Bundesministerium
für Bildung
und Forschung

Forschungsvorhaben in Ankopplung an Large-Scale-Assessments

Bildungsforschung Band 44
Forschungsvorhaben in Ankopplung an Large-Scale-Assessments

Inhalt

Einleitung

Bildungsvergleichsstudien bzw. Large-Scale-Assessments sind ein zentrales Element des Bildungsmonitorings. Sie leisten einen wichtigen Beitrag, um Ergebnisse von Bildungsprozessen zu verstehen. Darüber hinaus wird durch die systematische wissenschaftliche Beobachtung zuverlässiges und differenziertes Wissen über Stärken und Schwächen des Bildungwesens in Deutschland bereitgestellt.

Forschung zu Large-Scale-Assessments unterstützt ihre wissenschaftliche Weiterentwicklung und Aktualität. Auch ist eine im internationalen Vergleich hervorragende Forschung notwendig, um längerfristig Einfluss auf die internationalen Large-Scale-Assessments nehmen zu können.

Das Bundesministerium für Bildung und Forschung hat daher auf der Basis einer Förderbekanntmachung aus dem Jahr 2011 zehn Projekte in diesem Bereich von 2012 bis 2015 gefördert. Sie leisteten auf sehr unterschiedliche Weise wertvolle Beiträge für die Weiterentwicklung von Large-Scale-Assessments. In den Projekten wurden methodische Ansätze zur Erklärung von Kompetenzunterschieden geprüft, Tests für die Erfassung von Kompetenzen entwickelt, die bislang selten oder nicht getestet wurden, sowie die Forschung zur Vergleichbarkeit von Kompetenztests und deren Fairness und Validität weiterentwickelt.

In dem vorliegenden Band werden die Ergebnisse der Projekte vorgestellt. Er gibt somit einen guten Überblick über aktuelle Entwicklungen in diesem Forschungsfeld.

Heinz Reinders, Martin Fresow, Paulina Fresow

Kompetenzunterschiede und Bildungsgangwechsel bei Schülern mit Migrationshintergrund – Ergebnisse zur Vorhersage der Mathematikleistungen durch individuelle Voraussetzungen

1 Einleitung

Nach wie vor gilt in der öffentlichen Wahrnehmung die Zugehörigkeit zur Gruppe der Migrantinnen und Migranten als prinzipieller Nachteil für die Teilhabe an gesellschaftlichen Ressourcen. Migrantinnen und Migranten weisen ein geringeres Bildungsniveau auf, erreichen geringere Bildungsabschlüsse bzw. verlassen häufiger die Schule ohne Abschluss und sind häufiger von Arbeitslosigkeit oder prekärer Beschäftigung bedroht.

Tatsächlich sprechen aktuelle Statistiken für diese Wahrnehmung. Personen mit Migrationshintergrund haben seltener Hochschulabschlüsse und häufiger keinen Schulabschluss als die Vergleichsgruppe ohne Migrationshintergrund (Bildungsbericht, 2014, 40 f.). Auch nehmen Kinder aus Migrantenfamilien seltener Angebote der vorschulischen Bildung und Betreuung wahr als ihr gleichaltriges Pendant (Bildungsbericht, 2014, 56).

Diese Perspektive hat jedoch zwei Nachteile. Zum einen ist sie im Kern nicht zutreffend, und sie bietet zum anderen keine Ansatzpunkte für pädagogische Interventionen. Sie ist nicht zutreffend, weil das Merkmal Migrationshintergrund lediglich eine Proxy-Variable für komplexe Merkmalskombinationen darstellt. Sozioökonomischer Status, Bildungsnähe der Familie, regionale Wohnsituation, Kompetenzzuschreibungen durch Lehrkräfte etc. sind nur einige der Variablen, die Unterschiede zwischen Personen deutlich besser erklären als das dichotome Merkmal Migrationshintergrund (zusf. Stanat & Edele, 2015). Eine theoretische Dekomposition und damit Klassifikation der differenzierenden Merkmale findet sich bereits im Modell von Watermann und Baumert (2006), und zahlreiche Befunde wurden seither berichtet, die eine globale Wirkung von individueller oder familialer Wanderungserfahrung stark in Zweifel ziehen.

Die Perspektive bietet zweitens keine Ansatzpunkte für pädagogische Interventionen, weil – wie dargestellt – nicht das Merkmal an sich, sondern allenfalls korrespondierende Korrelate Unterschiede zwischen den Gruppen erklären. Hinzu kommt, dass die beiden vermutlich stärksten Begleiterscheinungen von Migration – Bildungsferne der Eltern und sozioökonomischer Status – kaum systematisch veränderbar sind. Zwar wäre es prinzipiell möglich und wird in Modellprojekten praktiziert, Migranteneltern in die Bildungsbiografie einzubinden, etwa durch Sprachkurse

an Schulen, Lernbegleiter in den Familien und dergleichen. Allerdings gestaltet sich die Offenheit von Schulen und Lehrkräften für ein derartiges Mesosystem schwierig, was häufig in der geringen Bereitschaft von Lehrkräften begründet ist, systematische Elternarbeit zu betreiben (Hillesheim, 2009). Wenngleich diese Kooperation zwischen Elternhaus und Schule unabdingbar ist, rücken gleichwohl stärker Individualmerkmale von Schülerinnen und Schülern mit Migrationshintergrund in den Vordergrund. Diese weisen nicht nur eine ungleich größere Varianz als soziale Merkmale der Familie oder der Herkunftsgruppe auf. Sie bieten darüber hinaus Ansatzpunkte für die Förderung von Kindern und Jugendlichen, die sich auch im Bildungsalltag der Schule umsetzen lassen. Einfach formuliert: Schülerinnen und Schüler sind den Tag über in der Schule und sind daher gut erreichbar für Lehrkräfte. Die Eltern, deren Bildungsabschluss und Geldbeutel sind es nicht.

Es ließe sich noch pointierter formulieren, dass für die Erziehung und Bildung von Heranwachsenden seit der Subjektwende in der Pädagogik das einzelne Kind gilt – mit seinen individuellen Bedürfnissen, Fähigkeiten und Fertigkeiten. Warum sollten dann nicht auch individuelle Merkmale von Migrantenkindern als Ansatzpunkt für deren Erziehung und Bildung sowie Bildungschancen gelten können?

Aus diesen beiden Gründen – geringer prognostischer Wert von Migrationshintergrund als Prädiktor sowie pädagogische Orientierung an individuellen Voraussetzungen – widmet sich das Projekt „KuBiS – Kompetenzunterschiede und Bildungsgangwechsel bei Schülern mit Migrationshintergrund“ der Frage, welchen Einfluss personale Merkmale von Schülerinnen und Schülern auf deren Bildungserfolg besitzen. Diese Frage wird allgemein für Schülerinnen und Schüler ausgiebig in der Lehr-Lern-Forschung behandelt (Gräsel & Gniewosz, 2015) und gewinnt auch innerhalb der Migrationsforschung bzw. der an Migration interessierten Bildungsforschung an Prominenz (zusf. Fresow et al., 2012).

Der vorliegende Beitrag berichtet erste Ergebnisse aus dem im März 2015 abgeschlossenen Projekt und gibt dabei einen Einblick in die vergleichenden Analysen der repräsentativen Datengrundlage aus PISA und einer eigenständig neu konzipierten Längsschnittstudie bei Schülerinnen und Schülern in Bayern.

2 Ziele des Projekts

Das Ziel des Projekts KuBiS ist es, individuelle Merkmale zu identifizieren, die jenseits von Migrationshintergrund und sozioökonomischem Status den Bildungserfolg von Schülerinnen und Schülern mit Migrationshintergrund erklären. Als Indikatoren für den Bildungserfolg werden der Wechsel auf den Mittlere-Reife-Zweig (M-Zweig) an bayerischen Mittelschulen einerseits sowie die Mathematikleistungen der Schülerinnen und Schüler andererseits herangezogen.

Vorausgesagt werden der Wechsel auf den M-Zweig bzw. die Mathematikleistungen durch die intrinsische Motivation, die Bildungsaspirationen sowie die Selbstwirksamkeitserwartung der Schülerinnen und Schüler. Diese haben sich aus theoretischer und empirischer Sicht als ertragreiche Bedingungen des Bildungserfolgs herauskristallisiert und waren zugleich auch in den Daten der PISA-Studie verfügbar.

Zur Prüfung des theoretischen Modells wurden im Projekt die Daten der PISA-Erhebung aus dem Jahr 2003 einer erneuten Analyse unterzogen und um eine Längsschnittstudie bei Schülerinnen und Schülern in Bayern ergänzt. Es wurden gleiche oder vergleichbare Instrumente in der Längsschnittstudie gewählt, um die in einer repräsentativen PISA-Studie identifizierten Befunde im Längsschnitt replizieren zu können und auf diese Weise die Plausibilität von Ergebnissen einer stark ausgelesenen Stichprobe zu erhöhen.

Letztlich werden die Ergebnisse dazu genutzt, die Triftigkeit der Annahmen über individuelle Voraussetzungen für den Bildungserfolg von Schülerinnen und Schülern mit Migrationshintergrund zu prüfen und daraus Folgerungen für die schulische Bildungspraxis abzuleiten.

3 Forschungsstand[1]

Für die Bildungskarrieren von Schülerinnen und Schülern mit Migrationshintergrund sind drei Merkmale kennzeichnend:

Erstens finden sich Schülerinnen und Schüler mit Migrationshintergrund überproportional häufig in den unteren Bildungsgängen des deutschen Schulsystems (vgl. Baumert & Schümer, 2002; Bos et al., 2007; Ditton, Krüsken & Schauenberg, 2005; Schofield et al., 2006). Erklärende Variable für den überproportional häufigen Besuch der Hauptschule ist die soziale Schicht, der die Migrantenschülerinnen und -schüler angehören (vgl. Bos et al., 2007; Ditton et al., 2005; Klieme et al., 2010; Steinbach & Nauck, 2004).

Zweitens sind Schülerinnen und Schüler mit Migrationshintergrund im Vergleich zu ihren Mitschülerinnen und -schülern ohne Migrationshintergrund häufiger von einem Schulartwechsel betroffen (vgl. Diefenbach, 2002; Konsortium Bildungsberichterstattung, 2006; Staatsinstitut für Schulqualität und Bildungsforschung, 2009). Bei Bildungskarrieren Jugendlicher mit Migrationshintergrund sprechen Schulze und Soja (2006) von „verschlungenen Bildungswegen". Die Bildungskarrieren dieser Population sind sehr häufig von Ab- oder Aufstiegen gekennzeichnet und weniger von gradlinigen Bildungsverläufen als bei der Kohorte der heranwachsenden deutschen Jugendlichen. So verzeichneten Schulze und Soja (2006), dass 18 Prozent der Bildungsverläufe befragter Migrantenjugendlicher in Köln von Ab- und Aufstiegen gekennzeichnet waren (Schulze & Soja, 2006, 198). Es zeigt sich sowohl für Deutschland insgesamt als auch für Bayern im Besonderen, dass Jugendliche mit Migrationshintergrund mehr Schulartwechsel vollziehen. Besonders hoch ist die Wechselquote zwischen der Haupt- und Realschule bzw. der Wirtschaftsschule (Diefenbach, 2002, 32; Staatsinstitut für Schulqualität und Bildungsforschung, 2009, 94). Bei diesem Wechsel zeigen sich mehr Aufstiege bei Migranten als bei Nichtmigranten (Diefenbach, 2002, 32; Staatsinstitut für Schulqualität und Bildungsforschung, 2009, 94).

Drittens erzielen Jugendliche mit Migrationshintergrund bei Kompetenzmessungen im Durchschnitt geringere Leistungen als Gleichaltrige ohne Migrationshinter-

1 Der vorliegende Beitrag basiert teilweise auf bereits veröffentlichte Projektskizzen (vgl. Fresow et al., 2012; Rettich et al., 2012). Bei dem hier berichteten Forschungsstand und der Theorie handelt es sich um Aktualisierungen und Überarbeitungen der ursprünglichen Darstellungen.

grund (vgl. Klieme et al., 2010; Walter & Taskinen, 2007). Jedoch nähern sich diese Kompetenzunterschiede nach Kontrolle von sozioökonomischem Status, Bildungsniveau der Eltern sowie familialem Sprachgebrauch an (vgl. Walter & Taskinen, 2007).

Kennzeichnend für alle drei Merkmale der Bildungskarrieren von Migranten ist, dass die Herkunft häufig als Proxy-Variable zur Erklärung von Niveauunterschieden herangezogen bzw. in jüngeren Studien der Migrationshintergrund tatsächlich als Herkunftseffekt des sozioökonomischen Status identifiziert wird. Ferner wird das Augenmerk regelmäßig auf den Vergleich zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund gelegt. Weiterführend für die Erklärung differenzieller Bildungskarrieren sind zudem Studien, die lernerfolgsrelevante Merkmale auf Personenseite in den Blick nehmen. Da im Projekt KuBiS die Bildungsaspirationen (Kapitel 3.1), die Lernmotivation (Kapitel 3.2) und die akademische Selbstwirksamkeit (Kapitel 3.3) als solche lernerfolgsrelevanten Merkmale im Mittelpunkt stehen, konzentriert sich die Darstellung des Forschungsstands auf diese erklärenden Variablen und wird abschließend einer kritischen Würdigung unterzogen (Kapitel 3.4).

3.1 Bildungsaspirationen

In verschiedenen Untersuchungen gibt es Evidenz dafür, dass Migrantenjugendliche nach Kontrolle von sozialem Hintergrund sowie Kompetenzen höher qualifizierende Bildungsgänge als Jugendliche ohne Migrationshintergrund anstreben (vgl. Becker, 2010; Organisation for Economic Co-operation and Development [OECD], 2006; Stanat & Christensen, 2006; Stanat, Segeritz & Christensen, 2010).

Gemäß der „Immigrant Optimism"-Hypothese resultieren die höheren Bildungsaspirationen der Jugendlichen mit Migrationshintergrund aus der Erwartung und dem Wunsch der Eltern, einen sozialen Aufstieg im Aufnahmeland zu realisieren (Kao & Tienda, 1995). Dies kann die Varianzen in den Bildungsaspirationen zwischen den verschiedenen Migrantengruppen und unterschiedlichen Migrantengenerationen erklären. So zeigten Stanat, Segeritz und Christensen (2010) in einer Analyse der PISA-2003-Daten, dass Jugendliche türkischer und polnischer Herkunft kontrolliert nach Leistung sowohl höhere schulische als auch höhere berufliche Aspirationen aufweisen als Gleichaltrige ohne Migrationsgeschichte.

Auch innerhalb der Migrantengruppe lassen sich Variationen in den Bildungsaspirationen ausmachen, die sich zumeist auf den sozioökonomischen Status zurückführen lassen: Jugendliche, die in einer Familie mit niedrigerem sozioökonomischen Status aufwachsen, tendieren eher dazu, geringere und zudem weniger stabile Aspirationen aufzuweisen (vgl. Kao & Tienda, 1998).

3.2 Lernmotivation

Auch bezüglich der Lernmotivation weisen Schülerinnen und Schüler mit Migrationshintergrund höhere mittlere Ausprägungen auf als jene ohne Migrationshintergrund (vgl. Helmke et al., 2002; OECD, 2006; Stanat & Christensen, 2006). Dies zeigt

sich im besonderen Maße bei den Kindern der ehemaligen „Gastarbeitergeneration“ sowie bei der Gruppe der sogenannten „Spätaussiedler“ – auch unter statistischer Kontrolle der Leistungen (Stanat, Segeritz & Christensen, 2010).

Aktuelle Ergebnisse einer Studie von Hartmann und McElvany (2013) bestätigen dieses Muster für die intrinsische Motivation im Bereich der Mathematik. Für Schülerinnen und Schüler der vierten Klasse mit türkischem Migrationshintergrund zeigten sich höhere mittlere Ausprägungen im Vergleich zu den Schülerinnen und Schülern ohne Migrationshintergrund. Demgegenüber zeigten sich für die Leistungen in Mathematik für die Migrantinnen und Migranten signifikant niedrigere Werte im Gruppenvergleich. Darüber hinaus wurde für Schülerinnen und Schüler ohne Migrationshintergrund von Hartmann und McElvany (2013) ein Zusammenhang zwischen intrinsischer Motivation und Leistung nachgewiesen, für die Gruppe der Schülerinnen und Schüler mit türkischem Migrationshintergrund jedoch nicht. Allerdings ließen sich diese Unterschiede in der Höhe des Zusammenhangs nicht statistisch absichern. Der sozioökonomische Status erwies sich in der Gruppe der Schülerinnen und Schüler ohne Migrationshintergrund als signifikant positiver Prädiktor für die Mathematikleistungen, in der Gruppe der Schülerinnen und Schüler mit türkischem Migrationshintergrund demgegenüber nicht. Gleichzeitig konnte für diese Gruppe auch kein Einfluss des sozioökonomischen Status auf die Motivation nachgewiesen werden, während dieser für die Schülerinnen und Schüler ohne Migrationshintergrund ebenfalls nicht signifikant ausfiel – lediglich auf dem eher unüblichen Signifikanzniveau von zehn Prozent.

Einschränkend muss jedoch berücksichtigt werden, dass die Ergebnisse von Hartmann und McElvany (2013) auf querschnittlichen Analysen und einem lediglich über den Sprachgebrauch in der Familie operationalisierten Migrationshintergrund basieren.

3.3 Selbstwirksamkeit

Allgemein wird die Selbstwirksamkeitserwartung in jüngerer Forschung zunehmend auch in ihrem Zusammenspiel mit weiteren personalen und kontextuellen Merkmalen untersucht und erweist sich dabei als zentraler Prädiktor für Leistungsdisparitäten bzw. den Kompetenzerwerb (vgl. Liem, Lau & Nie, 2008; Schunk & Mullen, 2012; Zuffianò et al., 2013).

Auf Basis einer Analyse der PISA-2003-Daten kommen Stanat und Christensen (2006) zu dem Ergebnis, dass es sich bei der Selbstwirksamkeit um einen „der stärksten Prädiktoren der Schülerleistungen“ (Stanat & Christensen 2006, 117) handelt.

„Die Ergebnisse zeigen, dass mit einem Anstieg von einer Indexeinheit (bzw. einer Standardabweichung) auf der Skala der Selbstwirksamkeit in Mathematik in den OECD-Erhebungsländern ein Anstieg in der Mathematikleistung [...] einhergeht. Dies entspricht fast einer ganzen Kompetenzstufe in Mathematik“ (Stanat & Christensen, 2006, 117).

Für Schülerinnen und Schüler mit Migrationshintergrund zeichnen Stanat und Christensen (2006) ein differenziertes Bild: In gut der Hälfte der Teilnehmerländer

zeigen sich im internationalen Vergleich für Migrantinnen und Migranten im Vergleich zu Schülerinnen und Schülern ohne Migrationshintergrund höhere oder vergleichbare Ausprägungen in der Selbstwirksamkeit (Stanat & Christensen, 2006, 116 ff.). In den anderen Ländern verfügen diese über eine niedrigere Selbstwirksamkeit – zu diesen Ländern gehört auch Deutschland (Stanat & Christensen, 2006, 116 ff.). Und dies gilt in diesem Fall sowohl für die erste als auch die zweite Generation der Migrantinnen und Migranten (Stanat & Christensen, 2006, 116 ff.). Darüber hinaus verweisen Stanat und Christensen (2006) in diesem Zusammenhang auf weitere, insbesondere für das Projekt KuBiS vielversprechende Details:

„In den meisten Ländern verschwinden diese Unterschiede allerdings nach Berücksichtigung des sozioökonomischen Hintergrunds der Schülerinnen und Schüler" (Stanat & Christensen, 2006, 117). Des Weiteren zeigen die Analysen, „dass die Schülerinnen und Schüler der ersten und zweiten Generation nach Berücksichtigung der Mathematikleistungen in den meisten Erhebungsländern eine signifikant stärkere Selbstwirksamkeit zeigen als ihre Altersgenossen aus einheimischen Familien" (Stanat & Christensen, 2006, 117).

Diese ersten querschnittlichen Analyseergebnisse auf Basis der PISA-2003-Daten liefern vielversprechende Ansatzpunkte bezüglich der Bedeutung der Selbstwirksamkeit bei der Genese und Erklärung von Leistungen und Leistungsdisparitäten.

3.4 Kritische Würdigung

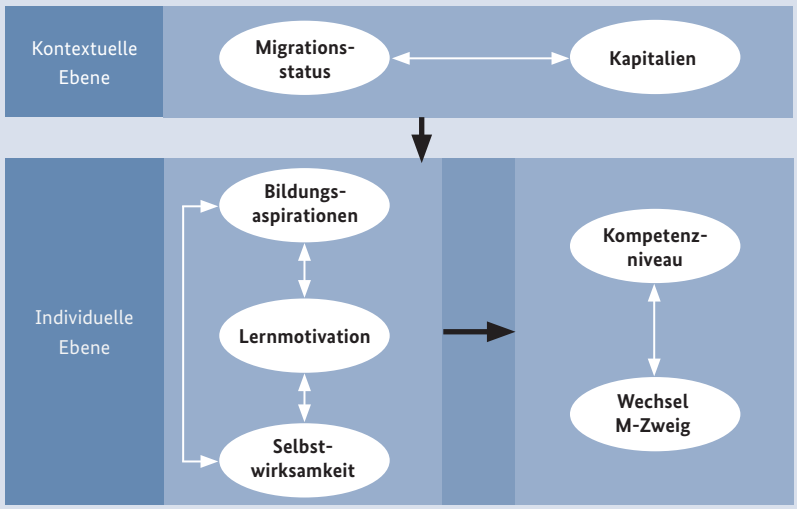
Ein wesentlicher Fortschritt in der Migrationsforschung ist es, bildungsbezogene Disparitäten nicht nur monokausal über ethnische Herkunft als einfache Proxy-Variable zu bestimmen, sondern die mit der Herkunft verknüpften Merkmale, wie den sozioökonomischen Status, als Erklärungen heranzuziehen und gleichzeitig die mit Kapitalausstattungen verknüpften kulturellen Praxen in der Familie in den Blick zu nehmen (Watermann & Baumert, 2006). Ferner macht bisherige Forschung deutlich, dass Migrantenjugendliche in den drei erklärenden Variablen Aspirationen, Selbstwirksamkeit und Leistungsmotivation zumeist höhere Ausprägungen aufweisen als Jugendliche deutscher Herkunft. Schließlich lassen sich Variationen zwischen den verschiedenen Migrantengruppen entdecken, die sich aus Unterschieden in der Kapitalausstattung und der kulturellen bzw. Bildungspraxis erklären lassen.

Neben dem Vergleich zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund bietet sich daher der Fokus auf Varianzen innerhalb von Migrantenjugendlichen an, um mögliche Zusammenhänge zwischen den unabhängigen Variablen Aspiration, Selbstwirksamkeit und Leistungsmotivation mit den abhängigen Variablen Bildungsgangwechsel und Kompetenzniveau innerhalb dieser Population zu entdecken. Studien, die einzelne Aspekte untersucht haben, lassen sich auf diese Weise gut durch empirische Analysen ergänzen, die alle drei Variablen simultan und vor dem Hintergrund der Kapitalausstattung von Migrantenschülerinnen und -schülern betrachten und den Einfluss auf Bildungsgangwechsel und Kompetenzniveaus theoriegeleitet bestimmen.

4 Theorie

Den theoretischen Rahmen für die Untersuchung bildet das Erwartungs-x-Wert-Modell sensu Eccles und Kollegen (Eccles & Wigfield, 2002; Eccles, 2005). Im Mittelpunkt des Projekts stehen Bildungsaspiration, akademische Selbstwirksamkeit sowie die Lernmotivation als Mediatoren zwischen den Kontextmerkmalen einerseits und den abhängigen Variablen Bildungsgangwechsel und Kompetenzerwerb bei Migrantenschülerinnen und -schülern andererseits. Die zentrale Annahme ist, dass die durch kontextuelle Ausprägungen erklärten Unterschiede im Kompetenzniveau und dem besuchten Bildungsgang über Aspirationen, Lernmotivation und Selbstwirksamkeit vermittelt werden können (vgl. Abbildung 1).

Abb. 1: Theoretisches Modell zur Vorhersage des Bildungserfolges von Schülerinnen und Schülern mit Migrationshintergrund



Dieses Modell ist in seiner Struktur kulturinvariant angelegt, also sollten die postulierten Zusammenhänge zwischen den einzelnen Modellkomponenten auf Schülerinnen und Schülern mit und ohne Migrationshintergrund anwendbar sein, obwohl sich die Ausprägungen der Komponenten selbst unterscheiden können.

Als wichtige Prädiktoren auf der kulturellen Ebene führt das Modell Variationen in der Ausstattung der Familien mit sozialem, kulturellem und ökonomischem Kapital an (vgl. Bourdieu, 1992), von denen bekannt ist, dass sie Leistungsunterschiede im deutschen Schulsystem zwischen Migrantinnen und Migranten und Nichtmigrantinnen und -migranten, aber auch innerhalb der Migrantenpopulation teilweise erklären können (Walter & Taskinen, 2007; Walter, 2009; Stanat & Edele, 2015). Hierunter fallen unter anderem die finanzielle Ausstattung der Familien, die erworbenen Bildungszertifikate, das Wissen über das Bildungssystem, die sozialen Einbindungen der Familien und das soziale Umfeld.

Gemäß dem Erwartungs-x-Wert-Modell werden diese kulturbezogenen, eher distalen Einflüsse und die Effekte der Variation bezüglich der Kapitalienausstattung auch innerhalb der Gruppe der Migranten auf den Bildungsgangwechsel und den Kompetenzerwerb vollständig über individuelle Variablen vermittelt. Als zentrale Prädiktoren akademischer Entscheidungen und Leistungen werden (1) der subjektive Wert, (2) die akademische Selbstwirksamkeit als Bereich allgemeiner Erfolgserwartungen sowie (3) Bildungsaspirationen beschrieben.

Zu (1): Viele Facetten des Lern- oder Schulmotivationsbegriffes finden sich in der Wertkomponente wieder, die sich in eine intrinsische (Spaß, Interesse), eine Nützlichkeits- und Wichtigkeitskomponente sowie in die wahrgenommenen Kosten aufgliedern lässt (Rheinberg, 2004). Werte werden als „guiding principles in people's lives" (Schwartz, 1996, 2) verstanden. Sie beeinflussen direkt Handlungsentscheidungen und stehen im Falle hoher Leistungswerte im Zusammenhang zur Volition und Persistenz beim schulischen Lernen (Hofer et al., 2007; Hofer, Reinders & Fries, 2010). Schülerinnen und Schüler, die Spaß beim Lernen empfinden, die ihren Tätigkeiten in der Schule einen Nutzen (extrinsische Komponente) beimessen, die die Beschäftigung mit schulischen Lerninhalten für ihre Person als wichtig und identitätsrelevant erachten und die die Kosten, die mit dem Lernaufwand verbunden sind, als angemessen ansehen, werden eher bildungsbezogene Entscheidungen treffen, die eine positive Leistungsentwicklung nach sich ziehen (Boudon, 1974; Eccles & Wigfield, 2002; Hofer, Reinders & Fries, 2010; Schmid et al., 2005).

Zu (2): Die zweite Hauptkomponente des Erwartungs-x-Wert-Modells ist die subjektive Erwartung, dass ein bestimmtes Verhalten zu einem positiv bewerteten Ergebnis führt. In diesem Kontext spielen Selbstwirksamkeitsüberzeugungen eine prominente Rolle (Bandura, 2001; Schunk & Pajares, 2005). Die Antezedenzien der Erfolgserwartungen sind Erfahrungen, stellvertretende Verstärkungen, verbale Bekräftigungen sowie körperliche und emotionale Reaktionen in den Leistungssituationen. Die Wahrscheinlichkeit dieser Erfahrungen kann wiederum durch den sozialen Kontext bzw. den sozialen und ethnischen Hintergrund bestimmt werden. Kollektive Identitäten oder die Erfahrungen, Mitglied einer sozialen Gruppe zu sein, die systematisch anders behandelt wird (vgl. Diskriminierung, differenzielle Verstärkungen im sozialen Kontext), bestimmen in nicht geringem Maße die wahrgenommenen Erfolgswahrscheinlichkeiten in schulischen Leistungssituationen (vgl. Aronson, Steele, Elliot & Dweck, 2005). Von diesen Erfahrungen kann angenommen werden, dass sie auch innerhalb der Gruppe der Migrantinnen und Migranten, z. B. nach Herkunftsland, variieren.

Die Selbstwirksamkeit fungiert als Mediator zwischen Kontexteinflüssen, deren Interpretationen und den Aufgabenwerten bzw. Erfolgserwartungen (vgl. Denissen, Zarrett & Eccles, 2007; Marsh et al., 2005; Skaalvik, 1999).

Zu (3): Im Erwartungs-x-Wert-Modell werden Bildungsaspirationen eng an die wahrgenommenen Erfolgserwartungen und den beigemessenen Wert eines solchen Bildungsabschlusses geknüpft (z. B. Wigfield & Cambria, 2010). Somit dienen die Bildungsaspirationen als wichtige Verbindung zu den schul- und bildungsbezogenen Entscheidungen, die sich wiederum auf die Schulleistung und den Erwerb von Bil-

dungszertifikaten auswirken. Die höheren Bildungsaspirationen von Schülerinnen und Schülern mit Migrationshintergrund werden auch über Variationen im Sozialkapital erklärt (Becker, 2010). Hieraus wird abgeleitet, dass Bildungsaspirationen ebenfalls die Einflüsse der kulturellen Ebene mediieren.

Darüber hinaus muss zwischen idealistischen (ohne Berücksichtigung des Kompetenzniveaus) und realistischen (unter Berücksichtigung des Kompetenzniveaus) Aspirationen unterschieden werden. Beide werden direkt durch das soziale und kulturelle Kapital prädiziert (Becker, 2010). Die Differenzierung zwischen übersteigerten idealistischen Bildungsaspirationen und angemessenen realistischen Aspirationen dient der Erklärung einer fehlenden Umsetzung positiver Aspirationen in gute Schulleistungen. Ebenso hat die Diskrepanz zwischen beiden Aspirationsarten einen Erklärungswert für „Aufsteiger" im Bildungssystem. Ist der Unterschied zwischen den Leistungen und den Aspirationen nur so groß, dass das anvisierte Bildungsziel auch (strukturell bedingt, etwa durch Zugangsbarrieren wie Notendurchschnitte) erreicht werden kann, wird dies aus zieltheoretischer Perspektive eine wichtige Voraussetzung für angemessene bildungsbezogene Entscheidungen sein (Kruglanski et al., 2002). Positive Auswirkungen auf die Schulleistung als wichtigste Voraussetzung für den Bildungsgangwechsel werden erwartet (vgl. Eccles, 2005; Eccles & Wigfield, 2002).

5 Stichprobe und Instrumente

Im Projekt wurden die Daten der PISA-I-Studie (Datensatz: Schülerdaten, 15-Jährige) aus dem Jahr 2003 (Ramm et al., 2006) als Grundlage herangezogen, um im Querschnitt Prozesse der individuellen Vorhersage schulischen Erfolgs bei Migrantenschülerinnen und -schülern zu identifizieren. Die Stichprobeneigenschaften, besonderen Merkmale und Instrumente der PISA-Studie sind gut dokumentiert und müssen hier nicht gesondert dargestellt werden.

5.1 Stichprobe

Bei der Längsschnittstudie an bayerischen Mittelschulen handelt es sich um eine Fragebogenstudie, deren erster Messzeitpunkt im Winter 2012/13 und dessen zweiter Messzeitpunkt im Februar bis April 2014 stattgefunden hat. Die Termine wurden so gewählt, dass Informationen über die Schülerinnen und Schüler vorlagen, als sie a) noch keine Hinweise darauf haben konnten, mit welchen schulischen Leistungen sie welchen Schulzweig (Regel- oder Mittlere-Reife-Zweig)[2] erreichen werden, und als sie b) den Übergang in einen der beiden Schulzweige bereits vollzogen haben. Hierdurch konnten Voraussetzungen für den faktischen Wechsel auf einen der beiden

2 An bayerischen Mittelschulen (ehemals Hauptschulen) ist der Erwerb des Hauptschul-, des qualifizierten Hauptschulabschlusses und an ausgewählten Schulen des Mittlere-Reife-Abschlusses möglich. Regelklassen gehen bis zur neunten Klasse, Mittlere-Reife-Klassen bis zur zehnten Jahrgangsstufe.

Bildungsgänge empirisch nachvollzogen und so das theoretische Modell im Längsschnitt geprüft werden.

Insgesamt wurden zu beiden Messzeitpunkten der Längsschnittstudie 818 Schülerinnen und Schüler (43,3 Prozent weiblich) an 34 bayerischen Mittelschulen mittels Fragebogen mit zumeist standardisierten Indikatoren befragt. Hiervon weisen 586 Schülerinnen und Schüler einen Migrationshintergrund auf (44,7 Prozent weiblich), die zu großen Teilen türkischer (31,4 Prozent) oder russischer (inkl. Teilrepubliken) (16,9 Prozent) Herkunft sind bzw. aus den Nachfolgestaaten des ehemaligen Jugoslawiens stammen (10,1 Prozent). Die verbleibenden Herkunftsländer und -regionen sind zu gering besetzt, als dass eine Aufschlüsselung nach Herkunft sinnvoll ist. Knapp jede fünfte Schülerin bzw. jeder fünfte Schüler entstammt der ersten (21,3 Prozent) bzw. dritten Migrationsgeneration (23,5 Prozent), den Großteil machen aber Schülerinnen und Schüler der sogenannten zweiten Generation aus (45,4 Prozent), bei der nicht die Eltern, wohl aber die Kinder selbst in Deutschland geboren wurden.

5.2 Instrumente

Wechsel in R- oder M-Klasse (abhängige Variable; MZP 2)
In der Gesamtstichprobe sind ab der siebten Klasse 592 Schülerinnen und Schüler in die Regelklasse (72,4 Prozent) und 226 in den M-Zweig (27,6 Prozent) gewechselt. Die Übertrittsquote in den M-Zweig ist bei den Migrantenschülerinnen und -schülern vergleichbar hoch. Hier haben 25,1 Prozent aller Migrantinnen und Migranten den Sprung in den M-Zweig geschafft (*N = 147).* Die verbleibenden 439 Schülerinnen und Schüler mit Migrationshintergrund sind im Regelangebot, das zum Hauptschulabschluss führt, verblieben (74,9 Prozent).

Mathematikleistung (abhängige Variable; MZP 2)
Die mathematischen Kompetenzen der Schülerinnen und Schüler wurden mittels DEMAT 6+ (Götz, Lingel & Schneider, 2013a; Götz, Lingel & Schneider, 2013b) erhoben, und mit den Entwicklern des DEMAT wurden mögliche Bodeneffekte in den Leistungsverteilungen vorab diskutiert. Diese Effekte sind letztlich nicht aufgetreten, weder bei der Gesamtstichprobe (*R = 1–30; M = 13.17; SD = 5.77)* noch bei der Substichprobe der Schülerinnen und Schüler mit Migrationshintergrund *(R = 1–30; M = 12.58; SD = 5.58).*

Sozioökonomischer Status nach HISEI (unabhängige Variable; MZP 1)
Die Berufsangaben der Schülerinnen und Schüler über ihre Eltern wurden anhand der ISCO-08-Klassifizierung codiert und liegen als Familien-HISEI vor. Somit ist die Erfassung des SES der Familie mit den Daten aus der PISA-Studie vergleichbar.

Intrinsische Motivation Mathematik (unabhängige Variable; MZP 1)
In Anlehnung an die Items zur Erfassung der intrinsischen Motivation im Fach Mathematik aus der PISA-Studie wurden den Schülerinnen und Schülern vier Items

vorgelegt, die somit in beiden Teilstudien verfügbar waren (z. B. „Mich interessiert das, was ich in Mathe lerne“; *Cronbachs* α = 0.84; M = 2.90; SD = 0.82).

Selbstwirksamkeitserwartung Mathematik (unabhängige Variable; MZP 1)
Mit ebenfalls vier Items aus der PISA-Studie wurde die Selbstwirksamkeitserwartung der Schülerinnen und Schüler für das Fach Mathematik erfasst (z. B. „In Mathe bin ich sicher, dass ich auch den schwierigen Stoff verstehen kann“; *Cronbachs* α = 0.86; *M = 2.91; SD = 0.75).*

Bildungsaspirationen (unabhängige Variable; MZP 1)
Die Bildungsaspirationen wurden in KuBiS wie in PISA auch über den Bildungsabschluss erfasst, den die Schülerinnen und Schüler zu erreichen glauben. Unterschiede ergeben sich zwischen der Gesamtstichprobe und der Teilgruppe der Migrantenschülerinnen und -schüler kaum (vgl. Tabelle 1).

Tabelle 1: Relative Verteilung erwarteter Bildungsabschlüsse bei der Gesamtstichprobe und der Teilstichprobe der Schülerinnen und Schüler mit Migrationshintergrund in Prozent

	Hauptschulabschluss	Qualifizierter Hauptschulabschluss	Mittlere Reife/M-Zweig	Lehre/ Berufsabschluss	Abitur/ Hochschulreife
Gesamt	17,7	10,6	60,8	4,0	6,9
Migrantinnen und Migranten	18,8	10,7	58,6	4,2	7,6

Den Schwerpunkt bildete zum ersten MZP die Erwartung, die mittlere Reife bzw. den M-Zweig erfolgreich zu absolvieren. Wie die faktischen Übertrittsquoten in den M-Zweig zeigen, überschätzen die Befragten zum ersten Messzeitpunkt ihre Bildungschancen in leichtem Maße.

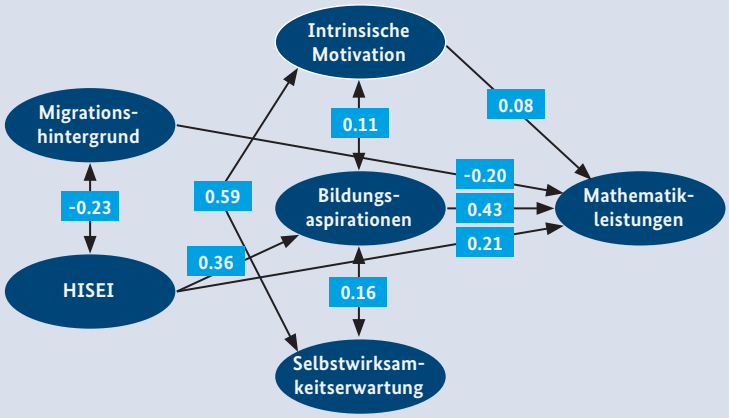
6 Ergebnisse

Das theoretische Modell wurde sowohl den Daten aus der PISA-Studie als auch den KuBiS-Daten selbst zugrunde gelegt. Durch die vergleichbaren Erhebungsinstrumente kann wenigstens geprüft werden, ob sich ähnliche Zusammenhangsmuster finden lassen. Dabei ist jedoch einschränkend festzustellen, dass die PISA-Daten quer- und die KuBiS-Daten längsschnittlich angelegt sind. Während also die PISA-Modelle interindividuelle Rangfolgen durch Persönlichkeitsmerkmale vorhersagen, erlauben die KuBiS-Daten prinzipiell, Veränderungsvarianzen zu prädizieren. Damit wären aber vergleichende Schlussfolgerungen nicht möglich, sodass bei den empirischen Modellen aus den KuBiS-Daten auf die Kontrolle der Eingangswerte für die abhängigen Variablen verzichtet wurde.

Insgesamt wurden vier Modelle auf Passung zum theoretischen Modell geprüft. Dies sind jeweils zwei Modelle aus den PISA- und den KuBiS-Daten, die separat für alle Probanden und nur für jene mit Migrationshintergrund bestimmt wurden.

Zunächst zeigt der Blick auf das Modell für alle Schülerinnen und Schüler der PISA-Studie, dass der Migrationshintergrund (ß = −0.20) und der HISEI (ß = 0.21) die Mathematikleistungen vorhersagen. Schülerinnen und Schüler mit Migrationshintergrund und Schülerinnen und Schüler mit geringerem SES erzielen im Durchschnitt schlechtere Leistungen in Mathematik als ihr jeweiliges Pendant (vgl. Abbildung 2).

Abb. 2: Vorhersage der Mathematikleistungen durch individuelle Voraussetzungen (alle Schülerinnen und Schüler, PISA-I 2003; dargestellt sind nur Pfade mit p < 0.05)



PISA – Gesamtsample, RMSEA = 0.037, SRMR = 0.022

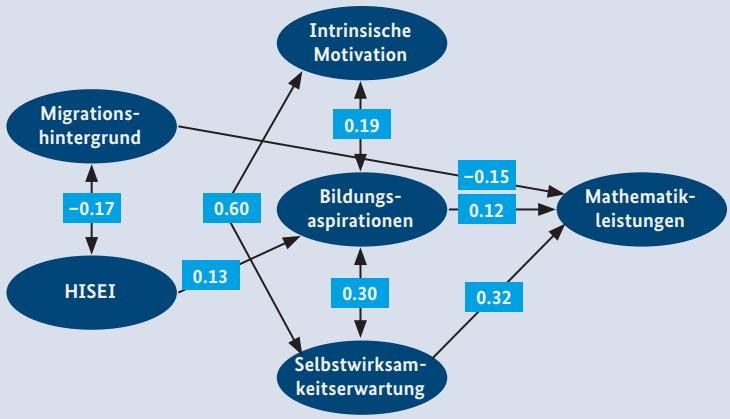
Darüber hinaus zeigt von den Individualmerkmalen lediglich die Bildungsaspiration der Schülerinnen und Schüler einen intensiven Zusammenhang zu den Matheleistungen. Schülerinnen und Schüler mit hohen Aspirationen schneiden in den Kompetenztests für das Fach Mathematik deutlich besser ab als solche mit geringen Erwartungen (ß = 0.43). Da die Aspirationen selbst in engem Zusammenhang zum HISEI stehen (ß = 0.36), repliziert sich hier der bereits bekannte Befund, wonach Heranwachsende aus besser gestellten Elternhäusern höhere Aspirationen haben.

Gegenüber diesem Einfluss der Bildungsaspirationen bleibt jener der intrinsischen Lernmotivation deutlich zurück (ß = 0.08), die Selbstwirksamkeitserwartung hat keinen Einfluss auf die Matheleistungen.

Ein hiervon abweichendes Bild zeigt sich im Modell für die anfallende Stichprobe der Schülerinnen und Schüler aus der KuBiS-Studie (vgl. Abbildung 3).

Zum einen haben der HISEI keinen und der Migrationshintergrund einen deutlich schwächeren Einfluss (ß = −0.15) auf die durch den DEMAT gemessenen Matheleistungen. Zum anderen ist der direkte Einfluss der Bildungsaspirationen geringer als im PISA-Modell (ß = 0.12) und wird deutlich überflügelt von jenem der Selbstwirksamkeitserwartung (ß = 0.32). Schülerinnen und Schüler, die zum ersten Messzeitpunkt eine hohe Selbstwirksamkeitserwartung äußern, weisen ein Jahr später höhere Matheleistungen auf als Gleichaltrige mit geringer Selbstwirksamkeitserwartung.

Abb. 3: Vorhersage der Mathematikleistungen durch individuelle Voraussetzungen (alle Schülerinnen und Schüler, KuBiS; dargestellt sind nur Pfade mit p < 0.05)



KUBIS – Gesamtsample RMSEA = 0.041 SRMR = 0.029

Vergleichbar ist hingegen das Muster, wonach der HISEI signifikant mit den Bildungsaspirationen korreliert ist (β = 0.13) und jener wiederum die Matheleistungen vorhersagt.

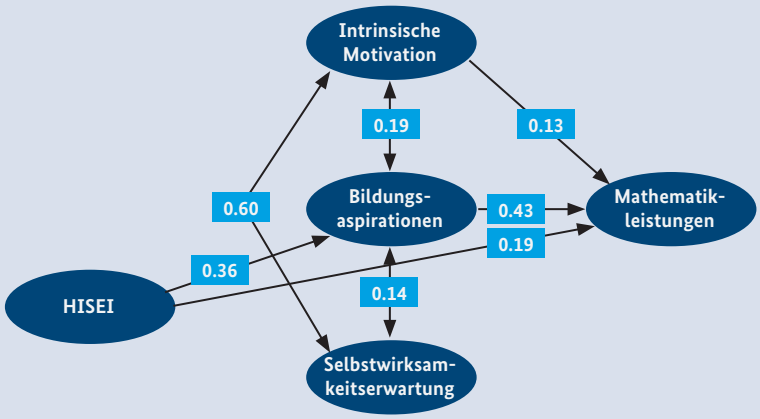
Für Schülerinnen und Schüler mit Migrationshintergrund wurden sodann gesonderte Modelle gerechnet, um die Binnenvarianz dieser Zielgruppe bestimmen zu können. Durch die reduzierte Varianz sind a priori zwar weniger signifikante Befunde erwartbar, dennoch gleichen die Muster jenen der Gesamtstichprobe. Dies gilt sowohl für die PISA- als auch für die KuBiS-Daten.

Wiederum sind es die Bildungsaspirationen, die die Mathekompetenzen der Schülerinnen und Schüler prädizieren (β = 0.43) und ihrerseits im Zusammenhang zum HISEI stehen (β = 0.36) (vgl. Abbildung 4). Das heißt, dass auch innerhalb der Gruppe der Schülerinnen und Schüler mit Migrationshintergrund eine Abhängigkeit des SES besteht. Migrantenschülerinnen und -schüler mit besser situiertem Elternhaus weisen (zumindest in Mathematik) einen größeren Schulerfolg auf als andere Schülerinnen und Schüler mit Migrationshintergrund.

Darüber hinaus zeigt der HISEI auch einen direkten positiven Pfad zu den Mathematikkompetenzen (β = 0.19). Interessant ist, dass sich der Zusammenhang zwischen der intrinsischen Lernmotivation und den Matheleistungen im Vergleich zum Gesamtmodell leicht erhöht (von β = 0.08 auf β = 0.13).

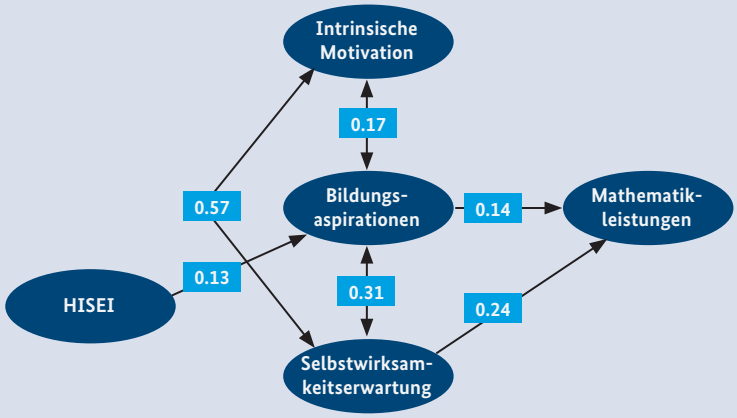
Inwieweit es sich hierbei um einen systematischen Unterschied handelt, lässt sich anhand der Daten nicht prüfen, zumal sich in den KuBiS-Daten der Zusammenhang des anderen Individualmerkmals nicht systematisch erhöht (vgl. Abbildung 5). Bei der Substichprobe der Migrantinnen und Migranten verringert sich der Zusammenhang zwischen der Selbstwirksamkeitserwartung und den DEMAT-Leistungen auf β = 0.24 (im Vergleich zu β = 0.32 in der Gesamtstichprobe).

Abb. 4: Vorhersage der Mathematikleistungen durch individuelle Voraussetzungen (nur Schülerinnen und Schüler mit Migrationshintergrund, PISA-I 2003; dargestellt sind nur Pfade mit p < 0.05)



PISA – Nur MigH, RMSEA = 0.001, SRMR = 0.027

Abb. 5: Vorhersage der Mathematikleistungen durch individuelle Voraussetzungen (nur Schülerinnen und Schüler mit Migrationshintergrund, KuBiS; dargestellt sind nur Pfade mit p < 0.05)



KUBIS – Nur MigH, RMSEA = 0.039, SRMR = 0.028

Im Kern erhalten bleibt jedoch auch in diesem Modell die Aussage, dass der HISEI in schwachem Zusammenhang zu den Bildungsaspirationen der Schülerinnen und Schüler steht ($\beta = 0.13$) und diese wiederum mit den Mathematikleistungen korrelieren ($\beta = 0.24$). Auch lässt sich bei diesem Teilsample kein signifikanter Pfad zwischen der intrinsischen Motivation und den Matheleistungen modellieren.

Insgesamt ist bei der Betrachtung der Modelle zu berücksichtigen, dass sich die Komposition der Stichproben einerseits sowie die Messung der abhängigen Variab-

len andererseits unterscheiden. Daher kann bei einer Schlussfolgerung und abschließenden Diskussion weniger der Vergleich von PISA und KuBiS im Mittelpunkt stehen, sondern vielmehr der Vergleich innerhalb der Studien.

7 Diskussion

Der vorliegende Beitrag berichtet erste Ergebnisse aus dem Projekt „KuBiS – Kompetenzunterschiede und Bildungsgangwechsel bei Schülern mit Migrationshintergrund". Das Ziel der Studie ist es, den Beitrag individueller Merkmale für den Bildungserfolg von Migrantenschülerinnen und -schülern zu bestimmen und im Vergleich zum Merkmal Migrationshintergrund selbst sowie dem assoziierten SES zu betrachten. Als Referenzrahmen werden hierzu auch Daten aus der PISA-Studie von 2003 herangezogen. Basierend auf dem Erwartungs-x-Wert-Modell wird erwartet, dass die Selbstwirksamkeitserwartung, die Bildungsaspirationen sowie die intrinsische Lernmotivation einen substanziellen Beitrag zu den Matheleistungen der Schülerinnen und Schüler leisten.

Die berechneten Pfadmodelle zeigen zum einen Unterschiede zwischen beiden Datenquellen an. Während in den PISA-Daten familiale Hintergrundmerkmale einen bedeutsamen Einfluss haben, reduziert sich dieser in der kleineren und fokussierten Mittelschulstichprobe aus Bayern. Auch ist in den PISA-Daten ein Zusammenhang der intrinsischen Lernmotivation zu verzeichnen, in KuBiS ist es hingegen die Selbstwirksamkeitserwartung, die mit den Matheleistungen korreliert ist. Gründe für diese Unterschiede liegen mutmaßlich zum einen in der unterschiedlichen Grundgesamtheit, zum anderen aber auch in der unterschiedlichen Erfassung der Mathematikleistungen und schließlich im Vergleich eines Quer- mit einem Längsschnittmodell.

Innerhalb der Teilstudien lassen sich aber ähnliche Muster finden, die darauf hindeuten, dass neben den sozioökonomischen Merkmalen für Migrantenschülerinnen und -schüler auch individuelle Voraussetzungen eine Rolle spielen. Ob es sich hier je nach Studie um das eine oder andere lernrelevante Merkmal handelt, ist dabei eher sekundär. Interessant ist vielmehr, dass sich auch innerhalb der Population der Migrantenschülerinnen und -schüler Variationen finden lassen, die den Schulerfolg jenseits struktureller Merkmale erklären.

Wenngleich diese Modelle erst eine grundlegende Annäherung darstellen und im Rahmen eines Beitrags kaum weiter gehende Befunde darstellbar sind, so spricht einiges dafür, den individuellen Eigenschaften von Kindern und Jugendlichen mit Migrationshintergrund in Zukunft mehr Gewicht beizumessen, um vor allem das Zusammenspiel struktureller und individueller Merkmale besser zu verstehen.

Das Erwartungs-x-Wert-Modell kann hierzu als erste Heuristik dienen, ist aber innerhalb der Variablenklassen noch ebenso unterspezifiziert wie hinsichtlich der Anwendbarkeit auf Schülerinnen und Schüler mit Migrationshintergrund. Da der Fokus auf Individualmerkmale bei Migrantenschülerinnen und -schülern aber noch vergleichsweise neu ist, werden theoretische Erweiterungen und empirische Prüfungen noch folgen.

Literaturverzeichnis

Aronson, J., Steele, C. M., Elliot, A. J. & Dweck, C. S. (2005). Stereotypes and the Fragility of Academic Competence, Motivation, and Self-Concept. In *Handbook of competence and motivation* (S. 436–456). New York: Guilford Publications.

Autorengruppe Bildungsberichterstattung (2014). *Bildung in Deutschland 2014. Ein indikatorengestützter Bericht mit einer Analyse zur Bildung von Menschen mit Behinderungen.* Bielefeld: W. Bertelsmann Verlag.

Bandura, A. (2001). Social Cognitive Theory: An Agentic Perspective. *Annual Review Psychology, 52,* 1–26.

Baumert, J. & Schümer, G. (2002). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb im nationalen Vergleich. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, G. Schiemer, P. Stenat, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000. Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 159–200). Opladen: Leske + Budrich.

Becker, B. (2010). *Bildungsaspirationen von Migranten: Determinanten und Umsetzung in Bildungsergebnisse. Arbeitspapiere – Mannheimer Zentrum für Europäische Sozialforschung, 137.* Mannheim.

Bos, W., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., Lankes, E.-M. et al. (2007). *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.* Münster: Waxmann.

Boudon, R. (1974). *Education, opportunity, and social inequality: Changing prospects in western society.* New York, NY: Wiley.

Bourdieu, P. (1992): Ökonomisches Kapital – Soziales Kapital – Kulturelles Kapital. In P. Bourdieu (Hrsg.), *Die verborgenen Mechanismen der Macht* (S. 49–79). Hamburg: VSA: Verlag.

Denissen, J. J. A., Zarrett, N. R. & Eccles, J. S. (2007). I Like to Do It, I'm Able, and I Know I Am: Longitudinal Couplings Between Domain-Specific Achievement, Self-Concept, and Interest. *Child Development, 78 (2),* 430–447.

Diefenbach, H. (2002). Bildungsbeteiligung und Berufseinmündung von Kindern und Jugendlichen aus Migrantenfamilien. Eine Fortschreibung der Daten des Sozio-Ökonomischen Panels (SOEP). In H. Diefenbach, G. Renner & B. Schulte (Hrsg.), *Migration und die europäische Integration: Herausforderungen für die Kinder- und Jugendhilfe* (S. 9–70). München: DJI.

Ditton, H., Krüsken, J. & Schauenberg, M. (2005). Bildungsungleichheit – der Beitrag von Familie und Schule. *Zeitschrift für Erziehungswissenschaft, 8,* 285–304.

Eccles, J. S. (2005). Subjective Task Value and the Eccles et al. Model of Achievement-Related Choices. In A. J. Elliot & C. S. Dweck (Hrsg.), *Handbook of competence and motivation* (S. 105–121). New York, NY: Guilford Publications.

Eccles, J. S. & Wigfield, A. (2002). Motivational Beliefs, Values, And Goals. *Annual Review Psychology, 53,* 109–132.

Fresow, M., Rettich, P., Gniewosz, B. & Reinders, H. (2012). Individuelle Bedingungen für erfolgreiche Bildungskarrieren bei Schülerinnen und Schülern mit Migrationshintergrund. *Diskurs Kindheits- und Jugendforschung, 7 (4),* 473–480.

Götz, L., Lingel, K. & Schneider, W. (2013a). Diagnostik mathematischer Kompetenzen in der Sekundarstufe I am Beispiel der Deutschen Mathematiktests für die fünften und sechsten Klassen (DEMAT 5+, DEMAT 6+). In M. Hasselhorn, A. Heinze, W. Schneider & U. Trautwein (Hrsg.), *Diagnostik mathematischer Kompetenzen. Tests und Trends, N. F. Bd. 11* (S. 241–260). Göttingen: Hogrefe.

Götz, L., Lingel, K. & Schneider, W. (2013b). DEMAT 6+. *Deutscher Mathematiktest für sechste Klassen.* Göttingen: Hogrefe.

Gräsel, C. & Gniewosz, B. (2015). Überblick Lehr-Lernforschung. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Lehrbuch Empirische Bildungsforschung. Bd. 2 Gegenstandsbereiche* (S. 19–24). Wiesbaden: SpringerVS.

Hartmann, R. M. & McElvany, N. (2013). Domänenspezifische Motivation und Mathematikleistungen in der Grundschule vor dem Hintergrund kultureller und sprachlicher Diversität. *Zeitschrift für Grundschulforschung, 6,* 142–157.

Helmke, A., Hosenfeld, I., Schrader, F.-W. & Wagner, W. (2002). Sozialer und sprachlicher Hintergrund. In A. Helmke & R. S. Jäger (Hrsg.), *Das Projekt MARKUS. Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext* (S. 71–153). Landau: Verlag Empirische Pädagogik.

Hillesheim, S. (2009). *Elternarbeit in der Schule. Ein Vergleich der Elternarbeit mit Migranteneltern an Halbtags- und Ganztagsschulen in Bayern. Schriftenreihe Empirische Bildungsforschung, Bd. 13.* Universität Würzburg.

Hofer, M., Reinders, H. & Fries, S. (2010). Wie sich Werte ändern. Ein zieltheoretischer Vorschlag zur Erklärung individuellen und gesellschaftlichen Wertewandels. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42 (1),* 26–38.

Hofer, M., Schmid, S., Fries, S., Dietz, F., Clausen, M. & Reinders, H. (2007). Individual values, motivational conflicts, and learning for school. *Learning and Instruction, 17 (1),* 17–28.

Kao, G. & Tienda, M. (1995). Optimism and Achievement: The Educational Performance of Immigrant Youth. *Social Science Quarterly, 76 (01),* 1–19.

Kao, G. & Tienda, M. (1998). Educational Aspirations of Minority Youth. *American Journal of Chicago, 106, 3,* 349–384.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M. et al. (2010). *PISA 2009. Bilanz nach einem Jahrzehnt.* Münster: Waxmann.

Konsortium Bildungsberichterstattung (2006). *Bildung in Deutschland: Ein indikatorgestützter Bericht mit einer Analyse zu Bildung und Migration.* Bielefeld: W. Bertelsmann.

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y. & Sleeth-Keppler, D. (2002). A theory of goal systems. J*ournal of Applied Psychology, 34,* 331–378.

Liem, A. D., Lau, S. & Nie, Y. (2008). The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. *Contemporary Educational Psychology, 33,* 486–512.

Marsh, H. W., Köller, O., Trautwein, U., Lüdtke, O. & Baumert, J. (2005). Academic Self-Concept, Interest, Grades, and Standardized Test Scores: Reciprocal Effects Models of Causal Ordering. *Child Development, 76 (2),* 397–416.

Organisation for Economic Co-operation and Development (2006). *Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003.* Paris: OECD.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J. & Schiefele, U. (Hrsg.) (2006). *PISA 2003. Dokumentation der Erhebungsinstrumente.* Münster: Waxmann.

Rettich, P., Fresow, M., Gniewosz, B. & Reinders, H. (2012). Das Projekt KuBiS. Kompetenzunterschiede und Bildungsgangwechsel bei Schülerinnen und Schülern mit Migrationshintergrund. *Zeitschrift für Soziologie der Erziehung und Sozialisationsforschung, 32 (03),* 330 f.

Rheinberg, F. (2004). *Motivation.* Stuttgart: Kohlhammer.

Schmid, S., Hofer, M., Dietz, F., Reinders, H. & Fries, S. (2005). Value orientations and action conflicts in students' everyday life: An interview study. *European Journal of Psychology of Education, 20,* 259–274.

Schofield, J. W., Alexander, K., Bangs, R. & Schauenburg, B. (2006). *Migrationshintergrund, Minderheitenzugehörigkeit und Bildungserfolg. Forschungsergebnisse der pädagogischen, Entwicklungs- und Sozialpsychologie. AKI-Forschungsbilanz 5.* Berlin: BMBF.

Schulze, E. & Soja, E.-M. (2006). Verschlungene Bildungspfade. Über Bildungskarrieren von Jugendlichen mit Migrationshintergrund. In G. Auernheimer (Hrsg.), *Schieflagen im Bildungssystem. Die Benachteiligung der Migrantenkinder* (S. 193–205). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schunk, D. H. & Mullen, C. A. (2012). Self-Efficacy as an Engaged Learner. In S. L. Christenson, A. L. Reschly & C. Wylie (Hrsg.), *Handbook of Research on Student Engagement* (S. 219–235). Boston: Springer.

Schunk, D. H. & Pajares, F. (2005). Competence Perceptions and Academic Functioning. In A. J. Elliot & C. S. Dweck (Hrsg.), *Handbook of competence and motivation* (S. 85–104). New York, NY: Guilford Publications.

Schwartz, S. H. (1996). Value priorities and behavior: Applying a theory of integrated value systems. In C. Seligman, J. M. Olson & M. P. Zanna (Hrsg.), *The psychology of values* (S. 1–24). Mahwah, NJ: Lawrence Erlbaum Associates.

Skaalvik, E. M. (1999). Relations Among Achievement, Self-Concept, and Motivation in Mathematics and Language Arts: a Longitudinal Study. *Journal of Experimental Education, 67,* 135–149.

Staatsinstitut für Schulqualität und Bildungsforschung (2009). *Bildungsbericht Bayern 2009.* München: Kastner.

Stanat, P. & Christensen, G. (2006). *Schulerfolg von Jugendlichen mit Migrationshintergrund im internationalen Vergleich. Eine Analyse von Voraussetzungen und Erträgen schulischen Lernens im Rahmen von PISA 2003.* Berlin: BMBF.

Stanat, P. & Edele, A. (2015). Zuwanderung und soziale Ungleichheit. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Lehrbuch Empirische Bildungsforschung. Bd. 2 Gegenstandsbereiche* (S. 215–228). Wiesbaden: Springer.

Stanat, P., Segeritz, M. & Christensen, G. (2010). Schulbezogene Motivation und Aspiration von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos,

E. Klieme & O. Köller (Hrsg.), *Schulische Lerngelegenheiten und Kompetenzentwicklung. Festschrift für Jürgen Baumert* (S. 31–57). Münster: Waxmann.

Steinbach, A. & Nauck, B. (2004). Intergenerationale Transmission von kulturellem Kapital in Migrantenfamilien. Zur Erklärung von ethnischen Unterschieden im deutschen Bildungssystem. *Zeitschrift für Erziehungswissenschaft, 7, 1,* 20–32.

Walter, O. (2009). Herkunftsassoziierte Disparitäten im Lesen, der Mathematik und den Naturwissenschaften: ein Vergleich zwischen PISA 2000, PISA 2003 und PISA 2006. In M. Prenzel & J. Baumert (Hrsg.), *Vertiefende Analysen zu PISA 2006.* (S. 149–168). Wiesbaden: VS-Verlag.

Walter, O. & Taskinen, P. (2007). Kompetenzen und bildungsrelevante Einstellungen von Jugendlichen mit Migrationshintergrund in Deutschland: ein Vergleich mit ausgewählten OECD-Staaten. In M. Prenzel, C. Artel, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 337–366). Münster: Waxmann.

Watermann, R. & Baumert, J. (2006). Entwicklung eines Strukturmodells zum Zusammenhang zwischen sozialer Herkunft und fachlichen und überfachlichen Kompetenzen. In J. Baumert, P. Stanat & R. Watermann (Hrsg.), *Herkunftsbedingte Disparitäten im Bildungssystem* (S. 61–94). Opladen: VS Verlag.

Wigfield, A. & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review, 30 (1),* 1–35.

Zuffianò, A., Alessandri, G., Gerbino, M., Luengo Kanacri, B. P., Di Giunta, L., Milioni, M. & Caprara, G. V. (2013). Academic achievement: The unique contribution of self-efficacy beliefs in self-regulated learning beyond intelligence, personality traits, and self-esteem. *Learning and Individual Differences, 23,* 158–162.

Nele McElvany, Franziska Schwabe,
Miriam M. Gebauer, Wilfried Bos

Prüfung der Testfairness ausgewählter Large-Scale-Assessments für zentrale Schülersubpopulationen

1 Einleitung

Der Umgang mit heterogenen Schülerschaften zählt zu einer der bedeutenden Aufgabenstellungen unseres Bildungssystems (z. B. Tillmann, 2008). Groß angelegte Schulleistungsuntersuchungen erfüllen in diesem Kontext die Aufgabe des Systemmonitorings mit Blick auf die Bildungsqualität im Sinne von Kompetenzerwerb. Diese Studien fokussieren hierbei die Gesamtheit der Schülerinnen und Schüler. Daraus ergibt sich vor dem Hintergrund der Unterschiedlichkeit innerhalb der Schülerschaft eine mögliche Unfairness der eingesetzten Verfahren zur Kompetenzmessung. Eine grundlegende Voraussetzung für die Fairness von Testaufgaben ist, dass keine Variable neben der eigentlich zu messenden Zielvariable die Testresultate beeinflusst (Camilli, 2006). Das Projekt „Prüfung der Testfairness ausgewählter Large-Scale-Assessments für zentrale Schülersubpopulationen" verfolgt das Ziel, mithilfe von Differential-Item-Functioning-Analysen (DIF) (Camilli, 2006) die Fairness der in Large-Scale-Assessments eingesetzten Tests für ausgewählte Schülersubgruppen im Bereich Lesen zu analysieren. Analysiert werden Testaufgaben aus der Internationalen Grundschul-Lese-Untersuchung (IGLU) und aus dem Programme for International Student Assessment (PISA). Aus dem Projekt resultieren Erkenntnisse, die dazu beitragen können, zukünftige nationale und internationale Large-Scale-Assessments unter Aspekten der Fairness zu optimieren. Außerdem können auf Basis der identifizierten spezifischen Stärken und Schwächen der verschiedenen Schülergruppen in der Bearbeitung von unterschiedlichen Aufgabenstrukturen und -inhalten Hinweise zur gezielten Unterstützung im Bereich Lesen für die Bildungspraxis bereitgestellt werden.

2 Theoretischer Hintergrund des Projekts

Grundschülerinnen und -schüler unterscheiden sich schon am Anfang ihrer schulischen Laufbahn deutlich. Heterogenität herrscht einerseits in den gezeigten Leistungen in den Schulfächern und Kompetenzen und andererseits in individuellen, sozialen und institutionellen Merkmalen, die ihren Schulerfolg und ihre Leistungen determinieren (Decristan et al., 2014). Diese Unterschiede bleiben über die Schullauf-

bahn hinweg bestehen (Ehmke & Baumert, 2008). Die Heterogenität der Schülerschaft in Deutschland, aber auch in anderen Ländern lässt sich anhand verschiedener Kriterien und Merkmale beschreiben (international: Deville & Chalhoub-Deville, 2011; Chatterji, 2013; national: Beutel, Bos & Porsch, 2013). Schülerinnen und Schüler unterscheiden sich hinsichtlich individueller, sozialer, ethnischer und kultureller Faktoren. Erweitert werden diese Dimensionen der Heterogenität um schulformspezifische, also institutionelle Bedingungen wie etwa vorgegebene Curricula (vgl. Systematisierung von Wenning, 2007). Für den zentralen Kompetenzbereich Lesen können vier Kategorien von Schülersubgruppen festgestellt werden, deren Betrachtung im Rahmen von Kompetenzmessungen eine besonders wichtige Aufgabe ist. Diese Merkmale aufseiten der Schülerinnen und Schüler sind in Tabelle 1 dargestellt.

Tabelle 1: Relevante Schülergruppen und Unterscheidungsmerkmale

Schülergruppen	Art des Unterschieds
Mädchen und Jungen	Geschlechtsspezifisch
Kinder und Jugendliche aus Familien mit unterschiedlichem sozioökonomischen Status	Demografisch
Kinder und Jugendliche aus Familien mit und ohne Migrationsgeschichte	Sprachlich
Kinder und Jugendliche unterschiedlicher Schulformen	Schulformspezifisch

Sollen beispielsweise die Lesekompetenzen aller Schülerinnen und Schüler eines Schulsystems oder auch vergleichend aus mehreren Staaten mit Testaufgaben gemessen werden, so kann und muss diese Leistungsmessung unter fairen Bedingungen für alle Subgruppen von Lernenden erfolgen (z. B. Xi, 2010). Unfairness kann in diesem Kontext durch Unterschiede in den Bedingungen oder Voraussetzungen, die das Testresultat beeinflussen, zwischen verschiedenen Personen(gruppen) verursacht werden. Ein internationaler Rahmen für die Untersuchung von Testfairness sind die Standards for Educational and Psychological Testing (AERA, APA & NCME, 1999; 2015). Eine grundlegende Anforderung an Testmaterialien und -prozeduren im Sinne von Testfairness ist, dass die Unverzerrtheit der Messung der Ausprägung in der Zielvariable gegeben ist. Folgendes Beispiel erläutert eine mögliche Verzerrung anhand einer hypothetischen Mathematikaufgabe (Beispiel entnommen aus Schwabe, 2014).

Abbildung 1 zeigt eine Textaufgabe in Mathematiktests, die das Auftreten potenziell verzerrender Variablen, in diesem Fall die sprachlichen Anforderungen, darstellt: Die Fähigkeit zur Lösung der mathematischen Fragestellung (hier Addition) ist die Zielvariable. Gleichzeitig erfordert die Aufgabenstellung das Erlesen der relevanten Informationen aus einem Textteil der Aufgabe. Unter der Annahme von Gruppenunterschieden innerhalb der Lesekompetenz oder aber auch dem Vorwissen über Orchester hat die lesestärkere Schülergruppe bzw. die Schülergruppe mit mehr Vorwissen eine größere Chance, die Mathematikaufgabe richtig zu beantwor-

Abb. 1: Hypothetische Mathematikaufgabe

Das Sinfonieorchester ist der übliche Klangkörper zur Wiedergabe von Orchesterwerken ab etwa der zweiten Hälfte des 18. Jahrhunderts. Die Streicher, deren Stimmen mehrfach besetzt sind, stehen den anderen, solistisch besetzten Instrumentengruppen gegenüber. Bei den Berliner Philharmonikern spielen z. B. 23 erste Violinen, 20 zweite Violinen, 16 Bratschen, 13 Celli und elf Kontrabässe.

Frage: Wie viele Augenpaare richten Streicher der Berliner Philharmoniker auf den Taktstock des Dirigenten?

ten. Die Ausprägung in der Zielvariablen bzw. in dem Testkonstrukt (Addition) wird durch Unterschiede zwischen den Gruppen in weiteren konstruktirrelevanten Variablen (Lesekompetenz, Vorwissen) beeinflusst. Bei Vorliegen einer solchen Verzerrung kann von einer falschen Antwort nicht zwingend der Rückschluss gezogen werden, dass die Kompetenz in der Zielvariable gering ist. Bei der Fairnessanalyse von Testaufgabe ist es also wichtig festzulegen, welche Variable(n) der Test messen soll, damit eine Unterscheidung zwischen tatsächlichen Differenzen in der Kompetenz und systematischer Benachteiligung einer oder mehrerer Gruppen möglich wird.

Zur Identifikation systematischer Verzerrungen können Analysen zu Differential Item Functioning (DIF; z. B. Angoff, 1993) eingesetzt werden. DIF tritt auf, wenn Menschen aus verschiedenen Gruppen mit denselben Persönlichkeitseigenschaften oder mit derselben Kompetenz eine unterschiedliche Wahrscheinlichkeit aufweisen, eine bestimmte Antwort auf eine Aufgabe in einem Fragebogen oder einem Test zu geben (für eine ähnliche Definition: Ferne & Rupp, 2007). Der Grund des DIF ist häufig, wie im oben beschriebenen Beispiel, ein (Kompetenz-)Unterschied in einem Bereich abseits der Variable, die mit der Testaufgabe gemessen werden soll. Aus mathematisch-statistischer Sicht existieren unterschiedliche Verfahren zur Bestimmung von DIF. Anhand ihrer Berechnungsgrundlage können die DIF-Analysemethoden in zwei Kategorien eingeteilt werden. Erstens gibt es Methoden, die auf beobachteten Testscores basieren. Zweitens beziehen sich Verfahren auf Modelle der Item Response Theory (IRT; Camilli, 2006). Ein systematischer Vergleich der Ansätze unter den Bedingungen im Large-Scale-Kontext steht noch aus.

Über ein Screening im Kontext von Test(un)fairness hinaus können DIF-Verfahren auch zur inhaltlichen Beschreibung von spezifischen Stärken und Schwächen bei Subgruppen herangezogen werden. In dieser Rahmung – d. h. im Sinne relativer Kompetenzunterschiede – werden in aktuellen Studien sowohl national als auch international DIF-Ergebnisse interpretiert (Haag, Heppt, Stanat, Kuhl & Pant, 2013; Koo, Becker & Kim, 2014; Li & Kolen, 2005; Schwippert, Bos & Lankes, 2004).

Schlussfolgernd stellen DIF-Analysen ein probates Mittel sowohl zur Analyse von Testfairness als auch zur Untersuchung von relativen Stärken und Schwächen im Kontext von Large-Scale-Assessments dar. Mit Blick auf die Heterogenität der Schülerschaft in Deutschland können DIF-Analysen folglich bedeutsame Informationen einerseits im Kontext von Leistungsmessung und andererseits auch für schulische Förderung generieren.

Um gezielte Hinweise auf Förderbedarfe bzw. spezifische Stärken bei Schülerinnen und Schülern im Lesen zu gewinnen und gleichzeitig auch Hinweise auf Optimierungspotenziale von Testaufgaben im Sinne von Testfairness auf Basis von Untersuchungen von Testaufgaben abzuleiten, ist es zentral, Merkmale der Aufgaben in den Blick zu nehmen. Im Bereich Lesen, d. h. bei Lesetestaufgaben, zählen die Aufgabenmerkmale

(1) **Antwortformat** (offene Antwortformate wie Constructed-Response-Aufgaben [CR] versus geschlossene Antwortformate wie Multiple-Choice-Aufgaben [MC]),
(2) **erforderlicher Leseprozess** (hierarchieniedrige Leseprozesse versus hierarchiehohe Leseprozesse),
(3) **Textgattung** (literarische Texte versus Sachtexte)

zu den zentralen Unterscheidungsmerkmalen der Aufgaben.

3 Übergeordnete Forschungsfragen des Projekts

Konkret werden die Fragen beantwortet, inwieweit DIF für zentrale Schülersubgruppen nach a) dem Geschlecht, b) dem sozioökonomischen Hintergrund, c) dem sprachlichen Hintergrund und d) der Schulformzugehörigkeit in Abhängigkeit von den Aufgabenmerkmalen i) Antwortformat, ii) Leseprozess und iii) Textgattung besteht. Außerdem wird im Rahmen einer Machbarkeitsstudie untersucht, welche DIF-Methoden im LSA-Kontext besonders geeignet sind.

4 Methodisches Vorgehen im Projekt

Zur Beantwortung der Forschungsfragen wurden sowohl die deutschen Teilstichproben der IGLU-Erhebung 2001 und 2011 als auch der PISA-2009-Studie verwendet. Folglich wurden jeweils mehrere Tausend Schülerinnen und Schüler der vierten Klassenstufe (IGLU) bzw. im Alter von 15 Jahren (PISA) in den Analysen berücksichtigt. Die Einteilung der Schülerinnen und Schüler in Subgruppen wurde auf Basis der Angaben in den Schüler- und/oder Elternfragebogen durchgeführt. Die Testaufgaben wurden anhand der Angaben, die in den Datensätzen verfügbar waren, kategorisiert. Zur Identifikation von DIF wurden als Analysestrategie insbesondere Item-Subset-Berechnungen mit GLM-Modellen (De Boeck et al., 2011) durchgeführt.

5 Ergebnisse des Projekts im Überblick

Zentrale Ergebnisse des Projekts sind, dass erwartungskonform bei Jungen, Kindern und Jugendlichen aus Familien mit geringerem sozioökonomischen Status, Schülerinnen und Schülern mit sprachlichem Migrationshintergrund und Kindern und Jugendlichen aus Schulformen ohne Abituroption eine spezifische Schwäche bei

der Bearbeitung von Lesetestaufgaben mit offenen Antwortformaten im Vergleich zu geschlossenen Antwortformaten gefunden wurde (vgl. für Details McElvany & Schwabe, 2013; Schwabe, McElvany & Trendtel, 2015a; Schwabe, McElvany & Trendtel, 2015b; Schwabe, McElvany, Trendtel, Gebauer & Bos, 2014). Das bedeutet z. B., dass Jungen im Vergleich zu Mädchen trotz eines ähnlichen Niveaus in der Lesekompetenz eine geringere Chance haben, eine Lesetestaufgabe richtig zu beantworten, wenn diese Aufgabe ein offenes Antwortformat hat. Dieser Zusammenhang gilt auch für die übrigen identifizierten Schülergruppen jeweils im Vergleich zu ihrer konträren Gruppe. Bei der Betrachtung relativer Stärken oder Schwächen zeigten sich hingegen in Abhängigkeit von den Aufgabenmerkmalen erforderlicher Leseprozess und Textgattung keine bedeutsamen oder nur geringe Unterschiede bei fähigkeitsgleichen Schülergruppen. Tabelle 2 gibt eine Übersicht über die gefundenen Effekte.

Tabelle 2: Übersicht über gefundene relative Schwächen in den empirischen Beiträgen nach Schüler- und Aufgabenmerkmalen

	Aufgabenmerkmal		
Schülermerkmal	**Offenes Antwortformat**	**Hierarchiehoher Leseprozess**	**Sachtext**
Migrationshintergrund (mit)	✓	X	X
Sozioökonomischer Status (schwach)	✓	n. u.	X
Geschlecht (männlich)	✓	n. u.	n. u.
Schulform (niedrig)	✓	n. u.	n. u.
Motivation (niedrig)	✓	n. u.	n. u.

Anmerkungen: ✓ = statistisch signifikante relative Schwäche; X = keine statistisch signifikante relative Schwäche; n. u. = nicht untersucht.

Der hervorstechende Befund ist, dass das Antwortformat bei den Schülersubgruppen, deren Relevanz im Kontext von Lesekompetenz theoriebasiert identifiziert wurde, zu relativen Unterschieden in den Lösungswahrscheinlichkeiten all dieser Schülergruppen führt. Methodisch erwiesen sich die Verfahren im Kontext der Item-Response-Theorie als besonders geeignet für die DIF-Analyse von LSA-Daten. Die Ergebnisse der inhaltlichen Forschungsfragen müssen mit Blick auf die Testfairness differenziert betrachtet werden: Es gibt in Bezug auf das Aufgabenmerkmal Antwortformat signifikante relative Stärken und Schwächen bei Kindern und Jugendlichen, die sich anhand verschiedener Merkmale unterscheiden, während keine differenziellen Kompetenzen bei anderen Aufgabenmerkmalen nachgewiesen wurden. Das Antwortformat kann als ein konstruktirrelevantes Aufgabenmerkmal eingestuft werden.

Dennoch kann nicht abschließend geschlussfolgert werden, dass das gefundene DIF zu Testunfairness führt, hierzu bedürfte es einer Festlegung der Zielvariablen.

6 Abschließende Bemerkung

Ausgehend von der Zusammenschau der Ergebnisse aus unterschiedlichen Stichproben unterschiedlichen Alters können Entwicklungsbedarfe hinsichtlich der Testkonstruktion und -interpretation unter Aspekten der Testfairness vorgeschlagen werden, die vor allem den sensiblen Umgang mit der Auswahl von Antwortformaten bei Lesetestaufgaben beinhalten. Darüber hinaus wurden aus den beobachteten subgruppenspezifischen Stärken und Schwächen in der Bearbeitung von unterschiedlichen Aufgabenstrukturen und -inhalten Implikationen zur gezielten Förderung identifiziert, die in Folgeprojekten genutzt werden können.

Literaturverzeichnis

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2015). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Hrsg.), *Differential Item Functioning* (S. 3–23). Hillsdale: Lawrence Erlbaum Associates.

Beutel, S. I., Bos, W. & Porsch, R. (Hrsg.). (2013). *Lernen in Vielfalt: Chance und Herausforderung für Schul-und Unterrichtsentwicklung. Münster:* Waxmann.

Camilli, G. (2006). Test fairness. In R. Brennan (Hrsg.), *Educational measurement* (4. Auflage, S. 221–256). Westport, CT: American Council on Education and Praeger.

Chatterji, M. (2013). *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*. Bingley: Emerald Group.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39 (12),* 1–28.

Decristan, J., Naumann, A., Fauth, B., Rieser, S., Büttner, G. & Klieme, E. (2014). Heterogenität von Schülerleistungen in der Grundschule: Bedeutung unterschiedlicher Leistungsindikatoren und Bedingungsfaktoren für die Einschätzung durch Lehrkräfte. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46 (4),* 181–190.

Deville, C. & Chalhoub-Deville, M. (2011). Accountability-assessment under No Child Left Behind: Agenda, practice, and future. *Language Testing, 28 (3),* 307–321.

Ehmke, T. & Baumert, J. (2008). Soziale Disparitäten des Kompetenzerwerbs und der Bildungsbeteiligung in den Ländern: Vergleiche zwischen PISA 2000 und 2006. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme et al. (Hrsg.), *PISA 2006 in Deutschland: Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (S. 319–342). Münster: Waxmann.

Ferne, T. & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4 (2),* 113–148.

Haag, N., Heppt, B., Stanat, P., Kuhl, P. & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 28,* 24–34.

Koo, J., Becker, B. J. & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing, 31 (1),* 89–109.

Li, D. & Kolen, M. (2005). *Exploring item characteristics associated with DIF in reading comprehension between English language learners (ELLs) and non-ELLs.* Vortrag auf dem Annual Meeting der National Council on Measurement in Education, Montreal, Canada.

McElvany, N. & Schwabe, F. (2013). Fairness von Lesetestaufgaben für Kinder aus Familien mit unterschiedlichem sozioökonomischem Status bei Large-Scale-Studien. In N. McElvany & H. G. Holtappels (Hrsg.), *Empirische Bildungsforschung. Theorien, Methoden, Befunde und Perspektiven. Festschrift für Wilfried Bos* (S. 219–233). *Münster: Waxmann.*

Schwabe, F. (2014). *Leseleistungsdifferenzen bei spezifischen Schülersubgruppen: DIF-Analysen von Large-Scale-Assessments.* Dissertation, Technische Universität Dortmund.

Schwabe, F., McElvany, N. & Trendtel, M. (2015a). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly, 50 (1),* 1–14.

Schwabe, F., McElvany, N. & Trendtel, M. (2015b). Reading skills of students in different school tracks: Systematic (dis)advantages based on item formats in large scale assessments. *Zeitschrift für Erziehungswissenschaft, 18 (4),* 781-801. doi: 10.1007/s11618-015-0645-3.

Schwabe, F., McElvany, N., Trendtel, M., Gebauer, M. M. & Bos., W. (2014). Vertiefende Analysen zu migrationsbedingten Leistungsdifferenzen in Leseaufgaben – Differenzielles Itemfunktionieren für Kinder mit und ohne Migrationshintergrund auf Basis der Internationalen Grundschul-Lese-Untersuchung (IGLU). *Zeitschrift für Pädagogische Psychologie, 28 (3),* 1–12.

Schwippert, K., Bos, W. & Lankes, E. M. (2004). Lesen Mädchen anders? Vertiefende Analysen zu Geschlechtsdifferenzen auf Basis der Internationalen Grundschul-Lese-Untersuchung IGLU. *Zeitschrift für Erziehungswissenschaft, 7 (2),* 219–234.

Tillmann, K.-J. (2008). Die homogene Lerngruppe – oder: System jagt Fiktion. In H.-U. Otto & T. Rauschenbach (Hrsg.), *Die andere Seite der Bildung. Zum Verhältnis von formellen und informellen Bildungsprozessen* (2. Auflage, S. 33–39). Wiesbaden: VS Verlag für Sozialwissenschaften.

Wenning, N. (2007). Heterogenität als Dilemma für Bildungseinrichtungen. In S. Boller, E. Rosowski & T. Stroot (Hrsg.), *Heterogenität in Schule und Unterricht. Handlungsansätze zum pädagogischen Umgang mit Vielfalt* (S. 21–31). Weinheim: Beltz.

Xi, X. (2010). How do we go about investigating fairness? *Language Testing, 27 (2),* 147–170.

Cathrin Becker, Wolfgang Schnotz, Johannes Naumann

Vorhersage von Testleistungen aus Aufgabenanforderungen und Bearbeitungsprozessen beim Lesen elektronischer Texte (TAUBE)

1 Einleitung

Das Lesen sogenannter elektronischer Texte[1] spielt im beruflichen, sozialen und privaten Alltag eine immer wichtigere Rolle. Auch in Schule, Ausbildung und Studium ist diese Form des Lesens im Rahmen der Internetnutzung zu einem selbstverständlichen Teil des Informationsangebots im Lehr-Lern-Prozess geworden. Elektronische Texte haben zahlreiche Gemeinsamkeiten mit gedruckten Texten: Beide erfordern vom Leser[2] Prozesse der Worterkennung, der Satzstrukturanalyse (parsing), der Konstruktion von Sinneinheiten (Propositionen), der Herstellung von Sinnzusammenhängen (Kohärenzbildung) sowie die Konstruktion von sogenannten mentalen Situationsmodellen des Textinhalts (Johnson-Laird, 1983; Kintsch, 1998; Schnotz, 1994). Darüber hinaus gilt es, die diskursive Funktion der jeweiligen Textsorte zu erkennen, indem der Leser z. B. erkennt, ob es sich um einen Zeitungs- oder Lehrbuchartikel, um eine Reklame oder ein Flugblatt handelt und ob dem Text eine eher expositorische oder eher persuasive Funktion zukommt (Graesser, Millis & Zwaan, 1997). Allgemeine Lesekompetenz kann als der Grad von Akkuratheit und Effizienz beschrieben werden, mit dem die jeweiligen Teilprozesse vom Individuum vollzogen werden können. Diese wiederum ermöglichen die Bewältigung weiter gehender praktischer Anforderungen.

Das Lesen elektronischer Texte erfordert jedoch spezifische Kompetenzen, die über die allgemeine Lesekompetenz hinausgehen. Elektronische Texte werden im Allgemeinen verwendet, um in einem sehr umfangreichen Angebot nach ganz bestimmten Informationen zu suchen und diese zu einem kohärenten Ganzen zu verknüpfen. Die Leser eines solchen Texts müssen spezifizieren, welche Art von Information relevant ist. Sie müssen sich während der Informationssuche dieses Ziels bewusst bleiben, also die Zielspezifikation im Arbeitsgedächtnis präsent halten, und sich nicht von irrelevanter Information ablenken lassen. Elektronische Texte sind meist als Hypertexte organisiert, besitzen also eine nicht lineare Struktur. Dementsprechend müssen Leser bei der Informationssuche in einem sehr komplexen Informationsraum navigieren, wobei der

1 Elektronische Texte sind computerbasiert dargebotene Hypertexte, in denen Textabschnitte in Knoten eines Informationsnetzwerks organisiert sind. Diese Knoten sind miteinander durch sogenannte Hyperlinks zu hierarchischen oder netzwerkartigen Strukturen verknüpft, was einen flexiblen Zugriff auf die jeweilige Textinformation von verschiedenen Seiten her ermöglicht.

2 Wir verwenden die grammatisch maskuline Form „Leser“ zur Bezeichnung für weibliche und männliche Leser.

Begriff der Navigation die Prozesse der Bewegung im Informationsraum, der Auswahl und der Sequenzierung von Textsegmenten durch die Leser umfasst (vgl. Lawless & Schrader, 2008). Das heißt, es sind vor allem die Leser, die entscheiden, welche Textbestandteile gelesen werden und in welcher Reihenfolge dies geschieht. Das Navigieren in einem komplexen Informationsraum erfordert Orientierung über die Struktur des vorhandenen Informationsraums sowie über die verfügbaren Navigationsinstrumente. Darüber hinaus müssen die Leser bei der Informationssuche die jeweils vorgefundene Information hinsichtlich ihrer Zielrelevanz sowie ihrer Glaubwürdigkeit bewerten. Erst dann kann die vorgefundene zielrelevante Information in der oben beschriebenen Weise semantisch weiterverarbeitet und in die Wissensstrukturen des Lesers integriert werden. Insgesamt spielen damit beim Lesen elektronischer Texte Prozesse der Zielsetzung, Orientierung, Navigation und Informationsbewertung eine deutlich größere Rolle als beim Lesen von üblicherweise linear strukturierten gedruckten Texten. Leser von elektronischen Texten benötigen folglich über die allgemeine Lesekompetenz hinaus auch Wissen über die Konventionen der Gestaltung und Funktionalität von Benutzeroberflächen sowie Wissen über Navigationstechniken, etwa die Verwendung von Hyperlinks, Scrollbars und der Möglichkeiten des Backtrackings.

Im Rahmen des OECD-Projekts PISA 2009 wurden die Kompetenzen von insgesamt 470.000 Schülerinnen und Schülern aus 65 Staaten in den Bereichen „Lesen", „Mathematik" und „Naturwissenschaften" erfasst. Die Teilnehmer waren zum Zeitpunkt der Untersuchung 15 Jahre alt. Innerhalb dieses Untersuchungsrahmens wurde in 19 Ländern außerdem das Lesen elektronischer Texte als ein spezifischer Kompetenzbereich (in Form einer sogenannten internationalen Option) computergestützt erfasst. Bei der Durchführung der Testung zum elektronischen Lesen wurden für jeden Teilnehmer Logfiles erstellt, die eine Rekonstruktion der Navigationspfade und damit eine Analyse des Informationssuchverhaltens sowie der abschließenden Aufgabenbeantwortung erlauben.

2 Bisheriger Forschungsstand

In verschiedenen Untersuchungen zum Lesen elektronischer Texte konnten Zusammenhänge zwischen der Art der Navigation und den Verstehensleistungen nachgewiesen werden (z. B. Naumann, 2010; Rouet, 2006; Rouet & Passerault, 1999). Studien zur Bedeutung der Struktur des Informationsraumes weisen darauf hin, dass hierarchisch strukturierte Hypertexte zu besseren Verstehens- und Lernleistungen führen als völlig nicht lineare, netzwerkartig strukturierte Hypertexte (z. B. Lin, 2003; McDonald & Stevenson, 1998a, 1998b; Shapiro, 1998). Auch scheint das Design des Bildschirms, der zugleich als Schnittstelle für die Interaktion zwischen Leser und elektronischem Text dient, eine wichtige Rolle zu spielen. Hierzu gehört unter anderem, ob Hyperlinks typisiert sind, ob also ihr Erscheinungsbild bereits auf eine bestimmte Funktion verweist, oder ob dies nicht der Fall ist. Ebenso gehört hierzu, ob Übersichtskarten über den Textinhalt vorhanden sind (vgl. z. B. DeStefano & LeFevre, 2007; Shapiro & Niederhauser, 2004) oder ob Signaling-Techniken angewandt werden (Lorch, 1989; Lorch & Lorch, 1995, 1996; Naumann, Richter, Flender, Christmann & Groeben, 2007).

Ob ein Navigationsverhalten kohärenzfördernd ist oder nicht, kann nicht einfach an der Zahl der Seitenaufrufe insgesamt oder der Zahl der Aufrufe anforderungsrelevanter Seiten festgemacht werden. Eine hohe Zahl aufgerufener Seiten bedeutet nicht notwendig, dass es sich dabei um anforderungsrelevante Seiten gehandelt hat. Aber selbst eine hohe Zahl aufgerufener relevanter Seiten und eine längere Bearbeitungszeit weisen nicht notwendig auf eine gründliche Textlektüre hin, bei der der Leser die Textinformation elaboriert verarbeitet und durch Inferenzbildung eine hochkohärente mentale Repräsentation konstruiert. Eine hohe Zahl aufgerufener (relevanter oder irrelevanter) Textseiten und eine längere Bearbeitungszeit können auch nur Ausdruck eines unsystematischen Umherirrens im Informationsraum sein, bei der die einzelnen Textinformationen nur oberflächlich verarbeitet werden (Naumann, 2010). In diesem Fall spiegelt das Navigationsverhalten eine Desorientierung des Probanden wider, die meist als „lost in hyperspace" bezeichnet wird und bei der der Proband nicht mehr in der Lage ist, die Struktur des Informationsraums im Blick zu behalten und für die Informationssuche fruchtbar zu machen (Schroeder & Grabowski, 1995; Smith, 1996).

Solche Orientierungsdefizite manifestieren sich häufig in Rücksprüngen zu bereits zuvor aufgerufenen Textseiten. In verschiedenen Untersuchungen wurden negative Zusammenhänge zwischen der Häufigkeit solcher Rücksprünge mit den erzielten Lern- und Verstehensleistungen gefunden (z. B. Cress & Knabel, 2003; Richter, Naumann & Noller, 2003). Gleichzeitig berichten Leser, die häufig solche Rücksprünge vollzogen haben, vermehrt das subjektive Erleben von Desorientierung (z. B. Herder & Juvina, 2004; Schroeder & Grabowski, 1995). Die Häufigkeit von Rücksprüngen muss auch differenziert bewertet werden in Abhängigkeit davon, wie viel Orientierungshinweise in Form von signaling (textuelle Signale) gegeben werden. Naumann et al. (2007) fanden in einem Hypertext mit weniger textuellen Signalen mehr Rücksprünge, wobei deren Zahl positiv mit den Lernerfolgsmaßen korreliert waren. In der Bedingung mit vielen textuellen Signalen hingegen war die Zahl der Rücksprünge nicht oder negativ mit verschiedenen Lernerfolgsmaßen korreliert. Ähnlich fanden Richter et al. (2005), dass bei einem mit vielen textuellen Signalen ausgestatteten Hypertext Rücksprünge negativ mit Lernergebnissen assoziiert waren, nicht dagegen, wenn dasselbe Textmaterial in Form eines elektronisch dargebotenen Lineartexts organisiert war, wo Rücksprünge einem einfachen Zurückblättern entsprechen.

Insgesamt gesehen scheinen Aufgaben beim Lesen mit elektronischen Texten insofern besondere Anforderungen zu stellen, als es zu weiten Teilen dem Leser oder der Leserin überlassen bleibt, aufgabenrelevantes Textmaterial auszuwählen, zu sequenzieren und kognitiv zu verarbeiten. Die bisherige Forschung hat sich somit zum einen auf die Auswirkung verschiedener Facetten der Gestaltung von Hypertexten in Verbindung mit bestimmten Aufgabenanforderungen auf das elektronische Lesen und Textverstehen konzentriert. Zum anderen hat sie Prozessmerkmale des Navigationsverhaltens beim elektronischen Lesen untersucht und in einen Zusammenhang mit dem Leseverständnis gestellt. Das Navigationsverhalten ist jedoch nicht nur eine Bedingung erfolgreichen elektronischen Leseverstehens, sondern auch Resultat der jeweiligen Aufgabenanforderungen sowie der Gestaltung des jeweiligen Hypertexts. Insofern bietet es sich an, das Zusammenspiel von Textmerkmalen, Aufgabenmerk-

malen, Personenmerkmalen und Merkmalen des Navigationsverhaltens simultan zu betrachten und entsprechend zu modellieren.

3 Fragestellung und Hypothesen

In den bisherigen Untersuchungen wurde meist untersucht, wieweit spezifische Einflussfaktoren wie z. B. bestimmte Textmerkmale, Itemmerkmale, Navigationsanforderungen oder das reale Navigationsverhalten für sich genommen prädiktiv für den Verstehenserfolg und die Itembeantwortung sind. Kaum untersucht wurden bislang jedoch die möglichen Wechselwirkungen zwischen diesen Faktoren, die zu Moderationseffekten führen können, indem manche Zusammenhänge nur unter bestimmten Bedingungen bestehen.

Ziel der vorliegenden Ausführungen ist, anhand der in PISA 2009 erhobenen Daten zur Aufgabenbearbeitung und Aufgabenbeantwortung mittels Logfile-Analysen die Frage zu beantworten, wieweit die Struktureigenschaften des Informationsraumes in Verbindung mit dem tatsächlichen Informationssuchverhalten die Verstehensleistungen beim Lesen elektronischer Texte erklären können. Dies geht über die bisher vorliegenden Analysen zum elektronischen Lesen bei PISA 2009 (siehe Kapitel 8 im Band VI des Berichts über PISA 2009), in denen bereits auf die Wichtigkeit der Informationssuche beim Lesen elektronischer Texte hingewiesen wurde, hinaus (Organisation for Economic Co-operation and Development [OECD], 2012). Angesichts des internationalen Charakters der Studie besteht darüber hinaus die Möglichkeit, etwaige Unterschiede zwischen den einzelnen Ländern im Hinblick auf die verschiedenen Zusammenhänge aufzudecken.

3.1 Einflüsse von Lesermerkmalen auf die Verstehensleistungen bei der Verarbeitung elektronischer Texte

Hinsichtlich der Leserseite sei zunächst auf den Einfluss von Personenmerkmalen auf die Kompetenz zum Lesen elektronischer Texte bzw. auf die Leseleistungen eingegangen. Was die Gemeinsamkeiten betrifft, so erfordern beide Varianten des Lesens allgemeine Lesekompetenz.

Dementsprechend wurde erwartet, dass die Kompetenz zum Lesen elektronischer Texte und die zum Lesen traditioneller gedruckter Texte hoch korrelieren bzw. dass die eine Kompetenz die andere in relativ hohem Maß vorhersagen kann.

3.2 Einflüsse von Merkmalen des Testmaterials beim Lesen elektronischer Texte auf die Itemschwierigkeiten

Hinsichtlich des Testmaterials sei zunächst auf den Einfluss von Textmerkmalen auf die Schwierigkeit der Items eingegangen. Leitfrage ist hier, wieweit sich die empirischen Itemschwierigkeiten des Tests zur Erfassung der Kompetenz zum Lesen

elektronischer Texte aus den Textmerkmalen vorhersagen lassen. Angesichts der Gemeinsamkeiten zwischen dem Lesen elektronischer Texte und dem Lesen von traditionellen gedruckten Texten kann davon ausgegangen werden, dass das Lesen längerer Texte mehr kognitiven Aufwand erfordert und damit zu einer höheren Itemschwierigkeit führt als das Lesen kürzerer Texte. Dementsprechend ist zu erwarten, dass Items, die das Lesen von längeren Texten erfordern, (ceteris paribus) schwieriger sind als Items, die nur das Lesen kürzerer Texte erfordern.

Außerdem müssten leichter lesbare Texte (d. h. Texte, die pro Textlängeneinheit weniger kognitiven Aufwand erfordern) zu einer geringeren Itemschwierigkeit führen als schwerer lesbare Texte.

Dementsprechend ist zu erwarten, dass Items, die das Lesen von lesbareren Texten erfordern, (ceteris paribus) leichter sind als Items, die das Lesen von weniger lesbaren Texten erfordern.

Was die Unterschiede zwischen beiden Textarten betrifft, so erfordert das Lesen elektronischer Texte, wie bereits erwähnt, (durch das Operieren mit einer Mensch-Computer-Schnittstelle) ein höheres Maß an Interaktivität als ein gedruckter Text. Dementsprechend stellt sich hier die Frage, wieweit sich die empirischen Itemschwierigkeiten aus den Navigationsanforderungen vorhersagen lassen. Hier ist zu erwarten, dass Items, die höhere Navigationsanforderungen stellen, (ceteris paribus) schwieriger sind als solche mit geringeren Navigationsanforderungen.

Auch hier stellt sich über die genannten Fragen nach der Existenz der einzelnen Zusammenhänge hinaus die Frage, welches relative Gewicht den verschiedenen Prädiktoren zukommt.

3.3 Interaktionen zwischen Lesermerkmalen und Merkmalen des Testmaterials auf den Prozess der Itembeantwortung

Neben den oben beschriebenen Haupteffekten der genannten Merkmale des Lesers und des Testmaterials ist noch mit verschiedenen Interaktionseffekten zu rechnen, da manche der oben genannten Variablen auch eine Moderatorfunktion hinsichtlich der Wirksamkeit anderer Einflussvariablen wahrnehmen können. Beispielsweise könnten höhere Navigationsanforderungen einerseits die Itemschwierigkeit erhöhen und damit die Leistungen bei der Itembeantwortung reduzieren, andererseits aber auch den Einfluss der Textlänge auf die Itembeantwortung moderieren: Höhere Navigationsanforderungen bedeuten, dass für die Lektüre der relevanten Textseiten mehr Clicks erforderlich sind, wodurch (ceteris paribus) die einzelnen Textseiten kürzer werden, die relevante Textinformation also in kleineren, kognitiv „leichter verdaubaren“ Texthäppchen dargeboten wird. Unter diesen Umständen könnte sich eine größere Textlänge nicht mehr so negativ auf den Verstehensprozess auswirken. Gewissermaßen im Gegenzug könnten die Navigationsanforderungen den Einfluss der Lesbarkeit der Textseiten auf die Itembeantwortung verstärken, da in der Situation der „häppchenweisen“ Darbietung der Lesbarkeit dieser relativ kurzen Seiten eine vergleichsweise wichtigere Rolle zukommt als in anderen Situationen, wo die Länge der Texte den Leser stärker beansprucht.

Was die Vorhersage des Navigationsverhaltens betrifft, so dürfte dies in hohem Maße von den Navigationsanforderungen abhängen. Höhere Navigationsanforderungen dürften die richtige Itembeantwortung erschweren. *Innerhalb* gegebener Navigationsanforderungen aber könnte ein intensiveres tatsächliches Navigieren im Sinne eines häufigeren Aufsuchens der jeweils relevanten Seiten zu besseren Leistungen bei der Itembeantwortung führen. Dieser Zusammenhang dürfte besonders dann ausgeprägt sein, wenn tatsächlich höhere Navigationanforderungen bestehen, es sich also um anspruchsvollere Hypertexte handelt, das intensivere Navigieren insofern also funktional für den Verstehensprozess ist. Wenn hingegen viel navigiert wird, obwohl nur geringe Navigationsanforderungen bestehen, so ist dies eher als Ausdruck von Desorientierung zu interpretieren. In diesem Fall ist kein positiver Zusammenhang zwischen der Intensität des Navigationsverhaltens und den Leistungen bei der Itembeantwortung zu erwarten.

4 Methode

4.1 Testmaterial

Zur Messung der Kompetenz für das Lesen elektronischer Texte wurden neun Testeinheiten mit insgesamt 28 Items verwendet. Die Testeinheiten handelten jeweils über ein bestimmtes Thema wie z. B. Geruch. Die verschiedenen Testeinheiten stellten eine Bandbreite unterschiedlicher Szenarien dar, um eine Vielfalt herzustellen. Zu jeder Einheit gehörten bestimmte Items, wobei die Zahl der Items zwischen einem und vier variierte. Nachdem ein Item bzw. eine Einheit bearbeitet worden war, konnten die Probanden nicht wieder zu diesem Item respektive dieser Einheit zurückspringen.

Das Testmaterial wurde am Bildschirm anhand von zwei Bereichen dargestellt – dem Browserbereich, in dem das Stimulusmaterial dargeboten wurde, und dem Aufgabenbereich bzw. Itembereich, in dem die Fragen zu lesen und auch in den meisten Fällen zu beantworten waren.

Die verschiedenen Items ermittelten die Fähigkeit, Texte zusammenzufassen, zu beschreiben und zu analysieren, die wichtigen Elemente einer Information herauszufinden, relevante Informationen zu identifizieren sowie die „Güte“ der gefundenen Informationen zu bewerten. Bei einigen Items waren die benötigten Informationen schon auf der Startseite zu finden. Bei anderen Items musste der Leser auf mehrere Seiten navigieren, da die notwendigen Informationen sich auf unterschiedlichen Seiten befanden. Das Testmaterial repräsentierte zum Zeitpunkt der Erhebung das Spektrum digitaler Texte mit Blogs, Webseiten, E-Mails, Foren. Navigationstools waren:

- Scrollbars, um sich innerhalb der Seite zu bewegen
- Tabs/Reiter für verschiedene Webseiten
- Auflistungen von Hyperlinks in einer Reihe, Spalte oder als Drop-down-Menü
- eingebettete Hyperlinks innerhalb eines Absatzes, einer Tabelle oder einer Information
- Sitemaps

Die Aufgaben bzw. Items wurden bewusst so konstruiert, dass überwiegend Navigation nötig war, um die volle Punktzahl zu erreichen. Folglich wurden die Schüler in manchen Items aufgefordert, eine bestimmte Anzahl an Seiten zu besuchen, um die Informationen, die sie zur Beantwortung des Items brauchten, zu finden oder Informationen von mindestens zwei verschiedenen Seiten zu integrieren. Bei 20 Items mussten die Schülerinnen und Schüler navigieren, um die richtige Antwort herauszufinden.

Für die Charakterisierung der Komplexität der Informationsräume wurden jeweils die Anzahl der Gesamtknoten, die höchste und die niedrigste Textlänge innerhalb dieser Knoten sowie die Gesamtzahl der Links betrachtet. Die Gesamtzahl der Knoten variierte zwischen 9 und 31, die höchste Textlänge lag zwischen 131 und 411 und die niedrigste Textlänge zwischen 17 und 124; die Gesamtzahl der Links lag zwischen 35 und 150.

Auf der Basis dieser Strukturanalysen wurde dann für die einzelnen Items zunächst die Zahl der für die Itembeantwortung relevanten Seiten[3] bestimmt. Darüber hinaus wurde die Höhe der Navigationsanforderungen festgelegt: Hierzu wurde die Anzahl der notwendigen Clicks identifiziert, die unter Verwendung des optimalen Navigationspfads notwendig waren, um die zur Itembeantwortung relevanten Seiten zu lesen.

Die folgenden Analysen zur Lesbarkeit bzw. zur Leichtigkeit der kognitiven Verarbeitung der zur Itembeantwortung relevanten Textseiten wurden jeweils anhand des englischsprachigen Textmaterials vorgenommen. Natürlich haben die verschiedenen Sprachen ihre morphologischen und syntaktischen Spezifika sowie ihre eigenen linguistischen Mechanismen zur Förderung der mentalen Kohärenzbildung bei den Lesern, wodurch es grundsätzlich problematisch ist, in einem solchen Projekt nur eine der Sprachen zu fokussieren und die dort gefundenen Merkmale auf die anderen Sprachen zu übertragen, indem dort die gleichen Verhältnisse zwischen den einzelnen Analysen angenommen werden. Wenn allerdings bei allen strukturellen Unterschieden zwischen den in PISA 2009 verwendeten Sprachen die Unterschiede zwischen den verwendeten Texten innerhalb einer Sprache bei der Übersetzung in eine andere Sprache erhalten bleiben (wenn ein Text A nicht nur im Englischen komplexer ist als ein Text B, sondern auch im Französischen oder im Japanischen), so kann die Fokussierung auf eine Sprache dennoch zu global brauchbaren Ergebnissen führen.

Für jedes Item wurden die zur Beantwortung relevanten Textseiten identifiziert. Damit wurde dann zum einen die Gesamtlänge des aus den itemrelevanten Seiten bestehenden Texts als Anzahl der enthaltenen Worte analysiert. Diese kann als Indikator für die zu leistende Menge an kognitiver Verarbeitung angesehen werden. Zum anderen wurde versucht, die Lesbarkeit dieser relevanten Textseiten zu bestimmen. Diese kann als Indikator für die (relative, auf eine bestimmte Textlängeneinheit bezogene) Leichtigkeit der Verarbeitung angesehen werden.

Um die syntaktische und morphologische Komplexität innerhalb eines Textes und die Texteigenschaften, die über die Betrachtung einzelner Sätze hinausgehen (indem sie z. B. auf den Aspekt der Kohärenzbildung abheben), zu erfassen, wurde

3 Als relevante Seiten gelten jene Seiten, die lösungsfördernde Information enthalten, unabhängig davon, ob diese unabdingbar (notwendig) oder nur lösungsfördernd sind.

mithilfe von Coh-Metrix je ein Index für die Leichtigkeit der Verarbeitung auf der Wortebene, der Satzebene und der Textebene bestimmt.

Die Berechnung des Index wurde mit der an der Universität Memphis TN im Institute for Intelligent Systems entwickelten Software Coh-Metrix vorgenommen (Graesser, McNamara & Kulikowich, 2011; Cai et al., 2004)[4]. Coh-Metrix ist ein computerbasiertes Tool zur Analyse englischsprachiger Texte, mit dem verschiedene Indizes errechnet werden können. Neben Maßen zur Lesbarkeit und Häufigkeit von Wörtern, Sätzen, Abschnitten und zur syntaktischen Komplexität werden auch Maße der Kohäsion und Kohärenz bestimmt. Coh-Metrix basiert auf verschiedenen Datenbanken und Lexika, wie unter anderem MRC Psycholinguistics Database, WordNet. Die eingesetzte Coh-Metrix-Version ist Version 3.0.

(1) **Wortebene:** Für die Leichtigkeit der Verarbeitung auf der Wortebene wurde die (durch korpuslinguistische Analysen definierte) durchschnittliche statistische Häufigkeit der enthaltenen Inhaltsworte bestimmt.
(2) **Satzebene:** Für die Leichtigkeit der Verarbeitung auf der Satzebene wurde die durchschnittliche syntaktische Einfachheit bestimmt, die ihrerseits durch geringe Satzlänge und einfache, vertraute syntaktische Strukturen charakterisiert wird.
(3) **Textebene:** Für die Leichtigkeit der Verarbeitung auf der Textebene im Hinblick auf die erforderliche mentale Kohärenzbildung wurde die durchschnittliche semantische Verknüpftheit zwischen aufeinanderfolgenden Sätzen im Sinne der von Haviland und Clark (1974) beschriebenen Given-new-Strategie mithilfe eines Algorithmus der Latenten Semantischen Analyse (LSA, Landauer et al., 2007) bestimmt.

4.2 Navigations- und Antwortverhalten

Das Navigationsverhalten der Probanden wurde per Logfile-Recording automatisch erfasst. Festgehalten wurden dabei die ausgeführten Clicks zusammen mit einem Zeitstempel. Auf der Grundlage der damit vorliegenden Sequenz-Rohdaten wurde für jeden Probanden bei jedem Item unter anderem die Zahl der Seitenbesuche itemrelevanter Seiten (inklusive wiederholter Besuche) erfasst. Außerdem wurde das Antwortverhalten der Probanden bei den einzelnen Items erfasst, wobei alle Antworten einheitlich in ein dichotomes Format (richtig/falsch) übersetzt wurden.[5] Außerdem wurde die für die Itembeantwortung benötigte Zeit erfasst.

4 Die Autoren danken Carl Zhiqiang Cai und Arthur C. Graesser für ihre freundliche Unterstützung bei der Analyse der Texte mithilfe von Coh-Metrix.

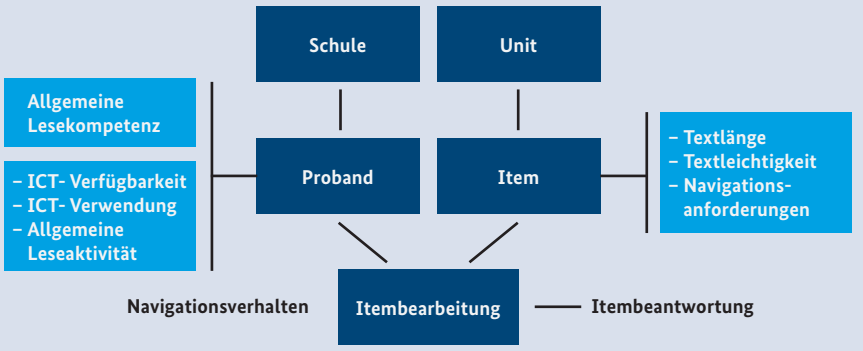
5 Bei drei Vierteln der Items war eine dichotome Kodierung (nicht beantwortet/beantwortet) und bei einem Viertel der Items war zunächst eine dreistufige Kodierung realisiert worden (nicht beantwortet/teilweise beantwortet/ganz beantwortet). Einheitlichkeitshalber wurde die dreistufige Codierung in eine dichotome Codierung transformiert, indem „teilweise beantwortet“ und „ganz beantwortet“ in „beantwortet“ übersetzt wurden.

4.3 Analyseverfahren

Die im Projekt erhobenen Daten stammen von Untersuchungseinheiten, die teilweise ineinander verschachtelt sind und damit unterschiedlichen Hierarchieebenen angehören. Die Items werden von Schülerinnen und Schülern bearbeitet. Da alle Teilnehmer mehrere Items bearbeiten, werden die Bearbeitungsprozesse dieser Items jeweils von den gleichen personalen Merkmalen des betreffenden Individuums beeinflusst. Die Schülerinnen und Schüler kommen ihrerseits aus bestimmten Schulen, wobei jeweils mehrere Teilnehmer die gleiche Schule besuchen und damit gemeinsam bestimmten schulischen Einflüssen ausgesetzt sind. Somit sind die Bearbeitungsprozesse in Teilnehmer und die Teilnehmer in Schulen geschachtelt, was einer hierarchischen Drei-Ebenen-Datenstruktur (Bearbeitungsprozess – Teilnehmer – Schule) entspricht. Im Folgenden wird dies als Personenhierarchie bezeichnet.

Die Bearbeitungsprozesse der Schülerinnen und Schüler sind aber zugleich die Bearbeitung bestimmter Items mit ihren spezifischen Materialmerkmalen und Anforderungen. Da jedes Item von vielen Teilnehmern bearbeitet wird, werden diese Bearbeitungsprozesse gemeinsam von den gleichen Materialmerkmalen und Anforderungen beeinflusst. Die Items gehören wiederum zu bestimmten Testeinheiten, sodass die Items einer bestimmten Einheit gemeinsam von den Merkmalen dieser Einheit beeinflusst werden. Somit sind die Bearbeitungsprozesse auch in Items und die Items in Testeinheiten geschachtelt, was ebenfalls einer hierarchischen Drei-Ebenen-Datenstruktur (Bearbeitungsprozess – Item – Testeinheit) entspricht. Im Folgenden wird dies als Itemhierarchie bezeichnet. Beide Hierarchien sind miteinander verschränkt, sodass eine gekreuzte hierarchische Mehrebenenstruktur (hierarchical and cross-classified structure) entsteht. Die Datenstruktur ist in Abbildung 1 dargestellt.

Abb. 1: Grafische Darstellung der gekreuzten hierarchischen Mehrebenenstruktur der vorliegenden Daten



Eine solche Datenstruktur ermöglicht Analysen aus verschiedenen Perspektiven, wobei manchmal nur eine Ebene und manchmal mehrere Ebenen einbezogen werden müssen.

5 Ergebnisse

Im Folgenden sollen Prozesse der Itembearbeitung simultan anhand beider Hierarchien vorhergesagt bzw. erklärt werden. Hierzu wird zunächst das Antwortverhalten bei der Bearbeitung der Items zur allgemeinen Lesekompetenz der Probanden und zu den Navigationsanforderungen der Items sowie zum Umfang der erforderlichen kognitiven Verarbeitung[6] und zur relativen Leichtigkeit der kognitiven Verarbeitung[7] in Beziehung gesetzt. Darüber hinaus wird auch das Navigationsverhalten als Prädiktor in die Analyse mit einbezogen.

Die Auswertung erfolgte jeweils durch Mehrebenenanalysen mit dem R-Paket. Die eingesetzte R-Version mit R-Studio war Rx64 3.0.3.[8] Da angesichts einer Fallzahl von über 580.000 selbst geringste Zusammenhänge statistisch signifikant von null verschieden sind, werden im Folgenden nur solche Zusammenhänge berichtet, deren standardisierte Regressionskoeffizienten einen Wert von 0.10 oder höher haben. Die berichteten Regressionskoeffizienten sind alle statistisch auf dem .01-Niveau signifikant.

5.1 Länderübergreifende Analyse

Das Modell zur Vorhersage der Richtigkeit der Itembeantwortung anhand der oben genannten Prädiktoren für alle Probanden bzw. für alle Länder ist in Abbildung 2 dargestellt. Prädiktor auf Ebene 1 ist das Navigationsverhalten. Prädiktor auf Ebene 2 der Probandenhierarchie ist die allgemeine Lesekompetenz. Prädiktoren auf Ebene 2 der Itemhierarchie sind Navigationsanforderungen sowie Textlänge und Textleichtigkeit. Das folgende Modell berücksichtigt sowohl die Navigationsanforderungen der Items (Ebene 2 der Itemhierarchie) als auch das tatsächliche Navigationsverhalten der Probanden bei der Bearbeitung der Items (Hierarchieebene 1).

Das Modell erklärt mit dem Navigationsverhalten als Prädiktor 10,9 Prozent der Itembeantwortungsvarianz innerhalb der Personenhierarchie. Verwendet man nur die allgemeine Lesekompetenz als Prädiktor, so erklärt diese 60,8 Prozent der Itembeantwortungsvarianz innerhalb der Personenhierarchie. Fügt man dem Navigationsverhalten als Prädiktor die allgemeine Lesekompetenz als Prädiktor hinzu, so erklären beide gemeinsam 66,3 Prozent der Itembeantwortungsvarianz innerhalb der Personenhierarchie. Innerhalb der Itemhierarchie erklärt zunächst das Navigationsverhalten als Prädiktor ebenfalls 10,9 Prozent der Itembeantwortungsvarianz. Fügt man dem Navigationsverhalten als Prädiktor die Textleichtigkeit als Prädiktor hinzu, so erklären beide zusammen 22 Prozent der Itembeantwortungsvarianz innerhalb der Itemhierarchie. Nimmt man zusätzlich noch Navigationsanforderungen auf, so

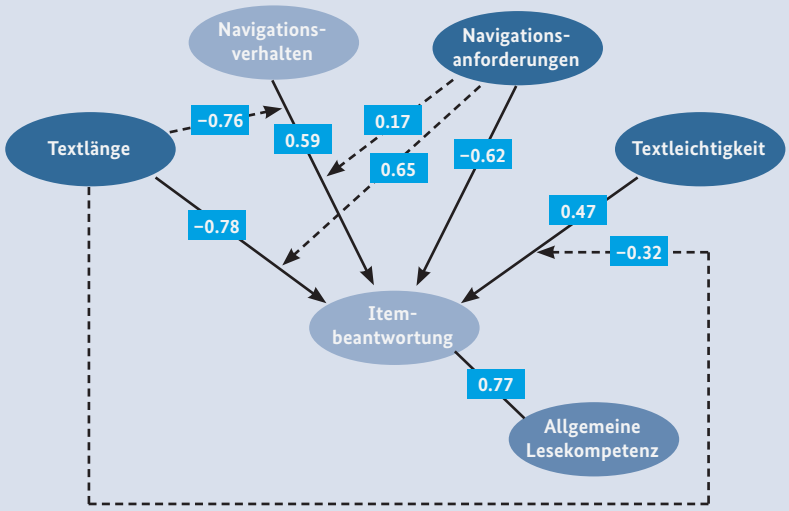
6 Der Umfang der erforderlichen kognitiven Verarbeitung wurde durch die Textlänge indiziert.

7 Die Leichtigkeit der kognitiven Verarbeitung wurde durch multiple Regression der Itembeantwortung auf Leichtigkeit auf der Wortebene, Leichtigkeit auf der Satzebene und Leichtigkeit auf der Textebene in Form einer entsprechenden Linearkombination definiert.

8 R-Version 3.0.3 (2014-03-06) – „Warm Puppy". © 2014 The R Foundation for Statistical Computing. Platform: x86_64-w64-mingw32/x64.

erhöht sich die aufgeklärte Itembeantwortungsvarianz auf 26 Prozent. Fügt man außerdem noch die Textlänge als Prädiktor hinzu, so ergibt sich eine insgesamt aufgeklärte Itembeantwortungsvarianz von 34 Prozent innerhalb der Itemhierarchie.

Abb. 2: Vorhersage der Itembeantwortung anhand von Navigationsverhalten, allgemeiner Lesekompetenz, Navigationsanforderungen sowie Textlänge und Textleichtigkeit



In diesem Modell ist die allgemeine Lesefähigkeit ein sehr guter Prädiktor für die Itembeantwortung (Beta des Haupteffekts = 0.77). Dies ist nicht überraschend, da auch elektronische Texte allgemeine Lesefähigkeit erfordern. Dieser Haupteffekt wird nicht durch andere Einflüsse moderiert, denn es fanden sich keine nennenswerten Interaktionseffekte. Auch hier überrascht nicht, dass kürzere und leichtere Texte die richtige Beantwortung von Items begünstigen. Für den Haupteffekt der Textlänge ergibt sich ein Beta von −0.78, für den der Textleichtigkeit ein Beta von 0.47.

Der Einfluss der Textlänge auf die Itembeantwortung wird einerseits durch die Navigationsanforderungen moderiert. Höhere Navigationsanforderungen führen zunächst (ceteris paribus) zu einer schlechteren Itembeantwortung (Beta des Haupteffekts = −0.62), was wieder nicht überrascht, da höhere Anforderungen das Arbeitsgedächtnis belasten und insofern mehr Fehlermöglichkeiten schaffen. Gleichzeitig schwächt sich aber der negative Einfluss der Textlänge auf die Itembeantwortung bei höheren Navigationsanforderungen ab. Für diese Zweifachinteraktion ergibt sich ein Beta von 0.65. Das heißt, bei höheren Navigationsanforderungen wird der negative Einfluss der Textlänge zunehmend aufgehoben.

Was die Moderation des Einflusses von Textlänge auf die Itembeantwortung durch die Navigationsanforderungen betrifft, so kann man auch hier annehmen, dass höhere Navigationsanforderungen (also mehr notwendige Clicks) bei gleicher Textlänge bedeuten, dass die einzelnen Textseiten kürzer sind. Die Lektüre verteilt

sich dadurch bei höheren Navigationsanforderungen gewissermaßen auf eine größere Zahl kleinerer Textabschnitte, die entsprechend leichter kognitiv verarbeitbar sind. Infolgedessen scheint sich die Textlänge nicht mehr so negativ auf die Itembeantwortung auszuwirken.

Der Einfluss der Leichtigkeit des Texts auf die Itembeantwortung wird durch die Textlänge moderiert. Für diese Zweifachinteraktion ergibt sich ein Beta von −0,32. Das heißt, je länger der Text, desto geringer wird der positive Einfluss der Textleichtigkeit auf die Itembeantwortung.[9]

Höhere Navigationsanforderungen gehen mit einer schlechteren Itembeantwortung einher (Beta von −0.62). Intensiveres tatsächliches Navigationsverhalten trägt jedoch zur korrekten Itembeantwortung bei: Ein häufigeres Aufsuchen der jeweiligen itemrelevanten Seiten geht mit einer besseren Itembearbeitung einher. Für diesen Haupteffekt ergibt sich ein Beta von 0.59. Das heißt, zwar erschweren höhere Navigationsanforderungen generell eine richtige Itembeantwortung, aber im Rahmen gegebener Navigationsanforderungen führt ein intensiveres tatsächliches Navigieren (im Sinne häufigeren Aufsuchens der jeweils relevanten Seiten) zu besseren Leistungen bei der Itembeantwortung.

Der Einfluss des Navigationsverhaltens auf die Itembeantwortung wird durch die Textlänge moderiert. Für diese Zweifachinteraktion ergibt sich ein Beta von −0,76. Das heißt, ein enger Zusammenhang zwischen Navigationsverhalten und Itembeantwortung ist vor allem bei kürzeren Texten zu beobachten; bei längeren Texten hingegen verschwindet der Effekt und kann sich in sein Gegenteil kehren.

Der Einfluss des Navigationsverhaltens auf die Itembeantwortung wird außerdem durch die Navigationsanforderungen moderiert: Bei höheren Navigationsanforderungen wird dieser Zusammenhang stärker. Für diese Zweifachinteraktion ergibt sich ein Beta von 0.17.

5.2 Länderspezifika

Abbildung 3 zeigt in grafischer Form die Beta-Gewichte für den Einfluss der allgemeinen Lesekompetenz auf die Itembearbeitung in den verschiedenen Ländern. In den asiatischen Ländern (Hongkong, Japan, Korea, Macau) scheint dieser Einfluss geringer zu sein als in den übrigen Ländern, insbesondere Australien und Neuseeland. Die Unterschiede in den Beta-Gewichten zwischen beiden Ländergruppen waren signifikant ($z = 5.406$; $p < .001$[10]). Insgesamt gesehen zeichnen sich demnach die asiatischen Länder (Hongkong, Japan, Korea, Macau) durch eine geringere Bedeutung der allgemeinen Lesekompetenz für die Itembeantwortung im Vergleich zu den übrigen Ländern aus.

Abbildung 4 zeigt die Beta-Gewichte für den Einfluss des tatsächlichen Navigationsverhaltens auf die Itembearbeitung in den verschiedenen Ländern. In den asia-

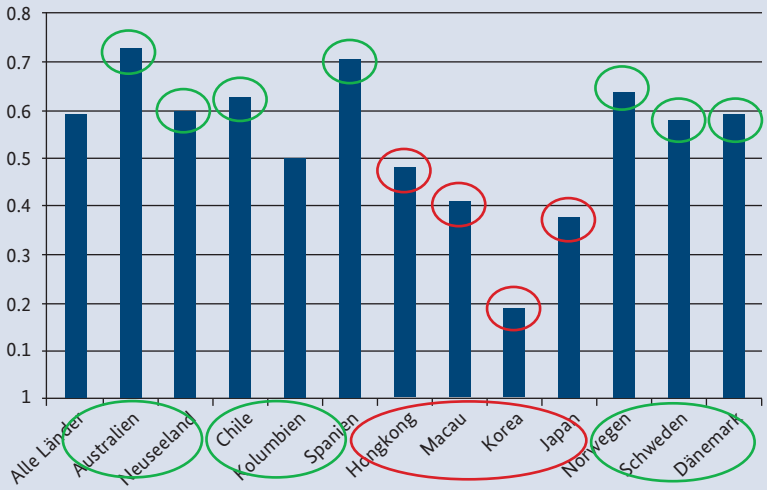
9 Dieser Interaktionseffekt ist schwer zu interpretieren: Man würde eigentlich erwarten, dass sich die Textleichtigkeit umso stärker auswirkt, je länger dieser (mehr oder weniger leicht/schwer zu verarbeitende) Text ist.

10 Der Signifikanzprüfung liegt der Vergleich zwischen Hongkong und Australien zugrunde.

Abb. 3: Haupteffekt – Allgemeine Lesekompetenz auf die Itembearbeitung

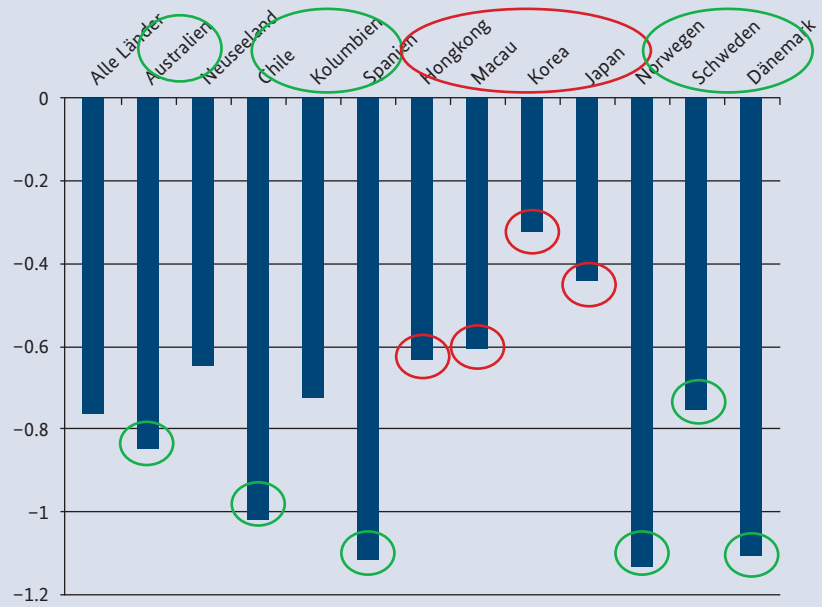
tischen Ländern (Hongkong, Japan, Korea, Macau) ist dieser Einfluss am geringsten. Die Häufigkeit des Besuchs relevanter Seiten scheint hier für eine erfolgreiche Itembearbeitung weniger bedeutsam zu sein als in den anderen Ländern. Zwischen den asiatischen Ländern (Hongkong, Japan, Korea, Macau) einerseits und den skandinavischen Ländern ergab sich ein signifikanter Unterschied ($z = 2.335$; $p = .019$).[11] Auch bestand ein signifikanter Unterschied zwischen den asiatischen Ländern (Hongkong,

Abb. 4: Haupteffekt – Navigationsverhalten auf die Itembearbeitung



11 Der Signifikanzprüfung liegt der Vergleich zwischen Hongkong und Schweden zugrunde.

Abb. 5: Interaktion – Navigationsverhalten und Textlänge auf die Itemverarbeitung



Japan, Korea, Macau) einerseits und Australien und Neuseeland andererseits (*z = 2.771; p = .006).*[12] Ferner ergab sich zwischen Korea, Macau und Japan einerseits und den Spanisch sprechenden Ländern (Chile, Kolumbien, Spanien) andererseits ein signifikanter Unterschied (*z = 2.530; p = .01).*[13]

Der Einfluss des tatsächlichen Navigationsverhaltens auf die Itembearbeitung wird durch die Textlänge moderiert. Das heißt, durch längere Texte wird der Einfluss des tatsächlichen Navigationsverhaltens auf die Itembearbeitung verringert. Abbildung 5 zeigt die Beta-Gewichte für diesen moderierenden Effekt der Textlänge. In den asiatischen Ländern (Hongkong, Japan, Korea, Macau) fällt dieser moderierende Einfluss der Textlänge am geringsten aus. Zwischen den asiatischen Ländern (Hongkong, Japan, Korea, Macau) einerseits und den skandinavischen Ländern (Dänemark, Norwegen, Schweden) andererseits ergab sich ein signifikanter Unterschied (*z = 5.226; p < .001).*[14] Auch bestand ein signifikanter Unterschied zwischen den asiatischen Ländern und Australien (*z = 8.996; p < .001).*[15] Ferner ergab sich zwischen den asiatischen Ländern (Hongkong, Japan, Korea, Macau) einerseits und Chile und Spanien andererseits ein signifikanter Unterschied (*z = 9.549; p < .001).*[16]

12 Der Signifikanzprüfung liegt der Vergleich zwischen Macau und Neuseeland zugrunde.

13 Der Signifikanzprüfung liegt der Vergleich zwischen Hongkong und Kolumbien zugrunde.

14 Der Signifikanzprüfung liegt der Vergleich zwischen Hongkong und Schweden zugrunde.

15 Der Signifikanzprüfung lag der Vergleich zwischen Hongkong und Australien zugrunde.

16 Der Signifikanzprüfung liegt der Vergleich zwischen Hongkong und Chile zugrunde.

6 Zusammenfassung und Ausblick

Die Ergebnisse der Mehrebenenanalyse der Itembeantwortungen in Abhängigkeit von allgemeiner Lesekompetenz sowie Textlänge, Textleichtigkeit, Navigationsanforderungen und Navigationsverhalten machen Folgendes deutlich. Die allgemeine Lesefähigkeit ist ein sehr guter Prädiktor für eine korrekte Itembeantwortung. Auch überrascht es nicht, dass kürzere und leichter zu lesende Texte eine korrekte Itembeantwortung begünstigen.

Das tatsächliche Navigationsverhalten folgt in hohem Maß den Navigationsanforderungen. Zwar erschweren höhere Navigationsanforderungen (ceteris paribus) generell die richtige Itembeantwortung. Andererseits führt aber im Rahmen gegebener Navigationsanforderungen ein intensiveres tatsächliches Navigieren (also ein häufigeres Aufsuchen der jeweils relevanten Seiten) zu besseren Leistungen bei der Itembeantwortung.

Der Einfluss des Navigationsverhaltens auf die Itembeantwortung wird ebenfalls durch die Textlänge moderiert. Das heißt, ein enger Zusammenhang zwischen Navigationsverhalten und Itembeantwortung ist vor allem bei kürzeren Texten zu beobachten; bei längeren Texten hingegen verschwindet der Effekt und kann sich in sein Gegenteil kehren.

Hinsichtlich länderspezifischer Unterschiede stehen die asiatischen Länder mit ihrem spezifischen Gewichtungsprofil den Spanisch sprechenden Ländern und Australien und Neuseeland oder den skandinavischen Ländern gegenüber. In den skandinavischen und Spanisch sprechenden Ländern sowie in Australien und Neuseeland haben die allgemeine Lesefähigkeit und das Navigationsverhalten ein stärkeres Gewicht für die Vorhersage der Itembeantwortung als in den asiatischen Ländern.

Durch längere Texte wird der Einfluss des tatsächlichen Navigationsverhaltens auf die Itembearbeitung verringert. In den asiatischen Ländern fällt dieser moderierende Einfluss der Textlänge am geringsten aus – im Gegensatz zu den skandinavischen Ländern, Chile, Spanien und Australien.

Über die Ursachen dieser Unterschiede können wir an dieser Stelle nur spekulieren. Zur Interpretation solcher kulturellen Unterschiede kommen unterschiedliche Ansätze infrage. Zum einen dürften sich die schulischen Curricula sowie die Lesegewohnheiten von Land zu Land unterscheiden. Zum anderen könnten aber auch Unterschiede in den jeweiligen Sprach- und Schriftsystemen eine Rolle spielen, wodurch die kognitiven Anforderungen beim Lesen elektronischer Texte jeweils unterschiedlich ausfallen. Zur Aufklärung dieser Fragen könnten kognitionswissenschaftliche, insbesondere kognitiv-linguistische Untersuchungen, einen wesentlichen Beitrag leisten.

Insgesamt gesehen hat die Analyse der vorliegenden Daten deutlich gemacht, dass das Lesen und Verstehen elektronischer Texte sehr komplexen Interaktionen zwischen Lesermerkmalen, Text- und Itemmerkmalen sowie den jeweiligen Eigenschaften des Informationsraums unterliegt. Um diese Interaktionen weiter aufzuklären, bietet sich ein interdisziplinäres Vorgehen an, bei dem Kognitionspsychologie, Lehr-Lern-Forschung, Kognitive Linguistik sowie Methodologie des Large-Scale-Assessments eng zusammenarbeiten und so zur Erweiterung der jeweiligen Forschungsperspektive beitragen.

Literaturverzeichnis

Cai, Z., McNamara, D. S., Louwerse, M., Hu, X., Rowe, M. & Graesser, A. C. (2004). NLS: Non-latent similarity algorithm. In K. Forbus, D. Gentner & T. Regier (Hrsg.), *Proceedings of the 26th Annual Cognitive Science Society* (S. 180–185). Mahwah, NJ: Erlbaum.

Cress, U. & Knabel, O. B. (2003). Previews in hypertext: Effects on navigation and knowledge acquisition. *Journal of Computer Assisted Learning, 19,* 517–527.

DeStefano, D. & LeFevre, J.-A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior, 23,* 1616–1641.

Graesser, A. C., McNamara, D. S. & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36,* 193–202.

Graesser, A. C., Millis, K. K. & Zwaan, R. A. (1997). Discourse comprehension. *Annual review of psychology, 48 (1),* 163–189.

Herder, E. & Juvina, I. (2004). Discovery of individual user navigation styles. In G. D. Magoulas & S. Y. Chen (Hrsg.), *Adaptive hypermedia AH 2004 workshop on individual differences in adaptive hypermedia.* Eindhoven: Springer.

Johnson-Laird, P. N. (1983). Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge, MA: Harvard University Press.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge University Press.

Landauer, T., McNamara, D. S., Dennis, S. & Kintsch, W. (2007). *Handbook of latent semantic analysis.* Mahwah, NJ: Erlbaum.

Lawless, K. A. & Schrader, P. G. (2008). Where do we go now? Understanding research on navigation in complex digital environments. In D. J. Leu & J. Coiro (Hrsg.), *Handbook of New Literacies* (S. 267–296). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lin, D.-H. M. (2003). Hypertext for the aged: Effects of text topologies. *Computers in Human Behavior, 19,* 201–209.

Lorch, R. F. (1989). Text-signaling devices and their effects on reading and memory processes. *Review of Educational Research, 1,* 209–234.

Lorch, R. F. & Lorch, E. P. (1995). Effects of organizational signals on text-processing strategies. *Journal of Educational Psychology, 87,* 537–544.

McDonald, S. & Stevenson, R. J. (1998a). Navigation in hyperspace: An evaluation of the effects of navigational tools and subject matter expertise on browsing information retrieval in hypertext. *Interacting With Computers, 10,* 129–142.

McDonald, S. & Stevenson, R. J. (1998b). Effects of text structure and prior knowledge of the learner on navigation in hypertext. *Human Factors, 40,* 18–27.

Naumann, J. (2010). *Log file analysis in hypertext research: An overview and two a meta-analyses.* Manuscript under revision (Educational Psychology Review, 2nd revision).

Naumann, J., Richter, T., Flender, J., Christmann, U. & Groeben, N. (2007). Signaling in expository hypertexts compensates for deficits in reading skill. *Journal of Educational Psychology, 99,* 791–807.

Organisation for Economic Co-operation and Development (2009a). *PISA 2009 Assessment Framework. Key competencies in reading, mathematics, and science.* Paris: OECD.

Organisation for Economic Co-operation and Development (2009b). *PIAAC Literacy: A Conceptual Framework.* (OECD Education Working Paper No. 34). Paris: OECD.

Organisation for Economic Co-operation and Development (2009c). *PIAAC Numeracy: A Conceptual Framework* (OECD Education Working Paper No. 35). Paris: OECD.

Organisation for Economic Co-operation and Development (2009d). *PIAAC Problem Solving in Technology-Rich Environments: A Conceptual Framework* (OECD Education Working Paper No. 36). Paris: OECD.

Organisation for Economic Co-operation and Development (im Druck). *PISA 2009 results Vol. 6: Students on line – reading and using digital information.* Paris: OECD.

Organisation for Economic Co-operation and Development (2012). *PISA 2009 Technical Report.* Paris: OECD. doi: 10.1787/9789264167872-en.

Richter, T., Naumann, J., Brunner, M. & Christmann, U. (2005). Strategische Verarbeitung beim Lernen mit Text und Hypertext. *Zeitschrift für Pädagogische Psychologie, 19,* 5–22.

Richter, T., Naumann, J. & Noller, S. (2003). LOGPAT: A semi-automatic way to analyze hypertext navigation behavior. *Swiss Journal of Psychology, 62,* 113–120.

Rouet, J. F. (2006). The skills of document use: From text comprehension to Web-based learning. Mahwah, NJ: Erlbaum.

Rouet, J. F. & Passerault, J. M. (1999). Analyzing learner-hypermedia interaction: An overview of online methods. *Instructional Science, 27 (3/4),* 201–219.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen.* Weinheim: Beltz.

Schroeder, E. E. & Grabowski, B. L. (1995). Patterns of exploration and learning with hypermedia. *Journal of Educational Computing Research, 13,* 313–335.

Shapiro, A. M. (1998). Promoting active learning: The role of system structure in learning from hypertext. *Human-Computer-Interaction, 13,* 1–35.

Shapiro, A. M. & Niederhauser, D. (2004). Learning from hypertext: Research issues and findings. In D. H. Jonassen (Hrsg.), *Handbook of research on educational communications and technology* (2. Auflage, S. 605–620). Mahwah, NJ: Erlbaum.

Smith, P. A. (1996). Towards a practical measure of hypertext usability. *Interacting with Computers, 4, 365–381.*

Christoph Niepel, Julia Rudolph, Frank Goldhammer, Samuel Greiff

Die Rolle transversaler Kompetenzen für schulisches Lernen: das Beispiel des komplexen Problemlösens

1 Problemlösen als transversale Kompetenz und seine Rolle in internationalen Large-Scale-Assessments

In den letzten beiden Dekaden wurde eine Reihe von Large-Scale-Assessments (LSA) auf internationaler Ebene initiiert. Zentrale Aufgabe solcher LSA ist das Bildungsmonitoring – untersucht wird also, ob und inwieweit politisch gesetzte Bildungsziele erreicht werden konnten bzw. in welchen zentralen Kompetenzbereichen besondere Stärken oder Schwächen zu beobachten sind. Konzipiert als groß angelegte Befragungen und Testungen von Schülerinnen und Schülern, geben LSA empirisch fundiertes Feedback zu Bildungssystemen auf nationaler und – zum Teil – internationaler Ebene. Dabei hatten und haben LSA einen großen Einfluss auf die Bildungspolitik und Bildungsforschung. Da LSA vornehmlich bildungspolitisch ausgerichtet sind, sind zusätzliche begleitende Forschungsanstrengungen nötig, um die gesammelten Daten für die Bildungsforschung tiefer gehend zu erschließen – Forschungsanstrengungen wie die des LSA004-Projektes zum komplexen Problemlösen im Kontext von LSA. Im Weiteren geben wir einen kombinierten und breiteren Überblick über Befunde zum komplexen Problemlösen aus der PISA-Studie und aus dem LSA004-Projekt.

Zu den bekanntesten LSA zählt das von der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (Organisation for Economic Co-operation and Development; OECD) in Auftrag gegebene *Programme for International Student Assessment* – besser bekannt unter dem Akronym PISA. PISA fokussiert auf die Erfassung von alltags- und berufsrelevanten Kompetenzen von 15-jährigen Schülerinnen und Schülern vor dem Ende ihrer allgemeinen Schulpflicht; Kompetenzen, die als grundlegend für eine vollwertige gesellschaftliche Teilnahme angesehen werden können (www.oecd.org/pisa; für einen deutschsprachigen Überblick zu PISA: www.oecd.org/berlin/themen/pisa-internationaleschulleistungsstudiederoecd.htm). Neben Kompetenzen werden dabei auch demografische Angaben sowie sozioaffektive Maße (wie z. B. Interesse oder Einstellungen) erfasst. Diese umfangreichen Fragebogen- und Testbatterien dienen dadurch nicht nur für Vergleiche verschiedener Bildungssysteme untereinander, sondern ermöglichen darüber hinaus Vergleiche verschiedener demografischer Gruppen innerhalb der teilnehmenden Länder und auch tiefer gehende wissenschaftliche Analysen – z. B. zum Wechselspiel zwischen Leistung und motivationalen Variablen.

PISA ist damit (neben anderen LSA) zu einem bedeutenden Impulsgeber für Politik und Wissenschaft avanciert. Der Einfluss auf die Bildungspolitik ist augenfällig; erinnert sei hier nur an die bundesdeutsche Debatte im Zuge des sogenannten PISA-Schocks Anfang der 2000er-Jahre, welcher die gesamtgesellschaftliche Bedeutung von Bildung sowie moderner Bildungspolitik nachhaltig in den Blickpunkt der öffentlichen Diskussion gerückt hat. Nachdem im Jahr 2000 die Leistungen deutscher Schülerinnen und Schüler in PISA weit hinter den allgemeinen Erwartungen zurückgeblieben sind (für eine deutschsprachige Zusammenfassung der zentralen Befunde von PISA 2000 siehe Artelt et al., 2001), führte dies in der Folge nicht nur zu kontrovers geführten Debatten, sondern auch zu einer Reihe konkreter Schulreformen (vgl. z. B. Smolka, 2002; Terhart, 2015). Darüber hinaus profitierte die empirische Bildungsforschung nachhaltig vom gewachsenen politischen und öffentlichen Interesse an ihrem Forschungsgegenstand. Zudem hat die bloße Verfügbarmachung von bislang kaum gekannten Datenmengen durch PISA (und anderen LSA) der empirischen Bildungsforschung enormen Auftrieb gegeben; deutlich wird dies z. B. an der über die vergangenen Jahre auszumachenden vermehrten Präsenz und Wahrnehmung bildungswissenschaftlicher Forschung in Wissenschaft und Praxis.

PISA wird in Dreijahreszyklen durchgeführt und ist eine der größten LSA weltweit: Im Jahre 2012 nahmen über 500.000 Schülerinnen und Schüler in über 60 Ländern teil (Organisation for Economic Co-operation and Development [OECD], 2013a; die Daten des 2015er-Zyklus werden 2016 veröffentlicht). In jedem PISA-Zyklus werden jeweils die Grundkompetenzen in den curricularen Domänen Mathematik, Naturwissenschaften und Lesen erfasst; hierbei wird jedes Mal ein anderer Schwerpunkt gesetzt – so lag der Fokus 2012 auf Mathematik und 2009 auf Lesen. Neben der Testung dieser curricularen Kompetenzen bestand 2012 für die teilnehmenden Länder zugleich die Möglichkeit, ergänzend eine domänenübergreifende, sogenannte transversale Kompetenz zu testen; namentlich: komplexes Problemlösen als Kompetenz, interaktive und dynamische Probleme weitgehend unabhängig von ihrer konkreten inhaltlichen Einbettung zu lösen.

Die Entscheidung der OECD, PISA um Testungen zum Problemlösen als transversale Kompetenz zu ergänzen, spiegelt die gestiegene – und vermutlich weiter ansteigende – gesellschaftliche Relevanz solcher domänenübergreifenden Kompetenzen wider. Die Welt im 21. Jahrhundert ist in weiten Teilen gekennzeichnet durch eine wachsende Komplexität beruflicher und alltäglicher Herausforderungen; also durch zunehmend intransparent werdende und neuartige Situationen, sich stetig wandelnde Prozesse und, damit einhergehend, in weiten Teilen immer kürzer werdenden Halbwertszeiten von vormals angeeignetem fachspezifischen Wissen (siehe hierzu z. B. die Analyse von veränderten Arbeitsanforderungen im ausgehenden 20. Jahrhundert von Cascio, 1995). In der Arbeitswelt wirkt sich dieser anhaltende Trend bereits großflächig aus: Prozesse der Automatisierung, wie der vermehrte Einsatz von Computern am Arbeitsplatz, haben Tätigkeitsfelder vieler Berufe nachhaltig verändert. Empirisch-sozialwissenschaftliche Analysen bestätigen diese Entwicklung. So haben die Ökonomen Autor, Levy und Murnane (2003) Tätigkeitsbeschreibungen des US-amerikanischen Arbeitsmarktes der vergangenen Jahrzehnte analysiert. Sie kamen dabei zu dem Schluss, dass in nahezu allen Branchen, ob nun im Industrie-

sektor, in Dienstleistungsberufen oder im Bildungswesen, die Bedeutung interaktiv-kommunikativer sowie neuartiger analytischer Aufgaben über die vergangenen Jahrzehnte stark zugenommen hat. Im selben Zeitraum, so fanden die Autoren heraus, hat dagegen die relative Bedeutung von Routinetätigkeiten stark abgenommen. Neben komplexem Problemlösen werden eine Reihe anderer transversaler Kompetenzen diskutiert, deren Bedeutung in den vergangenen Jahren zugenommen hat. Zu nennen sind hier etwa selbstregulative Kompetenzen wie Zeitmanagement oder Grundkompetenzen im Umgang mit Informationstechnologien (information and communication technology [ICT] literacy). Zudem steht gerade komplexes Problemlösen, dessen Kern es schließlich ist, mit bisher nicht gekannten intransparenten Situationen umzugehen und diese zu meistern, im Zentrum der derzeit diskutierten transversalen Kompetenzen, die heute verstärkt benötigt werden, um die Herausforderungen des 21. Jahrhunderts zu bewältigen. Dementsprechend erscheint es nur konsequent, dass der Ruf nach einer stärkeren Förderung domänenübergreifender, transversaler Problemlösekompetenzen in der schulischen Ausbildung – zusätzlich zum Lernen curricularen Wissens – immer lauter wird (siehe hierzu das Positionspapier von Greiff et al., 2014).

PISA 2012, welches komplexes Problemlösen unter dem Label „kreatives Problemlösen" (creative problem solving) in seine Testungen implementierte, definiert komplexes Problemlösen als die Kompetenz, „Prozesse kognitiv zu verarbeiten, um Problemsituationen zu verstehen und zu lösen, in denen die Lösungsmethode nicht unmittelbar auf der Hand liegt" (OECD, 2014a). Dabei umfasst es „die Bereitschaft, sich mit derartigen Situationen auseinanderzusetzen, um sein Potenzial als konstruktiver und reflektierender Bürger voll auszuschöpfen" (OECD, 2014a, S. 2). Diese Kompetenz stützt sich hierbei auf komplexe kognitive Prozesse wie das Planen von Handlungssequenzen, das Treffen von Entscheidungen und den Erwerb von Wissen, die allesamt koordiniert werden müssen, um eine spezifische und komplexe Problemlösesituation zu meistern (Funke, 2010; Raven, 2000). Schülerinnen und Schüler müssen also wissen, was sie zu tun haben, aber auch, was es zu unterlassen gilt. Mayer (1998) fasst dies anhand dreier Komponenten erfolgreichen Problemlösens zusammen: die Fähigkeit (skill), die Metafähigkeit (metaskill) und der Wille (will).

Neben PISA wurde Problemlösen als transversale Kompetenz auch in anderen LSA implementiert. Als Beispiel seien hier das Adult Literacy and Life Skills Survey (ALL; Statistics Canada & OECD, 2005) oder das Programme for the International Assessment for Adult Competencies (PIAAC; OECD, 2013b) genannt. PIAAC kann hierbei als eine Erweiterung von PISA verstanden werden, erfasst es doch weltweit Grundkompetenzen von Erwachsenen zwischen 16 und 65 Jahren. Dabei ist es wichtig zu beachten, dass sich die verschiedenen LSA in ihren konkreten Konzeptualisierungen – und damit in ihrer Erfassung – von Problemlösekompetenz durchaus unterscheiden. In PISA 2012 wurden die komplexen Problemlösekompetenzen von 15-jährigen Schülerinnen und Schülern computerbasiert mittels Testitems erfasst, die großteilig auf den beiden Aufgabentypen MicroDYN und MicroFIN basierten (nähere Informationen folgen im nächsten Abschnitt). Theoretisch fußen diese Testitems auf Dörners (1986) Theorie der operativen Intelligenz. Komplexes Problemlösen wird hier

operationalisiert als die Kompetenz, sich Sachverhalte und Funktionsweisen durch aktive Informationsgenerierung und der strategischen Exploration einer intransparenten Problemsituation selbstständig anzueignen (und entspricht hiermit der oben gegebenen Definition komplexen Problemlösens). Die teststatistische Zuverlässigkeit und Gültigkeit von MicroDYN- und MikroFIN-Testitems konnten in zahlreichen empirischen Studien nachgewiesen werden (Greiff et al., 2013; Schweizer, Wüstenberg & Greiff, 2013; Sonnleitner, Brunner, Keller & Martin, 2014; Sonnleitner, Keller, Martin & Brunner, 2013; Wüstenberg, Greiff & Funke, 2012). Im folgenden Teil werden die beiden Aufgabentypen MicroDYN und MicroFIN und die Ergebnisse der PISA-Studie zum komplexen Problemlösen eingehender beschrieben.

2 Problemlösen in PISA 2012: Konzept, Aufgaben und Ergebnisse

In PISA 2012 nahm eine Subgruppe von 44 der insgesamt über 60 partizipierenden Länder, bestehend aus OECD-Mitgliedsstaaten und assoziierten Staaten, an den computerbasierten Testungen zur Problemlösekompetenz teil, die großteilig auf den beiden Aufgabentypen MicroDYN und MicroFIN basierten. Der Problemlösetest, der von insgesamt ca. 85.000 Schülerinnen und Schülern bearbeitet wurde, war komplett computerbasiert und dauerte etwa 40 Minuten, wobei immer nur eine Teilmenge der Aufgaben von allen Schülerinnen und Schülern bearbeitet wurde. Der Problemlösetest wurde in Ergänzung zu den Testungen in den klassischen Domänen Mathematik, Lesen und Naturwissenschaften durchgeführt (OECD, 2014b). Das LSA004-Projekt war eng angekoppelt an die PISA-2012-Problemlösetestungen und sollte ergänzend zu der internationalen Erhebung wichtige Forschungsfragen zum Verhalten der Messinstrumente und deren Rolle für schulisches Lernen liefern, z. B. die Frage zur Messinvarianz in ausgewählten Ländern (nähere Informationen folgen später).

So war der zentrale Gedanke bei der Entwicklung des Problemlösetests in PISA 2012, 15-jährige Schülerinnen und Schüler mit Problemstellungen zu konfrontieren, die einerseits dem Erlebensalltag der Schülerpopulation möglichst nahekommen und diesen adäquat widerspiegeln sollten, andererseits aber nicht ausschließlich über bestehendes Vorwissen oder durch spezifische Kenntnisse aus den Inhaltsdomänen lösbar sein sollten. Die gewählten Problemstellungen bewegten sich daher in unterschiedlichen Kontexten mit variierendem Fokus. So konnten die konkreten Aufgaben entweder technologische Gerätschaften und Apparate beinhalten (technology setting) oder eben auch in einem nicht-technologischem Kontext angesiedelt sein (non-technology setting). Der Fokus der Problemstellung konnte sich darüber hinaus entweder eher auf das persönliche Umfeld des Problemlösers beziehen (personal focus) oder auf das breitere soziale Umfeld oder auf gesellschaftliche Problemstellungen im Allgemeinen (social focus).

Zentral für das Problemlösekonzept war dabei, dass die einzelnen Problemstellungen zwar durchaus Elemente aus den Inhaltsbereichen Mathematik, Lesen und Naturwissenschaften beinhalten konnten, aber über diese sowie andere Inhaltsbereiche streuten und zugleich kein spezifisches Wissen aus diesen Bereichen erforderlich

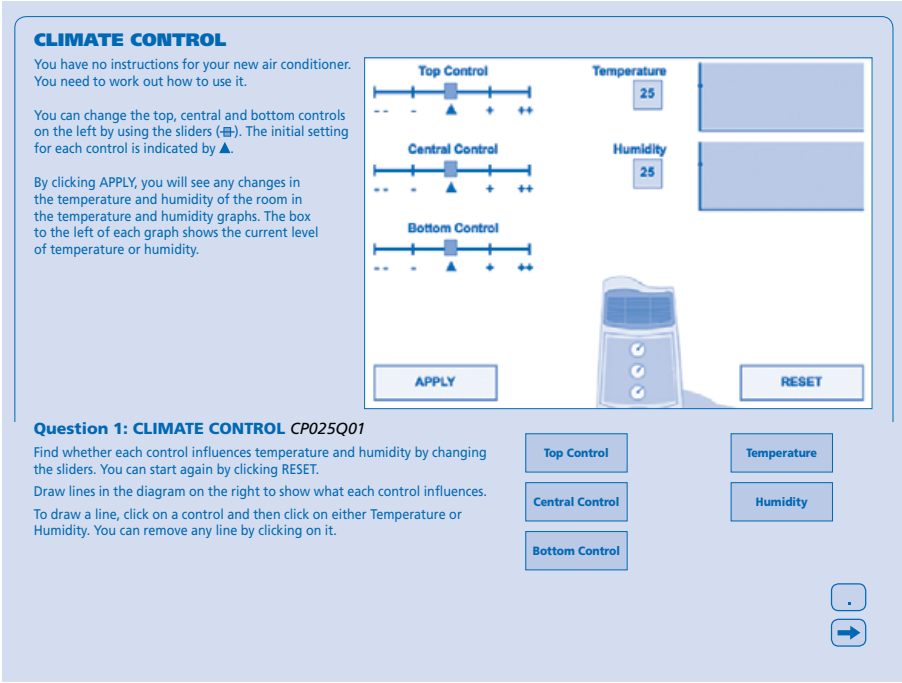
war, um die Aufgaben erfolgreich zu bearbeiten. Das Ziel dieser Konzeption war es, Problemlösekompetenz zu erfassen, die relevante Informationen über die Leistungsfähigkeit von Schülerinnen und Schülern beinhaltet, die nicht bereits in den gut und über viele Jahre etablierten Messungen in den drei zentralen Kernkompetenzen (Mathematik, Lesen und Naturwissenschaften) enthalten ist. Um das Problemlösekonzept weiter zu verdeutlichen und zu konkretisieren sollen nun zwei exemplarische Problemlöseaufgaben dargestellt werden: *Climate Control* und *Traffic*.

In der Aufgabe *Climate Control* wird die Notwendigkeit einer aktiven Informationsgenerierung und der strategischen Exploration einer intransparenten Problemsituation besonders deutlich. In Climate Control sollen die Schülerinnen und Schüler die Funktionsweise einer Klimaanlage, deren Handbuch verloren gegangen ist, durch systematisches Experimentieren und Ausprobieren erarbeiten. Dazu stehen den Schülerinnen und Schülern drei Regler zur Verfügung (top control, central control, bottom control), die sie in beliebiger Kombination variieren können. Es ist dabei Aufgabe der Schülerinnen und Schüler, die Auswirkungen dieser Reglervariationen auf die beiden Zielvariablen *Raumtemperatur* (temperature) und *Luftfeuchtigkeit* (humidity) herauszufinden und grafisch in einem Diagramm, einer sogenannten concept map, abzubilden. Ein Screenshot von Climate Control findet sich in Abbildung 1. Diese Aufgabe zeigt insbesondere die Möglichkeiten einer computerbasierten Problemlösediagnostik auf: Derart interaktive Aufgaben, in denen die Ausgangssituation zunächst intransparent ist und der Problemlöser durch gezielte Interventionen diese Intransparenz auflösen muss, lassen sich nicht über klassische Papier-Bleistift-Verfahren realisieren und stellen somit ein Novum in der PISA-2012-Erhebung dar (Greiff, Holt & Funke, 2013).

In PISA werden die Kompetenzen von Schülerinnen und Schülern sowie die Schwierigkeit der einzelnen Aufgaben auf einer gemeinsamen Skala verortet. Diese Skala wird definiert mit einem Mittelwert von 500 und einer Standardabweichung von 100. Dies bedeutet, dass eine Schülerin oder ein Schüler mit einem Wert von 500 eine durchschnittliche Problemlösekompetenz aufweist und dass eine Aufgabe mit einer Schwierigkeit von 500 eine durchschnittlich schwierige Problemlöseaufgabe ist. In *Climate Control* wurden die Schülerinnen und Schüler mit zwei zentralen Aufgaben konfrontiert: (1) Zunächst sollen sie die Verknüpfungen zwischen den Reglern auf der einen Seite und Temperatur und Luftfeuchtigkeit auf der anderen Seite erarbeiten (Wissenserwerb). Anschließend sollen sie die Regler in einer kurzen Sequenz von Interventionen so einstellen, dass vorgegebene Temperatur- und Luftfeuchtigkeitszielwerte erreicht werden (Wissensanwendung). Auf der Schwierigkeitsskala ist die erste Aufgabe des Wissenserwerbs mit einer Schwierigkeit von 523 etwa durchschnittlich schwierig, während die Wissensanwendung mit einem Schwierigkeitswert von 672 deutlich schwerer zu lösen war (und nur von wenigen Schülerinnen und Schülern gelöst werden konnte; OECD, 2014b).

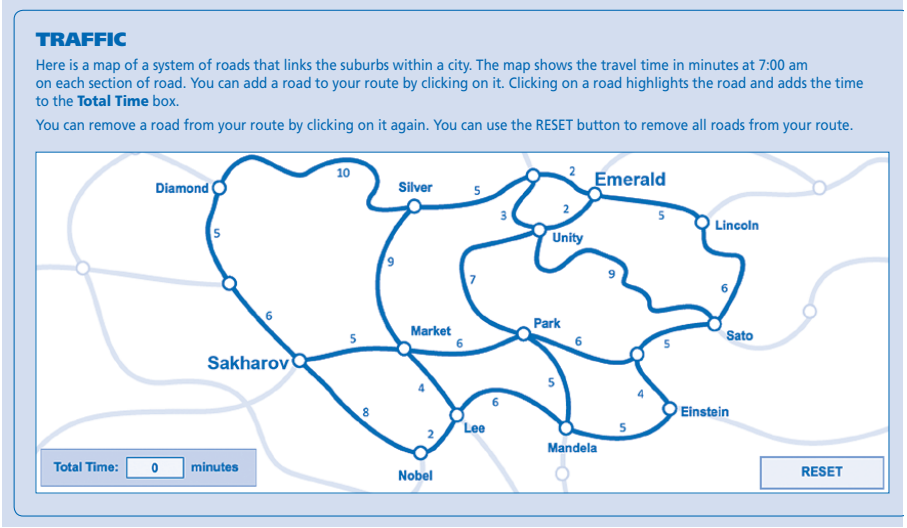
Eine weitere Veranschaulichung der Aufgaben zur Erfassung der Problemlösekompetenz von 15-jährigen Schülerinnen und Schülern in PISA 2012 ist die Aufgabe *Traffic*. In *Traffic* wird den Schülerinnen und Schülern eine Straßenkarte innerhalb einer Stadt vorgegeben, in der die Fahrzeiten zwischen den unterschiedlichen Orten angegeben sind. Die Straßenkarte mit den Fahrzeiten ist in Abbildung 2 dargestellt.

Abb. 1: Die PISA-2012-Problemlöseaufgabe Climate Control



Quelle: OECD, 2014b, S. 37 f.

Abb. 2: Die PISA-2012-Problemlöseaufgabe Traffic



Quelle: OECD, 2014b, S. 41

Wie man Abbildung 2 entnehmen kann, gibt es unterschiedliche Möglichkeiten von einem Ort (z. B. Diamond) zum anderen (z. B. Park) zu gelangen. Die Schülerinnen und Schüler werden nun aufgefordert, unterschiedliche Problemstellungen

anhand der vorliegenden Informationen zu bearbeiten, wie etwa den kürzesten Weg zwischen zwei Orten zu finden oder für mehrere Freunde, die sich treffen möchten, eine Route zu finden, die für alle eine etwa gleich lange Anreise bedeutet. In dieser Aufgabe steht nicht so sehr die Auflösung von intransparenten Zusammenhängen wie in Climate Control im Vordergrund, sondern vielmehr die Kompetenz, recht einfache Problemstellungen, in denen klare Problemvorgaben und eine begrenzte Menge von möglichen Zuständen vorgegeben sind, zu lösen. Dementsprechend liegen auch die Schwierigkeiten für die unterschiedlichen Aufgaben in Traffic zwischen 340 und 446 und damit im unterdurchschnittlichen Bereich. Bemerkenswert bei der Traffic-Aufgabe ist noch, dass ein computergesteuerter Taschenrechner (siehe untere linke Ecke in Abbildung 2) die benötigte Zeit entsprechend der gewählten Wegstrecke automatisch aufaddiert, sodass der Einfluss von Additionsfähigkeiten minimiert wird. Übrigens sind beide Aufgaben, Climate Control und Traffic, sowie einige weitere Aufgaben frei unter cbasq.acer.edu.au verfügbar und können dort ausprobiert werden, um ein Gefühl für die konkreten Aufgabenstellungen zu entwickeln.

Insgesamt beinhaltet der PISA-2012-Problemlösetest 15 Aufgaben, von denen die meisten eine aktive Exploration der Problemumgebung erfordern, einige wenige Aufgaben aber auch auf statische Problemlöseprozesse fokussieren, in denen Explorationsverhalten eine untergeordnete oder gar keine Rolle spielt (OECD, 2014b). In einem komplexen rotierten Design wurden die 15 Aufgaben administriert, sodass keiner der Schülerinnen und Schüler alle Aufgaben, sondern immer nur eine Teilmenge bearbeitete. Vergleichbar zu den Domänen Mathematik, Lesen und Naturwissenschaften werden dann auch entsprechende internationale Vergleiche von der OECD veröffentlicht, die einen Überblick über die Problemlösekompetenz 15-jähriger Schülerinnen und Schüler im internationalen Vergleich geben. Da es sich zugleich allerdings um ein recht neues Aufgabenformat und auch um eine innovative Konstruktdomäne handelt, ist die entsprechende Begleitforschung zu diesen Aufgaben, wie in dem LSA004-Projekt geschehen, von entscheidender Bedeutung, um z. B. zunächst einmal konzeptuell zu erarbeiten, welche Rolle allgemeine Problemlösekompetenzen für spezifisches schulisches Lernen spielen.

Die zentralen Ergebnisse des offiziellen OECD-Problemlösereports (OECD, 2014b) zeigen, dass ähnlich wie in den Domänen *Mathematik, Lesen* und *Naturwissenschaften* auch im Problemlösen eine Reihe asiatischer Länder im internationalen Vergleich am besten abschneiden. Ganz vorne in der Länderrangliste liegen Singapur (562 auf der genannten Skala mit Mittelwert 500 und Standardabweichung 100), Korea (561) und Japan (552). Unter den europäischen Ländern schneidet Finnland mit 523 am besten ab. Deutsche Schülerinnen und Schüler liegen mit einem Wert von 509 im oberen Mittelfeld und immerhin noch statistisch bedeutsam über dem OECD-Durchschnitt. Am unteren Ende der Problemlöseskala finden sich Kolumbien (399), Bulgarien (402) und Uruguay (403). Neben dieser quantitativen Verortung erlaubt die PISA-Skala eine Aussage darüber, wie viele Schülerinnen und Schüler sich auf unterschiedlichen inhaltlich beschriebenen Kompetenzniveaus bewegen (OECD, 2014b). Für die Problemlöseskala wurden dabei insgesamt sechs Kompetenzniveaus definiert. Schülerinnen und Schüler auf den Stufen 5 und 6 sind in der Lage, kompetent mit neuartigen komplexen Problemstellungen umzugehen und diese adäquat

zu explorieren, während Schülerinnen und Schüler auf Kompetenzstufe 1 lediglich sehr einfach strukturierte Probleme, die in der Regel keine Sequenz von aufeinanderfolgenden Schritten benötigen, lösen können.

Im Mittel erreichen 11,4 Prozent der Schülerinnen und Schüler in OECD-Mitgliedsstaaten die Kompetenzstufen 5 und 6 – aber dieser Anteil variiert beträchtlich zwischen den Ländern. Während er bei den Top-Performern bei nahezu 25 Prozent liegt, sind es z. B. in Bulgarien nur etwas über 5 Prozent der Schülerinnen und Schüler, die diese Kompetenzstufe erreichen. Umgekehrt liegt der Prozentsatz derjenigen Schülerinnen und Schüler, die lediglich Kompetenzstufe 1 oder darunter erreichen, in Singapur, Korea und Japan bei etwa 5 Prozent und in Kolumbien, Bulgarien und Uruguay bei etwa 40 Prozent. Knapp 14 Prozent der deutschen Schülerinnen und Schüler erreichen Stufe 5 oder 6 und etwa 10 Prozent liegen auf Stufe 1 oder darunter. Diese und andere Befunde – z. B. auch zum Zusammenhang zwischen sozioökonomischem Status und Problemlöseleistung oder zu Geschlechterunterschieden im Problemlösen – finden sich ausführlich im OECD-Report zu PISA 2012 Problemlösen (OECD, 2014b; für eine deutschsprachige Abhandlung zum Thema siehe auch OECD, 2014a). Eine interessante und an das LSA004-Projekt angekoppelte Arbeit zur Frage nach Leistungsunterschieden zwischen ungarischen und deutschen Schülerinnen und Schülern und die Rolle der Problemexploration in diesen Studien findet sich bei Wüstenberg, Greiff, Molnar und Funke (2014). In diesem Kontext konnte zudem die Messinvarianz der Problemlöseaufgaben zwischen deutschen und ungarischen Schülerinnen und Schülern nachgewiesen werden; ein wichtiger erster Nachweis dafür, dass komplexe Problemlösekompetenzen im Vergleich beider Länder gleichermaßen zuverlässig gemessen werden können – unabhängig von ihren nationalen Bildungssystemen.

Ein zentraler Aspekt im OECD-Report, der im Weiteren kurz aufgegriffen werden soll, ist die Frage nach dem Zusammenhang der Problemlösekompetenz auf der einen Seite und den drei klassischen Kompetenzen Mathematik, Lesen und Naturwissenschaften auf der anderen Seite, da es ja explizites Ziel der Problemlöseerhebung war, soweit möglich, andere Aspekte abzudecken als diejenigen, die sich bereits in den klassischen Kompetenzen finden. Tatsächlich war der Zusammenhang von Problemlösen zu Mathematik, Lesen und Naturwissenschaften substanziell geringer als die Zusammenhänge zwischen diesen drei Kompetenzen untereinander: Mathematik, Lesen und Naturwissenschaften zeigen eine hohe empirische Überschneidung von im Mittel 77 Prozent, d. h., dass Schülerinnen und Schüler, die gut in einem Bereich abschneiden, mit hoher Wahrscheinlichkeit auch gut in den anderen beiden Bereichen abschneiden. Für Problemlösen lag die mittlere Überschneidung bei 61 Prozent und damit deutlich niedriger.

Dieser Befund kann einerseits als Rechtfertigung für die Erhebung einer separaten Problemlösekompetenz in Studien wie PISA dienen, da sie andeuten, dass hinreichend neue Informationen über Mathematik, Lesen und Naturwissenschaften hinaus gewonnen werden, zugleich aber zeigen sich auch für das Problemlösen substanzielle Zusammenhänge zu den drei klassischen Domänen, sodass sich die Frage stellt – auch vor dem Hintergrund eines curricular ausgerichteten Bildungssystems –, was nun die Implikationen der Problemlöseresultate – auch und insbeson-

dere für Deutschland – sind. So zeigt sich also in PISA 2012 ein nicht zu vernachlässigender Zusammenhang zwischen Problemlösen auf der einen und naturwissenschaftlicher Kompetenz auf der anderen Seite. Konzeptuell ist Problemlösen, gerade vor dem Hintergrund der Bedeutsamkeit aktiver Exploration und zielgerichteter Interventionen, eng mit dem Prinzip der naturwissenschaftlichen Erkenntnisgewinnung, das in den deutschen Bildungsstandards verankert ist (Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010), verknüpft. Ein Teil dieses Zusammenhangs mag darauf zurückzuführen sein, dass Problemlösekompetenzen auch im fachspezifischen Bereich der naturwissenschaftlichen Erkenntnisgewinnung eine große Rolle spielen könnten. Diesem Zusammenhang, seinen möglichen Ursachen und damit der Bedeutung des Problemlösens für Unterricht sowie für schulisches Lernen soll im letzten Teil dieses Beitrags nachgegangen werden.

3 Komplexes Problemlösen und seine Rolle für den Kompetenzbereich der naturwissenschaftlichen Erkenntnisgewinnung

Komplexes Problemlösen hängt konzeptionell insbesondere mit Kompetenzen zusammen, die in den naturwissenschaftlichen Unterrichtsfächern Physik, Biologie und Chemie gefragt sind. Ein Grund hierfür sind die Nähe von komplexem Problemlösen und naturwissenschaftlicher Erkenntnisgewinnung: Beide erfordern, kurz gesagt, die Kompetenz, sich selbstständig Wissen durch Exploration anzueignen. Im Folgenden wird genauer dargelegt, worin die Überschneidungen und Abgrenzungen liegen und inwieweit die Überschneidungen dazu dienen können, Erkenntnisse und Forschungsmethoden aus dem Bereich der Problemlöseforschung für die Forschung zur naturwissenschaftlichen Erkenntnisgewinnung zu nutzen.

Der Begriff *naturwissenschaftliche Erkenntnisgewinnung* stellt im Rahmen der deutschen Bildungsstandards jeweils eine Dimension der Unterrichtsfächer Biologie, Chemie und Physik dar (die anderen Dimensionen sind in allen drei Unterrichtsfächern Fachwissen, Kommunikation und Bewertung; Kauertz et al., 2010). Die naturwissenschaftliche Erkenntnisgewinnung umschließt dabei eine Reihe von Regeln und Prinzipien zur Entwicklung, Anwendung und Prüfung naturwissenschaftlicher Theorien (Kauertz et al., 2010). Zentral ist hierbei – neben sehr spezifischen Fertigkeiten wie Mikroskopieren in Biologie – auch das kriterienbezogene Beobachten von naturwissenschaftlichen Phänomenen und das Experimentieren. Experimentieren erfordert zunächst ein Bewusstsein über den eigenen Wissensstand sowie die Kompetenz, Versuche systematisch durchzuführen und auszuwerten. Anschließend müssen die neu gewonnenen Erkenntnisse – insbesondere bei komplexeren Sachverhalten – verstanden und in abstraktere Modelle überführt werden. Zudem müssen die neuen Ergebnisse und Modelle in Bezug zu bestehenden naturwissenschaftlichen Theorien gesetzt werden (KMK, 2004a, 2004b, 2004c). Folglich benötigen Schülerinnen und Schüler zum Experimentieren sowohl Kompetenzen, die einen unmittelbaren Bezug zu domänenspezifischem Wissen aufweisen (wie etwa das Einordnen von Erkenntnissen in vorhandene Theorien), als auch Kompetenzen, welche unabhängig von domänenspezifischem Fachwissen vorliegen können. Hierzu zählen die mit Ex-

perimentieren verbundenen Kompetenzen, Sachverhalte systematisch zu untersuchen und die richtigen Schlussfolgerungen aus den Beobachtungen zu ziehen. Diese lassen sich durchaus fächerübergreifend bzw. fächerunabhängig erfassen. Hierzu müssten Aufgaben in Form von kleinen Experimenten eingesetzt werden, in denen Vorwissen nicht helfen würde. Durch eine solche fächerunabhängige Testung kann zudem sichergestellt werden, dass gelungenes Experimentieren nicht auf Vorwissen zurückzuführen ist. In diesem Falle könnte eine Schülerin oder ein Schüler mit tatsächlich vergleichsweise geringem Kompetenzniveau im Experimentieren trotzdem gut in Aufgaben zum Experimentieren abschneiden, eben weil sie oder er die dem Experiment zugrunde liegenden Sachverhalte und experimentellen Aufbauten bereits kennt (z. B. aus vorhergehendem Unterricht).

Ein Vergleich der naturwissenschaftlichen Erkenntnisgewinnung mit komplexen Problemlösen zeigt, dass sich komplexes Problemlösen möglicherweise gut eignet, um die von Vorwissen unabhängigen Facetten des Experimentierens zu testen. Wie weiter oben bereits beschrieben, stellt komplexes Problemlösen die Kompetenz dar, sich selbstständig Sachverhalte und Funktionsweisen durch aktive Informationsgenerierung und der strategischen Exploration einer intransparenten Problemsituation anzueignen. Darin weist komplexes Problemlösen große Schnittmengen mit Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung auf. So gilt z. B. für beide Bereiche, dass ein strategisch richtiges Vorgehen eine zentrale Rolle spielt: Nur wer mit einer (guten) Strategie an die Aufgaben herangeht, kann gute Leistungen erzielen – sowohl in der naturwissenschaftlichen Erkenntnisgewinnung als auch im Bereich des komplexen Problemlösens. Hierbei können ähnliche Strategien des Problemlösens und des Experimentierens gleichermaßen zielführend sein. So werden z. B. in beiden Kompetenzbereichen die Control of Variables Strategy (CVS; Chen & Klahr, 1999) oder auch Vary One Thing At a Time Strategy (VOTAT; Greiff, 2012) als zentrale Strategien verstanden. Beide Begriffe bezeichnen ein strategisches, schrittweises Vorgehens beim Explorieren, wobei mögliche Einflussfaktoren getrennt voneinander oder nacheinander untersucht werden. Die große Bedeutung dieses strategischen Vorgehens zeigt sich z. B. an einer klassischen Aufgabe aus der naturwissenschaftlichen Erkenntnisgewinnung (Chen & Klahr, 1999): Schülerinnen und Schüler werden gebeten herauszufinden, wodurch sich die Länge einer Metallfeder beeinflussen lässt. Hierbei stehen den Schülerinnen und Schülern Metallfedern mit unterschiedlichen Durchmessern und verschiedene Gewichte zur Verfügung, welche sich an die Metallfedern hängen lassen. Schülerinnen und Schüler, die die CVS anwenden, würden die möglichen Einflüsse des Gewichts und des Durchmessers getrennt voneinander untersuchen, indem z. B. zuerst ermittelt wird, welchen Einfluss die Gewichte haben (siehe Abbildung 3). Hierbei sollte immer die gleiche Feder genutzt werden. Erst wenn dieser Einfluss ermittelt wurde, sollte der Einfluss des Federdurchmessers untersucht werden. Hierbei sollte stets das gleiche Gewicht an die Federn gehängt werden. Eine ähnliche strategische Vorgehensweise zeigt sich bei der oben beschriebenen PISA-Aufgabe *Climate Control* (siehe Abbildung 1) als zielführend: Um die Auswirkungen zu explorieren, die das Verstellen eines bestimmten Reglers hat, sollte einer nach dem anderen verändert werden; ansonsten kann nicht sicher

abgeleitet werden, auf welchen Regler die Veränderungen in den Zielvariablen Raumtemperatur (temperature) und Luftfeuchtigkeit (humidity) zurückzuführen ist. Insgesamt zeigen die Aufgaben aus beiden Bereichen (naturwissenschaftliche Erkenntnisgewinnung und komplexes Problemlösen), dass Schülerinnen und Schüler, die eine solche Strategie (CVS/VOTAT) nicht anwenden (können), sowohl beim Durchführen von experimentellen Versuchen als auch beim Lösen komplexer Probleme eher schlecht abschneiden. Diese Ähnlichkeit deutet an, dass jeweils der eine Forschungsbereich (z. B. zum komplexen Problemlösen) großes Potenzial für den jeweils anderen (z. B. zur naturwissenschaftlichen Erkenntnisgewinnung) birgt. Ein verstärkter Austausch von Erkenntnissen und Ansätzen erscheint hier vielversprechend, um die Forschung in beiden Bereichen substanziell voranzubringen. Das LSA004-Projekt hat sich tiefer gehend damit auseinandergesetzt, inwiefern die Forschung zum komplexen Problemlösen auch einen Mehrwert für die bildungswissenschaftliche Forschung zu naturwissenschaftlichen Fächern bieten könnte, und weiterhin, wie diese adaptiert werden müsste, um ihren Nutzen im Bereich der Kompetenzmessung noch zu erhöhen.

So könnte die Forschung zu naturwissenschaftlicher Erkenntnisgewinnung durch computerbasierte Ansätze der Kompetenzmessung profitieren, wie sie auch im Kontext der PISA-Studie zum Messen komplexer Problemlösekompetenz eingesetzt wurden. In solchen MicroDYN- und MicroFIN-Testitems (wie oben beschrieben) bearbeiten Schülerinnen und Schüler ansprechend aufbereitete interaktive Aufgaben am Computer, wie die beschriebene Aufgabe Climate Control (siehe Abbildung 1). Diese Aufgaben sind so konzipiert, dass die Aufgaben nicht leichter zu lösen sind, wenn die Schülerinnen oder Schüler Vorwissen haben. Anders als z. B. bei der vorgestellten Metallfeder-Aufgabe haben bei diesen Aufgaben Schülerinnen und Schüler mit viel naturwissenschaftlichem Vorwissen keinen Vorteil bei Testungen zum strategischen Vorgehen. Tests, in denen Vorwissen nicht hilft, ermöglichen eine genaue Erfassung davon, welche Stärken und Schwächen eine Schülerin oder ein Schüler hat. So könnte bei der Metallfeder-Aufgabe möglicherweise nicht genau auseinandergehalten werden, ob Schülerinnen und Schüler gut bei dieser Aufgabe abschneiden, weil sie bereits wissen, wie sich die Zusammenhänge gestalten, oder weil sie hinreichend gut experimentieren. Da MicroDYN- und MicroFIN-Testitems so konzipiert wurden, dass der Einfluss von etwaigem Vorwissen auf die Testleistung minimiert wurde, sind diese wohl besonders gut geeignet, um zu messen, wie gut Schülerinnen und Schüler in der Lage sind, Zusammenhänge aufzudecken, ohne dass das Fachwissen – welches eine weitere eigenständige Komponente der Bildungsstandards abbildet – diese Testung beeinflusst.

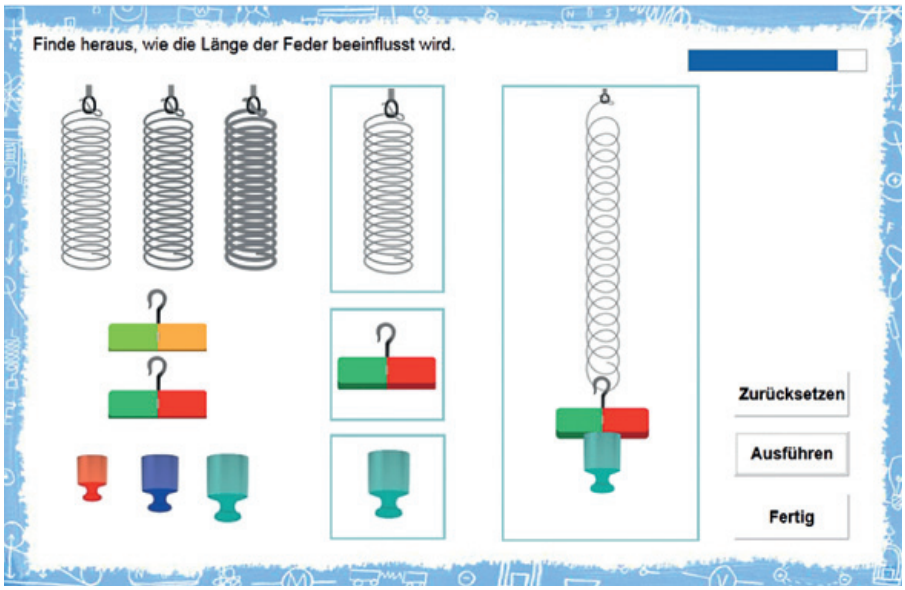
Weitere Potenziale für beide Forschungsstränge birgt eine Zusammenführung von naturwissenschaftlichem Experimentieren und komplexem Problemlösen, indem MicroDYN- und MicroFIN-Aufgabentypen in einen naturwissenschaftlichen Kontext eingebettet werden. Solche naturwissenschaftlichen Problemlöseaufgaben könnten z. B. eine Reihe an kleineren Experimenten abbilden (wie die Metallfeder-Aufgabe; siehe Abbildung 3). Die bisherigen Erfahrungen und technische Methoden bei der Entwicklung von Problemlöseitems können genutzt werden, um computerbasiert naturwissenschaftliche Problemlöseaufgaben zu entwickeln, die ebenfalls

das strategische Vorgehen beim Experimentieren automatisch erfassen können. Diese Aufgaben haben den Vorteil, dass sie unmittelbar naturwissenschaftliche Sachverhalte darstellen können und Schülerinnen und Schüler tatsächlich existierende Experimente am Computer direkt selbst durchführen können. Wie oben bereits erwähnt, hängt die Bearbeitung solcher Aufgaben unter anderem von dem Vorwissen der Schülerinnen und Schüler ab. Daher eignen sich solche Aufgaben insbesondere, um zu untersuchen, wie Vorwissen das strategische Verhalten der Schülerinnen und Schüler beeinflusst. Möglicherweise fällt es Schülerinnen und Schülern besonders leicht, ein gutes strategisches Vorgehen zu erlernen und die CVS-Strategie zu verinnerlichen, wenn sie ein Experiment durchführen, zu dem sie bereits Vorwissen besitzen, und können dann das erlernte strategische Vorgehen besser auf neue experimentelle Fragestellungen übertragen.

Ein weiterer Nutzen der beschriebenen computerbasierten Problemlöseaufgaben besteht darin, dass sie sich für den Unterricht weiterentwickeln lassen, um z. B. Schülerinnen und Schülern zu vergegenwärtigen, dass (gleiche oder ähnliche) Strategien in unterschiedlichen Kontexten (z. B. in verschiedenen PISA-Problemlöseaufgaben) angewandt werden können. Zudem wäre es auch in diesem Kontext möglich, auf Basis bisheriger Forschungsarbeiten und Erfahrungen zum komplexen Problemlösen ein individuelles, automatisches Feedback zu programmieren. Die bisher entwickelte Testsoftware kann automatisch erkennen, wenn Schülerinnen und Schüler keine oder schlechte Strategien anwenden. Diese könnte ausgebaut werden, sodass eine individuelle Rückmeldung erscheint und der Schülerin oder dem Schüler Hilfestellungen geboten werden, um ihr oder sein Explorationsverhalten zu optimieren. Wenn z. B. eine Schülerin bei Climate Control (siehe Abbildung 1) mehrere Regler gleichzeitig verschiebt, könnte automatisch ein Video erscheinen, in dem eine Lehrkraft (oder eine Gleichaltrige/ein Gleichaltriger) das Prinzip der CVS-Strategie erklärt. Ein solcher möglicher Ausbau der Problemlöseaufgaben zeigt, wie Forschung zu naturwissenschaftlichem Unterricht die Forschung zum komplexen Problemlösen und ihre Methoden nutzen kann, um herauszufinden, wie Lernsoftware zur Vermittlung von guten Strategien optimal gestaltet werden könnte.

Im LSA004-Projekt wurde der Grundstein für die Entwicklung solcher Aufgaben gelegt und die Machbarkeit einer solchen Idee geprüft. Derzeit werden solche Aufgaben an der Universität Luxemburg in der Arbeitsgruppe Computer Based Assessment (Leitung: Samuel Greiff) weiterentwickelt, um das Potenzial der Forschung zum komplexen Problemlösen für die Forschung zur naturwissenschaftlichen Erkenntnisgewinnung nutzbar zu machen und um den Einfluss von Vorwissen auf das Explorationsverhalten der Schülerinnen und Schüler zu untersuchen.

Abb. 3: Aufgabenbeispiel einer interaktiven und computerbasierten naturwissenschaftlichen Problemlöseaufgabe



Quelle: Rudolph, Niepel, Martin und Greiff (2015)

4 Schlussbemerkung

Problemlösen als transversale Kompetenz spielt eine zentrale Rolle für schulisches Lernen. Zugleich wird in einer sich immer schneller wandelnden Welt die Bedeutung komplexer Problemlösekompetenzen für die Bewältigung moderner Anforderungen im Alltag und Berufsleben voraussichtlich noch weiter zunehmen. Individualdiagnostische Ansätze zur zuverlässigen und validen Erfassung solcher Kompetenzen tief greifend zu verstehen und diese weiterzuentwickeln ist daher ein eminent wichtiges Ziel; nicht zuletzt, um wirksame Interventionen zur Förderung komplexer Problemlösekompetenzen im konkreten schulischen Alltag zu entwickeln. Zudem können Messmethoden zur Erfassung transversaler Problemlösekompetenzen auch fachspezifische, naturwissenschaftliche Testungen bereichern. So überschneidet sich komplexes Problemlösen sehr stark mit Anforderungen der naturwissenschaftlichen Erkenntnisgewinnung, welche nach den deutschen Bildungsstandards einen wesentlichen Bestandteil der Kompetenzen in den Unterrichtsfächern Biologie, Chemie und Physik ausmacht. Weiter gehend können Aufgaben zur Erfassung komplexen Problemlösens auch zu kleinen naturwissenschaftlichen Experimenten ausgebaut werden, die Schülerinnen und Schüler am Computer lösen können. Solche Aufgaben könnten z. B. in zukünftigen LSA eingesetzt werden. Außerdem könnte darauf aufbauend eine Software entwickelt werden, die es Lehrkräften vereinfacht, Kompetenzen, die hinter der naturwissenschaftlichen Erkenntnisgewinnung stehen, zu vermitteln.

Literaturverzeichnis

Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Schümer, G., Stanat, P., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2001). *PISA 2000. Zusammenfassung zentraler Befunde.* Berlin: Max-Planck-Institut für Bildungsforschung.

Autor, D. H., Levy, F. & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics, 118,* 1279–1333.

Cascio, W. F. (1995). Whither Industrial and Organizational Psychology in a changing world of work? *American Psychologist, 50,* 928–939.

Chen, Z. & Klahr, D. (1999). All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy [Alles andere bleibt gleich: Erlernen und Transfer der Variablen-Kontroll-Strategie]. *Child Developoment, 70 (5),* 1098–1120.

Dörner, D. (1986). Diagnostik der operativen Intelligenz. *Diagnostica, 32,* 290–308.

Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing, 11,* 133–142.

Greiff, S. (2012). *Individualdiagnostik komplexer Problemlösefähigkeit.* Münster: Waxmann.

Greiff, S., Holt, D. V. & Funke, J. (2013). Perspectives on problem solving in cognitive research and educational assessment: analytical, interactive, and collaborative problem solving. *Journal of Problem Solving, 5,* 71–91.

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C. & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review, 13,* 74–83.

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J. & Csapó, B. (2013). Complex problem solving in educational contexts – Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105,* 364.

Kauertz, A., Fischer, H. E., Mayer, J., Sumflet, E. & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften, 16,* 135–153.

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science, 26,* 49–63.

Organisation for Economic Co-operation and Development (2013a). *PISA 2012 assessment and analytical framework.* Paris: OECD.

Organisation for Economic Co-operation and Development (2013b). *OECD skills outlook 2013: First results from the survey of adult skills.* Paris: OECD.

Organisation for Economic Co-operation and Development (2014a). *PISA im Fokus: Sind 15-Jährige kreative Problemlöser?* Paris: OECD.

Organisation for Economic Co-operation and Development (2014b). *PISA 2012 results. Creative problem solving.* Paris: OECD.

Raven, J. (2000). Psychometrics, cognitive ability, and occupational performance. *Review of Psychology, 7,* 51–74.

Rudolph, J., Niepel, C., Martin, R. & Greiff, S. (2015). Die Erfassung naturwissenschaftlicher Kompetenzen bei Luxemburger Schülerinnen und Schülern. In Ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse, SCRIPT & Université du Luxembourg, FLSHASE (Hrsg.), *Bildungsbericht Luxembourg 2015. Bd. 2 Analysen und Befunde* (S. 8–14). Luxemburg: Ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse.

Schweizer, F., Wüstenberg, S. & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences, 24,* 42–52.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (Hrsg.). (2004a). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss.* München: Neuwied.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (Hrsg.). (2004b). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss.* München: Neuwied.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (Hrsg.). (2004c). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss.* München: Neuwied.

Smolka, D. (2002). Die PISA-Studie: Konsequenzen und Empfehlungen für Bildungspolitik und Schulpraxis. *Aus Politik und Zeitgeschichte, 41,* 3–11.

Sonnleitner, P., Brunner, M., Keller, U. & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology, 106,* 681.

Sonnleitner, P., Keller, U., Martin, R. & Brunner, M. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence, 41,* 289–305.

Statistics Canada & Organisation for Economic Co-operation and Development (2005). *Learning a living: First results of the Adult Literacy and Life Skills Survey.* Paris: OECD.

Terhart, E. (2015). Wie geht es weiter mit der Qualitätssicherung im Bildungssystem – 15 Jahre nach PISA? *Aus Politik und Zeitgeschichte, 65,* 3–10.

Wüstenberg, S., Greiff, S. & Funke, J. (2012). Complex problem solving – More than reasoning? *Intelligence, 40,* 1–14.

Wüstenberg, S., Greiff, S., Molnar, G. & Funke, J. (2014). Determinants of cross-national gender differences in complex problem solving competency. *Learning and Individual Differences, 29,* 18–29.

Richard Göllner, Wolfgang Wagner, Eckhard Klieme, Oliver Lüdtke, Benjamin Nagengast, Ulrich Trautwein

Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen: Chancen, Grenzen und Forschungsperspektiven

1 Einleitung

Schülerurteile stellen eine wichtige Datenquelle zur Messung lern- und leistungsförderlicher Qualitätsaspekte des Unterrichts dar (Clausen, 2002; Klieme, Schümer & Knoll, 2001; Fraser & Walberg, 1991; Seidel & Shavelson, 2007). Schülerinnen und Schüler gelten als Experten des Unterrichts. Sie sind in der Lage, Vergleiche mit anderen Lehrkräften anzustellen, sie können ihr Urteil auf einen ausgedehnten Beobachtungszeitraum stützen, und sie haben die Möglichkeit, selbst über seltene Ereignisse im Unterricht Auskunft zu geben. Zudem können mit Schülerbefragungen eine Vielzahl von Beurteilern im Hinblick auf einen Beurteilungsgegenstand befragt und somit eine hohe Informationsdichte (und gegebenenfalls -vielfalt) erzielt werden. Gerade in Hinblick auf große Schulleistungsstudien wie PISA, IGLU oder TRAIN sind Schülerurteile von potenziell hoher Bedeutung, da ihre Erfassung relativ kostengünstig ist und wenig Zeitressourcen in Anspruch nimmt (Clausen, 2002; Lüdtke, Trautwein, Kunter & Baumert, 2006).

Aber wie gut sind diese Schülerurteile? Nach wie vor ist zu wenig über die psychometrische Güte (wie Reliabilität und Validität) von Schülerurteilen über den Unterricht bekannt. Können Schülerinnen und Schüler theoretisch distinkte Facetten des Unterrichtsgeschehens zuverlässig und zutreffend einschätzen? Inwiefern sind die Einschätzungen von Schülerinnen und Schülern über unterschiedliche Kontexte (z. B. Unterrichtsfächer oder Schulklassen) hinweg vergleichbar? Inwiefern beeinflusst die sprachliche Gestaltung der Items den Beurteilungsprozess? Diese bislang wenig untersuchten Fragen standen im Mittelpunkt des durch das BMBF geförderten Projekts „Erfassung der Unterrichtsqualität in Large-Scale-Studien: Optimierung der Modellierung und Itemauswahl“ (01LSA008; Trautwein, Lüdtke, Klieme, Nagengast & Wagner, 2011). In dem vorliegenden Beitrag werden zentrale Ergebnisse des Projekts vorgestellt. Zunächst wird ein allgemeiner Überblick über die Erfassung der Unterrichtsqualität gegeben, um dann genauer auf Chancen und Grenzen von Schülerbeurteilungen des Unterrichts anhand von empirischen Ergebnissen des Projekts einzugehen.

2 Unterrichtsqualität und ihre Erfassung

Aktuelle Konzepte der Unterrichtsqualitätsforschung betrachten das Unterrichtsgeschehen als ein Zusammenspiel von Merkmalen, welche die Lerngeschichte von Schülern nachhaltig beeinflussen können (Hattie, 2009; Helmke, 2010; Klieme et al., 2008; Scheerens & Bosker, 1997). Inzwischen liegen sowohl national als auch international empirisch begründete Beschreibungssysteme relevanter Merkmale vor. Diese umfassen sogenannte Basismerkmale der Unterrichtsqualität, welche dann auf untergeordneten Hierarchien das konkrete Lehrerverhalten beschreiben. Zu den in der Literatur genannten Basismerkmalen bzw. Dimensionen gehören die Klassenführung, die Schülerorientierung und die kognitive Aktivierung. Diesen Modellen zufolge ist ein Unterricht lern- und leistungsförderlich, der klar strukturiert ist, die zur Verfügung stehende Lernzeit nutzt und dabei kognitiv anregend und individuell unterstützend gestaltet ist (Hamre & Pianta, 2010; Klieme, Schümer & Knoll, 2001).

Die Erfassung der Unterrichtsqualität erfolgt in der Unterrichtsforschung meist anhand von Schülerbeurteilungen, Selbstberichten der Lehrkräfte oder Beobachtungen durch externe Beobachter, die entweder unmittelbar in der Klasse oder auf der Basis von Videoaufnahmen eine Beurteilung vornehmen bzw. Unterrichtskonzepte oder Unterrichtsmaterialien bewerten (Clausen, 2002; Fraser & Walberg, 1991). Alle genannten Zugänge weisen ihre spezifischen Vor- und Nachteile auf, insbesondere dann, wenn es um ihren Einsatz in Large-Scale-Studien geht. So gelten Beobachtungen durch Experten zwar allgemein als der „Goldstandard" in der Lehr-Lern-Forschung, sie sind aber mit sehr hohem zeitlichen und finanziellen Aufwand verbunden, und die Experteneinschätzungen entsprechen nicht immer den üblichen psychometrischen Standards. Insbesondere die Erfassung sogenannter hoch inferenter Beobachtungen erfordert ein hohes Maß an Schulung der Beurteiler. Hoch inferente Beobachtungen verlangen von den Beobachtern, über das konkret beobachtete Verhalten hinaus auf abstrakte Sachverhalte oder allgemeine Verhaltenstendenzen zu schließen. Demgegenüber beschränken sich die sogenannten niedrig inferenten Beurteilungen auf spezifische, beobachtbare Verhaltensweisen, welche einfach und objektiv zu codieren sind (Rosenshine, 1970). Die Codierung und Auswertung einer Unterrichtsstunde beansprucht häufig ein Vielfaches der eigentlichen Beobachtungsdauer (Fauth, Decristan, Rieser, Klieme & Büttner, 2014). Trotz dieses enormen Aufwands wird eine zufriedenstellende Messgüte in vielen Fällen nur über eine Anpassung der untersuchten Unterrichtsmerkmale erreicht (Derry et al., 2010).

Deutlich weniger aufwendig ist die Verwendung von Lehrerselbstberichten (Clausen, 2002; Desimone, Smith & Frisvold, 2010). Selbstberichte stellen eine gut umsetzbare Möglichkeit dar, lern- und leistungsrelevante Qualitätsmerkmale des Unterrichts zu erfassen. Selbstberichte von Lehrkräften haben in den vergangenen Jahren wieder an Bedeutung gewonnen und sind inzwischen aus dem Forschungsfeld der Unterrichtsqualität, der Lehrerexpertise und des Professionswissens (Fachwissen, pädagogisches Wissen und fachdidaktisches Wissen) nicht mehr wegzudenken (Baumert & Kunter, 2006; Blömeke, 2004). Lehrkräfte sollten aufgrund ihrer beruflichen Qualifikation auch komplexe Sachverhalte ihres beruflichen Handelns beschreiben können. Allerdings weisen empirische Befunde der vergangenen Jahre auch auf eine Vielzahl von „Pro-

blemen“ hin (Clausen, 2002; Wubbels, Brekelmans & Hooymayers, 1992). Es findet sich eine nur geringe Übereinstimmung von Lehrerselbstberichten mit anderen Datenquellen (z. B. Beurteilungsdaten) und nur wenige Hinweise auf eine prädiktive Validität für relevante Zielkriterien des Unterrichts. Erklärt wird dies durch die Involviertheit der Lehrkraft in den täglichen Unterricht, die zu selbstwertschützenden Verzerrungen oder auch sozial erwünschten Beurteilungstendenzen führen kann (Clausen, 2002). Zudem ist aus dem Bereich der Persönlichkeitspsychologie bekannt, dass nicht alle Aspekte des eigenen Verhaltens einer Selbstbeschreibung in gleicher Weise zugänglich sind und entsprechend durch außenstehende Personen akkurater beschrieben werden können (z. B. Funder, 2001; Vazire & Solomon, 2015).

Eine dieser Außenperspektiven und die wohl am häufigsten eingesetzte Methode der Datengewinnung zur Erfassung der Unterrichtsqualität in Large-Scale-Assessments sind Schülerbeurteilungen des Unterrichts. Üblicherweise werden hierbei Fragebogenitems für theoretisch distinkte Dimensionen der Unterrichtsqualität verwendet, die keinen konkreten Stundenbezug aufweisen, sondern rückblickend auf einen längeren Unterrichtszeitraum (z. B. das vergangene Schuljahr) nach Merkmalen des Unterrichts fragen. Aus forschungspraktischer Perspektive stellen Schülerbeurteilungen des Unterrichts eine höchst effektive Form der Datengewinnung dar (Clausen, 2002; Fraser & Walberg, 1991). Neben dem geringeren Erhebungsaufwand und der ökonomischen Durchführung ist insbesondere die hohe Reliabilität von Schülerdaten aufgrund der Zusammenfassung mehrere Einzelurteile zu nennen (Lüdtke, Trautwein, Kunter & Baumert, 2006). Doch auch die Verwendung von Schülerurteilen wird mitunter kritisch gesehen. Schülerinnen und Schüler sind nicht nur Beurteiler des Unterrichts, sondern auch stark involvierte Akteure. Zudem verfügen sie trotz ihrer Erfahrungen nicht über eine didaktisch-pädagogische Expertise im engeren Sinne. Empirische Studien der frühen 1990er-Jahre zeigten beispielsweise, dass das Ausmaß einer positiven Beurteilung der Lehrkraft bzw. des Unterrichts im Kontext universitärer Lehre sowohl vom Geschlecht, dem Leistungsstand, dem Interesse oder auch individuellen Urteilstendenzen (z. B. Milde-Strenge-Effekte) der einzelnen Schülerinnen und Schüler abhängig ist (z. B. Gigliotti & Buchtel, 1990; Greenwald & Gillmore, 1997; Marsh & Roche, 1997). Einige Arbeiten weisen für die Schülersicht sogenannte Halo-Effekte nach, die als Überstrahlungseffekte zu einer weniger detaillierten Beschreibung verschiedener Qualitätsdimensionen führen können (Clausen, 2002; Fisicaro & Lance, 1990).

Darüber hinaus ist die Beantwortung der Frage, inwieweit Schülerinnen und Schüler die Unterrichtsqualität einer Lehrkraft zuverlässig und zutreffend beurteilen können, erheblich durch die Vielzahl an vorliegenden Instrumenten erschwert. Bereits eine oberflächliche Betrachtung der Items aus den bekannten Large-Scale-Studien (z. B. PISA, KESS, COACTIV, BIJU, DESI usw.) zeigt, dass sich die eingesetzten Fragebogeninstrumente nicht nur im Hinblick auf die Auswahl und Operationalisierung der Qualitätsdimensionen unterscheiden, sondern auch eine erstaunlich hohe Variabilität der konkreten Itemformulierungen aufweisen. Welche (unbeabsichtigten) Auswirkungen spezifische Itemformulierungen auf die psychometrische Qualität von Schülerbeurteilungen des Unterrichts haben, ist eine nach wie vor offene Frage, die im Rahmen des vom BMBF geförderten Projekts „Erfassung der Unter-

richtsqualität in Large-Scale-Studien: Optimierung der Modellierung und Itemauswahl“ untersucht wurde. In den folgenden Abschnitten erfolgt eine Zusammenfassung zentraler Befunde.

3 Die Validität von Schülerbeurteilungen des Unterrichts

Schülerbeurteilungen der Unterrichtsqualität versprechen eine hohe Effektivität für die Erfassung lern- und leistungsförderlicher Qualitätsmerkmale des Unterrichts, sie sind im Rahmen von Large-Scale-Assessments flexibel einsetzbar und zeigen in einer Vielzahl von Studien substanzielle Zusammenhänge mit Kriterien des Schulerfolges auf (Clausen, 2002; Fraser & Walberg, 1991; Klieme et al., 2008). Andererseits scheint die Verwendung von Schülerbeurteilungen nicht grundsätzlich eine bessere Erfassung relevanter Unterrichtsaspekte zu versprechen. Die Kritikpunkte umfassen die mangelnde Fähigkeit von Schülerinnen und Schülern zur Differenzierung verschiedener Facetten der Unterrichtsqualität sowie die Vergleichbarkeit (bzw. Messäquivalenz) von Schülerbeurteilungen über kontextuelle Rahmenbedingungen, wie etwa Schulfächer oder unterschiedliche Schulklassen. Beide Aspekte wurden im Rahmen einer Studie von Wagner, Göllner, Helmke, Trautwein und Lüdtke (2013) näher untersucht, deren Ergebnisse im Folgenden dargestellt werden.

3.1 Die Fähigkeit von Schülern zur dimensionalen Beschreibung der Unterrichtsqualität

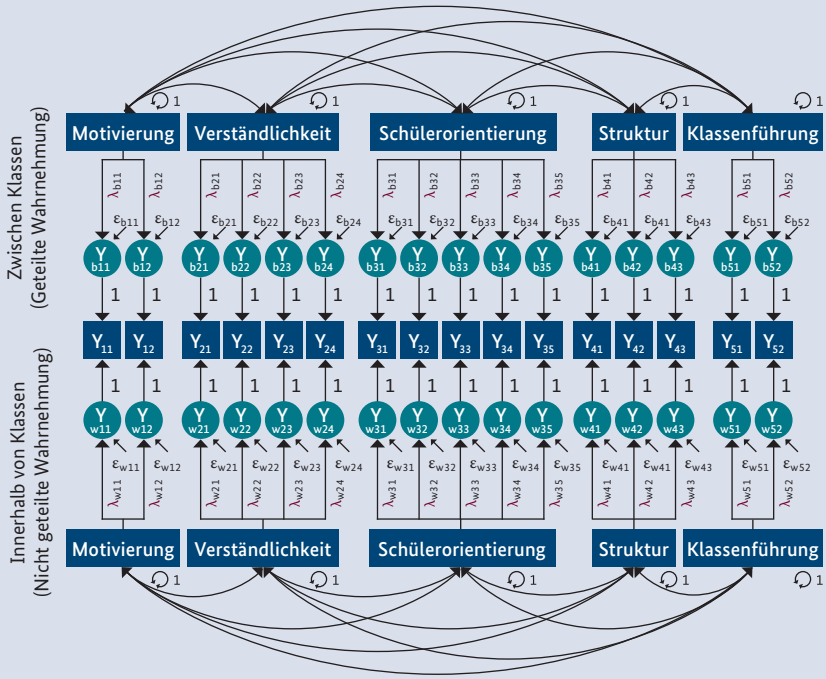
Im Allgemeinen ist die Frage nach der dimensionalen Struktur des Unterrichts eng mit der Anwendung faktorenanalytischer Verfahren verbunden (z. B. Nunnally, 1978). Faktorenanalytische Modelle geben Auskunft darüber, ob und wie sich ein Untersuchungsgegenstand anhand von Variablen (z. B. Fragebogenitems) mehrdimensional beschreiben lässt. Die entsprechenden Analysen erfolgen entweder erkundend im Sinne einer Exploration oder konfirmatorisch vor dem Hintergrund eines theoretischen Modells. Die für die empirische Bildungsforschung charakteristische Datenstruktur (Schülerinnen und Schüler innerhalb von Klassen und Schulen) erfordert die Anwendung sogenannter Mehrebenen-Strukturgleichungsmodelle (ML-SEM; Lüdtke, Marsh, Robitzsch & Trautwein, 2011; Marsh et al., 2013; Mehta & Neale, 2005). Diese Modelle ermöglichen es, die Faktorstruktur von Instrumenten zur Erfassung von Unterrichtsqualität simultan sowohl auf der Ebene der individuellen Schülerinnen und Schüler als auch auf Ebene der Klasse zu modellieren (Hox, Maas & Brinkhuis, 2010; Mehta & Neale, 2005), wobei im Falle von Schülerbeurteilungen der Unterrichtsqualität der Fokus der Analyse auf der Klassenebene liegt (Lüdtke, Robitzsch, Trautwein & Kunter, 2009).

Ein solches Mehrebenen-Strukturgleichungsmodell wurde in einer Studie von Wagner et al. (2013) auf Schülerbeurteilungen des Unterrichts in den Fächern Deutsch und Englisch angewandt. Grundlage der Untersuchung waren Daten der Studie Deutsch-Englisch-Schulleistungen International (DESI), die in den Jahren

2001 bis 2006 von einem interdisziplinären Konsortium unter der Federführung des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) durchgeführt wurde. DESI stellte eine bundesweite Untersuchung dar, an der sich allgemeinbildende Schulen aus sämtlichen Ländern der Bundesrepublik beteiligten. Ein Schwerpunkt des DESI-Projektes war die Untersuchung von Unterrichtsmerkmalen im Hinblick auf die Leistungsentwicklung von Schülerinnen und Schülern. Insgesamt nahmen 427 Klassen an der Untersuchung teil. Um im Rahmen der Analysen die Schülerbeurteilungen des Unterrichts für die Fächer Englisch und Deutsch parallel analysieren zu können, beschränkte sich die Auswertungsstichprobe auf 280 Schulklassen (N = 6.909 Schülerinnen und Schüler). Eingeschlossen wurden zudem nur Klassen, die in beiden Fächern durch unterschiedliche Lehrkräfte unterrichtet wurden. Das durchschnittliche Alter der Schülerinnen und Schüler betrug M = 15,8 Jahre. Das eingesetzte Instrument zur Beurteilung der Unterrichtsqualität aus Schülersicht umfasste eine große Zahl von Unterrichtsmerkmalen (siehe Klieme, Jude, Rauch, Ehlers, Helmke, Eichler et al., 2008). Aufgrund der Komplexität der Modellierung beschränkten sich die Analysen auf die fünf Unterrichtsdimensionen Schülermotivierung, Verständlichkeit, Schülerorientierung, Strukturiertheit und Klassenführung.

Das zugrunde liegende Faktormodell für beide Analyseebenen ist in Abbildung 1 dargestellt. Die Modellierung orientierte sich an dem von verschiedenen Autoren (z. B. Muthén, 1994) vorgeschlagenen Vorgehen.

Abb. 1: Mehrebenen-Strukturgleichungsmodelle (ML-SEM) zur Analyse von Schülerbeurteilungen des Unterrichts anhand von fünf Unterrichtsmerkmalen



Die Analyseergebnisse bestätigten eine dimensionale Struktur der zugrunde liegenden Schülerbeurteilungen des Unterrichts. Nicht nur auf der Ebene „innerhalb von Klassen", sondern auch „zwischen Klassen" waren die fünf Merkmalsdimensionen empirisch trennbar. Wenngleich die statistischen Zusammenhänge der einzelnen Unterrichtsmerkmale auf der Klassenebene vergleichsweise hoch ausfielen (Deutsch: $r = .70$ bis $r = .95$; Englisch: $r = .54$ bis $r = .94$), können Schülerurteile genutzt werden, um die Unterrichtsqualität einer Lehrkraft sehr spezifisch und detailliert beschreiben zu können. Dieser Befund wurde zusätzlich erhärtet durch die simultanen Analysen beider Unterrichtsfächer. Es zeigten sich nur mäßige Zusammenhänge der Unterrichtsmerkmale über beide Fächer (und unterschiedliche Lehrkräfte) hinweg. Die Zusammenhangsmaße auf der Klassenebene variierten für die fünf betrachten Merkmale zwischen $r = -.14$ und $r = .35$ (insgesamt 25 Korrelationen), wobei sich nur für drei Merkmale statistisch signifikante Zusammenhänge finden ließen. Demnach ist das geteilte Urteil von Schülerinnen und Schülern in einem Fach mehr oder weniger unabhängig von dem geteilten Urteil in dem anderen Fach (bezogen jeweils auf unterschiedliche Lehrkräfte). In anderen Worten spiegelt das Urteil von Schülerinnen und Schülern nicht primär stabile oder transkonsistente Anteile der Beurteiler wider, sondern beschreibt die spezifische Qualität einer Lehrkraft entlang verschiedener Qualitätsdimensionen.

3.2 Generalisierbarkeit von Schülerbeurteilungen des Unterrichts

Ein zweiter wichtiger Aspekt der Studie von Wagner et al. (2013) umfasste die Prüfung der Vergleichbarkeit von Schülerbeurteilungen des Unterrichts über unterschiedliche Kontexte (z. B. Fächer oder unterschiedliche Schulklassen) hinweg. Ohne Zweifel ist die Prüfung der faktoriellen Struktur von Konstrukten vor dem Hintergrund einer theoretischen Struktur ein unverzichtbarer erster Schritt zur Beantwortung der Frage, inwiefern Schülerinnen und Schüler die Unterrichtsqualität von Lehrkräften zutreffend bzw. valide beurteilen können. Ein sich daran anschließender und notwendiger zweiter Schritt umfasst die Frage nach der Vergleichbarkeit der Messung über verschiedene kontextuelle Bedingungen wie etwa Unterrichtsfächer oder Schulklassen. Im Kern ist zu klären, inwieweit Messergebnisse für diese unterschiedlichen Kontextbedingungen auf einem identischen Messmodell beruhen und folglich einen (unmittelbaren) Vergleich dieser Messwerte überhaupt erst rechtfertigen. So wäre bei Verletzung der Äquivalenzannahme beispielsweise ein Vergleich der Variabilität der Klassenführung von Lehrkräften in den Fächern Deutsch und Englisch nicht zulässig, da die Messungsunterschiede nicht sinnvoll auf Unterschiede bezüglich einer einzigen Dimension reduzierbar sind (sondern je nach Item mehr oder weniger stark variieren). Das Beispiel macht deutlich, wie stark auch die Unterrichtsqualität von Lehrkräften eine adäquate Messung im Sinne eines äquivalenten Messmodells voraussetzt.

Die Äquivalenzannahmen bei Schülerbeurteilungen der Unterrichtsqualität wurden im Beitrag von Wagner et al. (2013) sowohl für die an der Studie teilnehmenden Unterrichtsklassen als auch die beiden Unterrichtsfächer überprüft. Die Analysen

erfolgten in Anlehnung an das Verfahren multipler Gruppenanalysen, in der über eine sukzessive Fixierung verschiedener Parameter des Messmodells geprüft wird (Meredith, 1993; Widaman & Reise, 1997). Jedoch basiert die Grundidee der Äquivalenztestung über Klassen auf einem Vergleich der ebenenspezifischen Ladungsmuster (siehe Jak, Oort & Dolan, 2013; Wagner, 2008; Wagner et al., 2013). Inwieweit die Annahme eines äquivalenten Messmodells über Klassen hinweg haltbar ist, kann geprüft werden, indem Faktorladungen über Ebenen hinweg (deskriptiv) verglichen und mögliche Unterschiede inferenzstatistisch getestet werden. Die Anwendung eines solchen Verfahrens auf die Daten der DESI-Studie zeigte für die Merkmale Motivierung, Verständlichkeit und Schülerorientierung die größten Unterschiede. Für diese drei Merkmale ist ein Vergleich der Messwerte über verschiedene Schulklassen hinweg nur bedingt zu rechtfertigen. Vielmehr scheinen die entsprechenden Messungen durch jeweils klassenspezifische Eigenschaften der Schüler mitbestimmt. Hingegen bestätigte sich die Annahme äquivalenter Messmodelle für die Merkmale Strukturierung und Klassenführung. Für beide Merkmale können demnach Messwerte in sinnvoller und inhaltlich valider Weise über Klassen hinweg miteinander verglichen werden. Auf diesen ebenenspezifischen Ladungsmustern beruhend, wurde abschließend zusätzlich die fächerbezogene Messäquivalenz überprüft. Für die beiden Merkmale Strukturierung und Klassenführung konnte eine vollständige Ladungsäquivalenz über beide Analyseebenen (innerhalb und zwischen Klassen) und Fächer (Deutsch, Englisch) hinweg als gerechtfertigte Annahme bestätigt werden.

Zusammenfassend zeigen die Ergebnisse der Studie von Wagner et al. (2013), dass Schülerinnen und Schüler in differenzierter Weise das Unterrichtsgeschehen bzw. die Qualität des Unterrichts auf unterschiedlichen Dimensionen beurteilen können, wenngleich Einschränkungen hinsichtlich der Vergleichbarkeit über Klassen hinweg zu beachten sind. Insbesondere Merkmale, die durch einen hohen Schülerbezug (d. h. Schüler als Bestandteil eines gelingenden Unterrichts) gekennzeichnet sind (d. h. Motivierung, Schülerorientierung und Verständlichkeit), sind demnach nur bedingt für die Vergleiche verschiedener Lehrkräfte geeignet. Bei der Beurteilung solcher Unterrichtsmerkmale spielen offenbar Schülermerkmale eine nicht zu unterschätzende Rolle, die sich in verschiedenen Klassen unterscheiden.

4 Die Übereinstimmung unterschiedlicher Datenquellen

Neben der Überprüfung der faktoriellen Struktur und Messäquivalenz von Schülerbeurteilungen der Unterrichtsqualität sind weitere Kriterien zur Prüfung der Validität von Schülerbeurteilungen des Unterrichts formuliert worden. Hier ist in erster Linie die Übereinstimmung von Schülerbeurteilungen mit weiteren Datenquellen (z. B. Selbstberichte von Lehrkräften und Beobachtungen) zu nennen. Empirische Arbeiten zur Übereinstimmung unterschiedlicher Datenquellen sind rar und legen die Annahme lediglich moderater Zusammenhänge zwischen den Beurteilungsperspektiven nahe. So berichtete Clausen (2002) kleine bis mittlere relative Übereinstimmungskoeffizienten von $r = -.28$ bis $r = .42$ zwischen Schülerbeurteilungen und Lehrerselbstberichten, von $r = -.22$ bis $r = .45$ zwischen Schüler- und Videobeurteilern

und von $r = -.04$ bis $r = .43$ zwischen Lehrerberichten und Videobeurteilern. Untersuchungsergebnisse der Studie COACTIV zeigten nur unwesentlich höhere Übereinstimmungskoeffizienten zwischen Schülerinnen und Schülern und Lehrkräften ($r = .21$ bis $r = .31$; Kunter et al., 2008). Auch eine neuere Studie zur Unterrichtsbeurteilung von Grundschülern zeigt vergleichbare Ergebnisse für unterschiedliche Dimensionen der Unterrichtsqualität (Fauth et al., 2014). Worin liegen die Ursachen dieser vergleichsweise geringen Übereinstimmung unterschiedlicher Perspektiven? In Anlehnung an Clausen (2002) ist zunächst festzuhalten, dass keine der Datenquellen als prinzipiell „besser" oder „schlechter" zu betrachten ist. Vielmehr weisen alle Perspektiven spezifische Anteile auf, die je nach Beurteilungsgegenstand bzw. der betrachteten Unterrichtsmerkmale näher an der Unterrichtswirklichkeit liegen (Clausen, 2002; Kunter & Baumert, 2006).

Doch es gibt vermutlich weitere Erklärungsansätze, die insbesondere kognitive Anforderungen bei der Beurteilung als Ursache mangelnder Datenqualität identifizieren. Aus diesem Grund wurden im Rahmen des BMBF-Projektes verschiedene Anforderungsmerkmale von Fragebogenitems definiert und im Hinblick auf die Übereinstimmung verschiedener Urteilsperspektiven untersucht.

4.1 Anforderungsmerkmale von Items

Fraglos ist die Beantwortung eines Items für die Schülerinnen und Schüler nicht trivial. Es besteht Übereinkunft darin, dass die Beantwortung von Fragebogenitems entlang verschiedener Prozessschritte erfolgt, die je nach konkretem Iteminhalt teilweise anspruchsvolle kognitive und motivationale Anforderungen an den Beurteiler stellen (Lenske, 2011; Tourangeau, Rips & Rasinski, 2000). Diese Schritte umfassen unter anderem das Verständnis und die Interpretation des jeweiligen Iteminhalts, die Sammlung relevanter Informationen zu dessen Beantwortung, die Verdichtung der ermittelten Informationen zu einem Urteil und schließlich die Auswahl einer passenden Antwortalternative (z. B. Tourangeau et al., 2000). Bereits ein oberflächlicher Blick auf Fragebogeninstrumente zur Erfassung der Unterrichtsqualität zeigt, dass allein das sprachliche Verständnis und die Interpretation des Iteminhalts in hohem Maße anfordernd bzw. herausfordernd sein können. Aus theoretischer Perspektive und nach einer Systematisierung von Fragebogeninstrumenten aus großen deutschen Schulleistungsstudien können neben der sprachlichen (d. h. orthografischen, grammatikalischen und linguistischen) Komplexität drei weitere Anforderungsmerkmale unterschieden werden. Hierzu gehören a) der Adressatenbezug eines Items (bestimmt den Adressat des Lehrerverhaltens: individueller Schüler versus sämtliche Schüler der Klasse; den Brok, Brekelmans & Wubbels, 2006), b) die Wahrnehmungsperspektive (bestimmt die Sichtweise, aus der beurteilt wird: individuelle Sicht des Schülers oder Wir-Perspektive; den Brok, Brekelmans & Wubbels, 2006) und c) der Zeitbezug eines Items (bezieht sich das Urteil auf die letzte Stunde oder ist ein über mehrere Unterrichtsstunden aggregiertes Urteil verlangt; Tourangeau & Rasinski, 2000). Je nach Ausprägungsgrad stellt die Itembeantwortung eine mehr oder weniger hohe Anforderung an die Schülerinnen und Schüler dar. So müssen Schülerinnen

und Schüler in vielen Fällen nicht nur auf der Grundlage ihrer eigenen individuellen Erfahrungen mit der Lehrkraft urteilen (z. B. „Mein Lehrer achtet darauf, dass ich im Unterricht mitkomme"), sondern müssen in vielen Fällen die Erfahrungen der Mitschüler in ihr Urteil integrieren (z. B. „Der Lehrer unterstützt uns zusätzlich, wenn wir Hilfe brauchen"). Es ist zu vermuten, dass damit eine erhebliche Anforderung an die Schülerinnen und Schüler gestellt wird, welche die Güte der Beurteilung erheblich einschränken kann.

Eine im Rahmen des BMBF-Projektes durchgeführte Klassifikation von insgesamt 533 Fragebogenitems aus sieben deutschen Large-Scale-Studien zeigte, dass ein großer Teil von Items derartige Anforderungen stellt, wenngleich Bezüge und Perspektiven in vielen Fällen nicht eindeutig definiert sind und somit im Ermessen der einzelnen Schülerinnen und Schüler liegen (Göllner, Wagner, Klieme & Trautwein, 2014). Beispielsweise ist nur für einen kleinen Teil der gesichteten Schülerfragebogen der genaue Zeitbezug für die Beurteilung explizit gegeben. Die vorgenommenen Systematisierungen können als Basis für die Zusammenstellung von Items für zukünftige Schulleistungsstudien dienen, die beispielsweise eine größere Einheitlichkeit aufweisen als bisher genutzte Instrumente.

4.2 Die Bedeutung von Zeitbezügen für die Übereinstimmung von Schülern und Lehrkräften

Inwieweit der Zeitbezug der Beurteilungen der Unterrichtsqualität, der in den systematisch erfassten Items der deutschen Schulleistungsstudien erheblich variiert, die Übereinstimmung verschiedener Datenquellen beeinflusst, war Gegenstand einer vertiefenden Untersuchung von Wagner et al. (im Druck).

In vielen Fällen wird die Unterrichtsqualität von Lehrkräften als eine statische Größe behandelt, die Einfluss auf die Lern- und Leistungsentwicklung ausüben kann, aber selbst keiner Veränderung (z. B. im Laufe eines Schuljahres) unterliegt. Dennoch spielt der Faktor „Zeit" für den Beurteilungsprozess eine nicht unerhebliche Rolle. Erstens unterliegen mit großer Wahrscheinlichkeit auch Konstrukte wie das unterrichtliche Handeln einer Lehrkraft einer Veränderung. Lehrkräfte könnten beispielsweise im Laufe eines Schuljahres auf spezifische Bedürfnisse der Schülerinnen und Schüler reagieren und entsprechende Unterstützungsformen der Schülerorientierung oder die Strukturgebung stärker betonen. Belege für die Veränderbarkeit des Lehrerhandelns finden sich darüber hinaus in einer zunehmend größer werdenden Zahl von Interventionsstudien (z. B. Brown, Jones, LaRusso & Aber, 2010). Zweitens unterliegt vermutlich jegliche Form des Verhaltens natürlichen Schwankungen, die nicht im Sinne systematischer Veränderungen zu erklären, sondern der Spezifität einer bestimmten Situation oder eines Zeitpunktes geschuldet sind. So werden zwar Unterrichtsbeobachtungen oft als die beste Möglichkeit zur neutralen Erfassung des Unterrichts bezeichnet, sie können aber letztlich nur einen zeitlich begrenzten Ausschnitt eines zeitvariablen Verhaltens abbilden (Clausen, 2002). Schließlich ist der zeitliche Bezug in der eigentlichen Befragungssituation ein weiterer wichtiger Zeitaspekt. Der Umstand, dass der Beurteilungszeitraum bei den meisten Instrumenten zur

Erfassung der Unterrichtsqualität nicht explizit genannt wird, verschärft dieses Problem erheblich. Je nachdem, welchen zeitlichen Bezug die Schülerinnen und Schüler in der Befragungssituation wählen, könnten – unter der Annahme von sich über die Zeit veränderndem Lehrerhandeln – sehr unterschiedliche Beurteilungsergebnisse resultieren. Ein Teil der Schülerinnen und Schüler könnte seine Beurteilungen auf das bereits vergangene Schuljahr beziehen, während andere Schülerinnen und Schüler ihr Urteil auf den letzten Eindruck stützen. Beide Fälle würden im Ergebnis nicht nur dazu führen, dass Schülerinnen und Schüler gegebenenfalls zu sehr unterschiedlichen Beurteilungen gelangen, sondern auch die Übereinstimmung der Beurteilungen mit anderen Datenquellen (z. B. Lehrerbericht) beeinträchtigen.

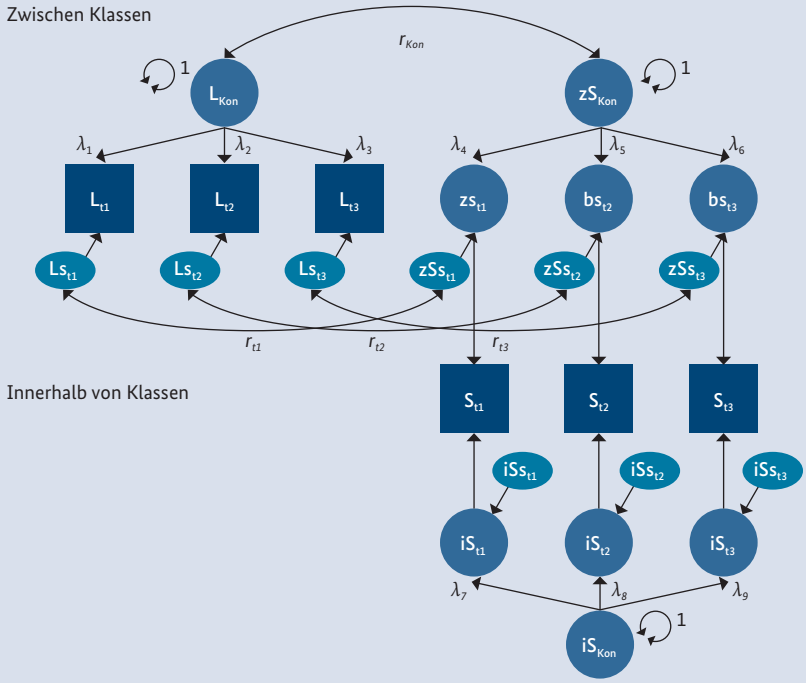
Die Bedeutung des Zeitbezuges bei der Beurteilung des Unterrichts wurde in dem BMBF-Projekt untersucht, indem Daten des ebenfalls vom BMBF geförderten Projekts „Lernen mit Plan“ (vgl. Ogrin, Keller, Friedrich, Trautwein & Schmitz, im Druck) reanalysiert wurden. Um eine möglichst genaue Abschätzung möglicher Zeiteffekte zu ermöglichen, wurde ein Fragebogendesign verwendet, welches den Zeitbezug der Beurteilung für Schülerinnen und Schüler genau definierte. Es wurden $N = 686$ Schülerinnen und Schüler der fünften Jahrgangsstufe sowie ihre Mathematiklehrer aus insgesamt 74 Hauptschulklassen wiederholt zur Unterrichtsqualität im Fach Mathematik befragt. Das mittlere Alter der Schüler betrug 11,8 Jahre (48.3 Prozent Mädchen). Die Untersuchung erfolgte zu drei Messzeitpunkten über 13 Wochen hinweg (T1: Mitte des Schuljahres; T2: nach sechs Wochen; T3: nach weiteren sechs Wochen). Zum ersten Messzeitpunkt wurden sowohl die Schülerinnen und Schüler als auch die Lehrkräfte zur Qualität des Mathematikunterrichtes der Lehrkraft befragt. Zu den beiden folgenden Messzeitpunkten bezog sich die Einschätzung der Unterrichtsqualität auf die Zeit seit der letzten Befragung. Erfasst wurden neben der Klassenführung die Merkmale Autonomieunterstützung und Zielklarheit des Unterrichts. Die Instruktion und die gewählten Zeitpunkte der Untersuchung sind aus zweierlei Gründen interessant. Zum einen macht der vergleichsweise kurze Zeitraum der Untersuchung zur Mitte des Schuljahres eine „wirkliche“ Veränderung der Unterrichtsqualität eher unwahrscheinlich. Die erste Messung fand nach Bekanntgabe der Halbjahresnoten statt, während die letzte Messung weit vor Ende des Schuljahres abgeschlossen wurde. Etwaige Einflussgrößen, wie die Unbekanntheit der Schülerinnen und Schüler mit der Lehrkraft oder auch eine mögliche Schwerpunktverschiebung des Unterrichts aufgrund bevorstehender Zeugnisnoten, können mit einiger Sicherheit ausgeschlossen werden. Zum Zweiten wurde der zu betrachtende Beurteilungszeitraum explizit definiert, um der Zeitspezifität der einzelnen Messzeitpunkte entsprechend gerecht werden zu können.

Die Ergebnisse eines State-Trait-Modells (siehe Abbildung 2) zeigten für die Schülerurteile (zSKon) mit Ausnahme der Klassenführung eine geringere Zeitspezifität als für Lehrerurteile (LKon). Nahezu die gesamte Variabilität in den Schülerbeurteilungen für die Merkmale Zielklarheit und Autonomieunterstützung auf Klassenebene konnte auf zeitkonsistente Urteilskomponenten zurückgeführt werden. Für das Merkmal Klassenführung lag der entsprechende Varianzanteil bei durchschnittlich 75 Prozent. Lehrerurteile variierten demgegenüber stärker über die Messzeitpunkte hinweg. Etwa drei Viertel der Unterschiede zwischen den Beurteilungen waren auf

zeitlich stabile Urteilskomponenten zurückzuführen, wobei die Unterschiede zwischen den Unterrichtsmerkmalen deutlich weniger variierten als bei den Schülerbeurteilungen. Mit Blick auf die Übereinstimmung der Beurteiler zeigten sich – mit Ausnahme der Autonomieunterstützung – höhere Beurteilerübereinstimmungen bezüglich der zeitkonsistenten Komponenten verglichen mit den „einfachen" messzeitpunktspezifischen Urteilen. Während sich für das Merkmal Klassenführung zu den einzelnen Messzeitpunkten mittlere bis hohe Korrelationen fanden ($r_{t1 =}$.72, r_{t2} = .58 und r_{t3} = .77), ergab sich für die zeitkonsistenten Urteilskomponenten eine nahezu perfekte Übereinstimmung von *r*Kon = 1.0. Ein ähnliches Bild zeigte sich für die Zielklarheit. Auch hier wiesen die zeitkonsistenten Urteilskomponenten der Lehrkräfte und Schülerinnen und Schüler höhere Übereinstimmungen (r_{Kon} = .50) als die einzelnen messzeitpunktspezifischen Urteile (r_{t1}= .45, r_{t2}= .35 und r_{t3}= .32) auf. Lediglich für das Merkmal Autonomieunterstützung konnten weder für die einzelnen Messzeitpunkte (r_{t1}= .05, r_{t2}= .11 und r_{t3}= .04) noch für die zeitkonsistenten Urteilskomponenten substanzielle Übereinstimmungen zwischen Schüler- und Lehrerperspektive nachgewiesen werden.

Zusammenfassend und mit Blick auf die Bedeutung des Zeitbezuges für die Unterrichtsbeurteilungen scheint die Erfassung zeitlich stabiler Unterrichtsmerkmale (im Sinne des „typischen" Unterrichts bei einer bestimmten Lehrkraft) durch Zusammenfassung mehrerer Messzeitpunkte vorteilhaft zu sein. Dies ist für die Unterrichtsforschung ausgesprochen aufschlussreich, da bei einer Vielzahl der bekannten Unterrichtsstudien die Erfassung des Unterrichts nur einmal erfolgte. Die Beurteilung des Unterrichts zum Zeitpunkt einer Messung beinhaltet demnach Messanteile, die für die Erfassung der postulierten Qualitätsmerkmale (z. B. generelle Strukturierung der einzelnen Unterrichtsstunden) weniger relevant sind. Dies zeigt sich sowohl für Lehrer- und Schülerbeurteilungen des Unterrichts, wenngleich für Schülerbeurteilungen das Ausmaß zeitspezifischer Anteile insgesamt geringer zu sein scheint. Die Annahme zeitspezifischer „Verunreinigungen" im Hinblick auf die Messung verschiedener Qualitätsaspekte des Unterrichts bestätigte sich im Rahmen weiterer Analysen zur Vorhersage relevanter Zielkriterien des Unterrichts (d. h. Entwicklung der Mathematikleistung und des Selbstkonzeptes über den Zeitraum der Untersuchung). Die Ergebnisse nachgeschalteter Analysen zeigten, dass beide Zielvariablen durch die drei untersuchten Unterrichtsmerkmale (insbesondere Schülerbeurteilungen) vorhergesagt werden konnten, wenngleich die Effekte in unsystematischer Weise über die Messzeitpunkte streuten. Es ließen sich keine „besseren" oder „schlechteren" Messzeitpunkte im Hinblick auf die Prädiktion der beiden Zielgrößen finden. Vielmehr wechselten sich die Messzeitpunkte sowohl in den Effektstärken als auch der statistischen Signifikanz in unsystematischer Weise ab. Demgegenüber ergab sich für die die zeitstabilen Beurteilungsanteile ein weitaus konsistenteres und theoriekonformeres Bild.

Abb 2: Latentes State-Trait-Modell zur Analyse von zeitkonsistenten und messzeitpunktspezifischen Urteilskomponenten



5 Die Bedeutung sprachlicher Komplexitätsmerkmale

Wird die Bedeutung des Iteminhalts für die psychometrische Qualität von Schülerbeurteilungen des Unterrichts in konsequenter Weise weitergedacht, bleiben mögliche Einflussgrößen nicht nur auf die Wahrnehmungsperspektive sowie den Adressaten- und Zeitbezug beschränkt, sondern umfassen alle schwierigkeitsgenerierenden Merkmale eines Fragebogenitems. So ist davon auszugehen, dass jede Form sprachlicher Komplexität mit der Güte von Schülerbeurteilungen assoziiert ist. Ein weiteres durch das BMBF gefördertes Projekt („Sprachliche Komplexitätsmerkmale von Fragebogenitems: Bedeutung für die psychometrische Qualität von Schülerbeurteilungen des Unterrichts aus Schülersicht und die Vorhersage des Lernerfolgs in Large-Scale-Assessments“, 01LSA1507) ermöglicht die Vertiefung und Erweiterung dieser Forschungsfrage.

5.1 Sprachliche Komplexitätsmerkmale

Modernen Konzepten der Leseforschung folgend, setzt auch die sprachliche Verarbeitung bzw. das sprachliche Verständnis eines Items die Fähigkeit voraus, die dargebotenen Informationen flüssig und sinnverstehend zu lesen. Zur Interpretation

muss der Itemtext in einzelne Elemente zerlegt, die einzelnen Wörter oder Chunks müssen decodiert und deren Bedeutung muss abgerufen werden. Anschließend sind die einzelnen Bestandteile entsprechend den grammatischen Relationen und semantischen Interpretationsprinzipien zu kombinieren, d. h., die Zusammenhänge zwischen den Satzteilen müssen erkannt werden. Diese Verarbeitungsschritte erfordern morphologische, syntaktische und semantische Kompetenzen sowie die Fähigkeit, das Item im Kontext adäquat pragmatisch zu interpretieren. Hierbei gilt es zu bedenken, dass Schriftsprache unabhängig von spezifischen Inhalten andere Chunks und Strukturen aufweist als gesprochene Sprache. Schriftsprachliches Material bzw. konzeptionell schriftliche Sprache zeichnet sich – im Gegensatz zur Alltagssprache bzw. zur konzeptionell mündlichen Sprache – unter anderem durch lange Nominalphrasen, die Verwendung von Passiv sowie lange und komplexe Sätze aus (für einen Überblick siehe Berendes, Dragon, Weinert, Heppt & Stanat, 2013).

Einige dieser Sprachmerkmale konnten in aktuellen Untersuchungen zur bildungssprachlichen Kompetenz von Schülerinnen und Schülern bereits als relevant für die Vergleichbarkeit von Testaufgaben identifiziert werden (z. B. Berendes, Wagner, Meurers & Trautwein, 2015). Es zeigte sich, dass für das Verständnis sprachlicher Inhalte, die in Situationen mit geringer sozialer und situativer Einbettung kommuniziert werden, bildungssprachliche Kompetenzen von direkter Relevanz sind. So fanden beispielsweise Shaftel, Belton-Kocher, Glasnapp und Poggio (2006) in einer Untersuchung mit Viert-, Siebt- und Zehntklässlern, dass Präpositionalphrasen und mehrdeutige Wörter in Testitems einen substanziellen Einfluss auf die Testleistung der Schülerinnen und Schüler ausüben (siehe auch Haag, Heppt, Stanat, Kuhl & Pant 2013).

5.2 Sprachmerkmale und psychometrische Güte von Schülerbeurteilungen des Unterrichts

Auch bei der Bearbeitung von Fragebogenitems zur Unterrichtswahrnehmung ist eher von einer geringen Kontextualisierung auszugehen. Zugleich enthalten Fragebogeninstrumente für Schülerbeurteilungen des Unterrichts durchaus komplexe morphologische Muster und Satzkonstruktionen sowie bildungssprachlich geprägte Begriffe. Auf der Grundlage der im Rahmen des Projektes durchgeführten Sichtung und Klassifikation von Fragebogenitems aus sieben deutschen Schulleistungsstudien wurde untersucht, inwieweit derartige Komplexitätsmerkmale mit der psychometrischen Güte von Schülerbeurteilungen des Unterrichts einhergingen (Göllner, Wagner, Meurers, Berendes & Trautwein, in Vorbereitung). Zu den für die Untersuchung relevanten Größen gehörten neben der relativen Übereinstimmung von Schülerinnen und Schülern weitere Maße der Itemstruktur, wie etwa die mittlere Iteminterkorrelation innerhalb der einzelnen Unterrichtsskalen. Die entsprechenden Indizes wurden anhand von Reanalysen der Originaldatensätze ermittelt. Zudem erfolgte eine Codierung potenziell relevanter Sprachmerkmale. Die Auswahl der Komplexitätsmerkmale orientierte sich an aktuellen Konzepten der Leseforschung sowie an Theorien des Textverstehens (Lenhard & Artelt, 2009). Da es sich bei Items um sehr

kurzes Sprachmaterial auf Satzebene handelt, war die Variationsbreite für eine Vielzahl bekannter Sprachmerkmale erheblich eingeschränkt. Für die Analysen wurden daher nur acht Sprachmerkmale für insgesamt 98 der 533 Items herangezogen. Dazu gehörten die Anzahl der Zeichen pro Wort, die Wortanzahl pro Satz, die mittlere Wortanzahl in Haupt- und Nebensätzen, die Anzahl der Nebensätze, die Anzahl von Nominalphrasen sowie die Anzahl von Hypernymen, Hyponymen und Synonymen pro Sinneinheit. Die Auswahl der Items erfolgte anhand zweier Kriterien. Erstens entstammten alle Items aus Studien mit dem Fokus Mathematik und kamen zudem in mindestens zwei Studien zur Anwendung. Die Ergebnisse zeigten, dass sich die Items im Hinblick auf die codierten Sprachmerkmale deutlich unterschieden. Das in der Analyse genutzte Gütekriterium von Schülerbeurteilungen war der Intraklassenkorrelationskoeffizient (ICC) als Maß der relativen Übereinstimmung von Schülerinnen und Schülern einer Klasse (Snijders & Bosker, 1999). Die mittlere ICC der 98 Items betrug .14 (*Min* = .02, *Max* = .31; *SD* = 0.07).

Im Anschluss daran wurden systematische Analysen zur Frage durchgeführt, ob sprachliche Komplexität das Antwortverhalten der Schülerinnen und Schüler beeinflusst. Hierbei wurde von der Annahme ausgegangen, dass sprachlich komplexe Items von den Schülerinnen und Schülern einer Klasse unterschiedlich gut verstanden und interpretiert werden. Sollten diese Unterschiede bestehen, würden bei entsprechend sprachlich komplexen Items die Schülerurteile innerhalb einer Klasse relativ unterschiedlich ausfallen, selbst wenn in Wirklichkeit die Schülerinnen und Schüler eine eher einheitliche Meinung zur abgefragten Qualitätsdimension haben. Eine hohe sprachliche Komplexität würde sich dementsprechend in relativ niedrigen ICCs ausdrücken. Tatsächlich konnte ein entsprechendes Muster für die Anzahl von Nebensätzen und die Anzahl von Nominalisierungen gefunden werden, nicht aber für die anderen Sprachmerkmale. Die beiden Komplexitätsmerkmale blieben auch dann prädiktiv, wenn für die Herkunftsstudie – und somit auch für Unterschiede bezüglich der Stichprobenzusammensetzungen – des Items statistisch kontrolliert wurde.

6 Zusammenfassung und Ausblick

Schülerurteile der Unterrichtsqualität bieten nicht nur eine hohe Effektivität bei der Erfassung bzw. Messung lern- und leistungsförderlicher Qualitätsmerkmale des Unterrichts, sondern versprechen auch eine umfangreiche und detaillierte Beschreibung des Unterrichtsgeschehens. Die Tatsache, dass Schülerbeurteilungen in vielen Fällen deutlich stärker mit verschiedenen Zielkriterien des Unterrichts assoziiert sind als Selbstberichte von Lehrkräften oder Beobachterdaten, unterstreicht zusätzlich die Bedeutung von Schülerbeurteilungen der Unterrichtsqualität in Large-Scale-Assessments. Dem wurde vielfach entgegengehalten, dass Schülerinnen und Schüler keine adäquate didaktisch-pädagogische Expertise besitzen, um komplexe Unterrichtsprozesse angemessen einordnen zu können, und die eigene Involviertheit von Schülerinnen und Schülern zu einer mangelnden Objektivität während der Beurteilung führen könnte (Aleamoni, 1999; Clausen, 2002; Kunter & Baumert, 2006).

Die Frage, wie angemessen Unterrichtsqualität mithilfe von Schülerurteilen erfasst werden kann, war entsprechend das zentrale Anliegen des vom BMBF geförderten Projekts „Erfassung der Unterrichtsqualität in Large-Scale-Studien: Optimierung der Modellierung und Itemauswahl". In dem Projekt entstanden eine Reihe von empirischen Studien, welche a) die Konstruktvalidität von Schülerbeurteilungen, b) die Bedeutung des Zeitbezugs für die Schüler-Lehrer-Übereinstimmung und c) die Bedeutung sprachlicher Komplexitätsmerkmale für die relative Übereinstimmung von Schülerbeurteilungen näher untersuchten.

In der Rückschau zeigen die Ergebnisse, dass Schülerinnen und Schüler – zumindest für eine Reihe von zentralen Qualitätsdimensionen – die Qualität des Unterrichts nicht nur zuverlässig, sondern auch differenziert in der dafür notwendigen Detailliertheit beschreiben können. Befragungen von Schülerinnen und Schülern ermöglichen somit eine Quantifizierung verschiedener Qualitätsaspekte des Unterrichts, die als zentrale Elemente eines lern- und leistungsförderlichen Unterrichts gelten. Dabei erlauben Schülerbeurteilungen der Unterrichtsqualität nicht nur, das Geschehen in seiner Breite zu erfassen, sondern können auch zur Erfassung von weniger gut beobachtbaren Unterrichtsmerkmalen dienen. So ist mit Blick auf die durchgeführte Untersuchung zur faktoriellen Struktur von Schülerbeurteilungen eine vergleichbar umfangreiche Erfassung relevanter Qualitätsdimensionen im Rahmen einer Videostudie kaum oder nur durch enormen Aufwand vorstellbar. Darüber hinaus bieten Schülerbeurteilungen der Unterrichtsqualität in Kombination mit neueren Verfahren der Datenanalyse weitreichende Möglichkeiten, die zentralen Gütemaßstäbe pädagogisch-psychologischer Forschung erweiternd zu prüfen. Beispielsweise besitzt das im Rahmen des Projektes entwickelte Analysevorgehen zur Äquivalenzprüfung von Schülerbeurteilungen über die Analyseeinheit der Klassen hohe Relevanz für die Überprüfung zukünftiger Fragebogeninstrumente in der Unterrichtsforschung.

Derartige Verfahren zeigen jedoch auch die Grenzen von Schülerbeurteilungen des Unterrichts und die Notwendigkeit weiterer Entwicklungsanstrengungen im Bereich der Unterrichtsforschung. Insbesondere sprachliche Anforderungs- und Komplexitätsmerkmale sind in ihrer Bedeutung für den Beantwortungsprozess zu berücksichtigen. Existierende Fragebogenverfahren weisen eine erstaunlich hohe Variation sprachlicher Anforderungsmerkmale auf, welche Einfluss auf die psychometrische Qualität von Schülerbeurteilungen des Unterrichts ausüben können. Beispielsweise lässt sich die Verletzung der Messäquivalenzannahme für einige der im Projekt untersuchten Unterrichtmerkmale nicht einseitig auf spezifische Charakteristika des Merkmals zurückführen, sondern könnte gleichfalls Ausdruck unterschiedlicher kognitiver Anforderungen während der Beurteilung sein. Die für die fünf Merkmale verwendeten Items zeichnen sich ausnahmslos durch einen hohen Inferenzgrad aus, welcher die Abhängigkeit des Messergebnisses unter Umständen deutlich erhöht.

Doch auch die Frage, ob die psychometrische Qualität von Schülerbeurteilungen darüber hinaus von weiteren Anforderungsmerkmalen abhängig ist, kann vorsichtig mit einem Ja beantwortet werden. Es wurde gezeigt, dass eine mangelnde Übereinstimmung zwischen Schülerinnen und Schülern und Lehrkräften in ihrer Beurteilung der Unterrichtsqualität zu einem substanziellen Anteil auf die Spe-

zifität des Messzeitpunkts zurückgeführt werden kann. Die geschieht sogar dann, wenn der Beurteilungszeitraum in der Fragebogeninstruktion explizit definiert ist. Die Situationsabhängigkeit einer Messung resultiert zumindest für einige der untersuchten Merkmale in einer Unterschätzung der Übereinstimmung und somit in einer weniger validen Messung. Die Tatsache, dass in der Mehrzahl angewendeter Fragebogeninstrumente kein Zeitbezug gegeben ist, verleiht notwendigen Anstrengungen zur Verbesserung von Fragebogeninstrumenten zusätzliches Gewicht. Diese Anstrengungen sollten sich jedoch nicht nur auf Anforderungsmerkmale wie etwa Perspektive, Adressaten- oder Zeitbezug in der Fragebogenkonstruktion beschränken, sondern auch weitere Komplexitätsmerkmale der Formulierung umfassen. Dies gilt insbesondere dann, wenn die Erfassung der Unterrichtsqualität nicht nur auf Schülerinnen und Schüler höherer Klassenstufen beschränkt bleiben soll. Die im Rahmen des bereits genannten Projektes „Sprachliche Komplexitätsmerkmale von Fragebogenitems" gefundenen Ergebnisse bezüglich der Urteilsübereinstimmung von Schülerinnen und Schülern lässt vermuten, dass die psychometrische Güte der Befragungsinstrumente bei jüngeren Schülerinnen und Schülern noch deutlich stärker beeinflusst ist. Entsprechende Weiterentwicklungen sind erforderlich, um Untersuchungsinstrumente für Schulleistungsuntersuchungen zu generieren, mit deren Hilfe noch differenzierter als bisher die Determinanten und Konsequenzen guten Unterrichts (vgl. Kunter & Trautwein, 2013) identifiziert werden können.

Literaturverzeichnis

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal Of Personnel Evaluation In Education, 13,* 153–166.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9,* 469–520.

Berendes, K., Dragon, N., Weinert, S., Heppt, B. & Stanat, P. (2013). Hürde Bildungssprache? Eine Annäherung an das Konzept „Bildungssprache" unter Einbezug aktueller empirischer Forschungsergebnisse. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Perspektiven aus Psychologie, Sprachwissenschaft und empirischer Bildungsforschung* (S. 17–41). Münster: Waxmann.

Berendes, K., Wagner, W., Meurers, D. & Trautwein, U. (2015). Grammatikverständnis von Kindern unterschiedlicher sprachlicher und sozioökonomischer Herkunft. *Frühe Bildung, 4 (3),* 126–134.

Blömeke, S. (2004). Empirische Befunde zur Wirksamkeit der Lehrerbildung. In S. Blömeke, P. Reinhold, G. Tulodziecki & J. Wildt (Hrsg.). *Handbuch Lehrerbildung* (S. 59–91). Bad Heilbrunn: Klinkhardt.

Brown, J. L., Jones, S. M., LaRusso, M. D. & Aber, J. L. (2010). Improving classroom quality: Teacher influences and experimental impacts of the 4Rs program. *Journal of Educational Psychology, 102 (1),* 153–167.

Clausen, M. (2002). *Qualität von Unterricht: Eine Frage der Perspektive?* Münster: Waxmann.

den Brok, P., Brekelmans, M. & Wubbels, T. (2006). Multilevel issues in reasearch using students' perceptions of learning environments: The case of the questionnaire on teacher interaction. *Learning Environments Research, 9,* 199–213.

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G. & Sherin, B. L. (2010). Conducting Video Research in the Learning Sciences: Guidance on Selection, Analysis, Technology, and Ethics. *Journal of The Learning Sciences, 19,* 3–53.

Desimone, L. M., Smith, T. M. & Frisvold, D. E. (2010). Survey Measures of Classroom Instruction: Comparing Student and Teacher Reports. *Educational Policy, 24,* 267–329.

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. *Zeitschrift für Pädagogische Psychologie, 28,* 127–137.

Fisicaro, S. A. & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14,* 419–429.

Fraser, B. J. & Walberg, H. J. (1991). *Educational environments: Evaluation, antecedents and consequences.* Elmsford, NY: Pergamon Press.

Funder, D. C. (2001). Accuracy in personality judgment: Research and theory concerning an obvious question. In B. W. Roberts, R. Hogan (Hrsg.), *Personality psychology in the workplace* (S. 121–140). Washington, DC: American Psychological Association.

Gigliotti, R. J. & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology, 82,* 341–351.

Göllner, R., Wagner, W., Klieme, E. & Trautwein, U. (2014). *Die Erfassung des Unterrichts aus Schülersicht: Ergebnisse einer Systematisierung nationaler Fragebogeninstrumente.* Vortrag auf der 2. Tagung der Gesellschaft für Empirische Bildungsforschung, Frankfurt.

Göllner, R., Wagner, W., Meurers, D. W., Berendes, K. & Trautwein, U. (2016). *When Questions Unintentially Shape the Answers: Psycholinguistic Item Features Predict Student Ratings of Instructional Quality.* Manuskript in Vorbereitung.

Greenwald, A. G. & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209–1217.

Haag, N., Heppt, B., Stanat, P., Kuhl, P. & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 28,* 24–34.

Hamre, B. K. & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. L. Meece & J. S. Eccles (Hrsg.), *Handbook of Research on Schools, Schooling and Human Development* (S. 25–41). New York, NY: Routledge.

Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement.* London: Routledge.

Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts.* Seelze: Klett-Kallmeyer.

Hox, J. J. & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8,* 157–174.

Jak, S., Oort, F. J. & Dolan, C. V. (2013). A Test for Cluster Bias: Detecting Violations of Measurement Invariance Across Clusters in Multilevel Data. S*tructural Equation Modeling, 20,* 265–282.

Klieme, E., Eichler, W., Helmke, A., Lehmann, R., Nold, G., Rolff, H. G., Schröder, K., Thomé, G. & Willenberg, H. (Hrsg.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI).* Weinheim: Beltz.

Klieme, E., Jude, N., Rauch, D., Ehlers, H., Helmke, A., Eichler, W., et al. (2008). Alltagspraxis, Qualität und Wirksamkeit des Deutschunterrichts. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 319–344). Weinheim: Beltz.

Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: Aufgabenkultur und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–57). München: Medienhaus Biering.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht.* Münster: Waxmann.

Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9,* 231–251.

Kunter, M. & Trautwein, U. (2013). *Psychologie des Unterrichts.* Stuttgart: UTB.

Lenhard, W. & Artelt, C. (2009). Komponenten des Leseverständnisses. In W. Lenhard & W. Schneider (Hrsg.), *Diagnose und Förderung von Leseverständnis und Lesekompetenz* (S. 1–18). Göttingen: Hogrefe.

Lenske, G. (2011). Pupils as raters of instructional quality: Does it work in primary school? In K. Ruhl (Hrsg.), *Das Poster in der Wissenschaft. Zum Stellenwert des Posters in der Nachwuchsförderung am Beispiel der Universität Koblenz-Landau.* Gießen: Johannes Herrmann.

Lüdtke, O., Marsh, H. W., Robitzsch, A. & Trautwein, U. (2011). A 2 x 2 Taxonomy of Multilevel Latent Contextual Models: Accuracy-Bias Trade-Offs in Full and Partial Error Correction Models. *Psychological Methods, 16,* 444–467.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings in multilevel modeling. *Contemporary Educational Psychology, 34,* 120–131.

Lüdtke, O., Trautwein, U., Kunter, M. & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment – A re-analysis of TIMSS data. *Learning Environments Research, 9,* 215–230.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47,* 106–124.

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52,* 1187–1197.

Mehta, P. D. & Neale, M. C. (2005). Peoples are variables too: Multilevel structural equation modeling. *Psychological Methods, 3,* 259–284.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22 (3),* 376–398.

Nunnally, J. C. (1978). *Psychometric theory* (2. Auflage). New York, NY: McGraw-Hill.

Ogrin, S., Keller, S., Friedrich, A., Trautwein, U. & Schmitz, B. (im Druck). Entwicklung und empirische Prüfung einer Lehrkräftefortbildung zur Förderung von Selbstregulationskompetenz und mathematischer Kompetenz bei Schülerinnen und Schülern der Haupt- und Werkrealschule („Lernen mit Plan"). In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals. Interdisziplinäre Betrachtungen, Befunde und Perspektiven.* Berlin: Springer Online.

Rosenshine, B. (1970). Evaluation of classroom instruction. *Review of Educational Research, 40,* 279–300.

Scheerens, J. & Bosker, J. (1997). The foundations of educational effectiveness. Oxford: Elsevier.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77,* 454–499.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11,* 105–126.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Tourangeau, R., Rips, L.J. & Rasinski, K. (2000). *The psychology of survey response.* New York, NY: Cambridge University Press.

Trautwein, U., Lüdtke, O., Klieme, E., Nagengast, B. & Wagner W. (2011). *Erfassung der Unterrichtsqualität in Large-Scale-Studien: Optimierung der Modellierung und Itemauswahl.* Unveröffentlichter BMBF-Antrag. Tübingen: Eberhard Karls Universitat Tübingen.

Vazire, S. & Solomon, B. C. (2015). Self- and other-knowledge of personality. In M. Mikulincer, P. R. Shaver, M. L. Cooper & R. Larsen (Hrsg.), *APA handbook of personality and social psychology, Vol. 4 Personality processes and individual differences* (S. 261–281). Washington, DC: American Psychological Association.

Wagner, W. (2008). *Methodenprobleme bei der Analyse der Unterrichtswahrnehmung und -wirksamkeit – am Beispiel der Studie DESI (Deutsch-Englisch-Schülerleistungen-International) der Kultusministerkonferenz.* Dissertation. Universität Koblenz-Landau, Campus Landau, Fachbereich Psychologie.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U. & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimen-

sionality and generalizability of domain-independent assessments. *Learning and Instruction, 28,* 1–11.

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B. & Trautwein, U. (im Druck). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*.

Widaman, K. F. & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Hrsg.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* Washington, DC: American Psychological Association.

Wubbels, T., Brekelmans, M. & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching And Teacher Education, 8,* 47–58.

Heike Theyßen, Horst Schecker, Martin Dickmann, Bodo Eickhorst, Knut Neumann

Messung experimenteller Kompetenz in Large-Scale-Assessments (MEK-LSA)

1 Hintergrund

Das Experiment spielt in den Naturwissenschaften als Methode der Erkenntnisgewinnung eine zentrale Rolle (vgl. Kircher, Girwidz & Häußler, 2015, 228). Experimentelle Kompetenz ist dementsprechend ein wichtiger Bestandteil naturwissenschaftlicher Bildung. In den Bildungsstandards für den mittleren Schulabschluss im Fach Physik (KMK, 2005a) und den Anforderungen für die Abiturprüfung (KMK, 2005b) wird die Entwicklung experimenteller Kompetenz explizit gefordert. Dies gilt auch für entsprechende Dokumente in anderen Ländern (z. B. NRC, 2012). Dabei werden unter experimenteller Kompetenz in der Regel die Fähigkeiten zur Planung und Durchführung physikalischer Experimente und zur Auswertung der damit gewonnenen Daten verstanden (siehe auch Schreiber, Theyßen & Schecker, 2009; für eine Übersicht vgl. Emden & Sumfleth, 2012).

Um zu überprüfen, in welchem Maße Schülerinnen und Schüler über diese Fähigkeiten verfügen, und um Grundlagen für gezielte Fördermaßnahmen zu schaffen, werden Testverfahren benötigt, die eine valide Erfassung experimenteller Kompetenz bei Schülerinnen und Schülern ermöglichen.

Bislang wird in Large-Scale-Erhebungen experimentelle Kompetenz in der Regel ausschließlich schriftlich erhoben. Die Schülerinnen und Schüler experimentieren nicht selbst, sondern versetzen sich in eine experimentelle Situation und beantworten Fragen dazu. Allerdings haben sich in der einschlägigen Forschung nur geringe Zusammenhänge zwischen den Leistungen von Schülerinnen und Schülern in schriftlichen Tests und ihren Leistungen in Experimentiertests mit Realexperimenten gezeigt (Schreiber, Theyßen & Schecker, 2014; Shavelson, Baxter & Gao, 1993). Dies gilt insbesondere für Fähigkeiten im Bereich der Durchführung von Experimenten, der deshalb in einigen Untersuchungen ausgeklammert wird (z. B. IQB, 2013). Experimentiertests mit Realexperimenten sind hingegen für den Large-Scale-Einsatz sehr aufwendig in Administration und Auswertung (Stecher & Klein, 1997) und werden deshalb nur vereinzelt und in kleineren Stichproben eingesetzt (z. B. TIMSS-Experimentiertest; Stebler, Reusser & Ramseier, 1997). Daraus resultiert eine Diskrepanz zwischen der Bedeutsamkeit des Erwerbs experimenteller Kompetenz in der Praxis und der angemessenen Berücksichtigung experimenteller Kompetenz in Large-Scale-Studien.

Es liegen jedoch aus empirischen Studien Hinweise darauf vor, dass eine valide Diagnostik experimenteller Kompetenz auch mithilfe von Experimentiertests

möglich ist, in denen die Schülerinnen und Schüler statt mit Realexperimenten mit interaktiven Simulationen dieser Realexperimente arbeiten (Schreiber, Theyßen & Schecker, 2014; Shavelson, Ruiz-Primo & Wiley, 1999). Insbesondere im Bereich der Durchführung von Experimenten ermöglicht der Einsatz von Simulationen eine handelnde Auseinandersetzung mit dem Experimentiermaterial, die hinsichtlich der Handlungsmöglichkeiten und der Rückkopplung durch die auf dem Bildschirm dargestellte Reaktion des Experimentiermaterials einer realen Experimentiersituation sehr viel näher kommt als die schriftliche Bearbeitung einer Aufgabe. Gleichzeitig ist ein computergestütztes Testverfahren im Large Scale sehr viel effektiver einsetzbar und auswertbar als Tests mit Realexperimenten.

2 Ziel des Projekts

Das hier vorgestellte Projekt „Messung experimenteller Kompetenz in Large-Scale-Assessments“ geht von der in Kapitel 1 dargestellten Notwendigkeit einer validen Messung experimenteller Kompetenz und dem Mangel an entsprechenden Testverfahren für den Einsatz in großflächigen Untersuchungen mit vielen Teilnehmenden aus. Ziel des Projektes war die Entwicklung und Erprobung eines Tests, der gleichzeitig eine valide Messung experimenteller Kompetenz erlaubt und im Large Scale effektiv einsetzbar ist. Die Anforderungen an einen solchen Test wurden wie folgt konkretisiert:

- Der Test soll die experimentellen Fähigkeiten von Schülerinnen und Schülern in den Bereichen „Planung“, „Durchführung“ und „Auswertung“ erfassen. Damit wird Anschlussfähigkeit an bestehende Modellierungen experimenteller Kompetenz hergestellt. Die explizite Berücksichtigung der Durchführung von Experimenten ermöglicht gegenüber anderen Testverfahren eine umfassendere und damit potenziell validere Messung der experimentellen Fähigkeiten. Zur Standardisierung der Aufgabenkonstruktion wurden die Aufgaben anhand eines Modells experimenteller Kompetenz entwickelt, das aus vorliegenden fachdidaktischen Modellen zur experimentellen Kompetenz abgeleitet wurde.
- Zielgruppe des Tests sollen Schülerinnen und Schüler am Ende der Sekundarstufe I sein. Damit wird unter anderem die Anschlussfähigkeit an andere Large-Scale-Erhebungen gesichert (z. B. PISA: Organisation for Economic Cooperation and Development [OECD], 1999; IQB-Ländervergleich: Pant, Stanat, Schroeders, Roppelt, Siegle & Pöhlmann, 2013). Aus der Zielgruppe leitet sich ab, dass die Aufgaben in Inhalten und Anforderungen typischen Schülerexperimenten der Sekundarstufe I entsprechen müssen.
- Der Test soll vollständig on-screen zu bearbeiten sein und anstelle von Realexperimenten interaktive Simulationen zur Planung und zum Aufbau von Experimenten sowie zur Durchführung und Auswertung von Messungen enthalten. Damit wird einerseits ein effektiver Einsatz im Large Scale ermöglicht und gleichzeitig größtmögliche Nähe zu realen Experimentiersituationen hergestellt. Letzteres lässt gegenüber rein schriftlichen Verfahren eine größere Validität der Ergebnisse erwarten (vgl. Kapitel 1).

3 Der MEK-LSA-Experimentiertest

Der im Rahmen des Projekts entwickelte Experimentiertest basiert auf einem speziell entwickelten Testverfahren. Dieses Testverfahren umfasst die experimentellen Aufgabenstellungen, das Format, in dem diese den Schülerinnen und Schülern zur Bearbeitung präsentiert werden, sowie den Bewertungsmaßstab für die Schülerlösungen.

3.1 Aufgabenstellungen

Damit die Inhalte und Anforderungen der experimentellen Aufgabenstellungen typischen Schülerexperimenten der Sekundarstufe I entsprechen, wurden sie auf Basis umfangreicher Lehrplan- und Schulbuchanalysen (Dickmann & Theyßen, 2013) sowie einer Befragung von Lehrkräften formuliert. Die Aufgabenstellungen decken alle drei Bereiche experimenteller Fähigkeiten – Planung, Durchführung und Auswertung – durch Teilaufgaben ab, insbesondere den bei anderen Tests im Large Scale häufig unterrepräsentierten Bereich der Durchführung. Das zugrunde liegende Aufgabenentwicklungsmodell stellt eine Synthese vorliegender Modelle experimenteller Kompetenz dar und konkretisiert die drei Bereiche experimenteller Fähigkeiten durch acht Teilbereiche (Abbildung 1) (für Details zur Aufgabenkonstruktion siehe Kapitel 3.2).

Abb. 1: Aufgabenentwicklungsmodell

Eine der Aufgaben betrifft beispielsweise die Untersuchung des Zusammenhangs von angehängter Masse und Ausdehnung bei einem Gummiband. Insgesamt wurden zwölf Aufgabenstellungen aus den Themenbereichen Mechanik, Optik und Elektrizitätslehre ausgewählt, deren Anforderungen eine hohe Passung zu typischen schulischen Anforderungen und zum zugrunde gelegten Aufgabenentwicklungsmodell (Abbildung 1) aufweisen.

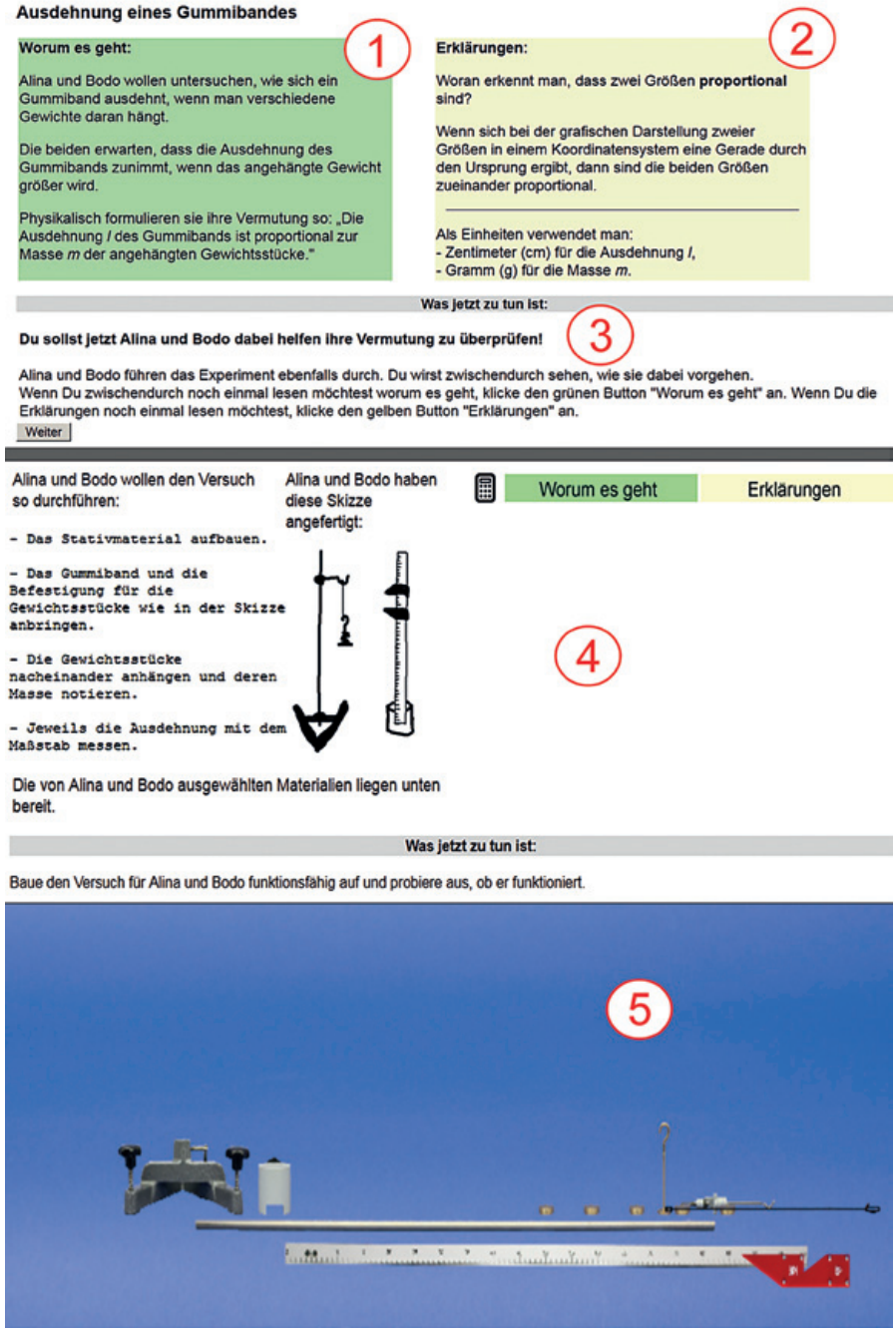
3.2 Aufgabenformat

Eine durchgängige Bearbeitung einer experimentellen Aufgabenstellung, die sowohl die Planung als auch die Durchführung und Auswertung eines Experiments verlangt (vgl. Abbildung 1), kann Schülerinnen und Schüler überfordern. Das führt zu Folgefehlern oder gar dem Abbruch von Aufgabenbearbeitungen. Dadurch wird die Beurteilung weiterer Fähigkeiten erschwert oder unmöglich gemacht, weil ein Schüler z. B. mangels eines funktionsfähigen Versuchsaufbaus keine Messungen mehr durchführen kann (Schreiber, 2012). Deshalb wurden die Aufgabenbearbeitungen durch Teilaufgaben vorstrukturiert. Um Folgefehler zu vermeiden, wird zu Beginn jeder Teilaufgabe eine Zwischenlösung angegeben, auf der die Schülerinnen und Schüler im Folgenden aufbauen sollen. Um die Präsentation der Zwischenlösungen zu motivieren, wurden zwei fiktive Personen eingeführt, Alina und Bodo, die dieselbe Aufgabe bearbeiten. Die Schülerinnen und Schüler sollen sich bei der Bearbeitung der Aufgabe in die Situation versetzen, dass sie einerseits Alina und Bodo helfen und andererseits jeweils mit deren Zwischenergebnissen weiterarbeiten (Abbildung 2: ③).

Eine Testaufgabe enthält einen Aufgabenstamm mit der übergeordneten Aufgabenstellung (Abbildung 2: ①), Fachinformationen zur Bearbeitung der Aufgabenstellung (Abbildung 2: ②) sowie jeweils sechs aufeinander aufbauende Teilaufgaben, deren Anforderungen den im Aufgabenentwicklungsmodell beschriebenen Teilfähigkeiten entsprechen. Die Teilaufgaben „Versuch aufbauen und testen“ und „Messung durchführen und dokumentieren“ (vgl. Abbildung 1) sind in jeder Testaufgabe enthalten. Sie werden gerahmt durch je zwei Teilaufgaben zur Planung und Auswertung. Abbildung 2 zeigt den Aufgabenstamm und die Teilaufgabe „Versuch aufbauen und testen“ für die Beispielaufgabe zur Ausdehnung eines Gummibandes. Die Merkmale des Tests werden im Folgenden anhand dieses Beispiels erläutert.

Je nach Teilaufgabe haben die Schülerinnen und Schüler unterschiedliche Handlungsmöglichkeiten. In der Teilaufgabe „Vorgehensweise angeben“ sollen sie z. B. die für das Experiment nötigen Geräte aus einer Materialbox auswählen (z. B. einen Maßstab und Stativmaterial), eine Skizze des Versuchsaufbaus anfertigen und ihre geplante Vorgehensweise in Stichworten beschreiben. Beim Aufbau wird die Vorgehensweise vorgegeben (Abbildung 2: ④), und die Schülerinnen und Schüler müssen den Versuch entsprechend aufbauen (Abbildung 2: ⑤). Sie können die Geräte frei – auch falsch – anordnen und sollen den Aufbau auf Funktionsfähigkeit testen. Bei der Teilaufgabe zur Durchführung der Messung ist der fertige Aufbau vorgegeben und in seinen Grundzügen nicht mehr zu verändern. Hier können nur noch die relevanten Variablen (im Beispiel die angehängte Masse und die Einstellung der Ablesemarken am Maßstab) verändert und die Messwerte abgelesen werden. Damit wird verhindert, dass durch zufällige Veränderungen am Aufbau Fehler entstehen, die eine Messung – und damit letztlich eine Beurteilung der Schülerfähigkeit zur Durchführung von Messungen – unmöglich machen.

Abb. 2: Aufgabenstamm (oben) und Teilaufgabe zum Aufbau (unten) der Aufgabe zur Ausdehnung eines Gummibandes. (1) Aufgabenstellung, (2) Fachinformation, (3) Kontext „Alina und Bodo“, (4) Zwischenlösung, (5) interaktive Simulation zum Aufbau



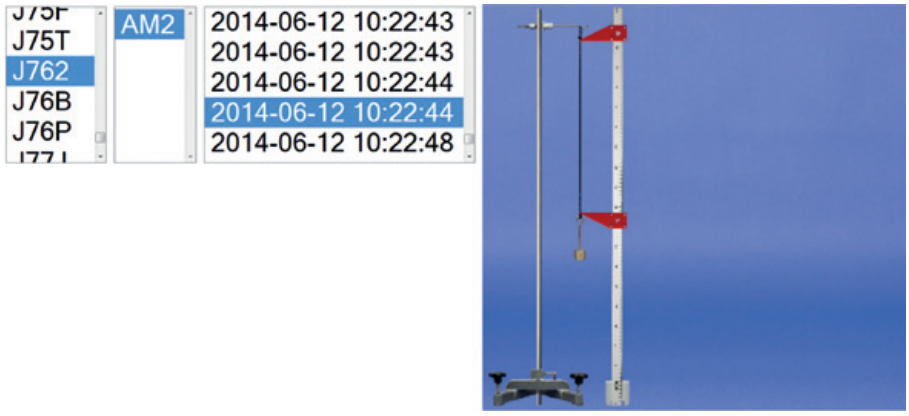
3.3 Bewertungsmaßstab

Während die Schülerinnen und Schüler die Aufgaben bearbeiten, interagieren sie mit den Simulationen, erstellen Skizzen und Diagramme und geben Messwerte oder Texte in Eingabefelder ein. Alle Eingaben werden automatisch in einer Datenbank gespeichert, auch alle Zwischenzustände der Simulationen, Skizzen und Diagramme. So ist anhand der gespeicherten Daten (Logdaten) z. B. nachvollziehbar, in welcher Reihenfolge auf- und gegebenenfalls umgebaut wurde, bei welchen Einstellungen Messwerte aufgenommen wurden oder wie ein Diagramm schrittweise erstellt wurde.

Basis für die Beurteilung der Schülerleistungen sind die oben beschriebenen detaillierten Logdaten, die während der Bearbeitung der Aufgaben automatisch auf dem Server gespeichert werden. Um die als reine Zeichenfolgen gespeicherten Logdaten zu veranschaulichen, wurde ein Softwaretool erstellt, mit dem für eine ausgewählte Teilaufgabe und eine ausgewählte Bearbeitung alle Zustände der Simulationen in chronologischer Folge als Screenshots betrachtet werden können. Abbildung 3 zeigt dieses Interface mit einer aus den Logdaten rekonstruierten Schülerlösung für die Teilaufgabe zum Aufbau aus Abbildung 2.

Die Logdaten erlauben neben dieser anschaulichen Rekonstruktion für die manuelle Auswertung auch eine teilautomatisierte Auswertung. Hierfür wurden für einzelne Teilaufgaben exemplarisch Auswertungsroutinen programmiert. So kann z. B. bei den Teilaufgaben zur Messung automatisch bewertet werden, ob die Anzahl der erhobenen Messwerte ausreicht, ob die in der Simulation erzeugten Messwerte zu den in die Tabelle eingetragenen passen und ob der Bereich der möglichen Messwerte ausreichend genutzt wurde.

Abb. 3: Benutzeransicht des Softwaretools zur Unterstützung der Bewertung von Schülerlösungen; links: Drop-down-Menü zur Auswahl von Schülercode (J762), Teilaufgabe (AM2) und Zeitpunkt (12.6.2014, 10:22 Uhr), rechts: ausgegebene zugehörige Momentaufnahme des Aufbaus



Die Bearbeitungen werden durchweg dichotom (richtig oder falsch bzw. ein oder null Punkte) bewertet. Teilaufgaben mit direkter Interaktion mit dem Material (z. B. Aufbau des Versuches) werden in drei Stufen bewertet (geeignet, teilweise geeignet,

ungeeignet bzw. zwei, ein oder null Punkte; vgl. Eickhorst, Dickmann, Schecker, Theyßen & Neumann, 2015). Die Teilaufgabe „Versuchsplan entwerfen“ wird z. B. als teilweise geeignet bewertet, wenn in dem Aufbau zwar ein grundsätzlich geeigneter experimenteller Ansatz zu erkennen ist, das Experiment aber so noch nicht durchführbar wäre. Ein Beispiel hierfür ist ein geschlossener Stromkreis mit korrekt eingebauter Glühlampe, in dem jedoch die Messgeräte für Spannung und Stromstärke falsch eingebaut sind, sodass – obwohl in der Aufgabenstellung verlangt – keine Messung dieser Größen möglich ist. Zwei Punkte werden vergeben, wenn der Aufbau für die Durchführung des Experiments einschließlich der Gewinnung von Messdaten geeignet ist. Um die Objektivität der Anwendung des Bewertungsmaßstabs sicherzustellen, wurden Manuale erstellt, in denen die Stufen detailliert beschrieben sind. In der Anwendung der Manuale geschulte Auswerter haben die Schülerlösungen codiert. In einer Studie zum Vergleich der Codierungen unterschiedlicher Auswerter zeigten sich zufriedenstellende Übereinstimmungen.

4 Güte des Testverfahrens

Eine Frage, die grundsätzlich an jedes Testverfahren gestellt werden muss, lautet, ob es valide misst, d. h., ob man mit dem Testverfahren tatsächlich das misst, was man messen möchte, in diesem Fall die in Kapitel 1 definierte experimentelle Kompetenz von Schülerinnen und Schülern am Ende der Sekundarstufe I. Diese Frage wurde im Rahmen des hier vorgestellten Projektes in mehreren Studien umfassend untersucht.

Eine erste Voraussetzung für eine valide Messung stellt die umfangreiche Absicherung der Aufgabenauswahl im Hinblick auf die Passung von Inhalten und Anforderungen zu typischen Schülerexperimenten der Sekundarstufe I dar (vgl. Kapitel 3.1). Die weiteren, hier exemplarisch vorgestellten Fragestellungen zur Validität beziehen sich überwiegend auf einen Vergleich des in Kapitel 3.2 beschriebenen computergestützten Aufgabenformats (on screen) mit einem auf inhaltlich gleichen Realexperimenten basierenden Aufgabenformat:

(1) Stellen die Schülerinnen und Schüler während der Bearbeitung der Aufgaben im On-Screen-Format überwiegend physikalisch-experimentelle Überlegungen an oder dominieren andere Überlegungen, z. B. zur Computerbedienung?
(2) Sind die Anteile physikalisch-experimenteller Überlegungen im On-Screen-Format vergleichbar mit den Anteilen, die bei der Bearbeitung entsprechender Realexperimente beobachtet werden?
(3) Ist die kognitive Belastung während der Bearbeitung der Aufgaben in beiden Formaten vergleichbar hoch?
(4) Werden in beiden Formaten vergleichbare Leistungen erzielt?

Die Studien zur Beantwortung dieser Fragestellungen und ihre Ergebnisse werden im Folgenden in Auszügen vorgestellt.

4.1 Überlegungen während der Aufgabenbearbeitung

Zur Beantwortung der Fragen (1) und (2) wurde eine Studie mit ca. 100 Schülerinnen und Schülern der Sekundarstufe I durchgeführt. Es wurden vier Aufgaben aus den verschiedenen Inhaltsbereichen im Detail untersucht. Die Schülerinnen und Schüler bearbeiteten die Aufgaben in Einzelarbeit und äußerten dabei ihre Überlegungen möglichst vollständig. Die Aufgabe zur Ausdehnung eines Gummibandes wurde für die Beantwortung von Frage (2) in ein Realexperiment mit identischen Teilaufgaben und Zwischenlösungen übertragen. Alle Aufgabenbearbeitungen wurden auf Video aufgezeichnet. Zur Auswertung der Aufgabenbearbeitungen wurde ein Kategoriensystem mit fünf Oberkategorien eingesetzt. Der Kategorie „physikalisch-experimentell" werden die eindeutig experimentbezogenen Überlegungen der Schülerinnen und Schüler zugeordnet (z. B. „Die Zeit muss ich hier nicht messen, daher brauche ich keine Stoppuhr"). Überlegungen, die sich auf die experimentelle Aufgabenstellung beziehen, aber keine neuen Gedankengänge der Schülerinnen und Schüler umfassen (z. B. Vorlesen der Aufgabenstellung, Mitsprechen einer Antwort beim Eintippen, ohne dabei neue Überlegungen zu ergänzen), werden in der Kategorie „Reproduktives" erfasst. Überlegungen, die sich ausschließlich auf die Bedienung der interaktiven Simulationen beziehen, gehören zur Kategorie „Computerbedienung" (z. B. „Wenn man hier draufklickt, kann man zoomen"). Analog werden bei der Analyse der Bearbeitung von Realexperimenten alle Überlegungen, die sich auf die Manipulation der Geräte beziehen, der Kategorie „manueller Umgang" zugeordnet (z. B. „Das wackelt hier ganz schön"). Alle weiteren Überlegungen werden in der Kategorie „Sonstiges" erfasst. In Tabelle 1 sind die Verteilungen der Überlegungen exemplarisch für die Aufgabe zur Ausdehnung eines Gummibandes (vgl. Kapitel 3.1) dargestellt. Man erkennt zum einen, dass physikalisch-experimentelle Überlegungen in beiden Testformaten (on screen bzw. reale Experimente) überwiegen, und zum anderen, dass sich die Anteile zwischen den Formaten kaum unterscheiden. Varianzanalysen bestätigen, dass der Anteil physikalisch-experimenteller Überlegungen nicht vom Testformat abhängt (Dickmann, Eickhorst, Theyßen, Schecker & Neumann, 2015).

Tabelle 1: Verteilung der Überlegungen bei der Bearbeitung der Aufgabe zur Ausdehnung eines Gummibands

	Kategorien				
	Physikalisch-experimentell	Reproduktives	Computerbedienung/ Manueller Umgang	Sonstiges	Keine Verbalisierung
Anteil on screen	64 %	9 %	7 %	7 %	13 %
Anteil real	61 %	12 %	4 %	7 %	16 %

4.2 Vergleich von kognitiver Belastung und erzielten Leistungen in beiden Formaten

Zur Beantwortung der Fragen (3) und (4) wurde eine Studie mit 42 Studierenden der Biologie herangezogen, weil der logistische Aufwand für die Bereitstellung der Realexperimente in Schulen zu hoch gewesen wäre. Die physikalische Kompetenz dieser Gruppe ist mit dem Stand am Ende der Sekundarstufe I vergleichbar. Die potenziell größten Unterschiede zwischen den Aufgabenformaten waren für Teilaufgaben mit interaktiven Simulationen zu erwarten. Deshalb wurden gezielt diese Teilaufgaben untersucht. Die Studierenden bearbeiteten bei fünf Aufgaben zunächst beispielsweise die Teilaufgabe „Versuch aufbauen und testen" im Realexperiment, anschließend im On-Screen-Experiment. Um einen Reihenfolgeneffekt, z. B. einen Lerneffekt aus der Bearbeitung des Realexperiments für die On-Screen-Bearbeitung, zu kontrollieren, wurde die Reihenfolge der Formate bei der Hälfte der Teilnehmenden getauscht. Nach der Bearbeitung der Teilaufgaben wurde mit einem Kurzfragebogen die von den Studierenden empfundene kognitive Belastung erhoben. Die Aufgabenbearbeitungen wurden beim On-Screen-Experiment über die Logdaten (vgl. Kapitel 3.3), beim Realexperiment durch Videoaufnahmen dokumentiert, und die Leistungen der Studierenden wurden für beide Formate mit dem in Kapitel 3.3 beschriebenen Bewertungsmaßstab beurteilt. Tabelle 2 zeigt die Korrelationen der im On-Screen-Format und im Realexperiment erzielten Leistungen.

Für die Teilaufgaben „Versuchsplan entwerfen" zeigen sich hohe Zusammenhänge zwischen der Bearbeitungsqualität in beiden Testformaten, die noch über den Werten liegen, die in bisherigen Studien zur Austauschbarkeit von On-Screen- und Realexperimenten gefunden wurden (Schreiber, Theyßen & Schecker, 2014; Shavelson, Ruiz-Primo & Wiley, 1999). Für die Teilaufgaben „Versuch aufbauen und testen" sowie „Messung durchführen und dokumentieren" zeigt sich über die einbezogenen Aufgaben hinweg ein uneinheitliches Bild. Analysiert man die Verteilungen genauer, so liegt die Ursache in der zu geringen Varianz der Bearbeitungsgüte. Insbesondere die Teilaufgaben zum Versuchsaufbau waren für diese Probandengruppe offenbar deutlich zu leicht, sodass die Bearbeitungen fast durchgängig als geeignet beurteilt wurden. Folglich lassen sich für diese Teilfähigkeiten keine statistisch bedeutsamen Zusammenhänge in Form von Rangkorrelationen nachweisen. Andererseits finden sich auch keine schlüssigen Hinweise, die gegen eine Äquivalenz der Testleistungen sprechen. Ganz im Gegenteil: Die Ergebnisse liefern Hinweise darauf, dass es möglich ist, in beiden Testformaten vergleichbare Leistungen zu erzielen. Die Ergebnisse zur empfundenen kognitiven Belastung (Frage (3)) zeigen, dass die durchschnittliche kognitive Belastung in beiden Testformaten eher gering ist und sie sich zwischen den Testformaten nicht unterscheidet.

Tabelle 2: **Korrelationen zwischen den Leistungen im On-Screen-Format und im Realexperiment (** signifikant auf dem Ein-Prozent-Niveau; * signifikant auf dem Fünf-Prozent-Niveau; n. s. = keine signifikante Korrelation)**

Aufgabe	Versuchsplan entwerfen	Versuch aufbauen und testen	Messungen durchführen
Reihenschaltung	.69**	.86**	.47*
Kennlinie	.66**	n. s.	n. s.
Brechung	.85**	n. s.	.66**
Brennweite	.56**	n. s.	n. s.
Gummiband	.63**	n. s.	.62**

5 Ergebnisse zur Stufung und weiteren Struktur experimenteller Kompetenz

Zur Erprobung im Large Scale und um Erkenntnisse über die Stufung und die weitere Struktur experimenteller Kompetenz zu gewinnen, wurde der Test in einer Erhebung mit 1.262 Schülerinnen und Schülern aus den Bundesländern Bremen, Niedersachsen, Nordrhein-Westfalen und Schleswig-Holstein eingesetzt. Die erhobenen Daten wurden entsprechend dem beschriebenen Bewertungsverfahren (Kapitel 3.3) codiert. Zur Prüfung, inwieweit der Test zur Erfassung der experimentellen Kompetenz von Schülerinnen und Schülern am Ende der Sekundarstufe I im Large Scale geeignet ist, wurde der Test in einem ersten Schritt auf der Grundlage eines einfachen (partial credit) Rasch-Modells skaliert (vgl. Neumann, 2014). Die daraus erhaltenen Aufgabenschwierigkeiten und Personenfähigkeiten wurden auf die PISA-Metrik (mit einem Mittelwert der Personenfähigkeiten von 500 Punkten und einer Standardabweichung von 100 Punkten bei einer Lösungswahrscheinlichkeit von 67 Prozent) transformiert. Die vergleichsweise geringe Abweichung der mittleren Aufgabenschwierigkeit um 37 Punkte von der mittleren Personenfähigkeit legt nahe, dass die Anforderungen des Tests im Wesentlichen den Fähigkeiten der Schülerinnen und Schüler entsprechen. Die Verteilung der Aufgabenschwierigkeiten zeigt, dass sowohl im mittleren als auch im unteren und oberen Fähigkeitsbereich ausreichend Aufgaben zur reliablen Einschätzung der Fähigkeiten in der Zielgruppe zur Verfügung stehen. Dies drückt sich auch in der sehr guten Reliabilität der Personenfähigkeitsparameter (WLE) von .84 aus. Der Test ist damit insbesondere auch für die Definition von Kompetenzstufen und die Setzung von Standards für experimentelle Kompetenz geeignet.

5.1 Standards experimenteller Kompetenz

Um die Aufgabenschwierigkeiten und die Fähigkeiten der Schülerinnen und Schüler in inhaltlich interpretierbare Aussagen über den Fähigkeitsstand zu übersetzen, bieten sich Kompetenzstufen an (vgl. Pant, Tiffin-Richards & Köller, 2010). Die Festlegung von Kompetenzstufen beruht auf den Schwierigkeiten der eingesetzten Testaufgaben, stützt sich aber wesentlich auf qualitative Analysen der damit verbundenen Anforderungen. Das Setzen der Schwellenwerte beim Übergang zwischen zwei Stufen und die Beschreibung der Kompetenzstufen wird als Standardsetting bezeichnet (Pant et al., 2010).

Im Projekt wurde hierfür das Bookmark-Verfahren verwendet (z. B. Karantonis & Sireci, 2006). Bei diesem Verfahren werden die Teilaufgaben des Tests in einem Heft in der Reihenfolge steigender Schwierigkeit abgedruckt (Ordered-Item-Booklet, OIB). Eine Gruppe von drei Experten erhielt den Auftrag, in der Folge von 74 Teilaufgaben aus der Elektrizitätslehre und der Optik an denjenigen Stellen im OIB Marken zu setzen, bei denen nach ihrer Einschätzung der Übergang zu einer höheren Fähigkeitsstufe erfolgt. Als Orientierung für die inhaltliche Beschreibung der damit verbundenen experimentellen Fähigkeiten diente den Experten das fünfstufige Kompetenzraster, das zum Kompetenzbereich „Erkenntnisgewinnung“ in der Normierungsstudie zu den nationalen Bildungsstandards für den mittleren Bildungsabschluss entwickelt worden war (Pant et al., 2013).

Die Experten kamen zu drei Schwellenwerten. Der Schwellenwert zu Stufe III liegt mit 481 Punkten knapp unter 500 Punkten, d. h. etwas unterhalb der mittleren Schülerfähigkeit. Diese Tatsache legt es zusammen mit der hohen Besetzungszahl nahe, Stufe III als Regelstandard zu betrachten. Die Schwellenwerte zu den Stufen II bzw. IV liegen mit 412 und 582 Punkten grob im Bereich einer Standardabweichung unterhalb bzw. oberhalb des Mittelwertes.

Die Experten ergänzten die oben genannten Fähigkeitsbeschreibungen und differenzierten sie an einigen Punkten. Die Ergänzungen betreffen insbesondere das Konzipieren und Aufbauen von Experimentieranordnungen, das Durchführen von Messungen und die Auswertung von Daten aus Experimenten. Kompetenzstufe III ist inhaltlich umschrieben als „Anwenden von experimentellen Methoden in einfachen fachlichen Zusammenhängen“. Dazu zählt unter anderem die Fähigkeit, nach gegebener Skizze einfache Experimente vollständig richtig aufzubauen (mit der Vorbereitung für Messungen) und Experimente quantitativ auszuwerten (z. B. durch Diagramme). Etwa 40 Prozent der Testteilnehmenden liegen auf den Stufen I und II und damit unter dem Regelstandard.

5.2 Struktur experimenteller Kompetenz

Die Daten aus der Large-Scale-Studie wurden weiterhin genutzt, um die Modellannahme empirisch zu untersuchen, dass es sich bei experimenteller Kompetenz um einen nach drei Bereichen gegliederten Komplex generischer, d. h. themenübergreifender Fähigkeiten handelt (siehe Kapitel 3.1).

Die vorliegenden Auswertungen zeigen, dass die postulierten drei Bereiche experimenteller Kompetenz (Planung, Durchführung und Auswertung) auch empirisch begründbar zu unterscheiden sind. Zunächst sind die EAP-Reliabilitäten der so gebildeten Teilskalen mit Werten über .70 ausreichend hoch, was auf eine hohe innere Konsistenz hinweist. Ein eindimensionales Rasch-Modell, in dem alle Teilaufgaben unter dem Konstrukt „experimentelle Kompetenz" zusammengefasst werden, zeigt im Vergleich eine signifikant schlechtere Passung zu den Daten aus dem Leistungstest als das aus drei Bereichen gebildete dreidimensionale Modell ($\chi^2(5, N = 1194) = 132.7, p < .001$).

In Verbindung mit dem Experimentiertest wurden in der Large-Scale-Studie auch das Fachwissen der Schülerinnen und Schüler sowie ihre kognitiven Fähigkeiten erhoben. Durch Regressionsanalysen zeigt sich, dass die Bereiche Fachwissen und kognitive Fähigkeiten nur einen geringen Teil der Varianz (unter 15 Prozent) im Experimentiertest aufklären. Daraus lässt sich schlussfolgern, dass die Leistung beim Experimentiertest zwar wie erwartbar mit Intelligenz und Fachwissen zusammenhängt, dass aber die Lösung experimenteller Aufgabenstellungen nicht von diesen Personenmerkmalen dominiert wird.

Die Befunde der Strukturanalysen deuten darauf hin, dass es sich bei experimenteller Kompetenz um einen abgrenzbaren und gegliederten Bereich physikalischer Kompetenz handelt, in dem Fachwissen für die Bearbeitung experimenteller Aufgaben mit heranzuziehen ist, der aber keine schlichte Anwendung von Fachwissen darstellt.

6 Zusammenfassung und Ausblick

Das zentrale Ziel des hier dargestellten Projektes war die Entwicklung eines Testverfahrens, mit dem die experimentelle Kompetenz von Schülerinnen und Schülern am Ende der Sekundarstufe I valide gemessen werden kann und das auch in großflächigen Erhebungen mit vielen Teilnehmenden effizient einsetzbar ist. Unter experimenteller Kompetenz wird dabei die Fähigkeit zur Planung und Durchführung physikalischer Experimente sowie zur Auswertung der damit gewonnenen Daten verstanden (vgl. Kapitel 2). Dieses Ziel wurde erreicht.

Die Validierungsstudien liefern zahlreiche Belege dafür, dass die mit dem computergestützten Testverfahren erhobenen Schülerleistungen zuverlässig als Ausdruck experimenteller Kompetenz interpretiert werden können. Im Gegensatz zu anderen Studien wird durch den Einsatz interaktiver Simulationen insbesondere der Bereich der Durchführung mit erfasst. Die Erprobung mit mehr als 1.200 Schülerinnen und Schülern hat gezeigt, dass mit dem vorliegenden Testverfahren sowohl die Durchführung der Erhebung als auch die Auswertung der Testbearbeitungen bei großen Stichproben praktikabel und effizient sind. Besonderen Anteil daran haben die im Projekt entwickelten Verfahren zur Erfassung und teilautomatisierten Auswertung detaillierter Logdaten. Darüber hinaus liefert die Auswertung der Large-Scale-Daten neue Erkenntnisse über die Struktur und Stufung experimenteller Kompetenz.

Ein weiterer wesentlicher Schritt besteht darin nachzuweisen, dass das Testverfahren ausreichend empfindlich misst, um die Entwicklung experimenteller Kompetenz auch auf Ebene der einzelnen Fähigkeiten sowie individualdiagnostisch nachzuzeichnen. In diesem Fall können gezielte Interventionen zur Förderung experimenteller Kompetenz in weiterführenden Studien hinsichtlich ihrer Wirkung vergleichend untersucht und auf Basis der Ergebnisse optimiert werden.

Bislang wurde der Status quo experimenteller Kompetenz innerhalb einer nicht repräsentativen Stichprobe erhoben. Erst der Einsatz in repräsentativen Stichproben, wie sie für PISA-Studien oder andere Vergleichsstudien herangezogen werden, kann differenziert Aufschluss über die Ausprägung experimenteller Kompetenz bei deutschen Schülerinnen und Schülern und über Förderbedarfe geben. Die Voraussetzungen dafür wurden mit den Ergebnissen des vorliegenden Projektes geschaffen.

Literaturverzeichnis

Dickmann, M., Eickhorst, B., Theyßen, H., Schecker, H. & Neumann, K. (2015). Testinstrument für experimentelle Kompetenz: Einfluss des Testformats auf konstruktbezogene Denkprozesse. In S. Bernholt (Hrsg.), *Heterogenität und Diversität – Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Bremen 2014* (S. 663–665). Kiel: IPN.

Dickmann, M. & Theyßen, H. (2013). Curriculare Validität von Units zur Messung experimenteller Kompetenz. In S. Bernholt (Hrsg.), *Inquiry-based Learning – Forschendes Lernen. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012* (S. 587–589). Kiel: IPN.

Eickhorst, B., Dickmann, M., Schecker, H., Theyßen, H. & Neumann, K. (2015). Messung experimenteller Kompetenz im Large-Scale: Bewertung experimenteller Aufgaben. In S. Bernholt (Hrsg.), *Heterogenität und Diversität – Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Bremen 2014* (S. 169–171). Kiel: IPN.

Emden, M. & Sumfleth, E. (2012). Prozessorientierte Leistungsbewertung des experimentellen Arbeitens. Zur Eignung einer Protokollmethode zur Bewertung von Experimentierprozessen. *Der mathematische und naturwissenschaftliche Unterricht, 65 (2)*, 68–75.

Institut zur Qualitätsentwicklung im Bildungswesen (IQB). (2013). *Kompetenzstufenmodelle zu den Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Kompetenzbereiche „Fachwissen" und „Erkenntnisgewinnung"* – Entwurf. Abgerufen am 11.07.2015 von https://www.iqb.hu-berlin.de/bista/ksm/KSM_Physik.pdf.

Karantonis, A. & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25 (1),* 4–12.

Kircher, E., Girwidz, R. & Häußler, P. (2015). *Physikdidaktik – Theorie und Praxis* (3. Auflage). Berlin: Springer.

National Research Council (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, DC: The National Academies Press.

Neumann, K. (2014). Raschanalyse naturwissenschaftsbezogener Leistungstests. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden der naturwissenschaftsdidaktischen Forschung* (S. 355–370). Berlin: Springer.

Organisation for Economic Co-operation and Development (1999). *Measuring Student Knowledge and Skills. A New Framework for Assessment.* Paris: OECD.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster: Waxmann.

Pant, H. A., Tiffin-Richards, S. P. & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment. *Zeitschrift für Pädagogik, Beiheft 56,* 175–188.

Schreiber, N. (2012). *Diagnostik experimenteller Kompetenz – Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells.* Berlin: Logos.

Schreiber, N., Theyßen, H. & Schecker, H. (2009). Experimentelle Kompetenz messen? *Physik und Didaktik in Schule und Hochschule, 5/3,* 92–101.

Schreiber, N., Theyßen, H. & Schecker, H. (2014). Diagnostik experimenteller Kompetenz: Kann man Realexperimente durch Simulationen ersetzen? *Zeitschrift für Didaktik der Naturwissenschaften, 20 (1),* 161–173. doi: 10.1007/s40573-014-0017-1.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005a). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss.* München: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005b). *Einheitliche Prüfungsanforderungen in der Abiturprüfung Physik.* München: Luchterhand.

Shavelson, R. J., Baxter, G. P. & Gao, X. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement, 30 (3),* 215–232.

Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. W. (1999). Notes on Sources of Sampling Variability in Science Performance Assessments. *Journal of Educational Measurement, 36 (1),* 61–71.

Stebler, R., Reusser, K. & Ramseier, E. (1997). Spitzenleistungen der Schweizer Siebtklässler im TIMSS-Experimentiertest 1721. *SLZ, 10,* 18–21.

Stecher, B. M. & Klein, S. P. (1997). The Cost of Performance Assessments in Large-Scale Testing Programs. *Educational Evaluation and Policy Analysis, 19 (1),* 1–14.

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

Entwicklung und Validierung eines historischen Kompetenztests zum Einsatz in Large-Scale-Assessments (HiTCH)

1 Einleitung

„Geschichte denken statt pauken" (Schreiber & Mebus, 2005; Ventzke, Mebus & Schreiber, 2010) ist das Credo der nationalen und internationalen Diskussion über die Ziele historischen Lernens bzw. der history education (Köster, Thünemann & Zülsdorf-Kersting, 2014). Derzeit ist zwischen Geschichtsdidaktikerinnen und -didaktikern in westlich geprägten Demokratien weitgehender Konsens darüber zu beobachten, dass im Geschichtsunterricht nicht primär fallbezogenes Faktenwissen, sondern vielmehr grundsätzliche Einsichten in historische Denkweisen und Methoden historischer Erkenntnis vermittelt werden sollen. Entsprechende Überlegungen dazu finden sich z. B. in den Vereinigten Staaten (Mandell & Malone, 2008; Stearns, 1998; VanSledright, 2014; Wineburg, 2001), in Kanada (Seixas, 2008; Seixas & Morton, 2013), in Europa (vgl. Überblick in Erdmann & Hasberg, 2011; van Drie & van Boxtel, 2008) und in Australien (Taylor & Young, 2003). Der Geschichtsunterricht in demokratischen Staaten soll die Lernenden dazu befähigen, mit den vielfältigen Geschichtsdeutungen reflektiert und (selbst-)reflexiv umgehen zu können.

Ziel des Projekts HiTCH (**Hi**storical **T**hinking: **C**ompetencies in **H**istory) war es, ausgehend von einem theoretisch ausformulierten Kompetenzstrukturmodell historischen Denkens, einen den üblichen psychometrischen Gütekriterien entsprechenden Test vorzulegen, mit dessen Hilfe auch in größeren Schulleistungsstudien überprüft werden kann, inwieweit Schülerinnen und Schüler der neunten Jahrgangsstufe aller Schulformen über historische Kompetenzen verfügen. Der Fokus lag auf der Entwicklung standardisierter Aufgaben, die in einem überschaubaren zeitlichen Rahmen bearbeitet und ausgewertet werden können. Hierfür war eine enge interdisziplinäre Zusammenarbeit zwischen Fachdidaktik Geschichte und empirischer Bildungsforschung notwendig. Antragsteller des Projekts waren das Hector-Institut

für Empirische Bildungsforschung an der Universität Tübingen, die Fachdidaktiken Geschichte an der Universität Hamburg und der Katholischen Universität Eichstätt-Ingolstadt sowie das Leibniz-Institut für Wissensmedien, Tübingen (Trautwein, Körber, Borries, Schreiber, Schwan & Bertram vom 16.05.2011, Zuwendungsbescheid für LSA006 am 26.01.2012). Im Verlauf des Projekts hat sich darüber hinaus eine enge Kollaborationsbeziehung mit Geschichtsdidaktikerinnen und -didaktikern der Ruhr-Universität Bochum um Nicola Brauch, der Pädagogischen Hochschule Salzburg um Christoph Kühberger und der Pädagogischen Hochschule der Fachhochschule Nordwestschweiz um Béatrice Ziegler und Monika Waldis entwickelt.

2 Kompetenzen historischen Denkens

2.1 Geschichtsdidaktische und geschichtstheoretische Grundlagen

Seit dem PISA-Schock im Jahr 2000 und der nachfolgenden Klieme-Expertise (2003) wurden auch in der Geschichtsdidaktik die Bemühungen verstärkt, Kompetenzmodelle zu entwickeln, die die Grundlage für eine vergleichende Leistungsmessung bieten. Die Narrativitätstheorie und das darauf aufbauende Konzept des Geschichtsbewusstseins stellen die theoretische Basis der vorgelegten Kompetenzmodelle historischen Denkens im deutschsprachigen Kontext dar (unter anderem Gautschi, 2009; Hasberg, 2006/2010; Heil, 2011; Körber, Schreiber & Schöner, 2007; Pandel, 2005; Verband der Geschichtslehrer Deutschland, 2007; vgl. die Diskussion in Barricelli, Gautschi und Körber, 2012). Einige vergleichbare Modellierungen historischen Denkens („Kognitionsmodelle"; Ercikan & Seixas, 2015) aus dem internationalen Kontext anerkennen Narrativität als Strukturprinzip und erkenntnistheoretische Prämisse von Geschichte und bestimmen sowohl den Ausgangspunkt als auch die Ziele historischen Lernens auf dieser Basis (z. B. Ercikan & Seixas, 2015; Seixas & Morton, 2013; Stearns, Seixas & Wineburg, 2000).

Die Narrativitätstheorie wurde entscheidend von Artur C. Danto (1965) geprägt und unter anderem von Paul Ricoeur (1983/1988) und in Deutschland maßgeblich von Jörn Rüsen (unter anderem 1983, 2013) ausformuliert. Ihr zufolge wird zwischen der „Vergangenheit" als der Realität in zurückliegender Zeit und der „Geschichte" als der Form späterer Darstellungen zu dieser bzw. diesen Vergangenheiten unterschieden. Die Auseinandersetzung mit Vergangenem nimmt als „Geschichte" notwendigerweise eine sprachliche Form an. In einer historischen „Narration" werden retrospektiv mindestens zwei verschiedene, zeitlich differente Ereignisse sinnhaft miteinander verknüpft, sodass eine sprachlich vermittelte Verlaufsstruktur entsteht (unter anderem Barricelli, 2012; Pandel, 2013; Rüsen, 1983). Eng mit der Narrativitätstheorie verknüpft ist das Konzept des Geschichtsbewusstseins, welches im deutschsprachigen Kontext seit den 1970er-Jahren diskutiert und erforscht wurde (unter anderem Borries, 1988; Jeismann, 1977, 1988; Pandel, 1987; Rüsen, 1983, 1994). Über die Rezeption der englischsprachigen Arbeiten von Rüsen (z. B. 2005) wurde das „concept

of historical consciousness“ (Clark, 2014, 84) teilweise auch im angelsächsischen und skandinavischen Raum wirksam (Jensen, 2003; Seixas, 2011). Geschichtsbewusstsein wird gemeinhin als ein die Grenzen des Unterrichtsfachs übersteigendes und die Gesellschaft adressierendes Konzept verstanden. Jeismann (1977) definierte das Geschichtsbewusstsein als den Zusammenhang zwischen Vergangenheitsdeutung, Gegenwartsverständnis und Zukunftserwartung. Hiermit stellte er – wie auch Rüsen (1983, 2013) oder Schreiber et al. (2006) – heraus, dass Geschichte Vergangenes von heute aus betrachtet und eine Orientierungsfunktion für die Gegenwart und Zukunft besitzt.

Die Fähigkeit zum kritischen historischen Denken basiert auf der begrifflichen Trennung von Vergangenheit und Geschichte sowie auf der Unterscheidung von und dem kritischen Umgang mit Quellen und Darstellungen. Im Geschichtsunterricht in Deutschland steht seit den 1970er-Jahren der regelgeleitete Umgang mit Quellen (Quellenkritik, Quellenanalyse und -interpretation) im Vordergrund (Schneider, 2010). In ihrem Alltag jedoch werden Schülerinnen und Schüler sehr viel häufiger mit historischen Narrationen (z. B. in Filmen, TV-Geschichtsdokumentationen, multimedialen Darstellungen im Internet, in Computerspielen, politischen Reden oder Gedenkveranstaltungen) konfrontiert. Um mit der sie umgebenden Geschichtskultur reflektiert und kritisch umgehen zu können und (selbst-)reflexiv nach deren Bedeutung für historische Orientierung zu fragen, sollten sich die Lernenden der epistemologischen Voraussetzungen von geschichtlichem Wissen bewusst sein und demzufolge befähigt werden, historische Narrationen zu dekonstruieren (Hasberg & Körber, 2003; Körber et al., 2007; Schreiber, 2002).

Die grundsätzlich perspektivische Verfasstheit von Geschichte wurde mit der Durchsetzung der narrativen Geschichtstheorie in Deutschland (Koselleck, Mommsen & Rüsen, 1977) für Quellen wie auch Darstellungen als konstitutives Element historischer Erkenntnis offengelegt (Baumgartner, 1997; Bergmann, 2007; Borries, 2000/2004; Lücke, 2012; Rüsen, 1983; Schöner, 2007). Das Prinzip der Perspektivität wird auf drei Ebenen bezogen (unter anderem Bergmann, 2007; Borries, 2000/2004): (1) Auf der Ebene der Quellen wird aufgrund der Sichtweisen der Menschen, die in der Vergangenheit in einen historischen Sachverhalt verstrickt waren, von einer Multiperspektivität gesprochen. (2) Auf der Ebene der Darstellungen entsteht die Kontroversität durch unterschiedliche Sichtweisen der retrospektiv auf einen historischen Sachverhalt blickenden Historiker, Ausstellungs- und Filmemacher, politische Redner usw. (3) Auf der Ebene der historischen Orientierung muss eine Pluralität der Urteile über einen historischen Sachverhalt ermöglicht werden. Die Triftigkeit bzw. Plausibilität historischer Darstellungen und Urteile (Rüsen, 1983, 2013; Schreiber et al., 2006) wird in einer pluralen, auf den kritischen Umgang mit Quellen und Gegenwartsdeutungen rekurrierenden Diskussion ausgehandelt. Die kritische Auseinandersetzung mit und um Geschichte erfolgt also perspektivisch, aber methodisch kontrolliert. Strukturell liegen dem Prozess die beiden Basisoperationen des historischen Denkens, die Re- und Dekonstruktion, zugrunde.

Die später als „FUER“[1] bezeichnete Projektgruppe von Geschichtsdidaktikerinnen und Geschichtsdidaktikern hat diese Grundlagen in der Modellierung der „Sechs-Felder-Matrix“ (Schreiber, 2002) und im Konzept des „Geschichtsbewusstsein dynamisch“ (Hasberg & Körber, 2003) zusammengefasst und zum Ausgangspunkt für die Entwicklung des Kompetenzstrukturmodells historischen Denkens gemacht, das den heuristischen Rahmen für die Entwicklung des HiTCH-Tests bildete.

2.2 Das FUER-Modell: ein Kompetenzstrukturmodell historischen Denkens

Das FUER-Modell basiert auf der Vorstellung eines Regelkreises des historischen Denkens. Die Kompetenzbereiche der historischen Frage-, Methoden-, Orientierungs- und Sachkompetenzen leiten sich aus diesem Konzept ab. Ausgangspunkt bildet die Feststellung, dass Menschen das Bedürfnis haben, sich in der Zeit zu orientieren (Hasberg & Körber, 2003; Jeismann, 1977; Körber et al., 2007; Rüsen, 1983). Durch Verunsicherung und Neugierde werden temporale Orientierungsprozesse in Gang gesetzt. Die Fähigkeit, Fertigkeit und Bereitschaft, diese in einer historischen Fragestellung zu bündeln, wird als historische Fragekompetenz bezeichnet. Sie umfasst sowohl die Formulierung eigener Fragen als auch den gegenläufigen Prozess, Fragestellungen anderer zu erschließen. Wenn sich eine Frage an die Vergangenheit richtet und zu ihrer Beantwortung Quellen und Darstellungen untersucht werden, beginnt ein Rekonstruktionsprozess. Nimmt die Frage hingegen vorliegende Narrationen analysierend in den Blick, initiiert sie einen Dekonstruktionsprozess. Die Fähigkeit, die beiden Basisoperationen der Re- und Dekonstruktion auszuführen, wird als Methodenkompetenz bezeichnet. Im Ergebnis führen Re- und Dekonstruktionsprozesse zu einer eigenen historischen Narration bzw. zu einer beurteilenden Stellungnahme zu vorhandenen Darstellungen. Die vorliegenden Ergebnisse können zur historischen Orientierung genutzt werden, indem das Welt-, Selbst- und Fremdverstehen erweitert wird, historisch fundierte Handlungsdispositionen entwickelt und die individuellen Vorstellungen von und Einstellungen zu einer Geschichte gegebenenfalls modifiziert werden. Die Fähigkeit zur Ausführung solcher Handlungen wird als historische Orientierungskompetenz bezeichnet. Im Rahmen des an verschiedenen Themen und Fragestellungen immer wieder durchlaufenen Prozesses historischen Denkens bilden sich die historischen Sachkompetenzen heraus, d. h., historisch Denkende verfügen in zunehmendem Maße über die Prinzipien, Konzepte und Skripts, die für den Umgang mit Geschichte relevant sind. Sie können inhalts-, theorie-, methoden- und subjektbezogen mit Begriffen und den in ihnen zum Ausdruck gebrachten Strukturierungen umgehen.

1 Das Akronym „FUER Geschichtsbewusstsein“ steht für eine internationale Gruppe von Geschichtswissenschaftler/-innen, Fachdidaktiker/-innen und Lehrer/-innen, die es sich zum Ziel gesetzt haben, die Förderung eines reflektierten und (selbst-)reflexiven Geschichtsbewusstseins durch Grundlagenforschung und Empirie zu präzisieren (Schreiber et al., 2006).

3 Erfassung von Kompetenzen historischen Denkens

In der empirischen Forschung zur Erfassung historischer Kompetenzen dominieren im angloamerikanischen (Baron, 2012; Rouet, Britt, Mason & Perfetti, 1996; Wineburg, 1991) wie auch im deutschsprachigen Raum bisher die qualitativen Methoden. Ein Blick in die Tagungsbände der seit 2007 im Zweijahresrhythmus stattfindenden Tagung „geschichtsdidaktik empirisch" (Hodel & Ziegler, 2009, 2011; Hodel, Waldis & Ziegler, 2013; Waldis & Ziegler, 2015) zeigt, dass in den empirischen Studien der Geschichtsdidaktik Daten meist in Form von Essays, Interviews oder Protokollen lauten Denkens erhoben und mit unterschiedlichen qualitativen Methoden ausgewertet werden (vgl. auch Köster, Thünemann & Zülsdorf-Kersting, 2014). Geschichte zählt zu den schwach strukturierten Domänen (unter anderem Langer-Plän & Beilner, 2006); historische Kompetenzen werden demzufolge als „schwer messbar" eingeschätzt (Körber et al., 2008). Tatsächlich gibt es spezifische Herausforderungen bei der Entwicklung standardisierter Tests zur Erfassung historischer Kompetenzen, die im Folgenden benannt werden.

3.1 Zentrale Herausforderungen

Konstruktcharakter von Geschichte. Dem Konstruktcharakter von Geschichte ist geschuldet, dass für Fragestellungen nicht einfach eine richtige Antwort gefunden werden kann. Quantitative Testverfahren, die auf einem Richtigkeitsstandard (vgl. Eid, Gollwitzer & Schmitt, 2010) beruhen, scheinen dieser konstruktivistischen Verfasstheit von Geschichte und den hieraus abgeleiteten historischen Kompetenzen zu widersprechen. Zugleich gilt aber, dass die erzählte(n) Geschichte(n) nicht beliebig sind, sondern in ihrer Triftigkeit/Güte/Plausibilität eingeschätzt werden können. Somit besteht die Herausforderung darin, standardisierte Aufgabenformate zu entwickeln, die reliabel und valide zeigen, inwiefern die Probanden über gesellschaftlich anerkannte Konzepte und Fähigkeiten zur historischen Orientierung verfügen.

Prozessuale Kompetenzen. Über welche Ausprägungen historischen Denkens Probanden verfügen, wird in actu am deutlichsten, wenn sie also historische Narrationen erstellen oder sich analytisch zu gegebenen Narrationen verhalten (Körber, Borries, Pflüger, Schreiber & Ziegler, 2008). Dies in einem standardisierten Test zu berücksichtigen bedeutet, Aufgaben zu konstruieren, die zu aktivem historischen Denken auffordern und ermöglichen, das Verfügen über benötigte Konzepte und Fähigkeiten zu messen.

Kontext- und Themengebundenheit der Aufgaben. Da sich die historischen Kompetenzen im Umgang mit konkreten historischen Fragen, im methodengeleiteten Umgang mit Quellen und Darstellungen sowie in der sorgfältigen Begründung von Orientierungen zeigen, müssen Testaufgaben zum historischen Denken bzw. Handeln herausfordern. Sie sollten demnach in einem historischen Kontext eingebettet sein, und es müssen Materialien mitgeliefert werden. Daraus ergibt sich bei der Aufgabenkonstruktion die Herausforderung, den Lese- und Zeitaufwand nicht zu hoch

werden zu lassen. Zudem sollten Effekte, die sich aus spezifischem Vorwissen und der Einstellung zu einzelnen Themen ergeben, begrenzt werden. Historische Kompetenz sollte daher an verschiedenen Themen erfasst werden. Das notwendige Kontextwissen muss in der Aufgabenstellung jeweils mitgeliefert werden, damit der Test erfassen kann, inwiefern die Lernenden über die für die Bearbeitung der Aufgaben notwendigen Konzepte und Struktureinsichten in die Natur historischen Denkens und dessen Prozesse und Verfahren verfügen (vgl. hierzu auch die Kompetenzdefinition von Weinert, 2001).

Curriculumvorgaben und Wissensvoraussetzungen. Die Leistung im HiTCH-Test soll – unter anderem in Hinblick auf einen (bundes-)länder- und schulartübergreifenden Einsatz – unabhängig davon sein, dass bestimmte Inhalte zuvor im Unterricht behandelt worden sind. Der Test soll vielmehr solche Einsichten, Konzepte und Fähigkeiten erfassen, die an durchaus unterschiedlichen Themen erworben und ausdifferenziert wurden.

Geschichtskultur. Die umgebende(n) Geschichtskultur(en) hat/haben Einfluss auf die Einstellungen der Probandinnen und Probanden. Es ist nicht die Aufgabe des HiTCH-Tests, derartige Einflüsse zu erfassen oder zu klassifizieren. Vielmehr geht es darum, bei der Themenauswahl und Itemkonstruktion geschichtskulturelle Aspekte zu berücksichtigen, indem beispielsweise unterschiedliche Erinnerungstraditionen zum Thema einer Dekonstruktionsaufgabe gemacht werden.

3.2 Erfahrungen mit der Erfassung historischer Kompetenz

Im deutschsprachigen Kontext haben standardisierte Tests im Fach Geschichte – nach einem kurzen Vorlauf in den 1970er-Jahren (z. B. Borries, 1974) – eine untergeordnete Rolle gespielt. In empirischen Untersuchungen der 1980er- und 1990er-Jahre zum Konzept des „Geschichtsbewusstseins" wurde dieses nicht als eine überprüfbare „Kompetenz" verstanden, über die die Schülerinnen und Schüler – vermittelt durch den Geschichtsunterricht – zu verfügen lernen. Vielmehr wurde das Konzept in „Sozialisationspraktiken" (Familien, Massenmedien, Schule), „Einstellungen" (Vorlieben, Überzeugungen, Selbstdefinitionen) und „Fähigkeiten" (Konzeptbeherrschung, Transferleistungen, Denkstrategien) aufgefächert. Allerdings wurden bezogen auf den letzten Teilbereich Aufgaben vorgelegt (unter anderem Angvik & Borries, 1997; Borries, Fischer, Leutner-Ramme & Meyer-Hamme, 2005), die als Vorstufe zur Messung von Kompetenzen historischen Denkens verstanden werden können, etwa Aufgaben zum Vergleich von Schulbuchauszügen oder von widersprüchlichen Darstellungen zum gleichen Thema (vgl. Dekonstruktionskompetenz), Aufgaben zu einem „experimentellen Perspektivenwechsel" (vgl. Rekonstruktionskompetenz) oder zu zentralen historischen Begriffen (vgl. Sachkompetenz).

Erst in den vergangenen Jahren wurden im deutschsprachigen Raum Instrumente vorgelegt, die Teilbereiche historischer Kompetenzen mit standardisierten Aufgabenformaten erfassen sollen. Die Studie von Hartmann (2008) nahm die historische Perspektivenübernahme in den Blick, während Bertram, Wagner und Trautwein (2013, 2014) auf die Einsicht der Lernenden in die epistemologischen Prinzipien und auf das

Begriffsverständnis als Facetten der Sachkompetenz fokussierten. Bisher fehlen jedoch Instrumente, die die Breite des Konstrukts „historisches Denken" adressieren.

In den Vereinigten Staaten haben Large-Scale-Assessments auch im Fach Geschichte eine sehr viel längere Tradition. Im Rahmen des National Assessment of Educational Progress (NAEP) werden seit den 1960er-Jahren regelmäßig Lernende verschiedener Jahrgangsstufen (vierte, achte und zwölfte Klasse) landesweit getestet. Dabei sollen nicht nur ihre Kenntnisse zur US-amerikanischen Geschichte, sondern auch die Denkoperationen des knowing und thinking history erfasst werden. Während unter knowing (genauer: historical knowledge and understanding) historisches Wissen, aber auch Sichtweisen verstanden werden, sollen unter thinking (genauer: historical analysis and interpretation) Kompetenzen historischen Denkens erfasst werden, wie z. B. die Herstellung von Kausalbeziehungen oder das Abwägen von Beweisen (http://nagb.org/publications/frameworks.htm). Bei einer genaueren Analyse dieser Aufgaben zeigt sich allerdings, dass trotz der anderslautenden Ansprüche meist Kenntnisse auf einem eher bescheidenen Reflexionsniveau abgefragt werden (Lazer, 2015; VanSledright, 2014).

Die Übersicht über die bisher vorliegenden Testinstrumente zeigt, dass zumindest für den deutschen Sprachraum ein Geschichtstest fehlt, der eine Messung historischer Kompetenzen mit standardisierten Aufgabenformaten erlaubt. Die Entwicklung des HiTCH-Tests, der sich diesen Anforderungen stellt, soll im Folgenden skizziert werden.

4 Methode

4.1 Aufgabengenerierung/cognitive labs

Für die Entwicklung des HiTCH-Tests, der historische Kompetenzen in der ganzen Bandbreite des FUER-Modells reliabel und valide erfassen sollte, wurden zur Vermeidung von construct underrepresentation (Messick, 1995) alle entsprechenden Kompetenzdimensionen mit ihren Unterfacetten adressiert. Mit leichteren und anspruchsvolleren Aufgaben wurde auf die heterogene Schülerschaft des Samples 15-Jähriger reagiert. Zudem wurde angestrebt, die Aufgaben und Items sowohl auf curriculumnahe als auch -ferne, sowohl auf geschichtskulturell eher präsente als auch eher vernachlässigte Themen zu beziehen. Durch die Variation des Themenbezugs und die Verwendung möglichst kurzer und nicht zu komplexer Materialien sollte gesichert werden, dass nicht Faktoren wie Lesekompetenz, Kenntnisse von, Vorstellungen über, Einstellungen zu oder Interesse an einem Thema im Sinne von construct-irrelevant variance (Messick, 1995) die Messung der Kompetenzen historischen Denkens gravierend beeinflussen.

Bei der Aufgabengenerierung wurden vorhandene Messinstrumente mitberücksichtigt (Bertram et al., 2013, 2014; Borries et al., 2005; Hartmann, 2008). Anregungen für die Formulierung von geschlossenen Aufgaben wurden auch aus qualitativen Studien gewonnen (Kraus, 2013; Zabold, im Druck). An standardisierten Aufgaben-

formaten wurden vor allem Ordnungsaufgaben (Zuordnung oder Umordnung) und Auswahlaufgaben (z. B. dichotome Aufgaben, in denen zwischen richtig oder falsch gewählt werden soll, Multiple-Choice- oder Single-Choice-Aufgaben) genutzt. Generell wurde bei der Wahl des Aufgabenformats die Einfachheit oder Komplexität der Testanweisung, der Zeitaufwand und die Ratewahrscheinlichkeit bei der Lösung der Aufgabe wie auch der Auswertungsaufwand gegeneinander abgewogen (Jonkisz, Moosbrugger & Brandt, 2012). Aufgaben mit freiem Aufgabenformat, bei denen die Antwort von der Testperson selbst formuliert wird (z. B. Kurzessays oder Ergänzungsaufgaben), spielten in der HiTCH-Testentwicklung eine untergeordnete Rolle.

Die neu generierten Aufgaben und Items wurden zunächst mit einzelnen Testpersonen oder in kleinen Gruppen aus der Zielpopulation erprobt, um die Verständlichkeit der Items und die Schwierigkeiten der Aufgaben zu evaluieren. Systematisierend wurden sogenannte cognitive labs eingesetzt, in denen die spontanen Äußerungen der Probanden durch situative Nachfragen (prompts) zum Verständnis der Aufgabe und Lösungsstrategien oder zur Begründung, warum eine bestimmte Antwortalternative gewählt wurde, ergänzt werden (Willis, 2005). Die Schüleräußerungen waren wichtige Indikatoren dafür, ob mit den geschlossenen Formaten tatsächlich historisches Denken angestoßen wurde (Werner & Schreiber, 2015).

4.2 Konstruktionslogiken

Die Konstruktionslogiken für die Aufgaben- und Itementwicklung können im Folgenden nur kurz und kursorisch vorgestellt werden. Ähnlich den Erfahrungen in der Mathematik erwies es sich auch für Geschichte schwierig, Aufgaben zu konzipieren, welche jeweils einen einzigen Kompetenzbereich (gemäß FUER-Modell) rein adressieren. Dass unterschiedliche Misch- und Beiladungen zu verzeichnen sind, ergibt sich unter anderem aus den Überlappungen zwischen den vier Kompetenzbereichen.

Die Überlappungsbereiche bilden jeweils Ausschnitte aus dem idealtypischen historischen Denk- und Erkenntnisgewinnungsprozess ab. Dies kann für die Aufgabenkonstruktion genutzt werden. Dem Überlappungsbereich von Frage- und Methodenkompetenz gehören etwa Aufgaben an, bei denen es darum geht, zu einer bereits feststehenden historischen Fragestellung zielführende Materialien auszuwählen oder einzuschätzen, welche historischen Fragestellungen mit vorliegendem Material beantwortet werden könnten. Bei der Aufgabenkonstruktion wurden durch die Formulierung der Items Informationen gegeben, die dem einen Kompetenzbereich zugehören. Die zu lösenden Aufgaben zielen auf die Messung der jeweils anderen Kompetenz des Überlappungsbereichs ab.

Eine andere Konstruktionslogik geht von dem – in der Sechs-Felder-Matrix kompakt visualisierten – Zusammenhang zwischen den „Basisoperationen historischen Denkens“ und den Fokussierungen auf Vergangenes, auf Geschichte und auf Gegenwart/Zukunft aus.

Auf der Ebene der Vergangenheit wird auf noch unverbundene empirische Begebenheiten und Gegebenheiten fokussiert. Auf der Ebene der Geschichte liegt das Augenmerk auf den Kontextualisierungen und Verknüpfungen, durch die Deutun-

Abb. 1: Sechs-Felder-Matrix

	Fokussierung auf Vergangenheit Vergangenes feststellen	**Fokussierung auf Geschichte** Vergangenes in Kontexte setzen und als Geschichte darstellen	**Fokussierung auf Gegenwart/Zukunft** Geschichte auf Gegenwart und Zukunft beziehen
Umgang mit Vegangenheit	*Vergangenes aus Quellen re-konstruieren*	*Vergangenes auf spezifische Weise in einer Geschichte darstellen*	*Durch Bezüge auf Vergangenes/Geschichte der eigenen Gegenwart/Zukunft historische Tiefe geben*
Re-Konstruktion von Vergangenheit	Feststellen von: • Daten, Ereignissen, Handlungen, Personen („Fakten“) • Bedeutungen, die in der Vergangenheit den Fakten zugewiesen wurden • Motiven, Kausalbeziehungen, die zugewiesen wurden	Berücksichtigung finden z.B. • Funktionen, die der Narration zugewiesen werden (Erklärung, Legitimation, Identitätsbildung …) • fachspezifische und überfachlich relevante Theorien • Spezifika der Adressaten • Besonderh. d. gewählten Darstellungsmediums *synchrone Kontextualisierungen (Zustände)* | *diachrone Kontextualisierungen (Zeitverläufe)*	Konstruieren von: • Kontinuitätsvorstellungen (Ursachen-, Sinnzusammenhänge) • Vortellungen von Wandel • historischer Identität • Orientierung für zukünftiges Handeln
	Basisoperationen des Geschichtsbewusstseins		
De-Konstruktion von Vergangenheit	• Feststellen, was die Narrationen über Vergangenes aussagen • Abklären der fachlichen Triftigkeit des Berichteten • Klären der Repräsentativität des über das historische Phänomen Berichteten	*synchrone Kontextualisierungen (Zustände)* | *diachrone Kontextualisierungen (Zeitverläufe)* Offenlegen • der zugrunde liegenden Fragestellungen • der Argumentationsstruktur der Narration • der angewandten Theorien und Alltagshypothesen • der äußeren Zwänge • der Perspektivität • des Standorts	Erschließen von • Botschaften • Orientierungsangeboten • Sinnbildungsmustern • Orientierungsfragen, die hinter der Darstellung stehen • Normen und Werten, die vertreten oder abgelehnt werden • kulturellen Prägungen
Umgang mit Geschichte	*Erheben, was historische Narrationen über Vergangenes aussagen*	*Feststellen, in welche Kontextualisierungen die Vergangenheitspartikel in der jeweiligen Geschichte gestellt werden, auf welche Weise die Geschichte erzählt wird*	*Feststellen, welche Gegenwartsbezüge in der Geschichte hergestellt werden, welche Orientierungsangebote gegeben werden*

Quelle: Schreiber et al., 2006, 53

gen erwachsen. Diesen wiederum werden im Fokus auf Gegenwart/Zukunft Bedeutungen oder auch konkrete Auswirkungen für die jeweilige Lebenspraxis des Autors bzw. der Rezipienten zugemessen.

Nach der der Sechs-Felder-Matrix innewohnenden Logik können Aufgaben konstruiert werden, die analytische, d. h. dekonstruierend evaluierende, bzw. synthetische, d. h. rekonstruierend narrativierende, Fähigkeiten messen. Zudem können Aufgaben generiert werden, die überprüfen, inwiefern die Probanden mit den Fokussierungen umgehen können, die in den Narrativen fassbar sind.

4.3 Pilotierungen

Das Testinstrument wurde in mehreren Erhebungen sukzessive weiterentwickelt. In der ersten Pilotierung des HiTCH-Tests, bei der 1.701 Schülerinnen und Schüler aus neunten Klassen aller Schulformen[2] befragt wurden, wurden die Aufgaben und Items in sieben Themenheften zusammengestellt („Scherbengericht im antiken Athen", „Untergang Roms", „Pest", „Hexen", „US-Geschichte", „Nürnberger Prozesse", „DDR").[3] Jeder Proband bearbeitete zwei Themenhefte. Hinzu kamen Aufgaben zur Testung der Sachkompetenz (z. B. Staatsformen kennen und zuordnen können, Einsicht in epistemologische Prinzipien). Neben den Kompetenzaufgaben wurde der persönliche Hintergrund der Lernenden erfasst (z. B. sozioökonomischer Hintergrund, Schulnoten).

Ziel der Datenauswertung war die Itemauswahl und -optimierung aufgrund statistischer Verfahren. Daher wurden die Items deskriptiv anhand ihrer Mittelwerte und Streuung beschrieben, unter anderem auch, um die Schwierigkeit der Items einzuschätzen. In Reliabilitätsanalysen wurden die interne Konsistenz der Skalen wie auch die Itemtrennschärfe und die korrelativen Zusammenhänge zwischen den Aufgaben überprüft. Zudem wurde untersucht, welchen Beitrag die einzelnen Aufgaben bzw. Items zur Erfassung historischer Kompetenzen leisteten. Hierfür wurde mit den – auf der Skalenebene im Hinblick auf ihre interne Konsistenz reliablen – Aufgaben eine Faktorenanalyse mit nur einem Faktor gerechnet, der den Globalfaktor „historische Kompetenz" abbildete. Aus der Ladung auf diesen einen Faktor ergab sich eine Reihung der eingesetzten Aufgaben im Hinblick auf ihren jeweiligen Beitrag zur Messung des Globalfaktors „historische Kompetenz".

Als Kriterien für die Aufgabenauswahl für die nächste Pilotierung (2014) wurden die statistischen Kennwerte wie auch fachdidaktische Überlegungen genutzt. Um darüber hinaus die Aufgaben auf der Itemebene zu optimieren, wurde jedes Item im Hinblick auf die Ladung auf dem Globalfaktor geschätzt, und potenzielle DIF-Effekte wurden überprüft, z. B. ob ein Item im Gymnasium oder Nichtgymnasium, im Osten oder Westen der Bundesrepublik andere Messeigenschaften aufwies. Auf Grundlage der Ergebnisse wurden einzelne Items selektiert bzw. umformuliert.

2 59,8 Prozent Gymnasium, 40,2 Prozent andere Schulformen; Durchschnittsalter: 14,89 Jahre; 51,2 Prozent Mädchen; Baden-Württemberg: 25,2 Prozent, Bayern: 31,1 Prozent, Hamburg: 30,3 Prozent; Nordrhein-Westfalen: 6,9 Prozent und Sachsen: 6,5 Prozent.

3 Die sieben „Themenhefte" wurden von jeweils ca. 10,0 bis 13,6 Prozent der Schülerinnen und Schüler bearbeitet ($170 < N < 231$).

Für die zweite Pilotierung wurden Items zu verschiedenen Themenbereichen kombiniert. Es wurde ein Design gewählt, das auch eine Überprüfung der Dimensionalität der historischen Kompetenz ermöglichte. In einen großen gemeinsamen Anker wurden pro Kompetenzbereich mindestens zwei bereits in der ersten Pilotierung erprobte Aufgaben aufgenommen. Unter anderem wurden von den Assoziationspartnern neu entwickelte Aufgaben zusätzlich und alternierend eingesetzt. Damit erweiterte sich das Themenspektrum um Aufgaben zur japanischen Geschichte, zu Jugendkrawallen in den 80er-Jahren, Anne Frank, Thukydides und König Artus. Zudem wurden anhand von kleinen Materialauszügen Dekonstruktionsaufgaben an verschiedenen Themen und Materialarten durchgespielt (z. B. Comics, Karikaturen, Denkmäler). An der zweiten Pilotierung (2014) nahmen 1.265 Schülerinnen und Schüler aller Schularten (außer Schulen mit besonderem Förderbedarf) der neunten Jahrgangsstufe in Deutschland, Österreich und der Schweiz teil.[4] Alle Schülerinnen und Schüler bearbeiteten den gemeinsamen Anker. Die Teilhefte mit den neu zu erprobenden Aufgaben wurden gleichmäßig auf die Stichprobe verteilt.

Belege für die Separierbarkeit der theoretisch plausiblen vier Kompetenzbereiche aus dem FUER-Modell, die auch bei der Konstruktion der Items Pate standen, konnten in exploratorischen wie auch konfirmatorischen Faktorenanalysen auf Basis der vorliegenden Daten nicht nachgewiesen werden. Mögliche Gründe hierfür könnten sein, dass (1) die Kompetenzbereiche in theoretischer Hinsicht als eng verbunden beschrieben sind, (2) eine etwaig vorhandene mehrfaktorielle Struktur der historischen Kompetenz von einem Generalfaktor „historische Kompetenz" oder von itemspezifischen Effekten überlagert wird, (3) die Items in aller Regel mehrere Aspekte messen und (4) ein kompetenzförderlicher Unterricht möglicherweise häufig alle Kompetenzbereiche gleichermaßen adressiert. Dies schließt jedoch nicht aus, dass eine solche Dimensionalität auf der Basis von (noch zu konstruierenden) Items, die jeweils sehr spezifisch einen der Kompetenzbereiche adressieren, sehr wohl gefunden werden könnten.

Neben der Überprüfung der Dimensionalität hatte die zweite Pilotierung das Ziel, die psychometrisch geeignetsten Aufgaben für die Haupterhebung zu identifizieren und die Aufgaben zu optimieren. Hierfür kamen eindimensionale Rasch-Modelle wie auch konfirmatorische Nested-Factor-Ansätze zur Anwendung. Für die nachfolgende Haupterhebung wurden 15 Aufgaben mit insgesamt 106 Items ausgewählt. Die eindimensionale Rasch-Analyse in ConQuest (Wu, Adams, Wilsons & Haldane, 2007) zeigte eine hohe Score-Reliabilität (WLE-Person Separation Reliability: .88). 39 Prozent der Varianz ließen sich durch die Klassenzugehörigkeit erklären, die – wie üblich bei Schulleistungstests in mehrgliedrigen Schulsystemen (vgl. z. B. Baumert, Trautwein & Artelt, 2003) – auf Schulartunterschiede zurückgeführt werden kann. Aufgrund fachdidaktischer und psychometrischer Überlegungen wurden Aufgaben für die Haupterhebung ausgewählt und optimiert.

4 51,9 Prozent Gymnasium, 48,1 Prozent andere Schulformen; Durchschnittsalter: 14,80 Jahre; 51,0 Prozent Mädchen; Bayern: 23,2 Prozent, Hamburg und Schleswig-Holstein: 29,3 Prozent; Nordrhein-Westfalen: 21,5 Prozent, Schweiz: 19,3 Prozent, Österreich: 6,5 Prozent.

4.4 Haupterhebung

In der abschließenden Haupterhebung bearbeiteten die Schülerinnen und Schüler neben dem HiTCH-Instrument (15 Aufgaben mit 106 Items) eines von fünf zusätzlichen Testheften: (1) die figurale und verbale Skala eines Tests für kognitive Grundfähigkeiten (KFT) (Heller & Perleth, 2000); (2) die Lesetestbatterie für die Klassenstufe 8–9 (Bäuerlein, Lenhard & Schneider, 2012) mit dem Lesegeschwindigkeitstest und einem Leseverständnistest, der sich entweder auf einen literarischen Text oder auf einen Sachtext bezog; (3) ein Heft (Testheft 3, 4, 5 bzw. 6) mit neuen bzw. überarbeiteten Aufgaben zur Erfassung historischer Kompetenzen. Ziel der Studie war es zu überprüfen, (1) ob das Kurzinstrument eine hinreichende Homogenität für eine eindimensionale Skalierung aufweist, (2) ob die Messeigenschaften über verschiedene Subgruppen hinweg variieren (DIF-Effekte) und (3) ob sich ein plausibles Muster von Korrelationen mit Außenkriterien finden lassen würde (Validität). Die Validität des Instruments wurde hinsichtlich der Abgrenzung von benachbarten Kompetenzbereichen (Lesekompetenz) und weiteren Kriterien (z. B. Intelligenz, Schulform, Schulnoten) untersucht. Zudem sollte (4) der Aufgabenpool für künftige Analysen nochmals erweitert werden; die neuen Aufgaben wurden bei den hier berichteten Analysen allerdings nicht berücksichtigt.

An der Haupterhebung nahmen 2.853 Schülerinnen und Schüler (50,6 Prozent Mädchen) der neunten Jahrgangsstufen (Durchschnittsalter: 14,41 Jahre) in allen Schulformen (außer Schulen mit besonderem Förderungsbedarf) teil (Gymnasium: 53,7 Prozent, Gesamt- bzw. Gemeinschaftsschulen: 22,4 Prozent, Realschulen: 15,9 Prozent, Hauptschulen: 7,9 Prozent). Die Erhebung fand in mehreren Ländern der Bundesrepublik Deutschland und in der Schweiz und Österreich statt (Baden-Württemberg: 22,9 Prozent, Bayern: 4,1 Prozent, Hamburg/Schleswig-Holstein: 25,3 Prozent, Nordrhein-Westfalen: 17,0 Prozent, Thüringen: 0,6 Prozent, Schweiz: 9,5 Prozent und Österreich: 20,6 Prozent). Neben dem gemeinsamen Anker (N = 2.853) bearbeiteten die folgenden Anteile der Probandinnen und Probanden einen Test zur Erfassung der kognitiven Grundfähigkeiten (N = 451), einen Lesekompetenztest (N = 616) und zusätzliche historische Kompetenzaufgaben (Zusatz-Kompetenztest 1: 453; Zusatz-Kompetenzheft 2: N = 432, Zusatzheft zum Geschichtsbewusstsein: N = 309, Zusatzheft zu Dekonstruktionsaufgaben: N = 588). Vier Schülerinnen und Schüler haben kein Zusatzheft bearbeitet.

4.5 Analysemethoden in der Haupterhebung[5]

Im ersten Schritt wurden sämtliche 106 Items des HiTCH-Instruments in einem eindimensionalen 2PL-Modell (mit einem mehrstufigen Item mit einer zusätzlichen Ausprägung für „teilrichtig“) untersucht. Ziel dabei war es, Items mit niedrigen Diskriminationsparametern von den weiteren Analysen auszuschließen, sofern diese

5 An dieser Stelle muss die Darstellung des methodischen Vorgehens wegen enger Seitenvorgaben sehr kompakt erfolgen. Für eine ausführlichere Version siehe die Grundlagenpublikation zu HiTCH (Trautwein et al., in Vorbereitung).

nicht aus theoretischen Gründen als besonders relevant erachtet wurden. Damit sollte einerseits die Effizienz des Tests erhöht werden (hohe Reliabilität bei möglichst geringer Itemanzahl), andererseits sollte die Variabilität der Diskriminationsparameter eingeschränkt werden, um eine möglichst gute Approximation an ein 1PL-Modell zu erhalten. Eine Itemselektion auf Basis der üblichen Infit-Statistiken eines Rasch-Modells hätte zu vergleichbaren Ergebnissen geführt, da diese erwartungsgemäß einen engen Zusammenhang mit der standardisierten Ladung aus dem 2PL-Modell aufwies ($r = -.98$).

Im nächsten Schritt wurden die latenten Interkorrelationen auf der Ebene der einzelnen Aufgaben basierend auf einem 15-dimensionalen Rasch-Modell (ConQuest) geschätzt, um mögliche Verletzungen der Annahme der lokalen stochastischen Unabhängigkeit zu prüfen. Zur Prüfung, in welchem Ausmaß neben einem Generalfaktor „historische Kompetenz" zusätzliche spezifische Faktoren für die Beantwortung der Items relevant sind, wurde ein Testlet-Modell mit zusätzlichen latenten Variablen für die theoretischen Subdimensionen in ConQuest berechnet.[6] Ein weiteres Modell mit sowohl aufgaben- als auch subdimensionsspezifischen latenten Variablen wurde in Mplus (Muthén & Muthén, 1998–2012) berechnet.[7]

Um Aussagen darüber treffen zu können, ob Items in bestimmten Subgruppen unterschiedlich funktionieren (DIF), wurden die Variablen Geschlecht, Schulform und häuslicher Buchbesitz (als Hinweis auf das soziokulturelle Kapitel) dummycodiert und in DIF-Analysen basierend auf eindimensionalen Modellen mit jeweils einer der genannten Gruppierungsvariablen in ConQuest untersucht.[8]

Die Zusammenhänge des HiTCH-Instruments mit den beiden Lesekompetenztests (Sachtext, literarischer Text), den beiden KFT-Subtests (verbal, figural), der Lesegeschwindigkeit und des Interesses an Geschichte wurden anhand eines mehrdimensionalen Modells in Mplus geschätzt. Das HiTCH-Instrument, die Lesekompetenztests sowie die KFT-Subskalen wurden dabei als latente Variablen (eindimensionale einparametrische Messmodelle) spezifiziert, während die Lesegeschwindigkeit sowie das Interesse an Geschichte als manifeste Variablen in die Analyse eingingen. Da eine komplette Interkorrelationsmatrix aufgrund des Multi-Matrix-Designs nicht schätzbar war, wurde das HiTCH-Instrument, das eine Überlappung mit sämtlichen Kriterien aufwies, als einzelner Prädiktor auf die Kriterien als abhängige Variable spezifiziert. Die standardisierten Regressionsgewichte sind somit im Sinne von Korrelationen zu interpretieren. Die Residuen wurden anhand von unkorrelierten latenten Variablen modelliert, wobei bei Überlappungen (z. B. KFT-V und KFT-F) aufgrund der Datenkonstellation schätzbare Zusammenhänge anhand von Regressionen zwischen diesen Residuen berücksichtigt wurden.

6 Sämtliche latente Variablen wurden in diesem Modell als unkorreliert spezifiziert (Wang & Wilson, 2005), was einem Nested-Factor-Modell entspricht (Gustafsson & Balke, 1993).

7 Dieses zweite Modell konnte nicht in ConQuest berechnet werden, da die Anzahl der Dimensionen zu groß war (ein Generalfaktor, 14 aufgaben- und 5 subdimensionsspezifische Faktoren, d. h. insgesamt 20 Dimensionen).

8 Dummy-Codierungen: Geschlecht: 0 = männlich, 1 = weiblich, Schulform: 1 = Gymnasium, 0 = andere Schulform, häuslicher Buchbesitz: dummycodiert, Median-Split.

5 Ergebnisse

In der folgenden Ergebnisübersicht werden zentrale Befunde des HiTCH-Projekts zusammengefasst. Neben den Kennwerten aus den psychometrischen Analysen, die unverzichtbar für eine Rezeption und kritische Bewertung durch die Fachkollegen sind und die in kompakter Form dargestellt werden, erfolgt am Ende jedes Abschnitts jeweils eine kurze inhaltliche Bewertung. Für eine ausführlichere Darstellung und Interpretation wird auf die Grundlagenpublikation zu HiTCH (Trautwein et al., in Vorbereitung) verwiesen.

5.1 Itemauswahl und Reliabilität des HiTCH-Instruments

Die Itemauswahl wurde auf der Basis von inhaltlichen und psychometrischen Überlegungen vorgenommen. Mit einer Ausnahme lagen im eindimensionalen 2PL-Modell alle standardisierten Ladungen im positiven Bereich. Als statistisches Auswahlkriterium wurde als untere Grenze des 95-Prozent-Konfidenzintervalls der standardisierten Ladung ein Wert von .30 festgelegt. Dies hätte zum Ausschluss von insgesamt 34 Items geführt. 19 dieser Items wurden allerdings aus inhaltlichen Gründen beibehalten. Vor der Itemauswahl lag der WMNSQ (weighted mean-square) im Bereich von 0.84 bis 1.37 bei einer Varianz der latenten Variable von 0.65. Bezogen auf die finale Itemauswahl der 91 Items aus dem HiTCH-Instrument lag der WMNSQ im Bereich von 0.84 bis 1.17, also relativ nahe am angestrebten Wert von 1.0. Die Varianz der latenten Variable lag bei 0.88. Wie Abbildung 2 zeigt, ergab sich bezogen auf die Itemschwierigkeiten eine gute Abdeckung des Leistungsspektrums der Schülerinnen und Schüler in der Stichprobe. Die WLE-person-separation-Reliabilität (WLE PSR) betrug *Rel = .91*.

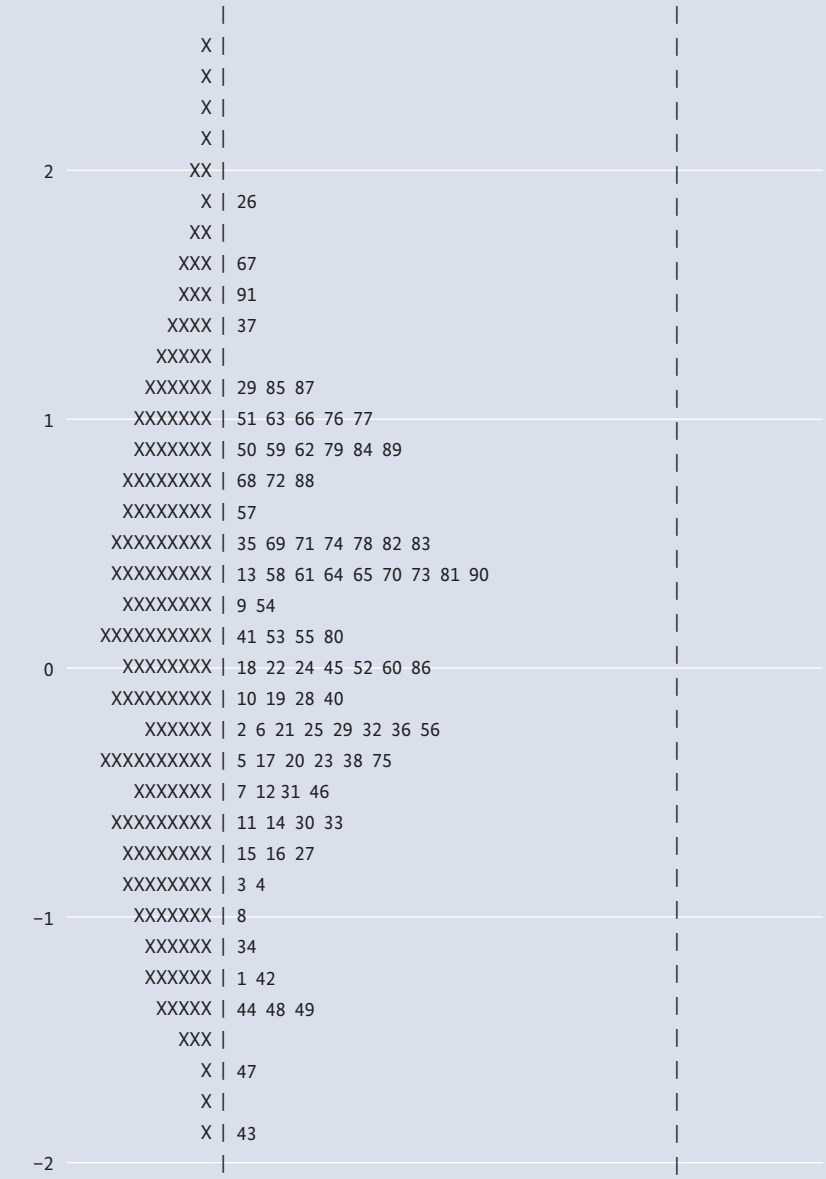
Insgesamt liegt mit dem HiTCH-Test damit – gemessen an den üblichen Standards in großen Schulleistungsstudien – ein reliables Instrument zur Erfassung historischer Kompetenzen vor, dessen Aufgaben in wünschenswerter Weise in ihrer Schwierigkeit streuen.

5.2 Überprüfung der Ein- bzw. Mehrdimensionalität

Die Interkorrelationen auf Aufgabenebene lagen deutlich unterhalb von 1.0, welche beim Vorliegen einer perfekten Eindimensionalität zu erwarten wäre. Als mögliche Ursachen für diese Heterogenität könnten die Subdimensionen des FUER-Modells (Methoden-, Sach-, Frage- und Orientierungskompetenz), der Einfluss spezifischer Themen oder die unterschiedlichen Aufgabenformate infrage kommen. Das Korrelationsmuster ergab nur sehr geringe Hinweise darauf, dass sich die theoretisch unterscheidbaren Subdimensionen in den Daten als eigene Faktoren identifizieren lassen würden. Eine exploratorische Faktorenanalyse auf der Basis der Interkorrelationsmatrix ergab einen Eigenwert für die erste Komponente von 9.6, wohingegen bereits die zweite Komponente mit einem Eigenwert von 0.9 den Wert 1.0 (Kaiser-Kriterium) unterschritt. Dieser Befund spricht dafür, dass auf Basis dieser Daten eine

Abb 2: Verteilungen der Kompetenzscores (WLE-Schätzer, mit „X“ gekennzeichnet) und der Itemschwierigkeiten (mit Itemnummer 1–91 gekennzeichnet) des HiTCH-Instrumentariums (91 Items) aus einem eindimensionalen 1PL-Modell

Der untere Bereich der Abbildung repräsentiert leistungsschwache Schülerinnen und Schüler bzw. Items mit niedriger Schwierigkeit, der obere Bereich leistungsstarke Schülerinnen und Schüler bzw. Items mit hoher Schwierigkeit.



multidimensionale Modellierung nicht sinnvoll erscheint. Der erste Faktor (der Generalfaktor „historische Kompetenz“) erklärte 64 Prozent der Gesamtvarianz. Einiges an Varianz verblieb auf der Einzelaufgabenebene.

Die Berechnung des Testlet-Modells zur Überprüfung, ob die theoretischen Subdimensionen oder die Aufgabeneffekte diese verbleibende Varianz erklären, führte zu dem Ergebnis, dass der Generalfaktor erwartungsgemäß die höchste Varianz aufwies (1.29), während die subdimensionsspezifischen Varianzen der latenten Variablen im Bereich von 0.21 bis 0.90 lagen.[9] In einem weiteren Modell wurden sowohl aufgaben- als auch subdimensionsspezifische latente Variablen berechnet. Hierbei zeigte sich, dass die subdimensionsspezifischen Komponenten, die im Bereich von 0 bis 5 Prozent (M = 2 Prozent) lagen, weniger bedeutsam waren als die aufgabenspezifischen Varianzkomponenten, die im Bereich von einem bis 42 Prozent lagen (M = 15 Prozent). Die deutlich größten Varianzanteile ließen sich erwartungsgemäß auf den Generalfaktor „historische Kompetenz" zurückführen, wobei die relativen Anteile zwischen 30 und 57 Prozent der Gesamtvarianz betrugen (M = 38 Prozent). Das HiTCH-Instrumentarium lässt sich dementsprechend im Wesentlichen als eindimensionale Struktur abbilden. Aufgrund des Vorgehens bei der Konstruktion des Tests, bei der die theoretisch als bedeutsam eingeschätzten Bestandteile von historischer Kompetenz berücksichtigt wurden, kann der Gesamttest deshalb beanspruchen, „historische Kompetenz" in wünschenswerter Weise in einem Gesamtwert zu erfassen; gleichzeitig finden sich Hinweise darauf, dass die einzelnen Aufgaben einen nicht komplett vernachlässigbaren Teil der Varianz binden, sodass vertiefende (fachdidaktische) Analysen zu den Eigenschaften der Aufgaben angezeigt sind.

Vertiefte Prüfungen der Dimensionalität erbrachten also Hinweise darauf, dass sich die im FUER-Modell theoretisch definierten vier Kompetenzbereiche – zumindest mit den bislang entwickelten Items – empirisch nicht trennen lassen. Dies ist bei Kompetenztests nicht ungewöhnlich: Ein ähnlicher Befund zeigte sich auch bei dem Test zur Erfassung der Bildungsstandards im Fach Mathematik, in dem die theoretisch trennbaren Kompetenzbereiche in den empirischen Daten eng miteinander korrelierten (Blum, Drüke-Noe, Hartung & Köller, 2006). Hinsichtlich des HiTCH-Instruments könnte es sein, dass die Aufgaben die verschiedenen Kompetenzbereiche noch nicht differenziert genug adressieren. Eine andere Erklärung könnte darin zu finden sein, dass die Subdimensionen der Kompetenz historischen Denkens sehr eng miteinander zusammenhängen. Es ist beispielsweise zu vermuten, dass Lernende, die eine differenzierte Einsicht in die epistemologischen Prinzipien von Geschichte haben und über eine hohe Strukturierungs- und Begriffskompetenz (Sachkompetenz) verfügen, auch die prozessualen Operationen besser ausführen können. Darüber hinaus kann spekuliert werden, dass ein kompetent durchgeführter, kompetenzorientierter Geschichtsunterricht alle Kompetenzbereiche adressiert und in ähnlich guter Weise fördert, sodass ein empirischer Nachweis einer etwaig tatsächlich vorhandenen Dimensionalität erschwert werden würde. Hier sind weitere Untersuchungen, gegebenenfalls auch unter Nutzung von weiteren Aufgaben/Items sowie homogeneren Stichproben, notwendig, um ein noch vertiefteres Verständnis der Dimensionalität zu erhalten.

9 Verglichen mit dem eindimensionalen Modell zeigte sich eine bessere Modellpassung für das komplexere Modell (Deviance-Differenz-Test: $\chi 2(5) = 2881.2$, $p < .001$).

5.3 DIF-Analysen

Hinsichtlich der Vergleichbarkeit der Itemschwierigkeiten in verschiedenen Subgruppen fanden wir – bezogen auf das Geschlecht – insgesamt geringfügige Unterschiede zwischen Mädchen und Jungen ($-0.57 \leq b_{\text{Jungen}} - b_{\text{Mädchen}} \leq 0.42$), wobei sich für 80 der insgesamt 91 Items absolute Differenzen kleiner als 0.30 ergaben. Etwas größere Unterschiede ergaben sich für die Schulform: Hier lagen die Differenzen der Itemschwierigkeiten im Bereich von $-0.66 \leq b_{\text{nichtgymnasiale Schulform}} - b_{\text{Gymnasium}} \leq 0.73$, wobei sich für 70 Items absolute Differenzen kleiner als 0.40 ergaben. Bezüglich des häuslichen Buchbesitzes ergaben sich insgesamt eher niedrige relative Schwierigkeitsunterschiede ($-0.43 \leq b_{\text{geringer Buchbestand}} - b_{\text{hoher Buchbestand}} \leq 0.54$).

Insgesamt weisen die DIF-Analysen (Analysen zur Prüfung, ob die Items bei den unterschiedlichen Gruppen von Untersuchungsteilnehmern ähnlich funktionieren) also darauf hin, dass der HiTCH-Test in den untersuchten Subgruppen relativ ähnliche Messeigenschaften aufwies und damit eine faire Erfassung historischer Kompetenzen ermöglicht.

5.4 Validität

Durch die enge Zusammenarbeit von Fachdidaktik und empirischer Bildungsforschung bei der Entwicklung und Überprüfung der Items konnte eine große inhaltliche Breite (Abdeckung der vier Bereiche des FUER-Modells) der im finalen Instrument berücksichtigten Items bei insgesamt akzeptablen bis guten Messeigenschaften des HiTCH-Tests erreicht werden. Wie erwartet worden war, fanden sich in den Kompetenzwerten substanzielle Mittelwertsunterschiede zugunsten der Gymnasiasten. Bei der Überprüfung der konvergenten und diskriminanten Validität durch Korrelationsanalysen zeigten sich auf latenter Ebene relativ deutliche Zusammenhänge des HiTCH-Instruments mit den beiden Lesekompetenztests und dem KFT-verbal-Test ($.83 \leq r \leq .85$). Etwas niedriger war der Zusammenhang des HiTCH-Instruments mit der figuralen Analogie-Facette des KFT ($r = .68$). Relativ niedrige Zusammenhänge fanden sich zwischen dem HiTCH-Instrument und der Leistung im Lesegeschwindigkeitstest ($r = .40$). Auch auf der Ebene der Einzelaufgaben wurden die Zusammenhänge mit den Außenkriterien (Lesekompetenz, kognitive Fähigkeiten) ermittelt, wie auch mit den jeweils verbleibenden Aufgaben des HiTCH-Instruments. Mit wenigen Ausnahmen zeigten sich erwartungsgemäß höhere Zusammenhänge der Einzelaufgaben mit den jeweils verbleibenden HiTCH-Aufgaben als mit den Außenkriterien.

Zusammenfassend lässt sich somit konstatieren, dass die vorliegenden Validitätsbelege darauf hinweisen, dass der HiTCH-Test tatsächlich das misst, was er messen soll (nämlich historische Kompetenz). Bei der Überprüfung der diskriminanten Validität zeigte sich, dass sich die im HiTCH-Test erfasste Kompetenz historischen Denkens von der Lesekompetenz wie auch von allgemeinen kognitiven Fähigkeiten trennen ließ. Jedoch erschien der Zusammenhang mit den Tests, die verbale Fähigkeiten adressieren, mit einer Korrelation von bis zu .85 relativ hoch. Um die Bedeutung des

hohen Zusammenhangs einschätzen zu können, lohnt sich erneut ein Vergleich mit anderen Schulleistungsstudien. So fanden sich in PISA für die mathematische und sprachliche literacy ähnlich hohe Korrelationen von rund .85, die nicht als Hinweis darauf gedeutet werden sollten, dass die PISA-Tests für die unterschiedlichen Domänen in Wirklichkeit nur eine dahinterliegende Fähigkeit messen würden (Baumert, Brunner, Lüdtke & Trautwein, 2007). Zudem ist ein relativ hoher Zusammenhang zwischen allgemeiner Lesekompetenz und historischer Kompetenz nicht erstaunlich, wenn man die auch theoretisch gegebenen Überschneidungen zwischen allgemeiner Lesekompetenz und historischer Lesekompetenz betrachtet, die sich unter anderem im Bereich der Methodenkompetenzen zeigen. Wer eine hohe Lesekompetenz aufweist, sollte auch Vorteile beim historischen Lesen haben, aber man sollte im Übrigen nicht von einer Einbahnstraße ausgehen: Es ist zu erwarten, dass ein guter, kompetenzorientierter Geschichtsunterricht auch substanzielle Rückwirkungen auf die allgemeine Lesekompetenz hat. Ohne Zweifel: In der Untersuchung des Zusammenhangs und Zusammenspiels von allgemeiner Lesekompetenz und historischer Lesekompetenz oder, anders gesagt, in der Frage nach dem spezifisch Historischen in historischen Leseaufgaben steckt viel Potenzial für zukünftige Forschung.

6 Zusammenfassung und Ausblick

Historische Kompetenzen sind von enormer individueller und gesellschaftlicher Bedeutung. Ihr Potenzial bei der Orientierung in einer modernen Welt ist so hoch wie das der MINT-Fächer, wenn auch anders gelagert: Historische Kompetenzen sind die Grundlage dafür, zentrale Herausforderungen der Moderne zu meistern. Historische Kompetenzen ermöglichen es, die historischen Dimensionen in gegenwärtigen Entwicklungen (z. B. Flüchtlingsproblematik) zu erkennen und zur Fundierung von Entscheidungen für die Zukunft zu nutzen. Zudem erschließen historische Kompetenzen einen Zugang zur Welt, wie er prototypisch für Kulturwissenschaften ist und einen Bestandteil einer humanistischen Bildung darstellt. Geschichte ist nicht nur ein Fach, sondern ein methodischer Zugang zur Welt. Historisch gedacht und gearbeitet wird ja nicht nur im Fach Geschichte, sondern unter anderem auch in den Sprachen und Literaturen, in Politik, Soziologie und Wirtschaft, Geografie, Recht und Kunst. Aber systematisch erklärt und eingeübt wird diese Erkenntnisweise eben vor allem in Geschichte. Die tatsächliche Bedeutung des historischen Denkens spiegelt sich also nicht in der Rolle des Faches Geschichte im Fächerkanon der Schule wider.

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

Das zentrale Ziel des HiTCH-Projekts war die Entwicklung eines Instruments, mit dem historisches Denken in psychometrisch „sauber" konstruierten standardisierten Instrumenten erfasst werden kann. Hierfür wurde ein großer Itempool entwickelt, der alle Kompetenzbereiche des FUER-Modells adressierte und curriculumnahe wie auch -ferne Themen umfasste. Auf der Basis der Ergebnisse von drei großen Studien, an denen insgesamt fast 6.000 Schülerinnen und Schüler teilnahmen, wurde der HiTCH-Test entwickelt. Das nach einer Entwicklungszeit von drei Jahren nun vorliegende HiTCH-Instrument mit 91 Items stellt ein standardisiertes, reliables und valides Instrument zur Erfassung historischer Kompetenz dar, das – wie es das erklärte Ziel des HiTCH-Projekts war – in Schulleistungsstudien eingesetzt werden kann.

Der vorliegende HiTCH-Test beruht auf einem narrativistisch-konstruktivistischen Verständnis des domänenspezifischen Prozesses historischen Denkens und erfasst generische Aspekte historischer Kompetenzen. Damit ist dieses Instrument weitgehend kompatibel mit anderen in der Geschichtsdidaktik diskutierten Kompetenzmodellen. Da sich die Testaufgaben nicht auf vorher im Unterricht behandelte Gegenstände und Unterrichtsthemen beziehen und die Aufgaben für alle Schularten einen angemessenen Schwierigkeitsgrad abbilden, kann in Zukunft in schulform- und (bundes-)länderübergreifenden Schulleistungsstudien auch das Fach Geschichte adressiert werden. Ausgehend vom HiTCH-Instrument könnten zudem zukünftig individualdiagnostische Aufgaben (z. B. für die Lehrkräfte), aber auch Aufgaben zur Kompetenzförderung entwickelt werden.

Der jetzt entwickelte HiTCH-Test ist ein fertiges Instrument, mit dessen Hilfe die Resultate kumulativen Kompetenzerwerbs im Bereich der Geschichte erfasst werden können. Gleichzeitig betrachten die Projektbeteiligten das Projekt nicht als abgeschlossen. Im Gegenteil: Aufgaben sollen neu- und weiterentwickelt werden, eine Reihe weiterer Validitätsstudien wurden bereits durchgeführt bzw. werden demnächst abgeschlossen. Es bestehen Überlegungen zur Identifikation von Kompetenzstufen, und es sind Studien geplant, in denen die Effekte bestimmter Unterrichtssettings bzw. der Unterrichtsqualität auf die Ausprägung historischer Kompetenzen untersucht werden sollen. Zur Fortsetzung der Forschungsarbeit wurde nach Auslaufen der Förderung des Projekts durch das BMBF das HiTCH-Konsortium gegründet (vgl. die HiTCH-Website: www.hitch-projekt.de).

Literaturverzeichnis

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

Baron, C. (2012). Understanding historical thinking at historic sites. *Journal of Educational Psychology, 104 (3),* 833–847. doi: 10.1037/a0027476.

Barricelli, M. (2012). Narrativität. In M. Barricelli & M. Lücke (Hrsg.), *Handbuch Praxis des Geschichtsunterrichts* (S. 255–280). Schwalbach/Ts.: Wochenschau.

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

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

Baumert, J., Brunner, M., Lüdtke, O. & Trautwein, U. (2007). Was messen internationale Schulleistungsstudien? – Resultate kumulativer Wissenserwerbsprozesse. Eine Antwort auf Heiner Rindermann. *Psychologische Rundschau, 58,* 118–127.

Baumert, J., Trautwein, U. & Artelt, C. (2003). Schulumwelten. In Deutsches PISA-Konsortium (Hrsg.): PISA 2000. *Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–331). Opladen: Leske + Budrich.

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

Bergmann, K. (2007). Multiperspektivität. In U. Mayer, H. J. Pandel & G. Schneider (Hrsg.), *Handbuch Methoden im Geschichtsunterricht* (S. 65–77). Schwalbach/Ts.: Wochenschau.

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

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

Blum, W., Drücke-Noe, C., Hartung, R. & Köller, O. (Hrsg.). (2006). *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen.* Berlin: Cornelsen Scriptor.

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

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

Borries, B. von (2004). Perspektivenwechsel und Sinnbildungsfiguren im Umgang mit der Geschichte. In B. von Borries (Hrsg.), *Lebendiges Geschichtslernen. Bausteine zu Theorie und Pragmatik, Empirie und Normfrage* (S. 236–287). Schwalbach/Ts.: Wochenschau (Erstabdruck in Deutsch-Technisches Forum der Frauen/Frauennetzwerk für Frieden e.V. [Hrsg.]. [2000]. *Deutsche und tschechische Frauen im zivilgesellschaftlichen Dialog über die Gestaltung der zukünftigen Beziehungen. Perspektivenwechsel im Umgang mit Vergangenheit – Gegenwart – Zukunft.* S. 8–27. Bonn).

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

Clark, P. (2014). History education research in Canada. A late bloomer. In M. Köster, H. Thünemann & M. Zülsdorf-Kersting (Hrsg.), *Researching History Education. International Perspectives and Disciplinary Traditions* (S. 81–103). Schwalbach/Ts.: Wochenschau.

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

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

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

Erdmann, E. & Hasberg, W. (2011). *Facing mapping bridging diversity. Foundation of a European discourse on history education.* Schwalbach, Ts.: Wochenschau.

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

Gustafsson, J. E. & Balke, G. (1993). General and Specific Abilities as Predictors of School Achievement. *Multivariate Behavioural Research, 28 (4),* 407–434.

Hartmann, U. (2008). *Perspektivenübernahme als eine Kompetenz historischen Verstehens* (Dissertation). Abgerufen am 23.01.2015 von https://ediss.uni-goettingen.de/handle/11858/00-1735-0000-0006-AD13-2.

Hasberg, W. (2010). Historiker oder Pädagoge? Geschichtslehrer im Kreuzfeuer der Kompetenzdebatte. *Zeitschrift für Geschichtsdidaktik (Jahresband 2010),* 159–179.

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

Heil, W. (2010). *Kompetenzorientierter Geschichtsunterricht. Geschichte im Unterricht, Bd. 1.* Stuttgart: Kohlhammer.

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

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

Hodel, J. & Ziegler, B. (Hrsg.). (2009). Forschungswerkstatt *Geschichtsdidaktik 07. Beiträge zur Tagung „geschichtsdidaktik empirisch 07".* Bern: hep.

Hodel, J. & Ziegler, B. (Hrsg.). (2011). Forschungswerkstatt Geschichtsdidaktik 09. *Beiträge zur Tagung „geschichtsdidaktik empirisch 09"*. Bern: hep.

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

Jeismann, K.-E. (1988). Geschichtsbewußtsein als zentrale Kategorie der Geschichtsdidaktik. In G. Schneider (Hrsg.), *Geschichtsbewußtsein und historisch-politisches Lernen* (S. 1–24). Pfaffenweiler: Centaurus.

Jensen, B. E. (2003). *Historie – livsverden og fag* (1. Ausgabe, 1. Auflage). Kopenhagen: Gyldendal.

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Auflage, S. 27–74). Berlin: Springer.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Abgerufen am 25.06.2015 von http://www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf.

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

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

Koselleck, R., Mommsen, W. & Rüsen, J. (Hrsg.). (1977). *Objektivität und Parteilichkeit in der Geschichtswissenschaft (Theorie der Geschichte. Beiträge zur Historik; Bd. 1).* München: dtv.

Köster, M., Thünemann, H. & Zülsdorf-Kersting, M. (Hrsg.). (2014). *Researching History Education. International Perspectives and Disciplinary Traditions.* Schwalbach/Ts.: Wochenschau.

Kraus, A. (2013). Kategoriale Inhalts- und Strukturanalyse zur Auswertung von Schüleräußerungen zu Zeitzeugen – Wirksamkeitsforschung für kompetenzorientierten Geschichtsunterricht an Hauptschulen. In W. Schreiber, A. Schöner & F. Sochatzy (Hrsg.), *Schulbuchanalysen – Grundlage empirischer Geschichtsdidaktik* (S. 194–210). *Stuttgart: Kohlhammer.*

Langer-Plän, M. & Beilner, H. (2006). Zum Problem historischer Begriffsbildung. In H. Günther-Arndt & M. Sauer (Hrsg.). *Geschichtsdidaktik empirisch. Untersuchungen zum historischen Denken und Lernen (Zeitgeschichte – Zeitverständnis. Bd. 14)* (S. 215–250). Berlin: LIT.

Lazer, S. (2015). A large scale assessment of historical knowledge and reasoning: NAEP U.S. history assessment. In K. Ercikan & P. Seixas, *New directions in assessing historical thinking* (S. 145–158). New York, NY: Routledge.

Lücke, M. (2012). Multiperspektivität, Kontroversität, Pluralität. In M. Barricelli & M. Lücke (Hrsg.), *Handbuch Praxis des Geschichtsunterrichts* (S. 281–288). Schwalbach/Ts.: Wochenschau.

Mandell, N. & Malone, B. (2008). *Thinking like a historian: Rethinking history instruction.* Chicago, IL: Wisconsin Historical Society Press.

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

Muthén, L. K. & Muthén, B. O. (1998–2012). *Mplus user's guide* (7. Auflage). Los Angeles, CA: Muthén & Muthén.

Pandel, H.-J. (1987). Dimensionen des Geschichtsbewusstseins. Ein Versuch, seine Struktur für Empirie und Pragmatik diskutierbar zu machen. *Geschichtsdidaktik. Probleme, Projekte, Perspektiven, 12 (2),* 130–142.

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

Pandel, H.-J. (2013). *Geschichtsdidaktik. Eine Theorie für die Praxis.* Schwalbach/Ts.: Wochenschau.

Ricœur, P. (1988). *Zeit und Erzählung, Bd. 1: Zeit und historische Erzählung.* München: Fink. (Erstausgabe 1983. Temps et récit. Paris: Éd. du Seuil).

Rouet, J. F., Britt, M. A., Mason, R. A. & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88 (3),* 478–493. doi: 10.1037/0022-0663.88.3.478.

Rüsen, J. (1983). *Historische Vernunft. Die Grundlagen der Geschichtswissenschaft. Grundzüge einer Historik I.* Göttingen: Vandenhoeck & Ruprecht.

Rüsen, J. (1994). *Historische Orientierung. Über die Art des Geschichtsbewusstseins, sich in der Zeit zurechtzufinden.* Köln: Böhlau.

Rüsen, J. (2005). *History. Narration – interpretation – orientation.* New York, NY: Berghahn Books.

Rüsen, J. (2013). *Historik. Theorie der Geschichtswissenschaft.* Köln: Böhlau.

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

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

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

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

Schreiber, W., Körber, A., Borries, B. von, Krammer, R., Leutner-Ramme, S., Mebus, S., Schöner, A., Ziegler, B. (2006). *Historisches Denken. Ein Kompetenz-Strukturmodell.* Neuried: ars una Verlag. Abgerufen am 10.11.2014 von http://blogs.epb.uni-hamburg.de/historischeslernen/files/2009/09/Sonderdruck_Kompetenzen_klein_1.pdf.

Seixas, P. (2008). *"Scaling Up" the benchmarks of historical thinking.* A report on the Vancouver meetings. Abgerufen von http://historicalthinking.ca/sites/default/files/Scaling%20Up%20Meeting%20Report.pdf.

Seixas, P. (2011). *Theorizing historical consciousness* (2011st ed.). University of Toronto Press.

Seixas, P. & Morton, T. (2013). *The big six historical thinking concepts.* Abgerufen von http://www.nelson.com/thebigsix/documents/The%20Big%20Six%20Sample%20Chapter%20with%20BLM_Aug%2030.pdf.

Stearns, P. N. (1998). *Why study history?* American Historical Association. Abgerufen von http://www.historians.org/about-aha-and-membership/aha-history-and-archives/archives/why-study-history-%281998%29,

Stearns, P. N., Seixas, P. & Wineburg, S. (Hrsg.). (2000). *Knowing, teaching, and learning history. National and international perspectives.* New York University Press.

Taylor, T. & Young, C. (2003). *Historical literacy. Making history: a guide for the teaching and learning of history in Australian schools.* Abgerufen von http://www.hyperhistory.org/images/assets/pdf/complete.pdf.

Trautwein, U., Bertram, C., Borries, B. von, Brauch, N., Hirsch, M., Klausmeier, K., ..., Zuckowski, A. (in Vorbereitung). *Kompetenzen historischen Denkens erfassen – Konzeption, Operationalisierung und Befunde des Projekts „Historical Thinking – Competencies in History" (HiTCH).* Stuttgart: Kohlhammer-Verlag.

Van Drie, J. & van Boxtel, C. (2008). Historical reasoning: Towards a framework for analyzing students' reasoning about the past. *Educational Psychology Review* 20, 87–110.

VanSledright, B. A. (2014). *Assessing historical thinking & understanding. Innovative designs for new standards.* New York, NY: Routledge.

Ventzke, M., Mebus, S. & Schreiber, W. (Hrsg.). (2010). *Geschichte denken statt pauken in der Sekundarstufe II. 20 Jahre nach der friedlichen Revolution: Deutsche und europäische Perspektiven im gymnasialen Geschichtsunterricht.* Radebeul: Sächsisches Bildungsinstitut. Abgerufen am 22.06.2015 von http://www.pedocs.de/volltexte/2012/6540.

Verband der Geschichtslehrer Deutschland (Hrsg.). (2006/2010). *Bildungsstandards Geschichte. Sekundarstufe I. Rahmenmodell Gymnasium. 5.–10. Jahrgangsstufe.* Abgerufen am 17.12.2014 von http://www.geschichtslehrerverband.org/fileadmin/images/pdf/bildungsstandards.pdf

Waldis, M. & Ziegler, B. (Hrsg.). (2015). *Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13".* Bern: hep.

Wang, W.-C. & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29 (2),* 126–149. doi: 10.1177/0146621604271053.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Werner, M. & Schreiber, W. (2015). Testfragen befragen. Pretesting und Optimierung des Large-Scale-Kompetenztests „HiTCH" durch Cognitive Labs. In M. Waldis & B. Ziegler (Hrsg.), *Forschungswerkstatt Geschichtsdidaktik 13. Beiträge zur Tagung „geschichtsdidaktik empirisch 13".* Bern: hep.

Willis, G. B. (2005). *Cognitive Interviewing. A Tool for Improving Questionnaire Design.* Thousand Oaks, CA: Sage.

Wineburg, S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83 (1),* 73–87.

Wineburg, S. (2001). *Historical thinking and other unnatural acts: Charting the future of teaching the past.* Philadelphia, PA: Temple University Press.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest version 2.0: generalized item response modelling software.* Camberwell, AUS: ACER Press.

Zabold, S. (im Druck). *Vor dem ersten Geschichtsunterricht: zur empirischen Erschließung des historischen Denkens junger Lerner.* Stuttgart: Kohlhammer.

Gabriel Nagy, Benjamin Nagengast, Andreas Frey, Michael Becker, Norman Rose

Itempositionseffekte in Large-Scale-Assessments

Das vorliegende Projekt widmet sich einem spezifischen Testkontexteffekt, der als (Item-)Positionseffekt bekannt ist und die systematische Variation von statistischen Eigenschaften von Einzelitems in Abhängigkeit ihrer Darbietungsposition in einem Leistungstest bezeichnet. Im Fokus der Betrachtung stehen individuelle Unterschiede in der Höhe von Positionseffekten, deren Korrelate und die Auswirkungen der Vernachlässigung von Positionseffekten insbesondere in Large-Scale-Assessments. Die auf Grundlage exitierender Datensätze von Large-Scale-Assessments durchgeführten Analysen dokumentieren systematische Zusammenhänge von Positionseffekten mit Merkmalen der Schülerinnen und Schüler, der Einzelschulen und der Erhebungszeitpunkte. Die Ergebnisse zeigen ferner, dass Positionseffekte zu ignorieren zu verzerrten Ergebnissen hinsichtlich Leistungsunterschieden, Korrelaten von Testleistungen und Leistungszuwächsen führen kann.

1 Hintergrund

Die Erfassung schülerseitiger Kompetenzen ist das Hauptanliegen nahezu aller bildungswissenschaftlicher Large-Scale-Assessments. Diese im Deutschen auch als Schulleistungsstudien bezeichneteten groß angelegten empirischen Untersuchungen sollen einen Einblick in die Verteilung der Kompetenzausprägungen und deren Entwicklung in einer oder mehrere Populationen von Schülerinnen und Schülern geben. Um dieses Ziel zu erreichen, wurden sophistizierte Messmodelle und Prozeduren entwickelt, die eine unverzerrte und gleichermaßen reliable Schätzung der Populationsverteilung der Kompetenzen ermöglichen sollen (z. B. Wu, 2005). Typische Ansätze der Kompetenzmessung sind in der Tradition der Item-Response-Theorie (IRT) verwurzelt. Ziel dieser Verfahren ist, die den beobachteten Itemantworten zugrunde liegenden Kompetenzausprägungen abzuschätzen.

Eine zentrale Idee des IRT-Ansatzes ist, dass die beobachteten Itemantworten eine Funktion feststehender Itemeigenschaften (z. B. Schwierigkeiten) und Personenmerkmale (d. h. Kompetenzausprägungen) sind. Bisherige Untersuchungen liefern aber eine Fülle von Belegen dafür, dass diese Annahme in vielen Schulleistungsstudien nicht erfüllt ist. Ein robuster Befund ist, dass die Wahrscheinlichkeit einer korrekten Itemantwort von der Position des Items in der dargebotenen Sequenz von Testitems abhängt. Je näher ein Item zum Testende hin positioniert wird, desto geringer fällt die Lösungswahrscheinlichkeit des Items typischerweise aus (Leary & Dorans, 1985). Derartige Itempositionseffekte können als eine Stör-

quelle verstanden werden, die sich auf die Ergebnisse der Leistungsmessung auswirkt.

Die Implikationen von Itempositionseffekten für die Validität der Rückschlüsse in Large-Scale-Assessments hängen maßgeblich von den Eigenschaften der Itempositionseffekte ab. Itempositionseffekte wurden lange Zeit als feststehende Effekte betrachtet, deren Ausprägung nicht zwischen Individuen, Schulen und Messzeitpunkten variiert (z. B. Meyers, Miller & Way, 2009). Insofern diese Sichtweise zutrifft, wirken sich Itempositionseffekte nicht zwingend auf Leistungsvergleiche zwischen Schülerinnen und Schülern, Gruppen von Schülerinnen und Schülern und Messzeitpunkten aus, da diese Effekte mittels eines geeigneten Testheftdesigns (Frey, Hartig & Rupp, 2009) ausbalanciert bzw. konstant gehalten werden können (Johnson, 1990). Neuere Arbeiten liefern jedoch Belege dafür, dass die Höhe von Itempositionseffekten zwischen Personen (Debeer & Janssen, 2013; Hecht, Weirich, Siegle & Frey, 2015), Schulen (Debeer, Buchholz, Hartig & Janssen, 2014) und Messzeitpunkten (Nagy, Lüdtke, Köller, Heine & Mang, 2015) variieren kann. Derartige Störeinflüsse können nicht durch das Testheftdesign kontrolliert werden, da die Höhe des Itempositionseffekts keine alleinige Funktion der verwendeten Testform ist.

Ausgehend von der Feststellung, dass die Ausprägung von Itempositionseffekten zwischen Personen variiert, wurden diese als Indikatoren der individuellen Bearbeitungspersistenz in einem Leistungstest vorgeschlagen (Debeer et al., 2014). Diese Sichtweise impliziert, dass die in Large-Scale-Assessments verwendeten Testwerte neben der zu erfassenden Kompetenzausprägung auch von der individuellen Bearbeitungspersistenz abhängen. Damit ergeben sich direkte Konsequenzen für die Validität von Rückschlüssen, die auf Grundlage von Testwerten getroffen werden. Diese betreffen unter anderem die Abschätzung (1) der Leistungsheterogenität in einer (Teil-)Population, (2) der Zusammenhänge individueller Kompetenzausprägungen mit anderen Merkmalen und (3) von Kompetenzzuwächsen über einen vorgegebenen Beschulungszeitraum.

Interindividuell variierende Itempositionseffekte führen zu einer Überschätzung der Leistungsheterogenität, da sie die Variabilität von Testwerten künstlich erhöhen. Ebenso gilt, dass Korrelationen zwischen Testwerten und Außenkriterien ein verzerrtes Abbild der entsprechenden Kriteriumskorrelationen der Kompetenzen darstellen, da die berechneten Zusammenhänge auch von der Korrelation der Bearbeitungspersistenz mit den untersuchten Kovariaten abhängen. Schließlich können sich Itempositionseffekte auf die Abschätzung von Leistungszuwächsen auswirken, da sich die Bearbeitungspersistenz der getesteten Schülerinnen und Schüler zwischen den Messzeitpunkten verändern kann.

Die aufgezählten Implikationen von Itempositionseffekten ergeben sich direkt aus deren Konzeption als variable Störgrößen. Bis dato finden sich aber kaum Arbeiten aus dem Schulleistungsbereich, die sich der Frage der Korrelate und der zeitlichen Veränderung von Itempositionseffekten gewidmet haben. Im vorliegenden Projekt werden diese Forschungslücken aufgegriffen und exemplarisch anhand existierender Datensätze untersucht.

2 Zielstellungen des Projekts

Das vorliegende Projekt gliedert sich in vier Teilstudien auf, die sich der Untersuchung unterschiedlicher Aspekte von Itempositionseffekten widmen. Studie 1 fokussiert auf die Erfassung individueller Unterschiede in Itempositionseffekten und deren Korrelate mittels eines mehrdimensionalen IRT-Modells. In Studie 2 werden Positionseffekte in der nationalen Erweiterung der PISA-2006-Studie untersucht. Konkret werden Prädiktoren von Positionseffekten auf der Schüler- und Schulebene betrachtet. In der dritten Studie werden Itempositionseffekte im Längsschnitt untersucht. Im Fokus steht die mittlere Veränderung sowie die Rangstabilität von Itempositionseffekten. Die vierte Studie widmet sich schließlich der Frage, inwieweit Itempositionseffekte mit anderen Testkontexteffekten (Brennan, 1992) interagieren.

Die Erfassung individueller Unterschiede in Itempositionseffekten setzt erstens die Verfügbarkeit von Testheftdesigns voraus, in denen die gleichen Items unterschiedlichen Schülerinnen und Schülern an unterschiedlichen Positionen eines Tests dargeboten werden. Zweitens setzen die verwendeten Verfahren eine ausreichend große Stichprobe voraus, da sie sich in der Regel mehrdimensionaler IRT-Verfahren bedienen (Debeer & Janssen, 2013). Diese Vorgaben implizieren, dass viele derzeit frei nutzbare Large-Scale-Datensätze nicht für die hier anvisierten Arbeiten in Betracht kamen, da sie keine ausreichende Variation der Positionen einzelner Items bieten. Die Durchführung einer eigens geplanten Primärdatenerhebung hat zwar einige Vorteile, wurde aber vor dem Hintergrund des zum Zeitpunkt der Projektplanung herrschenden Wissensstands als zu kostenintensiv eingeschätzt.

Die ausgewählten Datensätze ermöglichen die Modellierung individueller Unterschiede in Itempositionseffekten und erlauben die Behandlung unterschiedlicher Aspekte des Gegenstandbereichs, die sowohl grundlagenwissenschaftlich als auch anwendungsorientiert ausgerichtet sind. So liefert der Datensatz der TRAIN-Studie individuelle Hintergrundvariablen, die aus einer theoretischen Perspektive als zentrale Determinaten der Bearbeitungspersistenz betrachtet werden können. Dieser Datensatz ermöglicht zudem die längsschnittliche Untersuchung von Itempositionseffekten. Der PISA-2006-Datensatz ist von seiner Struktur her prototypisch für viele aktuelle Large-Scale-Assessments und ermöglicht es abzuschätzen, inwieweit Positionseffekte mit Hintergrundvariablen kovariieren, die im Fokus traditioneller Schulleistungsstudien stehen (Schulform, Geschlecht und Merkmale des familiären Hintergrunds). Ein weiterer Datensatz aus dem Projekt MaK-adapt (Messung allgemeiner Kompetenzenadaptiv), mit Daten von drei computerbasierten Tests, bietet schließlich die seltene Möglichkeit, die Interaktion von Itempositionseffekten mit anderen Testkontexteffekten (Effekte der Domänenabfolge) zu untersuchen.

Im vorliegenden Beitrag werden die zentralen Befunde der ersten drei Studien zusammengefasst. Die vorgenommene Fokussierung ist insofern konsistent, da sich die ausgewählten Studien eng an der Thematik individueller Unterschiede in Itempositionseffekten ausrichten, während diese Perspektive in Studie 4 eine untergeordnete Rolle einnimmt. In diesem Überblicksbeitrag beschränken wir uns auf die Darstellung der inhaltlich-substanzwissenschaftlichen Ergebnisse. Verzichtet wird auf die Darstellung der verwendeten mathematischen Modelle. Um die Nachvoll-

ziehbarkeit der Ergebnisse zu gewährleisten, setzen wir, wann immer möglich, grafische Hilfsmittel ein.

3 Studie 1: Itempositionseffekte in einem Leseverständnistest

Die erste Studie (Nagy, Rose & Trautwein, 2013) verfolgte zwei Ziele. Aus methodischer Perspektive wurde ein IRT-Modell vorgestellt, das zwei Arten von Positionseffekten umfasst: zum einen Positionseffekte, die auf die Itemschwierigkeiten einwirken und zwischen Personen variieren können, und zum anderen Itempositionseffekte, die auf die Diskriminationsleistung der Items einwirken und zu einer Reduktion der Sensitivität der Testitems für die zur Lösung benötigte Kompetenz führen. Das verwendete Modell ist eine Weiterentwicklung des Verfahrens von Debeer und Janssen (2013).

Das zweite Ziel der Studie war es, Zusammenhänge zwischen Positionseffekten und schülerseitigen motivationalen und kognitiven Merkmalen zu untersuchen, die aus einer theoretischen Perspektive als Determinaten der Testbearbeitungspersistenz in Betracht kommen. Die Auswertungen liefern wichtige Hinweise für die inhaltliche Bedeutung der untersuchten Itempositionseffekte und zeigen zudem, inwieweit sich Itempositionseffekte verzerrend auf die ermittelten Konstruktzusammenhänge auswirken können. Die in dieser Studie anvisierten Forschungsfragen wurden anhand eines Leseverständnistests, der in der fünften Jahrgangsstufe in der TRAIN-Studie (Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen) eingesetzt wurde, untersucht (N = 2.830 Schülerinnen und Schüler an Haupt-, Real- und Mittelschulen, 53,6 Prozent männlich, die jeweils 30 bis 32 Items bearbeitet haben). Zwei Kovariaten wurden zur Prädiktion der Testleistung und der Bearbeitungspersistenz herangezogen, nämlich eine Selbstberichtsskala zur Erfassung der Freude am Lesen und ein Instrument zur Messung der Decodiergeschwindigkeit. Als Hypothesen wurden angenommen, dass Schülerinnen und Schüler, die für den Bereich Lesen ein höheres Interesse berichten, eine höhere Persistenz bei der Bearbeitung von Textaufgaben aufweisen sowie hinsichtlich der Decodiergeschwindigkeit, dass sich diese förderlich auf die Bearbeitungspersistenz auswirkt, da der Prozess des Lesens für schnelle Decodierer weniger ermüdend ist.

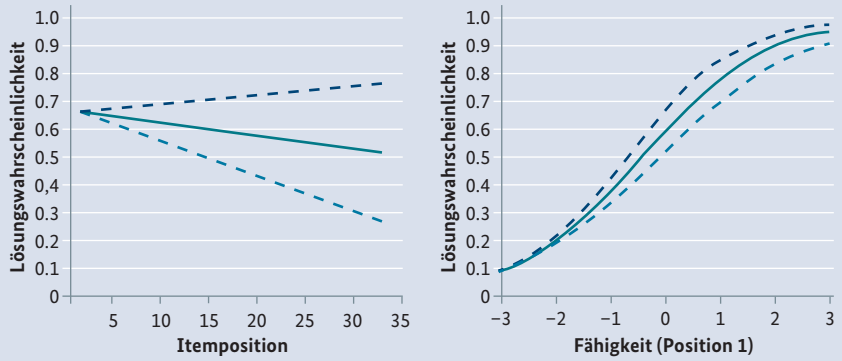
3.1 Zentrale Ergebnisse

Hinsichtlich der Positionseffekte wurde festgestellt, dass die Wahrscheinlichkeit korrekter Antworten mit der Itemposition assoziiert ist, wobei die Höhe der Abnahme zwischen Personen variiert. Ebenso wurde ein Rückgang der Sensitivität der Testitems zur Erfassung der im Fokus stehenden fokalen Kompetenz (d. h. des Leseverständnisses) ermittelt. Die entsprechenden Ergebnisse sind in Abbildung 1 dargestellt. Die linke Teilabbildung stellt den Rückgang der Lösungswahrscheinlichkeit eines prototypischen Items (d. h. mittlere Itemschwierigkeit und Itemdiskrimination) dar. Die Abbildung dokumentiert eine kontinuierliche Abnahme der Lösungswahrscheinlichkeit in Abhängigkeit der Itemposition. Individuelle Unterschiede in

der Testbearbeitungspersistenz manifestieren sich in der Variation der Leistungsabnahme zwischen Personen. Aus der Abbildung geht hervor, dass ein Teil der Schülerinnen und Schüler keinen Leistungsrückgang während der Testbearbeitung aufweisen, während für andere der Rückgang besonders stark akzentuiert ist.

In der rechten Teilabbildung 1 ist die Veränderung der Itemcharakteristikkurve eines prototypischen Items dargestellt. Die Abbildung verdeutlicht, dass der ogivenförmige Zusammenhang zwischen der Lösungswahrscheinlichkeit eines Items (y-Achse) und der individuellen Kompetenzausprägung (x-Achse) über die Itempositionen hinweg flacher wird. Dieser Befund indiziert, dass die zum Testende hin eingesetzten Items weniger sensitiv für individuelle Unterschiede in der zugrunde liegenden Kompetenz sind, wobei die Sensitivität für Unterschiede in der Testbearbeitungspersistenz zunimmt (ohne Abbildung).

Abb. 1: Linke Teilabbildung: Lösungswahrscheinlichkeit eines prototypischen Items in Abhängigkeit der Itemposition. Vorhersage für die mittlere Ausprägung des Positionseffekts und des Wertebereichs zwischen dem 10. und 90. Percentil der Verteilung des Positionseffekts. Rechte Teilabbildung: Itemchararakteristikkurven für ein prototyisches Item zu Beginn (obere gebrochene Linie), in der Mitte (durchgezogene Linie) und am Ende des Tests (untere gebrochene Linie)

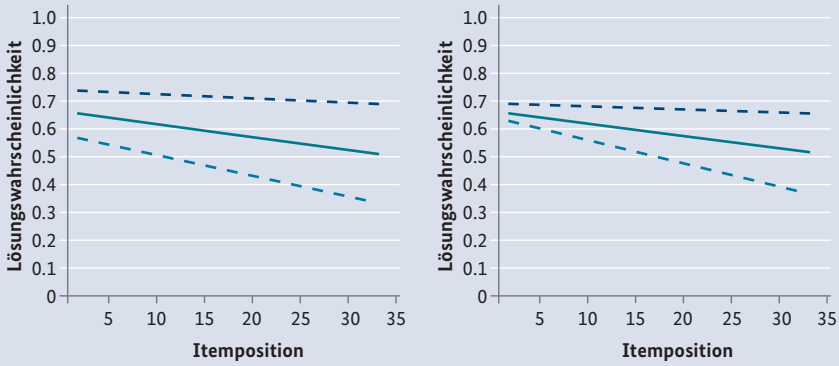


Das IRT-Modell wurde in einem zweiten Schritt um die Kovariaten Lesefreude und Decodiergeschwindigkeit erweitert. Dieses Modell ermöglicht es, die Zusammenhänge zwischen den Kovariaten und den Positionseffekten darzustellen. Es beantwortet somit die Frage, ob individuelle Unterschiede in der Testbearbeitungspersistenz mit den betrachteten Kovariaten in erwarteter Weise zusammenhängen. Auf Grundlage der Modellparameter können zudem Zusammenhänge zwischen der fokalen Testleistung und den Kovariaten getrennt für jede mögliche Itemposition berechnet werden. Damit ist es möglich, die Außenkorrelationen von Testwerten in Abhängigkeit unterschiedlicher Referenzpositionen darzustellen.

In dieser Anwendung zeigte sich, dass beide Kovariaten signifikant positiv mit dem Itempositionseffekt korreliert waren ($r = .21$; $SE = 0.04$; $p < .01$ und $r = .23$; $SE = 0.05$; $p < .01$ für Freude am Lesen und Dekodiergeschwindigkeit). Die Ergebnisse indizieren, dass Schülerinnen und Schüler mit einer höheren Lesefreude und einer

günstigeren Decodiergeschwindigkeit eine höhere Bearbeitungspersistenz aufweisen. Eine Folge dieses Zusammenhangs ist, dass die Leistungsunterschiede zwischen Schülerinnen und Schülern mit hohen und geringen Werten auf den Kovariaten in Abhängigkeit der Position der Leseverständnisitems zunehmen. Dieser Effekt ist in Abbildung 2 veranschaulicht. Eine Konsequenz dieses Ergebnismusters ist, dass die Zusammenhänge zwischen den Kovariaten und den Testergebnissen in Abhängigkeit der Position ansteigen. Die jeweiligen Kriteriumskorrelationen decken je nach Itemposition einen Wertebereich von $r = .10$ bis $.24$ (Lesefreude) bzw. $r = .27$ bis $.35$ (Decodiergeschwindigkeit) ab.

Abb. 2: Vorhergesagte Lösungswahrscheinlichkeit eines prototypischen Items in Abhängigkeit der Kovariaten Decodiergeschwindigkeit (linke Teilabbildung) und Freude am Lesen (rechte Teilabbildung). Durchgezogene Linie: Stichprobenmittelwert; gebrochene Linien: Vorhersagen für das 10. und 90. Percentil der Verteilung der Kovariaten



Ein weiterer wichtiger Befund, der sich in dieser Anwendung zeigte, ist, dass die aufgrund eines herkömmlichen IRT-Modells (ohne Berücksichtigung der Itempositionseffekte) ermittelten Konstruktzusammenhänge den aufgrund der im erweiterten Modell hinsichtlich der mittleren Itemposition erwarteten Zusammenhängen entsprachen. Dieser Befund dokumentiert, dass die in den meisten Studien ermittelten Zusammenhänge als Komposita der Kriteriumszusammenhänge mit dem intendierten Konstrukt (hier Leseverständnis) und mit der Testbearbeitungspersistenz zu verstehen sind.

3.2 Zusammenfassung

Studie 1 liefert wichtige methodische, grundlagenwissenschaftliche und anwendungsrelevante Ergebnisse. Der methodische Beitrag besteht in der Entwicklung eines flexiblen mehrdimensionalen IRT-Modells, das Itempositionseffekte erfasst, die auf Itemschwierigkeiten und die Diskriminationsleistung von Items einwirken. Das Modell stellt somit eine sinnvolle Erweiterung rezenter IRT-Ansätze zur Erfassung von Itempositionseffekten dar (z. B. Debeer & Janssen, 2013).

Aus einer inhaltlichen Perspektive tragen die Befunde zur Klärung des Konstruktstatus von Itempositionseffekten bei. Deren Interpretation als Indikatoren der Testbearbeitungspersistenz setzt voraus, dass diese mit motivationalen und kognitiven Ressourcen assoziiert sind. Die Ergebnisse der Studie 1 untermauern die Sichtweise, dass Itempositionseffekte einen Aspekt der Testbearbeitungspersistenz darstellen.

Studie 1 dokumentiert zudem die Bedeutung von Itempositionseffekten für die Abschätzung von Konstruktzusammenhängen. Kriteriumskorrelationen von (IRT-skalierten) Testwerten werden typischerweise als Zusammenhänge zwischen individuellen Kompetenzen und anderen Variablen interpretiert. Die Abhängigkeit der ermittelten Zusammenhänge von der Itemposition dokumentiert jedoch die Gefahr, die aus einer automatischen Gleichsetzung von Kompetenzen und Testwerten resultiert. Insgesamt liefern unsere Befunde Hinweise für die Annahme, dass die in Large-Scale-Assessments typischerweise verwendeten Testwerte als Indikatoren der Kompetenz mit der Testbearbeitungspersistenz konfundiert sind und dies auch für die ermittelten Kriteriumskorrelationen gilt.

4 Studie 2: Positionseffekte in der PISA-2006-Studie

Gegenstand von Studie 2 (Nagengast, Nagy, Rose & Becker, 2015) war die Untersuchung von Positionseffekten in einem für Large-Scale-Assessments prototypischen Datensatz. Zu diesem Zweck wurde die nationale Ergänzung der PISA-2006-Studie gewählt. Ein zentrales Anliegen von Large-Scale-Assessments ist die Beschreibung der Kompetenzverteilung in unterschiedlichen Teilpopulationen von Schülerinnen und Schülern. Typische Beschreibungen richten sich auf Populationen, die anhand der Schulform, des Geschlechts, des Migrationshintergrunds oder des sozioökonomischen Hintergrunds definiert sind. In Studie 2 wurde untersucht, inwieweit Positionseffekte zu verzerrten Rückschlüssen über Leistungsunterschiede zwischen Teilpopulationen führen können.

Eine zweite Fragestellung, die in dieser Studie angegangen wurde, richtet sich auf die Variabilität von Positionseffekten auf der Ebene von Schulen. Bis heute liegt lediglich eine Untersuchung vor, die für den Bereich Leseverständnis zwischenschulische Unterschiede in der Ausprägung von Positionseffekten dokumentiert (Debeer et al., 2014). In Studie 2 untersuchten wir deshalb diese Fragstellung für alle in PISA getesteten Kompetenzbereiche (Naturwissenschaften, Mathematik und Lesen). Obwohl die entsprechenden Auswertungen deutliche Hinweise für zwischenschulische Unterschiede in Positionseffekten erbracht haben, haben wir uns aus Platzgründen dafür entschieden, diese im vorliegenden Überblicksbeitrag nicht zu berichten.[1]

1 Das Befundmuster lässt sich grob wie folgt zusammenfassen: Positionseffekte in allen untersuchten Tests wiesen eine statistisch signifikante Varianz auf Schulebene auf. Ein Teil dieser Variabilität konnte auf Kompositionseffekte der Schülerschaft zurückgeführt werden (Geschlecht, Migrationshintergrund, sozioökonomischer Hintergrund und selbst berichtete Testanstrengung). Darüber hinaus waren die auf der Schulebene vorliegenden Positionseffekte mit der Schulform (alle Leistungsbereiche), dem Anteil der Schülerschaft mit Migrationshintergrund

Als individuelle Prädiktoren der Testleistung und des Positionseffektes wurden in dieser Untersuchung folgende Variablen berücksichtigt: Geschlecht, Migrationshintergrund (mindestens ein Elternteil im Ausland geboren), sozioökonomischer familiärer Hintergrund und die zum Ende des Tests selbst berichtete Testanstrengung (dichotomisiert). In der nachfolgenden Darstellung beschränken wir uns aus Platzgründen auf die Effekte des Geschlechts, des sozioökonomischen Hintergrunds und der besuchten Schulform. Wir haben uns für diese Variablen entschieden, da diese im Fokus der Berichtserstattung aller Large-Scale-Assessments stehen.

4.1 Zentrale Ergebnisse

Die Auswertung der Daten geschah anhand eines Verfahrens, das mehrere Analyseschritte umfasste. Wir haben uns gegen eine direkte Modellierung von Itempositionseffekten entschieden, da deren Bestimmung anhand eines nicht linearen mehrebenenanalytischen IRT-Modells (Schülerinnen und Schüler geschachtelt in Schulen) sich als rechnerisch nicht handhabbar erwies. Stattdessen haben wir die Positionseffekte auf Ebene der Positionen der in PISA eingesetzten Itemcluster erfasst (d. h. Clusterpositionseffekte). In PISA werden die Testitems in Itemcluster zusammengefasst, deren Bearbeitungszeit jeweils ca. 30 Minuten beträgt. Jeweils vier Itemcluster werden anschließend zu einem Testheft zusammengefasst. Die Verteilung der Itemcluster geschieht dabei auf Basis eines balancierten unvollständigen Block-Designs, das unter anderem gewährleistet, dass jedes Itemcluster jeweils genau einmal an jeder der vier möglichen Clusterpositionen eines Testhefts dargeboten wird (Frey, Carstensen, Walter, Rönnebeck & Gomolka, 2008). Da die durch das Testheftdesign spezifizierten Testhefte randomisiert den Schülerinnen und Schülern zugeteilt wurden, können (Cluster-)Positionseffekte anhand der zwischen den Clusterpositionen vorliegenden Leistungsunterschiede ermittelt werden. Dieses Verfahren setzt voraus, dass für jede Kombination, die sich aus der Kreuzung der Itemcluster und der Clusterposition ergibt, Testwerte vorliegen. In Studie 2 wurden die Testwerte mittels getrennter Rasch-Skalierungen auf Grundlage der Plausible-Value-Technik (PV; Mislevy, Beaton, Kaplan & Sheehan, 1992) ermittelt. Das in PISA 2006 verwendete Testheftdesign ist in Tabelle 1 wiedergegeben.

Die für jede Domäne ermittelten PVs wurden anschließend mittels eines neu entwickelten multivariaten Mehrebenenmodells ausgewertet. In diesem Modell werden die verwendeten PVs in zwei Komponenten zerlegt, nämlich die mittlere Ausprägung der PVs über alle Itemcluster und alle Clusterpositionen sowie die positions- und clusterpezifischen Abweichungen von diesem Mittelwert. In seiner einfachsten Form erlaubt dieses Modell die Zerlegung der Variabilität der Testwerte in Leistungsmittelwerte und Positionseffekte getrennt für die Schüler- und Schulebene. Das Modell

(Mathematik) und dem Anteil der Schülerschaft mit geringer Testanstrengung (Naturwissenschaften und Lesen) assoziiert. Mit Ausnahme der Schulformeffekte lassen sich diese Effekte im Sinne von Kontexteffekten interpretieren, die darauf hinweisen, dass die Testbearbeitungspersistenz in Abhängigkeit der Zusammensetzung der Schülerschaft (nach Kontrolle der Individualvariablen) variiert.

wurde in einem nächsten Schritt um die verwendeten Kovariaten erweitert. Hier berichten wir Befunde, die sich auf die Effekte der Schülermerkmale Geschlecht und Migrationshintergrund sowie des Schulmerkmals Schulform beziehen.

Tabelle 1: Testheftdesign von PISA 2006

	Booklet												
	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11	B12	B13
Position 1	S1	S2	S3	S4	S5	S6	S7	M1	M2	M3	M4	R1	R2
Position 2	S2	S3	S4	M3	S6	R2	R1	M2	S1	M4	S5	M1	S7
Position 3	S4	M3	M4	S5	S7	R1	M2	S2	S3	S6	R2	S1	M1
Position 4	S7	R1	M1	M2	S3	S4	M4	S6	R2	S1	S2	S5	M3

Anmerkungen: S = Science (Naturwissenschaften), M = Mathematics (Mathematik), R = Reading (Lesen)

In die Auswertung gingen insgesamt N = 33.480 Schülerinnen und Schüler ein, wobei wir eine kombinierte Stichprobe herangezogen haben, die aus den Populationen der 15-jährigen Schülerinnen und Schüler und der Neuntklässlerinnen und Neuntklässler zusammengesetzt war. Um eine interpretierbare Abschätzung der Schulformunterschiede zu ermöglichen, haben wir uns auf Schülerinnen und Schüler der alten Bundesländer konzentriert (mit Ausnahme des Saarlandes) und zwischen den traditionellen Schulformen des dreigliedrigen Sekundarschulsystems (d. h. Hauptschule, Realschule und Gymnasium) sowie der zusammengefassten Gruppe der integrierten Gesamtschule und der Schulen mit mehreren Bildungsgängen unterschieden. Schülerinnen und Schüler mit spezifischem Förderbedarf wurden aus den Auswertungen ausgeschlossen.

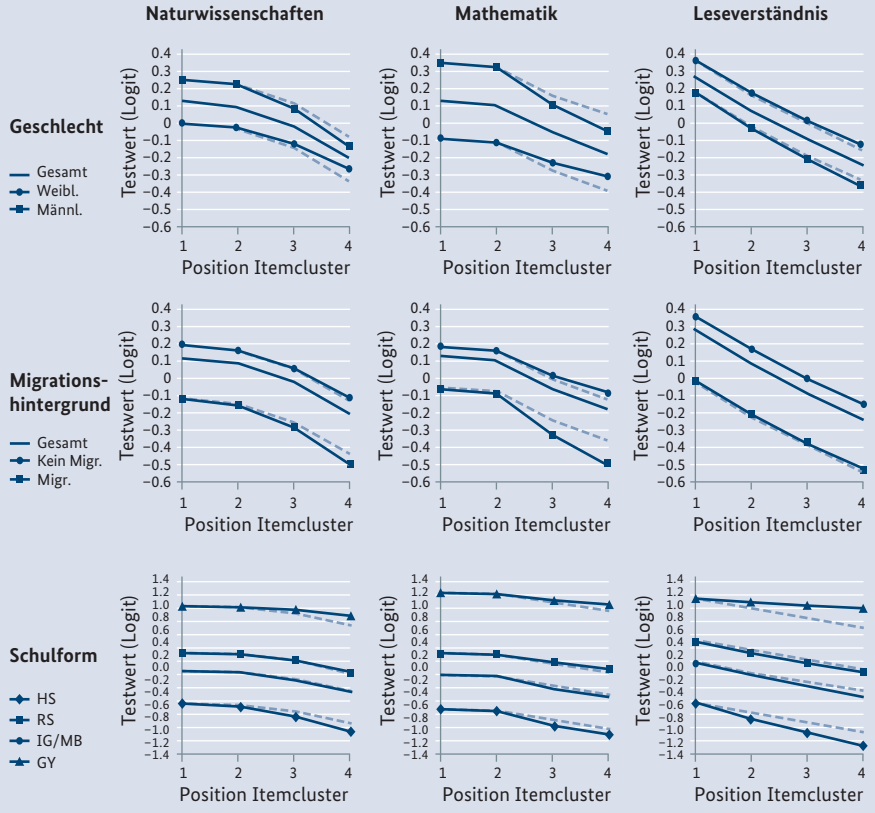
In Abbildung 3 sind die über die Itemcluster gemittelten Testleistungsverläufe in Abhängigkeit der Kovariaten Geschlecht und Migrationshintergrund sowie der Clusterposition dargestellt. Die grau dargestellten Linien beschreiben die mittleren Kompetenzverläufe in der Gesamtstichprobe. Die gestrichelten Linien geben den erwarteten Verlauf in jeder Gruppe unter der Annahme konstanter Positionseffekte wider. Abweichungen zwischen den gruppenspezifischen Kompetenzverläufen (schwarze durchgezogene Linien) und den gestrichelten Linien indizieren somit die Verzerrung, die sich aus der Ausblendung gruppenspezifischer Positionseffekte ergibt.

Aus Abbildung 3 geht hervor, dass sich die Gestalt der mittleren Positionseffekte (d. h. Kompetenzverläufe) zwischen den Domänen unterscheidet. Die Positionseffekte weisen im Bereich Lesen eine lineare Form auf, während sie in den Bereichen Naturwissenschaften und Mathematik erst in der zweiten Testhälfte deutlich zutage treten (ab Clusterposition 3). Hinzu kommt, dass sich die auf der Logit-Metrik ausgedrückte Höhe der Positionseffekte (Differenz zwischen der ersten und letzten Clusterposition) zwischen den Domänen unterscheidet. Die Positionseffekte fielen im Bereich Lesen am höchsten aus.

Abbildung 3 stellt die Positionseffekte in Abhängigkeit der Kovariaten dar. Jungen zeichneten sich im Vergleich zu Mädchen durch stärkere Abnahmen in allen

Kompetenzbereichen aus. Dieser Unterschied trat insbesondere für den Bereich Mathematik zutage, während er im Bereich Lesen vernachlässigbar erscheint. Stärkere Kompetenzrückgänge fanden sich zudem für Schülerinnen und Schüler mit Migrationshintergrund in den Bereichen Naturwissenschaften und Mathematik, wobei der Unterschied im Bereich Mathematik erneut am deutlichsten ausgeprägt war. Hinsichtlich der besuchten Schulform fanden sich für alle drei Domänen vergleichbare Befunde. Die Positionseffekte sind in den Gymnasien weitgehend vernachlässigbar und in den Hauptschulen am stärksten akzentuiert. Die Positionseffekte unterschieden sich im Bereich Lesen am deutlichsten zwischen den Schulformen.

Abb. 3: Testwerte nach Domänen (Spalten), Subgruppen (Zeilen) und Positionen der Itemcluster (x-Achse)



Die in Abbildung 3 dargestellten Zusammenhänge implizieren, dass die in Abhängigkeit der Hintergrundmerkmale ermittelten Kompetenzunterschiede von der Clusterposition abhängen. Die für die Bereiche Naturwissenschaften und Mathematik hinsichtlich der ersten Clusterposition ermittelten Vorteile der Jungen reduzieren

sich über den Verlauf der Testbearbeitung. Die Kompetenzunterschiede zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund nehmen hingegen zu (Naturwissenschaften und Mathematik), und ein vergleichbarer Befund findet sich für die Schulformunterschiede.

Tabelle 2: Mittelwertunterschiede (Logit-Metrik) für die erste (P1), letzte (P4) und gemittelte Itemposition (Mittel) sowie prozentuale Veränderung der Unterschiede relativ zur ersten Position

	Naturwissenschaft			Mathematik			Lesen		
	P1	Mittel	P4	P1	Mittel	P4	P1	Mittel	P4
*Geschlecht (**männlich** vs. weiblich)*									
Effekt	0.25	0.21	0.12	0.44	0.37	0.26	−0.18	−0.21	-0.24
% Veränderung		(−20%)	(−51%)		(−17%)	(−41%)		(16%)	(30%)
Migrationshintergrund (mit vs. ohne)									
Effekt	−0.31	−0.34	−0.39	−0.24	−0.31	−0.42	−0.18	−0.38	−0.37
% Veränderung		(10%)	(26%)		(33%)	(77%)		(−2%)	(−3%)
*Schulformenver**gleiche***									
Gy vs. IG/MB									
Effekt	1.03	1.11	1.22	1.32	1.38	1.47	1.00	1.22	1.43
% Veränderung		(7%)	(19%)		(5%)	(11%)		(22%)	(42%)
GY vs. RS									
Effekt	0.74	0.80	0.88	0.98	1.00	1.03	0.66	0.83	1.00
% Veränderung		(7%)	(19%)		(2%)	(5%)		(27%)	(52%)
GY vs. HS									
Effekt	1.55	1.67	1.86	1.87	1.95	2.07	1.66	1.94	2.20
% Veränderung		(8%)	(20%)		(5%)	(11%)		(17%)	(33%)
IG/MB vs. RS									
Effekt	−0.29	−0.31	−0.34	−0,34	−0.38	−0.44	−0.35	−0.39	−0.43
% Veränderung		(7%)	(17%)		(12%)	(28%)		(13%)	(25%)
IG/MB vs. HS									
Effekt	0.52	0.56	0.63	0.55	0.57	0.61	0.65	0.71	0.77
% Veränderung		(9%)	(22%)		(5%)	(11%)		(9%)	(18%)
RS vs. HS									
Effekt	0.81	0.87	0.97	0.89	0.96	1.04	1.00	1.11	1.20
% Veränderung		(8%)	(21%)		(7%)	(17%)		(11%)	(20%)

Anmerkungen: GY = Gymnasium, IG/MB = Integrierte Gesamtschule und Schulen mit mehreren Bildungsgängen, RS = Realschule, HS = Hauptschule

Aus einer praktischen Perspektive stellt sich somit die Frage nach dem Grad der Verzerrung der Effektstärken, die sich bei einer Ausblendung der Positionseffekte ergeben. In Tabelle 2 sind die Leistungsunterschiede an den Extrempositionen 1 und 4 sowie die über alle Positionen gemittelten Unterschiede abgetragen. Die Unterschiede an Position 1 können als näherungsweise frei von Positionseffekten betrachtet werden, während die Ergebnisse an Position 4 maximal vom Positionseffekt betroffen sind. Die gemittelten Unterschiede entsprechen näherungsweise den in Large-Scale-Assessments ermittelten Effekten. Tabelle 2 dokumentiert, dass sich die Kompetenzunterschiede zwischen den Positionen 1 und 4 zum Teil deutlich un-

terscheiden, während die gemittelten Unterschiede sich in vielen Fällen vom Betrag her nicht wesentlich von Ergebnissen an der ersten Clusterposition unterschieden. Größere Unterschiede finden sich jedoch für den Effekt des Migrationshintergrunds (Mathematik) und der Schulformunterschiede (Lesen).

4.2 Zusammenfassung

Gegenstand von Studie 2 ist die Untersuchung von Positionseffekten in einem prototypischen Large-Scale-Assessment. Die Ergebnisse dokumentieren, dass die Testbearbeitungspersistenz nicht nur zwischen Individuen, sondern auch zwischen Einzelschulen variiert. Ein weiterer wichtiger Befund ist, dass sich die in Large-Scale-Assessments häufig betrachteten Teilpopulationen von Schülerinnen und Schülern in ihrer Testbearbeitungspersistenz voneinander unterscheiden. Dieser Befund impliziert, dass die Validität von Kompetenzvergleichen unter Umständen gefährdet sein könnte, da die ermittelten Kompetenzunterschiede nicht ausschließlich auf die zugrunde liegenden Kompetenzausprägungen, sondern darüber hinaus von den gruppenspezifischen Ausprägungen der Testbearbeitungspersistenz abhängen. Für die in dieser Studie betrachteten Hintergrundvariablen zeigte sich jedoch, dass die Verzerrungen in vielen Fällen vergleichsweise gering ausfallen. Dieser Befund ist zwei Ursachen geschuldet. Erstens zeichnen sich die Positionseffekte in den Bereichen Naturwissenschaften und Mathematik durch einen nicht linearen Verlauf aus, wonach die erste Hälfte der Tests nur im geringen Maß von Positionseffekten betroffen ist. In der Konsequenz üben die zum Teil deutlichen Gruppenunterschiede in den Positionseffekten im Mittel (d. h. über alle Positionen hinweg) einen vergleichsweise schwachen Einfluss auf das Gesamtergebnis aus. Zweitens waren die hinsichtlich der ersten Itemclusterposition ermittelten Effekte (die per Definition nicht von Clusterpositionseffekten betroffen sind) bereits relativ stark ausgeprägt, sodass sich die über die nachfolgenden Testteile gemittelten Positionseffekte relativ dazu nur gering auf den Gesamtmittelwert auswirken.

An dieser Stelle mag sich der Eindruck einstellen, dass Positionseffekte eher vernachlässigbare Konsequenzen für die in Schulleistungsstudien erzielten Ergebnisse aufweisen. Jedoch sei darauf verwiesen, dass dies nicht unbedingt gelten muss. Erstens waren einige Vergleiche durchaus in substanziellem Umfang von Positionseffekten betroffen (z. B. Migrationshintergrund im Bereich Mathematik und Schulformvergleiche im Bereich Lesen), sodass die erzielten Ergebnisse nicht generell als robust gegenüber Positionseffekten gelten können. Zweitens sind viele in Schulleistungsstudien untersuchte Effekte weitaus schwächer ausgeprägt als die hier betrachteten Zusammenhänge (z. B. Effekte familiärer Interaktionsstile). Sofern diese Variablen ähnlich stark wie die hier untersuchten Merkmale mit dem Positionseffekt assoziiert sind, können sich größere Validitätsprobleme ergeben. Drittens gilt anzumerken, dass PISA auf ein rotiertes Testheftdesign mit balancierten Itemclusterpostionen setzt. In diesem Design sind die gemessenen Kompetenzen über die mittlere Position definiert. Im Gegensatz dazu greifen viele andere groß angelegte Schulleistungsstudien auf eine festgelegte Sequenz von domänenspezifischen Tests zurück (z. B. Na-

turwissenschaft an Position 1, Mathematik an Position 2 usw.). In solchen Designs sind die auf Grundlage der zum Ende der Sequenz dargebotenen Tests stärker vom Effekt der Testbearbeitungspersistenz betroffen. Dies führt beispielsweise dann zu erheblichen Problemen, wenn die Testergebnisse kriteriumsorientiert (z. B. bezüglich Kompetenzstufen) interpretiert werden sollen. In Abhängigkeit der Position, auf der eine Domäne vorgegeben wird, kommt man zu unterschiedlichen Schlüssen, welche Anteile der Population bestimmte Dinge wissen und können. Insgesamt ergibt sich somit die Einschätzung, dass die durch die Positionseffekte induzierten Verzerrungen in Abhängigkeit des verwendeten Testheftdesigns in verschärfter oder abgemilderter Form zutage treten können (Weirich, Hecht & Böhme, 2014).

5 Studie 3: Itempositionseffekte im Längsschnitt

Retestdesigns sind in der empirischen Bildungsforschung weit verbreitet. An solche Studien ist die Hoffnung geknüpft, dass sie die Erfassung individueller Lerngewinne und deren Korrelate erlauben. Bis heute ist aber faktisch nichts über die Rolle von Itempositionseffekten in Längsschnittstudien bekannt. Im Prinzip können Itempositionseffekte die Ergebnisse von Längsschnittstudien in verschiedener Weise beeinflussen. So wirken sich Zu- oder Abnahmen in der mittleren Ausprägung von Positionseffekten auf die Schätzung der mittleren Lerngewinne aus. Insofern die Änderungen in den mittleren Ausprägungen der Positionseffekte zwischen verschiedenen zu vergleichenden Teilpopulationen (z. B. Schulformen) unterscheiden, ist zudem die Validität von Rückschlüssen über Gruppenunterschiede in Lerngewinnen gefährdet.

Da bis dato keine empirischen Befunde zur zeitlichen Entwicklung von Itempositionsffekten vorliegen (für eine aktuelle Studie siehe aber Nagy et al., 2015), können kaum empirisch fundierte Erwartungen über die Veränderung der Testbearbeitungspersistenz in Längsschnittstudien formuliert werden. So erscheinen Abnahmen in der mittleren Ausprägung von Positionseffekten in vielen Schulleistungsstudien plausibel, da ältere Schülerinnen und Schüler über ein größeres Portfolio mentaler Ressourcen verfügen, die negativen Positionseffekten entgegenwirken können (z. B. Decodiergeschwindigkeit). Andererseits könnten sich negative Positionseffekte über die Zeit verstärken, da sich die wiederholte Teilnahme an einer Testung motivationshemmend auswirken könnte. Hinsichtlich der Reteststabilität von Positionseffekten erscheinen auch unterschiedliche Szenarien plausibel. So könnten diese eine relativ zeit- und situationsüberdauernde Größe repräsentieren oder sich durch eine hohe Situationsspezifität auszeichnen.

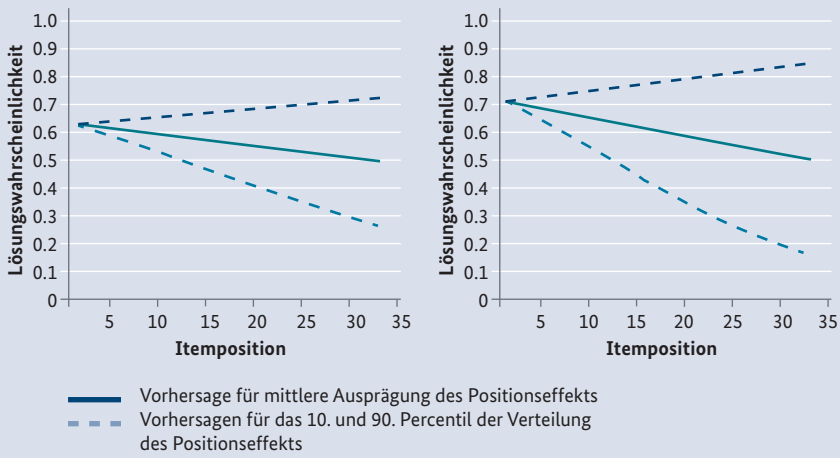
Die hier skizzierten Fragen stehen im Zentrum von Studie 3 (Nagy, Rose & Nagengast, 2015), die ebenso wie Studie 1 exemplarisch für den Bereich Leseverständnis durchgeführt wurde. Zu diesem Zweck wurde erneut auf die TRAIN-Daten zurückgegriffen (*N = 3.092),* wobei hier Leistungswerte der in den Klassenstufen 5 und 6 durchgeführten Erhebungen eingingen. In dieser Studie wurde die besuchte Schulform (Haupt-, Real- und Mittelschule) als Kovariate verwendet, da Schulformunterschiede in Lerngewinnen häufig im Fokus längsschnittlicher Schulleistungsstudien stehen.

5.1 Zentrale Ergebnisse

Die Auswertung der Daten geschah auf Grundlage einer Erweiterung des von Debeer und Janssen (2013) vorgeschlagenen IRT-Modells für Positionseffekte. Unsere Erweiterung vereinigt IRT-Modelle für wiederholte Messungen (von Davier, Xu & Carstensen, 2011) mit dem Modell für Positionseffekte. Das in Studie 1 verwendete Modell erlaubt zwar eine vollständigere Erfassung von Itempositionseffekten, seine Umsetzung für den Längsschnittfall erwies sich jedoch als problematisch, da es eine große Zahl nicht linearer Parameterrestriktionen beinhaltet.

Abbildung 4 beschreibt die messzeitpunktspezifischen Ausprägungen der Positionseffekte anhand der vorhergesagten Lösungswahrscheinlichkeiten für ein typisches Item (mittlere Schwierigkeit und Itemdiskrimination). Aus der Abbildung geht hervor, dass die mittlere Ausprägung der Itempositionseffekte in Klassenstufe 6 deutlich ansteigt. Ebenso wurde ein prägnanter Anstieg der Variabilität der Positionseffekte festgestellt. In diesem Modell zeigte sich, dass die individuellen Unterschiede in den Positionseffekten eine geringe Stabilität aufweisen (r = .22; SE = 0.11; p = .040). Dieser Befund indiziert, dass die individuelle Ausprägung der Itempositionseffekte eine hohe Situationsspezifität aufweist. Die hinsichtlich der ersten Itemposition definierte Traitvariable zeichnete sich hingegen durch ein hohe Stabilität aus (r = .74; SE = 0.03; p < .001).

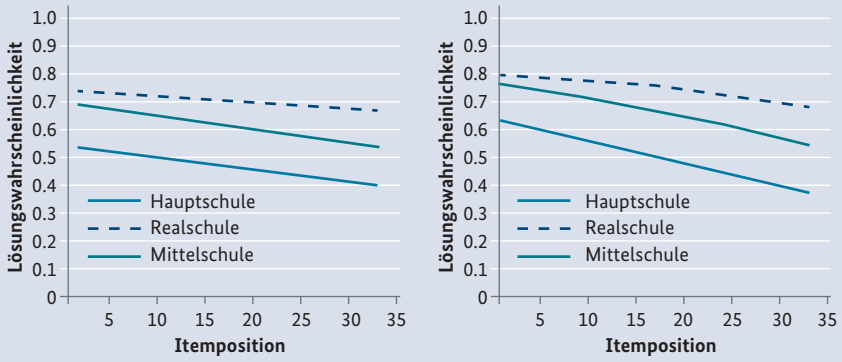
Abb. 4: Lösungswahrscheinlichkeit eines prototypischen Items in Abhängigkeit der Itemposition in den Klassenstufen 5 (linke Teilabbildung) und 6 (rechte Teilabbildung)



In Abbildung 5 werden die Schulformunterschiede anhand der Lösungswahrscheinlichkeit eines prototypischen Items für beide Messzeitpunkte dargestellt. Wie ersichtlich wird, unterschieden sich Leistungen zwischen den Schulformen zu beiden Messzeitpunkten deutlich voneinander. Ein interessanter Befund ist, dass die Leistungsunterschiede in Klassenstufe 5 nur im geringen Maß auf Unterschie-

de in Itempositionseffekten zurückzuführen sind (d. h. relativ parallel verlaufende Lösungswahrscheinlichkeiten über Positionen). Die Ergebnisse für den zweiten Messzeitpunkt unterscheiden sich davon, da die negativen Itempositionseffekte in der Gruppe der Schülerinnen und Schüler an Haupt- und Mittelschulen deutlicher ausgeprägt sind. In diesen Gruppen wurde ein markanter Anstieg der Itempositionseffekte ermittelt. Dieser Befund indiziert, dass die Ausblendung von Itempositionseffekten eventuell zu einer verfälschten Einschätzung über schulformspezifische Lerngewinne führen kann. So wurde der Lerngewinn in Hauptschulen ohne Berücksichtigung des Itempositionseffekts auf $d = 0.42$ (relativiert an der Ausgangsmessung) geschätzt. Die Effektstärke stieg nach Berücksichtigung des Positionseffekts um rund 25 Prozent auf $d = 0.52$ an.

Abb. 5: Lösungswahrscheinlichkeit eines Items mit mittlerer Schwierigkeit und Diskrimination in Abhängigkeit der Itemposition und der besuchten Schulform. Darstellungen für die Klassenstufen 5 (linke Teilabbildung) und 6 (rechte Teilabbildung)



5.2 Zusammenfassung

Gegenstand von Studie 3 ist die Untersuchung von Itempositionseffekten in einem längsschnittlichen Setting. Die Befunde sind für die angewandte Schulleistungsforschung von Bedeutung, da sie zeigen, dass die wiederholte Erfassung von Schülerinnen und Schülern zu Testartefakten führen kann. Der hier festgestellte Anstieg in den mittleren Itempositionseffekten kann sich, sofern er nicht berücksichtigt wird, in einer Unterschätzung der Lerngewinne niederschlagen. Ebenso können Zusammenhänge von Hintergrundmerkmalen mit den Itempositionseffekten zu einer falschen Einschätzung ihrer Effekte auf die Leistungsentwicklung führen.

Die in Studie 3 berichteten Befunde wurden in der Essenz kürzlich in einem unabhängigen Datensatz repliziert. Nagy und Kollegen (2015) konnten einen Anstieg der Positionseffekte in einer Large-Scale-Studie replizieren, wobei die Veränderung in nicht gymnasialen Schulformen besonders akzentuiert ausfiel und die Ausblendung von Positionseffekten sich deutlich auf die Abschätzung der Leistungszuwächse auswirkte.

An dieser Stelle kann festgehalten werden, dass die Auswirkungen von Positionseffekten in Längsschnittstudien besonders gravierend ausfallen können. Der Grund hierfür ist, dass Lerngewinne in späteren Phasen der Beschulung (z. B. in der gymnasialen Oberstufe) relativ gering ausfallen (z. B. Bloom, Hill, Black & Lipsey, 2008), sodass eine valide Interpretation der entsprechenden Effekte eine hohe Präzision der Zuwachsschätzung voraussetzt. Ebenso gilt, dass die Zusammenhänge der Zuwächse in den Testwerten mit anderen Kovariaten im Vergleich zu querschnittlichen Zusammenhängen gering sind, sodass Itempositionseffekte zu fehlerhaften Rückschlüssen über die Determinanten von Lerngewinnen führen könnten.

6 Abschließendes Resümee

Itempositionseffekte sind im Kontext von Low-Stakes-Large-Scale-Assessments eher die Regel als die Ausnahme. Derartige Effekte wurden bis vor kurzer Zeit statistisch als feste Effekte behandelt, die als eine Eigenschaft des verwendeten Tests und nicht der getesteten Schülerinnen und Schüler konzipiert sind. In Übereinstimmung mit der aktuellen Literatur (z. B. Debeer & Janssen, 2013) wurde im vorliegenden Projekt die Sichtweise eingenommen, dass Itempositionseffekte eher als ein auf der Personenseite lokalisiertes Phänomen zu behandeln sind, da sie im Sinne individueller Reaktionen auf Leistungstests zu verstehen sind. Ziel war es, den personenseitigen Aspekt von Itempositonseffekten genauer zu untersuchen. Konkret liefert das vorliegende Projekt Beiträge zur Erfassung von Itempositionseffekten, zu den Korrelaten von Itempositionseffekten, deren Auswirkungen auf die anhand herkömmlicher Testwerte vollzogenen Inferenzen und zur Rolle von Itempositionseffekten in Längsschnittstudien. Darüber hinaus umfasst das vorliegende Projekt eine weitere Teilstudie (Rose, Nagy, Nagengast, Frey & Becker, 2015), die sich den Interaktionen zwischen Positionseffekten und Auswirkungen der Abfolge von Testitems zu unterschiedlichen Inhaltsdomänen widmet (Brennan, 1992; Harris, 1991).

Die in diesem Beitrag zusammengefassten Studien liefern zusätzliche Evidenz für die Existenz individueller und schulischer Unterschiede in Positionseffekten. Die von uns durchgeführten Studien erweitern den bisherigen Kenntnisstand zu Positionseffekten in wichtiger Weise, da sie dokumentieren, dass diese systematisch mit anderen Personen- und Schulmerkmalen und Merkmalen der Testsituation assoziiert sind. Insgesamt unterstreichen unsere Befunde die Einschätzung, dass Itempositionseffekte als Indikatoren der Testbearbeitungspersistenz betrachtet werden können.

Die Erfassung der Korrelate der Testbearbeitungspersistenz ist nicht nur aus einer grundlagenwissenschaftlichen Perspektive interessant. Für anwendungsorientiert arbeitende Wissenschaftlerinnen und Wissenschaftler ist besonders wichtig, dass individuelle, gruppenspezifische und zeitliche Unterschiede in der Testbearbeitungspersistenz eine Gefährdung der Validität vieler Schlussfolgerungen darstellen. Auch wenn der Einfluss von Positionseffekten häufiger – aber eben nicht immer – vom Betrag her klein ausfällt, sind wir der Meinung, dass von Itempositionseffekten erhebliche Probleme bezüglich der Validität von Testwertinterpretationen bei Large-

Scale-Assessments ausgehen. Dies gilt insbesondere für Studien, die auf Testheftdesigns mit balancierten Itempositionen setzen, wie z. B. PISA. Unsere Ergebnisse deuten darauf hin, dass die Situation in Studien mit einem domänensequenziellen Design wahrscheinlich gravierender ausfällt (vgl. Tabelle 2). Nichtsdestotrotz plädieren wir dafür, dass die Auswirkungen von Positionseffekten auf die Ergebnisse von Leistungsvergleichen – wann immer es möglich ist – untersucht werden sollten.

7 Ausblick

Itempositionseffekte sind in allen aktuellen Arbeiten als graduelle Abnahmen der Lösungswahrscheinlichkeit von Items konzipiert (z. B. Debeer & Janssen, 2013). Statistisch gesehen werden Itempositionseffekte durch die Rate der Veränderung der Itemschwierigkeiten indiziert. Im vorliegenden Projekt sind wir diesem Ansatz gefolgt. Nichtsdestotrotz erscheinen neben der Rate der Veränderung auch andere Indikatoren der Testbearbeitungspersistenz sinnvoll. So könnte diese auch über die Position in einem Test, ab der eine Reduktion der Lösungswahrscheinlichkeit festzustellen ist, indiziert werden (z. B. Yamamoto & Gitomer, 1993). Der Vergleich unterschiedlicher Indikatoren der Testbearbeitungspersistenz ist ein wichtiges Thema, da andere Aspekte der Änderung des Lösungsverhaltens die vorliegenden Leistungsdaten unter Umständen besser als die Rate der Veränderung beschreiben und darüber hinaus höher mit individuellen Kovariaten korreliert sein könnten.

Ein weiterer Aspekt, der eine verstärkte Beachtung verdient, richtet sich auf die Gestaltung optimaler Testheftdesigns, die den Einfluss von Positionseffekten auf die Ergebnisse der Leistungsmessung minimieren und/oder die Identifikation und statistische Kontrolle von Itempositionseffekten optimieren (vgl. Hecht, Weirich, Siegle & Frey, 2015). Da eine nahezu vollständige Ausschaltung von Positionseffekten nur mithilfe vergleichsweise kurzer Tests zu bewerkstelligen ist und darüber hinaus nicht mit einer sequenziellen Darbietung von Tests zu unterschiedlichen Leistungsdomänen zu vereinen ist, ist eine designbedingte (approximativ) vollständige Kontrolle von Positionseffekten nicht möglich. Ein alternativer Ansatz könnte im Versuch bestehen, eine Balance zwischen von per Design relativ unbeeinflussten Testteilen und Testteilen, die eine statistische Kontrolle von Positionseffekten ermöglichen, zu erreichen.

Schließlich verdient die Betrachtung von Itempositionseffekten in Längsschnittsettings eine intensivere Beachtung. In den vergangenen Jahren ist auch im Bereich der Schulleistungsforschung eine verstärkte Zuwendung hin zu Längsschnittstudien festzustellen. An derartige Studien wird die Hoffnung geknüpft, dass sie einen detaillierten Einblick in Lernprozesse erlauben. Wie wir zeigen konnten, erscheinen längsschnittliche Schulleistungsstudien als anfällig für Positionseffekte. Die bisherigen Erfahrungen beschränkten sich aber auf einfache Retestdesigns mit zwei Erhebungspunkten. Zukünftige Forschung sollte die längerfristige Entwicklung der Testbearbeitungspersistenz untersuchen.

Literaturverzeichnis

Bloom, H. S., Hill, C. J., Black, A. R. & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research in Educational Effectiveness, 1*, 289–328.

Brennan, R. L. (1992). The Context of Context Effects. *Applied Measurement in Education, 5*, 225–264.

von Davier, M., Xu, X. & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*, 318–336.

Debeer, D., Buchholz, J., Hartig, J. & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*, 502–523.

Debeer, D. & Janssen, R. (2013). Modeling item position effects within an IRT framework. *Journal of Educational Measurement, 50*, 164–185.

Frey, A., Carstensen, C. H., Walter, O., Rönnebeck, S. & Gomolka, J. (2008). Methodische Grundlagen des Ländervergleichs. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006 in Deutschland: Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (S. 375–397). Münster: Waxmann.

Frey, A., Hartig, J. & Rupp, A. (2009). Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28,* 39–53.

Harris, D. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement, 15*, 247–256.

Hecht, M., Weirich, S., Siegle, T. & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*. Advance online publication.

Johnson, E. G. (1990). *National assessment of educational progress: Design of the 1992 assessment*. Princeton, NJ: Educational Testing Service.

Leary, L. F. & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387–413.

Meyers, J. L., Miller, G. E. & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*, 38–60.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161.

Nagengast, B., Nagy, G., Rose, N. & Becker, M. (2015). *Positionseffekte in den Leistungstests der nationalen Erweiterung der PISA 2006 Studie: Eine Mehrebenenstudie zu individuellen und kontextuellen Prädiktoren von Positionseffekten.* Vortrag auf der 12. Tagung der Fachgruppe Methoden & Evaluation der Deutschen Gesellschaft für Psychologie, Jena.

Nagy, G., Lüdtke, O., Köller, O., Heine, J.-H. & Mang. J. (2015). *IRT Skalierung der Leistungstests in der PISA-Längsschnittstudie 2012/2013: Konsequenzen von Positionseffekten auf die Abschätzung der Leistungsentwicklung.* Vortrag auf der 3. Tagung der Gesellschaft für Empirische Bildungsforschung, Bochum.

Nagy, G., Rose, N. & Nagengast, B. (2015). *Reteststabilität von Itempositionseffekten in einem Leseverständnistest: Mittelwertstabilität, und Entwicklung individueller Unterschiede von Klassenstufe 5 zu 6.* Vortrag auf der 12. Tagung der Fachgruppe Methoden & Evaluation der Deutschen Gesellschaft für Psychologie, Jena.

Nagy, G., Rose, N. & Trautwein, U. (2013). *Individuelle Unterschiede in der Testermüdung während der Bearbeitung eines Leseverständnistests: Eine Anwendung eines IRT-Modells zur Erfassung individueller Positionseffekte und ihrer Korrelate.* Vortrag auf der 1. Tagung der Gesellschaft für Empirische Bildungsforschung, Kiel.

Rose, N., Nagy, G., Nagengast, B., Frey, A. & Becker, M. (2015). *Multiple Itemkontexteffekte in Mehrdimensionalen IRT-Modellen: Modellierung von Effekten der Itemposition, der Blockposition und der Domänenabfolge.* Vortrag auf der 12. Tagung der Fachgruppe Methoden & Evaluation der Deutschen Gesellschaft für Psychologie, Jena.

Weirich, S., Hecht, M. & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*, 535–548.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*, 114–128.

Yamamoto, K. & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. In N. Fredriksen, R. Mislevy & I. Beijar (Hrsg.), *Test theory for a new generation of tests* (S. 275–296). Hillsdale, NJ: Erlbaum.

Ann-Katrin van den Ham, Timo Ehmke, Inga Hahn, Helene Wagner, Katrin Schöps

Mathematische und naturwissenschaftliche Kompetenz in PISA, im IQB-Ländervergleich und in der National Educational Panel Study (NEPS) – Vergleich der Rahmenkonzepte und der dimensionalen Struktur der Testinstrumente

1 Einleitung

Bildung gilt als eine wesentliche Voraussetzung für die aktive Partizipation als verantwortungsvolle(r) Bürgerin bzw. Bürger in einer demokratischen Gesellschaft. Bisher ist allerdings noch wenig über die kumulative Entwicklung von Kompetenzen über die Lebensspanne bekannt. Um Einsicht in den Bildungsprozess und die Kompetenzentwicklung zu erhalten, wurde in Deutschland die National Educational Panel Study (NEPS) ins Leben gerufen (Blossfeld, Maurice & Schneider, 2011). NEPS untersucht, wie sich Kompetenzen von Kindern, Jugendlichen und Erwachsenen über den Lebenslauf entwickeln, wie diese Kompetenzen die Bildungskarriere beeinflussen und inwiefern die Kompetenzentwicklung von Lerngelegenheiten mitbestimmt wird. Im Jahr 2010 wurde in NEPS erstmalig eine Kohorte von ca. 16.500 Neuntklässlerinnen und Neuntklässlern ausgewählt.

Im Rahmen des Bildungsmonitorings wird in Deutschland zusätzlich regelmäßig das Erreichen der nationalen Bildungsstandards am Ende der Sekundarstufe durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) geprüft (Pant, Stanat, Schroeders, Roppelt, Siegle & Pöhlmann, 2013). Außerdem werden die Kompetenzen von Schülerinnen und Schülern in Deutschland im internationalen Vergleich durch die regelmäßige Teilnahme am Programme for International Student Assessment (PISA: Organisation for Economic Cooperation and Development [OECD], 2013b; Prenzel, Sälzer, Klieme & Köller, 2013) erfasst.

Obwohl die Testkonzeptionen aus NEPS im Bereich der mathematischen und naturwissenschaftlichen Kompetenzmessungen am Ende der Sekundarstufe (NEPS-K9) große Überschneidungen mit dem IQB-Ländervergleich (LV 2012) und PISA 2012 aufweisen, sind die Ergebnisse der Studien aufgrund von unterschiedlichen Berichtsskalen nicht direkt vergleichbar. Eine Verlinkung der Berichtsskalen der Mathematik- und Naturwissenschaftsmessungen in NEPS-K9 mit den Mathematik- und Naturwissenschaftsmessungen im LV 2012 und in PISA 2012 würde jedoch eine Interpretation der Befunde aus NEPS in einem nationalen bzw. internationalen Referenzrahmen ermöglichen.

Eine Verlinkung von Berichtsskalen aus unterschiedlichen Large-Scale-Assessments erfordert nach Kolen und Brennan (2010) und van de Vijver (1998), dass sich die Studien hinsichtlich der (1) Schlussfolgerungen, (2) Population, (3) Messeigenschaften und -bedingungen sowie der (4) operationalisierten Konstrukte (konzeptionell, dimensional und skalenbezogen) hinreichend ähnlich sind (vgl. Ehmke et al., 2014). Es wird also vorausgesetzt, dass zwischen einem latenten, also nicht direkt beobachtbaren Konstrukt, den verwendeten Testmodellen und den eingesetzten Testitems gleiche Beziehungen bestehen. Dies wird erreicht, wenn neben der konzeptuellen Äquivalenz auch dimensionale Äquivalenz und Skalenäquivalenz gegeben sind.

Ziel des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projektes „PISA, Bildungsstandards und die National Educational Panel Study (NEPS). Vergleich der Rahmenkonzepte und Validierung der NEPS-Testinstrumente in den Naturwissenschaften und in der Mathematik“ war es daher, unter anderem die Naturwissenschafts- und Mathematiktests aus NEPS-K9 mit PISA 2012 und LV 2012 anhand der oben genannten Kriterien zu vergleichen und die Berichtsskalen gegebenenfalls zu verankern. In diesem Beitrag werden wesentliche Ergebnisse zur konzeptionellen und dimensionalen Äquivalenz der Tests vorgestellt.

2 Theoretischer Hintergrund – Überblick über die Studien

2.1 IQB-Ländervergleich (LV 2012)

Das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) wurde von der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) beauftragt, die Einhaltung der Bildungsstandards regelmäßig und flächendeckend zu überprüfen, die Bildungsstandards weiterzuentwickeln und Kompetenzstufenmodelle zu entwerfen. Am Ende der Sekundarstufe I findet daher alle drei Jahre, alternierend für die Bereiche Deutsch und erste Fremdsprache sowie Mathematik und Naturwissenschaften, der LV statt. Der LV bietet insbesondere die Möglichkeit, die Kompetenzniveaus von Schülerinnen und Schülern zwischen den Bundesländern zu vergleichen. In der Erhebung in 2012 wurden die Bereiche Mathematik und Naturwissenschaften erfasst. Insgesamt wurden ca. 45.000 Neuntklässlerinnen und Neuntklässler getestet (Pant et al., 2013).

Der Mathematiktest des LV 2012 basiert auf dem Rahmenkonzept der Bildungsstandards im Fach Mathematik am Ende der Sekundarstufe I (Roppelt, Blum & Pöhlmann, 2013) und unterscheidet drei Teilbereiche. So werden zum Lösen einer mathematischen Aufgabe Kenntnisse über mathematische Inhalte benötigt, die fünf mathematischen Inhaltsbereichen (Leitideen) zugeordnet werden können: Zahl, Messen, Raum und Form, Funktionaler Zusammenhang und Daten und Zufall. Außerdem müssen beim Lösen der Aufgaben bestimmte mathematische Prozesse angewendet werden, die sich durch sechs allgemeine mathematische Kompetenzen (Prozesse) beschreiben lassen: „Mathematisch argumentieren“, „Probleme mathematisch lösen“, „Mathematisch modellieren“, „Mathematische Darstellungen verwenden“,

„Formal-technisches Arbeiten“ und „Mathematisch kommunizieren“. Die mathematischen Aktivitäten werden auf unterschiedlichen kognitiven Anforderungsniveaus gefordert, die in drei Anforderungsbereichen beschrieben werden: „Reproduzieren“, „Zusammenhänge herstellen“ und „Verallgemeinern und Reflektieren“ (KMK, 2003; van den Ham et al., 2014).

Der Naturwissenschaftstest basiert auf dem Kompetenzmodell der Bildungsstandards in den Fächern Biologie, Physik und Chemie (Pant, Stanat, Schroeders et al., 2013). Die naturwissenschaftliche Kompetenz wird in vier Kompetenzbereiche unterteilt: „Umgang mit Fachwissen“ (fachspezifisch für Biologie, Chemie und Physik), „Erkenntnisgewinnung“, „Kommunikation“ und „Bewertung“. Weiterhin werden für jeden Kompetenzbereich fünf Komplexitätsstufen unterschieden: „ein Fakt“, „zwei Fakten“, „ein Zusammenhang“, „zwei Zusammenhänge“ und „übergeordnetes Konzept“. Eine weitere Dimension zur Untersuchung der naturwissenschaftlichen Kompetenz stellen die kognitiven Prozesse dar, die in „reproduzieren“, „selegieren“, „organisieren“ und „integrieren“ unterteilt werden (KMK, 2004a, 2004b, 2004c; Wagner, Schöps, Hahn, Pietsch & Köller, 2014).

2.2 Programme for International Student Assessment (PISA 2012)

PISA wird von der Organisation for Economic Co-operation and Development (OECD) geleitet. PISA erfasst international vergleichend Schülerleistungen in einem Zyklus von drei Jahren. Es werden die drei Domänen Lesen, Mathematik und Naturwissenschaften unterschieden, wobei an jedem Messzeitpunkt eine der Domänen alternierend schwerpunktmäßig erhoben wird. In der jeweiligen Hauptdomäne werden umfassendere Tests durchgeführt. Die Zielgruppe der PISA-Studie sind Jugendliche im Alter von 15 Jahren, wodurch die Ergebnisse Aussagen über grundlegende Kompetenzen von Schülerinnen und Schülern am Ende der Pflichtschulzeit ermöglichen. Zusätzlich werden Bezugspunkte für die Weiterentwicklung und eine Wissensbasis für die politische Steuerung des Bildungssystems gewonnen (OECD, 2013b; Prenzel et al., 2013).

In PISA 2012 wurden in Deutschland ca. 5.000 15-Jährige im Rahmen der Studie getestet (Prenzel et al., 2013). Die Hauptdomäne in 2012 war die mathematische Kompetenz der Schülerinnen und Schüler. Die zugrunde liegende mathematische Rahmenkonzeption basiert auf dem mathematischen Grundbildungskonzept (Mathematical Literacy). PISA definiert Mathematical Literacy wie folgt:

„Mathematical literacy is an individual’s capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain, and predict phenomena. It assists individuals to recognise the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens“ (OECD, 2013a, 25).

Diese Definition bringt zum Ausdruck, dass PISA auf eine enge Orientierung an Lehrplänen verzichtet und vor allem die funktionale Nutzung von Mathematik in den Vordergrund stellt. Mathematische Probleme sollen dementsprechend in unter-

schiedlichen Kontexten und Situationen gelöst werden (Neubrand et al., 2001). In der mathematischen Rahmenkonzeption werden die Dimensionen Inhalt, Prozess und Kontext beschrieben. Hierzu gehören vier mathematische Inhaltsbereiche: „Quantität", „Veränderung und Beziehungen", „Raum und Form" und „Unsicherheit und Daten", die mit drei mathematischen Prozessen verbunden sind, welche die grundlegenden mathematischen Fähigkeiten konkretisieren: „Situationen mathematisch formulieren", „mathematische Konzepte, Fakten, Prozeduren und Schlussfolgerungen anwenden" und „mathematische Ergebnisse interpretieren, anwenden und bewerten". Außerdem werden sieben fundamentale mathematische Fähigkeiten definiert, die zum Lösen einer Aufgabe benötigt werden (kommunizieren, mathematisieren, repräsentieren, argumentieren, Problemlösestrategien entwickeln, mit Mathematik symbolisch, formal und technisch umgehen sowie mathematische Hilfsmittel verwenden). Die Aufgaben werden zusätzlich in unterschiedlichen Kontexten verortet: persönliche, berufliche, gesellschaftliche und wissenschaftliche Kontexte (OECD, 2013a).

Die naturwissenschaftliche Rahmenkonzeption basiert auf dem naturwissenschaftlichen Grundbildungskonzept (scientific literacy). In PISA wird scientific literacy wie folgt definiert:

„For the purposes of PISA, scientific literacy refers to an individual's:

- Scientific knowledge and use of that knowledge to identify questions, acquire new knowledge, explain scientific phenomena and draw evidence-based conclusions about science-related issues.
- Understanding of the characteristic features of science as a form of human knowledge and enquiry.
- Awareness of how science and technology shape our material, intellectual and cultural environments.
- Willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen" (OECD, 2013a, 100).

In der Rahmenkonzeption werden für die Naturwissenschaften drei Teilkompetenzen unterschieden: „naturwissenschaftliche Fragestellungen erkennen", „naturwissenschaftliche Phänomene erklären" und „naturwissenschaftliche Evidenz nutzen". Die Basis für diese Teilkompetenzen bilden das (objektbezogene) naturwissenschaftliche Wissen und (Meta-)Wissen über die Naturwissenschaften. Der naturwissenschaftlichen Kompetenz werden vier Wissenssysteme untergeordnet: „Physikalische Systeme", „Lebende Systeme", „Erd- und Weltraumsysteme" und „Technologische Systeme". Dem Bereich „Wissen über die Naturwissenschaften" werden zwei Aspekte untergeordnet: „naturwissenschaftliches Forschen" und „naturwissenschaftliche Erklärungen". Die motivationalen Orientierungen und Einstellungen einer Person bilden eine weitere Komponente, auf der die Teilkompetenzen der naturwissenschaftlichen Grundbildung basieren. Die naturwissenschaftliche Kompetenz wird bei PISA situations- bzw. kontextgebunden untersucht. In der Rahmenkonzeption werden folgende fünf Kontexte differenziert: „Gesundheit", „natürliche Ressourcen", „Umwelt, Risiken/Gefahren" und zuletzt „Grenzen von Naturwissenschaften und Technik" (OECD, 2006; Prenzel et al., 2007; Wagner et al., 2014).

2.3 Das Nationale Bildungspanel (NEPS)

NEPS wird unter Federführung des Leibniz-Instituts für Bildungsverläufe (LIfBi) durchgeführt. Die Studie differenziert fünf inhaltliche Untersuchungsbereiche (Säulen) (Kompetenzentwicklung, Lernumwelten, Bildungsentscheidungen, Migrationshintergrund und Bildungsrenditen), die im Längsschnitt über acht Lebensetappen (von den Neugeborenen und der frühkindlichen Betreuung bis hin zur Erwachsenenbildung und dem lebenslangen Lernen) untersucht werden. Bezüglich der Kompetenzentwicklung werden die Fähigkeiten in den Bereichen Mathematik, Sprache (Orthografie, Lese- und Hörverstehen in deutscher und englischer Sprache, Kenntnisse in der Erstsprache bei Schülerinnen und Schülern mit Migrationshintergrund), Naturwissenschaften, ICT-Literacy und Problemlösen erfasst. Die Daten werden mithilfe eines Multi-Kohorten-Sequenz-Designs erhoben und der Wissenschaft in Form von scientific use files zu Forschungszwecken zur Verfügung gestellt (Blossfeld, Maurice & Schneider, 2011; Weinert, Artelt, Prenzel, Senkbeil, Ehmke & Carstensen, 2011). Im Herbst 2010 startete zum ersten Mal eine Kohorte von ca. 15.000 Schülerinnen und Schülern der neunten Jahrgangstufe (NEPS-K9).

In NEPS wird mathematische Kompetenz als das Ausmaß beschrieben, „in dem Schülerinnen und Schüler, aber auch Erwachsene, die in der Schule gelernte Mathematik in problemhaltigen, vorwiegend außermathematischen Situationen flexibel anwenden können“ (Ehmke et al., 2009, 317). Diese Definition basiert auf dem mathematischen Grundbildungskonzept, wie es im Rahmen von PISA definiert wurde (Weinert et al., 2011). Damit steht auch im NEPS-Mathematiktest die Bearbeitung kontextbezogener Probleme im Vordergrund. Die mathematische Rahmenkonzeption differenziert zwischen einer inhaltlichen und einer prozessbezogenen Komponente. Es werden vier mathematische Inhaltsbereiche definiert: „Quantität“, „Veränderung und Beziehungen“, „Raum und Form“ und „Daten und Zufall“. Die prozessorientierte Komponente umfasst mathematische und kognitive Denkanforderungen, die beim Lösen mathematischer Aufgaben erforderlich sein können. Zudem wird zwischen sechs kognitiven Komponenten mathematischer Denkprozesse differenziert (vgl. Niss, 2003; KMK, 2003): „mathematisch argumentieren“, „mathematisch kommunizieren“, „modellieren“, „mathematische Probleme lösen“, „repräsentieren“ (Darstellungen verwenden) und „technische Fertigkeiten einsetzen“ (Ehmke et al., 2009).

Das Rahmenkonzept der Naturwissenschaften basiert auf dem naturwissenschaftlichen Grundbildungskonzept, welches von der American Association for the Advancement of Science (AAAS, 1993, 2009) und PISA (OECD, 2006) definiert wurde. Dabei wird davon ausgegangen, dass grundlegendes Verständnis von naturwissenschaftlichen Konzepten und Prozessen für das alltägliche Leben relevant ist und die Basis für lebenslanges Lernen formt: „Rather than focusing on memorized knowledge, scientific literacy reflects the ability to apply one’s existing scientific knowledge in different everyday life contexts and situations“ (Hahn et al., 2013). Die naturwissenschaftliche Kompetenz wird ähnlich wie in PISA in die objektbezogene Komponente (naturwissenschaftliches Wissen) und prozessbezogene Komponente (Wissen über die Naturwissenschaften) unterteilt. Das „naturwissenschaftliche Wis-

sen“ umfasst in der NEPS-Rahmenkonzeption die inhaltsbezogenen Komponenten „Stoffe“, „Systeme“, „Entwicklung“ und „Wechselwirkungen“. Innerhalb des „Wissens über die Naturwissenschaften“ werden die prozessbezogenen Komponenten „Naturwissenschaftliche Denk- und Arbeitsweisen“ unterschieden. Die Erfassung der naturwissenschaftlichen Kompetenz erfolgt eingebettet in die ausgewählten Kontexte: „Gesundheit“, „Umwelt“ und „Technologie“ (Hahn et al., 2013).

2.4 Vergleich der Rahmenkonzeptionen

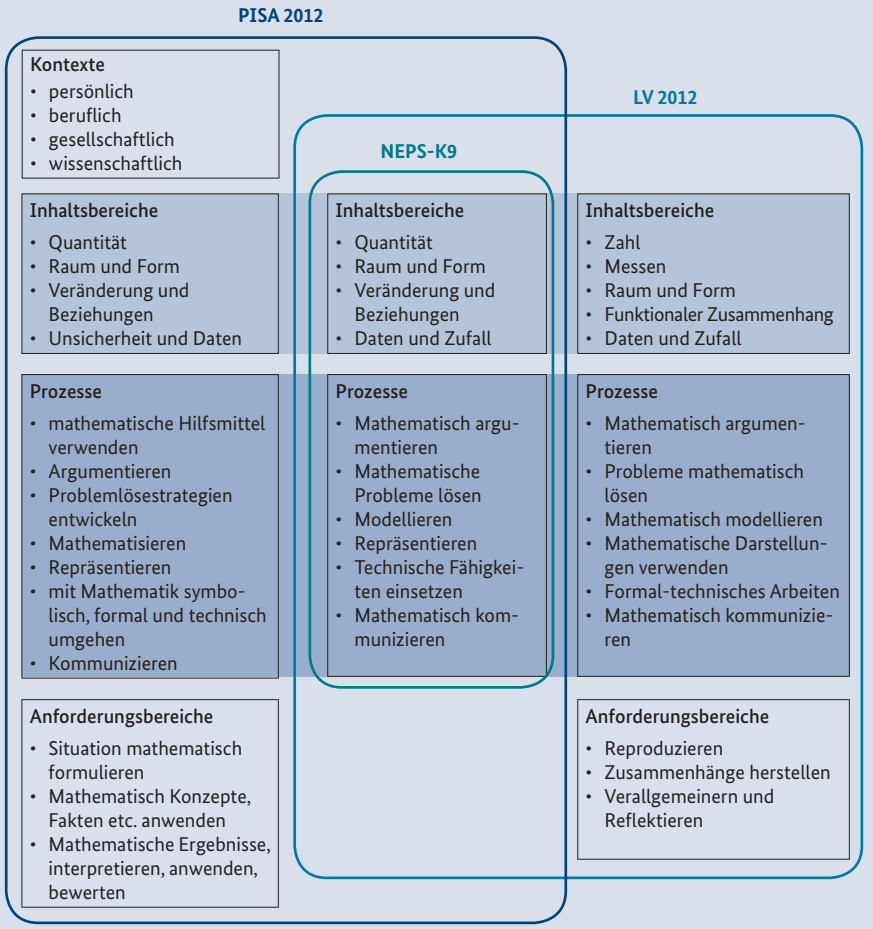
Alle drei Studien beanspruchen für sich, sowohl die mathematische als auch die naturwissenschaftliche Kompetenz von Schülerinnen und Schülern am Ende der Sekundarstufe I zu messen. Es lassen sich jedoch Unterschiede in den Zielstellungen und den damit verbundenen Schlussfolgerungen zwischen den Studien NEPS-K9, PISA 2012 und LV 2012 finden. So überprüft der LV 2012, inwieweit die in den Bildungsstandards formulierten Lernziele für Mathematik und die Naturwissenschaften am Ende der Sekundarstufe I erreicht wurden, und ermöglicht einen Vergleich der Kompetenzniveaus der Schülerinnen und Schüler zwischen den Bundesländern. PISA hingegen schafft eine Verortung der Kompetenzen von Schülerinnen und Schülern aus Deutschland an einem internationalen Referenzmaßstab (im Vergleich zwischen OECD-Staaten). Durch die zyklischen Durchführungen ermöglichen beide Studien außerdem Aussagen zu Trendentwicklungen (LV erst ab 2018). Im Unterschied dazu erlaubt NEPS bislang keine Verortung an Referenzmaßstäben. Das Ziel dieser Studie ist eine längsschnittliche Analyse der Entwicklung von Kompetenzen der Stichprobe von Schülerinnen und Schülern. Während anhand von NEPS Aussagen zu den längsschnittlichen Entwicklungen getroffen werden sollen, lassen die PISA-Ergebnisse Aussagen auf Ebene der Staaten und die LV-Ergebnisse Aussagen auf Ebene der Länder in Deutschland zu. Unterschiede in den Stichproben zwischen den Studien kommen dadurch zustande, dass diese im LV 2012 und in NEPS klassenbasiert sind, in PISA 2012 dagegen altersbasiert gezogen werden (vgl. van den Ham et al., 2014). Wobei die Mehrheit der Neuntklässlerinnen und Neuntklässler etwa 15 Jahre alt ist und es dadurch zwischen den Zielpopulationen einen hohen Überlappungsbereich gibt (Sälzer, Prenzel & Klieme, 2013).

Sowohl die mathematischen als auch die naturwissenschaftlichen Rahmenkonzepte aus PISA 2012 und NEPS-K9 basieren auf einem Literacy-Ansatz, der explizit nicht curriculumorientiert ist. Die Definition von mathematischer und naturwissenschaftlicher Kompetenz im LV 2012 basiert hingegen auf den KMK-Bildungsstandards. Diese geben vor, welche Kompetenzen etwa am Ende der Sekundarstufe I entwickelt sein sollten, und sind damit Richtschnur für die Kerncurricula der Länder (Pant et al., 2013; Köller, 2008).

Das Konstrukt der mathematischen Kompetenz wird sowohl in PISA 2012 als auch im LV 2012 und in NEPS-K9 durch die Definition von Teilbereichen weiter ausdifferenziert (vgl. Abbildung 1). Die NEPS-Rahmenkonzeption Mathematik orientiert sich bei der Gliederung der Inhaltsdimension vor allem an dem Rahmenkonzept von PISA. Der LV definiert einen Inhaltsbereich („Messen“) mehr als NEPS und

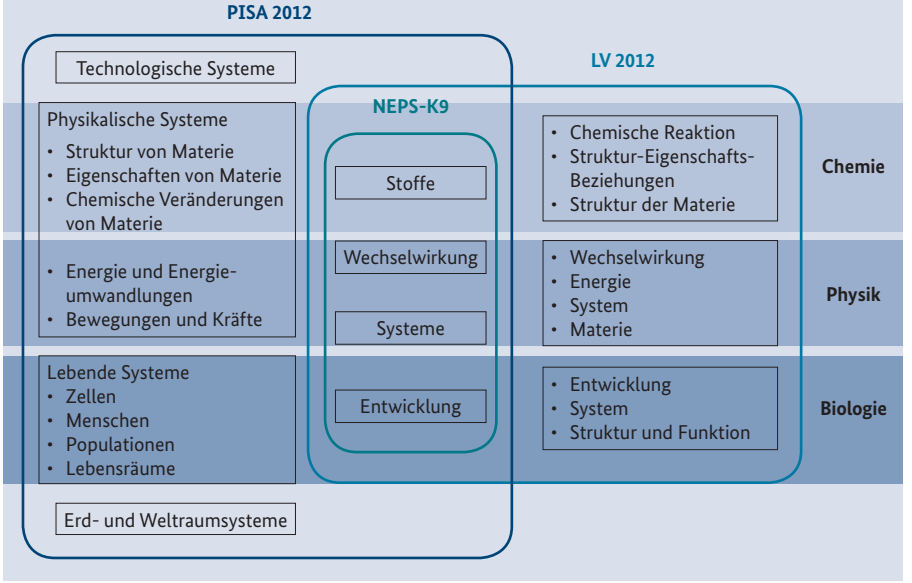
PISA. Die Inhalte des Bereichs „Messen“ fallen in NEPS unter den Bereich „Zahl“. Bei der Prozessdimension greift das NEPS-Rahmenkonzept die Bereiche aus dem LV auf. Sowohl die PISA-Rahmenkonzeption als auch die LV-Rahmenkonzeption definieren außerdem sogenannte Anforderungsniveaus. Diese beschreiben die kognitiven Anforderungen einer Aufgabe. PISA unterscheidet zudem unterschiedliche Aufgabenkontexte. Diese werden weder in der Rahmenkonzeption des LV noch in der Rahmenkonzeption der NEPS-Studie bestimmt.

Abb. 1: Überschneidung der mathematischen Rahmenkonzeption aus NEPS mit den Rahmenkonzeptionen aus PISA und dem LV



Die NEPS-Rahmenkonzeption für den Bereich naturwissenschaftlicher Kompetenz wurde in Anlehnung an die PISA-Rahmenkonzeption und die Konzeption der Bildungsstandards entwickelt (Hahn et al., 2013). Abbildung 2 zeigt die inhaltliche Überschneidung der Rahmenkonzeption aus NEPS mit der Rahmenkonzeption aus PISA im Bereich des naturwissenschaftlichen Wissens und mit der Rahmenkonzeption aus dem LV im Kompetenzbereich Fachwissen. Demnach kann das NEPS-Konzept „Stoffe" in den Umgang mit Fachwissen „Chemie" (Bildungsstandards) und „Physikalische Systeme" (PISA) eingeordnet werden. Ein Äquivalent für das Konzept „Wechselwirkungen" aus NEPS findet sich in der Rahmenkonzeption der Bildungsstandards im Kompetenzbereich Umgang mit Fachwissen „Physik" und in der Rahmenkonzeption aus PISA im Wissenssystem „Physikalische Systeme". In der NEPS-Rahmenkonzeption umfasst das Konzept „Systeme" biologische und technologische Systeme. Aus diesem Grund steht dieses Konzept an der Schnittstelle zwischen dem Umgang mit Fachwissen „Biologie" und „Physik" in den Bildungsstandards und den „Physikalischen und lebenden Systemen" bei PISA. Schließlich kann auch eine Überschneidung des NEPS-Konzepts Entwicklung mit dem Kompetenzbereich Umgang mit Fachwissen „Biologie" (Bildungsstandards) einerseits und dem Wissenssystem „Lebende Systeme" (PISA) andererseits postuliert werden (vgl. Wagner et al., 2014).

Abb. 2: Überschneidung der naturwissenschaftlichen Rahmenkonzeption aus NEPS mit den Rahmenkonzeptionen aus PISA und aus dem LV



Fazit: Die mathematischen und naturwissenschaftlichen Rahmenkonzeptionen aus den Studien PISA 2012, NEPS-K9 und LV 2012 weisen große inhaltliche Überschneidungen auf. So sind die Definition der mathematischen Inhaltsbereiche und kognitiven Prozesse sehr ähnlich, und es gibt deutliche Überschneidungen der NEPS-Rahmenkonzeption mit den Beschreibungen des Bereiches naturwissenschaftliches

Wissen in PISA 2012 und mit der Rahmenkonzeption aus dem LV 2012 im Kompetenzbereich Fachwissen. Daraus allein lässt sich jedoch nicht unmittelbar schlussfolgern, dass die drei Studien dieselben mathematischen bzw. naturwissenschaftlichen Kompetenzen messen.

3 Fragestellungen

In diesem Beitrag soll daher untersucht werden, ob die konzeptionellen und dimensionalen Voraussetzungen für ein Linking der Berichtsskalen aus den drei Studien in beiden Domänen gegeben sind. Folgende Forschungsfragen lassen sich ableiten:

(1) Inwieweit sind die mathematischen und naturwissenschaftlichen Testkonzeptionen in PISA 2012, im LV 2012 und in NEPS-K9 konzeptionell vergleichbar?
(2) Inwieweit ist die dimensionale Struktur für die latenten Konstrukte „mathematische Kompetenz“ und „naturwissenschaftliche Kompetenz“ in PISA 2012, im LV 2012 und in NEPS-K9 ähnlich?

4 Methode

4.1 Vergleichbarkeit der mathematischen und naturwissenschaftlichen Testkonzeptionen

4.1.1 Expertenreview der Testaufgaben in Mathematik

Für die Beantwortung der ersten Forschungsfrage wurden die 22 Mathematikaufgaben aus dem NEPS-Mathematiktest anhand eines Expertenreviews klassifiziert.[1] Drei in der Aufgabenkonstruktion erfahrene Expertinnen und Experten aus der Mathematikdidaktik wurden gebeten, alle 22 Items des NEPS-K9-Mathematiktests in die Rahmenkonzeptionen des LV-Mathematiktests 2012 und des PISA-Mathematiktests 2012 einzuordnen. Für diese Einordnung wurden Fragebögen erstellt, welche die Zuordnung der NEPS-Items zu den jeweiligen Dimensionen (Inhaltsbereiche, Prozesse, Anforderungsbereiche und Kontexte) vom LV 2012 und von PISA 2012 erfassen. Des Weiteren wurden umfangreiche Informationen zu den Hintergründen und Rahmenkonzeptionen der Studien zur Verfügung gestellt. Um zu überprüfen, inwieweit die Expertinnen und Experten in ihren Klassifikationen übereinstimmen, wurde die prozentuale Übereinstimmung zwischen den Ratern berechnet. Der Median der prozentualen Übereinstimmung zwischen den drei Ratern beträgt über alle Items 80 Prozent und ist damit als gut zu bewerten. Zusätzlich wurde das Cohens κ (Median der κ-Werte der drei Raterpaare) als zufallskorrigiertes Übereinstimmungs-

1 Befunde in diesem Abschnitt und in Abschnitt 5.1.1 sind publiziert in van den Ham et al. (2014).

maß berechnet. Die κ-Werte liegen im Mittel bei $\kappa = 0.5$ und sind damit ebenfalls in einem akzeptablen Bereich (Wirtz & Caspar, 2002). Anschließend wurden die 22 NEPS-Aufgaben in die Teilbereiche der PISA- und LV-Rahmenkonzeptionen eingeordnet. Um dies zu ermöglichen, wurde zuerst für jedes NEPS-K9-Mathematikitem von drei Expertinnen und Experten eingeschätzt, in welchen Inhaltsbereich, Prozess, Anforderungsbereich und Kontext das Item in der PISA-Rahmenkonzeption und in welchen Inhaltsbereich, Prozess und Anforderungsbereich das Item in der LV-Rahmenkonzeption fallen würde. Die Items wurden schließlich dem Teilbereich der PISA-2012- bzw. LV-2012-Rahmenkonzeption zugeordnet, welcher von mindestens zwei Expertinnen und Experten für dieses Item eingeschätzt wurde. Auf diese Weise konnten alle Items den Teilbereichen zugeordnet werden. Die so entstehenden Verteilungen der NEPS-Aufgaben in den PISA- und LV-Rahmenkonzeptionen wurden den Verteilungen der NEPS-Aufgaben in den NEPS-Inhaltsbereichen gegenübergestellt. Aufgrund der nicht vorliegenden Zuordnung der NEPS-Aufgaben zu den NEPS-Prozessen konnte für diese kein Vergleich zwischen den Studien durchgeführt werden.

4.1.2 Expertenreview der Testaufgaben in den Naturwissenschaften

Im Bereich Naturwissenschaften wurden die 28 Items des NEPS-K9-Naturwissenschaftstests von sieben Expertinnen und Experten in die PISA- und LV-Rahmenkonzeptionen eingeordnet.[2] Da das Ziel der Studie darin bestand, NEPS-Items anhand ihrer naturwissenschaftlichen Inhalte einzuschätzen, war es wichtig, Personen mit einem entsprechenden fachlichen und methodischen Hintergrund zu finden. Dementsprechend wurden fünf Expertinnen und Experten mit einem fachdidaktischen Hintergrund, eine Person mit einem erziehungswissenschaftlichen und psychologischen Hintergrund und eine Person mit einem Lehramtshintergrund ausgewählt. Weiterhin war es wichtig, dass die ausgewählten Personen über umfangreiche Kenntnisse hinsichtlich mindestens einer der Rahmenkonzeptionen und eines Kompetenztests verfügten. So waren vier der Expertinnen und Experten besonders mit PISA vertraut und drei von ihnen mit dem LV.

Zur Einordnung der NEPS-Items erhielten die Expertinnen und Experten alle nötigen Hintergrundinformationen zu den Studien sowie ein Reviewsheet, anhand dessen sie die Zuordnung der NEPS-Items zu den jeweiligen Wissens- und Kompetenzbereichen der anderen Studien vornehmen konnten (Wagner et al., 2014). Um zu überprüfen, inwieweit die Expertinnen und Experten in ihren Klassifikationen übereinstimmen, wurde die sogenannte Generalisierbarkeitstheorie angewandt (Cronbach, Gleser, Nanda & Rajaratnam, 1972). Die Auswertung wurde mit dem Programm G-String IV (Bloch & Norman, 2011) sowie mit dem Programm IASGA (Li & Lautenschlager, 1999) vorgenommen. Der berechnete Generalisierbarkeitskoeffizient darf als Cohens κ interpretiert werden. Hierbei werden Werte ≥ 0.6 als substanzielle Übereinstimmung zwischen den Ratern beurteilt (Landis & Koch, 1972). Bis auf

2 Befunde in diesem Abschnitt und in Abschnitt 5.1.2 sind publiziert in Wagner et al. (2014).

einen Bereich (Bildungsstandards, Bereich Komplexität), für den die Konsistenz bei gerade 0.6 liegt, zeigen alle Bereiche Konsistenzwerte zwischen 0.75 und 0.92, sodass die Expertenurteile in ihrer Übereinstimmung als substanziell bezeichnet werden können. Dieses Ergebnis stellt die Grundlage für die Beurteilung und Interpretation der weiteren Ergebnisse der Studie dar.

4.2 Dimensionale Zusammenhänge

Die Basis für die Analyse der dimensionalen Zusammenhänge bildeten Daten der Validierungsstudie, welche im Rahmen des Projektes im Frühjahr 2012 durchgeführt wurde.

Die Stichprobe bestand aus N = 1.965 Neuntklässlerinnen und Neuntklässlern aus 80 Schulen. Am ersten Testtag bearbeiteten N = 1.679 Schülerinnen und Schüler Mathematik- und Naturwissenschaftsaufgaben aus PISA. Am zweiten Testtag beantworteten N = 825 Testpersonen Mathematikaufgaben und N = 862 Naturwissenschaftsaufgaben aus dem Ländervergleich. Anschließend wurde der NEPS-Mathematiktest von N = 1.330 und der NEPS-Naturwissenschaftstest von N = 1.335 Schülerinnen und Schülern bearbeitet. Insgesamt nahmen N = 634 Schülerinnen und Schüler an allen drei Mathematiktests und N = 673 Schülerinnen und Schüler an allen drei Naturwissenschaftstests teil.

Die Zusammenhänge der mit den verschiedenen Tests erfassten mathematischen und naturwissenschaftlichen Kompetenzen wurden als latente Korrelationen jeweils in dreidimensionalen Rasch-Modellen mit der Software ConQuest (Wu, Adams & Haldane, 2007) geschätzt. In einem ersten Modell wurden die mathematische Kompetenz aus NEPS, die mathematische Kompetenz aus PISA und die mathematische Kompetenz aus dem LV auf jeweils einer separaten Dimension modelliert. In einem zweiten Modell wurden die naturwissenschaftlichen Kompetenzen aus NEPS, PISA und dem LV auf jeweils eigenen Dimensionen spezifiziert.

5 Ergebnisse

5.1 Ergebnisse zur konzeptionellen Äquivalenz der Mathematiktests aus NEPS-K9, PISA 2012 und dem LV 2012

In einem ersten Schritt wurde die Vergleichbarkeit der Kompetenzmessungen in Mathematik und in den Naturwissenschaften in NEPS-K9, PISA 2012 und dem LV 2012 auf der konzeptionellen Ebene untersucht. Dazu wurden die Testkonzeptionen analysiert: a) hinsichtlich der Einordbarkeit der Aufgaben aus dem Mathematik- und Naturwissenschaftstest aus NEPS-K9 in die entsprechenden Teildimensionen der Mathematik- und Naturwissenschaftstests aus PISA 2012 und LV 2012 sowie b) hinsichtlich der Unterschiede zwischen den Einordnungen der NEPS-K9-Aufgaben in die Rahmenkonzeptionen aus NEPS-K9 sowie aus PISA 2012 bzw. des LV 2012.

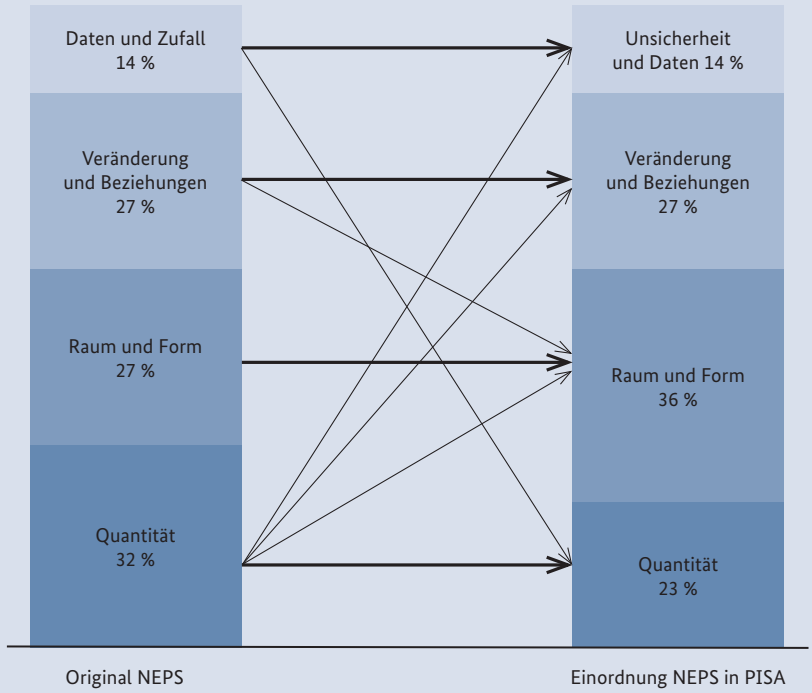
5.1.1 Vergleich der mathematischen Testkonzeptionen aus NEPS-K9 und PISA 2012

Das Expertenreview zeigt, dass die Expertinnen und Experten alle Aufgaben aus NEPS-K9 sowohl in einen Inhaltsbereich, in mindestens einen mathematischen Prozess- und Anforderungsbereich als auch in einen Aufgabenkontext aus der PISA-Rahmenkonzeption einordnen können. Es gibt also keine NEPS-Aufgaben, die einen Inhaltsbereich, Prozess oder Kontext erfassen, welcher nicht in der PISA-Rahmenkonzeption definiert wird.

Die prozentuale Verteilung der NEPS-K9-Aufgaben in den Inhaltsbereichen aus NEPS unterscheidet sich nicht signifikant von der durch die Expertinnen und Experten vorgenommenen Einordnung der NEPS-Aufgaben in die gleichnamigen PISA-Inhaltsbereichen ($\chi^2 = 1.8$, $df = 3$, *ns).* Tendenziell werden jedoch einige Unterschiede zwischen den Inhaltsbereichen deutlich (siehe Abbildung 3). Der größte Unterschied befindet sich zwischen den Inhaltsbereichen „Quantität" in NEPS und „Quantität" in PISA. Die Expertinnen und Experten ordnen drei der sieben NEPS-Aufgaben aus dem Bereich „Quantität" anderen PISA-Inhaltsbereichen zu. Ein weiterer Unterschied zeigt sich beim Inhaltsbereich „Raum und Form". Hier werden zwar alle „Raum und Form"-Aufgaben aus dem NEPS-K9-Test auch dem Bereich „Raum und Form" in PISA zugeordnet. Allerdings wird auch jeweils eine Aufgabe aus den NEPS-Inhaltsbereichen „Quantität" und „Veränderung und Beziehungen" dem PISA-Inhaltsbereich „Raum und Form" zugeordnet. Die Aufgaben aus den NEPS-Inhaltsbereichen „Veränderung und Beziehungen" und „Daten und Zufall" werden größtenteils den gleichnamigen Bereichen in PISA zugeordnet, nur jeweils eine Aufgabe wird einem anderen Inhaltsbereich zugesprochen (vgl. van den Ham et al., 2014).

In den NEPS-K9-Mathematikaufgaben lassen sich außerdem fünf der sechs PISA-Prozesse identifizieren. Keine der 22 NEPS-Mathematikaufgaben wird von den Expertinnen und Experten dem Prozess „mathematische Hilfsmittel verwenden" zugeordnet. Dieser Befund stimmt mit der NEPS-Rahmenkonzeption überein. In dieser wird explizit angegeben, dass mit dem NEPS-Test keine separate Kompetenz zum Umgang mit mathematischen Hilfsmitteln gemessen werden soll. Auf Basis des Expertenreviews werden zudem alle PISA-Anforderungsbereiche durch mindestens vier und alle Kontexte durch mindestens drei NEPS-Aufgaben abgedeckt.

Abb. 3: Prozentuale Verteilung der NEPS-K9-Aufgaben auf die mathematischen Inhaltsbereiche in NEPS sowie auf die mathematischen Inhaltsbereiche der PISA-Rahmenkonzeption



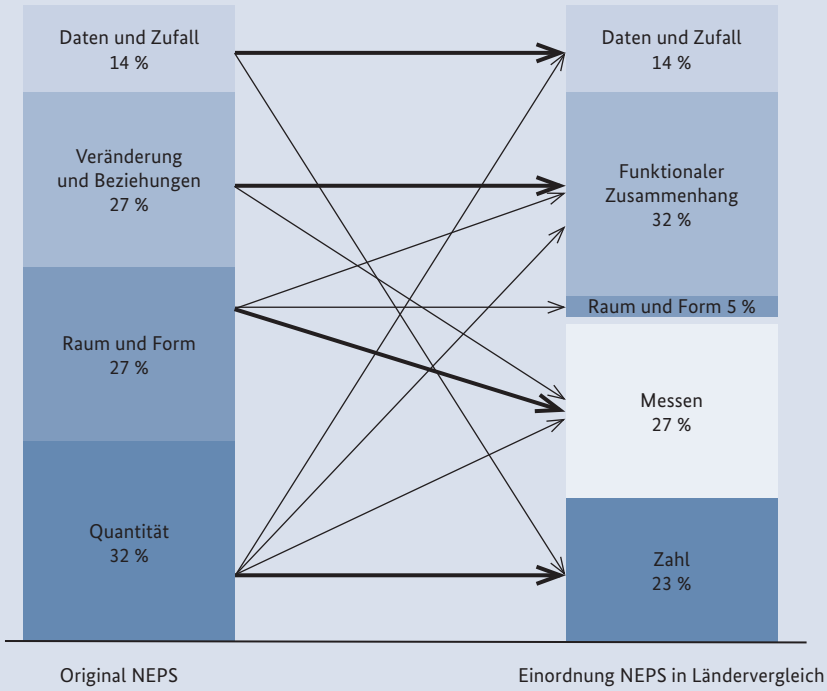
5.1.2 Vergleich der mathematischen Testkonzeptionen aus NEPS-K9 und LV 2012

Die Rahmenkonzeption des LV 2012 definiert einen Inhaltsbereich mehr als die NEPS-K9-Rahmenkonzeption. Die Inhalte dieses Bereiches „Messen" fallen in der NEPS-Rahmenkonzeption unter den Inhaltsbereich „Quantität". Bei der Einordnung der NEPS-Aufgaben in die LV-Rahmenkonzeption fällt jedoch auf, dass die meisten Aufgaben aus dem Bereich „Raum und Form" von den Expertinnen und Experten in den Bereich „Messen" der LV-Rahmenkonzeption eingeordnet werden. Bei diesen NEPS-Aufgaben müssen vor allem Flächen bzw. Winkel berechnet werden. Dementsprechend fallen im Vergleich weniger Aufgaben unter den Inhaltsbereich „Raum und Form" nach der Definition des LV. Dem Inhaltsbereich „Quantität" in NEPS werden zwei Aufgaben mehr zugeordnet als dem Inhaltsbereich „Zahl". Die Inhaltsbereiche „Veränderung und Beziehungen" aus NEPS und „Funktionaler Zusammenhang" aus dem LV sowie „Daten und Zufall" aus NEPS und „Daten und Zufall" aus dem LV unterscheiden sich kaum voneinander. Die Unterschiede in der Zuordnung lassen sich jedoch nicht statistisch absichern (vgl. van den Ham et al., 2014).

In allen NEPS-K9-Mathematikaufgaben konnten die Expertinnen und Experten Prozesse und Anforderungsbereiche der LV-Rahmenkonzeption identifizieren. Dabei werden alle Prozesse aus der LV-Rahmenkonzeption auf Basis des Expertenreviews

durch mindestens sechs Aufgaben abgedeckt. Die Anforderungsbereiche „Reproduzieren“ und „Zusammenhänge herstellen“ werden durch acht respektive 14 Aufgaben abgedeckt. Jedoch wurde keine der NEPS-Aufgaben dem Anforderungsbereich „Verallgemeinern und Reflektieren“ zugeordnet.

Abb. 4: Prozentuale Verteilung der NEPS-K9-Aufgaben auf die mathematischen Inhaltsbereiche in NEPS sowie auf die mathematischen Inhaltsbereich der LV-Rahmenkonzeption

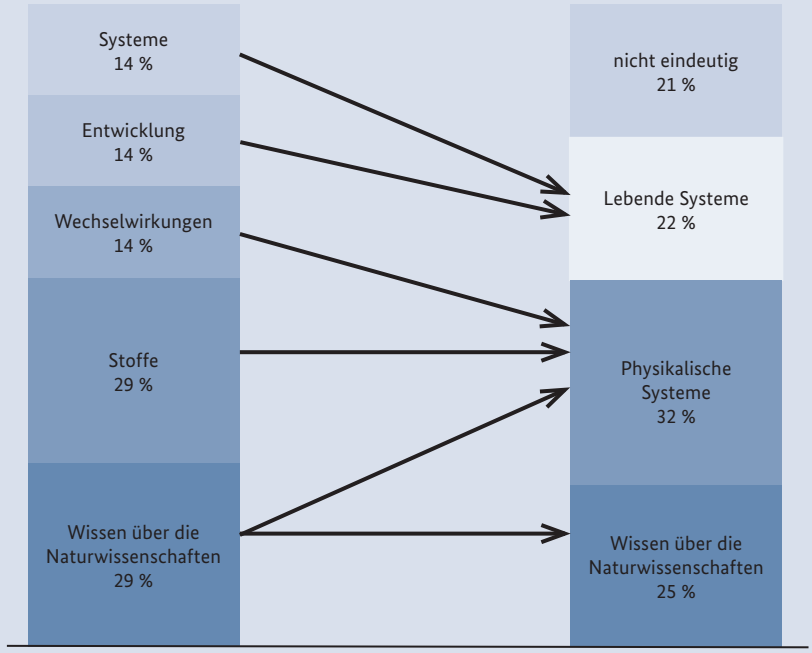


5.2 Vergleich der naturwissenschaftlichen Testkonzeptionen aus NEPS-K9 und PISA 2012

Das Expertenreview der Aufgaben des NEPS-K9-Naturwissenschaftstests zeigt, dass sich die NEPS-Items über die Wissenssysteme „lebende Systeme“, „physikalische Systeme“ und den Wissensbereich „Wissen über die Naturwissenschaften“ verteilen (siehe Abbildung 5). Die Zuordnungsrate der NEPS-Items zu den PISA-Wissensbereichen fällt mit 79 Prozent sehr hoch aus. Die in Abbildung 5 dargestellten Pfeile geben darüber Auskunft, wohin jeweils die Mehrzahl der Rater die meisten Items zugeordnet hat. Die Items der inhaltsbezogenen Komponenten „Systeme“ und „Entwicklung“ können vorwiegend in den PISA-Bereich „lebende Systeme“ eingeordnet werden. Weiterhin werden die Komponenten „Wechselwirkungen“ und „Stoffe“ beide vorwiegend dem PISA-Bereich der „physikalischen Systeme“ zugeordnet. Die Items der prozessbezogenen NEPS-Komponente „Wissen über die Naturwissenschaften“

fallen sowohl in die „physikalischen Systeme“ als auch in den inhaltlich äquivalenten Bereich des „Wissens über die Naturwissenschaften“ in PISA. Hinsichtlich der Zuordnung der NEPS-K9-Items auf die Teilkompetenzen „naturwissenschaftliche Fragestellungen erkennen“, „naturwissenschaftliche Phänomene beschreiben, erklären und vorhersagen“ sowie „naturwissenschaftliche Evidenz nutzen, um Entscheidungen zu treffen“ konnte eine Zuordnungsrate von 96 Prozent erreicht werden. Die meisten NEPS-Items (67 Prozent der eingeordneten Items) werden dabei der Kompetenz „naturwissenschaftliche Phänomene erklären“ zugeordnet. Für die Mehrheit der NEPS-Komponenten (vier von fünf) ist eine eindeutige Zuordnung zu den Inhalten der Rahmenkonzeptionen von PISA möglich. Dabei werden die NEPS-Konzepte jeweils meist ähnlich bezeichneten Kompetenz- und Wissensbereichen zugeordnet. Für den NEPS-Inhaltsbereich „Systeme“ fällt eine eindeutige Einordnung bezüglich der PISA-Rahmenkonzeption schwer.

Abb. 5: Prozentuale Verteilung der NEPS-K9-Aufgaben auf die naturwissenschaftlichen Inhaltsbereiche in NEPS sowie auf die naturwissenschaftlichen Wissenssysteme in PISA

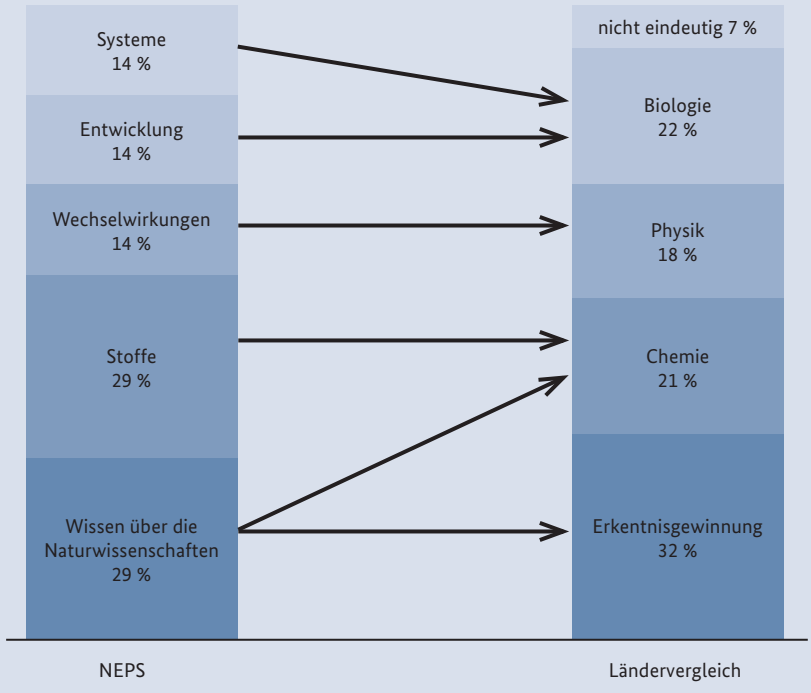


5.2.1 Vergleich der naturwissenschaftlichen Testkonzeptionen aus NEPS-K9 und dem LV 2012

Vier und mehr der sieben Rater, die die NEPS-K9-Items einordneten, konnten 93 Prozent der Items einem bestimmten Teilbereich des „Umgangs mit Fachwissen“ oder der „Erkenntnisgewinnung“ zuordnen.

Die Pfeile in Abbildung 6 geben erneut die Kompetenzbereiche wieder, in welche die Mehrzahl der Rater die meisten Items zugeordnet hat. Insgesamt verteilen sich die NEPS-Items relativ gleichmäßig auf die Kompetenzbereiche der Bildungsstandards. Lediglich der Fachbereich „Physik“ ist leicht unterrepräsentiert. Analog zur Zuordnung in die PISA-Rahmenkonzeption werden die NEPS-Items der inhaltsbezogenen Komponenten „Systeme“ und „Entwicklung“ dem Fachbereich „Biologie“ zugeordnet. Die Items der NEPS-Komponente „Wechselwirkungen“ können eindeutig dem Fachbereich Physik und die Items der NEPS-Komponente „Stoffe“ ebenso eindeutig dem Fachbereich „Chemie“ zugeordnet werden. Im Hinblick auf die kognitiven Prozesse werden 90 Prozent der zugeordneten NEPS-Items den Prozessen „Organisieren“ und „Integrieren“ zugeordnet. Die Zuordnungsrate zur Dimension „kognitive Prozesse“ liegt insgesamt bei 75 Prozent. Auf die Dimension „Komplexität“ können 64 Prozent der NEPS-Items zugeordnet werden. Mit 72 Prozent liegt der Fokus der Items auf der Stufe „ein Zusammenhang“. Auch bei der Einordnung der NEPS-Aufgaben in die Rahmenkonzeption des LV gilt, dass für vier der fünf NEPS-Komponenten eine eindeutige Zuordnung zumeist ähnlich bezeichneter Kompetenz- und Wissensbereiche möglich war. Eine eindeutige Zuordnung des NEPS-Inhaltsbereiches Systeme bezüglich der LV-Konzeption fällt jedoch schwer.

Abb. 6: Prozentuale Verteilung der NEPS-K9-Aufgaben auf die naturwissenschaftlichen Inhaltsbereiche in NEPS sowie auf die Kompetenzbereiche der Bildungsstandards

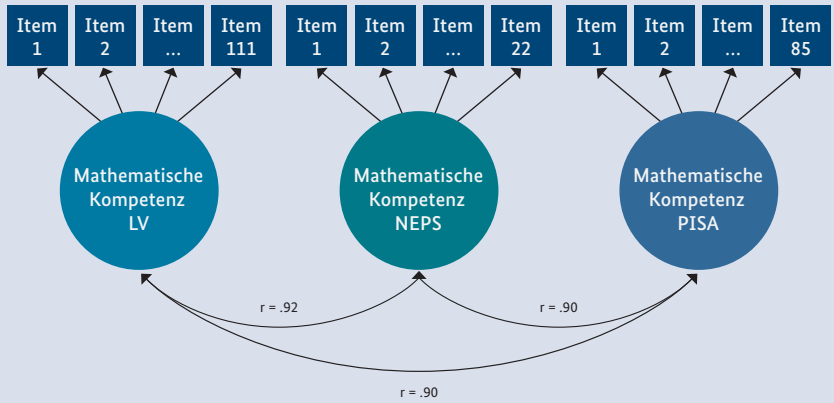


5.3 Ergebnisse zur dimensionalen Äquivalenz

5.3.1 Zusammenhänge zwischen den mathematischen Kompetenztests aus PISA 2012, LV 2012 und NEPS-K9

Die aus der dreidimensionalen Modellierung entstehenden Zusammenhänge zwischen den Mathematiktests aus NEPS, LV und PISA werden in Abbildung 7 dargestellt. Die Höhe der Korrelationen deutet auf einen substanziellen Zusammenhang zwischen den Tests. Die latenten Korrelationen zwischen den drei Mathematikskalen fallen in vergleichbarer Höhe aus.

Abb. 7: Latente Korrelationen zwischen den Mathematiktests aus NEPS, LV und PISA in einem dreidimensionalen Modell

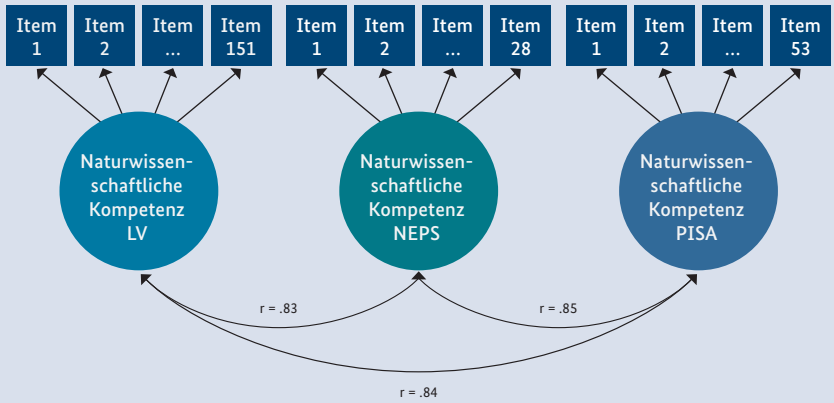


5.3.2 Zusammenhänge zwischen den Naturwissenschaftstests aus PISA 2012, LV 2012 und NEPS-K9

Um den korrelativen Zusammenhang der drei Naturwissenschaftstests aus dem Ländervergleich, aus PISA 2012 und aus NEPS-K9 zu bestimmen, wurde ebenfalls eine dreidimensionale Skalierung durchgeführt, in der jeder Kompetenztest als eigene Dimension angelegt wurde. Die Ergebnisse in Abbildung 8 zeigen substanzielle Korrelationen zwischen den drei Naturwissenschaftsskalen. Die Korrelation zwischen dem NEPS-Test und dem Ländervergleichstest ist in vergleichbarer Höhe wie die Korrelation zwischen dem NEPS-Test und dem PISA-Test.

Im Vergleich der Korrelationsmuster zwischen den Domänen Mathematik und Naturwissenschaften kann festgehalten werden, dass die Korrelationsmuster in Mathematik geringfügig höher ausfallen als in den Naturwissenschaften.

Abb. 8: Latente Korrelationen zwischen den drei Naturwissenschaftstests aus dem Ländervergleich, aus PISA und aus NEPS



6 Zusammenfassung und Diskussion

Dieser Beitrag hatte zum Ziel, die konzeptionelle und dimensionale Äquivalenz der in NEPS-K9, im LV 2012 und in PISA 2012 erfassten mathematischen und naturwissenschaftlichen Kompetenzen zu untersuchen. Dafür wurde geprüft, (1) inwieweit die mathematischen und naturwissenschaftlichen Testkonzeptionen in PISA 2012, in den länderübergreifenden Bildungsstandards und in NEPS konzeptionell vergleichbar sind und (2) inwieweit die dimensionale Struktur für die latenten Konstrukte „mathematische Kompetenz“ und „naturwissenschaftliche Kompetenz“ der Tests in PISA 2012, in den länderübergreifenden Bildungsstandards und in NEPS vergleichbar ausfällt.

Hinsichtlich der konzeptionellen Äquivalenz kann festgehalten werden, dass es große Übereinstimmungen zwischen den Testkonstrukten gibt. Für den NEPS-K9-Mathematiktest gilt, dass alle Aufgaben in das PISA- und LV-Rahmenkonzept eingeordnet werden können. Im NEPS-K9-Mathematiktest werden keine Aufgaben eingesetzt, die nicht diesen Rahmenkonzepten entsprechen. Lediglich ein Prozess der PISA-Rahmenkonzeption und ein Anforderungsbereich der LV-Rahmenkonzeption werden nicht durch die NEPS-Aufgaben repräsentiert. Bezüglich des NEPS-Naturwissenschaftstests kann festgehalten werden, dass für die Mehrheit der NEPS-Komponenten eine eindeutige Zuordnung zu den Inhalten der PISA- und LV-Rahmenkonzeption möglich ist. Die Zuordnungsrate von mindestens 79 Prozent zu den Kompetenz- bzw. Wissensbereichen der LV- und PISA-Rahmenkonzeption deutet auf eine hohe Vergleichbarkeit der theoretischen Konstrukte.

Die korrelativen Zusammenhänge bestätigen sowohl für die mathematische als auch für die naturwissenschaftliche Kompetenz eine deutliche Überschneidung zwischen den Tests. Dabei lassen sich die größeren Übereinstimmungen zwischen den mathematischen Rahmenkonzeptionen in höheren Korrelationen zwischen den mathematischen Konstrukten ($.90 \leq r \leq .92$) als zwischen den naturwissenschaftlichen Konstrukten ($.83 \leq r \leq .85$) wiederfinden.

Dennoch weisen die gefundenen Unterschiede zwischen den Rahmenkonzepten und auch die Höhe der Korrelationen darauf hin, dass es sich nicht um austauschbare Konstrukte handelt. Nach Kolen und Brennan (2010) ist jedoch auch eine Verknüpfung von unterschiedlichen Tests möglich. Die gefundenen Übereinstimmungen dieser Studien bilden eine wichtige Voraussetzung für eine mögliche Verlinkung der Berichtsskalen dieser Studien. Die aufgedeckten Unterschiede weisen auf die Relevanz tiefer gehender Analysen hin, da diese zu Unterschieden in den Testergebnissen und somit in den Ergebnissen eines Linking führen können (Kolen & Brennan, 2010). Es ist daher von Belang, die faktorielle Struktur der Tests noch intensiver auf ihre Vergleichbarkeit zu prüfen, z. B. durch einen Vergleich der Korrelationen zwischen den Subdimensionen innerhalb der Tests und zwischen den Tests. Ein weiterer wichtiger Schritt wäre es, zudem die Gruppeninvarianzen und deren Einfluss auf mögliche Entschlüsse, basierend auf einem Linking (Kolen & Brennan, 2010), zu analysieren. Die gefundenen Unterschiede in der sprachlichen Schwierigkeit der Tests (van den Ham et al., 2014) weisen vor allem auf die Wichtigkeit einer Analyse möglicher Invarianzen bezüglich der Subgruppe von Schülerinnen und Schülern mit Deutsch als Zweitsprache. Nur bei hinreichender Äquivalenz auf dimensionaler und skalenbezogener Ebene ist eine Verankerung der Skalen sinnvoll zu interpretieren. Zusammenfassend kann geschlossen werden, dass eine Verlinkung der mathematischen und naturwissenschaftlichen Berichtsskalen möglich ist, sofern die gefundenen Differenzen sowie mögliche Linkingfehler berücksichtigt werden, eine hierfür angemessene Linking-Methode gewählt wird und adäquate Interpretationen getätigt werden.

Literaturverzeichnis

American Association for the Advancement of Science (1993). *Benchmarks for science literacy. Project 2061.* New York, NY: Oxford University Press.

American Association for the Advancement of Science (2009). *Benchmarks for science literacy. Project 2061.* New York, NY: Oxford University Press. Abgerufen am 16.07.2015 von http://www.project2061.org/publications/bsl/online/index.php.

Bloch, R. & Norman, G. (2011). *G String IV (Version 6.1.1). User Manual.* Abgerufen am 16.07.2015 von http://fhsperd.mcmaster.ca/g_string/download/g_string_4_manual_611.pdf.

Blossfeld, H.-P. (2008). *Education as a Lifelong Process. A Proposal for a National Educational Panel Study (NEPS) in Germany. Part B: Theories, Operationalizations and Piloting Strategies for the Proposed Measurements.* Unveröffentlichter BMBF-Antrag. Bamberg: Universität Bamberg.

Blossfeld, H.-P., Maurice, J. von & Schneider, T. (2011). The National Educational Panel Study: need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14,* 5–18.

Cronbach, L., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurement. Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Ehmke, T., Duchhardt, C., Geiser, H., Grüßing, M., Heinze, A. & Marschick, F. (2009). Kompetenzentwicklung über die Lebensspanne – Erhebung von mathematischer Kompetenz im Nationalen Bildungspanel. In A. Heinze & M. Grüßing (Hrsg.), *Mathematiklernen vom Kindergarten bis zum Studium. Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht* (S. 313–327). Münster: Waxmann.

Ehmke, T., Köller, O., Nissen, A. & van den Ham, A.-K. (2014). Äquivalenz von Kompetenzmessungen in Schulleistungsstudien. *Unterrichtswissenschaft, 42 (4),* 290–300.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., Delahefte, J. M. & Prenzel, M. (2013). Assessing scientific literacy over the lifespan – A description of the NEPS science framework and the test development. *Journal for Educational Research Online, 5 (2),* 110–138.

Kolen, M. J. & Brennan, R. L. (2010). *Test equating, scaling, and linking: Methods and practices.* New York, NY: Springer.

Köller, O. (2008). Bildungsstandards in Deutschland: Implikationen für die Qualitätssicherung und Unterrichtsqualität. In M. A. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* (S. 47–59). Wiesbaden: VS Verlag für Sozialwissenschaften.

Landis J. R. & Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Li, M. N. & Lautenschlager, G. J. (1999). IASGA: a SAS MACRO program for interrater agreement studies of qualitative data via generalizability approach. *Educational and Psychological Measurement, 59 (3),* 532–537.

Neubrand, M., Biehler, R., Blum, W., Cohors-Freseborg, E., Flade, L., Knoche, N., Lind, D., Löding, W., Möller, G. & Wynands, A. (2001). Grundlagen der Ergänzung des internationalen PISA-Mathematiktests in der deutschen Zusatzerhebung. *Zentralblatt für Didaktik der Mathematik – Berichtsteil, 33,* 45–59.

Niss, M. (2003). Mathematical competencies and the learning of mathematics: The Danish KOM project. In A. Gagatsis & S. Papastavridis (Hrsg.), *3rd Mediterranean Conference on Mathematical Education. Athens – Hellas 3–5 January 2003* (S. 115–124). Athen: The Hellenic Mathematical Society.

Organisation for Economic Co-operation and Development (2006). *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006.* Paris: OECD.

Organisation for Economic Co-operation and Development (2013a). *PISA 2006 Science Competencies for Tomorrow's World.* Paris: OECD.

Organisation for Economic Co-operation and Development (2013b). *What students know and can do: Student performance in mathematics, reading and science* (Bd. I, überarbeitete Auflage, Februar 2014). Paris: OECD.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (Hrsg.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster: Waxmann.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E. & Pekrun, R. (Hrsg.). (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie.* Münster: Waxmann.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (Hrsg.). (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland.* Münster: Waxmann.

Roppelt, A., Blum, W. & Pöhlmann, C. (2013). Beschreibung der untersuchten mathematischen Kompetenzen. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 23–37). Münster: Waxmann.

Sälzer, C., Prenzel, M. & Klieme, E. (2013). Schulische Rahmenbedingungen der Kompetenzentwicklung. In M. Prenzel, C. Sälzer, O. Köller & E. Klieme (Hrsg.), *PISA 2012. Fortschritte und Herausforderungen in Deutschland* (S. 155–187). Münster: Waxmann.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2003). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. Beschluss vom 4.12.2003.* Neuwied: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004a). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* Neuwied: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004b). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* Neuwied: Luchterhand.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004c). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss vom 16.12.2004.* Neuwied: Luchterhand.

van den Ham, A.-K., Nissen, A., Ehmke, T., Sälzer, C. & Roppelt, A. (2014). Mathematische Kompetenz in PISA, IQB-Ländervergleich und NEPS – Drei Studien, gleiches Konstrukt? *Unterrichtswissenschaft, 42 (4),* 321–341.

van de Vivjer, F. J. R. (1998). Towards a theory of bias and equivalence. In J. A. Harkness (Hrsg.), *ZUMA-Nachrichten Spezial* (Bd. 3, S. 41–65). Mannheim: ZUMA.

Wagner, H., Schöps, K., Hahn, I., Pietsch, M. & Köller, O. (2014). Konzeptionelle Äquivalenz von Kompetenzmessungen in den Naturwissenschaften zwischen NEPS, IQB-Ländervergleich und PISA. *Unterrichtswissenschaft, 42 (4),* 301–320.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T. & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14,* 67–86.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.

Wu, M., Adams, R. & Haldane, S. (2007). *ConQuest [computer software].* Melbourne: Australian Council for Educational Research.

S. Franziska C. Wenzel, Lena Engelhardt, Katja Hartig, Kathrin Kuchta, Andreas Frey, Frank Goldhammer, Johannes Naumann, Holger Horz

Computergestützte, adaptive und verhaltensnahe Erfassung informations- und kommunikationstechnologiebezogener Fertigkeiten (ICT-Skills) (CavE-ICT)

1 Einleitung

2011 startete das Bundesministerium für Bildung und Forschung (BMBF) die Initiative zur Förderung von Forschungsvorhaben in Ankopplung an Large-Scale-Assessments (LSA). Large-Scale-Assessments werden als wichtige Instrumente des Bildungsmonitorings angesehen (z. B. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK], 2006). Mit der Ausschreibung der Förderinitiative werden Forschungsaktivitäten unterstützt, die neben Fragestellungen zur Kompetenzdiagnostik Testkonzeption und Methoden ständig zu verbessern suchen, weitere Domänen, Kompetenzen und Bedienungsfelder erschließen und in angemessener Weise modellieren wollen.

Das Projekt „Computergestützte, adaptive und verhaltensnahe Erfassung informations- und kommunikationstechnologiebezogener Fertigkeiten (ICT-Skills)“ (CavE-ICT) war in der genannten Initiative angesiedelt. Es widmete sich der Entwicklung eines computergestützten Testverfahrens zur simulationsbasierten Erfassung von informations- und kommunikationstechnologiebezogenen Fertigkeiten. Im vorliegenden Beitrag werden die zentralen Ergebnisse der Testentwicklung dargelegt. Der Einsatz des entwickelten Testinstruments kann im Kontext von Large-Scale-Assessments, aber auch zur Individualdiagnostik erfolgen. Der Test ist zudem potenziell adaptiv einsetzbar und beruht auf einer soliden theoretischen Basis.

CavE-ICT war ein Kooperationsprojekt zwischen der Goethe-Universität Frankfurt am Main, dem Deutschen Institut für Internationale Pädagogische Forschung in Frankfurt am Main und der Friedrich-Schiller-Universität Jena. Die übergeordnete „Vision“ des Projekts bestand in der Entwicklung eines computerbasierten, verhaltensnahen sowie adaptiven Instruments zur Messung von ICT-Skills, das bei künftigen (internationalen) Large-Scale-Assesments, wie etwa PISA, dem nationalen Bildungspanel (NEPS; Artelt, Weinert & Carstensen, 2013) oder der International Computer and Information Literacy Study (ICILS; Bos et al., 2014), Verwendung finden kann.

Die Struktur des vorliegenden Beitrags orientiert sich am Projektablauf. Zunächst wird die theoretische Rahmenkonzeption vorgestellt (Kapitel 2), auf deren Basis der computerisierte Test entwickelt wurde. Anschließend wird der Prozess der Itementwicklung skizziert (Kapitel 3) und schließlich die Erprobung der entwickelten Items im Rahmen der Kalibrierungsstudie dargestellt (Kapitel 4). Abschließend wird ein kurzes Fazit gezogen und ein Ausblick gegeben (Kapitel 5).

2 Hintergrund

Computer sind aus unserem Alltagsleben nicht mehr wegzudenken, sie kommen tagtäglich und in fast allen Lebensbereichen zum Einsatz. Auch in der Schule werden Rechner genutzt und zunehmend Fertigkeiten im Umgang mit Computertechnologien vermittelt. In vielen nationalen und internationalen Large-Scale-Assessments hält das computerbasierte Testen zunehmend Einzug. So wurde z. B. bei PISA auf internationaler Ebene erstmals im Jahr 2006 ein computerbasierter Test zur Kompetenzerfassung im naturwissenschaftlichen Bereich eingesetzt (Organisation for Economic Co-operation and Development [OECD], 2010). 2009 kam eine größere Aufgabenbatterie zum Einsatz, um das Lesen digitaler Texte zu erfassen, und 2012 folgte ein computerbasierter Test zur Erfassung dynamischer Problemlösefähigkeit sowie zur Erfassung mathematischer Kompetenz. Für künftige PISA-Erhebungen ist ein vollständiger Wechsel zu computerbasierter Testung geplant, wobei ein adaptives Testen angestrebt wird. Im Programme for the International Assessment of Adult Competencies (PIACC; OECD, 2012) ist die vollständige computergestützte Administration bereits in der ersten Welle 2012 umgesetzt worden. Sowohl der Hintergrundfragebogen als auch die Kompetenzmessungen erfolgten computerbasiert. Auch in NEPS wird der Einsatz von computerbasierten Tests angestrebt.

Die künftige Testrealität bei Large-Scale-Assessments liegt somit in dynamischen, interaktiven und adaptiven elektronisch administrierten Erhebungen. Damit spielt letztlich auch die Erfassung von Fertigkeiten im Umgang mit Computertechnologien eine immer wichtigere Rolle.

Im Rahmen der ICILS-Studie, die zeitlich parallel zum CavE-ICT-Projekt verlief, wurden Schülerinnen und Schüler der achten Jahrgangsstufe hinsichtlich ihrer computer- und informationsbezogenen Fertigkeiten im internationalen Vergleich computerbasiert untersucht (Bos et al., 2014). Eingesetzt wurden einerseits interaktive Items, die beispielsweise das Klicken eines Links erforderten, zum anderen aber auch nicht interaktive Items, wie z. B. Multiple-Choice-Fragen. Im Projekt CavE-ICT steht die theoriegeleitete Entwicklung und empirische Erprobung eines computerbasierten, verhaltensnahen und interaktiven sowie potenziell adaptiven Tests zur Messung von ICT-Skills im Vordergrund.

Im Folgenden wird zunächst die Ausrichtung des Projekts CavE-ICT kurz dargestellt (Kapitel 2.1). Im Anschluss wird konkret auf die erarbeitete Rahmenkonzeption zur Beschreibung von ICT-Skills eingegangen (Kapitel 2.2), da diese den Ausgangspunkt der Testentwicklung darstellt.

2.1 Ausrichtung des Projekts CavE-ICT

Schwerpunkt des Projekts ist die Ausdifferenzierung und Erweiterung von Konstrukten und Erhebungsverfahren im Kontext der internationalen PISA-Option ICT familiarity questionnaire. Für den darin enthaltenen Teil, der ICT-Skills als Selbstbericht erfasst (OECD, 2011), sollte eine stärkere theoretische Verankerung geschaffen und entsprechende valide Indikatoren entwickelt und empirisch erprobt werden. Basierend auf der im Projektverlauf entwickelten theoretischen Rahmenkonzeption wird ein Assessmentframework mit entsprechenden Indikatoren für die zu konstruierenden Testaufgaben (Items) abgeleitet. Das Assessmentframework stellt die Grundlage für die Entwicklung und empirische Erprobung von Items dar. Derzeit werden bei PISA computerbezogene Kompetenzen 15-jähriger Schülerinnen und Schüler im Sinne der Vertrautheit mit Informations- und Kommunikationstechnologien, einschließlich ICT-Skills, lediglich anhand des Schülerfragebogens ermittelt. Bis PISA 2009 bestand die Fragenbatterie in Selbsteinschätzungen hinsichtlich der eigenen Kompetenz im Umgang mit Computeranwendungen. Aufgrund der infrage zu stellenden Validität von Selbsteinschätzungen für die Ableitung von Aussagen über individuelle Kompetenzunterschiede wurde bei PISA 2012 der Versuch unternommen, diese durch szenariobasierte Wissensfragen zu ersetzen.

Nach Klieme (2004) gehört zu kompetentem Handeln neben Wissen aber auch Können, weshalb ein reiner Wissenstest dem Anspruch der Messung von ICT-Skills nicht gerecht werden würde. Demnach steht auch im vorliegenden Beitrag nicht nur das reine Wissen im Fokus, sondern vielmehr die individuellen Nutzungsmöglichkeiten von zu messenden Fertigkeiten. Vor diesem Hintergrund wird ein Messinstrument entwickelt, das valide ist, weil es verhaltensbasierte Messungen von ICT-Skills ermöglicht, und das besonders reliabel ist, weil es eine adaptive Anwendung erlaubt, die eine computergesteuerte, dem Kenntnisstand der jeweils untersuchten Person angemessene Vorgabe von Testaufgaben ermöglicht. Das computerbasierte Messinstrument ist so angelegt, dass es in Large-Scale-Assessments eingesetzt werden kann. Das neu entwickelte Instrument wird psychometrisch überprüft, um schließlich auf dieser Grundlage einen adaptiven Testalgorithmus erstellen zu können.

2.2 Rahmenkonzeption von ICT-Skills

Aufbauend auf einer Synthese aktueller Rahmenkonzeptionen von ICT-Skills wurden erste Überlegungen zu einer Rahmenkonzeption gemeinsam mit internationalen Experten für ICT-Literacy, Large-Scale-Assessments, Wissenserwerb mit Hypermedia und Leseverständnis diskutiert. Zu Beginn stand die Frage im Raum, inwiefern Rahmenkonzepte für ICT-Skills entwickelt werden können, die trotz ständiger technologischer Weiterentwicklung und Veränderung nicht sofort veralten. Um dem Rechnung zu tragen, wurde die Domäne nicht basierend auf den technischen, sich stetig verändernden Werkzeugen und Anwendungen strukturiert, sondern es wurde ein aufgabenzentrierter Ansatz basierend auf kognitiven Anforderungen gewählt. Diese Aufgaben und die daraus resultierenden Anforderungen bleiben bestehen,

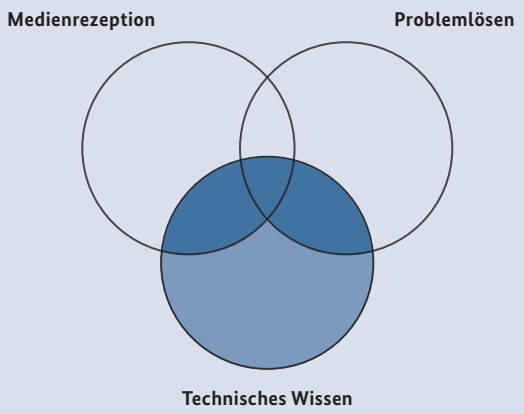
auch wenn sich die Anwendungen und Technologien weiterentwickeln oder durch andere ersetzt werden.

ICT-Skills werden definiert als Wissen und Fertigkeiten zur Nutzung von Informations- und Kommunikationstechnologien, die für die erfolgreiche Teilhabe in modernen Wissensgesellschaften relevant sind. ICT-Skills umfassen die Rezeption, Anwendung und Produktion multimedialer Information zur erfolgreichen Bearbeitung informationsbezogener Aufgaben.

Der Umgang mit Informationen spielte schon vor der Verbreitung von ICT eine Rolle, und deshalb werden auch die hierfür benötigten Fertigkeiten als nicht vollkommen neu verstanden. Es wird davon ausgegangen, dass sich ICT-Skills aus traditionellen Fertigkeiten wie Problemlösen und Medienrezeption, worunter auch Lesekompetenz fällt, entwickelt haben und auf technischem Wissen basieren.

Die Zusammenhänge zwischen technischem Wissen, Problemlösen und Medienrezeption werden im Abbildung 1 dargestellt. Die für das Konstrukt ICT-Skills als relevant angesehenen Aspekte wurden blau markiert. Somit wird Problemlösen unter Nutzung von technischem Wissen beispielsweise ICT-Skills zugeordnet, wohingegen dies bei Problemlösen ohne Rückgriff auf technisches Wissen nicht der Fall ist. Der Abbildung ist weiterhin zu entnehmen, dass reines technisches Wissen ebenfalls den ICT-Skills zugeordnet wurde.

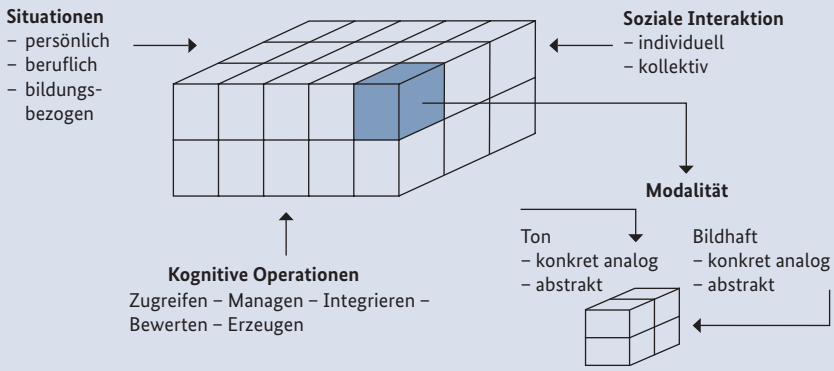
Abb. 1: Gegenstandsbereich von ICT-Skills



Zu dem Spektrum der informationsverarbeitenden Aufgaben zählen die fünf kognitiven Operationen des Internationalen ICT Literacy Panel (2002), die eine zentrale Rolle einnehmen und die erste Facette der Aufgabenbeschreibung darstellen. Diese sind das Zugreifen (access), Managen (manage), Integrieren (integrate), Bewerten (evaluate) und Erzeugen (create) von Informationen. Die Aufgaben können entweder individuell bearbeitet werden oder in der Kommunikation mit anderen zu einer kollektiven Aufgabe werden (Facette „soziale Interaktion“). Gleichzeitig soll zwischen

verschiedenen Kontexten unterschieden werden, in denen diese Anforderungen auftreten können. Diese sind persönliche, bildungsbezogene und berufliche Situationen (Facette „Situationen"). Um den verschiedenen Repräsentationsformen und den daraus resultierenden Anforderungen Rechnung zu tragen, wird unterschieden, ob Informationen visuell oder auditiv präsentiert werden, wobei diese Informationen jeweils konkret analog oder abstrakt sein können (Facette „Modalität"). Eine Darstellung aller Facetten ist Abbildung 2 zu entnehmen. Das primäre Ziel der Facettierung besteht in der vollständigen und strukturierten Beschreibung des Messgegenstandes. Die einzelnen Facetten und ihre Ausprägungen werden dabei nicht als psychometrisch trennbare Dimensionen angesehen. Es wird also nicht davon ausgegangen, dass die einzelnen Facetten vollständig unabhängig voneinander sind, vielmehr bilden sie den Rahmen, in dem ein kompetenter Umgang mit Informations- und Kommunikationstechnologien erforderlich ist. Entsprechend dieser Zielsetzung wurden Aufgaben entwickelt, die bestimmte Kombinationen der Facettenausprägungen (im Sinne von „Würfeln") operationalisieren.

Abb. 2: Theoretisches Framework für ICT-Skills



Die kognitiven Anforderungen (Zugreifen, Managen, Integrieren, Bewerten, Erzeugen) in diesen Aufgaben wurden basierend auf Theorien aus den Domänen Problemlösen und Medienrezeption abgeleitet und für die Itementwicklung genutzt (Engelhardt et al., eingereicht).

3 Itementwicklungsprozess

Auf Grundlage des theoretischen Frameworks wurde ein Assessmentframework entwickelt, welches im Wesentlichen der in Abbildung 2 dargestellten Struktur entspricht. Es definiert das sogenannte Itemuniversum, also jene Menge an Testaufgaben, die zur Messung des Konstrukts verwendet werden kann. So wurde versucht,

Items entsprechend den verschiedenen Ausprägungen jeder Facette zu entwickeln. Beispielsweise sollte die Schwierigkeit einer Aufgabe, die das „Zugreifen" auf Information abbilden soll, tatsächlich in der Informationsbeschaffung liegen, nicht aber in der Bewertung der Information hinsichtlich ihrer Glaubwürdigkeit, da dies unter die Facettenausprägung „Bewerten" fallen würde.

Weiterhin wurden mithilfe einer Onlinebefragung Daten zu ICT-spezifischen Problemen von Jugendlichen bei der Nutzung von ICT erhoben. Die gesammelten Informationen wurden zur genaueren Beschreibung des Problemfeldes und zur Generierung von Aufgabenstellungen im Rahmen der Itemkonstruktion genutzt. Eine Zielsetzung der Itementwicklung war es, mit den entwickelten Items das Assessmentframework möglichst umfassend und balanciert abzubilden. Für die Itemkonstruktion galt die zentrale Annahme, dass mit direkt am Computer zu bearbeitenden und auf Simulationen basierenden Items Verhaltensunterschiede gemessen werden können, die das interessierende Merkmal direkt repräsentieren.

Um Items einheitlich, orientiert am Assessmentframework, entsprechend üblichen Leitlinien zur Itementwicklung (vgl. Osterlind, 1998) und unter Verwendung einer speziellen Software (CBA ItemBuilder; Rölke, 2012) zu erstellen, fand ein dreitägiger Itemwriting-Workshop statt.

Die Itementwicklung umfasste folgende Schritte:

(1) Generierung einer Itemidee und detaillierte schriftliche Dokumentation dieser Idee
(2) Überprüfung und Überarbeitung hinsichtlich Inhalt und Umsetzbarkeit der Itemidee mit mehreren Rückmeldeschleifen
(3) computerbasierte Umsetzung der Itemidee unter Verwendung des CBA ItemBuilder
(4) Bewertung (Scoring) von korrekten und inkorrekten Lösungen und Teillösungen
(5) Erprobung der Items in Cognitive Laboratories[1]
(6) abschließende Prüfung der fertigen computerbasierten Items

Bereits bei der Generierung der Itemidee musste dargelegt werden, in welchem Teil des Assessmentframeworks das Item zu verordnen ist (z. B. Zugreifen, individuell, bildungsbezogen, bildhaft konkret analog). Hiermit wurde eine zielgerichtete Entwicklung sichergestellt und bewirkt, dass Items nicht mehrere Ausprägungen einer Facette gleichzeitig abbilden. Neben diesen inhaltlichen Überlegungen wurden auch Konkretisierungen auf technischer Ebene zur Umsetzung und automatischen Bewertung der Items festgehalten, sowie Ideen zu Aufbau und Layout des computerisierten Items vorgenommen. Die detaillierte Dokumentation der Itemideen erlaubte eine Überprüfung nicht nur auf inhaltlicher Ebene, sondern auch bezüglich der technischen Umsetzbarkeit der Items. Auf diese Weise war es möglich, Items zu vereinfa-

1 Cognitive Laboratories beinhalten Methoden, mit deren Hilfe Informationen über kognitive Prozesse während des Frage-Antwort-Prozesses gesammelt werden können (vgl. Prüfer & Rexroth, 1996). Im Projekt CavE-ICT kamen im Rahmen der Cognitive Labs vor allem die Methoden Think Aloud (der Befragte wird aufgefordert, bei der Bearbeitung eines Items laut zu denken, um sämtliche Gedankengänge, die zur Lösung führen, erfassen zu können) und Probing (eine gegebene Antwort wird hinterfragt, um weitere Informationen zu erhalten) zur Anwendung.

chen oder Anforderungen an den CBA ItemBuilder frühzeitig zu identifizieren und die Weiterentwicklung dieser Software voranzutreiben.

Für den Prozess der Itemimplementierung wurden zunächst Vorlagen entwickelt, durch die ein einheitliches Design für wiederkehrende Elemente in den Items festgelegt wurde. So wurde beispielsweise ein E-Mail-Postfach erstellt, das übliche E-Mail-Funktionalitäten umfasst, aber ausreichend von gängigen E-Mail-Postfächern abstrahiert ist und so keine Vorteile für bestimmte Nutzergruppen aufgrund des Designs bietet. Diese Vorlage wurde für alle Items mit E-Mail-Bezug verwendet.

Um Items hinsichtlich inhaltlicher, aber auch technischer Aspekte zu prüfen, wurden im Prozess der Itementwicklung stetig Cognitive-Laboratory-Studien eingesetzt. So konnten Unstimmigkeiten in der Aufgabenlogik oder untypische technische Umsetzungen schon im Voraus überarbeitet werden. Neben einer notwendigen Itemüberarbeitung konnte dadurch außerdem die Bearbeitungszeit der Items abgeschätzt und überprüft werden, ob die intendierten Schwierigkeitsniveaus angemessen sind.

Um Testteilnehmerinnen und Testteilnehmer in die Lage zu versetzen, die ICT-Items potenziell bearbeiten zu können, galt es, für eine ausreichende Vertrautheit mit der Simulationsumgebung zu sorgen. Zu diesem Zweck wurde ein etwa zehnminütiges Training erstellt, welches vor Beginn der Testung von den Schülerinnen und Schülern bearbeitet werden musste. Dieses Training führte in die Besonderheiten der Simulationsumgebung ein und machte die Teilnehmerinnen und Teilnehmer mit der Art der Aufgabenstellungen im Test vertraut.

Bereits bei der Itemimplementierung wurden für jedes Item automatische Bewertungsregeln entwickelt, welche nicht nur festhielten, ob das entsprechende Item richtig oder falsch gelöst wurde, sondern die gleichzeitig auch Teilprozesse der Aufgabenlösung abbildeten. Dies erlaubt die nachträgliche Analyse alternativer Lösungswege oder häufig erreichter Teillösungen. Analysen dieser Art können zum einen ein Nach- oder Umbewerten besonders schwerer oder besonders leichter Items ermöglichen, zum anderen aber auch Impulse zur Konstruktion neuer Items liefern.

4 Kalibrierungsstudie

In der Kalibrierungsstudie wurden die entwickelten ICT-Items empirisch erprobt. Im Folgenden werden die Zielsetzungen (Kapitel 4.1), das gewählte methodische Vorgehen (Kapitel 4.2) sowie erste Ergebnisse der Kalibrierungsstudie (Kapitel 4.3) beschrieben.

4.1 Zielsetzung

Ziel der Kalibrierungsstudie war es zum einen, die entwickelten ICT-Items hinsichtlich zentraler psychometrischer Aspekte zu analysieren, und zum anderen zu untersuchen, ob sich mit den entwickelten Items ein zuverlässiges, d. h. genau messendes Testinstrument zur Bestimmung von ICT-Skills konstruieren lässt.

4.2 Methode

Dieses Teilkapitel gibt einen Überblick über die Planung der Datenerhebung, die Zuordnung der zu analysierenden ICT-Items entsprechend der Rahmenkonzeption von ICT-Skills, das angewendete Testheftdesign, die Testdurchführung, die untersuchte Personenstichprobe und die statistischen Methoden zur Analyse der im Rahmen der Kalibrierungsstudie gesammelten Daten.

4.2.1 Erhebungsplanung

Die Planung und Organisation der Datenerhebung im Rahmen der Kalibrierungsstudie erfolgte in Kooperation mit dem IEA Data Processing Center (DPC) in Hamburg. Neben der Bestimmung eines günstigen Erhebungszeitraums mussten Genehmigungsanträge bei den Kultusministerien verschiedener Bundesländer gestellt und danach entsprechende Schulen durch Ansprache und Information rekrutiert werden. Zudem wurde festgelegt, welche Merkmale neben ICT-Skills erhoben werden sollten. Die Fragen zur Erhebung der ausgewählten Hintergrundvariablen sowie weitere Validierungsinstrumente wurden anschließend computerisiert, um sie gemeinsam mit dem ICT-Test computerbasiert auszuliefern.

In der Erhebungsplanung galt es zudem verschiedene technische Aspekte zu adressieren, wie die Prüfung technischer Gegebenheiten an den Schulen, die an der Studienteilnahme interessiert waren. Dazu wurde ein Computerprogramm zur Systemdiagnose entwickelt, welches an den Schulen im Vorfeld der eigentlichen Untersuchung eingesetzt wurde. Die Ergebnisse der Systemdiagnose wurden genutzt, um einen technisch reibungslosen Testverlauf zu gewährleisten. Das heißt, es wurden nur Schulen für die Erhebung ausgewählt, die die erforderlichen technischen Voraussetzungen in Bezug auf die Hardware erfüllten.

4.2.2 ICT-Items

In der Kalibrierungsstudie kamen insgesamt 70 ICT-Items zum Einsatz. Diese Items lassen sich entsprechend der in Kapitel 2 vorgestellten theoretischen Rahmenkonzeption von ICT-Skills verschiedenen Facetten beziehungsweise deren Ausprägungen zuordnen. Die in Kapitel 3 skizzierte Itemkonstruktion zielte darauf ab, Items zu erarbeiten, die exklusiv jeweils einer kognitiven Operation, einer Situation und einer sozialen Interaktion zugeordnet sind. In Tabelle 1 werden die Itemanzahlen nach Facetten und Facettenausprägung aufgeschlüsselt dargestellt.

Die simulationsbasierten Items wurden automatisch entsprechend den im Itementwicklungsprozess festgelegten Bewertungsregeln als richtig oder falsch codiert. Damit liegt ein dichotomes Antwortformat vor.

Tabelle 1: Itemanzahl nach Facetten und Facettenausprägung des Frameworks zur Messung von ICT-Skills

	Soziale Interaktion						Summe
	Individuell			Kollektiv			Item-anzahl
Kognitive Operationen	Situationen						
	Persönlich	Beruflich	Bildungs-bezogen	Persönlich	Beruflich	Bildungs-bezogen	
Zugreifen	3	1	5	2	2	0	13
Managen	4	4	1	2	6	7	24
Integrieren	1	1	1	2	3	2	10
Bewerten	5	1	4	1	2	0	13
Erzeugen	0	3	2	1	1	3	10
	13	10	13	8	14	12	**70**

4.2.3 Studiendesign

Bei der Festlegung des Designs der Kalibrierungsstudie galt es ebenso festzulegen, welche Merkmale zusätzlich zu den ICT-Skills erhoben werden sollten. Zielsetzung der Erhebung war, neben der Kalibrierung der ICT-Items auch empirische Evidenz für Testwertinterpretationen im Sinne der intendierten Nutzung der ICT-Skala und somit für die Validität dieser testwertbezogenen Interpretationen zu gewinnen. In diesem Sinne galt es für die Untersuchung ein Testdesign zu spezifizieren, welches den Einsatz der einzelnen Instrumente festlegt und damit ein ICT-Kalibrierungs- und Validierungsdesign integriert. In Abbildung 3 ist das resultierende Design der Kalibrierungsstudie veranschaulicht. Im Folgenden wird nun genauer auf die Erstellung des ICT-Testhefts-designs und die Zusammensetzung des Validierungsdesigns eingegangen.

Abb. 3: Studiendesign der Erhebung zur empirischen Überprüfung der ICT-Items und Validierung

<table>
<tr><th colspan="2">Teil 1 (60 min)</th><th rowspan="2">Pause (ca. 10 min)</th><th colspan="3">Teil 2 (60 min)</th></tr>
<tr><td rowspan="2">Training</td><td>ITC-Skills</td><td rowspan="3">Nutzungs-häufigkeit von ICT</td><td>Zu Validierungs-zwecken veränder-te ICT-Items</td><td>Interesse an ICT, Einstellungen und Selbstkonzept, technisches Wissen</td></tr>
<tr><td>Zu Validierungs-zwecken verän-derte ICT-Items</td><td>Kognitive Grundfähigkeit</td><td>Interesse an ICT, Einstellungen und Selbstkonzept</td></tr>
<tr><td></td><td></td><td>Lesefähigkeit, Problemlöse-fähigkeit</td><td>Technisches Wissen</td></tr>
</table>

Die geplante Testzeit betrug insgesamt, inklusive einer etwa zehnminütigen Pause zwischen der Erhebung des ICT-Tests und der Validierungsinstrumente, drei Schulstunden und somit etwa 135 Minuten.

ICT-Testheftdesign

Um im geplanten zeitlichen Umfang der Testung alle 70 ICT-Items vorlegen zu können, wurde ein balanciertes unvollständiges Testheftdesign (z. B. Frey, Hartig & Rupp, 2009) verwendet. Dies bedeutet, dass einzelne Schülerinnen und Schüler jeweils nur einen Teil der ICT-Items zur Bearbeitung vorgelegt bekamen. Die verfügbare Testzeit lag bei 60 Minuten, wobei etwa 10 Minuten zu Beginn der Testung für das Tutorial verwendet wurden. In den verbliebenen 50 Minuten konnten etwa 31 ICT-Items bearbeitet werden. Damit enthielt jedes Testheft etwas weniger als die Hälfte aller entwickelten ICT-Items. Das gewählte Testheftdesign wird in Tabelle 2 dargestellt. Demnach bestand ein Testheft aus fünf Itembündeln, die wiederum jeweils aus durchschnittlich sechs Items zusammengestellt wurden. Insgesamt wurden elf ICT-Testhefte erstellt.

Tabelle 2: Zuordnung der Itembündel zu den elf ICT-Testheften

		ICT-Testheft										
Position	Minuten	1	2	3	4	5	6	7	8	9	10	11
1	10	2	3	4	5	6	7	8	9	10	11	1
2	10	4	5	6	7	8	9	10	11	1	2	3
3	10	5	6	7	8	9	10	11	1	2	3	4
4	10	6	7	8	9	10	11	1	2	3	4	5
5	10	10	11	1	2	3	4	5	6	7	8	9

Bei der Zusammenstellung der Bündel wurde zudem darauf geachtet, dass jedes ICT-Testheft Items enthielt, die möglichst ausgewogen die verschiedenen Ausprägungen der Facette der kognitiven Operationen und verschiedene technische Applikationen (z. B. Webbrowser, Ordnerstruktur, Textverarbeitungsprogramm) abbilden. Des Weiteren wurden Schätzungen der Bearbeitungszeit der einzelnen Items aus den im Itementwicklungsprozess durchgeführten Cognitive-Laboratory-Studien genutzt, um die Testhefte im Hinblick auf die zur Beantwortung erforderliche Zeit möglichst homogen zu gestalten. Dies war erforderlich, da die Bearbeitungszeiten zwischen den einzelnen Items variieren.

Die einzelnen Items konnten entsprechend dem Testheftdesign in das erprobte Auslieferungssystem eingebettet werden. Mit einer Stichprobengröße von 766 Schülerinnen und Schülern liegen ungefähr 290 Bearbeitungen jedes Items vor. Damit können zuverlässige Schätzungen der Itemschwierigkeiten vorgenommen werden.

Validierungsdesign

Für Validierungszwecke wurden bereits vorliegende evidenzbasierte Instrumente, wie der Lesegeschwindigkeits- und -verständnistest (LGVT; Schneider, Schlagmüller & Ennemoser, 2007), der Berliner Test zur Erfassung fluider und kristalliner Intelligenz (BEFKI; Wilhelm, Schroeders & Schipolowski, 2014), ein Test zur Erfassung von theoretischem Computerwissen (TECOWI) aus dem Inventar zur Computerbildung (INCOBI-R; Richter, Naumann & Horz, 2010), und der Complex Problem Solving Test MicroDYN (CPS; Greiff & Funke, 2009) eingesetzt. Während LGVT, BEFKI und TECOWI zunächst computerisiert werden mussten, lag das Instrument zur Erfassung komplexer Problemlösefähigkeit bereits computerisiert vor und konnte direkt ins Testsystem aufgenommen werden. Außerdem wurden Fragen zur Selbsteinschätzung vorgelegt. Hier gab es Fragen zu Nutzungshäufigkeiten von ICT in Freizeit und Schule, aber auch Fragen zu computerbezogenem Interesse, Einstellungen und Selbstkonzept. Sie wurden basierend auf dem ICT familiarity questionnaire (OECD, 2011) zusammengestellt.

Darüber hinaus wurden zu Validierungszwecken einige der ICT-Items verändert. So wurden zum einen ICT-Herausforderungen gezielt eliminiert, zum anderen wurden bei einigen Items spezifische ICT-Herausforderungen verändert (z. B. in ihrer Schwierigkeit) und schließlich ausgewählte Items zu nicht interaktiven Testitems im Multiple-Choice-Format abgewandelt. Bei der Vorgabe der Validierungsvariablen kam wiederum ein unvollständiges Versuchsdesign zum Tragen, da, mit Ausnahme der Items zur Nutzungshäufigkeit von ICT, nicht alle Schülerinnen und Schüler alle Validierungsinstrumente vorgelegt bekamen.

4.2.4 Testdurchführung

Die empirische Erprobung der entwickelten ICT-Items wurde in Kooperation mit dem DPC durchgeführt. Die Testzeit betrug, wie bereits ausgeführt, insgesamt 135 Minuten, inklusive einer etwa 10- bis 15-minütigen Pause zwischen der Erhebung des ICT-Tests (ca. 60 Minuten) und der Validierungsinstrumente (ca. 60 Minuten).

Zur Administration der ICT-Items, wie auch der Validierungsinstrumente, wurde ein Testsystem entwickelt. Dieses System wurde an ausgewählten Schulen vorerprobt und entsprechend optimiert. Die Testauslieferung wurde über USB-Sticks realisiert.

Die Testleiterinnen und -leiter, welche die Erhebung an den Schulen durchführten, wurden vorab intensiv geschult. Ausgewählt und koordiniert wurden sie vom DPC. Die Schulung und die Herstellung von Manualen für die Testleiterinnen und -leiter wurden in Zusammenarbeit mit dem DPC vom CavE-ICT-Projektteam konzipiert. Zudem stand ihnen während des gesamten Erhebungszeitraums eine Hotline für technische Unterstützung zur Verfügung, die durch das Projektteam abgedeckt wurde.

4.2.5 Stichprobe

Die Stichprobe der Kalibrierungsstudie umfasste 983 Schülerinnen und Schüler. Da ein Teil der zu Validierungszwecken veränderten ICT-Items nicht mit den eigentlichen ICT-Items zusammen vorgegeben werden konnte, bearbeiteten nur 766 der 983 Schülerinnen und Schüler die entwickelten Items zur Messung von ICT-Skills. Die übrigen 217 Schülerinnen und Schüler bearbeiteten parallel ICT-Items mit veränderten ICT-spezifischen Herausforderungen. Insgesamt wurden Daten in 33 Schulen in Baden-Württemberg und Rheinland-Pfalz erhoben. 10 der Schulen waren Gymnasien, die anderen 23 Schulen setzen sich aus Realschulen, Hauptschulen und Gesamtschulen zusammen. 71 Prozent der Befragten besuchten die neunte Klasse, 26 Prozent gingen zum Zeitpunkt der Testung in die zehnte Klasse (3 Prozent machten keine Angabe). Die Schülerinnen und Schüler waren im Mittel 15,21 Jahre alt (SD = 0.57). 51 Prozent der Testpersonen waren männlich, 46 Prozent weiblich (3 Prozent machten keine Angabe zum Geschlecht).

4.2.6 Datenauswertung

Mit dem im Projekt CavE-ICT entwickelten ICT-Test soll von den erhobenen Schülerantworten auf die individuelle Ausprägung von ICT-Skills der Schüler geschlossen werden. Die Verortung von Schülerinnen und Schülern auf der Kompetenzskala und die Bestimmung der Itemschwierigkeiten mithilfe eines statistischen Modells wird Skalierung oder auch Kalibrierung genannt. Im Projekt wird für die Bestimmung individueller Kompetenzwerte und Itemschwierigkeiten auf statistische Modelle der Item-Response-Theorie (IRT; van der Linden & Hambleton, 2013) zurückgegriffen. Ein Vorzug der IRT besteht darin, dass sich die Schwierigkeit eines Items und die Kompetenz einer Person auf derselben Skala abbilden lassen. Die Anwendung von IRT-Modellen bietet sich auch aufgrund des verwendeten unvollständigen Testheftdesigns an, bei dem nicht alle Testpersonen alle ICT-Items bearbeiteten. Sind eine ausreichend große Stichprobe und eine hinreichende Überlappung der Testheftinhalte gegeben, ermöglicht die IRT-Skalierung einen Vergleich der Schülerleistungen trotz unterschiedlicher Testinhalte. Aus diesen und weiteren Gründen werden IRT-Modelle bei Large-Scale-Assessments standardmäßig zur Skalierung eingesetzt. Die Berechnungen erfolgten mit dem Computerprogramm ConQuest (Wu, Adams, Wilson & Haldane, 2007).

Die Datenauswertung erfolgte in zwei Schritten. Zunächst wurde die psychometrische Güte der ICT-Items geprüft, um ungeeignete Items identifizieren und ausschließen zu können (Kapitel 4.3.1). Anschließend erfolgte mit dem verbliebenen Datensatz die abschließende Skalierung sowie die Berechnung der Kompetenzverteilung (Kapitel 4.3.2). Vor allem im Kontext von Large-Scale-Assessments ist es oft das Ziel, Kompetenzen zu erfassen, die sich gemäß ihrer theoretischen Konzeption aus anderen, spezielleren Subkompetenzen zusammensetzen. Auch im Projekt CavE-ICT wurde eine derart komplexe Kompetenzmessung angestrebt. Ein gängiger (z. B. bei PISA) genutzter Ansatz bei der Berechnung aller für die Berichtlegung gewünschten Kompetenzskalen ist die Analyse der Daten mithilfe eines eindimensionalen Modells (zur Skalierung der Globalskala) und eines mehrdimensionalen Modells (zur Skalie-

rung der Subskalen). Dieses Vorgehen wurde auch im vorliegenden Fall gewählt, sodass im Zuge der Itemselektion zunächst ein eindimensionales Rasch-Modell zum Einsatz kam. Für die zweite Skalierung mit jeweils einer Subskala für die fünf kognitiven Prozesse wurden die Itemparameter auf die Werte aus der ersten Skalierung fixiert und ein fünfdimensionales Rasch-Modell angewendet. Zudem wurden das ein- und fünfdimensionale Basismodell um ein Hintergrundmodell erweitert. Auf dessen Basis wurden plausible Werte (plausible values, PVs; Wu, 2005) gezogen. Die Hintergrundinformationen schlossen unter anderem Variablen zu Alter, Geschlecht oder Computernutzung ein, wodurch eine unverfälschte Schätzung varianzabhängiger Populationsparameter ermöglicht wurde (von Davier & Sinharay, 2009).

4.3 Ergebnisse der Kalibrierungsstudie

In diesem Teilkapitel werden die Ergebnisse der Kalibrierungsstudie dargestellt. Zunächst wird die Auswahl der Items für die Skalierung skizziert (Kapitel 4.3.1). Danach werden die Ergebnisse der abschließenden Skalierung beschrieben (Kapitel 4.3.2).

4.3.1 Itemauswahl für die abschließende Skalierung

Ziel der Itemselektion war es, Items auszuwählen, die eine psychometrisch abgesicherte, aber auch eine theoriekonforme Erfassung von ICT-Skills ermöglichen. Neben der Itemschwierigkeit wurden die Passung einzelner Items mit dem Rasch-Modell (anhand des WMNSQ und entsprechenden t-Wertes) und die Korrelation des Itemtestwertes mit dem Testwert der Gesamtskala als Kriterien für die psychometrische Güte berücksichtigt. Zudem wurde analysiert, ob Männer oder Frauen systematisch bei der Beantwortung von bestimmten Items bevorzugt oder benachteiligt werden. Diese Effekte des Geschlechts auf die Lösungswahrscheinlichkeit einzelner Items werden auch differenzielles Itemfunktionieren (DIF; Osterlind & Everson, 2009) genannt. Zusätzlich wurden bei statistisch auffälligen Items unabhängige Inhaltsanalysen durchgeführt.

In begründeten Ausnahmefällen wurden kleine Abweichungen von den Kriterien toleriert. Ein Item wurde von keiner Person korrekt gelöst, woraufhin dieses Item eingehend hinsichtlich technischer Funktionalität, Instruktionsgenauigkeit und Bewertung geprüft wurde. Auf Basis dieser Überprüfung konnte die Bewertung des Items neu gestaltet und es dem endgültigen ICT-Itempool hinzugefügt werden.

Für die abschließende Skalierung konnten schließlich 64 der 70 ICT-Items ausgewählt werden. Bei zwei der sechs ausgeschlossenen Items wurden zu niedrige Korrelationen beobachtet. Insgesamt vier Items zeigten in der psychometrischen und inhaltlichen Analyse Hinweise auf DIF bezüglich des Geschlechts. Eines dieser Items wies zudem eine zu niedrige Modellpassung auf.

Die ausgeschlossenen Items können zum Teil überarbeitet und unter Umständen in einer späteren Untersuchung wieder eingesetzt werden. Die nachfolgend dargestellten Ergebnisse zur abschließenden Skalierung beziehen sich auf die 64 selektierten Items, die den endgültigen ICT-Itempool bilden.

4.3.2 Abschließende Skalierung

Im Folgenden werden die Ergebnisse der abschließenden Skalierung der Daten der Kalibrierungsstudie dargestellt. Hierzu wird zunächst auf die geschätzten Itemparameter sowie die Reliabilität der ICT-Skala und der kognitiven Subskalen eingegangen. Anschließend wird die Passung von Personen- und Itemparametern dargestellt, um schließlich die latenten Korrelationen zwischen den verschiedenen kognitiven Operationen zu berichten.

Itemkennwerte

Zunächst wurden die 64 selektierten Items ohne Verwendung eines Hintergrundmodells nochmals skaliert und Itemparameter geschätzt. Bei der abschließenden Skalierung mit dem Hintergrundmodell wurden die Itemschwierigkeitsparameter auf Itemparameter aus der vorherigen Skalierung fixiert. Die Schwierigkeiten der ICT-Globalskala nehmen Werte zwischen −2.96 und 4.24 Logits mit einem Mittelwert von M = 0.34 und einer Standardabweichung von SD = 1.58 Logits an. Die Betrachtung der mittleren Itemschwierigkeiten nach Ausprägungen der Facette „kognitive Operation“ (Tabelle 3) weist darauf hin, dass die Itemschwierigkeiten für die verschiedenen kognitiven Operationen unterschiedlich ausfallen.

Tabelle 3: Deskriptive Statistiken der Itemschwierigkeiten für die ICT-Gesamtskala und die Subskalen für die kognitiven Operationen

	Itemschwierigkeit			
Kognitive Operationen	Min	Max	M	SD
Zugreifen	−1.39	3.47	0.66	1.48
Managen	−2.33	4.24	0.54	1.71
Integrieren	−0.77	3.54	0.91	1.46
Bewerten	−2.04	1.67	−0.26	1.22
Erzeugen	−2.96	2.52	−0.39	1.73
ICT-Global	−2.96	4.24	0.34	1.58

Die Korrelationskoeffizienten liegen im Bereich von .11 bis .56, wobei acht Items mit Werten < .20 einen geringen Zusammenhang mit der ICT-Globalskala aufweisen. Die nachfolgenden Ergebnisdarstellungen zu Reliabilität und der Passung von Item- und Personenparametern (hier die ICT-Fertigkeiten) beziehen sich auf die abschließende Skalierung mit dem Hintergrundmodell.

Reliabilität

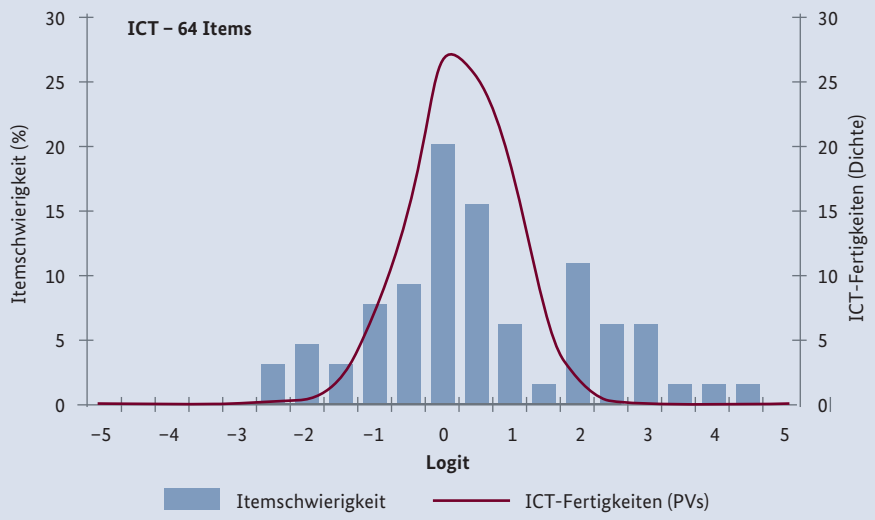
Der Begriff der Reliabilität bezeichnet die Zuverlässigkeit oder Messgenauigkeit eines Messinstruments. Die Reliabilität entspricht dem Anteil der Varianz, der durch Unterschiede im zu messenden Merkmal und nicht durch Messfehler oder andere Merkmale erklärt werden kann. Im vorliegenden Fall wird sie als Anteil der geschätzten latenten Varianz an der Varianz der individuellen EAP-Punktschätzer der ICT-Fertigkeiten in der untersuchten Stichprobe verstanden. Diese Reliabilität wird als EAP/PV-Reliabilität bezeichnet und bei Large-Scale-Assessments üblicherweise verwendet (Adams, 2005).

Für die ICT-Globalskala ergibt sich ein Reliabilitätskoeffizient von .80. Die Reliabilitäten der Subskalen für die kognitiven Operationen liegen zwischen .64 und .73 (Zugreifen: .73, Managen: .72, Integrieren: .64, Bewerten: .64 und Erzeugen: .70) und erreichen damit nur teilweise einen akzeptablen Wert.

Passung Item- und Personenparameter

Eine gute Passung zwischen Itemschwierigkeiten und ICT-Fertigkeiten ist im Hinblick auf hohe Differenzierungsfähigkeit bei hoher Messeffizienz erstrebenswert. Bei Betrachtung der ICT-Globalskala zeigt sich, dass die für die Skalierung ausgewählten Items den Merkmalsbereich gut abdecken (Abbildung 4).

Abb. 4: Itemschwierigkeit und ICT-Fertigkeiten (plausible values) der für die abschließende Skalierung mit dem Hintergrundmodell selektierten ICT-Items



Es wird aber auch deutlich, dass eine Vielzahl von Items im höheren Schwierigkeitsbereich liegt und sich damit deutlich über der Verteilung der ICT-Fertigkeiten befindet. Im Hinblick auf die Subskalen für die kognitiven Operationen zeigt sich, dass trotz der recht geringen Anzahl von Items pro kognitive Subskala eine verhältnismäßig gute Abdeckung der Fähigkeitsverteilung durch die Items gelungen ist.

Korrelationen zwischen den kognitiven Subskalen

Die latenten Korrelationen zwischen den verschiedenen kognitiven Operationen liegen in einem Bereich von .48 bis .72 und sind in Tabelle 4 zu finden. Die Zusammenhänge bewegen sich damit im für dieses Konstrukt erwarteten Bereich. Zudem weisen die Anteile erklärter Varianz an Gesamtvarianz der Facettenausprägung mit 22,66 Prozent bis 51,55 Prozent darauf hin, dass die kognitiven Operationen zwar Gemeinsamkeiten aufweisen, darüber hinaus aber auch spezifische Inhalte abbilden.

Tabelle 4: Latente Korrelationen und Varianzen der ICT-Fertigkeiten

	Kognitive Operation				
	Zugreifen	Managen	Integrieren	Bewerten	Erzeugen
Zugreifen					
Managen	.72				
Integrieren	.59	.72			
Bewerten	.48	.59	.53		
Erzeugen	.51	.63	.58	.50	
Varianz	.42	.53	.61	.36	.53

5 Fazit und Ausblick

Im Folgenden soll zum Abschluss ein kurzes Fazit der Testentwicklung gezogen (Kapitel 5.1) und ein abschließender Ausblick gegeben werden (Kapitel 5.2).

5.1 Fazit

Die Testentwicklung im Projekt CavE-ICT hat sich an dem bei Large-Scale-Assessments etablierten Prozess der Testentwicklung und den damit verbundenen hohen Qualitätsstandards orientiert. Zunächst wurde an der theoretischen Basis, der Definition, der Verortung und internen Struktur gearbeitet, wodurch eine theoriegeleitete Testentwicklung möglich war. Mehrere Rückmeldeschleifen an verschiedenen Stel-

len im Entwicklungsprozess stellten die Qualität und Zielgerichtetheit jedes einzelnen Items sicher, um Ausschuss zu vermeiden. Dies ist sicher ein wesentlicher Grund dafür, dass nur sechs der konstruierten Items eliminiert werden mussten. Da der Entwicklungsaufwand für die simulationsbasierten Items ausgesprochen hoch war, war eine geringe Ausschussquote wichtig für den Erfolg des Projekts. Durch den hohen Grad der Strukturierung des Itementwicklungsprozesses konnte weiterhin sichergestellt werden, dass die verschiedenen seitens des Assessmentframeworks spezifizierten Facetten und deren Ausprägungen durch die entwickelten Items umfassend über einen breiten Schwierigkeitsbereich abgedeckt werden. Dies ist im Hinblick auf die Möglichkeit, die Testergebnisse valide in Bezug auf das theoretische Framework zu interpretieren, unerlässlich. So sollen die Testergebnisse im Sinne einer erfolgreichen Teilhabe an der Gesellschaft interpretiert werden, weshalb nicht nur die Bearbeitung von Aufgaben aus dem Freizeitbereich, sondern auch Ergebnisse aus bildungsbezogenen und beruflichen Aufgaben in den Testwert eingehen. Auf diese Weise kann für den Kompetenzbereich ICT-Skills, der vor allem in Bezug auf Alltagsanforderungen und nicht curricular definiert ist, der Inhalt der Items abgeleitet werden und auf dieser Basis eine Interpretation der Testwerte erfolgen.

Das entwickelte ICT-Testinstrument umfasst 64 simulationsbasierte Items. Nach der abschließenden Skalierung können Ergebnisinterpretationen auf Ebene der globalen ICT-Skala, aber auch differenziert nach kognitiver Operation vorgenommen werden. Die Reliabilität der ICT-Skala liegt bei .80 und ist als gut zu bewerten. Die Reliabilitäten der Subskalen fallen niedriger aus, sind aber für Aussagen auf Populationsebene als ausreichend einzustufen. Items, die für den endgültigen ICT-Itempool nicht ausgewählt werden konnten, wurden bereits auf Möglichkeiten der Überarbeitung geprüft. Es zeigte sich, dass auch diese Items prinzipiell für ein neues Verfahren überarbeitet und wieder zum Einsatz kommen können. Beispielsweise ermöglichen verhaltensbasierte Items durch die detaillierte Verfolgung von Bearbeitungswegen die Identifikation alternativer Lösungen und deren nachträgliche Umbewertung. Möglichkeiten der Itemmanipulation wurden im Projektkontext im Rahmen des Validierungskonzepts bereits eruiert. Die Items wurden so entwickelt, dass alle enthaltenen Textteile editierbar sind. Dies führt zu dem Vorteil, dass die Items für andere Altersgruppen angepasst oder sogar in andere Sprachen übersetzt werden können. Die entwickelten Aufgaben sind also auch für andere Studienkontexte weiterverwendbar und können mit geringem Arbeitsaufwand sogar für internationale Studien nutzbar gemacht werden. Um die Validität der testwertbezogenen Interpretationen überprüfen zu können, wurden mehrere Strategien verfolgt. Wie aus dem in Kapitel 4.2.3 beschriebenen Studiendesign hervorgeht, wurden nicht nur verwandte Domänen wie Problemlösefähigkeit, Lesekompetenz, technisches Wissen und kognitive Grundfähigkeiten erhoben, um den Zusammenhang mit dem Konstrukt ICT-Skills zu prüfen, sondern es wurden auch gezielt Items manipuliert, um die Konstruktrepräsentation zu untersuchen. Zudem bietet die Datenlage auch die Verknüpfung von ICT-Skills zu Selbstberichten der Nutzungshäufigkeit, des Selbstkonzeptes, des Interesses und der Einstellungen, jeweils bezogen auf den ICT-Bereich.

5.2 Ausblick

Bei einer Domäne, die sich ständig weiterentwickelt und verändert, stellt sich natürlich die Frage, wie schnell ein entwickeltes Instrument veraltet. Hier muss zunächst zwischen dem entwickelten theoretischen Framework und den tatsächlichen Items unterschieden werden.

Das theoretische Framework ist nicht an spezifische Inhalte gebunden, sondern beschreibt Anforderungen, die bei der Nutzung von Technologien auf kognitiver Ebene entstehen. Wird z. B. eine Internetsuche ausgeführt, ist der Zugriff auf Informationen nicht völlig abhängig von dem Medium, das sich zweifelsohne ständig weiterentwickelt. Ob eine Suche auf dem Computer oder mit dem Smartphone durchgeführt oder vielleicht sogar in das Telefon gesprochen und nicht eingegeben wird, sollte nichts an der Schwierigkeit ändern, die passenden Suchbegriffe auszuwählen. Es ist nicht auszuschließen, dass das Framework künftig möglicherweise ergänzt und erweitert werden muss, so wie schon dieses Framework eine Synthese aus verschiedenen Frameworks darstellt und beispielsweise die kognitiven Operationen (International ICT Literacy Panel, 2002) um eine kollektive Facette und die Unterscheidung verschiedener Modalitäten ergänzt. Das theoretische Framework enthält bereits die Unterscheidung zwischen auditiv und visuell, während das Assessmentframework keine auditive Facette enthält, da diese eine eher zukunftsgerichtete Facette darstellt. Die einzelnen Items sind so konzipiert, dass sie nicht an aktuelle Computerprogramme, Betriebssysteme oder deren Versionen gebunden sind. Ein Update einer bestimmten gängigen Software führt damit nicht dazu, dass die Items nicht mehr eingesetzt werden können. Daher stellt eine gewisse Abstraktion von Items gegenüber einer tatsächlichen Software nicht nur einen Vorteil hinsichtlich der Testfairness dar, sondern schützt auch vor einer schnellen Alterung der Items. Dennoch muss die Itemmenge künftig unter Umständen an Änderungen der Nutzungsgewohnheiten angepasst werden. Die stärkere Internetnutzung über Smartphones und die hieraus entstehenden spezifischen Herausforderungen können beispielsweise in neueren Itementwicklungen ein verstärktes Gewicht eingeräumt bekommen. Der vorliegende ICT-Itempool besteht aus 64 Items, was zu einer Testzeit von fast zwei Stunden führen würde. Um diesem Problem zu begegnen, wurde eine verkürzte Skala mit einem Umfang von 25 ICT-Items zusammengestellt. Ziel der Testzusammenstellung war es, bei minimaler Testlänge das theoretische Framework zur Beschreibung von ICT-Skills möglichst umfassend abzubilden, die Vielfältigkeit üblicher Computer-Applikationen darzustellen und das Schwierigkeitsspektrum von ICT-Items angemessen zu repräsentieren. Der so erstellte Test beansprucht etwa 50 Minuten Testzeit und kann beispielsweise in computerbasierten Messungen als Kontrollvariable zum Einsatz kommen. Zudem zeigen Monte-Carlo-Simulationsstudien, dass er sich auch als reliables Instrument zur Individualdiagnostik eignet. Als weitere Möglichkeit, die Testreliabilität zu erhöhen und zudem die benötigte Testzeit noch weiter zu verringern, wird unter Nutzung des 64 Items umfassenden Pools ein adaptives Instrument entwickelt. In Monte-Carlo-Simulationen wurde bereits ein entsprechender adaptiver Algorithmus spezifiziert, der in einem weiteren Schritt empirisch erprobt werden kann.

Zusammenfassend wurde mit dem Projekt CavE-ICT das Konstrukt ICT-Skills theoretisch spezifiziert, dessen Überlappung mit anderen Konstrukten geklärt und in einer internationalen Standards entsprechenden Testentwicklung ein computerbasiertes, Simulationen nutzendes, innovatives Testverfahren entwickelt, erprobt und kalibriert. Mit dem resultierenden Test liegt nun ein Verfahren vor, das zur zeitgemäßen, theoriebasierten und präzisen Messung von ICT-Skills bei Large-Scale-Assessments und anderen Studien eingesetzt werden kann.

Literaturverzeichnis

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31,* 162–172.

Artelt, C., Weinert, S. & Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) – Editorial. *Journal for Educational Research Online,* 5, 5–14.

Bos, W., Eickelmann, B., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K., Senkbeil, M., Schulz-Zander, R. & Wendt, H. (2014). *ICILS 2013. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern in der 8. Jahrgangsstufe im internationalen Vergleich.* Münster: Waxmann.

Davier, M. von & Sinharay, S. (2009). Analytics in International Large-Scale Assessments: Item Response Theory and Population Models. In L. Rutkowski, M. von Davier & D. Rutkowski (Hrsg.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis.* New York, NY: CRC Press.

Engelhardt, L., Naumann, J., Goldhammer, F., Horz, H., Hartig, K., Frey, A. & Wenzel, S. F. C. (2016). *A Framework for the assessment of ICT literacy.* Zur Veröffentlichung eingereicht.

Frey, A., Hartig, J. & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28,* 39–53.

Greiff, S. & Funke, J. (2009). Measuring Complex Problem Solving – The MicroDYN approach. In F. Scheuermann (Hrsg.), *The transition to computer-based assessment – lessons learned from large-scale surveys and implications for testing.* Luxembourg: Office for Official Publications of the European Communities.

International ICT Literacy Panel (2002). *Digital Transformation: A Framework for ICT Literacy.* Princeton, NJ. Abgerufen von http://www.ets.org/research/policy_research_reports/publications/report/2002/cjik.

Klieme, E. (2004). Was sind und wie misst man Kompetenzen? *Pädagogik, 56 (6),* 10–13.

Organisation for Economic Co-operation and Development (2011). *PISA 2009 Assessment Framework. Key competencies in reading, mathematics, and science.* Paris: OECD.

Organisation for Economic Co-operation and Development (2010). *PISA Computer-Based Assessment of Student Skills in Science.* Paris: OECD.

Organisation for Economic Co-operation and Development (2012). *Literacy, Numeracy and Problem Solving in Technology-rich Environments: Framework for the OECD Survey of Adult Skills.* Paris: OECD.

Organisation for Economic Co-operation and Development (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy.* Paris: OECD.

Osterlind, S. J. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats.* Boston, MA: Kluwer Academic.

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning* (Bd. 161). Sage Publications.

Prüfer, P. & Rexroth, M. (1996). *Verfahren zur Evaluation von Survey-Fragen: Ein Überblick*. Mannheim: ZUMA.

Richter, T., Naumann, J. & Horz, H. (2010). Das Inventar zur Computerbildung (revidierte Fassung). *Zeitschrift für Pädagogische Psychologie, 24,* 23–37.

Rölke, H. (2012). The ItemBuilder: A Graphical Authoring System for Complex Item Development. In T. Bastiaens & G. Marks (Hrsg.), *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2012* (S. 344–353). Chesapeake, VA: Association for the Advancement of Computing in Education.

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2007). *Lesegeschwindigkeits- und -verständnistest für die Klassenstufen 6–12.* Göttingen: Hogrefe.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring.* München: Wolters Kluwer.

van der Linden, W. J. & Hambleton, R. K. (Hrsg.). (2013). *Handbook of modern item response theory.* New York: Springer Science & Business Media.

Wilhelm, O., Schroeders, U. & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe*. Göttingen: Hogrefe.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31,* 114–128.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER ConQuest 2.0: General item response modelling software* [computer program manual]. Camberwell, Australia: Australian Council for Educational Research.

Autorinnen und Autoren

Cathrin Becker (Universität Koblenz-Landau)

Michael Becker (Universität Potsdam)

Christiane Bertram (Eberhard Karls Universität Tübingen)

Bodo von Borries (Universität Hamburg)

Wilfried Bos (Institut für Schulentwicklungsforschung, TU Dortmund)

Nicola Brauch (Ruhr-Universität Bochum)

Martin Dickmann (Universität Duisburg-Essen)

Timo Ehmke (Leuphana Universität Lüneburg)

Bodo Eickhorst (Universität Bremen)

Lena Engelhardt (Deutsches Institut für Internationale Pädagogische Forschung [DIPF])

Martin Fresow (Universität Würzburg)

Paulina Fresow (Universität Würzburg)

Andreas Frey (Friedrich-Schiller-Universität Jena)

Miriam M. Gebauer (Institut für Schulentwicklungsforschung, TU Dortmund)

Frank Goldhammer (Deutsches Institut für Internationale Pädagogische Forschung [DIPF])

Richard Göllner (Eberhard Karls Universität Tübingen)

Samuel Greiff (Universität Luxemburg)

Inga Hahn (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik [IPN])

Katja Hartig (Goethe-Universität Frankfurt)

Matthias Hirsch (Katholische Universität Eichstätt-Ingolstadt)

Holger Horz (Goethe-Universität Frankfurt)

Kathrin Klausmeier (Ruhr-Universität Bochum)

Eckhard Klieme (Deutsches Institut für Internationale Pädagogische Forschung [DIPF])

Andreas Körber (Universität Hamburg)

Kathrin Kuchta (Goethe-Universität Frankfurt)

Christoph Kühberger (Pädagogische Hochschule Salzburg Stefan Zweig)

Oliver Lüdtke (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik [IPN] und Zentrum für internationale Bildungsvergleichsstudien [ZIB])

Nele McElvany (Institut für Schulentwicklungsforschung, TU Dortmund)

Martin Merkt (Leibniz-Institut für Wissensmedien)

Johannes Meyer-Hamme (Universität Paderborn)

Benjamin Nagengast (Eberhard Karls Universität Tübingen)

Gabriel Nagy (Eberhard Karls Universität Tübingen)

Johannes Naumann (Goethe-Universität Frankfurt)

Knut Neumann (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel)

Herbert Neureiter (Pädagogische Hochschule Salzburg Stefan Zweig)

Christoph Niepel (Universität Luxemburg)

Heinz Reinders (Universität Würzburg)

Norman Rose (Eberhard Karls Universität Tübingen)

Julia Rudolph (Universität Luxemburg)

Horst Schecker (Universität Bremen)

Wolfgang Schnotz (Universität Koblenz-Landau)

Katrin Schöps (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik [IPN])

Waltraud Schreiber (Katholische Hochschule Eichstätt-Ingolstadt)

Franziska Schwabe (Institut für Schulentwicklungsforschung, TU Dortmund)

Stephan Schwan (Leibniz-Institut für Wissensmedien, Tübingen)

Heike Theyßen (Universität Duisburg-Essen)

Ulrich Trautwein (Eberhard Karls Universität Tübingen)

Ann-Katrin Van den Ham (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik [IPN])

Wolfgang Wagner (Eberhard Karls Universität Tübingen)

Helene Wagner (Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik [IPN])

Monika Waldis (Pädagogische Hochschule der Fachhochschule Nordwestschweiz)

S. Franziska C. Wenzel (Goethe-Universität Frankfurt)

Michael Werner (Katholische Hochschule Eichstätt-Ingolstadt)

Béatrice Ziegler (Pädagogische Hochschule der Fachhochschule Nordwestschweiz)

Andreas Zuckowski (Universität Hamburg)

Impressum

Herausgeber
Bundesministerium für Bildung und Forschung (BMBF)
Referat Bildungsforschung
11055 Berlin

Bestellungen
schriftlich an
Publikationsversand der Bundesregierung
Postfach 48 10 09
18132 Rostock
E-Mail: publikationen@bundesregierung.de
Internet: http://www.bmbf.de
oder per
Tel.: 030 18 272 272 1
Fax: 030 18 10 272 272 1

Stand
August 2016

Druck
Silber Druck oHG, Niestetal

Gestaltung
W. Bertelsmann Verlag, Bielefeld

Text
siehe Autorenverzeichnis

ISBN
978-3-88135-335-9

www.bmbf.de