerald (2010): Forschungsmethoden in Psychologie und Sozialwissenschaft. Heidelberg: Springer.

Kaya, Maria/Himme, Alexander (2007): Möglichkeiten der Stichprobenbildung. In: Albers, Sönke et al. (Hrsg.): Methodik der empirischen Sozialforschung. Wiesbaden: Springer, 79-88.

Mooneyham, Benjamin W./Schooler, Jonathan W. (2013): The Costs and Benefits of Mind-Wandering: A Review. In: Canadian Journal of Experimental Psychology, 67, 1, 11-18.

Poscheschnik, Gerald (2015): Planung eines Forschungsprojekts. In: Hug, Theo/Poscheschnik, Gerald: Empirisch Forschen. Die Planung und Umsetzung von Projekten im Studium. Konstanz: UVK, 61-98.

Schnell, Rainer/Hill, Paul B./Esser, Elke (2013): Methoden der empirischen Sozialforschung. München: Oldenbourgh.

Settinieri, Julia (2014): Planung einer empirischen Studie. In: Settinieri, Julia et al.: Empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache: Eine Einführung. Stuttgart: Schöningh, 57-70.

Anhang:
Checkliste zur Vorbereitung empirischer Forschungsvorhaben

Vorbereitung empirischer Forschungsvorhaben	
Projektausrichtung	
Zielsetzung des Forschungsprojekts	
Konkrete Fragestellung	
(Potenzielle) Hypothesen	
Empirisches Forschungssetting	
Forschungsparadigma	
Forschungsansatz	
Untersuchungsdesign	
Erhebungsverfahren	
Auswertungsverfahren	
Planung der empirischen Erhebung	
Erhebungsstichprobe (Größe, Zusammensetzung, Alter, wichtige Prädiktoren…)	
Erhebungsort (Bildungseinrichtungen, [städtische] Freizeitangebote…)	
Erhebungszeitraum (einzelne Stunden, Tage, Wochen…)	
Erhebungsinstrumentarium (entwickelte Fragebögen/Leitfäden, standardisierte Tests…)	
Auswertungsinstrumentarium (Kategoriensysteme, Manuals…)	
Potenzielle Störvariablen	

Leitfragen zur konkreten Planung der empirischen Erhebung
Pilotierung
Wie wird das Erhebungsinstrumentarium oder das Forschungssetting pilotiert?
Welche Konsequenzen ergeben sich aus den Pilotierungsergebnissen hinsichtlich des Erhebungsinstrumentariums, des Erhebungsrahmens oder des Forschungssettings?
Personelle Ressourcen
Kann die Erhebung von dem Versuchsleiter selbst durchgeführt werden oder wird beispielsweise aufgrund der Stichprobe ein Versuchsleiterteam benötigt?
Welche Inhalte müssen den Teammitgliedern in einer Schulung vermittelt werden? Welche Materialien werden hierfür benötigt? Wie viel Zeit wird hierfür benötigt?
Rücksprachen mit externen Unterstützungssystemen
Welche Rücksprachen bestehen mit den externen Beteiligten der empirischen Erhebung, wie beispielsweise der betreffenden Lehrkraft/Betreuern oder der Schulleitung (zeitliche, räumliche oder materielle Ansprüche des empirischen Forschungssettings)?
Wurden alle Beteiligten der empirischen Erhebung über das Vorgehen und die Intention des Forschungsprojekts rechtzeitig informiert (beispielsweise Schulleitung, Lehrkräfte, Kinder, Eltern … evtl. Regierungspräsidien)?
Welche Maßnahmen zur Einhaltung des Datenschutzes werden im Forschungssetting vorgenommen?

DIANA MAAK

Personenbezogene Daten

Nicht *en passant*, sondern *en détail* und *bien conçu*.

1. Einleitung

Bei personenbezogenen Daten handelt es sich um „Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten […] natürlichen Person" (BDSG §3(1)), z.B. um das Alter oder um den Geburtsort. Der Erhebung solcher Daten kommt im Rahmen von Forschung häufig eine bedeutende Rolle zu: zur Stichprobenbeschreibung (Döring 2013, 95) oder wenn vorab formulierte Hypothesen zu Gruppenunterschieden (Jungen vs. Mädchen, Schülerinnen und Schüler mit Migrationshintergrund vs. Schülerinnen und Schüler ohne Migrationshintergrund) überprüft werden sollen.[1] Ein erstes Qualitätsmerkmal stellt bei der Erhebung dieser Daten die Gewährleistung von Nachvollziehbarkeit und Transparenz dar.[2] Demnach sollten eingesetzte Erhebungsinstrumente und zugrunde gelegte

[1] In Einzelfällen stellt die Erhebung solcher Daten sogar den Fokus von Forschungsprojekten dar. Beispielhaft wären hier die Entwicklung von sprachbiographischen Fragebögen zu erwähnen, welche vor allem der Erfassung der in der eigenen Biographie gemachten Lehr-/Lernerfahrungen und Kontakten mit Sprachen dienen (Maak 2018, 184). Dieses Ziel verfolgten etwa die home language surveys (Fürstenau/Gogolin/Yağmur 2003; Extra et al. 2001 und die Diskussion in Chlosta/Ostermann 2006), SPREEG (vgl. z.B. Chlosta/Ostermann 2006; Chlosta/Ostermann/Schroeder 2003), eine Untersuchung in Freiburg (vgl. z.B. Decker/Schnitzer 2012) und MaTS (vgl. z.B. Ahrenholz/Maak 2013).
Ziel war es vor allem, „grundlegende Daten über Sprachenvielfalt" (Chlosta/Ostermann 2005, 56) zu erheben. Alle angeführten Studien belegen eine (viel größer als bisher angenommene) Vielfalt hinsichtlich der Mehrsprachigkeit von Schülerinnen und Schülern und erzeugen damit ein realistischeres Bild der mehrsprachigen Schülerschaft in Deutschland. Die Erhebung von personenbezogenen Daten – fokussiert auf die Mehrsprachigkeit von Schülerinnen und Schülern – bildete folglich das zentrale Erkenntnisinteresse dieser Studien.

[2] Für Hinweise zu Gütekriterien der Klassischen Testtheorie sei auf die Ausführungen in Bortz/Döring 2006, 195ff. verwiesen, für Ausführungen zur Gütekriterien qualitativer Forschung auf Steinke 2006. Vgl. auch die beiden Beiträge zu Gütekriterien von Schmidt in diesem Band. Für allgemeine Hinweise zu Gütekriterien und für allgemeine Hinweise zu Gütekriterien ergänzt um Spezifika der Fremd- und Zweitsprachenforschung siehe Schmelter 2014.

Definitionen von Konstrukten wie Migrationshintergrund dargestellt und begründet werden. Eine solche Vorgehensweise ist gleichzeitig auch eine wichtige Voraussetzung für einen Vergleich unterschiedlicher Studien(ergebnisse).

Damit sind die beiden inhaltlichen Schwerpunkte des Beitrags bereits angedeutet. Erstens wird davon ausgegangen, dass in vielen – insbesondere kleineren – Forschungsvorhaben personenbezogene Daten eher *en passant* und nicht in jedem Fall wohl überlegt (*bien conçue*) erhoben werden. Der Erarbeitung und Qualität von durch die Probandinnen und Probanden zu bearbeitenden Aufgaben, Fragen, Items etc., die den Kern des Untersuchungsgegenstands betreffen, wird zumeist mehr Aufmerksamkeit gewidmet, was sich u.a. in mehr Bedenken sowie einem höheren Zeit- und Arbeitsaufwand diesbezüglich äußert. Das ist dann akzeptabel, wenn sich bei der Erhebung personenbezogener Daten an gängigen Standards orientiert wird, die ‚lediglich' auf das eigene Erkenntnisinteresse abgestimmt werden müssen. Entsprechende Hinweise auf Standards sowie grundlegende Überlegungen zur Erhebung personenbezogener Daten liefert der Beitrag in Kapitel 2.

Zweitens zeigt aber bereits ein erster Blick in die Erhebung gängiger personenbezogener Daten, wie z.B. des Geschlechts, dass selbst dieses scheinbar einfach zu erhebende Merkmal durchaus unterschiedlich operationalisiert werden kann. Eine eingehende Auseinandersetzung mit Definitionen zentraler Begriffe sowie bewusste Entscheidungen im spezifischen Untersuchungskontext sind daher entscheidend. Insbesondere, wenn auf Basis solcher Kriterien später etwa Gruppenunterschiede in Stichproben geprüft werden sollen. Diese kritische Auseinandersetzung sowie die Reichweite unterschiedlicher Entscheidungen wird in Kapitel 3 anhand der Merkmale Geschlecht sowie Migrationshintergrund und Mehrsprachigkeit diskutiert. Hier zeigt sich auch noch einmal, dass auch die Frage danach, wie detailliert die Erfassung von personenbezogenen Merkmalen erfolgen soll, ausreichend Aufmerksamkeit erhalten sollte.

Der Beitrag wendet sich insbesondere an Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler, die im Rahmen ihrer Forschungsvorhaben auch personenbezogene Daten erheben wollen und damit noch keine Erfahrungen haben bzw. diese Daten bislang eher *en passant* erhoben haben. Den Abschluss bildet in Kapitel 4 das Fazit.

2. Personenbezogene Daten

Das Kapitel gliedert sich in die Auseinandersetzung mit den Fragen, welche Daten als personenbezogen anzusehen sind und im Forschungsfeld Schule eventuell spezifische Relevanz haben (Kapitel 2.1). Daran schließt sich sie Auseinandersetzung mit Standards und Ressourcen an, die als Orientierung und Grundlage für die Erhebung von personenbezogenen Daten dienen können (Kapitel 2.2).

2.1 Welche Daten sind als personenbezogen anzusehen?

Zunächst sollen folgende Fragen geklärt werden: Welche Merkmale sind als personenbezogene Angaben anzusehen und insbesondere bei Untersuchungen im Schulkontext von Interesse?

Der vorliegende Beitrag widmet sich vor allem dem *Wie* der Erhebung von personenbezogenen Daten. Dabei sei bereits an dieser Stelle darauf hingewiesen, dass aus forschungsethischer Perspektive so wenige personenbezogene Daten wie möglich erhoben werden sollten (BDSG §3a). Ob die gewünschten Daten überhaupt erhoben werden dürfen, ist eine ebenso zentrale Frage, der aber an dieser Stelle nicht nachgegangen wird. Für entsprechende Ausführungen sei auf den Beitrag von Iberer in diesem Band verwiesen.

Wie bereits erläutert, ist die Orientierung an Standards – sowohl bezüglich der Begriffsdefinition als auch hinsichtlich der konkreten Erhebung – zu empfehlen. Erste entsprechende Anhaltspunkte liefert die regelmäßig aktualisierte Empfehlung zur Erhebung von Demographischen Standards, die gemeinsam vom ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V., der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V. (ASI) und dem Statistischen Bundesamt verantwortet wird (Statistisches Bundesamt 2010, 1999 u.a.). Ziel ist es, „[…] sozialstrukturelle Erhebungsmerkmale in Befragungen zu vereinheitlichen, um eine mögliche Vergleichbarkeit zwischen einzelnen Umfragen zu erzielen." (Statistisches Bundesamt 2016, 5). Eine einheitliche Erhebung erweitert ferner die Verknüpfungsmöglichkeiten von unterschiedlichen Datensätzen (Statistisches Bundesamt 2016, 3). Es werden insgesamt sechszehn Kernvariablen angegeben, hierunter z.B. das Geschlecht, das Alter, die Staatsangehörigkeit (optional liegt ein erweiterter Fragenkatalog zur Erfassung des Migrationshintergrunds vor), der Familienstand, der höchste allgemeinbildende Schulabschluss, die Erwerbssituation, die hauptsächlich ausgeübte berufliche Tätigkeit, die Telekommunikationsmöglichkeiten des Haushalts, die Haushaltsgröße und das Haushaltsnettoeinkommen (Statistisches Bundesamt 2016, 8).

Insbesondere Alter und Geschlecht gelten als personenbezogene Angaben, die in so gut wie allen Untersuchungen von allen Probandinnen und Probanden erhoben werden. Die Auflistung zeigt jedoch bereits, dass in Abhängigkeit von Forschungskontext und Erkenntnisinteresse davon abgesehen sehr viele weitere Merkmale von Bedeutung sein können und die genannten im schulischen Kontext wiederum nicht unbedingt relevant sein müssen. Hierbei gestaltet sich im Einzelfall die Grenze zwischen personenbezogenen und nicht personenbezogenen Daten fließend bzw. ist nicht eindeutig zu ziehen. Da im schulischen Kontext unterschiedliche Akteure – Schulleiterinnen und Schulleiter, Lehrerinnen und Lehrer, Schülerinnen und Schüler, Eltern etc. – an Studien teilnehmen, sollten dieselben Merkmale zudem adressatenspezifisch erhoben werden. Eine Anpassung an die

jeweilige Zielgruppe ist entscheidend, damit die Verständlichkeit und angemessene Beantwortung gewährleistet werden kann.[3] Weitere Orientierungen hierzu können z.B. größere Schulleistungsstudien wie *PISA* liefern.

Tabelle 1 stellt personenbezogene Angaben, die im Rahmen von *PISA* 2009 (Hertel et al. 2014) und im Rahmen von StEG[4] erhoben worden sind, vergleichend dar. Dabei kommen unterschiedlichen Bezeichnungen – Schülermerkmale, Hintergrundvariablen, demografische Angaben u.a. – zur Anwendung.

Tab. 1: Überblick über (ausgewählte) Hintergrundvariablen getrennt nach Personengruppen im Rahmen von *PISA* 2009 und StEG

	***PISA* 2009** (Hertel et al. 2014)	***StEG*** (Quellenberg 2009)
Schülerinnen und Schüler	Schülermerkmale: • Klassenstufe • Besuchte Schulart • Geburtsdatum • Geschlecht • Besuch von Kindergarten/Vorschule • Alter bei Einschulung in die Grundschule • Rückstufung/Wiederholung einer Klasse	Ausgewählte Hintergrundvariablen (Sekundarstufe): • Geschlecht • Alter • Familienstruktur • Ganztagsteilnahme • Klassenstufe • Migrationshintergrund
Schulleiterinnen und Schulleiter	Keine personenbezogenen Angaben erhoben	Ausgewählte Hintergrundvariablen: • Geschlecht • Alter • SchulleiterIn seit • Schulart • Siedlungstyp • Gemeindegröße • Bundesland • Ganztagsschultyp

[3] Exemplarisch sei auf eine eingehende Diskussion der Angemessenheit eines sprachbiographischen Fragebogens für Grundschulkinder von Maak/Zippel/Ahrenholz 2013 verwiesen. Hinweise zu Spezifika der Befragung von Kindern finden sich z.B. in Walper/Tippelt 2010; Lange/Mierendorff 2009; Grunert/Krüger 2006; Kränzl-Nagel/Wilk 2000; Lipski 2000. Allgemeine Hinweise zur Fragebogenerstellung finden sich in Porst 2009.

[4] „Seit 2005 wird mit der ‚Studie zur Entwicklung von Ganztagsschulen' (StEG) ein länderübergreifendes Forschungsprogramm zur Entwicklung von Ganztagsschulen […] angeboten [und] gefördert." (https://www.projekt-steg.de/content/Über-steg (letzter Zugriff: 01.08.2018)); Änderungen vorgenommen von D.M.). Das Projekt befindet sich zum Zeitpunkt der Erstellung des Beitrags in der dritten Förderphase.

	***PISA* 2009** (Hertel et al. 2014)	***StEG*** (Quellenberg 2009)
Lehrkräfte	Demografische Angaben: • Geschlecht • Altersgruppe • Dienstjahre (gesamt) • Dienstjahre an der Schule • Lehramtsabschluss • Fachstudium Deutsch im Rahmen des Lehramtsstudiengangs • Deutsch als Haupt-/Leitfach oder Erweiterungsfach • Teilnahme an Fortbildungen	Ausgewählte Hintergrundvariablen: • Alter • Geschlecht
Eltern	Allgemeine Angaben: • Antwortende Person Hintergrund der Person • Abgeschlossene Berufsausbildung des Vaters/der Mutter • Schulabschluss des Vaters/der Mutter • Beschäftigungsstatus des Vaters/der Mutter • Beruf des Vaters/der Mutter • Ausgeübte Tätigkeit des Vaters/der Mutter • Berufliche Stellung der Eltern • Vorgesetztenfunktion der Eltern bei der Arbeit • Jährliches Haushaltseinkommen • Ausgaben für Bildungseinrichtungen • Anzahl der im Haushalt lebenden Kinder	Ausgewählte Hintergrundvariablen: • Angabe zu den ausfüllenden Personen („Dieser Fragebogen wird ausgefüllt von …“: Mutter/Partnerin, Vater/Partner, Mutter und Vater, Andere etc.) • Höchster Ausbildungsabschluss • Höchster Abschluss (ISCED) • Zeitraum mit Betreuungsbedarf

Die Gegenüberstellung verdeutlicht einerseits, dass in Abhängigkeit von der befragten Personengruppe zum Teil sehr unterschiedliche personenbezogene Angaben erhoben werden. Schulleiterinnen und Schulleiter werden vornehmlich zur Struktur und Organisation, zur Schüler- und Lehrerschaft sowie zu Ressourcen der Schule befragt. Personenbezogene Angaben treten hier eher in den Hintergrund. Ein Merkmal, das bei Lehrkräften häufig erhoben wird, ist die Anzahl der Dienstjahre. Und von Eltern werden so gut wie alle Kernvariablen der demographischen Standards (Statistisches Bundesamt 2016) erhoben. Allerdings ist anzumerken, dass es sich hierbei nicht um eine vollständige Übersicht handelt. Denn die Schülerinnen und Schüler werden etwa im Rahmen von *PISA* zu zahlreichen weiteren Merkmalen, z.B. aus dem Bereich *Familie und Zuhause* und *Familiärer Hintergrund und Sprachgewohnheiten* befragt (Hertel et al. 2014, 5ff.), die ebenfalls zum Teil personenbezogene Daten darstellen.

Die Auflistung in Tabelle 1 zeigt auch, dass bei der Erhebung personenbezogener Daten teilweise sehr sensible Daten erhoben werden, z.B. Informationen über das Einkommen. Dabei ist das Empfinden darüber, was von den Befragten als sensible Daten oder gar Tabuthemen wahrgenommen wird, sehr unterschiedlich. Beispielsweise können Fragen zur Familie und zum Sprachgebrauch zu Hause potentiell belastend sein; dies wird insbesondere diskutiert, wenn es um die Befragung von unbegleiteten minderjährigen Flüchtlingen geht. Porst (2009, 143) weist darauf hin, dass demographische Daten von Befragten oft nicht gerne bearbeitet werden, obwohl sie in der Regel recht leicht zu beantworten seien. Daher empfiehlt er, solche Fragen an das Ende eines Fragebogens zu stellen:

> Kommt es aufgrund von demographischen Fragen zur Lustlosigkeit oder gar Verärgerung, sollte sich das […] möglichst erst am Ende auswirken, wenn man den größten Teil des Fragebogens in hoffentlich konstruktiver Weise schon abgearbeitet hat. (Porst 2009, 143)

Dieses Zitat zeigt einerseits, dass die Erhebung demographischer Daten eher nachgeordnet wird. Es zeigt andererseits, dass Fragen dazu potentiell belastend sein können und sowohl zum Abbruch der Befragung als auch zu veränderten Ergebnissen aufgrund von veränderten Gemütszuständen führen können. Letzteres spricht erneut für eine sorgfältige Auswahl und Abwägung.

2.2 Standards der Erhebung personenbezogener Daten

Eine Orientierung an demografischen Standards (Statistisches Bundesamt 2016) oder an erprobten Instrumenten wie den im Rahmen von *PISA* (Hertel et al. 2014) eingesetzten, liefert, neben einer generellen Qualitätssicherung, weitere Vorteile. Die Hinweise zu den demografischen Standards des Statistischen Bundesamts beinhalten nicht nur Vorschläge dazu, welche Kernvariablen zu erheben sind, sondern auch, wie diese konkret zu erheben sind; d.h., dass konkrete Frageformulierungen auch in Abhängigkeit der Erhebungsform (mündlich vs. schriftlich) vorgelegt sowie Hinweise zur Erhebungsdurchführung gegeben werden. Liegen ferner Skalenhandbücher vor, so enthalten diese zumeist weitere wichtige Hinweise für eigene Untersuchungen. Skalenhandbücher beinhalten in der Regel Skalen zur Erfassung bestimmter (latenter) Merkmale (Rössler 2011, X) und stellen wichtige Referenzwerke für die eigene Forschung dar. Denn:

> Ein fortschreitender, auf frühere Arbeiten aufbauender Erkenntniszuwachs ist im Feld der empirischen Forschung im Grunde nur möglich, wenn die Messung der für die Theorie relevanten Konstrukte in vergleichbarer, möglichst einheitlicher Form erfolgt. (Rössler 2011, IX)

Beispielsweise können die Skalenhandbücher von großen Schulleistungsstudien entsprechend genutzt werden (z.B. Hertel et al. 2014 für PISA 2009).[5] Sie enthalten u.a. folgende Informationen:

[5] Relevante Hinweise können sich allerdings auch in anderen Fachgebieten finden. So könnten in Abhängigkeit des Erkenntnisinteresses die Hinweise zur Erfassung von

- Wortlaut für Items
- Antwortformate sowie -kategorien für Items
- Hinweise zur Aufbereitung der Daten
- Literaturhinweise (Ursprung von Skalen)
- deskriptiv-statistische Ergebnisse der betreffenden Studie, die als Vergleichswerte für eigene Analysen dienen können
- statistische Kennwerte zur Güte der Skalen (z.B. Cronbachs α)

Idealerweise liegen im Rahmen der Ergebnisdarstellung zu Referenzstudien eine plausible Fallzahl, der Originaltext der Fragestellung sowie auch die verwendeten Antwortvorgaben und eine Grundauszählung aller Items mit Mittelwerten und Standardabweichung vor. Außerdem sollte bei Dimensionsreduktion das genaue faktorenanalytische Verfahren[6], die erklärte Varianz, Eigenwerte und alle relevanten Factor-Scores angegeben sein. Schließlich ist bei Indices die Angabe der exakten Berechnungsvorschriften zusammen mit einer Grundauszählung der Indexwerte wünschenswert (Rössler 2011, XIIf.).

Als weitere zentrale Anlaufstellen können GESIS, DIPF und IQB dienen.[7] Beispielsweise findet sich auf den Seiten des Fachportals für Pädagogik die Datenbank zur Qualität von Schule, kurz DaQS, die folgende Services bietet:

> DaQS ist ein Datenbankangebot von und für die quantitative empirische Bildungsforschung, das am Deutschen Institut für Internationale Pädagogische Forschung (DIPF) entwickelt wurde. Im Rahmen dieses Services werden auf der Basis von Fragebogen und Skalenhandbüchern einschlägiger Studien Instrumente zur Erfassung von Schul- und Unterrichtsqualität dokumentiert und aufbereitet. Dies ermöglicht differenzierte Einsichten in die Anlage der dokumentierten Studien sowie Vergleiche der eingesetzten Instrumente und Skalen anhand gängiger Kennwerte. Das Angebot ist frei zugänglich und kostenlos nutzbar. (DaQS 2018)

Neben einer einfachen und einer erweiterten Suche können auch die Konstruktliste, die das DaQS-eigene Konstruktschema abbildet sowie die Studienliste, die

---

Mediennutzung im Skalenhandbuch Kommunikationswissenschaft (Rössler 2011) ebenfalls von Bedeutung sein.

6 Faktorenanalysen dienen der Datenstrukturierung und -reduzierung. Es wird hierbei versucht, Beziehungszusammenhänge in einem großen Variablenset zu strukturieren, indem Gruppen von Variablen identifiziert werden, die hoch miteinander korreliert sind. Die Gruppen von jeweils hoch korrelierten Variablen bezeichnet man als Faktoren (vgl. Backhaus et al. 2008 für weiterführende Informationen).

7 GESIS – das Leibniz-Insitut für Sozialwissenschaften ist die größte deutsche Infrastruktureinrichtung und steht Forscherinnen und Forschern auf allen Ebenen ihrer Forschungsvorhaben mit Expertise und Dienstleistungen beratend zur Seite. Das Deutsche Institut für Internationale Pädagogische Forschung, kurz DIPF, unterstützt Forschung, Politik und Praxis im Bildungsbereich. Das Institut zur Qualitätsentwicklung im Bildungswesen, kurz IQB, ist eine wissenschaftliche Einrichtung der Länder und als An-Institut der Humboldt-Universität zu Berlin angesiedelt. U.a. werden Datensätze nationaler und internationaler Bildungsstudien mit Kompetenzmessungen archiviert und für Sekundäranalysen zur Verfügung gestellt (https://www.forschungsdaten-bildung.de/projekt?la=de (letzter Zugriff: 01.08.2018)).

in DaQS dokumentierte Studien auflistet, einen Zugang zu den gesuchten Aspekten ermöglichen. Die Suchergebnisse liefern – in Abhängigkeit der verfügbaren Daten – ähnliche Informationen wie Skalenhandbücher sowie zum Teil auch Verlinkungen zu den Originalinstrumenten und relevanten Publikationen.

Selbstverständlich sollten vorhandene Instrumente nicht unreflektiert übernommen werden. In der Regel sind Kürzungen und Zusammenfassungen ebenso wie die Ergänzung um weitere zu erfassende Merkmale sinnvoll oder gar notwendig (Rössler 2011, XIV, Statistisches Bundesamt 2016, 6). Auch erfolgreiche Standardisierungen können bald veraltet sein und damit aktuellen (gesellschaftlichen, wissenschaftlichen u.a.) Entwicklungen nicht mehr gerecht werden (Rössler 2011, XIV). Zu beachten ist allerdings, dass statistische Kennwerte zur Güte von Items, die eine validierte Skala abbilden, ihre Relevanz bzw. Gültigkeit für die eigene Studie verlieren, wenn die Items bzw. deren Zusammenstellung o.Ä. verändert werden.

3. Von der scheinbar einfachen Erhebung personenbezogener Daten

Nachfolgend soll anhand zweier Beispiele verdeutlicht werden, welche Bedeutung der wohl durchdachten Erhebung auch scheinbar *einfach* und *eindeutig* zu erhebender personenbezogener Daten zukommt.

3.1 Geschlecht

Bei den nachfolgenden Ausführungen handelt es sich im Wesentlichen um eine verkürzte Darstellung der umfassenden und lesenswerten Diskussion zur Operationalisierung von Geschlecht im Fragebogen von Döring (2013), die um aktuelle Entwicklungen ergänzt wurde. Das Geschlecht wird eigentlich immer erhoben. Dabei dient es als soziodemographische Variable zur Stichprobenbeschreibung, als Filtervariable zur Auswahl passender Fragen, als Kontrollvariable zur Verhinderung von Geschlechtsblindheit[8] bei der Auswertung und schließlich als theoretisch relevante Variable zur Hypothesenüberprüfung oder Hypothesenbildung (Döring 2013, 95). Insbesondere wenn die Erhebung des Geschlechts der letztgenannten Funktion dient, kommt der wohl überlegten Operationalisierung eine zentrale Bedeutung zu. Es sollte zunächst klar sein, ob das biologische oder das soziale Geschlecht erhoben werden soll. Das biologische Geschlecht bezieht sich auf angeborene körperliche Merkmale, das soziale Geschlecht hingegen auf geschlechtsbezogene Verhaltens- und Erlebensweisen (Döring 2013, 104). Insbesondere problematisch erscheint die Erhebung des biologischen Geschlechts als Stellvertreter für das soziale Geschlecht, wenn das biologische Geschlecht dann

[8] Wenn im Rahmen der Datenauswertung die Unterscheidung zwischen Männern und Frauen nicht routinemäßig einbezogen wird, dann wird dies als geschlechtsblind kritisiert. Denn eine solche Auswertung könnte dazu führen, dass geschlechtsspezifische Benachteiligungen ignoriert oder verleugnet werden (Döring 2013, 95).

als Erklärungsfaktor im Rahmen von Ursache-Wirkungs-Verhältnissen angeführt wird (Döring 2013, 104). Schnell, Hill und Esser (2005, 70f.) erläutern dies am Beispiel des Untersuchungsergebnisses „Frauen wählen häufiger als Männer christliche Parteien", wobei hier Geschlecht und Wahlverhalten in einen kausalen Zusammenhang gesetzt werden:

> Obgleich z.B. die Variable ‚Geschlecht' im Allgemeinen über die Feststellung der biologischen Geschlechtszugehörigkeit ermittelt wird, würde jedoch kein Sozialwissenschaftler argumentieren, dass diese biologische Geschlechtszugehörigkeit das Wahlverhalten determiniert. Dies würde nämlich bedeuten, dass die Ausstattung mit bestimmten primären und sekundären Geschlechtsmerkmalen eine bestimmte politische Wahlentscheidung verursacht. Vielmehr würde man argumentieren, dass das weibliche Rollenverständnis, die weibliche Identität, die männlichen Wertvorstellungen o.a. die Ursache für eine bestimmte Handlung (hier: eine Wahlentscheidung) darstellen. (Schnell/Hill/Esser 2005, 70)

Für die Operationalisierung in Form von Einzelitems ergeben sich neben der in Tabelle 2, Zeile 1 abgedruckten *klassischen* Variante noch zahlreiche weitere Möglichkeiten – jeweils mit eigenen Vor- und Nachteilen. Für die *klassische* Variante sei zunächst angemerkt, dass diese weit verbreitet ist. Allerdings wird keine Differenzierung nach biologisch und sozial vorgenommen, sodass das zu erhebende Merkmal nicht eindeutig definiert ist. Ferner müssten sich, um Exklusivität zu gewährleisten, die Antwortalternativen wechselseitig ausschließen, wobei es jedoch durchaus möglich ist, dass Personen sich als sowohl männlich als auch weiblich oder weder männlich noch weiblich definieren. Schließlich werden mit dieser Itemform nicht alle möglichen Merkmalsausprägungen abgedeckt, die Exhaustivität ist also nicht gegeben, wie die weiteren Beispielitems verdeutlichen (Döring 2013, 98).

Auch alternative Erhebungsvarianten sind nicht frei von Nachteilen bzw. abzuwägenden Aspekten. So sind aus Perspektive der Befragten ‚gelungene' Fragen an sie grundlegend für die Motivation zur Teilnahme an der Studie. Fühlen sich Personen also durch die Art der Frageformulierung ausgeschlossen oder diskriminiert, dann kann das zu „negativen Einstellungen gegenüber dem Forschungsprojekt – und schlimmstenfalls sogar gegenüber empirischer Sozialforschung allgemein" (Döring 2013, 102) führen. In der konkreten Befragungssituation kann es zu Abbrüchen der Befragung oder systematischer Modifizierung des eigenen Antwortverhaltens kommen (Döring 2013, 102). Es könnten sich z.B. Personen diskriminiert fühlen, die das 2. Item aus Tabelle 2 vorgelegt bekommen, weil diese Präsentation der Antwortalternativen als „symbolische Affirmation der gesellschaftlich vorherrschenden Geschlechter-Hierarchisierung" (Döring 2013, 101) gelesen werden kann. Das Weibliche folgt auf das Männliche und wer sich nicht eindeutig zuordnen kann oder möchte, der ist ‚anders' und kommt ‚zum Schluss' (Döring 2013, 101). Für bezüglich der Vielfalt von Geschlecht nicht sensibilisierte Personen kann jedoch eben dieses Item eine Provokation bedeuten und Irritation hervorrufen (Döring 2013, 103). Dies kann folglich zu einer negativen Einstellung gegenüber dem Forschungsprojekt bzw. -team führen, ebenso wie zu dem Eindruck, es würden ‚unsinnige' Fragen gestellt und die ganze Studie sei

dubios (Döring 2013, 103). Allerdings sind auch forschungsethische Aspekte zu berücksichtigen. So geht Döring (2013, 103, 107) davon aus, dass die Anzahl der Personen, die sich einem nicht-binären biologischen Geschlecht zuordnen, verhältnismäßig gering ist. Dies könnte einerseits zur Gefährdung der Anonymisierung führen und andererseits dazu, dass aufgrund der geringen Fallzahl die Daten dieser Personen zum Teil gar nicht ausgewertet werden, was forschungspraktisch unökonomisch wäre.

Tab. 2: Einzelitems zur Operationalisierung des Konstrukts *Geschlecht* (in Anlehnung an Döring 2013)

1.	Geschlecht: ☐ männlich ☐ weiblich
2.	Biologisches Geschlecht: ☐ männlich ☐ weiblich ☐ anderes, und zwar: _____
3.	Biologisches Geschlecht: ☐ männlich ☐ weiblich ☐ Mann-zu-Frau-transsexuell/transident ☐ Frau-zu-Mann-transsexuell/transident ☐ intersexuell, und zwar: _____ ☐ anderes, und zwar: _____
4.	Soziales Geschlecht ☐ feminin ☐ maskulin ☐ sowohl maskulin als auch feminin ☐ weder maskulin noch feminin ☐ anderes, und zwar: _____
5.	Soziales Geschlecht _____

Döring (2013, 108) plädiert für die Erhebung des Geschlechts mittels psychometrischer Skalen.[9] Allerdings sind diese häufig sehr umfangreich und es liegen

[9] Bei einer Skala wird ein Merkmal über einen Satz von Items operationalisiert, die jeweils auf Ratingskalen (z.B. „nie“ bis „immer“ oder „trifft gar nicht zu“ bis trifft völlig zu“) zu beantworten sind, wobei die Antworten auf die einzelnen Items zu einem intervallskalierten Skalenwert (Summen- oder Durchschnittswert) zusammengefasst werden, sodass für die statistische Auswertung dann z.B. ein Skalenwert im Wertebereich 0 bis 100 pro Person vorliegt (Döring 2013, 108). Psychometrische Skalen verfügen über geprüfte Testgütekriterien. Messtheoretisch sind sie somit Einzelitems generell überlegen. Jedoch verlängern sie Fragebögen in der Regel deutlich. Personenbezogene Angaben werden sowohl über Einzelitems als auch über Skalen erhoben –

kaum etablierte Skalen vor, die dem gesellschaftlichen Wandel mit Blick auf Geschlechterrollen Rechnung tragen.[10]

Damit sind Schwierigkeiten der Operationalisierung des scheinbar einfach zu erhebenden Merkmals Geschlecht umrissen. Ein Blick in die Entwicklung der Erhebung der Variable Geschlecht im Rahmen der bereits erwähnten Demographischen Standards zeigt, dass in der Ausgabe von 1999 (7) kein Kommentar zur Variable zu finden ist und diese dichotom erfasst wird (1. Zeile, Tabelle 2). 2010 wird auf die Notwendigkeit der Erhebung hingewiesen: „Ohne eine Unterscheidung nach Geschlecht sind die Lebensbedingungen, Verhaltensweisen, Einstellungen von Frauen und Männern soziologisch, wirtschaftlich oder psychologisch nicht zu analysieren." (2010, 8) Ferner wird angemerkt: „Auf Intersexualität oder Transsexualität wird nicht eingegangen. Hierzu gibt es keine besonderen Empfehlungen." (2010, 8) Dies wird 2016 folgendermaßen erweitert:

> Auf Intersexualität oder Transsexualität wird nicht eingegangen, obwohl ‚intersexuell' als drittes, in Umfragen zu berücksichtigendes Geschlecht in der Diskussion ist. In die Demographischen Standards wird diese Kategorie jedoch erst eingeführt, wenn sie auch in der Referenzstatistik – dem Mikrozensus – berücksichtigt wird. (2016, 8f.)

Mit dem Beschluss des Bundesverfassungsgerichts vom 10. Oktober 2017 (Bundesverfassungsgericht 2017) wurde eine Neuregelung des binären Personenstands bis zum 31.08.2018 notwendig. Da der Zuordnung zu einem Geschlecht für die individuelle Identität eine herausragende Bedeutung zukommt und das allgemeine Persönlichkeitsrecht auch die geschlechtliche Identität schützt, stellt der Geschlechtseintrag, der vom Personenstandsrecht verlangt wird, einen Eingriff in das Persönlichkeitsrecht dar, wenn Menschen sich weder dem männlichen noch dem weiblichen Geschlecht dauerhaft zuordnen. Auch die Alternative „keine Angabe" ist hier problematisch, wenn Menschen nach eigenem Empfinden ein Geschlecht jenseits von männlich oder weiblich haben. Da ferner das Grundgesetz nicht festlegt, dass der Personenstand hinsichtlich des Geschlechts ausschließlich binär zu regeln ist, durch eine dritte einheitliche Bezeichnung ferner keine zusätzlichen Zuordnungsprobleme entstehen, sind entsprechende Änderungen vom Ge-

---

meist sind beide Varianten möglich (vgl. Diskussion zur Erfassung des Geschlechts in Teilkapitel 3.1).

[10] Exemplarisch sei zur Bestimmung des sozialen Geschlechts das Bem-Sex-Inventory (Bem 1981, 1974: deutschsprachige Fassung von Schneider-Düker/Kohler 1988) angeführt. Im Rahmen der Befragung erhalten die Teilnehmerinnen und Teilnehmer insgesamt 60 Eigenschaften vorgelegt, die als maskulin (hat Führungseigenschaften), als feminin (romantisch) oder sozial erwünscht (gesellig) eingestuft sind. Die Befragten schätzen sich diesbezüglich auf einer siebenstufigen Skala (von „die Eigenschaft trifft nie zu" bis zu „die Eigenschaft trifft immer zu") ein (http://www.txkoeln.de/infothek/lexikon/bsri.htm (letzter Zugriff: 01.08.2018)). Als Ergebnis erfolgt die Zuordnung zu einer von vier Gruppen: maskulin, feminin, androgyn oder undifferenziert. Die Beispieleigenschaften zeigen bereits, dass das Bem-Sex-Inventory anfällig für gesellschaftlichen und historischen Wandel diesbezüglich ist.

setzgeber vorzunehmen. Neben der Möglichkeit, auf den personenstandsrechtlichen Geschlechtseintrag generell zu verzichten besteht die Möglichkeit, eine weitere positive Bezeichnung eines Geschlechts einzuführen, wobei die konkrete Bezeichnung dem Gesetzgeber überlassen bleibt (Bundesverfassungsgericht 2017). Es ist zu erwarten, dass sich diese gesetzliche Neuerung auch generell auf die Erfassung des Merkmals Geschlecht auswirken wird.

3.2 Migrationshintergrund, Staatsangehörigkeit und Mehrsprachigkeit

Auch der Migrationshintergrund ist eine Frage der Definition (Kemper 2010). Ziel war es mit Hilfe dieses Konstrukts, „den migrationsbedingten demografischen Wandel genauer zu erfassen" (Kemper 2010, 316), da das Merkmal Staatsangehörigkeit dies nicht vermochte. Beispielhaft seien zunächst die Definitionen des BAMF und des Statistischen Bundesamtes zitiert:

> Eine Person hat dann einen Migrationshintergrund, wenn sie selbst oder mindestens ein Elternteil nicht mit deutscher Staatsangehörigkeit geboren ist. (BAMF 2018)

> Zu dieser Bevölkerungsgruppe zählen im Mikrozensus alle seit 1950 nach Deutschland Zugewanderten und alle im Inland mit fremder Staatsangehörigkeit Geborenen sowie die hier geborenen Deutschen, die mit zumindest einem Elternteil im selben Haushalt leben, der zugewandert ist oder als Ausländer in Deutschland geboren wurde. (Statistisches Bundesamt 2011)

Kemper (2010, 315f.) zeigt auf, dass sowohl in den amtlichen Statistiken als auch in der Bildungsforschung unterschiedliche Merkmale (Staatsangehörigkeit, Geburtsort/-land der Eltern, ggf. auch Großeltern, Zuwanderungsalter/Datum der Zuwanderung sowie zum Teil Religionszugehörigkeit) miteinander kombiniert werden, um den Migrationshintergrund zu erfassen und stellt Probleme dar, die sich u.a. mit Blick auf die Vergleichbarkeit von Ergebnissen infolge unterschiedlicher Definitionen ergeben.

Wird das Geburtsland als relevante Größe angesehen, so wird einer Person in der Regel dann ein Migrationshintergrund zugeschrieben, wenn mindestens eines ihrer Elternteile und/oder die Person selbst im Ausland geboren ist. Die nachfolgenden Abbildungen stellen zwei Varianten zur Erhebung der Geburtsländer der befragten Kinder bzw. Jugendlichen sowie von deren Eltern dar.

6. Sind du und deine Eltern in Deutschland geboren?

Mach bitte in jeder Zeile ein Kreuz!

	Nein	Ja	Weiß ich nicht
Ist dein **Vater** in Deutschland geboren?	☐	☐	☐
Ist deine **Mutter** in Deutschland geboren?	☐	☐	☐
Bist **du** in Deutschland geboren?	☐	☐	☐

Abb. 1: Geburtsland im Rahmen des StEG-Projekts (StEG o.J.)[11]

3. In welchem Land…

…bist du geboren? ☐☐☐ ○ Das weiß ich nicht.

…ist deine Mutter geboren? ☐☐☐ ○ Das weiß ich nicht.

…ist dein Vater geboren? ☐☐☐ ○ Das weiß ich nicht.

Abb. 2: Geburtsland im Rahmen des MaTS-Projekts (Maak/Zippel/Ahrenholz 2013, 113)

Beide Varianten ermöglichen Auswertungen hinsichtlich der Frage, ob das Kind bzw. der/die Jugendliche sowie die Eltern in Deutschland geboren sind oder nicht. Ferner kann auf den Migrationshintergrund geschlossen werden, wenn die Definition zugrunde gelegt wird, dass eine Person einen Migrationshintergrund hat, wenn sie selbst und/oder mindestens ein Elternteil nicht in Deutschland geboren ist. In beiden Fällen besteht die Möglichkeit „Das weiß ich nicht." bzw. „Weiß ich nicht" anzugeben. Das im MaTS-Projekt verwendete Antwortformat erschließt sich Außenstehenden jedoch nicht unmittelbar. So ist eventuell unklar, wozu die Abbildung des Teddybären oder aber die drei Kästen dienen. Hierzu sind spezifische Informationen zur Studie bzw. zur Erhebung notwendig. Der Fragebogen ist für Grundschülerinnen und Grundschüler gedacht, die sowohl hinsichtlich des Lesens der Fragen als auch des Schreibens von (längeren) Antworten überfordert sein könnten. Daher dient der Teddybär der Orientierung, denn die Erhebung wurde in der Regel von zwei bis drei geschulten Personen im Klassenplenum durchgeführt. Die drei Kästen dienen den Anfangsbuchstaben des jeweiligen Landes, z.B. DEU für Deutschland. Dieses System wurde mit den Schülerinnen und Schülern zu Beginn besprochen und geübt. Somit können im MaTS-Projekt zu den Geburtsländern wesentlich spezifischere Angaben gemacht werden, was auch Ziel dieses Projekts gewesen ist. Diese Beispiele verdeutlichen, dass selbst bei ähnlicher oder gar gleicher Definition auch die Art und Weise der Frage- bzw. Itemformulierung sowie generell Antwortformate und das Layout des

[11] Anzumerken ist, dass im StEG-Fragebogen von 2012 lediglich danach gefragt wird, ob der Schüler/die Schülerin selbst in Deutschland geboren ist oder nicht (vgl. https://www.projekt-steg.de/sites/default/files/uploads/StEG_A_Ansichtsexemplar_SFB_0.pdf (letzter Zugriff: 01.08.2018)).

Fragebogens – abhängig auch von den Adressatinnen und Adressaten des Fragebogens – durchaus stark variieren können.

Bereits der Terminus Migrationshintergrund ist folglich nicht eindeutig. Erschwerend kommt hinzu, dass ursprünglich nicht-deutsche Staatsangehörigkeit, Migrationshintergrund, Deutsch als Zweitsprache und Mehrsprachigkeit gleichgesetzt zu werden schienen. So wurde der Migrationshintergrund beispielsweise in Nordrhein-Westfalen als Grundlage bzw. Indikator für die Berechnung des Sprachförderbedarfs von Schülerinnen und Schülern genutzt – in der Annahme, dass ein (Groß)Teil der Schülerinnen und Schüler Deutsch als zweite Sprache gelernt hat und über sprachliche Defizite verfügt:

> Das nordrhein-westfälische Landesamt für Datenverarbeitung und Statistik etwa erhebt die Staatsangehörigkeit und den Status „Aussiedler" – die Schuladministration leitet daraus z.T. den Bedarf an muttersprachlichem Unterricht oder fördernden Maßnahmen in Deutsch als Zweitsprache ab. (Chlosta/Ostermann 2005, 55)

Ferner wurde teilweise davon ausgegangen, dass Schülerinnen und Schüler, die über einen Migrationshintergrund verfügen, auch mehrsprachig sind in dem Sinne, dass sie zu Hause neben bzw. außer Deutsch eine weitere Sprache sprechen. Jedoch belegten Untersuchungen wie SPREEG (Chlosta/Ostermann 2005, 60) und MaTS (Ahrenholz/Maak 2013), dass es Kinder mit Migrationshintergrund gibt, die keine andere Sprache als Deutsch in der Familie sprechen und andererseits mehrsprachige Kinder, die keinen Migrationshintergrund haben. So wird der Aspekt des Sprachverlusts häufig nicht berücksichtigt. Ähnlich wie in anderen Studien konnte z.B. für die Selbsteinschätzung der Schülerinnen und Schüler im Rahmen des MaTS-Projekts gezeigt werden, dass die Kompetenzen in der Herkunftssprache tendenziell geringer eingeschätzt werden als im Deutschen, wobei die Fähigkeiten in der Herkunftssprache im Mündlichen besser als im Schriftlichen eingeschätzt werden (Ahrenholz/Maak 2013, 60ff.).

Außerdem zeigt sich, dass der Terminus Mehrsprachigkeit nicht eindeutig verwendet wird. So kann man davon ausgehen, dass allein aufgrund des Fremdsprachenunterrichts an deutschen Schulen spätestens in der Sekundarstufe I alle Schülerinnen und Schüler zumindest in gewissem Maße mehrsprachig sind. Wenn von ‚mehrsprachigen' Kindern gesprochen wird, ist jedoch oft eine spezifische Form, nämlich die migrationsbedingte oder lebensweltliche Mehrsprachigkeit gemeint. Um migrationsbedingte Mehrsprachigkeit handelt es sich, wenn der Spracherwerb unmittelbare Folge einer Migration ist. Lebensweltliche Mehrsprachigkeit wird in der Regel nicht näher definiert. Allerdings verweist sie in vielen Publikationen auf mehrsprachige Schülerinnen und Schüler in Deutschland, die zu Hause bzw. im privaten Umfeld (auch) andere Sprachen als Deutsch sprechen. Ganz allgemein könnte darunter aber auch der alltägliche Kontakt mit anderen Sprachen verstanden werden; etwa wenn in Grenzregionen mehrere Sprachen verwendet werden wie z.B. das Dänische an der norddeutschen Grenze zu Dänemark. In der Regel wird für lebensweltliche Mehrsprachigkeit eine „natürliche" – also nicht

künstlich geschaffene – Kontaktsituation, einhergehend mit hoher kommunikativer Relevanz angenommen (Maak 2018). So wurden im Rahmen des MaTS-Projekts in Anlehnung an Chlosta/Ostermann (2010, 19) zunächst all diejenigen Schülerinnen und Schüler als mehrsprachig definiert, die „angeben, mit mindestens einem Familienmitglied (auch) eine weitere Sprache außer Deutsch zu sprechen (Ahrenholz/Maak 2013, 30). Die Analyse etwa für die Daten aus Erfurter Grundschulen ergab zunächst mit 20,5% aller Befragten eine viel größere Anzahl an Schülerinnen und Schülern, die dementsprechend als mehrsprachig zu definieren wären, als vorab angenommen worden war. Auffällig häufig wurde Englisch angegeben. Aufgrund der Zuwanderungsgeschichte des Bundeslandes Thüringen kann jedoch lediglich für einen kleinen Anteil der Englisch-Nennungen angenommen werden, dass dieses eine zentrale Rolle im Rahmen der familiären Alltagskommunikation spielt (Ahrenholz/Maak 2013, 31). Vielmehr ist anzunehmen, dass die Grundschülerinnen und Grundschüler davon ausgegangen sind, dass hier auch punktuelle Verwendung des Englischen anzugeben ist. Auf der anderen Seite zeigen die Untersuchungen von Brizić (2007, 2006), dass mehrsprachige Schülerinnen und Schüler bei Befragungen nicht angeben, dass sie zu Hause andere Sprachen als Deutsch sprechen oder gar andere Sprachen angeben als sie tatsächlich sprechen (z.B. Türkisch statt Kurdisch). Die vorgestellten und diskutierten Beispiele zeigen u.a., welche Schwierigkeiten sich bei der konkreten Itemerstellung ergeben und welche Folgen Formulierungen haben, die von den Adressatinnen und Adressaten nicht bzw. anders verstanden werden als von den Forscherinnen und Forscher gedacht.

Zusammenfassend ist bezüglich der Erfassung von Migrationshintergrund, Staatsangehörigkeit und Mehrsprachigkeit anzumerken, dass insbesondere für die Gruppen vergleichende Auswertung von Daten entscheidend ist, welche Operationalisierung zugrunde gelegt wird. Je nachdem, ob nicht-deutsche und deutsche Schülerinnen und Schüler, Schülerinnen und Schüler mit Migrationshintergrund und ohne Migrationshintergrund oder mehrsprachige und nicht mehrsprachige Schülerinnen und Schüler gegenübergestellt werden, können sich sehr unterschiedliche Gruppen(größen) ergeben und damit unterschiedliche Ergebnisse erzielt werden.

4. Fazit

An dieser Stelle soll noch einmal dafür plädiert werden, dass auch personenbezogene Daten nicht *en passant*, sondern *bien conçue* erhoben werden. Für Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler stehen in diesem Zusammenhang zahlreiche Vorschläge zur Verfügung, von denen im Beitrag auch einige vorgestellt worden sind. Die Orientierung an Standards ist so gut wie immer spontanen und eigenen Itemformulierungen sowie -gestaltungen überlegen. Aber auch vorhandene Standards sollten nicht unkritisch und unreflektiert übernommen werden. Darüber hinaus sind aktuelle Entwicklungen stets zu berücksichtigen – wie auch die veränderte Gesetzeslage bzgl. des Geschlechts belegt.

Anhand der ausgewählten Merkmale *Geschlecht*, *Migrationshintergrund* sowie *Mehrsprachigkeit* wurde exemplarisch aufgezeigt, dass nicht nur die bewusste Auswahl von zu erhebenden personenbezogenen Merkmalen von Bedeutung ist, sondern auch die konkrete Ausgestaltung der Erhebung sorgfältig umzusetzen ist. Für die Vorbereitung eigener Forschungsvorhaben bietet es sich an, auch vom Ziel her zu denken: Welche Auswertungen sollen am Ende vorgenommen werden? Welche personenbezogenen Merkmale sollten also erhoben werden – und mit welcher Genauigkeit soll dies erfolgen?

Literatur

Ahrenholz, Bernt/Maak, Diana (2013): Zur Situation von SchülerInnen nicht-deutscher Herkunftssprache in Thüringen unter besonderer Berücksichtigung von Seiteneinsteigern. Abschlussbericht zum Projekt ‚Mehrsprachigkeit an Thüringer Schulen (MaTS)', durchgeführt im Auftrag des TMBWK. Unter Mitarbeit von Fuchs, Isabel/Hövelbrinks, Britta/Ricart Brede, Julia/Zippel, Wolfgang. http://www.daz-portal.de/images/Berichte/bm_band_01_mats_bericht_20130618_final.pdf (letzter Zugriff: 01.08.2018).

Backhaus, Klaus/Erichson, Bernd/Wulff, Plinke/Weiber, Rolf (2008): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Berlin: Springer.

BAMF (2018): Migrationshintergrund (Definition), https://www.bamf.de/DE/Service¬/Left/¬Glossary¬/function-/glossar.html?lv3=3198544 (letzter Zugriff: 01.08.2018).

BDSG (Bundesdatenschutzgesetz): online unter https://www.gesetze-im-internet.de-/bdsg_1990/inhalts_bersicht.html (letzter Zugriff: 01.08.2018).

Bem, Sandra L. (1974): The measurement of psychological androgyny. Journal of Consulting and Clinical Psychology, 42, 155-62.

Bem, Sandra L. (1981): Gender schema theory: A cognitive account of sex typing source. Psychological Review, 88, 4, 354-364.

Bortz, Jürgen/Döring, Nicola (2006): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 4., überarb. Aufl. Heidelberg: Springer.

Brizić, Katharina (2007): Das geheime Leben der Sprachen. Gesprochene und verschwiegene Sprachen und ihr Einfluss auf den Spracherwerb in der Migration. Münster: Waxmann.

Brizić, Katharina (2006): Das geheime Leben der Sprachen. Eine unentdeckte migrantische Bildungsressource. In: Kurswechsel, 2, 32-43.

Bundesverfassungsgericht (2017): Personenstandsrecht muss weiteren positiven Geschlechtseintrag zulassen. Pressemitteilung Nr. 95 vom 08. November 2017. Beschluss vom 10. Oktober 2017, https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/DE/2017/bvg17-095.html (letzter Zugriff: 01.08.2018).

Chlosta, Christoph/Ostermann, Torsten (2010): Grunddaten zur Mehrsprachigkeit im deutschen Bildungssystem. In: Ahrenholz, Bernt/Oomen-Welke, Ingelore (Hrsg.): Deutsch als Zweitsprache. Baltmannsweiler: Schneider Hohengehren, 17-30.

Cholsta, Christoph/Ostermann, Torsten (2005): Warum fragt man nach der Herkunft, wenn man die Sprache meint? Ein Plädoyer für eine Aufnahme sprachbezogener Fragen in demografische Untersuchungen In: Bundesministerium für Bildung und Forschung (Hrsg.): Migrationshintergrund von Kindern und Jugendlichen: Wege zur Weiterentwicklung der amtlichen Statistik, 55-65.

Chlosta, Christoph/Ostermann, Torsten (2006): Zur Gestaltung und Begleitung einer fragebogengestützten Erhebung bei Grundschulkindern. In: Ahrenholz, Bernt/Apeltauer, Ernst (Hrsg.): Zweitspracherwerb und curriculare Dimensionen. Empirische Untersuchungen zum Deutschlernen in Kindergarten und Grundschule. Tübingen: Stauffenburg, 55-72.

Chlosta, Christoph/Ostermann, Torsten/Schroeder Christoph (2003): Die ‚Durchschnittsschule' und ihre Sprachen: Ergebnisse des Projekts Sprachenerhebung Essener Grundschulen (SPREEG). EliSe: Essener Linguistische Skripte, 3, 1, 43-139.

Decker, Yvonne/Schnitzer, Katja (2012): FreiSprachen – Eine flächendeckende Erhebung der Sprachenvielfalt an Freiburger Grundschulen. In: Ahrenholz, Bernt/Knapp, Werner (Hrsg.): Sprachstand erheben – Spracherwerb erforschen. Beiträge aus dem 6. Workshop ‚Kinder mit Migrationshintergrund', 2010. Stuttgart: Fillibach bei Klett, 95-112.

DaQS (2018): Datenbank zur Qualität von Schule. Online unter: http://daqs.fachportal-paedagogik.de (letzter Zugriff: 01.08.2018).

Döring, Nicola (2013): Zur Operationalisierung von Geschlecht im Fragebogen: Probleme, Lösungsansätze aus Sicht von Mess-, Umfrage-, Gender- und Queer-Theorie. In: Gender, 2, 94-113.

Extra, Guus/Aarts, Rian/van der Avoird, Tim/Broeder, Peter/Yaĝmur, Kutlay (2001): Meertaligheid in Den Haag: De status van allochtone talen thuis en op school. Amsterdam: European Cultural Foundation.

Fürstenau, Sara/Gogolin, Ingrid/Yaĝmur, Kutlay (2003): Mehrsprachigkeit in Hamburg. Ergebnisse einer Sprachenerhebung an den Grundschulen in Hamburg. Münster: Waxmann.

Grunert, Cathleen/Krüger, Heinz-Hermann. (2006): Kindheit und Kindheitsforschung in Deutschland – Forschungszugänge und Lebenslagen. Opladen: Budrich.

Hertel, Silke/Hochweber, Jan/Mildner, Dorothea/Steinert, Brigitte/Jude, Nina (2014): PISA 2009 Skalenhandbuch. Münster: Waxmann.

Kemper, Thomas (2010): Migrationshintergrund – eine Frage der Definition! In: Die deutsche Schule, 102, 4, 315-326.

Kränzl-Nagl, Renate/Wilk, Lieselotte (2000): Möglichkeiten und Grenzen standardisierter Befragungen unter besonderer Berücksichtigung der Faktoren soziale und personale Wünschbarkeit. In: Heinzel, Friederike (Hrsg.): Methoden der Kindheitsforschung. Ein Überblick über Forschungszugänge zur kindlichen Perspektive. Weinheim: Juventa, 59-75.

Lange, Andreas/Mierendorff, Johanna (2009): Methoden der Kindheitsforschung. Überlegungen zur kindheitssoziologischen Perspektive. In: Honig, Michael-Sebastian. (Hrsg.): Ordnungen der Kindheit. Problemstellungen und Perspektiven der Kindheitsforschung. Weinheim: Juventa, 183-210.

Lipski, Jens (2000): Zur Verläßlichkeit der Angaben von Kindern bei standardisierten Befragungen In: Heinzel, Friederike (Hrsg.): Methoden der Kindheitsforschung. Ein Überblick über Forschungszugänge zur kindlichen Perspektive. Gefälligkeitsübersetzung: Childhood research methods. An overview of research access points to the child's perspective. Weinheim: Juventa, 77-86.

Maak, Diana (2018): „Manchmal ist viel auch besser!“ Nutzung von mehrsprachigen Produktbeschreibungen in Schule und Hochschule zur Auseinandersetzung mit Mehrsprachigkeit. In: Maak, Diana/Ricart Brede, Julia (Hrsg.): Wissen, Können, Wollen – sollen?! (Angehende) LehrerInnen und äußere Mehrsprachigkeit. Münster: Waxmann, 197-230.

Maak, Diana (2018): Sprachliche Merkmale des fachlichen Inputs im Fachunterricht Biologie. Eine konzeptorientierte Analyse der Enkodierung von Bewegung. (= Deutsch als Zweitsprache, Mehrsprachigkeit und Migration 14). Dissertationsschrift. Berlin/Boston: De Gruyter.

Maak, Diana/Zippel, Wolfgang/Ahrenholz, Bernt (2013): ‚Manche fragen wahren schwer aber sonst war es okey‘ – Methodische Aspekte der Befragung von GrundschülerInnen am Beispiel des Projekts Mehrsprachigkeit an Thüringer Schulen (MaTS). In: Decker-Ernst, Yvonne/Oomen-Welke, Ingelore (Hrsg.): Deutsch als Zweitsprache: Beiträge zur durchgängigen Sprachbildung. Stuttgart: Fillibach, 95-118.

Porst, Rolf (2009): Der Fragebogen. Wiesbaden: VS Verlag.

Quellenberg, Holger (2009): Studie zur Entwicklung von Ganztagsschulen (StEG). Ausgewählte Hintergrundvariablen, Skalen und Indices der ersten Erhebungswelle. In Zusammenarbeit mit dem StEG-Konsortium und den Mitarbeiter/innen des StEG-Teams. In: Materialien zur Bildungsforschung, 24, Frankfurt a.M.: DIPF u.a.

Rössler, Patrick (2011): Skalenhandbuch Kommunikationswissenschaft. Wiesbaden: VS Verlag.

Schmelter, Lars (2014): Gütekriterien In: Settinieri, Julia/Demirkaya, Sevilen/Feldmeier, Alexis/Gültekin-Karakoç, Nazan/Riemer, Claudia (Hrsg.): Empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache. Eine Einführung. Paderborn: Schönigh, 33-45.

Schneider-Düker, Marianne/Kohler, André (1988): Die Erfassung von Geschlechtsrollen – Ergebnisse zur deutschen Neukonstruktion des Bem Sex-Role-Inventory. Diagnostica, 34, 3, 256-270.

Schnell, Rainer/Hill, Paul B./Esser, Elke (2005): Methoden der empirischen Sozialforschung. 7., völlig überarb. u. erw. Aufl. München/Wien: Oldenbourg.

Statistisches Bundesamt (2016): Demographische Standards. Ausgabe 2016, https://www.destatis.de/DE/Publikationen/StatistikWissenschaft/Band17_DemographischeStandards1030817169004.pdf?__blob=publicationFile (letzter Zugriff: 01.08.2018).

Statistisches Bundesamt (2011): Ein Fünftel der Bevölkerung in Deutschland hatte 2010 einen Migrationshintergrund, https://www.presseportal.de/pm/32102/2118662 (letzter Zugriff: 01.08.2018).

Statistisches Bundesamt (Hrsg.) (2010): Demographische Standards. Ausgabe 2010. Statistik und Wissenschaft, Band 17, https://www.destatis.de/DE/Publikationen/StatistikWissenschaft/Band17_DemographischeStandards1030817109004.pdf?__blob=publicationFile (letzter Zugriff: 01.08.2018).

Statistisches Bundesamt (1999): Demografische Standards. Ausgabe 1999. Methoden – Verfahren – Entwicklungen. Materialien und Berichte.

StEG (oJ): Fragebogen für Schülerinnen und Schüler der Jahrgangsstufe 5, https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&cad=rja&uact=8&ved=0ahUKEwjauP3NrKjZAhXEIVAKHYzfC3gQFghAMAQ&url=https%3A%2F%2Fdaqs.fachportal-paedagogik.de%2Fdownload%2Findex%2Ffile_id%2F15&usg=AOvVaw2YqWSG1lXk_Cv-8mxOztql (letzter Zugriff: 01.08.2018).

Steinke, Ines (2006): Gütekriterien qualitativer Forschung In: Flick, Uwe/Kardoff, Ernst von/Steinke, Ines (Hrsg.): Qualitative Forschung. Ein Handbuch. 5. Aufl. Reinbek: Rowohlt, 319-331.

Walper, Sabine/Tippelt, Rudolph. (2010): Methoden und Ergebnisse der quantitativen Kindheits- und Jugendforschung. In: Krüger, Heinz-Hermann/Grunert, Cathleen (Hrsg.): Handbuch Kindheits- und Jugendforschung. Wiesbaden: VS Verlag für Sozialwissenschaften, 205-243.

CHRISTOPH BRÄUER & JESSICA-CATHERINE VAUPEL

Wissenschaftsethos und Forschungsethik

Implikationen für eine empirische Deutschdidaktik[1]

1. Forschungsethik in der Deutschdidaktik

Die Deutschdidaktik kann sicher nicht mehr als junge, aber doch noch als eine sich konsolidierende wissenschaftliche Disziplin bezeichnet werden: Als solche findet und festigt sie sich in Hinblick auf ihr fachlich-professionelles Gegenstandsfeld. Als Disziplin, die auch empirisch forscht, verständigt sie sich über ihre theoretischen und methodologischen Grundlagen, diskutiert ihre wissenschaftsethischen Prinzipien, ihre wissenschaftlichen Methoden und Praktiken. Eine Debatte über forschungsethische Verantwortung in der Deutschdidaktik steht dagegen noch weitgehend aus und soll aus diesem Grund durch den vorliegenden Beitrag angestoßen werden. Da sich die Deutschdidaktik in ihrer empirischen Forschung in hohem Maße auf sozialwissenschaftliche Methoden und Praktiken bezieht, teilt sie mit den Sozialwissenschaften auch deren ethische Probleme einer Forschung an und mit Menschen. Im Anschluss an die definitorische Schärfung grundlegender Begrifflichkeiten skizziert der Beitrag entsprechend die forschungsethische Debatte in den Sozialwissenschaften, bevor er deren Herausforderungen auf die deutschdidaktische Forschung überträgt. Exemplarisch wird dabei auf Unterrichtsvideografie in der empirischen Professions- und Unterrichtsforschung fokussiert, da sich die ganze Bandbreite ethischer Herausforderungen hier besonders anschaulich beschreiben lässt. Es sei jedoch ausdrücklich darauf hingewiesen, dass jede Methode der Erforschung sozialer Prozesse ethische Dimensionen beinhaltet, die es zu berücksichtigen gilt. Die diskutierten Herausforderungen lassen sich folglich auch auf andere deutschdidaktische Untersuchungsfelder, etwa die Erforschung von Lehr-Lern-Prozessen, übertragen. Während der

[1] Wir möchten Irene Pieper sehr herzlich für ihre weiterführenden und einschlägigen Anregungen danken.
Dieser Beitrag ist im Rahmen des Teilprojekts *Reflexionskompetenzen und Videografie* des Schlözer Programm Lehrerbildung unter der Leitung von Prof. Dr. Christoph Bräuer im Handlungsbereich *Lehrerkompetenzen entwickeln* entstanden. Das Schlözer Programm Lehrerbildung wird unter der Fördernummer 01JA1617 im Rahmen der gemeinsamen ›Qualitätsoffensive Lehrerbildung‹ von Bund und Ländern aus Mitteln des *Bundesministeriums für Bildung und Forschung* gefördert.

Beitrag das Ziel verfolgt, forschungsethische Fragestellungen aufzugreifen und in Ansätzen zu diskutieren, bleibt es die Aufgabe der Disziplin, diese Debatte weiterzuführen und in Forschung und Lehre zu integrieren.

2. Verantwortung und Freiheit in der Wissenschaft

Moderne Wissenschaft kann bei all ihrer Varianz als Methode und Praxis einer systematisch strukturierten und methodisch kontrollierten Wissensbildung beschrieben werden (Özmen 2015, 65-67). Sie entwickelt ihr Wissen über die menschliche Lebenswelt, indem sie deren natürliche und soziale Phänomene theorieförmig zu beschreiben, zu erklären und zu deuten sucht. Ihre Erkenntnisse beanspruchen im Rahmen ihrer Reichweite und Aussagekraft überpersönliche und objektive Gültigkeit. Schon diese knappe und allgemeine Beschreibung moderner Wissenschaft weist darauf hin, dass Wissenschaft in ihrem Handeln selbst auferlegten Regeln genügen muss, um zu allgemeingültigen Erkenntnissen zu gelangen. Sie steht hier in der Verantwortung gegenüber der *scientific community*. In der Wahl ihrer Erkenntnisgegenstände ist die Wissenschaft frei, und diese Freiheit gilt als unhintergehbare Voraussetzung und Grundlage moderner Wissenschaft. Inwieweit sie für ihre Erkenntnisse und deren Folgen zur Verantwortung gezogen werden kann, ist und bleibt jedoch seit ihren Anfängen ein kontrovers diskutiertes Problem.

„Probleme der Verantwortlichkeit“ lassen sich entsprechend als zentrale ethische Herausforderung der modernen Wissenschaft beschreiben (Lenk 1991, 54). Sie berühren die Grenzen von Wissenschaftsverantwortung wie die Grenzen von Wissenschaftsfreiheit gleichermaßen: Verantwortung umfasst sowohl die obligatorischen Pflichten als auch die meritorischen Pflichten, durch die verantwortliches Handeln erst zu einer verdienstlichen Mehrleistung wird (Heidbrink 2017, 5): „Damit gerät der Begriff der Verantwortung von vornherein in ein Spannungsverhältnis von Schuldigkeit und Zuständigkeit, von regelkonformem Verhalten und freiwilliger Selbstbindung“ (ebd.). Als ein relationaler Begriff verlangt Verantwortung ein Subjekt (die Wissenschaftlerin/den Wissenschaftler), das für ein Objekt (die Forschungshandlung, das Forschungsergebnis bzw. die Erkenntnis, die Erkenntnisfolgen) vor einer Instanz (der Forschungsgemeinschaft, der Allgemeinheit, der Justiz) gegenüber einem Adressaten (den unmittelbar oder mittelbar Betroffenen) auf der Grundlage normativer Kriterien (etwa Satzungs- oder Gebrauchsnormen) verantwortlich ist (Loh 2017, 53). Verantwortlichkeit bezeichnet also die Rechenschaftspflicht für (übertragene) Aufgaben und deren Folge für Einzelne oder die Allgemeinheit gegenüber einer Kontrollinstanz.

Hans Lenk schlägt vor diesem Hintergrund vor, Verantwortung in zwei Perspektiven zu diskutieren: In der ersten Perspektive ist nach der wissenschaftsinternen Verantwortung, in der zweiten nach der externen Verantwortung der Wissenschaftlerinnen und Wissenschaftler zu fragen (Lenk 1991, 56). Die interne Verantwortung der Wissenschaftlerinnen und Wissenschaftler gilt gegenüber der eigenen wissenschaftlichen Zunft als Kontrollinstanz und umfasst die Regeln guter

wissenschaftlicher Praxis, fairer Konkurrenz und bestmöglicher „Wahrheitssuche und -sicherung“ (ebd., 58). Sie wird auch als Wissenschaftsethos bezeichnet und liegt in vielen Teilen kodifiziert vor (siehe auch die Beiträge von Schmidt in diesem Band).

Die externe Verantwortung verweist nach Lenk zum einen auf die unmittelbaren Wirkungen des wissenschaftlichen Arbeitens auf die Betroffenen; zum anderen bezieht sich der Begriff auf die mittelbaren Folgen, die die gewonnen wissenschaftlichen Erkenntnisse für die Betroffenen wie für die Allgemeinheit haben. Sie kann als Forschungsethik bezeichnet werden, die in Hinblick auf die unmittelbaren Wirkungen auf Probandinnen und Probanden kaum umstritten und in Teilen ebenfalls kodifiziert ist (ebd., 57f.) (siehe auch den Beitrag von Iberer in diesem Band). In Hinblick auf die sozialen Folgen ihrer Erkenntnisse gegenüber der Allgemeinheit wird die Folgeverantwortung in den Sozial- und Geisteswissenschaften dagegen nach wie vor kontrovers diskutiert (Graumann 2006, 255).

Lenk fordert, dass die „interne (ethos) und die externe Perspektive (Ethik) [...] bei Verantwortlichkeitsfragen“ strikt unterschieden werden müssten, wenngleich diese Trennung analytisch erfolge und beide Perspektiven anschließend auch wieder aufeinander bezogen werden müssten (Lenk 1991, 58f.). Diese Unterscheidung wird im weiteren Verlauf des Beitrags aufgegriffen. Gerade in Hinblick auf die Probleme der gesellschaftlichen Folge(n)verantwortung berühren die forschungsethischen Fragen immer auch die Ein- und Beschränkung wissenschaftlicher (Forschungs-)Freiheit, treten ethische Fragen doch schnell in Konflikt mit dem wissenschaftsgrundlegenden Anspruch auf Forschungsfreiheit: Aufgrund einer forschungsethischen Folge(n)verantwortlichkeit könnten Erkenntnisse der Allgemeinheit vorenthalten werden (müssen) oder bestimmte Gegenstände erst gar nicht einer wissenschaftlichen Erforschung unterzogen werden (dürfen). Dieser Aspekt wird in seiner Problematik für die bildungswissenschaftliche und fachdidaktische Forschung in Kapitel 4 diskutiert. Andererseits lässt sich aus der Forderung nach freier Forschung oder der Berufung auf wertfreie Grundlagenforschung gerade auch für eine anwendungsorientierte empirische Forschung in der Deutschdidaktik keine „totale Neutralität“ ethischen Verpflichtungen gegenüber ableiten (Lenk 1991, 71; vgl. Maring 2011, 167). Sich unter Berufung auf das Forschersein jeglicher Verantwortungsdebatte oder gegebenenfalls einer Mitverantwortung entziehen zu wollen (Maring 2011, 167), erscheint in einem modernen Wissenschaftsverständnis hochproblematisch: Die Reichweite wissenschaftlicher Erkenntnisse beschränkt sich längst nicht mehr auf eine reine Anschauung, ein „zweckfreies Wissenwollen“ einer Bildungselite. Alle Wissenschaftsbereiche sind nicht nur zur praktischen Umsetzungsmöglichkeit theoretischen Wissens wie zur technischen Nutzbarmachung aufgefordert; sie zielen im wissenschaftlichen Wettbewerb um Anerkennung und Ressourcen letztlich selbst auf die Umsetzung, Anwendung und Nutzbarmachung ihres Wissens (Özmen 2015, 65).

Bevor in Kapitel 3 die Debatte in den Sozialwissenschaften um die forschungsethische Verantwortung und deren Grenzen skizziert wird, soll die eben eingeführte Differenzierung zwischen Wissenschaftsethos und Forschungsethik geschärft werden.

2.1 Wissenschaftsethos – „gute wissenschaftliche Praxis"

Das epistemische Ethos folgt einem breiten Konsens in der Wissenschaftsgemeinschaft und bildet „seit Beginn der Neuzeit die Grundlage für das Vertrauen, das die Wissenschaft und die Wissenschaftler in der Öffentlichkeit erfuhren" (Deutscher Ethikrat 2014, 58). Es beruht auf Selbstverpflichtung und Selbstkontrolle eines institutionalisierten Normengefüges, das Kriterien für Wissenschaftlichkeit und wissenschaftliches Arbeiten bestimmt und die Forschungspraxis funktional anleitet (Özmen 2015, 67). Kodifiziert finden sich die Hinweise zu „guter Wissenschaft" bzw. „guter wissenschaftlicher Praxis" in den Darstellungen zentraler Wissenschaftseinrichtungen sowie in den eigenen Verpflichtungen der Hochschulen und Universitäten (vgl. DFG 2013, HRK etc.): Hier gelten für die überpersonale, objektive Gültigkeit der Erkenntnisse Prinzipien der systematischen Widerspruchsfreiheit und der internen Kohärenz, der Reproduzierbarkeit, Klarheit und Genauigkeit sowie der begrifflichen Sparsamkeit und argumentativen Eleganz (Brendel 2011; Graumann 2006, 254) als unhintergehbar.

Der Wissenschaftssoziologe Robert K. Merton formulierte schon in den vierziger Jahren des vergangenen Jahrhunderts einen Normen-Kodex als Standesethos. Er verpflichtet die Wissenschaftsgemeinschaft auf vier Grundsätze: Das Streben nach Verallgemeinerbarkeit (‚Universalismus'), den systematischen Zweifel und die Begründungspflicht eigener Geltungsansprüche (‚organisierter Skeptizismus'), die Zurückstellung privater oder persönlicher Motive (‚Desinteressiertheit') und das Recht auf öffentlichen Zugang zum und jedermanns Teilhabe am wissenschaftlichen Wissen (‚communalism') (Deutscher Ethikrat 2014, 57; vgl. Lenk 1991, 56). Hans Mohr formulierte sogar konkrete Maximen: „Sei fair! Manipuliere nie die Daten! Sei präzise! Sei fair hinsichtlich der Priorität von Daten und Ideen deines Rivalen! Mache keine Kompromisse, sondern versuche ein Problem zu *lösen*!" (zit. nach Lenk 1991, 56; vgl. Maring 2011, 166).

Wissenschaftshistorische und -journalistische Recherchen, Betrugs- und Plagiatsfälle wie zuletzt die Diskussion um ‚Fake-Science' und die Veröffentlichung gegen Bezahlung (vgl. F&L 2018, 645) verweisen dabei auf Funktionsmängel der wissenschaftsinternen ethischen Kontrollmechanismen (schon Lenk 1991, 57). Sie liegen nicht nur in der schwierigen Kontrolle und Durchsetzung von Selbstverpflichtungen begründet. Die moderne Hochschulpolitik schafft durch ihr rigoroses Konkurrenzprinzip bei der Verteilung von Ressourcen, bei der Anerkennung von Forschungsleistungen oder der Verteilung von Teilhabemöglichkeiten (etwa die Bedeutung von Kennzahlen bei der Auswahl von (Nachwuchs-)Wissenschaftlerinnen und Wissenschaftler, durch die beispielsweise die Quantität ihrer Publikationen schnell mehr zählt als deren Qualität; vgl. Strohschneider et al.

2018, 668-670) auch Anreize, gegen das selbst auferlegte Standesethos zu verstoßen.

Vor diesem Hintergrund gilt es umso mehr, die Freiheit der Wissenschaft so weit als möglich aufrecht zu erhalten. Wissenschaftlerinnen und Wissenschaftler können nicht gänzlich frei von politischen und gesellschaftlichen Interessenslagen handeln, da sie ihnen als Individuen selbst unterliegen. Dennoch ist eine weitgehende Freiheit der Wissenschaft von hochschulpolitischen und gesellschaftlichen Interessenslagen notwendig (vgl. Özmen 2015, 67). Diese Freiheit hat nicht nur eine funktionale und epistemische Funktion, sondern auch eine ethische: Eine weitgehende Wissenschaftsautonomie eröffnet der Wissenschaft die Freiheit, sich an die selbst auferlegten ethischen Regeln zu halten. Allerdings darf die

> Verteidigung der Autonomie der Wissenschaft nicht mit der ganz anders gelagerten These der Autarkie der Wissenschaft verwechselt werden. Zwar folgt Wissenschaft ihren eigenen Zwecken [...], aber sie ist nicht ethisch neutral oder indifferent, ohne Verantwortung für ihre möglichen technischen Anwendungen und gesellschaftlichen Folgen. (ebd., 68)

Diese forschungsethische Perspektive soll im Weiteren skizziert werden.

2.2 Forschungsethik – gesellschaftliche Verantwortung

Externe Verantwortung im Sinne einer Rechenschaftspflicht für die möglichen Wirkungen und Folgen ihrer Forschung tragen die Wissenschaftlerinnen und Wissenschaftler gegenüber den unmittelbar wie mittelbar Betroffenen. Dazu können einzelne Beforschte, aber auch die *scientific community* oder die Allgemeinheit zählen. Die externe Verantwortung zeigt sich besonders deutlich in Forschungsarbeiten, bei denen unmittelbar Menschen zu Objekten der Forschung werden (Lenk 1991, 57f.; Maring 2011, 166). Gleichwohl bleibt die Reichweite der ethischen Verantwortung in der Wissenschaft umstritten, birgt jede Verantwortungsübernahme das Risiko der Gefährdung der Autonomie und der Verletzung des epistemischen Ethos (Maring 2011, 167). Diskutiert wird ebenfalls, wer die Kontrolle der Verantwortung wahrnehmen soll, ob sie der individuellen Wissenschaftlerin/dem individuellen Wissenschaftler und ihren/seinen ethischen Maßstäben übertragen wird, oder ob sie durch externe Akteure übernommen werden soll (Özmen 2015, 70). Beispiele für Letzteres sind etwa die Gründung des Deutschen Ethikrates im Jahr 2001 oder die Einrichtung von Ethikbeauftragten und -kommissionen. Als Adressat eines universellen Verantwortungsimperativs ist zunächst die wissenschaftliche Gemeinschaft angesprochen: „das Prinzip Verantwortung wird auf die Fähigkeit und den Willen zur verantwortlichen Selbstbeschränkung von Forschung und Wissenschaft gegründet“ (Özmen 2015, 70). Zur Stärkung der externen Verantwortung der einzelnen Wissenschaftlerinnen und Wissenschaftler wird oft eine Selbstverpflichtung vorgeschlagen (Lenk 1991, 59). So formuliert Hans Jonas (1979) als „Prinzip Verantwortung“ etwa:

> Handle so, dass die Wirkungen deiner Handlungen verträglich sind mit der Permanenz echten menschlichen Lebens auf Erden. Handle so, dass die Wirkungen deiner Handlung nicht zerstörerisch sind für die künftige Möglichkeit solchen

> Lebens. Gefährde nicht die Bedingungen für den indefiniten Fortbestand der Menschheit auf Erden. (Jonas 1979, 36)

Wie schon bei der wissenschaftsethischen Selbstverpflichtung zeigt sich jedoch auch hier, dass forschungsethische Verantwortungsimperative wenig wirksam, kontrollierbar und durchsetzbar sind: „Das Problem der ethischen Kontrolle ist durch einen Eid allein nicht zu lösen – zumal in das Karrieresystem der Wissenschaften Anreize zur Verletzung ethischer Normen geradezu eingebaut sind" (Lenk 1991, 59f.). Eine institutionelle Unterfütterung durch stützende Maßnahmen oder ideelle Sanktionen wiederum berge die Gefahr einer Verrechtlichung oder Quasiverrechtlichung der Moral (ebd., 60f.).

In der Debatte um eine verantwortungsbewusste Forschungsethik führen Positionen, die entweder eine einseitig deontologische[2] oder eine einseitig teleologisch-konsequentialistische Ethik vertreten, schnell in dilemmatische Situationen, in denen der Schutz grundlegender Werte gegen die Leistung wissenschaftlicher Erkenntnis ausgespielt werden (vgl. Özmen 2015, 72). Eine Alternative könnte in der Ausarbeitung einer diskursiv auszuhandelnden Forschungsethik bestehen, die Vorsorge- und Verantwortungsprinzipien verbindet. Um in ethischen Konflikten zu verantwortungsbewussten Lösungen zu gelangen, könnte die Wahrung deontologischer Prinzipien wie Selbstbestimmung oder Menschenwürde, Gemeinwohl oder auch Wissenschaftsfreiheit mit der Prüfung etablierter teleologisch-konsequenzialistischer Prinzipien wie Zweck-Mittel-Abwägung, Schadensvermeidung sowie Risikoabschätzung verbunden werden. Auf der Grundlage von Prioritätsregeln könnte sodann eine Lösung gefunden werden, die einen Ausgleich zwischen dem Anspruch auf wissenschaftlicher Erkenntnis und dem Schutz von Persönlichkeitsrechten, zwischen der möglichen Förderung des Gesamtwohls und der Minimierung von negativen Folgen und Risiken anstrebt (Lenk 1991, 60-66).

Bereits auf der theoretisch-konzeptionellen Ebene einer allgemeinen Forschungsethik zeigt sich ein Widerstreit der Argumentationen, die sich aus einem Spannungsverhältnis zwischen dem Streben nach ethisch verantwortungsvollem Forschertum, den Härten des ökonomisierten Wissenschaftsbetriebs und dem individuellen Erfolgswunsch ergibt. Dass diese Ambivalenz in der Forschungspraxis bestehen bleibt, auch wenn man sie durch die Einigung auf Ethikkodizes zu glätten versucht, soll im Folgenden anhand markanter Aspekte der sozialwissenschaftlichen Debatte um die Diskrepanz zwischen Forschungsideal und Forschungsrealität veranschaulicht werden.

---

[2] In deontologischen Positionen liegt der Beurteilung von Handlungen und Handlungsmotiven ein Pflichtgedanke zugrunde; die moralische Qualität bemisst sich danach, ob das Handeln einer normativen Verpflichtung folgt (z.B. dem unabdingbaren Schutz der Menschenwürde). Teleologisch-konsequentialistische Positionen betrachten die Handlungen als Mittel zum Zweck; Handlungen werden ausgehend von ihren (erwartbaren) Folgen moralisch beurteilt (z.B. der mögliche Nutzen überwiegt die möglichen Schäden) (vgl. Broad 1930; vgl. Prechtl 1996, 101).

3. Forschungsethik als Grundlage empirischer Forschung

Ein Blick auf die internationale Geschichte der medizinischen Forschung genügt, um sich insbesondere der schwerwiegenden externen Verantwortung von Wissenschaft bewusst zu werden. Die Tuskegee-Syphilis-Studie[3] (vgl. Jones 1993) oder die umstrittene SUPPORT-Studie an frühgeborenen Babys[4] (vgl. Cortés-Puch et al. 2016) zeigen eindrücklich, dass wissenschaftlichem Erkenntnisinteresse strikte moralische Grenzen gesetzt werden müssen, um die Grundrechte des Individuums zu schützen. Dies gilt jedoch nicht nur für die medizinische Forschung. Hella von Unger konstatiert für die Sozialwissenschaften, dass Experimente – anders als etwa in der Medizin – generell keine unmittelbar erkennbaren Folgen für das Leben der Beforschten hätten – aber „nur, weil mögliche Schädigungen nicht so offensichtlich Leben kosten, ist die sozialwissenschaftliche Forschung nicht davor gefeit, Menschen Schaden zuzufügen" (von Unger 2014, 16). Um solchen Schäden vorzubeugen, haben die Deutsche Gesellschaft für Soziologie (DGS) sowie der Berufsverband Deutscher Soziologinnen und Soziologen (BDS) Anfang der 1990er Jahre einen gemeinsamen Ethik-Kodex erstellt,[5] in dem die Merkmale intern- und extern-verantwortungsvollen Handelns für die soziologische und damit auch generell für die sozialwissenschaftliche Forschung expliziert werden. Auch die Deutsche Gesellschaft für Erziehungswissenschaften (DGfE) veröffentlichte 1999 ihr Verständnis von korrekter wissenschaftlicher Praxis in Form eines Ethik-Kodex, der 2006 um Richtlinien des ethisch sensiblen Umgangs mit qualitativen Daten ergänzt und 2010 nochmals aktualisiert wurde (DGfE 2010). In Deutschland, wo institutionell festgelegte Richtlinien und rege öffentliche Diskussionen vor allem im medizinisch-psychologisch orientierten Wissenschaftsspektrum Antworten auf ethische Fragen der Forschung am Menschen geben, ist ein Ethik-Kodex ein wichtiger Schritt, um auch Wissenschaftlerinnen und Wissenschaftler anderer Disziplinen eine Orientierungshilfe zu bieten. Dass das Streben nach einem ethischen Konsens jedoch auch immense Herausforderungen birgt, wird in der derzeitigen sozialwissenschaftlichen Debatte deut-

[3] Im Rahmen dieser rassistisch motivierten Studie führte der US-amerikanische Öffentliche Gesundheitsdienst von 1932 bis 1972 nicht-therapeutische Untersuchungen an über 400 afroamerikanischen Syphilispatienten durch, um den natürlichen Verlauf der unbehandelten Krankheit bei Schwarzen Männern dokumentieren zu können. Die Teilnehmer wurden weder über ihre Erkrankung und deren Folgen informiert, noch mit wirksamen Medikamenten behandelt; die Untersuchungen erfolgten unter dem Deckmantel der freien Gesundheitsversorgung.

[4] In dieser Studie wurden extrem frühgeborenen Babys zur Ermittlung der optimalen Sauerstoffdosis randomisiert hoch- oder niedrigdosierte Sauerstoffgaben verabreicht. Von einem ethischen Standpunkt wird kritisch diskutiert, ob die Eltern im Rahmen der informierten Einwilligung ausreichend über die Folgen (u.a. Tod, Erblindung) einer niedrigen bzw. hohen Sauerstoffgabe aufgeklärt wurden.

[5] Der Ethik-Kodex der DGS und des BDS wurde mit der Verabschiedung in beiden Verbänden am 10.06.2017 in einer aktualisierten Form in Kraft gesetzt.

lich (vgl. Miethe 2013; von Unger 2014; Hopf 2016), welche vor allem die Rigidität wissenschaftlicher Ethik-Kodizes kritisch reflektiert. Bevor auf die lösungsorientierte Diskussion der Probleme eingegangen wird, die sich bei einer strengen Befolgung der kodifizierten Grundsätze ergeben können, erfolgt zunächst eine kurze Darstellung der beitragsrelevanten Handlungsempfehlungen.

In der Präambel ihrer Ethik-Kodizes erklären es sowohl die DGfE als auch die DGS und der BDS zu einem zentralen Ziel, Forscherinnen und Forscher für ethische Probleme zu sensibilisieren und die eigene berufliche Praxis kritisch zu reflektieren. Um dieses Ziel zu verwirklichen, finden sich in den Kodizes Handlungsempfehlungen. Im Bereich des Berufsethos gelte es im Sinne einer internen Verantwortung, bei der Arbeit nach „wissenschaftlicher Integrität und Objektivität" (DGS/BDS 2017, §1,1) zu streben. Dabei solle in der erziehungswissenschaftlichen Forschung die „Präsentation oder Publikation soziologischer Erkenntnisse […] ohne verfälschende Auslassung von wichtigen Ergebnissen" (DGS/BDS 2017, §1,2) erfolgen. Darüber hinaus verweisen die Kodizes auf die soziale (externe) Verantwortung, die Wissenschaftlerinnen und Wissenschaftler tragen: „Ihre Empfehlungen, Entscheidungen und Aussagen können das Leben ihrer Mitmenschen beeinflussen" (DGfE 2010, §1,5; DGS/BDS 2017, §1,7). Es wird vor dem Missbrauch der potenziell einflussreichen Position wissenschaftlicher Akteurinnen und Akteure gewarnt und zugleich an den Weitblick der Forschenden appelliert, kritische Momente zu antizipieren, die zu unethischem Verhalten führen könnten (ebd.).

Im Umgang mit Probandinnen und Probanden werden Wissenschaftlerinnen und Wissenschaftler in den Ethik-Kodizes dazu angehalten, die gesetzlich verankerten Persönlichkeitsrechte der beforschten Individuen zu wahren (DGfE 2010, §4,1; DGS/BDS 2017, §2,2). Soziologinnen und Soziologen werden speziell darauf verwiesen, in ihrem Verhalten immer auch die übrige *scientific community* zu berücksichtigen, da „das Forschungshandeln den zukünftigen Zugang zu einer Untersuchungspopulation für den gesamten Berufsstand oder verwandte Berufsgruppen einschränken oder verschließen" könne (DGS/BDS 2017, §2,1). Zentraler Punkt bei der Arbeit mit Probandinnen und Probanden ist grundsätzlich die informierte Einwilligung: Teilnehmende müssen sorgfältig über die Ziele und Methoden des Forschungsvorhabens sowie über ihre Rechte aufgeklärt werden, wobei es in Bezug auf Individuen mit geringem Bildungsstand, niedrigem Sozialstatus oder sprachlichen Barrieren besonderer Anstrengungen bedürfen, um eine Zustimmung auf freiwilliger und informierter Basis zu ermöglichen (DGfE 2010, §4,2; DGS/BDS 2017, §2,4). Neben der Prämisse, die Anonymität (DGS/BDS 2017, §2,5) sowie die Integrität der Forschungsteilnehmenden zu wahren (DGfE 2010, §4,3), sind Wissenschaftlerinnen und Wissenschaftler dazu angehalten, über Risiken aufzuklären und die Individuen keinen Nachteilen oder gar Gefahren (DGS/BDS 2017, §2,5) auszusetzen.

Schon dieser Kurzüberblick macht deutlich, wie ernst das „Prinzip der Nicht-Schädigung" (Hopf 2016) genommen wird, zeigt aber auch, wie anfällig die verschiedenen Bereiche der sozialwissenschaftlichen Forschungsarbeit für (bewusst

oder unbewusst) ethisch fragwürdige Handlungen sein können. Zwar vermittelt der normative Empfehlungscharakter der Ethik-Kodizes den Eindruck, die Forschenden seien gegen unmoralisches Verhalten gefeit, wenn sie sich nur an die genannten Richtlinien hielten. Tatsächlich stehen ethische Grundsätze und methodisches Vorgehen in der sozialwissenschaftlichen Forschungsrealität jedoch oftmals in einem Spannungsverhältnis (vgl. von Unger 2014, 22). Den forschungsethischen Diskussionen im derzeitigen sozialwissenschaftlichen Diskurs folgend, ergeben sich forschungsethische Dilemmata vor allem im Bereich der informierten Einwilligung, der Anonymisierung von Forschungsdaten sowie bei der Rückmeldung und Publikation von wissenschaftlichen Erkenntnissen.

3.1 Prinzip der informierten Einwilligung

Das Prinzip der informierten Einwilligung (*informed consent*) birgt sowohl hinsichtlich der Informiertheit von Probandinnen und Probanden als auch in Bezug auf das erteilte Einverständnis moralisches Konfliktpotenzial. Zunächst stellt sich die definitorische Frage, ab wann eine Einwilligung als informiert bzw. ‚informiert genug' gilt. Zwar wird in den Ethik-Kodizes betont, dass Individuen mit Verstehensbarrieren besonderer Aufklärungsbemühungen bedürfen; jedoch stellt sich auch bei allen anderen Teilnehmenden die Frage, inwieweit sie die Methoden, Ziele und möglichen Auswirkungen eines Projekts verstehen: Der Erfolg eines kommunikativen Prozesses hängt nicht nur von der verständlichen Formulierung der Informationen ab, sondern auch davon, wie der Adressat diese Informationen vor dem Hintergrund seines Vorwissens, seiner Werte und Erfahrungen deutet. Verschiedene Formen des (Miss-)Verständnisses sowie resultierende Fehlinterpretationen und -vorstellungen sind folglich ein Risiko, das bei der Arbeit mit Menschen generell bedacht werden muss (vgl. Hopf 2016, 198).

Die Frage, ab wann eine Person informiert genug ist, um eine fundierte Entscheidung über die Teilnahme an einem Forschungsprojekt zu fällen, stellt sich darüber hinaus in Situationen, in denen sich die Fragestellungen auf sensible Erkenntnisse erstrecken. Christel Hopf führte beispielsweise eine Studie durch, in der sie mit ihrer Arbeitsgruppe die rechtsextreme Orientierung junger Frauen untersuchte (vgl. Hopf 2016, 198). Um negative Abwehrreaktionen der Teilnehmerinnen zu vermeiden, wurde bei der vorherigen Informierung aller Beteiligten auf den Gebrauch des Begriffs *rechtsextrem* verzichtet (vgl. Projektgruppe 1996, 89ff. zit. nach Hopf 2016, 198). Ist die Auslassung bzw. euphemistische Umschreibung eines anstößigen Wortes bereits eine vorsätzliche Täuschung und damit ethisch kritikabel oder wird sie durch das Produkt – den Zugang zu einer Gemeinschaft, deren Verhalten und Motivation von erheblichem gesellschaftspolitischen Interesse sind – ausreichend legitimiert?

Hinsichtlich einer ausführlichen Informierung ist zudem denkbar, dass die Probandinnen und Probanden wenig bis kein Interesse an Informationen zum Forschungsprojekt haben (vgl. Kelly 2005, 101). Um in diesem Fall eine fundiert informierte Einwilligung zu gewährleisten, wäre eine forcierte Aufklärung gegen

den Willen der Individuen nötig, was nicht nur die Teilnahmebereitschaft gefährden und die Persönlichkeitsrechte verletzen, sondern auch gegen die Maxime der Freiwilligkeit verstoßen würde. Kelly spricht sich hier für die Position aus, das Prinzip der informierten Einwilligung in Kombination mit anderen ethischen Überlegungen zu betrachten, statt es als ein „overriding principle" zu verstehen (ebd.). Doch selbst im Idealfall einer Einwilligung, die auf dem vollständigen Verständnis aller Studieninformationen basiert, besteht insbesondere in der Feldforschung das Restrisiko, dass die Probandinnen und Probanden im längerfristigen persönlichen Kontakt zu den Forschenden die sachorientierte Motivation der Zusammenarbeit ‚vergessen' (vgl. von Unger 2014, 27).

Neben der inhaltlich-zwischenmenschlichen Herausforderung einer informierten Einwilligung ergibt sich bei qualitativ-explorativen Methoden zusätzlich eine organisatorische Schwierigkeit, da es „aufgrund der größeren Flexibilität und der eingeschränkten Planbarkeit […] oft nicht möglich [ist], zu Beginn genau zu bestimmen, wie der Forschungsprozess verlaufen und zu welchen Resultaten er führen wird" (ebd., 26). Dies trifft darüber hinaus für die Nachnutzung (beispielsweise Sekundäranalysen) erhobener Daten zu, über die die Probanden nicht informiert werden und in Folge dessen auch nicht einwilligen konnten.

Die Lösung liegt möglicherweise in einem erweiterten Verständnis der informierten Einwilligung, welches speziell im anglophonen Ethikdiskurs vertreten wird: Dem Prinzip des *process consent* folgend, versichern sich die Wissenschaftlerinnen und Wissenschaftler durch Rücksprache in jeder Stufe des Forschungsprozesses, ob die Studienteilnehmenden weiterhin partizipieren möchten (vgl. Ellis 2007, 23).

Die mündliche Abfrage einer erneuerbaren Einwilligung bietet viele Vorteile. Dennoch ist eine einmalige schriftliche Erklärung des Individuums zur Teilnahme am Forschungsprojekt sinnvoll, da sie nicht nur eine praktische Möglichkeit darstellt, in größeren Untersuchungspopulationen Zustimmungen schnell und individuell zu prüfen; sie dient auch der rechtlichen Absicherung der Forschenden. Die Tatsache, dass diese Form der schriftlichen Einwilligung in der Regel durch eine Unterschrift und unter Angabe des vollständigen Namens erfolgt, führt jedoch zu einem neuen Dilemma, da sie die Teilnehmenden identifizierbar macht. Dies stellt strenggenommen einen Verstoß gegen die Maxime der Anonymität dar (vgl. Miethe 2013, 929).

3.2 Prinzip der Anonymisierung

Auch wenn das Streben nach anonymisierten Forschungsdaten ein wichtiges Instrument ist, um Individuen vor negativen Konsequenzen ihrer Untersuchungsteilnahme zu schützen, so stellt es die qualitativ-empirische Methodik vor erhebliche Herausforderungen. „Das Forschungsziel qualitativer Forschung besteht darin, die Prozesse zu rekonstruieren, durch die die soziale Wirklichkeit in ihrer sinnhaften Strukturierung hergestellt wird" (Lamnek/Krell 2016, 44) – und da diese Prozesse innerhalb individueller Sozialgefüge ergründet werden, ist die qualitative Forschung darauf angewiesen, gewonnene Erkenntnisse im Kontext der

untersuchten Gruppierung darzustellen. Die resultierende Gefahr, dass Probandinnen und Probanden durch Kontextualisierungen identifizierbar werden, wird durch die steigende Leistungsfähigkeit von Suchmaschinen (vgl. von Unger 2014, 25) sowie neueren technischen Entwicklungen wie Gesichtserkennung und Lokalisierungen zusätzlich verstärkt. Das Prinzip der Anonymisierung, welches in quantitativ-empirischen Ergebnisstatistiken praktikabel ist, zwingt die qualitative Forschung folglich dazu, grundlegende Informationen zurückzuhalten, die für eine intersubjektiv nachvollziehbare Präsentation der Ergebnisse möglicherweise unabdingbar sind. Hinzu kommt, dass der Ethik-Kodex der DGfE empfiehlt, „[g]rundsätzlich […] solche Verfahren […] [zu nutzen], die eine Identifizierung der Untersuchten ausschließen, also Anonymität gewährleisten“ (DGfE 2010, §4,3). Diese Forderung ist höchst problematisch, weil sie eine Bevorzugung quantitativer Methoden impliziert (vgl. Miethe 2013, 930f.). Trotz der Prämisse, dass „sowohl der Schutz der individuellen Privatsphäre wie auch wissenschaftliches Erkenntnisinteresse gleichermaßen wichtige Güter“ (ebd., 931) sind, stellt sich daher die Frage, ob das Prinzip der Anonymisierung von qualitativ Forschenden in seiner vorgesehenen Form umgesetzt werden kann (vgl. Tilley/Woodthorpe 2011).

Eine weitere Herausforderung bleibt der frühestmögliche Zeitpunkt der Anonymisierung: In der qualitativen Forschung ist eine Anonymisierung vor bzw. während der Datenerhebung kaum möglich. Sie kann bei der Datenaufbereitung und -auswertung umgesetzt werden, wird häufig aufgrund des zusätzlichen Arbeitsaufwandes aber auch erst mit Blick auf die Publikation realisiert. Eine frühestmögliche Anonymität der Beforschten schützt nicht nur deren Persönlichkeitsrechte, sondern eröffnet den Wissenschaftlerinnen und Wissenschaftlern auch einen „unbefangenen Forschungszugriff“, in dem es der qualitativ-empirischen Forschung nicht um Aussagen über konkrete Personen geht, sondern um die Analyse sozialer Phänomene (vgl. Rademacher/Wernet 2015, 143).

3.3 Rückmeldung und Publikation von Forschungserkenntnissen

Die Restriktionen, die für die qualitative Forschung aus dem Prinzip der Anonymität resultieren, stehen ferner der Empfehlung entgegen, dass Wissenschaftlerinnen und Wissenschaftler „[b]ei der Präsentation und Publikation soziologischer Erkenntnisse […] die Resultate ohne verfälschende Auslassung von wichtigen Ergebnissen“ darstellen sollen (DGS/BDS 2017, §1,2). Ingrid Miethe stellt hier die berechtigte Frage, ob Forschende nicht auch in der Pflicht sind, Ergebnisse zu veröffentlichen, die zwar dem Grundsatz der Anonymität zuwiderlaufen, aber soziale Probleme vor allem in institutionellen Kontexten aufdecken und zu deren Lösung beitragen können (z.B. sexistische oder rassistische Einstellungen bei Lehrerinnen und Lehrern, rechtliche Zuwiderhandlungen gegenüber Kindern und Jugendlichen) (vgl. Miethe 2013, 932). Hier könnte eine utilitaristische Perspektive, nämlich der Fokus auf den hohen Nutzen der Information für die Allgemeinheit, stärker wiegen als die deontologische Setzung, nach der die Anonymität der Untersuchten unter allen Umständen geschützt werden muss.

Eine Abwägung des Schadens und Nutzens ist auch bei der Rückmeldung gewonnener Forschungsergebnisse an beteiligte Probandinnen und Probanden gefragt. Einige Wissenschaftlerinnen und Wissenschaftler vertreten die Meinung, die Rückspiegelung von Ergebnissen sei „eigentlich eine Selbstverständlichkeit" (Flick 1995, 170), zumal sich dadurch die Informiertheit, auf deren Grundlage die Einwilligung zur Teilnahme erteilt wurde, nicht nur auf den Untersuchungsprozess, sondern auch auf die dadurch gewonnenen Erkenntnisse erstreckt. Probandinnen und Probanden über die Forschungsergebnisse und -analysen in Kenntnis zu setzen und ihnen damit die Wahl zwischen Publikation oder Nicht-Publikation zu ermöglichen, entspricht in gewisser Weise einer maximal informierten Einwilligung. Jedoch steht durch diese Option zugleich das Gelingen sowie die Erkenntnis eines gesamten Forschungsprojekts und der darin investierten Arbeit auf dem Spiel.

Zudem wird mit einer solchen Rückmeldung im Falle von Nachnutzungen auch auf Anonymität zwischen Forscherinnen und Forschern und Beforschten verzichtet, was einschränkende Konsequenzen für ein rücksichtsloses Erkenntnisinteresse hat (vgl. Rademacher/Wernet 2015, 143): Da das Erkenntnisinteresse ein hohes Kränkungspotenzial birgt, können Untersuchte u.U. in ihrer emotionalen Stabilität beschädigt werden (vgl. Miethe 2003; Miethe 2013, 933). Die Gefahr der Kränkung der Beforschten wiederum kann die Forscherin/der Forscher in ihrer kritischen Analyse hemmen.

Vor dem Hintergrund der hier nur in Kürze dargestellten Positionen kommen diverse Sozialwissenschaftlerinnen und Sozialwissenschaftler zu der Ansicht, dass die allgemein formulierten Vorgaben der Ethik-Kodizes in der Anwendung quantitativer Verfahren möglicherweise umsetzbar, für die qualitative Forschung jedoch in ihrer Rigidität impraktikabel sind (vgl. Hopf 2016; Miethe 2013; von Unger 2014). Stattdessen sollen „andere, für sozialwissenschaftliche Untersuchungen passende Verfahren [gefunden werden], mit deren Hilfe die Rechte der untersuchten Personen, Gruppen und Einrichtungen geschützt werden können, ohne die Logik, die spezifischen Vorgehensweisen und Gegenstände der sozialwissenschaftlichen Forschung aus dem Blick zu verlieren" (von Unger 2014, 36). Um eine solche Entwicklung zu ermöglichen, ist jedoch die Fähigkeit unerlässlich, das eigene Denken und Handeln kritisch zu reflektieren, und somit die Entscheidungsmomente im Forschungsprozess zu erkennen, die für eine verantwortungsvolle Wissenschaftspraxis maßgeblich sind. Dieser Reflexionsprozess sollte bereits im Studium angestoßen werden, um das Bewusstsein über die eigene forschungsethische Verantwortung möglichst früh zu fördern (vgl. McGinn/Bosacki 2004).

Was für die Sozialwissenschaften gilt, muss so auch für eine empirische Deutschdidaktik gelten: Die Anwendung qualitativ-sozialwissenschaftlicher Methoden ermöglicht eine profunde Untersuchung des Forschungsfeldes ‚Deutschunterricht', zieht jedoch vor dem Hintergrund der dargestellten Diskussion zwangsläufig auch die Übertragung forschungsethischer Überlegungen in den deutschdidaktischen Relevanzbereich nach sich. Die Unterrichtsvideografie als eine Art der

‚fossilierten' teilnehmenden Beobachtung nähert sich der simultanen Vielschichtigkeit des Unterrichtsgeschehens in ihrer Datenfülle und ist damit eine Methode, die sich zur Darstellung forschungsethischer Herausforderungen besonders anbietet.

4. Forschungsethische Herausforderungen einer deutschdidaktischen Professions- und Unterrichtsforschung

Ein wesentlicher Grund, warum forschungsethische Fragen auch in der Deutschdidaktik relevant werden, liegt in der disziplinären Entwicklung der vergangenen zwei Jahrzehnte. Neben der zunehmenden Ausrichtung am Paradigma einer qualitativ- und quantitativ-empirischen Forschung (vgl. Boelmann 2016) ist die Modellierung von Kompetenzen in Hinblick auf das Lehren und Lernen im Fach Deutsch gefragt, was wiederum zur verstärkten Beforschung von Lehrenden im Unterrichtskontext führt (vgl. Bräuer/Winkler 2012; Bräuer/Wieser 2015). Ein Verfahren der Datengewinnung, das in solchen Untersuchungen zunehmend eingesetzt wird, ist die Unterrichtsvideografie. Der Datenreichtum, den die videografische Aufzeichnung einer Deutschstunde umfasst, hat gegenüber audiografischen Mitschnitten und daraus gefertigten Transkripten den Vorteil, dass Kommunikation in ihrer Multimodalität[6] erfasst wird: Neben den verbalen Äußerungen werden Mimik, Gestik, Körpersprache sowie die Bedeutung des Raumes für das Ge- oder Misslingen kommunikativer Prozesse aufgezeichnet und somit nachvollziehbar; Praktiken, die sich in Raum und Zeit in Wechselwirkung zwischen Subjekten und Objekten vollziehen, werden rekonstruierbar. Dieser Vorzug der Unterrichtsvideografie ist zugleich jedoch mit einer erhöhten ethischen Verantwortung verknüpft, die weitgehend ähnlich dilemmatische Verflechtungen aufweist, wie sie für die Sozialwissenschaften bereits diskutiert wurden.

4.1 Herausforderungen der vollständigen Anonymisierung

Die Aufzeichnung realer Personen im geschützten Schulraum bringt umfangreiche datenschutzrechtliche Anforderungen mit sich. Es gilt nicht nur rechtliche Anforderungen zu erfüllen und dabei die aktuellen Datenschutzgesetze der Bundes- und Landesverfassungen zu berücksichtigen,[7] sondern sich darüber hinaus der Sensibilität der Daten bewusst zu werden. Denn im Falle von Videografien stellen nicht nur die Aufzeichnung der Klarnamen und die persönliche Unterschrift auf der Einwilligungserklärung ein Problem für eine umgehende Anonymisierung der Beforschten dar; die Lehrkräfte und die Schülerschaft werden in

[6] Dennoch bildet auch Videografie nicht die unterrichtliche Realität ab – jede Unterrichtsaufzeichnung konstatiert einen Ausschnitt und manifestiert eine Perspektive.

[7] Vgl. Bundesdatenschutzgesetz in der Fassung der Bekanntmachung vom 14.1.2003, zuletzt geändert durch Art. 10, 2 des Gesetzes vom 31.10.2017. Einzusehen unter http://www.gesetze-im-internet.de/bdsg_1990/index.html (letzter Zugriff: 01.08.2018).

ihren Worten und Taten, in ihrer individuellen Physiognomie und Prosodie aufgezeichnet und bleiben somit leicht identifizierbar. Zwar besteht die Möglichkeit, Gesichter nachträglich per Verpixelung unkenntlich zu machen oder die Stimme zu verfremden. Durch diese Eingriffe gehen jedoch die Mimik und Stimme als grundlegende Kommunikatoren in der sozialen Interaktion verloren mithin Informationsquellen, die die Videografie als Datengrundlage besonders informativ machen. Der Datenreichtum der Unterrichtsvideografie steht somit zugleich der Forderung des DGfE-Kodexes nach wissenschaftlichen Verfahren entgegen, die die Möglichkeit einer Identifizierung der Informanten ausschließen.

4.2 Herausforderung des minimalinvasiven Unterrichtseingriffs

Ein weiterer forschungsethisch kritischer Aspekt ist, dass das Aufzeichnen realen, alltäglichen Unterrichts die Anwesenheit fremder Personen mit Ton- und Kameraequipment verlangt. Deren Positionierung bzw. Bewegungen im Raum stellen einen Eingriff in den geschützten Schulraum dar. Das forschungsmotivierte Eindringen kann Lehrkräfte wie Schülerinnen und Schüler psychisch belasten und Auswirkungen auf Inhalt, Format und Erfolg des Lehr- und Lernprozesses haben. In aller Regel sind Unterrichtsaufnahmen je nach Landesvorgaben daher auch von den zuständigen Schulbehörden zu genehmigen (vgl. z.B. Niedersächsisches Kultusministerium 2015, Punkt 2). In Niedersachsen beispielsweise verlangt die Landesschulbehörde neben einer umfassen Projektdarstellung unter anderem die Vorlage der Informationsschreiben an Eltern und Schülerinnen und Schüler sowie Auskunft über die Personen, die im Weiteren Zugriff auf die Aufzeichnungen haben (sollen). Die Landesschulbehörde prüft auf dieser Grundlage, ob das Projektziel ggf. mit bereits bestehenden Daten und Forschungsergebnissen erreicht werden kann, inwieweit der Unterrichtsverlauf durch die geplante Erhebung gestört werden könnte und ob neben der Freiwilligkeit der Teilnahme die datenschutzrechtlichen Vorschriften eingehalten werden (vgl. ebd., Punkt 3). Die umfangreiche Antragstellung soll eine gute forschungsethische Wissenschaftspraxis schon im Vorfeld der Arbeit mit Unterrichtsvideografien sicherstellen.

4.3 Herausforderung der informierten Einwilligung

Dennoch treten trotz aller Vorsorge im Erhebungs- und Forschungsprozess immer wieder forschungsethische Herausforderungen auf. Schon die Informierung der Teilnehmenden, die einer Einwilligung vorausgeht, gestaltet sich schwierig: Teilnehmerinnen und Teilnehmer sind im Falle einer Unterrichtsvideografie neben der Lehrkraft die Schülerinnen und Schüler; das ausschlaggebende Einverständnis, das die Teilnahme am Projekt ermöglicht, erteilen bis zur Volljährigkeit jedoch die Eltern.[8] Die Frage ist demnach, inwiefern von einer informierten Einwilligung gesprochen werden kann, wenn zum einen nicht sicher ist, inwiefern

[8] Dies gilt z.B. in Niedersachsen für Kinder und Jugendliche bis zu dem Alter, ab dem sie laut Runderlass des niedersächsischen Kultusministeriums „einwilligungsfähig" sind, d.h. „die Bedeutung und die Tragweite der Einwilligung und deren rechtliche

die Schülerinnen und Schüler die Tragweite und die Folgen ihrer Teilnahme einschätzen können, und es zum anderen meist nicht die Teilnehmenden selbst, sondern dritte, nicht durch die Aufzeichnungen Betroffene sind, die über die Teilnahme am Projekt entscheiden. Darüber hinaus kann die vorliegende Genehmigung einer Behörde und der Schulleitung sowie der Teilnahmewunsch der Lehrkraft bereits illegitimen Druck auf Eltern und ihre Kinder ausüben. Aufgrund des hierarchischen Drucks könnten sich Eltern und Kinder dazu gedrängt sehen, wider eigener Überzeugung zuzustimmen, was gegen die forschungsethische Maxime der Freiwilligkeit verstoßen würde.

Eine weitere Schwierigkeit, die das Prinzip des *informed consent* für die Aufzeichnung von Unterrichtsvideografien birgt, ist die Unvorhersehbarkeit von sozialer Interaktion und die damit verbundene eingeschränkte Planbarkeit des Forschungsverlaufs. Eine gängige Art der Arbeit mit Videografie ist, dass die Forschenden den Unterricht zunächst aufzeichnen, um dann auf der Grundlage des gewonnenen Materials interessante Interaktionsphänomene und individuelle Handlungen zu identifizieren, die sich während der Stunde emergent ereignet haben. Ähnlich wie es bereits für qualitativ-sozialwissenschaftliche Methoden beschrieben wurde, ist es folglich schwierig, im Vorhinein genau festzulegen, welche Aspekte der Unterrichtsstunde analysiert werden sollen, d.h. welche konkrete Richtung der Forschungsprozess einschlagen wird. Dies gilt besonders für eine Nachnutzung des Materials: Bei Audio- und mehr noch bei Videodaten können Informationen zu Untersuchungsgegenständen werden, die zuvor nicht als solche fokussiert waren und über die aus diesem Grund nicht informiert werden konnte. Der analytische Frei- und Spielraum, den die videografische Methode braucht, um produktiv zu sein, steht somit dem Recht der Teilnehmenden auf profunde Informiertheit entgegen.

Des Weiteren können sich durch die Unvorhersehbarkeit sozialer Interaktionen praktische Probleme ergeben, indem z.B. während der Aufnahme einer Unterrichtsstunde eine fremde Lehrkraft kurzfristig das Klassenzimmer betritt, um etwas mit der/dem videografierten Kollegin/Kollegen zu besprechen. Diese Situation, die im Schulalltag unproblematisch ist und durchaus vorkommen kann, stellt die Filmenden jedoch vor die spontane Entscheidung, ob die Aufnahme an dieser Stelle aufgrund der fehlenden informierten Einwilligung der fremden Lehrkraft unterbrochen wird – trotz des Risikos, dass durch das fehlende Videofragment interessante Geschehnisse im Klassenzimmer nicht aufgezeichnet werden – oder ob die Aufnahme weiterläuft und zu hoffen bleibt, dass die fremde Lehrkraft ihre Zustimmung nach einer vollständigen Informierung zu einem späteren Zeitpunkt erteilt (vgl. auch von Unger 2014, 26).

Folgen […] erfassen und ihren Willen hiernach […] bestimmen [können]". Einwilligungsfähigkeit liegt nach dieser Definition in der Regel mit Besuch des 9. Jahrgangs vor (vgl. Niedersächsisches Kultusministerium 2015, Punkt 3.4.3.b). Ab diesem Zeitpunkt muss die Einverständniserklärung sowohl von den Eltern als auch von den betroffenen Schülerinnen und Schülern unterschrieben werden.

Ein ähnlicher, wenn auch moralisch dringlicherer Konflikt kann sich beim Videografieren von sensiblen Unterrichtssituationen ergeben, die von den gefilmten Personen u.U. als unangenehm empfunden werden. Dies kann insbesondere in Kontexten wie dem Klassenrat der Fall sein, in denen nicht der inhaltliche Lernstoff im Mittelpunkt steht, sondern sensible persönliche Themen besprochen werden. Vor allem bei der Lösung klasseninterner Konflikte kann sich die beobachtete Situation durch Streit und Vorwürfe schnell in eine Richtung entwickeln, die mehr sensible Informationen preisgibt als geplant und die Beteiligten angreifbar oder verletzlich macht. Unangenehme Gefühle können hierbei nicht nur bei Probandinnen und Probanden, sondern auch auf Seiten der Filmenden hervorgerufen werden, weil sie unbeabsichtigt voyeuristische Beobachterinnen oder Beobachter eines privaten Konflikts werden, der für das Forschungsprojekt zwar höchst aufschlussreich sein kann, sich aber in einem geschützten Raum abspielt, der unter normalen Umständen für Beobachtungen verschlossen ist. Hier stellt sich die Frage, ob die Kamera ausgeschaltet bzw. die Aufnahme gelöscht werden sollte, um die Privatsphäre der Teilnehmerinnen und Teilnehmer zu schützen, zumal das Prinzip der Anonymität aus den bereits genannten Gründen nur schwer umsetzbar ist. Es besteht jedoch die Chance, dass durch die Beschreibung und Analyse prekärer sozialer Situationen Erkenntnisse gewonnen werden, die den pädagogischen Umgang mit Konflikten innerhalb der Klasse optimieren könnten, was wiederum profitabel für angehende Lehrkräfte wäre. Auch hier stehen folglich eine deontologische und eine konsequentialistische Perspektive in einem forschungsethischen Spannungsverhältnis.

4.4 Herausforderungen im Umgang mit ethisch problematischen Beobachtungen

Vor dem Hintergrund der Beobachtung konfliktgeladener Interaktionen muss gefragt werden, inwiefern Forschende unmittelbar eingreifen dürfen bzw. sollen, wenn sie Zeugen ethisch oder juristisch inakzeptablen Handelns gegen Beforschte werden (z.B. institutionelle Ungerechtigkeiten, Rassismus, Sexismus, gewaltvolle Kommunikation etc.). In Hinblick darauf, dass es sich bei Schülerinnen und Schülern in der Mehrzahl um Minderjährige handelt, muss überlegt werden, ob auch ein mittelbares Eingreifen, etwa durch eine Mitteilung an die Schulleitung oder eine Behörde, als moralische Pflicht gerechtfertigt erscheint; etwa dann, wenn das Verschweigen beobachteter Ungerechtigkeit zugunsten der Anonymitätswahrung ein größerer forschungsethischer Verstoß wäre, als das Akzeptieren bestehender grundrechtsgefährdender Strukturen (vgl. Miethe 2013, 932).

4.5 Herausforderungen der Publikation von individualkritischen Forschungserkenntnissen

Eine weitere Möglichkeit des Eingreifens in eine problematische Praxis ist ihre Bearbeitung in einer wissenschaftlichen Veröffentlichung. Auch hier bleibt aus forschungsethischer Perspektive zu überlegen, wie in einer solchen Untersuchung die Situation und die an ihr Beteiligten ethisch angemessen dargestellt werden

können. Die kritische Analyse sozialer Phänomene birgt das Risiko, die Beforschten in ihrer Persönlichkeit anzugreifen, zumindest zu verletzten – selbst, wenn es der Forschung nicht um die Personen, sondern um die sozialen Praktiken oder Strukturen geht. Dies ist auch der Fall, wenn die konkreten Personen anonym bleiben, sich aber in der kritischen Darstellung selbst wiedererkennen und getroffen fühlen müssen. Rademacher und Wernet weisen in einem Beitrag durchaus in forschungsethischer Perspektive darauf hin, dass die „Rücksichtslosigkeit" kritischer Analyse „im Zeichen eines Erkenntnisinteresse[s]" stehe (Rademacher/Wernet 2015, 143). Eine solche Position zieht Kritik auf sich (vgl. Reichertz 1986; Bude 1994) und sollte unter forschungsethischer als auch erkenntnistheoretischer und methodologischer Perspektive diskutiert werden – so mögen die Binnenlogiken methodologischer Zugänge, wie etwa der Objektiven Hermeneutik oder der Dokumentarischen Methode, Wissenschaftlerinnen und Wissenschaftler in Hinblick auf forschungsethische Fragen vor unterschiedliche Herausforderungen stellen.

Individualkritische Forschungserkenntnisse werden auch auf deutschdidaktischen Tagungen oder in Bezug auf deutschdidaktische Publikationen (etwa als ‚Lehrer-Bashing') kritisch diskutiert. Es bleibt der forschungsethische Diskurs zu führen, wie zwischen einem kritisch-analytischen Erkenntnisinteresse und einer konstruktiv-reflexiven Lehrerbildung vermittelt werden kann – ein Rückzug auf eine Position wissenschaftlicher Autarkie erscheint für die Deutschdidaktik nicht nur aus forschungsethischen, sondern auch aus disziplinären Aspekten problematisch: Für die deutschdidaktische Forschung gilt, dass die Akteurinnen und Akteure im Handlungsfeld zugleich Objekte der Beforschung und Subjekte der Aufnahme und Anwendung von wissenschaftlichen Erkenntnissen sind. Sie sind dadurch – zumindest in einem anwendungsorientierten Verständnis empirischer Forschung in der Deutschdidaktik – nicht nur Informantinnen und Informanten, sondern zugleich auch Adressatinnen und Adressaten. Oder anders formuliert: Als Teil des beobachteten Problems sind Lehrerinnen und Lehrer immer auch Teil von dessen potentieller Lösung. Die mittelbaren Folgen eines ‚rücksichtslosen' Umgangs mit Unterrichtsvideografien sind für die Forschung demnach risikobehaftet: Sie können das Misstrauen von Lehrkräften nähren, nur als Objekte der Forschung zu dienen, und den zukünftigen Zugang zum Praxisfeld und die Kooperation mit Lehrenden erheblich erschweren. Vor diesem Hintergrund bleibt zu fragen, wie in einem forschungsethisch-sensiblen Diskurs die Probleme der sozialen Prozesse im Handlungsfeld analytisch erforscht und dargestellt werden können.

4.6 Herausforderungen bei Rückmeldung an beteiligte Informanten

Ein abschließender Aspekt, der im Umgang mit erhobenen Videodaten berücksichtigt werden muss, ist die Rückspiegelung der Erkenntnisse an Beteiligte. Wie in der sozialwissenschaftlichen Debatte stellt sich auch hier die Herausforderung, zwischen Nutzen und Schaden abzuwägen. Möchte z.B. eine videografierte Lehrkraft die Aufnahme ihres Unterrichts anschauen, um ihren Lehrstil von außen zu betrachten, so könnte dies sowohl ihr als auch ihren Schülerinnen und Schülern

nützen: Durch die Selbstbetrachtung ausgelöste Reflexionsprozesse können dazu beitragen, potenziell bestehende Probleme in der Kommunikation oder Unterrichtsführung zu erkennen, zu lösen und so die Unterrichtsqualität zu verbessern. Der Nutzen kann jedoch schnell zu einem Nachteil für die Schülerschaft werden, wenn sich der Blick der Lehrkraft auf die Beobachtung des übrigen Klassengeschehens ausweitet. Da bei Unterrichtsvideografien oftmals aus mehreren Perspektiven gefilmt wird – eine Kamera richtet sich etwa auf die Lehrperson, eine oder mehrere Kamera/s filmt/filmen die Lernenden – besteht die Gefahr, dass die Lehrkraft über die Kamera einen Einblick in die Persönlichkeitssphäre von Schülerinnen und Schüler erhält, die diese der Lehrkraft nicht gewähren möchten und in die die Schülerinnen und Schüler auch nicht eingewilligt haben. Solche Einblicke können u.U. negative Auswirkungen auf die allgemeine Einschätzung der jeweiligen Person und auch auf die Leistungsbewertung haben. In solchen Fällen hätte die Rückspiegelung der Unterrichtsaufnahme als Grundlage des wissenschaftlichen Erkenntnisprozesses einen Nachteil für die beforschten Schülerinnen und Schüler.

5. Ethische Implikationen für deutschdidaktische Forschung

Wie die exemplarische Skizzierung forschungsethisch kritischer Momente in der Arbeit mit Unterrichtsvideografien zeigt, bedeutet das weitgehende Fehlen eines forschungsethischen Diskurses in der Deutschdidaktik nicht, dass es hier keine solchen Herausforderungen gäbe. Der dringende Bedarf an einer Debatte über eine gute Wissenschaftspraxis, die sich sowohl mit ihren internen als auch mit ihren externen Verantwortlichkeiten auseinandersetzt und differenzierte Orientierungshilfen gibt, ergibt sich zum einen daraus, dass die deutschdidaktische Forschung mit Kindern und Jugendlichen arbeitet, die als nicht-mündige, moralische Objekte gerade im Schulkontext einen besonderen Schutz erfahren müssen. Zum anderen muss berücksichtigt werden, dass die Adressaten deutschdidaktischer Erkenntnis immer auch die Beforschten selbst sind.

Darüber hinaus herrscht Diskussionsbedarf bei der Frage nach verantwortungsvoller Datenerhebung: Eine aussagekräftige empirische Professions- und Unterrichtsforschung verlangt nach einer ausreichend großen Datengrundlage, um zu belastbaren Ergebnissen gelangen zu können. Es bestehen jedoch objektive und subjektive Zumutbarkeitsgrenzen für die Erhebung von realen Unterrichtsdaten an Schulen. Ein Ausweg könnte in der nachhaltigen Nutzung von bereits erhobenen Forschungsdaten liegen, die Didaktikerinnen und Didaktikern zur wissenschaftlichen Verfügung gestellt werden (vgl. das Projekt „Verbund Forschungsdaten Bildung“ am Deutschen Institut für internationale pädagogische Forschung)[9]. Der Vorteil läge darin, dass bestehendes Material von unterschiedlichen Forschungsgruppen auf die eigene Fragestellung hin untersucht und die Beforschten somit entlastet werden könnten. Allerdings ist die Aufbereitung der Daten für

[9] Einzusehen unter: https://www.forschungsdaten-bildung.de (letzter Zugriff: 01.08.2018).

derartige Sekundäranalysen aus zwei Gründen ethisch kritikabel: Die freiwillige und informierte Einwilligung ist bei einer Folgenutzung von Forschungsdaten kaum einholbar. Hinzu kommt, dass das individuelle Verantwortungsempfinden in Hinblick auf eine sensible Nutzung der Daten durch die Distanz zwischen Forschenden und Beforschten sehr gering sein kann.

Ein kontinuierlicher forschungsethischer Diskurs ist für die Deutschdidaktik zudem erstrebenswert, weil – wie in anderen Disziplinen – auch hier mehr und mehr in Projektzusammenhängen und Verbünden geforscht wird (z.B. *IGLU*, *DESI*, *TEDS-LT*, *FALKO-D*, *PERLE*). An die Stelle der von einzelnen Forscherinnen und Forschern betriebenen, selbstständigen und selbstverantworteten Wissenschaftspraxis tritt folglich die größtenteils weisungsgebundene Forschung (vgl. Özmen 2015, 66) in Gruppen, die „hochgradig kollektiv und arbeitsteilig organisiert [ist], was eine individuelle Zuschreibung von Verantwortung erschwert" (Graumann 2006, 255). Ein weiterhin fehlender Diskurs birgt vor diesem Hintergrund die Gefahr, dass spontan zu treffende Entscheidungen, die in ihrer Vielzahl die forschungsethische Klarheit des Arbeitsprozesses bestimmen, aufgrund von Uninformiertheit oder Unsicherheit nicht kompetent selbst getroffen, sondern auf andere Kolleginnen und Kollegen oder Vorgesetzte übertragen werden, sodass der ethisch gebotene individuelle Einsatz für eine moralische Forschungspraxis am Ende ausbleiben kann.

So erscheint die disziplinäre Verständigung über ethische Grundlagen und Prioritätsprinzipien noch dringlicher, auch wenn eine Einigung auf praktisch anwendbare Normen in Form eines Ethik-Kodex möglicherweise keine sofortige Heilung bringt, sondern – im Gegenteil – zunächst einmal neue prinzipielle Fragen über die Grenzen von Verantwortung und Wissenschaftsfreiheit aufwirft. Andererseits mag dadurch eine kritische Reflexion der eigenen Forschungshandlungen zusammen mit der Einsicht evoziert werden, dass es keinen ‚goldenen' Weg zu intern und extern verantwortlichem Handeln in der Wissenschaft gibt; vielmehr muss in jeder kritischen Forschungssituation eine moralisch-sensible Entscheidung getroffen werden, die die spezifischen Umstände berücksichtigt und das Ergebnis eines differenzierten und verantwortungsbewussten Reflexionsprozesses ist. Voraussetzung hierfür ist eine konstruktiv-kritische Diskussion über die ethischen Herausforderungen in der Deutschdidaktik, die durch diesen Beitrag am Beispiel der Unterrichtsvideografie in Ansätzen vorgenommen wurde.

Literatur

Boelmann, Jan M. (Hrsg.) (2016): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. Baltmannsweiler: Schneider Hohengehren.

Bräuer, Christoph/Winkler, Iris (2012): Aktuelle Forschung zu Deutschlehrkräften. Ein Überblick. In: Didaktik Deutsch, 18, 33, 74-91.

Bräuer, Christoph/Wieser, Dorothee (2015): Lehrende im Blick. Empirische Lehrerforschung in der Deutschdidaktik. Wiesbaden: Springer VS.

Brendel, Elke (2011): Wissenschaft. In: Kolmer, Petra/Wildfeuer, Armin G. (Hrsg.): Neues Handbuch philosophischer Grundbegriffe, Band 3. Freiburg i.Br.: Karl Alber, 2588-2601.

Broad, Charlie Dunbar (1930): Five Types of Ethical Theory. London.

Bude, Heinz (1994): Das Latente und das Manifeste: Aporien einer ‚Hermeneutik des Verdachts'. In: Garz, Detlef/Kraimer, Klaus (Hrsg.): Die Welt als Text: Theorie, Kritik und Praxis der objektiven Hermeneutik. Frankfurt a.M.: Suhrkamp, 114-124.

Bundesministerium für Justiz und Verbraucherschutz (2003): Bundesdatenschutzgesetz, http://www.gesetze-im-internet.de/bdsg_1990/index.html (letzter Zugriff: 01.08.2018).

Cortés-Puch, Irene/Wesley, Robert A./Carome, Michael A./Danner, Robert L./Wolfe, Sidney M./Natanson, Charles (2016): Usual Care and Informed Consent in Clinical Trials of Oxygen Management in Extremely Premature Infants. In: PLoS ONE 11(5): e0155005. https://doi.org/10.1371/journal.pone.0155005 (letzter Zugriff: 01.08.2018).

Deutsche Gesellschaft für Erziehungswissenschaften (2010): Ethik-Kodex der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), http://www.dgfe.de/fileadmin/OrdnerRedakteure/Satzung_etc/Ethikkodex_2010.pdf (letzter Zugriff: 01.08.2018).

Deutscher Ethikrat (2014): Biosicherheit – Freiheit und Verantwortung in der Wissenschaft. Stellungnahme, https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-biosicherheit.pdf (letzter Zugriff: 01.08.2018).

Deutsche Gesellschaft für Soziologie/Berufsverband Deutscher Soziologinnen und Soziologen (2017): Ethik-Kodex der Deutschen Gesellschaft für Soziologie (DGS) und des Berufsverbandes Deutscher Soziologinnen und Soziologen (BDS), http://www.soziologie.de/fileadmin/user_upload/DGS_Redaktion_BE_FM/DGSallgemein/Ethik-Kodex_2017-06-10.pdf (letzter Zugriff: 01.08.2018).

Ellis, Carolyn (2007): Telling Secrets, revealing lives: Relational ethics in research with intimate others. In: Qualitative Inquiry, 13, 1, 3-29.

Flick, Uwe (1995): Stationen des qualitativen Forschungsprozesses. In: Flick, Uwe/von Kardorff, Ernst/Keupp, Heiner/von Rosenstiel, Lutz/Wolff, Stephan (Hrsg.): Handbuch Qualitative Sozialforschung. Grundlagen, Konzepte, Methoden und Anwendungen. 2. Aufl. Weinheim: Beltz Psychologie, 147-173.

Forschung & Lehre (2018): Tausende Wissenschaftler publizieren in Pseudo-Zeitschriften. In: ebd., 645.

Graumann, Sigrid (2006): Forschungsethik. In: Düwell, Marcus/Hübenthal, Christoph/Werner, Micha H. (Hrsg.): Handbuch Ethik. Stuttgart/Weimar: J.B. Metzler, 253-257.

Heidbrink, Ludger (2017): Definitionen und Voraussetzungen der Verantwortung. In: Heidbrink, Ludger/Langbehn, Claus/Loh, Janina (Hrsg.): Handbuch Verantwortung. Wiesbaden: Springer VS, 3-33.

Hopf, Christel (2016): Forschungsethik und qualitative Forschung. In: Hopf, Wulf/Kuckartz, Udo (Hrsg.): Schriften zu Methodologie und Methoden qualitativer Sozialforschung. Wiesbaden: Springer VS, 195-205.

Jonas, Hans (1979): Das Prinzip Verantwortung. Frankfurt a.M.

Jones, James H. (1993): Bad blood. The Tuskegee Syphilis experiment. New York: The Free Press.

Kelly, Alison (2005): Education or Indoctrination? The Ethics of School Based Action Research. In: Burgess, Robert G. (Hrsg.): The ethics of educational research. New York: Falmer Press, 93-105.

Lamnek, Siegfried/Krell, Claudia (2016): Qualitative Sozialforschung. 6., überarb. Aufl. Weinheim/Basel: Beltz.

Lenk, Hans (1991): Zu einer praxisnahen Ethik der Verantwortung in den Wissenschaften. In: Ders. (Hrsg.): Wissenschaft und Ethik. Stuttgart: Reclam, 54-75.

Loh, Janina (2017): Strukturen und Relata der Verantwortung. In: Heidbrink, Ludger/Langbehn, Claus/Loh, Janina (Hrsg.): Handbuch Verantwortung. Wiesbaden: Springer VS, 35-56.

Maring, Matthias (2011): Forschungs- und Wissenschaftsethik. In: Stoecker, Ralf/Neuhäuser, Christian/Raters, Marie-Luise (Hrsg.): Handbuch Angewandte Ethik. Stuttgart/Weimar: J. B. Metzler, 165-169.

McGinn, Michelle K./Bosacki, Sandra L. (2004): Research Ethics and Practitioners: Concerns and Strategies for Novice Researchers Engaged in Graduate Education. In: Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 5, 2, http://nbn-resolving.de/urn:nbn:de:0114-fqs040263 (letzter Zugriff: 01.08.2018).

Miethe, Ingrid (2003): Das Problem der Rückmeldung. Forschungsethische und -praktische Erfahrungen und Konsequenzen in der Arbeit mit hermeneutischen Fallrekonstruktionen. In: Zeitschrift für qualitative Bildungs-, Beratungs- und Sozialforschung, 4, 2, 223-240.

Miethe, Ingrid (2013): Forschungsethik. In: Friebertshäuser, Barbara/Langer, Antje/Prengel, Annedore (Hrsg.): Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft. Unter Mitarbeit von Heike Boller und Sophia Richter. 4., durchg. Aufl. Weinheim/Basel: Beltz Juventa, 927-937.

Niedersächsisches Kultusministerium (2015): Umfragen und Erhebungen in Schulen. Runderlass der Ministerkonferenz, http://www.nds-voris.de/jportal/portal/t/1ct3/page/bsvorisprod.psml?doc.hl=1&doc.id=VVND-VVND000036294&documentnumber=1&numberofresults=1&doctyp=vvnd&showdoccase=1&doc.part=F&paramfromHL=true#focuspoint (letzter Zugriff: 01.08.2018).

Özmen, Elif (2015): Wissenschaft. Freiheit. Verantwortung. Über Ethik und Ethos der freien Wissenschaft und Forschung. In: Ordnung der Wissenschaft, 2, 65-72.

Prechtl, Peter (1999): Deontologisch. In: Prechtl, Peter/Burkhard, Franz-Peter (Hrsg.): Metzler Philosophie Lexikon. Stuttgart/Weimar: J. B. Metzler, 101.

Projektgruppe „Soziale Beziehungen in der Familie, geschlechtsspezifische Sozialisation und die Herausbildung rechtsextremer Orientierungen“ (1996): Dokumentation und Erläuterung des methodischen Vorgehens. Hildesheim: Institut für Sozialwissenschaften der Universität Hildesheim.

Rademacher, Sandra/Wernet, Andreas (2015): Pädagogik zwischen Selbsterhöhung, Missachtung und Verklärung. In: Rademacher, Sandra/Wernet, Andreas (Hrsg.): Bildungsqualen: Kritische Entwürfe wider den pädagogischen Zeitgeist. Wiesbaden: Springer VS, 139-165.

Reichertz, Jo (1986): Probleme der qualitativen Sozialforschung: Zur Entwicklungsgeschichte der Objektiven Hermeneutik. Frankfurt a.M./New York: Campus.

Strohschneider, Peter/Hacker, Jörg/Lohse, Martin/Krull, Wilhelm (2018): Wie sich die Qualität der Auswahl verbessern lässt. Ein Vorschlag. In: Forschung & Lehre, 8, 668-670.

Tilley, Liz/Woodthorpe, Kate (2011): Is it the end for anonymity as we know it? A critical examination of the ethical principle of anonymity in the context of 21st century demands on the qualitative researcher. In: Qualitative Research, 11, 2, 197-212.

Verbund Forschungsdaten Bildung (2015): Checkliste zur Erstellung rechtskonformer Einwilligungserklärungen mit besonderer Berücksichtigung von Erhebungen an Schulen. Version 1.1. fdbinfo Nr. 1, https://www.forschungsdaten-bildung.de/get_files.php?action=get_file&file=fdbinfo_1.pdf (letzter Zugriff: 01.08.2018).

von Unger, Hella (2014): Forschungsethik in der qualitativen Forschung: Grundsätze, Debatten und offene Fragen. In: von Unger, Hella/Narimani, Petra/M'Bayo, Rosaline (Hrsg.): Forschungsethik in der qualitativen Forschung. Reflexivität, Perspektiven, Positionen. Wiesbaden: Springer VS, 15-39.

ULRICH IBERER

Datenschutz

Was Forscherinnen und Forscher wissen sollten

Forscherinnen und Forscher lieben ihre Aufgabe: Die volle Aufmerksamkeit auf ihr Thema gerichtet, ganz im Flow nach neuem Wissen und Erkenntnissen. Sie schätzen es, in Offenheit und Unabhängigkeit den Dingen auf den Grund zu gehen. Als lästig oder gar widrig werden jene Verpflichtungen wahrgenommen, die entgegen diesem freien Handeln zum Einhalten von Normen, Auflagen und Regeln verpflichten. Im Zeitalter der Digitalisierung werden insbesondere datenschutzrechtliche Anforderungen immer bedeutsamer, auch bzw. gerade im Kontext empirischer Forschung. Gewiss bildet der Datenschutz keine Lieblingsaufgabe unter den Forschungsakteuren. Eher wird versucht, sich schnellstmöglich des Themas zu erledigen, zu marginalisieren oder gar zu ignorieren. Die folgenden Erläuterungen wollen Aufklärungsarbeit dahingehend leisten, welchen Rechten und Pflichten des Datenschutzes Forscherinnen und Forscher unterliegen. Die Ausführungen nehmen gezielt solche Aspekte in den Blick, die in empirischen Forschungsprojekten besonders häufig auftreten.

1. Was sind generelle datenschutzrechtliche Verpflichtungen für Forscherinnen und Forscher?

1.1 Konflikt zweier Grundrechte

„Kunst und Wissenschaft, Forschung und Lehre sind frei“ (GG § 5): In den grundgesetzlich verankerten Grundrechten nimmt die Wissenschafts- und Forschungsfreiheit einen hohen Stellenwert ein, Forscherinnen und Forscher können und sollen unabhängig an ihren Fragestellungen arbeiten. In einer besonderen Konstellation befinden sich die Sozial- und Geisteswissenschaften. Sie haben den Menschen zum Forschungsgegenstand, er wird zum Objekt von Forschung. Wissenschaftlerinnen und Wissenschaftler erheben in ihren Arbeiten sehr präzise, mitunter besonders sensible Daten von konkreten Personen. Damit kann die zitierte Wissenschafts- und Forschungsfreiheit in Konflikt zu einem anderen Grundrecht geraten: dem allgemeinen Persönlichkeitsrecht (GG § 3). Nach Rechtsprechung des Bundesverfassungsgerichts muss jeder Einzelne jederzeit grundsätzlich selbst

über die Preisgabe und Verwendung seiner persönlichen Daten bestimmen können – das bekannte Recht auf informationelle Selbstbestimmung (vgl. BVerfG 65,1; 15.12.1983, sog. *Volkszählungsurteil*). Einschränkungen dieses Rechts bedürfen einer verfassungsmäßigen, gesetzlichen Grundlage; die Grundsätze des überwiegenden Allgemeininteresses und der Verhältnismäßigkeit müssen dabei gewahrt bleiben. Bereits aus dieser hohen normativen Verankerung des Datenschutzes kann eine der wichtigsten Grundregeln für Forscherinnen und Forscher abgeleitet werden: Das Erheben, Verarbeiten, Speichern und Übermitteln von Daten, die auf einzelne Personen beziehbar sind, ist nur dann zulässig, wenn entweder ein Gesetz dies erlaubt oder die betroffenen Personen hierzu explizit eingewilligt haben (vgl. Kapitel 2.2).

1.2 Datenschutzgesetze

Um die Arbeit von Wissenschaftlern zu vereinfachen, hat der Gesetzgeber verschiedene Regelungen geschaffen. Zunächst gibt es kein allein bestimmendes Datenschutzgesetz; vielmehr regeln verschiedene Europa-, Landes- und Bundesgesetze, insbesondere die EU-Datenschutzgrundverordnung (EU-DSGVO), das Bundesdatenschutzgesetz (BDSG), die jeweiligen Landesdatenschutzgesetze und zahlreiche spezialgesetzliche Bestimmungen (z.B. Schulgesetze, Sozialgesetze, Verwaltungsbestimmungen) die Voraussetzungen, unter denen personenbezogene Daten für Forschungszwecke verarbeitet werden dürfen. Es gilt auch in diesen Gesetzen das Prinzip, dass die Persönlichkeitsrechte weitgehend gewahrt bleiben und zwischen widerstreitenden Positionen ein Ausgleich gefunden werden muss, beispielsweise durch eine nachträgliche Information an die Betroffenen. Wie viele andere Gesetze regeln die Datenschutzgesetze nicht den konkreten Einzelfall, sondern geben allgemeine Anforderungen zu den inhaltlichen und formalen Anforderungen an die Verarbeitung personenbezogener Daten aus. Für Forscherinnen und Forscher bedeutet dies: Generell müssen die datenschutzrechtlichen Hauptprinzipien beachtet werden: Datensparsamkeit, Datenvermeidung, Datenerforderlichkeit und Zweckbindung. Des Weiteren gilt es immer die konkreten Umstände im Einzelfall zu berücksichtigen.

1.3 Verantwortung für Datenschutz an der Hochschule

In den Gesetzen wird die „der Verantwortliche" als die für den Datenschutz organisationale Einheit genannt (Art. 4 (7) EU-DSGVO). In Forschungsprojekten ist dies typischerweise die Universität, Hochschule oder Forschungsinstitut. Generell verantwortet die offizielle Leitungsebene die erfolgreiche Anwendung der Datenschutzregelungen in der gesamten Organisation. An der Hochschule ist dies gewöhnlich die Hochschulleitung, das Rektorat bzw. der Direktor oder die Direktorin. Da in aller Regel eine Führungsperson alleine die Projekte und Aktivitäten einer Hochschule nicht überblicken kann, delegiert es die ordnungsgemäße Anwendung des Datenschutzes in die jeweiligen Fachbereiche, informiert und unterrichtet bei Änderungen, stellt organisationsweite Regelungen und Standards auf und kontrolliert diese über ein Datenschutz-Managementsystem (vgl. EU-DSGVO § 5, Absatz 2). In vielen Fällen wird bei der Zusage oder Genehmigung

von Ressourcen die Leitung eines Forschungsprojektes entsprechend verpflichtet. Die Verantwortung für die sachgemäße Berücksichtigung der Datenschutzbestimmungen trägt dann jede beteiligte Forscherin und jeder beteiligte Forscher. Zudem ist jedes Hochschulmitglied dazu verpflichtet, in seinem anvertrauten Arbeitsgebiet und mit seiner Fachkompetenz auf die Einhaltung der Datenschutzregeln hinzuwirken. Beispielsweise kann sich eine Doktorandin nicht auf Anweisungen oder Anordnungen verlassen, sondern muss in ihrem Dissertationsprojekt von sich aus für datenschutzkonforme Abläufe sorgen und ggf. übergeordnete Stellen auf Lücken oder Probleme hinweisen. Eine Ausnahme stellen Studierende dar: Da sie formal nicht in die Arbeitsstrukturen der Hochschule eingebunden und nicht weisungsgebunden sind, können sie auch nicht für Datenschutz-Aufgaben in einem Forschungsprojekt verpflichtet bzw. verantwortet werden, es sei denn, ein Arbeitsvertrag liegt vor. Umgekehrt agieren Studierende bei eigenen Forschungsprojekten, beispielsweise in einem Studienprojekt oder bei einer Abschlussarbeit, selbstverantwortlich und müssen ihrerseits Sorge tragen, dass ihr Handeln den rechtlichen Datenschutzanforderungen genügt.

1.4 Generelle Anforderungen

Das Erheben, Verarbeiten, Übermitteln und Speichern personenbezogener Daten zu Forschungszwecken muss stets auf das zum Erreichen des angegebenen Zwecks erforderliche Minimum beschränkt bleiben, das sogenannte Gebot zur Datensparsamkeit. In der Planung und Anlage eines Forschungsprojektes muss daher geprüft werden, ob das Ziel nicht auch mit einem gleich geeigneten, aber für die Betroffenen weniger belastenden Methode erreicht werden könnte. Dieses scheinbar selbstverständliche Gebot wird in der Forschungspraxis mal bewusst, mal unbewusst unterlaufen: Explorative Studien, die eine möglichste breite Datenerhebung anstreben; Daten-Repositorien, die in Anschlussprojekten wiedergenutzt werden sollen; Spontan- und Kleinforschungsprojekte, in denen zunächst Daten erhoben werden und erst in späteren Schritten über eine Ausweitung des Forschungsprojektes entschieden wird. Wissenschaftlerinnen und Wissenschaftler begeben sich hierbei in den Verdacht, gesetzeswidrig Datenvorratsspeicherung zu betreiben, auch wenn sie subjektiv empfinden, für die Forschung „doch etwas Gutes zu tun“. Die Problematik liegt nicht nur in mangelnden Rechtsgrundlagen, z.B. den fehlenden Einwilligungen der Betroffenen. So steigt zum Beispiel ohne entsprechende technische und organisatorische Maßnahmen das Risiko des unkontrollierten Datenverlusts und Datenmissbrauchs durch Dritte.

1.5 Problem der Zweckbindung

Ein gegenwärtig noch nicht gelöstes Problem des Datenschutzes in der Forschung liegt im Gebot der Zweckbindung (vgl. Art. 5 (1 lit. b) EU-DSGVO). Dies macht eine Löschung von Forschungsdaten spätestens dann notwendig, wenn ein Forschungsprojekt faktisch beendet ist, zum Beispiel mit der Publikation einer Dissertation, Abschlussbericht eines Forschungsprojektes. Für viele Forschungsprojekte von heute entsteht damit ein diametraler rechtlicher Konflikt, wenn die För-

derauflagen von privaten und zunehmend auch öffentlichen Geldgebern verlangen, dass alle im Projekt erhobenen (Primär-)Daten über das Projektende hinaus für potentielle Folgeforschungen bereitgestellt werden müssen. Wissenschaftlerinnen und Wissenschaftler, die ein mit solchen Förderauflagen verknüpftes Projekt durchführen, gleichzeitig weitere Genehmigungsauflagen, z.B. Urheber- und Nutzungsrechte, Verwendungsbeschränkungen, sowie die Löschungspflichten aus dem Datenschutz befolgen sollen, sehen sich dem Dilemma ausgesetzt, darüber zu entscheiden, gegen welche der Auflagen sie letztlich verstoßen wollen. Es liegt nahe, dass faktisch am ehesten die Auflagen des Zuwendungsgebers befolgt werden.

2. Grundsätzliche datenschutzrechtliche Klärung

2.1 Die „Null-Prüfung“: Ist das Forschungsvorhaben tatsächlich datenschutzrechtlich bedeutsam?

Datenschutzrechtliche Bestimmungen finden dann und nur dann nur Anwendung, wenn für ein Forschungsprojekt personenbezogene Daten benötigt werden. Forschung ohne jeglichen Personenbezug oder mit sicher anonymisierten Daten ist jederzeit ohne datenschutzrechtliche Einschränkungen möglich. Ein erstes Beispiel hierfür wären Datensätze von Schülerantworten aus standardisierten Prüfungsarbeiten über mehrere Jahrgänge, die keine weiteren Angaben zu soziodemografische Merkmalen enthalten.

Was bedeutet *sicher anonymisiert*? Nach den Bestimmungen der EU-Datenschutzgrundverordnung gelten die Grundsätze des Datenschutzes für alle Informationen, "die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen" (EU-DSGVO, Erwägungsgrund 26). Eine sichere Anonymität meint damit nicht nur das Entfernen von Angaben, die unmittelbar eine konkrete Person aufzeigen wie beispielsweise ein Name oder ein Gesicht. Es sind auch solche Angaben auszuschließen, die mittelbar über die Kombinationen von Einzelangaben eine Zuordnung auf eine konkrete Person ermöglichen könnten. Beispielsweise können bei geschlossenen Frageformen zu eng differenzierte Antwortkategorien dazu führen, dass in der Zusammenschau mit anderen Antworten und weiterem Kontextwissen offensichtlich wird, wer sich hinter einem Datensatz verbirgt.

Wenn eine datenschutzrechtliche Bedeutsamkeit vermieden werden soll, empfiehlt es sich bereits bei der Entwicklung der Erhebungsmethodik, z.B. bei der Konstruktion des Fragebogens, all jene Merkmaleauszuschließen, die personenbezogene Merkmale erfassen. Erhebt eine Forscherin bzw. ein Forscher Informationen von einer Person, so sind diese Angaben zunächst personenbezogen. Sicher anonymisiert sind sie dann, wenn weder Forscherin bzw. Forscher noch andere auf die konkrete Person rückschließen können. Die erhobenen Daten dürfen auch nachträglich nicht auf eine bestimmte Person rückbeziehbar sein, etwa durch einen Handschriftenvergleich, und müssen frei von Zusatzinformationen sein, die eine bestimmte Person identifizieren würden.

Ist ein Personenbezug nicht definitiv auszuschließen, so können Forschungsdaten nur auf der Grundlage eines Gesetzes oder mit der Einwilligung der Betroffenen erhoben werden, wie im Folgenden ausgeführt wird.

2.2 Die „Normen-Prüfung“: Auf welcher gesetzlichen Grundlage können Daten erhoben und verarbeitet werden?

Das Datenschutzrecht ist funktional ein *Verbot mit Erlaubnisvorbehalt*, d.h. es untersagt das Erheben, Nutzen und Speichern von personenbezogenen Daten grundsätzlich, sofern nicht durch ein Gesetz oder die Einwilligung der Person erlaubt wird. Sollen in einem Forschungsvorhaben Daten von Personen erhoben werden, so ist dies unter folgenden Bedingungen gestattet:

Entweder: Ein Gesetz oder eine andere Rechtsvorschrift erlaubt dies bzw. schreibt es vor.

Zum Zwecke der wissenschaftlichen Forschung hat der Gesetzgeber sogenannte „Besondere Verarbeitungssituationen“ geschaffen (vgl. BDSG, Teil 2, Abschnitt 2). Diese schränken das informationelle Selbstbestimmungsrecht ein und ermöglichen ein Erheben und Verarbeiten von personenbezogenen Daten auch ohne Einwilligung. In § 27 BDSG wird eine Möglichkeit zur „Datenverarbeitung zu wissenschaftlichen oder historischen Forschungszwecken und zu statistischen Zwecken“ eingeräumt. Dies ist jedoch an strenge Voraussetzungen geknüpft, z.B. dürfen die schutzwürdigen Belange des Betroffenen nicht beeinträchtigt werden, ein öffentliches Interesse an dem Forschungsvorhaben muss schwerer wiegen als die Belange des Betroffenen und der Forschungszweck wäre ansonsten nicht oder nur mit unverhältnismäßigem Aufwand zu erreichen. Für solche Fälle muss vorab eine ausführliche Abwägung zwischen dem Forschungsinteresse und den Interessen der Betroffenen vorgenommen und dokumentiert werden.

Oder: Die betreffende Person bzw. die betreffenden Personen, um deren Daten es geht, willigen ausdrücklich ein.

Für die allermeisten Forschungsprojekte ist die sogenannte *informierte Einwilligung* der Weg, personenbezogene Daten rechtswirksam zu erheben und zu verarbeiten. Die Form einer Einwilligung ist prinzipiell unbestimmt, d.h. Forscherinnen und Forscher können diese entsprechend ihrer konkreten Konstellation selbst gestalten. Jedoch sind auch hier Anforderungen des Datenschutzrechts zu beachten (siehe Kapitel 3).

3. Was muss bei empirischer Forschung geleistet werden?

3.1 Welche Maßnahmen müssen unternommen werden?

Wenn die im vorigen Kapitel genannten rechtlichen Voraussetzungen – begründet auf Rechtsvorschrift oder informierte Einwilligung – gegeben sind, dann können personenbezogene Daten prinzipiell erhoben und verarbeitet werden. Allerdings ist dies an weitere Vorgaben und bestimmte Bedingungen gebunden:

- *Datensicherheit*: Im gesamten Forschungsprozess müssen technische und organisatorische Maßnahmen getroffen werden, die die Sicherheit der Daten jederzeit gewährleisten. Datenschutz in der empirischen Forschung beschränkt sich folglich nicht das einmalige Einholen von Unterschriften. Vielmehr verlangt es von Forscherinnen und Forschern, ihre Forschungsaktivitäten so zu gestalten, dass die Unversehrheit und Integrität der Daten gewahrt bleiben. Das Netzlaufwerk der Hochschule mit individuellen Zugriffsregelungen, Bearbeitungshistorien und Backup-Funktionen kann eine solche Datensicherheit leisten. Demgegenüber ist bei ungebundenen Datenspeichern (z.B. Laptop, USB-Stick) oder werbefinanzierte Speicherplätzen bei Nicht-EU-Firmen (z.B. Online-Cloud) ein hohes Risiko gegeben, dass unbefugte Personen Zugriff auf die sensiblen Daten erhalten. Entsprechende Maßnahmen zur Datensicherheit korrespondieren mit anderen bekannten Anforderungen an professionelle Forschungsarbeit, zum Beispiel Maßnahmen zum Schutz vor Verlust oder zur Qualitätssicherung.
- *Anonymisierung von Daten*: Aus dem Grundsatz der Erforderlichkeit folgt für empirische Forschungsvorhaben die Anforderung, personenbezogene Daten zu anonymisieren, sobald der Zweck der Forschung erfüllt ist. Stehen einer vollständigen Anonymisierung der Daten forschungsmethodische Gründe entgegen, z.B. wenn in Längschnitt-Analysen Probanden mehrfach kontaktiert werden sollen, dann muss zumindest der Personenbezug für Außenstehende weitgehend verhindert werden. Oftmals werden in solchen Fällen Techniken zur Pseudonymisierung angewendet, die die Echtnamen von Personen durch willkürlichen Zeichen ersetzen. Nur demjenigen Forscher, der über das Schlüssel-Dokument verfügt, ist bekannt, welche Person sich hinter welchem Pseudonym verbirgt. Datenschutzrechtliche Probleme in Forschungsprojekten entstehen vor allem immer dann, wenn über die Angaben und Daten von Individuen bislang noch unbekannte Phänomene erschlossen werden sollen. Vor allem qualitative Forschungsansätze sind daran interessiert, Daten möglichst *unmittelbar am Menschen* zu erheben und daraus neue Erkenntnisse zu generieren. Hier bedarf es einer besonders sorgfältigen datenschutzrechtlichen Prüfung und passenden technischen und organisatorischen Schutzmaßnahmen. Diese müssen sowohl die Anonymität der Probanden gewährleisten, als auch die Anonymität solcher Personen, die bei der Datenerhebung von den Probanden zufälligerweise genannt werden.

- *Nachvollziehbarkeit und Rechenschaftspflicht*: Zu guterletzt sind Regelungen zutreffen, welche Personen wann auf welche Daten Zugriff haben sollen. Dort, wo Forschungsprojekte arbeitsteilig in Kooperationen und Forschungsnetzwerken realisiert werden, sind für die datenschutzkonforme Verwendung von personenbezogenen Daten explizite Rollenkonzepte und Datenablaufpläne erforderlich. Diese müssen vor Projektbeginn so organisiert und formuliert werden, dass sie bei Anfrage den Aufsichtsbehörden vorgelegt werden können. Bei Kooperationen mit externen Personen und Organisationen können Nutzungsverträge die Rechte und Pflichten der einzelnen Partner regeln.

Fachstellen innerhalb der eigenen Universität und in hochschulübergreifenden Organisationen bieten für diese Aufgaben Unterstützung und Beratung (siehe Kapitel 5). In gut dokumentierten vorherigen Forschungsprojekten sind möglicherweise Materialien und Konzepte entstanden, die für ein neues Projekt erneut aufgegriffen und genutzt werden können.

3.2 Welche Maßnahmen sind geeignet?

Die wichtigsten Maßnahmen bei der Erhebung, Verarbeitung und Nutzung von personenbezogenen Daten für Forschungszwecke sind folgende Techniken:

- Mittels einer Einwilligungserklärung wird die Zustimmung von betroffenen Personen eingeholt und damit die rechtliche Grundvoraussetzung für die Erhebung und Verarbeitung personenbezogener Daten geschaffen. Wird die Einwilligung von minderjährigen Personen benötigt, so ist zusätzlich die Zustimmung der Eltern notwendig (siehe Kapitel 4).
- Mit einer Nutzungsvereinbarung werden Personen im Forschungsprozess auf die datenschutzkonforme Handhabe von Forschungsdaten verpflichtet. Eine solche Vereinbarung kann auch eine Erklärung zur Wahrung des Datengeheimnisses beinhalten.
- Ein Datenplan mit Löschkonzept stellt sicher, dass die erhobenen Daten ihrem definierten Zweck zugeführt werden, frühzeitig Maßnahmen zur Anonymisierung oder Pseudonymisierung erfolgen und nicht mehr benötigte Daten entfernt werden. Bei Pannen, im Konflikt- oder Schadensfall kann über einen Datenplan festgestellt werden, wer wann Zugriff auf welche Daten hatte.
- Eine Verfahrensbeschreibung (früher: *Verfahrensverzeichnis*) dokumentiert die rechtlichen Grundlagen, technischen und organisatorischen Maßnahmen, Zugriffsrechte und Verantwortlichkeiten bei der Erhebung und Verarbeitung personenbezogener Daten in einem Dokument. Eine Verfahrensbeschreibung zu erstellen ist zunächst Aufgabe der Forscherinnen und Forscher. Bei Bedarf unterstützen weitere Fachstellen (z.B. Rechenzentrum) oder der Datenschutzbeauftragte der Hochschule.

4. Die informierte Einwilligung

Die Zustimmung von betroffenen Personen ist rechtlich nur dann wirksam, wenn bestimmte Anforderungen bei der Durchführung und Formulierung der Einwilligungserklärungen eingehalten werden.

Grundsätzlich muss die Zustimmung als *informierte Einwilligung* erfolgen, d.h. der Betroffene muss vorab darüber informiert werden, in was er anschließend einwilligt. Ein Anschreiben zu einer Einwilligung sollte daher mindestens folgende Angaben beinhalten:

- eine präzise Bestimmung des Verantwortlichen der Datenerhebung, der gleichzeitig der Adressat der Einwilligungserklärung ist, an den der bzw. die Betroffene ihre Zustimmung richtet;
- konkrete Angaben über Zweck, Umfang und Art der erhobenen Daten sowie über den weiteren Weg der Datenverarbeitung;
- den ausdrücklichen Hinweis auf die Freiwilligkeit zur Mitwirkung an der Datenerhebung sowie das Recht auf Widerruf.

Der Betroffene muss darüber informiert werden, dass aus einer Verweigerung oder einem Widerruf keine weiteren Nachteile entstehen. Desweiteren ist zu erläutern, dass sich ein Widerruf nur auf die zukünftige Nutzung der Daten auswirkt, bisherige Nutzungen aber nicht rückgängig gemacht werden können. Bereits ausgewertete oder publizierte Daten sind von einem Widerruf nicht betroffen.

Einwilligungen sind nur dann wirksam, wenn sie vor der Datenerhebung erfolgt. Generell ist auch eine mündliche Einwilligung möglich; es empfiehlt sich aber, Einwilligungen schriftlich einzuholen, da sie im Datenmanagement einfacher und effizienter zu handhaben sind.

Die informierte Einwilligung besteht idealerweise aus drei Textbausteinen:

(1) einem Informationsteil, in dem das Forschungsvorhaben beschrieben und das Anliegen, die Bitte um Einwilligung, vorgetragen wird;

(2) den Hinweisen zum Datenschutz, die darüber aufklären, für welche Daten für welchen Zweck erhoben werden und wie diese Daten im weiteren Forschungsverlauf verwendet werden sollen;

(3) der Einverständniserklärung, die von den Betroffenen zu unterschreiben ist.

Es sei nochmal daran erinnert, dass in einer Einwilligung ausschließlich das Einverständnis der Betroffenen bezüglich ihrer personenbezogenen Daten im konkret genannten Forschungszweck geregelt wird. Sind zum Start des Forschungsprojektes bzw. zum Zeitpunkt der Datenerhebung weitergehende Nutzungen beabsichtigt, so müssen auch diese in der Einwilligung explizit gemacht werden. Typische weitergehende Zwecke sind der Einsatz in der Lehre und Weiterbildung, die Präsentation in der Öffentlichkeit, Weitergabe an andere Forscher sowie Folgenutzung und Bereitstellung in Datenbanken.

Dies kann über folgende Varianten realisiert werden:

- In der Einwilligung werden neben dem engeren Forschungszweck alle folgenden, angedachten Nutzungszwecke genannt und soweit möglich erläutert. Der Vorteil dieses Vorgehens ist offensichtlich: Die Forschenden erlangen eine weitreichende Flexibilität in der Datennutzung ohne weitere datenschutzrechtliche Abklärungen vornehmen zu müssen. Inwieweit Betroffene diesen Bedingungen folgen oder ihre Mitwirkung verweigern, bleibt eine Risikoabwägung.
- Datenschutzfreundlicher ist es, die Einwilligung nach Nutzungsformen abzustufen. Auch hier werden alle folgenden Nutzungszwecke genannt, die der Betroffene dann jeweils explizit zustimmt – oder eben auch nicht. Die Organisation und Handhabe solcher gestufter Einwilligungserklärungen ist aufwändiger, da jeder einzelne Datensatz geprüft werden muss, ob das Einverständnis vollständig oder nur teilweise gegeben wurde. Es empfiehlt sich bei dieser Variante, nach der Datenerhebung die Datensätze in zweckunterschiedliche Datenspeicher zu sortieren, z.B. Analysedatensätze, Datensätze für Einsatz in der Lehre usw.
- Die Einwilligung beschränkt sich zunächst bewusst auf den engeren Forschungszweck. In den erhobenen Daten werden Kontaktdaten mitgeführt, so dass nach einer Erstanalyse die Probanden nochmals angesprochen und um eine zusätzliche Einwilligung geben werden können. In dieser zweiten Stufe werden nur diejenigen Personen kontaktiert, deren Datensätze für weitere Nutzungen in Frage kommen, z.B. wegen besonders interessanter Statements, hoher Aussagekraft, beispielhafter Äußerung.

Um eine Einwilligungserklärung für das eigene, konkrete Forschungsprojekt zu gestalten, bieten verschiedene Quellen, etwa die Internetseite https://www.forschungsdaten-bildung.de des Deutschen Instituts für Internationale Pädagogische Forschung, Anregungen und Textbausteine an. Welche Angaben – über die hier genannten Aspekte hinaus – zwingend enthalten sein müssen, ist von den bundeslandspezifischen Datenschutzgesetzen und Spezialgesetzen in bestimmten Praxisfeldern abhängig. Die in diesen Quellen enthaltenen Formulierungen müssen daher immer geprüft und auf den Einzelfall angepasst werden. Auch können die Bausteine anders angeordnet und miteinander kombiniert werden. Auch wenn die Vorlagen zum schnellen Copy&Paste verleiten: Es bleibt letztendlich die Aufgabe der Forscherin und des Forschers, die für ihre bzw. sein Forschungsprojekt passende Einwilligungserklärung zu gestalten und auch alle folgenden Schritte mitzudenken, vor allem die Aufbewahrung der Erklärungen und das Prozedere bei einem Widerruf.

5. Welche Anforderungen treten im Kontext von Schulforschung hinzu?

Viele pädagogische und fachdidaktische Studien können nur dann die gebotene forschungsmethodische Validität erreichen, wenn im unmittelbaren Schul- bzw. Unterrichtskontext entsprechende Primärdaten erhoben und Effekte getestet werden. Erst hier werden, anders als in Laborsituationen, jene Einflussgrößen sichtbar, die im sozialen Gefüge einer Schul- oder Klassengemeinschaft bedeutsam sind. Wenn empirische Forschungen in Schulen und Klassenzimmern durchgeführt werden sollen, sind neben den generellen Datenschutzbestimmungen weitere gesetzliche Anforderungen und Abhängigkeiten zu beachten. Insbesondere die je nach Bundesland unterschiedlich gehaltenen Schulgesetze nennen mal mehr und mal weniger präzise Vorgaben, unter denen mündliche oder schriftliche Befragungen, Unterrichtsbeobachtungen oder fachdidaktische Versuche durchgeführt werden können. Die Bandbreite reicht vom einfachen Einverständnis durch den Schulleiter bis hin zur Genehmigungspflicht in der Schulverwaltung. Dabei ist es zunächst unerheblich, ob ein größeres Forschungsvorhaben oder eine lediglich kleine Studie im Rahmen einer studentischen Forschungsübung an einer Schule stattfinden soll. Die folgenden Angaben erläutern die wesentlichen Spezifika. Die konkreten Anforderungen müssen aus den jeweiligen Schul- und sonstigen Landesgesetze sowie ggf. weiteren Rechtsgrundlagen entnommen werden. Eine gute Übersicht über die länderspezifischen Besonderheiten für Befragungen an Schulen gibt die bereits oben genannte Plattform *Forschungsdaten Bildung*.

5.1 Moment *minderjährige Probanden*

Voraussetzung für die informierte Teilnahme an einer empirischen Erhebung ist die sogenannte Einwilligungsfähigkeit. Die Rechtsgebung und Rechtssprechung gehen davon aus, dass erst ab dem Alter von 14 Jahren junge Menschen über ausreichend Einsichtsfähigkeiten verfügen, um die Ziele einer wissenschaftlichen Studie zu verstehen, um in diese selbst einwilligen zu können. Desweiteren gilt bis zur Volljährigkeit das elterliche Erziehungsrecht. Nach den meisten Landesgesetzen ist sowohl die Einwilligung des minderjährigen Schülers bzw. der minderjährigen Schülerin als auch die Einwilligung beider Erziehungsberechtigter notwendig. Die Einwilligung der Erziehungsberechtigten verpflichtet das minderjährige Kind jedoch nicht zur Teilnahme. Verweigert eine Schülerin oder ein Schüler ihre bzw. seine Mitwirkung an der Studie, so überstimmt dies das Elternvotum. Mitunter kann auch bei volljährigen Personen die Verständnisfähigkeit offensichtlich nicht gegeben sein, beispielsweise bei geistigen Einschränkungen oder starken sprachlichen Hindernissen. In solchen Fällen ist es ratsam, entweder die Informationen zielgruppenspezifisch anzupassen, in mehrsprachigen Ausführungen bzw. sogenannter *einfacher Sprache* vorzuhalten, oder die Einwilligung und anschließende Erhebung von vertrauten Personen durchführen zu lassen, z.B. durch sogenannte *Stellvertreter-Interviews*.

5.2 Moment *Freiwilligkeit*

Neben der Einwilligungsfähigkeit bildet die Freiwilligkeit der Teilnahme eine Voraussetzung, um eine rechtsgültige Einwilligung zu erhalten. Bei Erhebungen an Schulen sehen sich die Schülerinnen und Schüler sowie gleichermaßen auch die dort unterrichtenden Lehrerinnen und Lehrer verschiedenen Abhängigkeiten ausgesetzt. Forscherinnen und Forscher müssen dafür Sorge tragen, dass die Probanden tatsächlich eine echte Wahl haben, die Datenerhebung zu verweigern, ohne selbst einen Nachteil zu erleiden. Wenn einzelne Eltern einer Datenerhebung nicht zustimmen, müssen die betroffenen Kinder so von der Untersuchung ausgenommen werden, dass diese dem Unterricht weiter folgen können. Beispielsweise ist bei Videoaufnahmen die Kamera so im Raum zu positionieren, dass diese Kinder nicht erfasst werden. Gleichermaßen ist dann aber auch zu gewährleisten, dass diese Kinder nicht durch soziale oder räumliche Ausgrenzung benachteiligt werden.

5.3 Moment *Hausrecht*

Die meisten Schulgesetze sind so gestaltet, dass empirische Erhebungen an einer Schule vom verantwortlichen Schulleiter bzw. der Schulleiter vorab genehmigt werden müssen. Sie müssen gewährleisten, dass die Schule und die Schulaufsichtsbehörden nicht durch ein Übermaß an Datenerhebungen in ihrem originären Bildungsauftrag gehindert und in ihren organisatorischen Möglichkeiten überfordert werden. Wenn ein Schulleiter die ordnungsgemäße Durchführung des Unterrichts oder den Schulfrieden gefährdet sieht, kann dieser die Zustimmung zur Erhebung jederzeit zurückziehen.

Selbstredend müssen bei einer Veröffentlichung sämtliche Namen und Personenbezeichnungen anonymisiert werden. Ebenso dürfen Name und Ort der Schule nicht mehr in anschließenden Publikationen erscheinen. Einzelne Schulgesetze beinhalten mit Hinweis auf den Datenschutz Vorgaben zur Löschung der erhobenen Daten.

5.4 Moment *Persönlichkeitsrechte*

Immer beliebter werden Erhebungsmethoden, die in der Unterrichtspraxis Bild- bzw. Videodaten erheben (*Videografie*). Da solche Daten besonders viele personenbezogene Informationen festhalten, sind besonders dezidierte organisatorische und technische Maßnahmen erforderlich. So müssen alle (potentiell) abgebildeten Personen bzw. deren Erziehungsberechtigten vorab schriftlich zustimmen, insbesondere wenn angedacht ist, Bildmaterial über den Unterrichts- bzw. Analysekontext hinaus weiterzuverwenden, beispielsweise in wissenschaftlichen Publikationen oder zur Öffentlichkeitsarbeit des Forschungsprojektes bzw. der Hochschule. Mit der Schule muss abgestimmt werden, ob die Einwilligungserklärungen dort oder an der Hochschule aufbewahrt werden.

5.5 Moment *sensible und konfliktäre Ergebnisse*

In aller Regel unterstützen Schulleitung, Lehrkräfte und Eltern einer Schule das vorgetragene Forschungsvorhaben, da sie mitunter auch selbst an den Daten und Ergebnissen interessiert sind. Die Datenschutzgesetze sehen zwar ein Recht auf Dateneinsicht der Betroffenen vor, hier die Schülerinnen und Schüler, nicht aber weiterer anderer Personen (z.B. Eltern). Wenn ein kooperatives Forschungskonzept vorsieht, die Ergebnisse auch für die Unterrichts- und Schulentwicklung vor Ort zu nutzen, empfiehlt es sich daher, dies auch als weiteren Zweck in der Einwilligungserklärung aufzunehmen. Bisweilen spiegeln Wissenschaftler die erhobenen Daten an die Beforschten zurück, um die Güte der gewonnenen Erkenntnisse zu überprüfen, bekannt als *kommunikative Validierung*. Hier ist zu beachten, dass ungünstige Ergebnisse, Schwächen oder Negativbeschreibungen im Datenmaterial, die an eine Schule kommuniziert werden, die Persönlichkeitsrechte von Einzelnen tangieren können. Es empfiehlt sich daher, Ergebnisse nicht zu voreilig rückzumelden und bei der Rückmeldung jegliche Personenrückbezug auszuschließen, beispielsweise durch Schwärzen oder Anonymisieren.

Die Genehmigungsverfahren sind in den einzelnen Bundesländern verschieden umfangreich und unterschiedlich organisiert. Es empfiehlt sich, bereits vor der Kontaktaufnahme ein Grobkonzept der organisatorischen und technischen Maßnahmen vorzubereiten und dieses dann mit der zuständigen Stelle zu besprechen; in den meisten Fällen ist das die Schulleitung. Wichtig für die Genehmigung sind präzise Angaben über das wissenschaftliche Interesse und den Zweck der Erhebungen, die geplante Art und Vorgehensweise der Untersuchung in Form von Ablaufplänen und beispielhaft skizzierten Instrumenten, sowie nicht zuletzt Angaben zu den Projektverantwortlichen und ggf. weiteren Beteiligten.

6. Empfehlungen zur Weiterarbeit

Die hier vorgestellten Prinzipien sollen vor allem der Orientierung und Erstinformation beim Start in ein Forschungsprojekt geben. Die folgenden praxisorientierten Empfehlungen wollen aufzeigen, dass Datenschutz in Forschungsprojekten kein Hexenwerk darstellt:

- Unbedingt:
  - Holen Sie vor ausnahmslos jeder Erhebung die Einwilligung der Betroffenen ein.
  - Geben Sie in Informationsschreiben und Einwilligungserklärungen die verantwortliche Stelle sowie die Kontaktdaten des Projektverantwortlichen an.
  - Achten Sie jederzeit auf geregelten und geschützten Zugang und Zugriff der Daten.

- Hilfreich und empfehlenswert:
  - Weniger Daten vereinfachen den Forschungsprozess: Fokussieren Sie bereits bei der Datenerhebung Ihr primäres Forschungsinteresse.
  - Nutzen Sie Standardverfahren und Methoden des Datenschutzes, die ihre Hochschule für Sie schon bereitstellt.
  - Bedenken Sie bereits bei der Planung der Datenerhebung auch zukünftige (Einsatz-)Szenarien bis hin zur Frage, wann die Daten spätestens gelöscht werden sollen.
  - Erstellen Sie einen Plan zur Datenverarbeitung: „Von wem werden wann welche Daten bearbeitet?“
- Keinesfalls:
  - Sammeln Sie keine Daten, wovon Sie nicht wissen, wie sie weiter analysiert werden sollen.
  - Transportieren Sie keinesfalls Forschungsdaten auf ungeschützten oder ungebundenen Datenspeichern.

Welche Stellen können weitere Informationen und Beratung geben?

- Vor Ort: Verwaltungsleitungen und Justiziare in der Verwaltungsebene von Universitäten und Hochschulen sowie Expertinnen und Experten in Servicestellen zur Forschungsförderung und an den Rechenzentren sind häufig mit datenschutzrechtlichen Fragen konfrontiert. Sie verfügen über entsprechendes Fachwissen und Erfahrungswerte, v.a. zu Spezifika an Ihrer Hochschule. Darüber hinaus verfügt jede Organisation über einen Datenschutzbeauftragten, der unabhängig agiert und Anfragen vertraulich behandelt.
- Überinstitutionell: Eine Orientierung für generelle Aspekte des Datenschutzes wie auch vertieftes Wissen zu besonderen Aspekten im Forschungskontext bieten verschiedene Datenschutzfachstellen von Hochschulverbünden, Forschungszentren und wissenschaftlichen Fachgesellschaften. Über ihre Online-Plattformen, Kommunikationskanäle und Weiterbildungsangebote vermitteln sie Datenschutzgrundlagen und informieren zu veränderten Gesetzen, neu auftretenden Phänomenen, Good-Practice-Projekten usw. Beispiele: Zentrale Datenschutzstelle der baden-württembergischen Universitäten, GESIS Datenarchiv, Arbeitskreis der Datenschutzbeauftragten der außeruniversitären Forschungseinrichtungen.

Informieren Sie Ihre Vorgesetzten und Hochschulleitung frühzeitig über Ihr Forschungsvorhaben und fragen Sie nach Vorgaben und Standards zum Datenschutz. Mitunter liegen aus vorherigen Projekten erprobte Instrumente bereit, die Sie nutzen können. Im Zweifel oder Konfliktfall steht der Datenschutzbeauftragte Ihrer Organisation für vertrauliche Gespräche bereit.

Literatur

Forschungsdaten Bildung: https://www.forschungsdaten-bildung.de. Frankfurt a.M: Deutsches Institut für Internationale Pädagogische Forschung (DIPF) (letzter Zugriff: 01.08.2018).

Gebel, Tobias/Grenzer, Matthias/Kreusch, Julia/Liebig, Stefan/Schuster, Heidi/Tscherwinka, Ralf/Watteler, Oliver/Witzel, Andreas (2015): Verboten ist, was nicht ausdrücklich erlaubt ist: Datenschutz in qualitativen Interviews. In: Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 16, 2. Online unter: http://www.qualitative-research.net/index.php/fqs/article/view/2266 (letzter Zugriff: 01.08.2018).

Kinder-Kurlanda, Katharina/Watteler, Oliver (2015): Hinweise zum Datenschutz. Rechtlicher Rahmen und Maßnahmen zur datenschutzgerechten Archivierung sozialwissenschaftlicher Forschungsdaten. In: GESIS-Papers 1/2015. Köln: GESIS – Leibniz-Institut für Sozialwissenschaften. Online unter: http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_papers/GESIS-Papers_2015-01.pdf (letzter Zugriff: 01.08.2018).

Rat für Sozial- und Wirtschaftsdaten (2017): Handreichung Datenschutz. Berlin: RatSWD. Online: https://doi.org/10.17620/02671.6 (letzter Zugriff: 01.08.2018).

Wellbrock, Rita/Metschke, Rainer (2003): Datenschutz in Wissenschaft und Forschung. Herausgegeben vom Berliner Beauftragter für Datenschutz und Informationsfreiheit und vom Hessischen Datenschutzbeauftragten. Online unter: http://www.datenschutz-berlin.de/attachments/47/Materialien28.pdf (letzter Zugriff: 01.08.2018).

ANN-CHRISTIN BUTTLAR

Transkription

Analytische Aufbereitung gesprochener Sprache für die empirische Sozialforschung

1. Einleitung

Gesprochene Sprache ist ein wichtiges Mittel zur Organisation sozialer Interaktion, sowohl in privaten als auch institutionellen Kontexten. Daher ist gesprochene Sprache vielfach Gegenstand empirischer Untersuchungen. Die Merkmale mündlicher Kommunikation wie z.B. ihre Flüchtigkeit machen aber einige Schritte der Datenerfassung und -aufbereitung nötig. So muss gesprochene Sprache stets in Form von Audio- oder Videodaten aufgezeichnet werden, um sie rückholbar zu machen. Allerdings reicht diese Form der Datenaufzeichnung nicht aus, um gesprochene Sprache den verschiedenen Forschungsfragen der empirischen Sozialforschung zugänglich zu machen. Die Transkription ist daher für die Wissenschaft ein elementares Werkzeug zur Datenaufbereitung. Der vorliegende Artikel beschäftigt sich mit der Transkription gesprochener Sprache und ihrem Nutzen für wissenschaftliche Analysen.

Dabei ist anzumerken, dass es nicht das *eine* richtige Transkript gibt. Es gibt viele verschiedene Erscheinungsformen und Notationen, die je nach Fragestellung und Forschungsrichtung variieren. Da in der Kürze des Artikels unmöglich alle Varianten vorgestellt werden können, werde ich mich auf die Notation des gesprächsanalytischen Transkripts nach GAT 2 (Selting et al. 2009) konzentrieren. Diese Notation bietet zum einen für die Gesprächsanalyse einen detaillierten und umfassenden Zugriff auf die Transkription gesprochener Sprache, da vor allem sprachlich-formale und sequenzielle Aspekte erfasst und im Transkript sicht- und greifbar dargestellt werden können. Zum anderen liefert diese Notation aber auch Nicht-Gesprächsanalytikern einen angemessenen Zugang zu Formen mündlicher Kommunikation, auch wenn sich diese mehr auf inhaltliche denn auf sprachliche Aspekte beziehen, wie ich in Kapitel 2 zeigen werde. In Kapitel 3 werde ich anschließend vorstellen, wie ein (gesprächsanalytisches) Transkript angefertigt wird. Auch hier werde ich mich auf die Konventionen von GAT 2 konzentrieren. Das Vorgehen ist aber auf jegliche andere Konvention übertragbar.

2. Grundsätzliches zur Transkription: Was ist ein Transkript und wofür ist es gut?

Sprache ist in empirischen Forschungsrichtungen ein zentrales Medium zur Datenerhebung, selbst wenn der Fokus nicht auf den sprachlichen Äußerungen selbst liegt, sondern allein auf der inhaltlichen Ebene: Schriftlich oder mündlich erhobene Daten sind die Grundlage empirischer Forschung. Während aber schriftliche Ergebnisse oft unmittelbar verwendet werden können, muss gesprochene Sprache zunächst in eine Form überführt werden, die eine Untersuchung möglich macht. Gesprochene Sprache ist flüchtig und in diesem Sinn auch nicht rückholbar, denn sie unterliegt den Bedingungen von Raum und Zeit (Dürscheid 2006). Aus diesem Grund muss sie in irgendeiner Form konserviert werden. In der Regel geschieht das zunächst durch Audio- oder Videographie. Die Aufnahme gesprochener Sprache mit Audio- oder Videogeräten macht nun zwar eine Wiederholung möglich, aber dennoch bleibt das Gesprochene flüchtig und für die Analyse wenig greifbar. Daher folgt in den meisten Fällen noch ein weiterer Schritt: die Verschriftung (Transkription) der aufgenommenen Daten.

Der Begriff der Transkription beschränkt sich nicht ausschließlich auf die wissenschaftliche Aufarbeitung gesprochener Sprache, sie dient in vielen Fällen auch der Verständnissicherung, indem undeutliche Fernsehinterviews untertitelt oder handgeschriebene Texte ‚abgetippt' werden (vgl. Hausendorf 2017). Das Transkribieren gesprochener Sprache hat in empirischen Forschungskontexten aber eine gänzlich andere Funktion: „Weder wird dadurch etwas verfügbar gemacht, was wir vorher nicht verstanden hätten, noch wird dabei etwas ‚umgeschrieben', was schon vorher geschrieben stand" (ebd., 217). Vielmehr geht es beim Transkribieren um eine möglichst genaue Verschriftung des Gesprochenen, die einen neuen Zugang zur Analyse anbietet. Hausendorf (ebd., 219f.) schreibt dem angefertigten Transkript den Status von etwas Neuem zu, das „nicht nur konservierenden, sondern eben auch produktiv-schöpferischen Charakter" hat:

> Man weiß inzwischen in der Phonetik, aber auch in der Linguistik der Alphabetschrift (und des Schriftspracherwerbs) sehr gut darüber Bescheid, wie sehr mit der (Alphabet-)Schrift die auditive Wahrnehmung von Sprachschall nicht nur geschult, sondern überhaupt erst kreiert wird. (Hausendorf 2017, 219f.)

Die schriftliche Form mündlicher Äußerungen ist in diesem Sinn eine Verfremdung, die es ermöglicht, bestimmte Dinge anders wahrzunehmen. So werden vor allem unbewusste Prozesse sicht- und greifbar. Damit ist für jede wissenschaftliche Publikation ein wichtiges Beleginstrument geschaffen. Zunächst hilft die Verfremdungsfunktion der Transkription aber bei der Analyse gesprochener Daten. Deppermann (2008, 40) hält dazu fest:

> Transkripte bieten auch für die Auswertungspraxis Vorteile. Sie ermöglichen die extensive und beliebig oft wiederholbare Analyse eines Datensegments, während AV-Materialien aufgrund ihrer zeitlichen Dynamik und der Flüchtigkeit der Wiedergabe umständlicher zu handhaben (Vor- und Zurückspulen) und mehr von schwankenden Aufmerksamkeits- und Gedächtnisleistungen der Analytiker ab-

hängig sind. Transkripte bieten einen leichteren Überblick über Verläufe und ermöglichen es, ein Datensegment beliebig lange in bezug auf [sic!] unterschiedliche Gesichtspunkte in verschiedenen Auflösungsniveaus zu untersuchen. Zudem kann man verschiedene Textstellen simultan vergleichen oder zu Vergleichszwecken zusammenstellen. Schließlich zwingt die Transkription dazu, sich exakt darüber Rechenschaft abzulegen, wie dasjenige Merkmal zu beschreiben ist, das für eine Interpretation ausschlaggebend ist. Diese Explikation ist die Voraussetzung dafür, daß [sic!] Annahmen über Eigenschaften und Zusammenhänge in Gesprächsprozessen wissenschaftlich kommuniziert werden können (und nicht nur Eindrücke und Anmutungen bleiben, deren Grundlage nur empfunden, nicht aber erkannt ist).

Das folgende Transkript soll Deppermanns Ausführungen veranschaulichen. Es stammt aus dem abgeschlossenen DFG-Projekt DASS (Diskursstile als sprachliche Sozialisation) unter Leitung von Uta Quasthoff und wurde u.a. in Buttlar (2017) diskutiert.

Beispiel 1:

```
001      XX   ((Unruhe))
002      L    ja;
003           kannst DU uns jetzt vielleicht auch äh_jetz sAgen?
004           oder die ANderen kInder?
005           vielleicht können die uns AUCH sagen;=
006           =vielleicht kann_es_uns der tOm (.) AUCH sagen?
007           so_n tIEr (du)/das: KANN man ja nicht nur einfach so in
                  der wOhnung halten,=
008           =oder in der KLASse?
009           das B[RAUCH ja/    ]
010      Tom       [NEIN (diese?)]
011      L    wir wollen uns ja auch noch überLEgen;=
012           =was das vielleicht alles !BRAU:[CHT!.]
013      AK?                                  [nein.]
014      L    Oder?
015           VORher-
016           hab ich gAnz verGESsen?
017           solltn wir uns vielleicht noch drÜber überLEgen-
018           was/ wie das so AUSsieht,
019           (.) <<zu Tom> hAttest du nicht auch von kaNINchen was
                  gesAgt?
020           geRAde?>
021      Tom  ich hab ja OFT kanInchen.
022      L    RICHtig.
023           AUCH gut.
024      Tom  kaNI[Nchen.
025      L         [Okay.
026           dann gucket euch doch mal das kaninchen hierdrauf AN?
027           auf dem [BILD?   ]
028      Ni?          [ICH brau]ch nicht-
029      L    [und/]
030      Ni?  [weil] ich SELber kaninchen hab im gArten.
031      L    ja PRIma;
032           dann SAG uns das doch mal,
033           wie das AUSsieht.
034           zum BEIspiel.=
035           =wAs HAT es (hier als).
036           (.) NIna.
037      Tom  einen RÜcken,
```

```
038           ein BAUCH,
039           BEIne-
040      L    (.) kOmm mal rAUs und ZEIG_S uns.
041      Tom  ((geht zur Tafel))
042           (--)
043           BEIne?
044      L    RICHtig.
045      Tom  äh: RÜcken?
046           o[der BAUCH;]
047      L     [JA:::?    ]
048           RICHtig,=
049           =RÜcken.
050           GEnau.
051           du hast GRAD noch was gesAgt?
052           BAUCH?
053           wo ist DER?
054      Tom  und die ohren sind HIER?
055      L    <<bestätigend> hm hm?>
056      X    ((ruft etwas dazwischen, unv.))
057      Tom  und die HAsenhAAre.
058      L    R[ICHti::g. ]
059      X     [(a:::ha*-)]
060      AK   ((meldet sich))
061      L    der hat ja HAAre. =ne?
062           da wollte die ANNkristin was zu sA[gen   .]
063      AK                                     [eh_he*;]
064      L    DANN (mohamed).
065           ja?
066      X    ((unverständlich))
067      AK   UNser HA[se ist EIgentlich*-
068      Tom?         [frau SCHMIDT-=
069           DARFST du mir schon ein pUnkt drauf machen?
070      L    ja SPÄter.=
071           jEtzt NICHT.
072      AK   UNser HAse ist !EI!gentlich NUR brAU:n?
073      L    <<bestätigend> hm hm?>
074      AK   [((unverständlich))
075      XX   [((große Unruhe))
076      L    psch:::t.
077           ((sorgt für Ruhe))
078      XX   [((weiter Unruhe))
079      L    [ANNkristin ist dran.
080      AK   im WOHNzimmer und im flUr frei RUM.
081           (der KOMMt) unterm sessel HER?
082           und SPRINGT?
083      Tom  äh KArin.
084      AK   auf meine BEIne?
085           auf EINmal,
086      Tom  KA[rin.
087      AK     [DENK ich (es fast nicht).
088      Tom  ((flüstert mit jemandem))
089      L    ((sieht Annkristin an, dreht aber Tom nach vorne))
090      AK   [is da so/ was IST das (so DAS denn);
091      Tom  [((flüstert weiter, während L ihn festhält))
092      AK   dann sieht er mich die ganze zeit AN?=
093           =und SCHNUPpert.
094      L    mh:::;
095           geNAU.=
096           =ihr habt schon geSAGT;=
097           =ein SCHNURRbart hat der?
```

```
098           der HAse?
099           und?
100           du hast geSAGT?
101      XX   [((Unruhe, einige Kinder rufen durcheinander))
102      L    [der sieht BRAUN aus?
103           was IST denn brAUn hier bei dem (.) kaNINchen.
```

Abgedruckt ist ein Transkript im gesprächsanalytischen Format nach den Konventionen von GAT 2[1] (Selting et al. 2009). Es ist etwas länger, damit einige Dinge veranschaulicht werden können, ohne im Detail auf die Analysen selbst einzugehen[2]. Wir können den Abschnitt leicht in mehrere klar zu benennende Sequenzen aufteilen, indem auf die Zeilenangaben des Transkripts verwiesen wird.

- Von Z. 001-061 wird zunächst der Gegenstand des Unterrichtsgesprächs verhandelt, erste Schüleräußerungen werden getätigt und von der Lehrperson evaluiert. Anhand der lehrerseitigen Formulierungen lässt sich gut erkennen, dass der Arbeitsauftrag *online*, d.h. mehr oder weniger spontan produziert wird und die Lehrperson an die Schüleräußerungen anzuknüpfen versucht: Die Aussagen sind von Abbrüchen, Reformulierungen und überlappenden Sequenzen mit den Schülerinnen und Schülern gekennzeichnet. Vieles folgt in dieser Gesprächssequenz schnell aufeinander. Ohne das Transkript wäre es schwer, die Gesprächsstruktur nachzuvollziehen.
- Von Z. 060-093 erhält Annkristin den Turn, den sie über eine längere Zeit hält, der aber auch von großer Unruhe in der Klasse begleitet wird. Ohne eine Transkription wäre es sicher kaum möglich, die von Annkristin produzierte Erzählung herauszugreifen und einer Analyse zugänglich zu machen. So aber können die Textteile gleichsam ‚aus der Zeit herausgenommen'und von den störenden Zwischenrufen ‚bereinigt' werden.
- In Z. 094-102 lässt sich die Rückmeldung der Lehrperson auf Annkristins Beitrag und auf die vorherigen Schülerrückmeldungen rekonstruieren. Es fällt auf, dass die Erzählung als solche nicht aufgegriffen wird, stattdessen bleibt einzig der Zusatz „und? / du hast geSAGT? / der sieht BRAUN aus?", Z. 099+100+102). Das lässt sich sehr leicht durch die entsprechende Textstellenbelege demonstrieren.

Ungeübten Leserinnen und Lesern wird es sicher nicht leicht fallen, alle Einzelheiten der Transkriptkonventionen exakt zu deuten, daher soll dies im Folgenden anhand der ersten Zeilen dieses Transkripts verdeutlicht werden.

Zunächst fällt auf, dass zur Transkription nicht die Lautschrift IPA (Internationales phonetisches Alphabet) genutzt wird, die ‚normale Schrift' reicht für die Erfassung des Geäußerten aus. Allerdings wird nach GAT 2 alles kleingeschrieben,

---

[1] Die Abkürzung steht für „*G*esprächs*A*nalytische *T*ranskriptionssystem", die „2" zeigt an, dass es sich dabei um die überarbeitete zweite Version der Konventionen handelt, die in Selting et al. 2009 ausführlich vorgestellt wird.

[2] Das Beispiel wird unter gesprächsanalytischer Perspektive lehrerseitigen Umgangs mit Schüleräußerungen ausführlich in Buttlar 2017 diskutiert.

die Großschreibung ist für die Notation von Akzenten reserviert. Bei dem Fokusakzent, d.h. der am stärksten betonten Silbe wird diese komplett großgeschrieben, Nebenakzente werden zuweilen, d.h. wenn sie besonders deutlich wahrgenommen werden können, durch Großschreibung der Vokale gekennzeichnet („kannst DU uns jetzt vielleicht auch äh_jetzt sagen?“, Z. 003). Die Unterstriche (‚_‘) zeigen an, dass einige Wörter verschliffen werden. Der schnelle Anschluss zwischen den Äußerungssequenzen wird durch die Gleichheitszeichen am Ende des einen und am Anfang des schnell anschließenden Äußerungssegments markiert (Z. 005-006). Die Schrägstriche (‚/‘, z.B. Z. 007) zeigen an, dass ein Äußerungssegment abgebrochen wird. Eckige Klammern zeigen wiederrum an, welche Äußerungen verschiedener Gesprächsteilnehmerinnen und Gesprächsteilnehmer sich überlappen. Auch Pausen werden notiert. Je nach Länge gibt es hier unterschiedliche Notationen: In Z. 006 ist eine Mikropause zu sehen, die mit der Zeichenfolge *(.)* notiert wird. Diese Pause wird geschätzt, sie dauert bis zu 0.2 Sekunden. Erst ab ca. 0.5 Sekunden Pause werden diese gemessen und genau notiert, z.B. *(2.0)*. Eine Übersicht der Pausennotationen ist in Selting et al. (2009, 391) zu finden.

Mit diesen Konventionen kann das Unterrichtsgespräch im Detail analysiert werden: In Z. 009-010 führt der überlappende Zwischenruf eines Schülers z.B. dazu, dass die Lehrerin ihre Äußerung unterbricht. In Z. 012-013 kommt es dagegen nicht zu einem lehrerseitigen Abbruch, eventuell, da Annkristins Zwischenruf sehr leise ist: Das ‚<p> steht wie in der Musik als Metakommentar für ‚piano‘, also für ‚leise‘, die äußeren spitzen Klammern zeigen dabei die Reichweite des Metakommentars an, in diesem Fall erstreckt sich das Merkmal über die gesamte kurze Äußerung („<<p> nein.>“).

Eine solche detaillierte Transkription ist für Gesprächsanalytiker unerlässlich. Über die sprachliche Oberfläche wird aus Sicht der Konversationsanalyse soziale Wirklichkeit erst hergestellt, die sprachlichen Mittel übernehmen dabei vielfältige Funktionen (Bergmann 1994)[3]. Daher muss die Transkription den analytischen Zugriff darauf anbieten, also möglichst detailgetreu abbilden, was auf der sprachlichen Oberfläche sichtbar ist.

Dennoch halte ich diese Form der Transkription auch für andere empirische Fragestellungen, die eher auf den Inhalt fokussieren, für ein wichtiges Werkzeug. Eine ‚bereinigte‘ Abschrift der Sequenz ohne Pausen, Überlappungen, Betonungen etc. können den Eindruck des Gesagten leicht verfälschen. Deppermann (2013) macht dies am Beispiel eines Interviews deutlich, das – anders als die oben gezeigte Unterrichtsinteraktion – häufig als ‚mündlicher Text‘ behandelt wird. Dabei wird außer Acht gelassen, dass jegliche Form von mündlicher Kommunikation eine Interaktion darstellt (Hausendorf 2007), d.h. auch ein Interview eine

[3] Ziel des Artikels ist es, die Transkription als solches vorzustellen, daher verzichte ich an dieser Stelle auf eine detaillierte Ausführung analytischer Prinzipien der Konversationsanalyse, da sie dem Prozess des Transkribierens generell nachgeordnet sind. Für einen Überblick über diese Methode empfehle ich Heller/Morek 2016.

Form von Interaktion darstellt und die Antworten der Interviewten in Abhängigkeit der vorherigen Äußerungen entstehen.

Deppermann zeigt zur Veranschaulichung eine Äußerung einer Interviewten als „typische[r] Darstellung einer Befragtenantwort“ (Deppermann 2013, 8) innerhalb der Sozialforschung, die sehr entschlossen und zielstrebig wirkt:

Beispiel 2, aus Deppermann (2013, 8):

> ‚Ich selber bin son Typ, ich bin zwar spontan und geh auch auf die Leute zu, aber nicht so, dass ich nun den herauskehren würde: Ich bin der Ossi und nun habt alle Mitleid mit mir und ich möcht jetzt was geschenkt haben. Ich hab mir das mehr so aus der Distanz angesehen und ich bin son Typ, ich möchte das lieber alles selber allein beziehungsweise selber so durch die Tiefen und Höhen gehen. Ich brauch da keinen, der sich da meiner erbarmt und sacht: Hach und naja und wolln wir doch mal. Das ist nicht so.‘
>
> #1 Aus einem Interview zum Erleben der Wende 1989, Berliner Wende-Korpus, Institut für Deutsche Sprache (IDS), BWO11

Wenn man nun aber die Pausen und Gesprächspartikel hinzunimmt, zeigt sich ein anderes Bild: Die Probandin hat erhebliche Probleme mit dem Antwortbeginn und der Formulierung, was die Pausen, die Redeannahmepartikel am Anfang sowie das Ein- und Ausatmen deutlich zeigen. Dadurch erscheint ihre Äußerung weniger als ein klares Statement, es lässt sich der Prozess der Entstehung nachvollziehen.

Beispiel 3, aus Deppermann (2013, 9):

```
011            (4.0)
012   BW011:    mh-
013            (2.8)
014   BW011:    ja also;
015            (2.3)
016   BW011:    ick selba bin SO:_n typ, (.)
017             ick bin zwar sponTAN und-
018             geh och auf die lEUte ZU:-
019             (0.3) aba nicht SO:-
020             (0.7) dass ick nun (.)
021             dEn heRAUSkehren würde ick bin der Ossi,
022             und nu [habt alle ] <<h> mitLEID mit mir;>
023   Int:              [<<p> mhm.>]
024   BW011:    <<all> und ick möchte jetz wat jeSCHENKT ha:m;>
025             (0.7) ick hab mir mit dit mehr so aus der dis (.) distAnz
                ANjesehn und-
026             (1.6)
027   BW011:    ich bin SO:_n typ,
028             (0.4) ick möchte det lieba allet selBA::;
029             (2.3)
030   BW011:    alLEEN beziehungsweise,
031             selba so durch die TIEFN und HÖHN gehen;-=
032             =<<all> ick brauch da KEEN->
033             (1.7)
034   BW011:    der sich da mir (.) meiner erBARMT und sacht,
035             !HA:CH!
036             und na!JA::!
037             Und (0.4) WOLLN wa doch mA:l-
038             und .h
```

```
039             (1.6)
040   BW011:    ((atmet tief ein und aus))
041             det is nicht SO:;
042             (2.8)
```

Beispiel 2 und 3 sind beides Transkriptionen mündlicher Gesprächsdaten, doch der Einblick in die Strukturen wird durch eine gesprächsanalytische Transkription deutlicher. Die Antwort der Probandin wirkt nun weniger forsch und selbstsicher. Auch wenn es in Forschungskontexten außerhalb der Gesprächsforschung eher um inhaltliche Aspekte des Interviews geht, färbt die Darstellung der sprachlichen Oberfläche doch auf die Interpretation ab, da die sprachliche Form Hand in Hand mit dem Inhalt geht. Es lässt sich z.B. besser herausarbeiten, worauf die Probandin den Fokus legt (siehe Betonungen, markiert durch Großschreibung). Deppermann geht sogar noch einen Schritt weiter und verweist auf die Frage der Interviewerin, so dass der interaktive Einfluss auf die Interviewte deutlich wird und die Antwort im Kontext besser zu verstehen ist:

Beispiel 4, aus Deppermann (2013, 10):

```
001   BW011:    det wird Immer noch irgendwie-
002             (0.4) ne kluft GEBN zwischen ost und west .hh
003             (1.8)
004   INT:      und das is auch ein LANGwieriger schwerer prozess, (.)
005             und;
006             .hh aber !DA!mals-
007             (.) äh sind auch die: wEssis ge!KOM!men,
008             und ham sie da herzlichkeit gespürt;
009             (.) oder !FREU!de- .hh
010             (0.4) oder (.) war_s NU:R (.) zurückhaltung.
011             (4.0)
```

An den gezeigten Ausschnitten wird der interaktive Zusammenhang gut deutlich: Die Opposition ‚Ossi – Wessi' wird durch die Interviewerin etabliert (Z. 007), vorher war nur von einer Kluft zwischen Ost und West die Rede (Z. 002). Auf diese Etablierung sozialer Kategorien geht die Interviewte in ihrer Antwort ein (Z. 019-022), sie legt den Fokus nicht selbst. Da in diesem Artikel die Transkription als solche im Vordergrund steht, soll nicht weiter auf die Analyse eingegangen werden. Es lässt sich aber gut erkennen, wie eng die Wahl der Transkriptionsnotationen (‚bereinigte Abschrift' vs. GAT 2-Transkript) mit der Analyse zusammenhängen.

Man muss außerhalb der Gesprächsforschung nicht alle Feinheiten eines gesprächsanalytischen Transkripts abbilden, ich plädiere aber aus den dargestellten Gründen dafür, mindestens Überlappungen, Pausen und Betonungen für jede Art empirischer Fragestellung mitzutranskribieren, um dem Charakter der Interaktion gerecht zu werden und Äußerungen angemessen analysieren zu können. Eine möglichst detailgetreue Transkription bietet also einige Vorteile für die empirische Sozialforschung.

3. Transkribieren: Eine Anleitung

In Kapitel 2 wurde der Fokus auf Zweck und Nutzen eines Transkripts gelegt. Dabei wurden besonders die Vorteile einer möglichst genauen Transkription vorgestellt, wie sie die Konventionen von GAT 2 ermöglichen. Im Folgenden steht nun die Anfertigung eines Transkripts im Fokus. Dabei orientiere im mich wieder an den Konventionen von GAT 2, die Anleitung lässt sich aber mühelos auf jede Art von Transkription übertragen.

3.1 Transkriptionskonventionen: Die Auswahl eines geeigneten Transkriptionssystems

Die Eigenschaften gesprochener Sprache unterscheiden sich in einigen Punkten von denen geschriebener Sprache (Hausendorf 2017), daher ist eine Übertragung nicht unmittelbar möglich. Stattdessen kann die Transkription als Schritt der *Wirklichkeitsreduzierung* und als *Interpretation* sozialer Wirklichkeit gesehen werden (Brinker und Sager 2010). Aus diesem Grund sollte man sich im Voraus des Transkribierens geeignete Konventionen auswählen, die man entweder selbst festlegt oder die man an bestehenden Konventionen ausrichtet. Die Verwendung eines bereits ausgearbeiteten ‚Regelwerks' bietet dabei den Vorteil, dass man sofort mit der Transkription anfangen kann und die meisten auftretenden Phänomene bereits markiert werden können (Pausen, Äußerungsüberlappungen, Betonungen etc.). Sollten einmal aber nicht alle Phänomene abgedeckt sein, können die Konventionen bei Bedarf ergänzt, abgeändert oder erweitert werden.

Die Auswahl der Konventionen richtet sich primär nach der Forschungsfrage: Welche Informationen muss die Transkription enthalten, um die Daten für die Analyse angemessen aufbereiten zu können? Stehen inhaltliche Aspekte im Vordergrund, muss das Transkript sicher weniger ‚Metainformationen' enthalten als bei einer Analyse der sprachlichen Formen und sequenziellen Strukturen. Wie in Kapitel 2 erläutert, sollten aber Pausen, Überlappungen und Betonungen stets berücksichtigt werden, da sie auch für die rein inhaltliche Betrachtung von Nutzen sein können.

Aus meiner Sicht bietet GAT 2 hier einen guten Zugriff. Die Konventionen schlagen ein Zwiebelprinzip vor, d.h. es gibt drei aufeinander aufbauende Formen eines Transkripts, die sich im Grad der Feinheit voneinander unterscheiden[4]:

- das *Minimaltranskript* (sequenzielle Struktur, Atmen, sofern es eine kommunikative Bedeutung trägt, Pausen, Verschleifungen, Lachen und Weinen, Re-

[4] Eine Übersicht über die einzelnen Notationsregeln findet sich online frei verfügbar im Anhang von Selting et al. 2009, 391-393: http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf. Es empfiehlt sich, diese Liste ausgedruckt neben den Computer zu legen, um während des Transkribierens permanenten Zugriff darauf zu haben.

zeptionssignale, Metainformationen wie para- und außersprachliche Handlungen und Ereignisse, Doppelklammer für die Kodierung von Auslassungen im Transkript und unverständliche Passagen),

- das *Basistranskript* (Konventionen des Minimaltranskripts, Markierung schneller Äußerungsanschlüsse und Dehnungen sowie von Abbrüchen, Fokusakzentuierung, Tonhöhenbewegung am Ende von Intonationseinheiten, interpretierende Kommentare mit Reichweite)
- und das *Feintranskript* (Konventionen von Minimal- und Basistranskript, Fokus- sowie Nebenakzentuierung, Ton- und Akzenttonhöhenbewegungen, Lautstärke und Sprechgeschwindigkeit, Stimmqualität und Artikulationsweise).

Je nach Bedarf können die Konventionen gewählt werden. Es empfiehlt sich dabei stets, mit dem Minimaltranskript anzufangen und die Details je nach Intensität der Analyse zu verfeinern. Das ist mit den Notationstechniken von GAT 2 jederzeit möglich. In Fachdisziplinen der empirischen Sozialforschung außerhalb der Gesprächsforschung ist es sicher gewöhnungsbedürftig, alles kleinzuschreiben. Hier können die Konventionen abgeändert werden, indem man die Standardschreibung beachtet und Betonungen z.B. durch Unterstreichungen deutlich macht. Die Konventionen für Pausen u.ä. können aber trotzdem von GAT 2 übernommen werden.

GAT 2 ist ein Notationssystem, das sich problemlos in üblichen Textverarbeitungsprogramme wie z.B. Word umsetzen lässt. Dadurch ist sowohl die Anfertigung als auch die Einbettung sequenzieller Ausschnitte aus dem Datenkorpus in wissenschaftliche Publikationen ohne Weiteres möglich[5]. Dennoch empfiehlt es sich aus meiner Sicht, mit speziellen Transkriptionssoftwareprogrammen wie z.B. Transana oder f4[6] zu arbeiten. Der Hauptgrund ist die verfeinerte Playerfunktion: Diese Programme ermöglichen ein zielgenaues Zurück- und Vorspulen von wenigen Sekunden bis hin zu größeren Zeitabständen sowie die Möglichkeit der verlangsamten Wiedergabe, was besonders bei akustischen Problemen wichtig ist.

Im Folgenden sollen nun die Schritte einer Transkription erläutert werden[7].

---

[5] Andere Transkriptionsnotationen sind als Partitur – ähnlich der Notation von Musikstücken – zu verfassen, wie z.B. das Halbinterpretative Arbeitstranskript (HIAT, http://exmaralda.org/de/hiat/), das hauptsächlich in der funktionalen Pragmatik verwendet wird. Für die Anfertigung werden in der Regel Softwareprogramme wie EXMARaLDA (http://exmaralda.org/de/) oder ELAN (https://tla.mpi.nl/tools/tla-tools/elan/) gebraucht.

[6] Informationen siehe https://www.transana.com/ und https://www.audiotranskription.de/f4.

[7] An dieser Stelle sei auch auf das Online-Tutorial GAT-TO der Universität Freiburg verwiesen, dass die einzelnen Transkriptionsschritte ausführlich und unter Zuhilfenahme von Audiodateien erläutert (http://paul.igl.uni-freiburg.de/gat-to/ (letzter Zugriff: 01.08.2018)).

3.2 Die Tätigkeit des Transkribierens: Eine Anleitung am Beispiel von GAT 2

Das Transkribieren ist gerade anfangs eine langwierige Aufgabe, daher sollte keine unnütze Energie darauf verschwendet werden, die gesamten Daten von Anfang an im Detail zu transkribieren. Stattdessen empfiehlt es sich zunächst, einen Überblick über die Daten mit Hilfe von Gesprächsinventaren zu verschaffen (Deppermann 2008, 32ff.) und so für die Analyse interessante Sequenzen auswählen zu können.

Vor dem Start der Transkription sollten die Notationsregeln klar und in Form einer Liste griffbereit sein[8]. Ich empfehle die Verwendung von Kopfhörern, da man damit wesentlich mehr wahrnehmen kann. Für die beteiligten Interaktanten sollten zudem Pseudonyme ausgewählt werden, deren Silben denen des Originalnamens entsprechen, damit die Intonationsnotation in Fällen von Namensnennung nicht beeinträchtigt wird (z.B. von ‚Dennis' zu ‚Daniel'). Liegen mehrere Aufzeichnungen derselben Personen vor, empfiehlt sich eine Pseudonymliste, um den gleichen Personen im Datenkorpus stets die gleichen Pseudonyme zuzuordnen.

Bei der Transkription selbst sollte man sich zunächst auf die reinen Äußerungen beschränken und versuchen, den Wortlaut ohne Akzentuierungen, Pausen etc. zu transkribieren. Die Kleinschreibung von GAT 2 entlastet dabei, die Konzentration liegt allein auf der genauen Abfolge der Worte. Das mehrmalige Abhören von Äußerungen ist in den meisten Fällen nötig, um alle Details zu erfassen. Es empfiehlt sich, Äußerung für Äußerung vorzugehen. In der Regel wird die orthographisch korrekte Schreibweise verwendet, außer (regionale) Abweichungen sind deutlich hörbar.

Wenn der Wortlaut transkribiert wurde, sollten Pausen, Überlappungen, Akzentuierungen etc. vorgenommen werden, sofern mehr als ein Minimaltranskript angestrebt wird. Die Verfeinerung der Transkription kann aber auch zu späteren Zeitpunkten nachgeholt werden.

Die segmentale Gliederung des Transkripts sollte unmittelbar im Prozess beachtet werden, um eine gewisse Ordnung zu wahren. Ein Transkript gliedert sich nach Sprecherbeiträgen. In GAT 2 wird die Segmentierung der Sprecherbeiträge nach Intonationsphrasen vorgenommen, die jeweils eine eigene Zeile erhalten (siehe Beispiel in Kapitel 2). Intonationsphrasen sind Einheiten, die einen hörbaren Intonationsverlauf aufzeigen:

> Die Intonationsphrase wird durch einen als kohäsiv wahrgenommenen Tonhöhenverlauf als eine zusammenhängende Einheit gestaltet. Dabei kommt dem intonationsphrasenfinalen Tonhöhenverlauf, d.h. der letzten Tonhöhenbewegung vor dem Ende der Intonationsphrase, eine besondere Bedeutung sowohl für die Wahrnehmung der Intonationsphrasengrenze als auch für die interaktive Funktion der jeweiligen Intonationsphrase zu. (Selting et al. 2009, 370)

[8] Für GAT 2 siehe Link in Fußnote 4.

In der Regel ist pro Intonationsphrase eine Fokusakzentuierung enthalten. Darüber hinaus kann man sich an syntaktischen Grenzen orientieren, wobei diese nicht immer eine Intonationsphrase kennzeichnen müssen. Nebensätze werden in der Regel aber als eigenständige Phrasen aufgefasst, d.h. sie erhalten eine eigene Zeile. Auch Pausen oder visuell wahrnehmbare Tätigkeiten werden in eine eigene Zeile geschrieben. Die Notation dieser Merkmale braucht oft etwas Übung, um die Segmentierung sicher vornehmen zu können. Wenn der Fokus allein auf inhaltlichen Aspekten liegt, kann über die Segmentierung des Transkripts in Intonationsphrasen aber auch hinweggesehen werden. So arbeitet man sich Stück für Stück vorwärts, bis die gesamte zu transkribierende Datenmenge verschriftlicht ist.

Nonverbale Handlungen (Gestik, Mimik oder nicht-sprachliche Äußerungen wie Husten oder Räuspern) sollten im Transkript ebenfalls notiert werden, allerdings nur, sofern sie die Handlung beeinflussen. Der Fokus der Forschungsfrage ist hier ebenfalls entscheidend: Wer sich für Gesten und ihre kontextuelle Einbettung interessiert, sollte diese im Transkript sehr genau abbilden, während für andere Fragestellungen nur sehr auffällige Gesten transkribiert werden sollten. Nach GAT 2 kann man diese Informationen in eine eigene Zeile in Doppelklammern als Metakommentar notieren, dort gilt dann auch die übliche Schreibweise: ((räuspert sich)). Findet die nonverbale Handlung parallel zu einem Sprecherbeitrag statt, kann man sie entweder in der gleichen Zeile durch spitze Klammern begleitend notieren oder durch senkrechten Strich in eine neue Zeile schreiben:

Beispiel 5:

```
005   Leh   <<auf die Tafel zeigend> das ist RICHtig.>
```

Beispiel 6:

```
005   Leh   |das ist RICHtig.       |
            |((zeigt auf die Tafel))|
```

Letzteres bietet sich vor allem dann an, wenn häufig nonverbale Handlungen transkribiert werden. Dadurch wird das Transkript übersichtlicher.

Zur angemessenen Umsetzung der Notationen ist es in GAT 2 von Anfang an wichtig, auf eine entsprechende Formatierung zu achten. Um das Transkript vor Formatierungsschwächen zu schützen, wird auf die Nutzung von Tabstopps verzichtet, stattdessen wird mit Leerzeichen gearbeitet. Die Schriftart Courier New, wie sie in den bisherigen Transkriptionsbeispielen zu sehen ist, gewährt dabei ein einheitliches Erscheinungsbild, da jedes Zeichen den gleichen Platz beansprucht (‚nichtproportionale Schrift'). Die Zeilennummern sind dabei Bestandteil der Transkription. Nach der Zeilennummer folgen drei Leerzeichen, dann kommt die Sprechersigle (nur einmal bis zum Sprecherwechsel), danach folgen wieder drei Leerzeichen, bis die Äußerungsnotationen beginnen, die nach Intonationsphrasen gegliedert sind.

Die genannten Prinzipien sind in einem Ausschnitt aus obigem Transkriptionsbeispiel gut zu erkennen (die Punkte symbolisieren die Leerzeichen und sind im Transkript nicht also solche zu erkennen):

Beispiel 7:

```
003...Leh...kannst DU uns jetzt vielleicht auch äh_jetz sAgen?
004.........oder die ANderen kInder?
005.........vielleicht können die uns AUCH sagen;=
006.........=vielleicht kann_es_uns der tOm (.) AUCH sagen?
007.........so_n tIEr (du)/das: KANN man ja nicht nur einfach so in
            der wOhnung halten,=
```

Sollten einmal Teile von Äußerungen oder auch ganze Äußerungen nicht verstanden worden sein, so kann das folgendermaßen verschriftet werden: ((unverständlich, ca. 8 Sekunden)). Größere Auslassungen sind durch ((…)) zu kennzeichnen, mitunter kann in die Doppelklammern auch die Zeit der Auslassung geschrieben werden oder der Grund: ((Auslassung, ca. 3 Minuten)) oder ((Unruhe in der Klasse, ca. 2 Minuten)).

Am Ende sollte das Transkript mit einem Transkriptionskopf ausgestattet werden, um alle möglichen Informationen gebündelt zugänglich zu machen:

Transkriptname:	Unterricht_Grundschule_1
Inhalt der Aufnahme:	Frontalunterricht
Aufnahmeort und -tag:	Bielefeld, 18.04.2018
Gesamtdauer der Aufnahme:	43,08 Minuten
Zeitindex / Dauer der Ausschnitte:	10,30-15,40 / 5,10; 16,00-16,48 / 0,48
Transkribierende/r:	Ann-Christin Buttlar
Namen der beteiligten Personen:	Lehrerin (Leh), Mark (Mar), Lisa (Lis)
Kommentare/besondere Informationen:	*Klasse von Aufnahmegerät abgelenkt.*

Man sollte sich von dem anfänglichen Aufwand bei der Transkription nicht entmutigen lassen. Im Schnitt benötigen Anfänger ca. eine Stunde für eine Minute, abhängig von der Komplexität der Daten (Monolog, Interview, Gruppendiskussion). Je mehr man aber transkribiert, desto schneller geht der Prozess voran. In diesem Fall gilt in der Tat: Übung macht den Meister. Die Mühe bei der Transkription zahlt sich für die Analyse maximal aus: Je besser die Transkription, desto besser ist der Zugriff auf die Daten und damit die Analyse selbst.

3.3 Einige Praxistipps zum Schluss

Im vorliegenden Artikel wurde der Prozess des Transkribierens detailliert vorgestellt und viele Hinweise zur Durchführung gegeben. Am Ende dieses Artikels sollen nun noch einmal einige Praxistipps zusammengefasst werden, die für Anfänger/innen nützlich sein können.

- Transkribieren gesprochener Daten ist ein aufwändiger Prozess. Aus diesem Grund sollten die Datensätze gut ausgewählt werden, die transkribiert werden sollen. Je nach Forschungsfragen und Datenmaterial bietet es sich oft an, nicht alles zu transkribieren, sondern lediglich die Sequenzen, die für die Analyse interessant erscheinen. Zur Identifikation solcher Stellen können z.B. sogenannte *Gesprächsinventare* erstellt werden, wie Deppermann (2008) sie beschreibt.

- Des Weiteren sollte man zunächst mit den Konventionen eines Minimal- oder Basistranskripts beginnen (siehe Kapitel 3.1). Sequenzen, die detailliert analysiert werden sollen, lassen sich so finden und werden in einem zweiten Schritt fein(er)transkribiert. Der Grad der Feintranskription ist dabei von den jeweiligen Forschungsfragen abhängig.
- Gerade zu Anfang erscheinen die Konventionen oft noch fremd. Um die Einheitlichkeit des Transkripts zu gewährleisten und so Fehlinterpretationen vorzubeugen, lohnt es sich, die Übersicht der Konventionen zunächst gut zu lesen und dann ausgedruckt neben den Computer zu legen (im Falle von GAT 2 siehe Selting et al. 2009, 391-393).
- Beim Transkribieren ist es wichtig, möglichst genau zu hören, was wann und wie gesagt wird. Daher empfiehlt es sich, die Audio- oder Videodaten mit (hochwertigen) Kopfhörern anzuhören. Außerdem wird empfohlen, die Transkription in einer möglichst ruhigen Umgebung anzufertigen.
- Interviewdaten oder Dialoge sind einfacher zu transkribieren als Gruppengespräche oder Unterrichtsstunden: Je mehr Überlappungen auftreten, desto schwieriger wird die Transkription. Überlappungen erfordern eine große Aufmerksamkeit, um alle Sprecherinnen- und Sprecherbeiträge wahrzunehmen. Hier sollte man schrittweise vorgehen und sich zunächst auf eine/n Sprecher/in konzentrieren und die reinen Äußerungen fokussieren. Danach konzentriert man sich auf die nächste Person usw., bis (möglichst) alle Äußerungen erfasst sind. Anschließend kann man markieren, welche Äußerungen sich wann überlappen.

Zum Schluss sei noch auf diverse Softwareprogramme hingewiesen, die die Transkription wesentlich erleichtern und darüber hinaus auch diese für die Analyse besser zugänglich machen (siehe den Beitrag von König zum Softwareeinsatz in diesem Band). Gerade die Funktion von Timecodes, die das Transkript mit der Video- oder Audioaufnahme verknüpfen, ist ein gutes Hilfsmittel. Denn obwohl man Gesprächsdaten sehr fein transkribieren kann, bleibt das Transkript letztlich eine Reduzierung der Wirklichkeit und ersetzt nicht den Blick in die Gesprächsaufnahmen. Dennoch hilft die Transkription bei der Datenaufbereitung und ist insbesondere für die transparente Darstellung der Analyse im wissenschaftlichen Kontext ein unverzichtbares Mittel.

Literatur

Bergmann, Jörg (1994): Ethnomethodologische Konversationsanalyse. In: Fritz, Gerd/Hundsnurscher, Franz (Hrsg.): Handbuch der Dialoganalyse. Tübingen: M. Niemeyer, 3-16.

Brinker, Klaus/Sager, Sven F. (2010): Linguistische Gesprächsanalyse. Eine Einführung. 5., neu bearb. Aufl. Berlin: Erich Schmidt.

Buttlar, Ann-Christin (2017): Erwartungen von Lehrpersonen und Äußerungen von Schülerinnen und Schülern. Sequentielle Analysen zur Ko-Konstruktion von Angemessenheit im Unterrichtsdiskurs. Unveröffentlichte Dissertation. Technische Universität Dortmund.

Deppermann, Arnulf (2008): Gespräche analysieren. Eine Einführung. 4. Aufl. Wiesbaden: VS.

Deppermann, Arnulf (2013): Interview als Text vs. Interview als Interaktion. In: Forum Qualitative Sozialforschung, 3, 14, 1-40. http://www.qualitative-research.net/index.php/fqs/article/view/2064 (letzter Zugriff: 01.08.2018).

Dürscheid, Christa (2006): Einführung in die Schriftlinguistik. (= Studienbücher zur Linguistik, 8) 3., überarb. Aufl. Göttingen: Vandenhoeck & Ruprecht.

Hausendorf, Heiko (Hrsg.) (2007): Gespräch als Prozess. Linguistische Aspekte der Zeitlichkeit verbaler Interaktion. (= Studien zur deutschen Sprache, Bd. 37) Tübingen: G. Narr.

Hausendorf, Heiko (2017): Warum wir transkribieren. Anmerkungen aus der Welt der linguistischen Gesprächsanalyse. In: Text: Kritische Beiträge, 15, 217-230.

Heller, Vivien/Morek, Miriam (2016): Gesprächsanalyse. Mikroanalytische Beschreibung sprachlicher Interaktion in Bildungs- und Lernzusammenhängen. In: Boelmann, Jan M. (Hrsg.): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. 2., durchges. Aufl. Baltmannsweiler: Schneider Hohengehren, 223-246.

Selting, Margret/Auer, Peter/Barth-Weingarten, Dagmar/Bergmann, Jörg/Bergmann, Pia/Birkner, Karin et al. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion, 10, 353-402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf (letzter Zugriff: 01.08.2018).

LISA KÖNIG

Softwareeinsatz

Digitale Hilfen im Umgang mit empirischen Daten durch SPSS, MAXQDA und Co.

Die konkrete Verarbeitung und Aufbereitung erhobener Datensätze gestaltet sich zumeist als komplexe Herausforderung des Auswertungsprozesses, auch wenn zu Beginn der empirischen Erhebung das Forschungsdesign explizit festgelegt und eine Wahl der im Anschluss zu verwendeten Auswertungsmethoden getroffen wurde. Um diese Aufgabe zu bewältigen, wurden zahlreiche auf verschiedene Forschungstypen abgestimmte, computergestützte Verfahren entwickelt, welche den Verarbeitungs- und Auswertungsprozess empirischer Forschungsvorhaben erleichtern. Unabhängig von der Software selbst ersetzt deren Verwendung jedoch nicht die selbstständige Auseinandersetzung mit forschungstheoretischen, analytischen oder statistischen Verfahren, sodass die Anwendung digitaler Hilfen lediglich die Nutzung der eigenen Fähigkeiten des Forschenden unterstützt. Die für das Forschungsprojekt richtige Wahl der Programme erweist sich aufgrund der Fülle des Softwareangebots jedoch als gar nicht so leicht: Werden beispielsweise statistische Verfahren für die Auswertung benötigt? Müssen Codings vorgenommen werden, um die Daten zu strukturieren? Wie können die Forschungsergebnisse möglichst anschaulich dargestellt werden, sodass Außenstehende die Erkenntnisse auch erfassen können?

Die vorliegenden Beschreibungen sollen daher als Orientierungshilfe die bekanntesten Softwarangebote für verschiedene Verarbeitungsschwerpunkte vorstellen. Hierbei wurden einzelne, in Forschungskontexten weiter verbreitete Programme gewählt, welche exemplarisch für deren jeweiliges Einsatzgebiet bezüglich der Anwendungsmöglichkeiten, Potenziale sowie Grenzen erläutert werden und Literaturempfehlungen für eine weitere Auseinandersetzung mit der jeweiligen Software erfolgen. Neben denen hier ausgewählten Programmen lohnt sich die Sichtung alternativer Softwareangebote, welche ggf. für das spezifische Forschungsprojekt geeigneter erscheinen:

- Verarbeitung der Daten mittels statistischer Analyse — SPSS
- Verarbeitung der Daten mittels qualitativer Analyse — MAXQDA
- Aufbereitung der Daten mittels Transkription — f4/f5
- Aufbereitung der Daten mittels Visualisierung — Origin
- Verwahrung der Daten mittels Verschlüsselung — VeraCrypt

IBM SPSS Statistics
(IBM)

Kurzbeschreibung
Software zur statistischen Weiterverarbeitung und Auswertung empirischer Datensätze, um statistisch relevante Zusammenhänge aufzuzeigen sowie erste oder auch weiterführende Analysen vorzunehmen.

Technische Konfigurationen

Versionen	aktuelle Version: IBM SPSS Statistics 24 (15.03 2016) Diverse Testversionen (Zeitraum: 30 Tage) der früheren Programmversionen finden sich unter anderem auf der Homepage des Anbieters.
Kompatibilität	Microsoft Windows, Mac OS X, GNU/Linux

Anwendungsmöglichkeiten
IBM SPSS Statistic bietet die Möglichkeit, bereits kodierte Datensätze einzugeben, diese mittels statistischer Verfahren zu untersuchen sowie anschließend die aufbereiteten Ergebnisse in Tabellen oder Diagrammen darzustellen. Die einzusetzenden Analyseverfahren sind hierbei vielfältig angelegt, sodass neben deskriptiven Methoden der Datenaufbereitung, wie Häufigkeitsverteilungen oder Mittelwertanalysen, weiterführende Untersuchungen anhand der Berechnung neuer Variablen, Zusammenhangsanalysen, Voruntersuchungen zu bestimmenden Faktoren aber auch Modellanalysen durchgeführt werden können.

Potenziale und Grenzen

• umfassende Darstellungs-, Analyse- sowie Weiterverarbeitungs-optionen der erhobenen Daten mithilfe der gängigsten statistischen Verfahren • Analysetools für jede Phase des Forschungsprozesses (von der Eingabe bis zur Präsentation) • sowohl für kleine als auch umfangreiche Datensätze geeignet	• wenig intuitiver Aufbau des Programms erfordert Einarbeitung • Software setzt Vorwissen bezüglich der Anwendung und Funktion statistischer Analyseverfahren voraus

Literaturempfehlungen

- Braunecker, Klaus (2016): How to do Empirie, how to do SPSS: eine Gebrauchsanleitung. Heidelberg: Springer.
- Budischewski, Kai/Kries, Katharina (2015): SPSS für Einsteiger: Einführung in die Statistiksoftware für die Psychologie. Weinheim/Basel: Beltz.
- Wittenberger, Reinhard/Cramer, Hans/Vicari, Basha (2014): Datenanalyse mit IBM SPSS Statistics: Eine syntaxorientierte Einführung. Stuttgart: UTB.

Alternative Programme
PSPP (GNU), R (R Core Team), Stata (StataCorp), STATISTICA (StatSoft)

MAXQDA
(VERBI)

Kurzbeschreibung
Software für die qualitative Daten- und Textanalyse von quantitativen, qualitativen oder auch in Mixed-Methods-Verfahren erhobenen Datensätzen.

Technische Konfigurationen

Versionen	aktuelle Version: MAXQDA 2018 (01.12 2017) Eine Testversion des aktuellen Programms (Zeitraum: 14 Tage) findet sich auf der Website des Herstellers. Lizenzunabhängig ist es möglich, mithilfe des kostenlosen *MAXQDA-Readers* die in dem Programm erstellten Datensätze einzusehen und zu präsentieren.
Kompatibilität	Microsoft Windows, Mac OS X

Anwendungsmöglichkeiten
MAXQDA ermöglicht die qualitative Analyse diverser Datensätze sowie von Textdokumenten, welche durch simple Übertragungsvorgänge in das Programm integriert oder bereits in die Anwendung eingetragen werden können. Mithilfe von Codes können die Daten kategorisiert, systematisiert und anschließend hinsichtlich qualitativer Zusammenhänge untersucht werden. Anhand des umfassenden Ordnungssystems können Fokusgruppen gebildet oder eine Typologie generiert werden. Weiterhin besteht die Möglichkeit die qualitativen Ergebnisse deskriptiv darzustellen und in bildgebenden Verfahren abzubilden oder auch bereits die Transkriptionen innerhalb des Programms vorzunehmen.

Potenziale und Grenzen

• Möglichkeit, verschiedenste Datensätze (Textdatei, Excel-Tabelle, Bilder oder auch SPSS-Auswertungen) zu integrieren • Programminhärente Transkription von Audio- oder Videodateien und automatische Erkennung unterschiedlicher Probanden in Fokusgruppendiskussionen oder Gruppengesprächen	• wenig intuitiver Aufbau des Programms erfordert Einarbeitung • Software setzt Vorwissen bezüglich der Anwendung und Funktion qualitativer Auswertungsmethoden voraus

Literaturempfehlungen
- Wagner, Britta (2005): Arbeiten mit MAXQDA: Kurze Einführung in die computergestützte Analyse qualitativer Daten. Bamberg: UNIBA.
- Woolf, Nicholas (2017): Qualitative analysis using MAXQDA: the five-level QDA method. London: Routledge.

Alternative Programme
ATLAS.ti (altlas.ti), Dedoose (dedoose), NVivo (QSR), Transana (Spurgeon)

f4/f5 transcript
(Audiotranskription)

Kurzbeschreibung
Software für die Erstellung von Teil- oder Volltranskriptionen, welche entweder orthographisch korrigiert und phonologisch orientiert verfasst werden können.

Technische Konfigurationen

Versionen	aktuelle Version: f4/f5 2017 (01.11 2017) Eine Testversion des aktuellen Programms (Zeitraum: 10 Tage) findet sich auf der Website des Herstellers. Außerdem ist es möglich, Kurzzeitlizenzen zu erwerben und diese individuell auf die Dauer des Forschungsvorhabens anzupassen.
Kompatibilität	f4: Microsoft Windows, GNU/Linux; f5: Mac OS X

Anwendungsmöglichkeiten
f4/f5 transcript ist eines der gängigsten Programme für die Transkription von Audiodateien, welche während mündlichen Befragungen oder Diskussionen erhoben wurden. Innerhalb des Programms ist es möglich, die Audiodateien zu importieren, anschließend direkt zu verschriftlichen und die abgeschlossenen Transkripte in Textverarbeitungsprogramme zu exportieren. Hierbei stehen dem Forschenden verschiedene Tools zur Auswahl, mit welchen die Abspielgeschwindigkeit angepasst, die verschiedenen Sprecher markiert, Zeitmarken gesetzt, Kommentare eingefügt sowie die Tonspuren durch die Abbildung der Wellenform oder auch der Zeitmessung visualisiert werden können. Auch können häufig verwendete Formulierungen hinterlegt und mithilfe des Programms an den entsprechenden Stellen zeitsparend eingesetzt werden.

Potenziale und Grenzen

• Anpassung der Abspielgeschwindigkeit ohne Stimmverzerrung oder Stimmlagenänderung • Automatische Erkennung unterschiedlicher Stimmen, Sprecherwechseln und Pausenlängen • Verwendung eines Fußpedals zur Notation von Pausen möglich	• Einige Dateiformate lassen sich nicht ohne weitere Probleme abspielen, sodass WMA-Formate umgewandelt und für die Verwendung von MP3-Dateien eine feste Bitrate von 192 kbit/s vorliegen muss.

Literaturempfehlungen
- Dittmar, Norbert (2009): Transkription: ein Leitfaden mit Aufgaben für Studenten, Forscher und Laien. Wiesbaden: VS Sozialwissenschaft.
- Fuß, Susanne/Karbach, Ute (2014): Grundlagen der Transkription: eine praktische Einführung. Opladen: Budrich.

Es finden sich zudem zahlreiche Übungsbeispiele auf der Homepage des Anbieters.

Alternative Programme
easytranscript (GNU), ELAN (TLA), Express Scribe (NCH), Praat (praat)

Origin

(OoriginLab)

Kurzbeschreibung

Software zur komplexen statistischen Analyse und professionellen Visualisierung von bereits ausgewerteten Datensätzen in 2D- und 3D-Modellen.

Technische Konfigurationen

Versionen	aktuelle Version: Origin 2017 (09.11 2016) Eine Testversion des aktuellen Programms (Zeitraum: 21 Tage) findet sich auf der Website des Herstellers, welche je nach Einsatzgebiet im studentischen oder akademischen Kontext angepasst werden kann.
Kompatibilität	Microsoft Windows

Anwendungsmöglichkeiten

Origin ermöglicht die professionelle Visualisierung relevanter Datensätze in 2D- oder 3D-Modellen, welche anhand einer einfach anzuwendenden Exportierungsfunktion in die gängigsten Bildformate (u.a. EPS, JPEG, GIF, TIFF, PDF, WMF) übertragen werden können. Außerdem bietet die Software eine Auswahl an Analysetools, welche insbesondere auf mathematische Unterschiede innerhalb differenter Erhebungsgruppen oder Erhebungszeiträumen ausgelegt sind und anschließend in die generierten Modelle übertragen werden können. Auch besteht die Möglichkeit, komplexe statistische Berechnungen (Regressionen, nicht-lineare Kurvenanpassung etc.) vorzunehmen. Neben der Standardversion *Origin* liegt die Erweiterung des Herstellers *OriginPro* vor, in welcher der Korpus der Analyse- und Darstellungsmöglichkeit nochmals erweitert wurde.

Potenziale und Grenzen

• Flexibler Umgang: Möglichkeit, sowohl die Analyse als auch die Visualisierung innerhalb eines Programms vorzunehmen oder ausschließlich die Darstellungsfunktionen von bereits mithilfe anderer Software ausgewerteten Datensätzen zu nutzen	• ausschließlich unter Microsoft Windows nutzbar

Literaturempfehlungen

Neben Werken, welche die verschiedenen Darstellungsoptionen von Datensätzen erläutern, kann für eine erste Auseinandersetzung mit der konkreten Software folgendes Einsteigerwerk empfohlen werden: Beneke, Thomas W./Schwippert Wolfgang W. (1997): Datenanalyse und Präsentation mit ORIGIN: Anwendungsbeispiele und Lösungsvorschläge aus der Praxis. Bonn: Addison-Wesley-Longman.

Alternative Programme

Excel (Microsoft), Gephi (gephi)

VeraCrypt
(IDRIX)

Kurzbeschreibung
Software zur vollständigen oder auch partiellen Datenverschlüsselung von Festplatten, Datenträgern oder Cloudspeicherinhalten, auf welchen empirische Forschungsdaten verwahrt oder transportiert werden sollen.

Technische Konfigurationen

Versionen	aktuelle Version: VeraCrypt 1.21 (21.08 2017) Die Open Source Lizenz steht als freier Download auf der Website des Herstellers oder auch der programmeigenen Homepage zur Verfügung.
Kompatibilität	Microsoft Windows, Mac OS X, GNU/Linux

Anwendungsmöglichkeiten
VeraCrypt ist eines der am häufigsten verwendeten Datenverschlüsselungssysteme und stellt als freie Lizenzsoftware eine Alternative zu dem zuletzt 2014 aktualisierten TrueCrypt oder Bitlocker von Microsoft dar. Mithilfe der Verschlüsselungssoftware können ganze Systeme, Partitionen oder empirische Forschungsdaten innerhalb eines gesicherten Containers gespeichert werden, welche ausschließlich durch die Dekodierungsfunktion des Programms und einem zusätzlichen Sicherheitspasswort eingesehen sowie bearbeitet werden können. Nach der Entschlüsselung werden die Container als virtuelle Laufwerke angezeigt, welche nach der Bearbeitung der Datensätze wieder verschlüsselt und vor Datenmissbrauch geschützt werden.
In der Zwischenzeit wird *VeraCrypt* auch im schulpraktischen Kontext zur Verwahrung von Schülerinnen- und Schülerdaten verwendet.

Potenziale und Grenzen

• leicht umzusetzendes Sicherungssystem für ein nachhaltiges und Datenschutzrichtlinien entsprechendes Forschungsdatenmanagement • laufende Sicherheitsupdates, um die bestmögliche Verschlüsselung der Daten zu gewährleisten	• Daten sind lediglich auf Geräten ent- und dekodierbar, welche die Software direkt installiert haben • kurze Einarbeitung hinsichtlich der versteckten Containerdarstellung nötig • Passwortsicherheit muss eingehalten werden

Literaturempfehlungen
Neben Werken zur IT-Sicherheit und des Datenschutzes liegt eine Begleitsoftware zu *VeraCrypt* vor, welche die Handhabung erläutert und praktische Schritt-für-Schritt Anleitungen bereithält: RS Distribution GmbH (2017): VeraCyrpt für Einsteiger.

Alternative Programme
Eperi Cloud Data Protection (eperi), ESET Endpoint Encryption (ESET), TruPax (Coder's Lagoon)

ANKE SCHMITZ

Statistische Grundlagen

Anforderungen der empirischen Datenauswertung sukzessive meistern

1. Bedeutung des Erwerbs von statistischen Grundlagen

Empirische Methoden gewinnen in der deutschdidaktischen Forschung immer mehr an Bedeutung (Boelmann 2016). Studierende und Promovierende stehen heutzutage weitaus häufiger vor der Herausforderung, ihre Forschungsfragen in wissenschaftlichen Arbeiten durch empirische Analysen zu beantworten, als es vor einigen Jahren der Fall war. Die statistische Auswertung von empirischen Daten wird dabei oft mit dem Bild einer unvorstellbar großen, mit Ängsten behafteten Hürde gleichgesetzt. Eine unüberschaubare Vielzahl an statistischen Ratgebern und Nachschlagewerken aus Nachbarwissenschaften – wie die Psychologie und die Sozialwissenschaften – und deren unzureichende Passung zu deutschdidaktischen Forschungsthemen und eigenen Vorwissensbeständen verstärkt das Unwohlsein gegenüber Statistik und verringert die Motivation, sich mit statischen Auswertungsverfahren auseinanderzusetzen. Doch auch wenn das Thema Statistik keinen oder nur einen geringen Bestandteil im Studium eingenommen hat, ist eine Aneignung von statistischen Grundlagen durchaus zu bewältigen. Sie sind aufgrund ihres hohen Erkenntniswerts vor allem lohnenswert, da eigens erhobene empirische Daten durch statistische Analysen erst transparent werden und formulierte Fragestellungen systematisch beantwortet werden können. Zudem steuern statistische Grundlagenkenntnisse forschungsrelevante, methodische Vorüberlegungen, wie die zielgerichtete Konzeption von Erhebungsdesigns und die auf Auswertungsverfahren ausgerichtete Konstruktion von Instrumenten (vgl. Betz/Kröger-Bidlo/Schmitz/Schuttkowski 2018).

Neben einer überblicksartigen Einführung in grundlegende statistische Begrifflichkeiten und Analyseverfahren der quantitativen Forschung sowie einer kurzen Darstellung von ersten, sukzessiven Analyseschritten, soll dieser Beitrag einen Zugang zur quantitativen Datenanalyse schaffen und Ängste vor der Statistik abbauen. Diesbezüglich kann an dieser Stelle bereits erwähnt werden, dass eine Einarbeitung in statistische Grundlagen kein detailliertes Verstehen von mathematischen Operationen erfordert. Mittels einer Statistiksoftware wie *SPSS Statistics* lassen sich Berechnungen ohne tiefe statistische Vorkenntnisse durchführen.

Wichtig für die Auswahl der richtigen Berechnungen ist aber ein basales Verständnis von zentralen Begrifflichkeiten und das Nachvollziehen von Prinzipien und Nutzen ausgewählter statistischer Auswertungsverfahren. Das Verstehen von mathematischen Berechnungen wird erst dann bedeutsam, wenn komplexere Auswertungsverfahren für die Datenanalyse erforderlich werden, etwa in größeren quantitativen Forschungsprojekten.

2. Überblick über statistische Grundbegriffe und Analyseverfahren

Im Folgenden wird in einem ersten Schritt beleuchtet, was unter dem Begriff Statistik zu verstehen ist und was Statistik leisten kann. Anschließend wird stichpunktartig aufgelistet, welche statistischen Begriffe und Auswertungsverfahren für einen ersten Zugang zur empirischen Datenanalyse relevant sind und wobei es sich eher um fortgeschrittene Kenntnisse handelt. Ausgewählte Literaturempfehlungen für Einsteigerinnen und Einsteiger sowie Fortgeschrittene werden aufgeführt, um die hier in Auszügen präsentierten Grundlagen und in Kapitel 3 beschriebenen Auswertungsschritte näher betrachten und bei Bedarf vertiefen zu können.

2.1 Grundlegende Begrifflichkeiten und statistische Analysen

Als Statistik (lat. *status*, ‚Stand', ‚Stellung', ‚Zustand') bezeichnet man ein mathematisches Fachgebiet, in welchem Methoden zum systematischen Umgang mit quantitativen Informationen behandelt werden. Was Statistik kennzeichnet, lässt sich insbesondere durch ihre Funktion verdeutlichen (vgl. Müller-Benedict 2011, 17). Ziel ist es, empirisch gewonnene Informationen in Form von Variablen systematisch auszuwerten, zusammenzufassen, zu veranschaulichen und zu interpretieren (vgl. Reinders/Gniewosz 2015, 131). Die Statistik wird in zwei Bereiche unterteilt: (1) Die deskriptive, also beschreibende Statistik dient der Ordnung von Informationen und bezieht sich dabei nur auf einen ausgewählten Datensatz, und (2) die Inferenzstatistik, welche sich mit der Wahrscheinlichkeit der Verallgemeinerung von Interpretationen bzw. Schlussfolgerungen beschäftigt und somit über den einzelnen Datensatz hinausgeht (vgl. Müller-Benedict 2011, 22). Darüber hinaus werden univariate und multivariate statistische Verfahren unterschieden. Erstere behandeln die Analyse von einer singulären erhobenen empirischen Information (zum Beispiel die Verteilung einer einzelnen Variablen). Multivariate Verfahren hingegen beschäftigen sich mit mehreren Informationen (wie den Zusammenhang von mehreren Variablen) (ebd.). Die folgenden statistischen Begrifflichkeiten und Auswertungsverfahren werden zur besseren Übersicht in überwiegend grundlagenorientierte, deskriptive und fortgeschrittene, inferenzstatistische Bereiche unterteilt. Zur Veranschaulichung werden die Begriffe und Analyseverfahren mithilfe von Beispielen erläutert.

(1) Grundlagenkenntnisse in Bezug auf Begriffe und Analyseverfahren der deskriptiven Statistik (vgl. Betz et al. 2018; Müller-Benedict 2011; Reinders/Gniewosz 2015):

- *Grundgesamteinheit*: Menge an statistischen Einheiten, über die man mit einer Erhebung Aussagen machen möchte (z.B. alle Gymnasiasten in Köln).
- *Stichprobe*: Teilmenge einer Grundgesamteinheit, welche für eine Studie ausgewählt wird (z.B. eine ausgewählte Anzahl von Gymnasiasten in Köln).
- *Variable (abhängig und unabhängig)*: Statistisch messbares Merkmal, z.B. die Lesemotivation oder die verbale Grundfähigkeit eines Gymnasiasten. Abhängige Variablen sind erklärte (z.B. Leseleistung) und unabhängige Variablen erklärende, ursächliche Variablen (z.B. Vorwissen, Motivation, verbale Grundfähigkeit).
- *Skalenniveau*: Unterschieden werden in Bezug auf die Skalierung kategoriale Skalen wie Nominalskalen (z.B. das Geschlecht und Schulformen), Ordinalskalen (z.B. Schulnoten) sowie metrische Skalen wie Intervall- und Likertskalen (z.B. Testverfahren wie Intelligenztests oder Fragebögen mit Einteilungen von 1 = unwichtig bis 7 = sehr wichtig) und Verhältnisskalen (mit einem absoluten Nullpunkt wie das Alter).
- *Häufigkeiten*: Absolute Häufigkeiten repräsentieren eine Anzahl (z.B. 60 Mädchen und 50 Jungen), relative Häufigkeiten stellen ihren Anteil an der Gesamtstichprobe dar (z.B. 54,4%).
- *Lageparameter*: Maße bzw. Werte in einem Datensatz, die Auskunft über die zentrale Lage bzw. Tendenz eines Merkmals in einer Stichprobe geben, hierzu zählen der Modus (häufigster Wert in einer Stichprobe), Median (Wert, der eine Stichprobe mittig teilt) und Mittelwert (z.B. Durchschnittswert). Beträgt der Notendurchschnitt in einer fiktiven Klasse 2,7, handelt es sich hierbei um den Mittelwert. Wurde die Note „befriedigend“ am häufigsten von der Lehrkraft gegeben, ist dies der Modus. Der Median zeigt darüber hinaus, welche Note die gesamte Klasse in zwei Hälften teil.
- *Streuungsparameter*: Maße bzw. Werte in einem Datensatz, die anzeigen, wie sich die Daten um einen zentralen Wert verteilen (streuen). Spannweite (Abstand zwischen größtem und kleinstem Messwert), Varianz (Wert, der die quadrierte mittlere Abweichung aller abweichenden Werte vom Mittelwert angibt) und Standardabweichung (Wert, der die durchschnittliche Abweichung aller Messwerte vom Mittelwert anzeigt). Geht man von einem Lesekompetenztest aus, bei welchem der beste Lesekompetenzwert bei 36,0 Punkten liegt und der schlechteste Wert 12,0 Punkte beträgt, so beläuft sich die Spannweite auf 24,0 Punkte. Die Spannweite gibt also Auskunft über die Größe des Bereichs der Messwerte, jedoch bleibt offen, wie die einzelnen Messwerte variieren. Varianz und Standardabweichung berücksichtigen jeden einzelnen Messwert. Die Varianz ist ein wichtiges Maß für die Streuung

von Messwerten um den Mittelwert. Da die Varianz allerdings aufgrund einer Quadrierung aller einzelnen Werte nur schwer zu interpretieren ist, wird darauf basierend die Standardabweichung betrachtet. Die Standardabweichung besteht aus der Wurzel der Varianz und zeigt an, wie stark Messwerte um den Mittelwert streuen. Liegt der Mittelwert im Kompetenztest bei 18,0 Punkten und beträgt die Standardabweichung 3,0 Punkte, so weichen die Messwerte im Durchschnitt 3,0 Punkte vom Mittelwert ab.

- *Normalverteilung*: Häufigkeitsverteilung in Form einer Glockenkurve (z.B. der Intelligenzquotient mit vielen Ausprägungen im mittleren Bereich und Abflachung bei sehr hohen und sehr niedrigen Quotienten).
- *Korrelation*: Zusammenhang zwischen Variablen (z.B. zwischen Lesemotivation und verbaler Grundfähigkeit).

(2) Die folgenden Begriffe und Analyseverfahren lassen sich der fortgeschrittenen Inferenzstatistik zuordnen (Müller-Benedict 2011; Reinders/Gniewosz 2015; Schäfer 2011):

- *Signifikanzniveau (α)*: Höhe der Wahrscheinlichkeit, dass eine Hypothese irrtümlich angenommen wird (z.B. wäre eine Annahme bei einem Signifikanzniveau von $p < .05$ zu 95 Prozent richtig und nur 5 Prozent betrüge das Risiko einer Fehlannahme).
- *Effektstärke*: Stichprobenunabhängiges Maß für die statistische Größe eines Effekts (z.B. die Stärke eines Trainings oder eines Mittelwertsunterschiedes. Je größer die Streuung in einer Stichprobe nach einem Lesetraining ausfällt und je kleiner der Unterschied zwischen Experimental- und Kontrollgruppe, desto geringer ist der Effekt des Trainings).
- *t-Test*: Testverfahren, um Mittelwerte zwischen zwei Gruppen in Relation zu den Streuungen innerhalb der beiden Gruppen zu betrachten. Mit einem *t*-Test kann beispielsweise untersucht werden, ob sich die Differenz des Notendurchschnitts von Mädchen (1,7) und Jungen (2,0) in einer Klasse systematisch unterscheidet oder zufällig entstanden ist. Mittelwertsunterschiede werden z.B. durch die Effektgröße Cohens *d* ausgedrückt.
- *Varianzanalyse*: Vergleich von Mittelwerten zwischen mehr als zwei Gruppen in Relation zu den Streuungen innerhalb der Gruppen. Mit einer Varianzanalyse ließe sich beispielsweise ermitteln, ob sich die Mittelwerte in der Lesekompetenz von Schülerinnen und Schülern in Abhängigkeit der Schulform (Gruppen: Hauptschule, Realschule, Gymnasium, Gesamtschule) signifikant unterscheiden. Die Unterschiede werden durch die Effektgröße Eta-Quadrat ausgedrückt.
- *Regressionsanalyse*: Verfahren, um Werte einer abhängigen Variable (Kriterium) durch mehrere unabhängige Variablen (Prädiktoren) schätzen zu lassen. Das Kriterium Vorhersage von Leseleistungen kann beispielsweise durch die Prädiktoren Vorwissen, verbale Grundfähigkeit und Motivation vorhergesagt werden.

2.2 Literaturtipps zum Einstieg und zur Vertiefung

Neben dem vorliegenden Band, der auf empirische Forschung in der Deutschdidaktik fokussiert, gibt es weitere Einführungswerke, welche die Analyse von empirischen Daten nachvollziehbar machen. Folgende Literaturtipps dienen dem Einlesen in die Thematik und/oder dem Erwerb von Grundkenntnissen über statistische Verfahren und gewähren einen Einblick in die praktische Datenauswertung:

Betz/Kröger-Bidlo/Schmitz/Schuttkowski (2018): Empirische Methoden in der Sprachdidaktik. Ausgewählte Erhebungs- und Auswertungsverfahren.

Boelmann (Hrsg.) (2016): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung.

Rasch/Friese/Hofmann/Naumann (2014a): Quantitative Methoden 1. Einführung in die Statistik für Psychologen und Sozialwissenschaftler.

Field (2013): Discovering statistics using IBM SPSS Statistics.

Diese Werke dienen der weiterführenden Einarbeitung in komplexere Auswertungsverfahren, auch in Hinblick auf das Nachvollziehen von mathematischen Prinzipien.

Bortz/Döring (2006): Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler.

Bortz/Schuster (2010): Statistik für Human- und Sozialwissenschaftler.

Rasch/Friese/Hofmann/Naumann (2014b): Quantitative Methoden 2. Einführung in die Statistik für Psychologen und Sozialwissenschaftler.

Schäfer (2011): Statistik II. Inferenzstatistik.

Field (2013): Discovering statistics using IBM SPSS Statistics.

Die folgenden Literaturempfehlungen veranschaulichen anwendungsbezogene Auswertungsverfahren mithilfe eines etablierten Statistikprogramms:

Brühl (2016): SPSS 23. Einführung in die moderne Datenanalyse.

Field (2013): Discovering statistics using IBM SPSS Statistics.

3. Erste ausgewählte Analyseschritte

Mit quantitativen Auswertungsverfahren lassen sich umfangreiche, empirische Daten, z.B. aus Fragebogenstudien oder standardisierten Leistungstests, systematisch verarbeiten und Hypothesen verifizieren (bestätigen) oder falsifizieren (widerlegen). Im Folgenden werden ausgewählte deskriptive Analyseschritte am Beispiel einer Fragebogenerhebung vorgestellt, die für eine erste Systematisierung bzw. Beschreibung von empirischen Datensätzen relevant sind. Der Begriff Signifikanzniveau wird trotz der vorherigen Zuordnung zu den fortgeschrittenen Kenntnissen erläutert, da das Konzept für die Analyse von Zusammenhängen, der Korrelationsanalyse, zentral ist und eine Art Bindeglied zwischen deskriptiver

Statistik und Inferenzstatistik darstellt. Da die multivariate Statistik weitaus umfangreicherer Erläuterungen bedarf, soll an dieser Stelle nur ein kurzer Exkurs auf das Prinzip von ausgewählten Verfahren verweisen.

3.1 Überblick: Lageparameter und Verteilung von Variablen betrachten

Um die Lage und die Verteilung von Variablen exemplarisch zu veranschaulichen, soll eine fiktive Studie dienen. Es wurde eine Fragebogenerhebung durchgeführt, um die Bedeutung der Schulform für die Lesemotivation und ihren schulischen Leistungen von Schülerinnen und Schülern zu ermitteln. Für eine Weiterverarbeitung der erhobenen Daten werden die Antworten aus einem solchen Fragebogen zunächst in Zahlen überführt (vgl. Reinders/Gniewosz 2015, 131). Hierbei kann es sich einerseits um angekreuzte Antworten auf geschlossene Skalen von beispielsweise 1 (trifft nicht zu) bis 7 (trifft völlig zu) handeln, aber ebenso denkbar ist die Quantifizierung von Antworten auf offene Fragen, aus Ratings, Interviewdaten oder Inhaltsanalysen (Heins 2016; Kleinbub 2016; Uhl 2016). Nach dieser Transformation ist ein erster Überblick über die Verteilung der erhobenen Daten erforderlich, um das umfangreiche Datenmaterial zu sichten und zu strukturieren. Insbesondere nach einer langen Feldphase, in welcher der Fokus auf der praktischen Datenerhebung lag, bietet ein Überblick über die Daten eine mentale Reaktivierung der zurückliegenden theoretischen Vorüberlegungen. Welche Probanden haben an meiner Studie teilgenommen, welche Auffälligkeiten zeigen sich bei welchen Skalen?

Im Fragebogen wurden verschiedene Aspekte erfragt und mit verschiedenen Skalenniveaus erfasst. Für die deskriptiven Analysen ist es bedeutsam, sich zunächst das Skalenniveau der Variablen zu vergegenwärtigen, da in Abhängigkeit dessen unterschiedliche statistische Analysemöglichkeiten für die Berechnung von Lageparametern denkbar sind. Die Erhebung von Geschlechtern wird beispielsweise als Nominalskala operationalisiert. Solche Skalen beinhalten unterscheidbare Kategorien, deren Zahlenzuweisungen beliebig sind. Die Zuweisung dient lediglich der Ordnung der Daten und somit als Etikett, d.h. für die spätere Datenanalyse ist es unerheblich, ob das männliche oder das weibliche Geschlecht mit der Zahl 1 oder 2 kodiert wird (vgl. Bortz/Schuster 2010, 13). Ungeachtet der Zuweisung eines Etiketts zu einer Merkmalsausprägung, lassen sich für Nominalskalen absolute und relative Häufigkeiten bestimmen. Bei der absoluten Häufigkeit wird eine konkrete Anzahl errechnet (z.B. 80 Mädchen und 60 Jungen), die relative Häufigkeit hingegen stellt den Anteil einer Merkmalsausprägung an der Gesamtstichprobe dar (z.B. 57,14% Mädchen). Für nominalskalierte Variablen lässt sich zudem bestimmen, welche Merkmalsausprägung am häufigsten in der Stichprobe vorkommt (Reinders/Gniewosz 2015, 132). Hierbei handelt es sich um den Modus (*Mo*), der Auffälligkeiten hinsichtlich des Geschlechts aufzeigen würde, welches am häufigsten von den Schülerinnen und Schülern angekreuzt wurde. Es kann durchaus vorkommen, dass es mehrere Modi in einer Verteilung gibt, wenn mehr als ein Merkmal gleich häufig in einer Stichprobe auftritt.

Ein anderes Skalenniveau besitzen ordinalskalierte Daten, die es erlauben, Aussagen über Größer-Kleiner-Relationen oder Rangfolgen zu tätigen. Mit einem solchen Skalenniveau lassen sich Sortierungen, Ordnungen und Reihenfolgen von Variablen erstellen, die zwar eine natürliche Reihenfolge aufweisen, deren Abstände aber nicht identisch sind. Ein Beispiel dafür sind Schulnoten. Sie geben eine Reihenfolge vor, die Abstände zwischen den Schulnoten sind aber nicht gleich, da zweimal die Note „gut" kein „ausreichend ergibt. Weisen Variablen dieses Skalenniveau auf, so lässt sich zusätzlich zur Häufigkeitsanalyse und der Bestimmung des Modus der Median (*Md*) berechnen (ebd.). Der Median repräsentiert die Mitte eine Stichprobe, indem er sie hälftig teilt in 50% Stichprobenwerte, die kleiner und 50% Werte, die größer als der Median ausfallen. Erfasst man beispielsweise das Alter von Probanden und erhält folgende geordnete Reihung 11-13-15-16-19 mit einer ungeraden Anzahl an Ziffern, läge der zentrale Wert bei 15. Im Falle von Werten wie 11-13-15-18-18-19 mit einer geraden Anzahl umfasst der Median 16,5. Soll zudem die Distanz zwischen größtem und kleinstem Wert ermittelt werden, dann handelt es sich um die Spannweite, z.B. im Hinblick auf das Alter betrüge die Spannweite 8 (Differenz zwischen den Werten 11 und 19).

Möchte man Einstellungen oder Überzeugungen erheben, ist dies mittels Intervallskalen möglich. Intervallskalen stellen eine noch höhere Form des Messens dar als Ordinalskalen, da sie die Berechnung des arithmetischen Mittels (Mittelwert) und der Standardabweichung erlauben. Ein Beispiel für eine Intervallskala ist, wenn Schülerinnen und Schüler mit mehreren Items zu ihrer Lesemotivation auf einer Likert-Skala[1] von 1 (stimme überhaupt nicht zu) bis 5 (stimme völlig zu) befragt werden. Gefragt werden könnte beispielsweise mit mehreren Items danach, ob sie gerne zu ihrem Vergnügen lesen oder ob Lesen eines ihrer Hobbies ist. Zu jedem einzelnen Item kann ein Mittelwert und Standardabweichung gebildet werden.[2] Der Mittelwert (*M*) repräsentiert den Durchschnittswert einer Merkmalsausprägung von allen befragten Personen und die Standardabweichung (*SD*) stellt die durchschnittliche Entfernung aller Messwerte vom Stichprobendurchschnitt dar (ebd.). Beträgt der Mittelwert für das erste Item „Lesen zum Vergnügen" $M = 2{,}10$, so heißt dies, dass die Schülerinnen und Schüler eine weniger ausgeprägte Motivation vorweisen. Liegt die Standardabweichung bei $SD = 0{,}90$, verweist dies darauf, dass die Angaben in der Stichprobe durchschnittlich 0,90 Einheiten vom Mittelwert abweichen. Die Standardabweichung basiert auf dem Maß der Varianz, welches nicht die Entfernung von Messwerten vom Stichprobenmittelwert anzeigt, sondern wie stark jeder einzelne Wert um den Mittelwert

[1] Ob Likert-Skalen als Ordinal- oder Intervallskalen behandelt werden sollten, ist umstritten. Da eine Behandlung von Likert-Skalen als Intervallskalen umfangreichere Analysen erlaubt, wird diese Skalierung oft präferiert.

[2] Eine weitere Möglichkeit besteht darin, einzelne Items einer Skala zu einem neuen Gesamtmittelwert zusammenzufassen. Eine Zusammenfassung von mehreren Items ist nach erfolgter Reliabilitätsanalyse möglich, z.B. wenn verschiedene Items zur Gesamtskala *Lesemotivation* transformiert werden (Betz et al. 2018).

verteilt ist (streut). Median, Mittelwert, Standardabweichung und Varianz lassen sich mit der Software SPSS relativ zügig sowie anwenderfreundlich berechnen. Einen Einblick in die mathematischen Berechnungen bieten Reinders und Gniewosz (2015, 134) sowie Rasch et al. (2014a).

Während die zuvor benannten statistischen Lageparameter Maße der zentralen Tendenz darstellen und die erhobenen Daten mittels Zahlen charakterisieren, ist eine grafische oder tabellarische Veranschaulichung zusätzlich aufschlussreich, um sich Verteilungen und Größenverhältnisse zu vergegenwärtigen. Bei kategorial erhobenen Informationen ist eine Kreuztabelle zielführend, um die gemeinsame Häufigkeitsverteilung zweier Variablen zu sichten. Dabei werden die Häufigkeitsverteilungen zweier kategorialer Variablen gegenübergestellt und in Kombination zueinander betrachtet, z.B. wie viele Mädchen und Jungen in den Schulformen vertreten waren. Würde das Geschlecht als unabhängige Variable gelten, so würden die Häufigkeiten in die Spalten der Tabelle eingetragen. Die Häufigkeiten der abhängigen Variable Schulform hingegen würden in die Zeilen der Tabelle überführt. Folglich werden in der Kreuztabelle die Häufigkeiten für diese zwei Variablen als verknüpfte Merkmale dargestellt. Für Ordinalskalen und Intervallskalen ist eine grafische Visualisierung in Form eines Balkendiagramms und zudem ein Vergleich der Verteilung mit einer Normalverteilungskurve anzuraten (vgl. Abb. 1; Schmitz 2016, 382). In diesem exemplarischen Diagramm wird die Verteilung des auf einer Skala von 1 bis 5 metrisch erfassten Interesses, einen Text zu lesen, in Form von Balken dargestellt und mit der Normalverteilung abgeglichen.

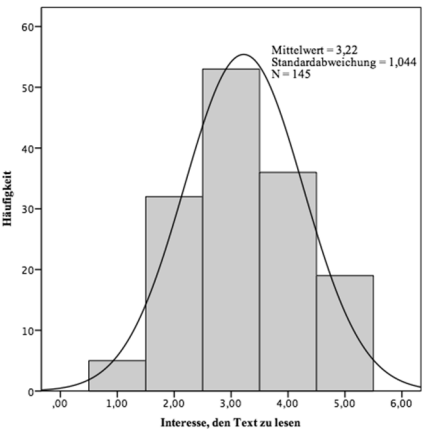

Abb. 1: Balkendiagramm und Normalverteilung (aus: Schmitz 2016, 382).

Die Normalverteilung besitzt nach Carl Friedrich Gauß einen glockenförmigen und symmetrischen Verlauf, der für eine Reihe von mindestens intervallskalierten Merkmalen gültig ist (z.B. für die Körpergröße und den Intelligenzquotienten). Mithilfe eines Balkendiagramms lässt sich augenscheinlich betrachten, wie die tatsächliche Verteilung eines Merkmals in der Datendatei ausfällt und in welcher

Relation sie zur Form einer Normalverteilung steht (vgl. ebd.). Abbildung 1 verdeutlicht, dass die Verteilung des Interesses geringfügig von der Normalverteilung abweicht (ebd., 382). Die Normalverteilung von Daten ist deshalb ein wesentliches Kriterium für Datenanalysen, da Signifikanztests in fortgeschrittenen, inferenzstatistischen Testverfahren auf normalverteilten Daten basieren. Die Prüfung, ob eigens erhobene Daten normalverteilt sind, ist daher sowohl ein wichtiger Schritt der grundlegenden, deskriptiven Datenanalyse als auch eine Schnittstelle zu fortgeschrittenen Analyseverfahren. Für eine statistische Überprüfung, inwieweit eine Verteilung von einer Normalverteilung abweicht, lässt sich der Kolmogorov-Smirnov-Verteilungstest durchführen. Um eine Normalverteilung für ein Merkmal annehmen zu können, ist ein nicht-signifikantes Ergebnis erforderlich (ebd.). Für eine ausführliche Recherche, wie mit nicht-normalverteilten, schiefen oder gewölbten Verteilungsformen umzugehen ist, welche abweichenden Verteilungen es gibt, welche kritischen Grenzwerte existieren und inwiefern sich Verteilungen transformieren lassen, ist das Werk von Bortz/Schuster (2010) zu empfehlen.

3.2 Vertiefung: Zusammenhänge zwischen Variablen analysieren

Nach erfolgter Analyse von Lageparametern und Verteilungsformen beginnt die vertiefte statistische Analyse wodurch weitere Erkenntnisse und Informationen über die erhobenen Daten gewonnen werden. Ein Analyseschritt besteht darin, die Beziehungen (Korrelation) zwischen ausgewählten Variablen eines Datensatzes zu betrachten. Eine Korrelation ist gegeben, wenn die Variation einer Variablen (z.B.: Lesemotivation) mit der Variation einer anderen Variablen (z.B.: Lesekompetenz) zusammenhängt (vgl. Betz et al. 2018; Reinders/Gniewosz 2015, 137). Unterschieden werden positive und negative Merkmalszusammenhänge. Ein positiver Zusammenhang besteht, wenn hohe Werte eines Merkmals x mit hohen Werten eines anderen Merkmals y einhergehen oder wenn niedrige Werte eines Wertes x mit niedrigen Werten eines Wertes y zusammenhängen (vgl. Rasch et al. 2014a, 121). Ein positiver Zusammenhang liegt beispielsweise vor, wenn Schülerinnen und Schüler mit einer ausgeprägten Lesemotivation auch eine bessere Lesekompetenz aufweisen (je mehr – desto mehr bzw. je weniger – desto weniger). Ein negativer Zusammenhang liegt vor, wenn hohe Werte einer Variablen x mit niedrigen Werten einer Variablen y einhergehen und, im umgekehrten Fall, niedrige Werte des Merkmals x mit hohen Werten des Merkmals y einhergehen (je mehr – desto weniger bzw. je weniger – desto mehr) (ebd.; Betz et al. 2018). Eine negative Korrelation liegt beispielsweise vor, wenn weniger Hausaufgaben zu besseren Lernleistungen führen. Bedeutsam an der Interpretation von Zusammenhangsmaßen ist, dass es sich nicht um Ursache-Wirkungs-Zusammenhänge handelt, da die Variation des einen Merkmals durch die Variation des anderen Merkmals und umgekehrt versursacht werden kann (vgl. Müller-Benedict 2011, 267).

In Abhängigkeit vom Skalenniveau kommen unterschiedliche statistische Zusammenhangsanalysen in Frage. Für die Betrachtung von Zusammenhängen zwischen nominalen oder ordinalen Skalen sowie zwischen Variablen mit unterschiedlichen Skalenniveaus gibt es eine Reihe von speziellen Verfahren und Zusammenhangsmaßen, von denen je nach Kombination von zu korrelierenden Skalen eine entsprechende Auswahl zu treffen ist (vgl. Field 2013; Müller-Benedict 2011, 208). Exemplarisch lässt sich aufzeigen, dass ein Zusammenhang zwischen zwei nominalskalierten Variablen über die Darstellung einer Kreuztabelle und den Phi-Koeffizienten ermittelt werden kann. Mittels der punktbiserialen Korrelation werden Zusammenhänge zwischen metrischen und nominalen Skalen berechnet, die Rangkorrelation nach Spearman ist zielführend bei zwei ordinalskalierten Variablen. Im Folgenden wird ausführlicher auf die Produkt-Moment-Korrelation nach Pearson eingegangen, da es sich um das gebräuchlichste Verfahren für die Analyse des Zusammenhangs von metrischen Variablen handelt. Soll beispielsweise untersucht werden, ob Lesemotivation und verbale Grundfähigkeit zusammenhängen, vorausgesetzt diese Variablen wurden intervallskaliert erhoben und sind normalverteilt, so ist die Produkt-Momentkorrelation nach Pearson zielführend. Sie ermittelt mithilfe eines Korrelationskoeffizienten die Höhe eines Zusammenhangs und dessen Richtung. Der Korrelationskoeffizient variiert zwischen $r = -1{,}00$ (perfekter negativer Zusammenhang), $r = 0{,}00$ (kein nachweislicher Zusammenhang) und $r = +1{,}00$ (perfekter positiver Zusammenhang). Die Entscheidung, ab welchem Wert ein Korrelationskoeffizient als inhaltlich bedeutsam interpretiert werden sollte, hängt von der Fragestellung, der Art der Erhebung (Feldstudie oder Experiment) und dem Fachgebiet ab (vgl. Rasch et al. 2014, 126). Als Konvention für die Interpretation von Korrelationskoeffizienten gilt, dass eine geringe Korrelation durch den Wert $r = 0{,}10$ gegeben ist, ein mittlerer Zusammenhang bei einem Wert von $r = 0{,}30$ vorliegt und ein starker Zusammenhang durch $r = 0{,}50$ ausgedrückt wird (vgl. Rasch et al. 2014a, 133; Schäfer 2011, 81). Näheren grafischen Aufschluss über die Korrelation zwischen metrischen Variablen gibt außerdem ein Streudiagramm (Punktewolke) (vgl. Abb. 2; Schmitz 2016, 383). In diesem werden die beobachteten Werte der Variablen x auf der x-Achse (z.B. thematisches Vorwissen) und die Werte der Variablen y auf der y-Achse (Mittelwert im Textverständnis) abgetragen.

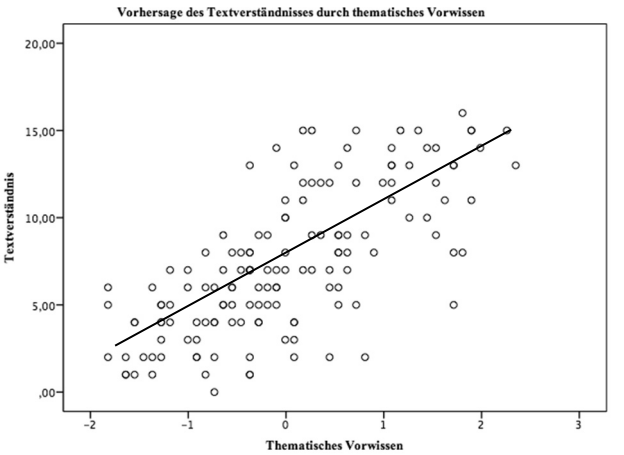

Abb. 2: Streudiagramm (aus: Schmitz 2016, 383).

Im Falle eines positiven, linearen Zusammenhangs würden sich die beobachteten x- und y-Werte treffen und eine exakte Gerade bilden (vgl. Rasch et al. 2014, 121). Ein solch perfekter Zusammenhang ist in empirischen Studien allerdings nahezu nicht nachweisbar wie auch die Abbildung 2 verdeutlicht (vgl. Bortz/Schuster 2010, 153). Ein Grund dafür ist, dass empirische Erhebungen immer durch einen Grad an Ungenauigkeit gekennzeichnet sind, weshalb die Verifizierung oder Falsifizierung einer Hypothese auf einer Wahrscheinlichkeitsrechnung beruht (Reinders/Gniewosz 2015, 135). Würde der Korrelationskoeffizient anzeigen, dass Lesemotivation und Lesekompetenz miteinander korrelieren, gibt die Wahrscheinlichkeitsschätzung zusätzlich an, wie fehlerbehaftet die Annahme dieses Zusammenhangs sein könnte. Neben dem Korrelationskoeffizienten wird daher das Signifikanzniveau (α) ausgewiesen. Etabliert hat sich der kritische Wert $p < .05$, was bedeutet, dass die Wahrscheinlichkeit eines Irrtums bei weniger als 5 Prozent liegt und die Annahme zu 95 Prozent richtig ist. Noch geringer ist die Irrtumswahrscheinlichkeit bei Werten wie $p < .01$ (1 Prozent) und $p < .001$ (0,1 Prozent). Welches Signifikanzniveau in der eigenen empirischen Arbeit fokussiert wird, hängt von der Fragestellung und der Stichprobengröße ab und ist im Hinblick auf die möglichen Konsequenzen, welche aus einer irrtümlichen Annahme folgen würden, abzuwägen.

3.3 Exkurs: Fortgeschrittene Verfahren der Inferenzstatistik

Inferenzstatistische Verfahren zielen auf die Analyse der Wahrscheinlichkeit der Verallgemeinerung von Schlussfolgerungen (vgl. Müller-Benedict 2011, 22), z.B. ob zwischen Merkmalsausprägungen systematische Unterschiede vorliegen oder diese Unterschiede nur zufällig entstanden sind (vgl. Schäfer 2011, 107). Mithilfe der Inferenzstatistik kann beispielsweise untersucht werden, ob systematische Mittelwertsunterschiede in der Skala Lesemotivation (abhängige Variable) durch das Merkmal Geschlecht (unabhängige Variable) erklärt werden können. Bei dem

Vergleich von Mittelwerten zweier unabhängiger Gruppen wird betrachtet, ob sich die Merkmale zufällig (Fehlervarianz) oder statistisch bedeutsam voneinander unterscheiden (systematische Varianz). Mit dem sogenannten *t*-Test lassen sich die Mittelwerte zwischen zwei Gruppen in Relation zu den Streuungen innerhalb der Gruppen vergleichen (vgl. Rasch et al. 2014a; Schäfer 2011). Je größer der Abstand der Mittelwerte zwischen den Gruppen und je geringer die Streuung innerhalb der Gruppen, desto wahrscheinlicher ist es, dass der Unterschied systematisch und nicht zufällig entstanden ist (ebd., 44; Reinders/Gniewosz 2015, 139). Umso größer ist auch die Effektstärke, die sich in Form eines Abstandsmaßes nach Cohen (*d*) errechnen lässt (Schäfer 2011, 111). Verglichen werden bei einem *t*-Test für unabhängige Stichproben ein kritischer *t*-Wert mit einem empirischen *t*-Wert (Reinders/Gniewosz 2015, 138). Überschreitet der empirisch gemessene Wert den kritischen Wert der sogenannten *t*-Verteilung, so ergibt sich „dass der Unterschied zwischen zwei Gruppen größer ist, als bei alleiniger Wirksamkeit des Zufalls erwartet werden kann“ (ebd.). Auf das vorherige Beispiel bezogen würde dies bedeuten: Der Mittelwertsunterschied in der Lesemotivation zweier Gruppen würde signifikant durch das Geschlecht verursacht werden. Das Signifikanzniveau (α) verweist zudem auf die Höhe der Irrtumswahrscheinlichkeit, wenn fälschlicherweise von einem systematischen Mittelwertsunterschied ausgegangen würde (vgl. Kapitel 3.1). Wesentlich dabei ist, dass der *t*-Test auf bestimmten Voraussetzungen basiert, als dass die abhängige Variable intervallskaliert ist, sie eine Normalverteilung aufweist und die Streuungen innerhalb der Gruppen gleich groß sind. Außerdem ist bedeutsam, dass die zu vergleichenden Gruppen unabhängig voneinander sind, sich also nicht gegenseitig beeinflusst haben (vgl. Schäfer 2011, 113).

Sollen darüber hinaus Mittelwerte von mehr als zwei Gruppen systematisch miteinander verglichen werden, ist eine Varianzanalyse (ANOVA: Analysis of Variance) erforderlich. Im Gegensatz zum *t*-Test, bei welchem Mittelwerte in Abhängigkeit von einer unabhängigen Variable mit der Ausprägung in zwei Gruppen (z.B. Geschlecht unterteilt in männlich und weiblich) verglichen werden, kann bei der Varianzanalyse die Variation durch mehr als zwei Variablen untersucht werden (vgl. ebd. 2011, 115). Ein Beispiel wäre, zu ermitteln, ob sich die Lesemotivation in Abhängigkeit von Schulformen mit vier Ausprägungen (Gruppen: Haupt-, Real-, Gesamtschule und Gymnasium) unterscheidet. Hierbei handelt es sich um eine einfaktorielle Varianzanalyse, wohingegen bei einer mehrfaktoriellen Varianzanalyse mehrere unabhängige Variablen mit ihren Ausprägungen einbezogen werden können. Das Signifikanzniveau wird auch bei der Varianzanalyse berechnet und analog zum *t*-Test lässt sich eine Effektgröße für den Mittelwertsunterschied ausweisen, die sich Eta-Quadrat nennt (ebd., 134). Mit einer Varianzanalyse lassen sich darüber hinaus auch Wechselwirkungen, sogenannte Interaktionen, zwischen unabhängigen kategorialen Variablen berechnen. Eine Wechselwirkung liegt vor, wenn sich der Effekt einer unabhängigen Variable durch die andere unabhängige Variable verändert (ebd., 128). Ein Beispiel wäre, wenn sich bei Mädchen ein Effekt des Buchbesitzes auf die Lesemotivation zeigt, bei Jungen hingegen nicht.

Im Gegensatz zu den zuvor benannten Analyseverfahren, ist in größeren Forschungsprojekten mit komplexen, mehrfaktoriellen Erhebungsdesigns statistische Feinarbeit erforderlich, die jedoch auch einen entsprechend hohen Erkenntnisgewinn erzeugt. Möchte man den Einfluss von einer oder mehreren metrischen Variablen (welche Prädiktoren genannt werden) auf eine abhängige Variable (Kriterium) vorhersagen, stellt die Regressionsanalyse ein etabliertes, zielführendes Verfahren dar (vgl. Rasch et al. 2014b; Schmitz 2016). Mittels der Regressionsanalyse lässt sich das relative Gewicht einer unabhängigen Variablen für die Ausprägung einer abhängigen Variable unter gleichzeitiger Berücksichtigung von weiteren unabhängigen Variablen schätzen. Beispielsweise kann überprüft werden, zu welchem prozentualen Anteil schulische Leistungen durch Interesse, Motivation und soziale Herkunft vorhergesagt werden können und welchen Einfluss jeder einzelne Prädiktor oder sämtliche Prädiktoren in Kombination auf die Leistung nimmt. Für die Einarbeitung in die Regressionsanalyse sei auf Schäfer (2011) und Rasch et al. (2014b) hingewiesen. Neben der Regressionsanalyse bieten die Pfadanalyse sowie die Modellierung von Strukturgleichungsmodellen spezifische Möglichkeiten, eine Reihe von Variablen mit verschiedenen Skalenniveaus zu berücksichtigen und ihren Einfluss auf eine oder mehrere abhängige Variablen zu schätzen (modellieren). Da es sich hierbei um ein äußerst komplexes Analyseverfahren handelt, die auf dem Verstehen der zuvor benannten Testverfahren aufbauen, sollen sie an dieser Stelle nicht näher behandelt werden.

4. Schlussbemerkungen

Der Beitrag bot einen kurzen Überblick über statistische Begrifflichkeiten und Verfahren, die für die Datenanalyse von kleineren empirischen Projekte relevant sind und die keine besonderen mathematischen Vorkenntnisse über die dahinterliegenden Prinzipien erfordern. Der Erwerb der benannten Grundlagen ist deshalb so wichtig, um eigens erhobene Daten systematisch zu beschreiben und auszuwerten sowie um eine Basis für die Analyse von Daten aus größeren empirischen Projekten und das Durchführen von komplexeren Verfahren zu schaffen (vgl. Kapitel 3.3). Für die eigene Einarbeitung und weitere Recherche in die quantitative Statistik ist zu betonen, dass die Analyseverfahren und statistischen Kennwerte in den Nachbardisziplinen, überwiegend in der Psychologie, entwickelt wurden. Gängige Kennwerte, wie die Reliabilität eines Testinstruments, Korrelationskoeffizienten, Effektstärken oder das Signifikanzniveau, wurden in artifiziellen Testbedingungen generiert und sind mitunter in deutschdidaktischen Forschungsfeldern aufgrund des Feldzugangs und der dort vorherrschenden nicht vollends kontrollierbaren Einflüsse nur bedingt erreichbar bzw. entsprechend zu relativieren. Dennoch sollte sich die deutschdidaktische Forschung an den etablierten Kennwerten und Interpretationskonventionen orientieren und jene Werte, die einen Grenzwert völlig aushebeln, kritisch hinterfragen und zur Diskussion stellen. Die forschungsmethodischen Zugänge in den Nachbardisziplinen sollten im Hin-

blick auf die Möglichkeiten, aber auch ihre Grenzen, somit durchaus wahrgenommen werden, um die Diskussion über und die Entwicklung von empirischen deutschdidaktischen Methoden voranzutreiben.

Literatur

Betz, Anica/Kröger-Bidlo, Hanna/Schmitz, Anke/Schuttkowski, Caroline (2018, i.Dr.): Empirische Methoden in der Sprachdidaktik. Ausgewählte Erhebungs- und Auswertungsverfahren. In: Rothstein, Björn/Müller, Claudia (Hrsg.): Kernbegriffe der Sprachdidaktik Deutsch. Ein Handbuch. Thema Sprache 1. Handbücher für den Unterricht. Baltmannsweiler: Schneider Hohengehren.

Boelmann, Jan M. (Hrsg.) (2016): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. Baltmannsweiler: Schneider Hohengehren.

Bortz, Jürgen/Döring, Nicola (2006): Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler. 4. Aufl. Heidelberg: Springer.

Bortz, Jürgen/Schuster, Christoph (2010): Statistik für Human und Sozialwissenschaftler. Berlin: Springer.

Brühl, Achim (2016): SPSS 23. Einführung in die moderne Datenanalyse. Hallbergmoos: Pearson.

Field, Andy (2013): Discovering statistics using IBM SPSS Statistics. London u.a.: Sage Publications.

Heins, Jochen (2016): Qualitative Inhaltsanalyse. In: Boelmann, Jan M. (Hrsg.): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. Baltmannsweiler: Schneider Hohengehren, 305-324.

Kleinbub, Iris (2016): Kriteriengeleitetes Rating. In: Boelmann, Jan M. (Hrsg.): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. Baltmannsweiler: Schneider Hohengehren, 343-360.

Müller-Benedict, Volker (2011): Grundkurs Statistik in den Sozialwissenschaften. Wiesbaden: Springer.

Rasch, Björn/Friese, Malte/Hofmann, Wilhelm/Naumann, Ewald (2014a): Quantitative Methoden 1. Einführung in die Statistik für Psychologen und Sozialwissenschaftler. Wiesbaden: Springer.

Rasch, Björn/Friese, Malte/Hofmann, Wilhelm/Naumann, Ewald (2014b): Quantitative Methoden 2. Einführung in die Statistik für Psychologen und Sozialwissenschaftler. Wiesbaden: Springer.

Reinders, Heinz/Gniewosz, Burkhardt (2015): Quantitative Verfahren. In: Reinders, Heinz/Ditton, Hartmut/Gräsel, Cornelia/Gniewosz, Burkhardt (Hrsg.): Empirische Bildungsforschung. Strukturen und Methoden. Wiesbaden: Springer, 131-153.

Schäfer, Thomas (2011): Statistik II. Inferenzstatistik. Wiesbaden: Springer.

Schmitz, Anke (2016): Interaktionsanalysen mittels multipler Regression. Veranschaulichung des Analyseverfahrens anhand einer Studie zur Wirkung der globalen Textkohäsion auf das Textverständnis. In: Boelmann, Jan M. (Hrsg.): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. Baltmannsweiler: Schneider Hohengehren, 377-392.

Uhl, Benjamin (2016): Quantitative Inhaltsanalyse. In: Boelmann, Jan M. (Hrsg.): Empirische Erhebungs- und Auswertungsverfahren in der deutschdidaktischen Forschung. Baltmannsweiler: Schneider Hohengehren, 325-342.

Wie wird aus einer ersten Idee ein konkretes Forschungsprojekt? Der vorliegende Band widmet sich im ersten Teil den theoretischen Grundlagen empirischen Forschens: Was unterscheidet qualitative und quantitative Forschung? Welcher Forschungsansatz ist für meine Fragestellung der richtige und was sind die Vor- und Nachteile verschiedener Untersuchungsdesigns? Im zweiten Teil werden praxisrelevante Hilfestellungen für die Organisation, Durchführung und Auswertung der Erhebung gegeben, wobei der gesamte Prozess von der Auswahl der Probanden über Datenschutzfragen, Softwareeinsatz bis hin zu statistischen Grundlagenkenntnissen berücksichtigt wird.

Ein Grundlagenwerk für Studierende und Promovierende.

Dr. Jan M. Boelmann ist Professor für Literatur- und Mediendidaktik.
Seine Arbeits- und Forschungsschwerpunkte liegen im medienübergreifenden literarischen Lernen, der Modellierung literarischer Kompetenz sowie in der Kinder- und Jugendliteratur.

9783834018816

Schneider Verlag Hohengehren