Lesekompetenzmessung und ihre Herausforderungen in der Testkonstruktion für Large-Scale-Assessments

Inauguraldissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

lic. phil. hist., dipl. HLA Karin Gehrer

geboren in St. Gallen, Schweiz

Bamberg, den 02.01.2017

Tag der mündlichen Prüfung: 10.02.2017

Dekan: Universitätsprofessor Dr. Stefan Hörmann

Erstgutachterin: Universitätsprofessorin Dr. Cordula Artelt

Zweitgutachterin: Universitätsprofessorin Dr. Sabine Weinert

Zusammenfassung

Gegenstand der vorliegenden Arbeit ist die Erfassung von Lesekompetenz (z.B. Artelt, Stanat, Schneider & Schiefele, 2001; Bos, Valtin, Voss, Hornberg & Lankes, 2007; Drechsel, 2010; Groeben & Hurrelmann, 2006; OECD, 1999) und verschiedene spezielle Aspekte, welche mit der Lesekompetenzmessung und ihren Herausforderungen für Large-Scale-Assements verbunden sind. Insbesondere wird der Testkonstruktion zur Lesekompetenzerfassung über die Lebensspanne hinweg und ihren besonderen Anforderungen Beachtung geschenkt. Ein zweiter Schwerpunkt ist der schwierigkeitsangemessenen Kompetenzmessung von (hoch)kompetenten Personen in akademisierten Kontexten gewidmet.

In diesem Zusammenhang wird als erstes die Auswahl geeigneter Stimulus- und Lesetexte ausführlich betrachtet (Artikel 1 und 2, sowie Artikel 4), da diese als Ausgangsbasis eines Testinstrumentes zur Erfassung von Lesefähigkeit eine tragende Rolle spielen (z. B. Augst & Pohl, 2007; Nold & Willenberg, 2007). Die Auswahl von Textsorten, welche über die Lebensspanne und damit über verschiedene Lesealter hinweg eine Verankerung über eine gemeinsame Rahmenkonzeption ermöglicht (Artikel 1 und 2), scheint im Spannungsfeld zu stehen zu einer Auswahl von Textsorten zur Messung von Lese- und weiteren Sprachkompetenzen von (hoch)kompetenten Personen bzw. im akademisierten Kontext (Artikel 4). Mit den Fragen der Testkonstruktion einhergehend, werden auch die kognitiven Anforderungen des Leseprozesses verschiedener Textsorten (Artikel 1; z. B. Kintsch 1994) sowie der Testaufgaben oder -fragen (Items; Artikel 2 und 3) diskutiert. Die Aufgabenformate, welche ein weiteres bestimmendes Merkmal eines validen und reliablen Testinstrumentes sind (u.a. Rost, 2004; Kubinger, 2009), werden bezüglich ihres schwierigkeitsgenerierenden Anteils diskutiert (Artikel 3). Verschiedene Arten von geschlossenen Formaten (Multiple-Choice, Entscheidungstabellen, Zuordnungs-aufgaben) werden empirisch unter einer experimentellen Kontextbedingung in einer

computerbasierten Entwicklungsstudie für (junge) Erwachsene und Studierende untersucht (Artikel 3). Demgegenüber steht die Verwendung auch von offenen Formaten (Kurz- und Langantworten) für (hoch)kompetente Personen im akademisierten Kontext (Artikel 4). Kausalen Einflüssen auf die Lesefähigkeiten von Personen mit unterschiedlicher Familiensprache bzw. unterschiedlicher Sprachsozialisation im (schweizer-)deutschen Sprachraum wird abschließend Aufmerksamkeit geschenkt (Artikel 4).

Der erste Artikel widmet sich den besonderen Merkmalen unterschiedlicher Textsorten, nimmt vor dem theoretischen Hintergrund der kognitiven Anforderungen von Leseprozessen eine Einordnung und Definition ausgewählter Textsorten vor und skizziert deren Möglichkeiten und Limitationen für eine Verwendung in der Messung von Lesekompetenzen. Die konkrete Verwendung der Textsorten (literarischer Text, Sachtext, kommentierender Text, Anleitung und Werbung) in den für das Nationale Bildungspanel (NEPS; Blossfeld, Roßbach & von Maurice, 2011) entwickelten Lesekompetenztests wird im zweiten Beitrag wieder aufgegriffen.

Der zweite Artikel stellt insbesondere die Rahmenkonzeption der längsschnittlichen Kompetenzerfassung in NEPS (Weinert et al., 2011) für die Domäne Lesen dar und nimmt eine theoretische Verortung und Abgrenzung gegenüber anderen wichtigen nationalen und internationalen Large-Scale-Assessments wie PISA, IALS oder DESI vor. Für die Textsorten bedeutet dies insbesondere eine nähere Betrachtung des Merkmals kontinuierlicher versus diskontinuierlicher Text. Empirische Ergebnisse zur Dimensionalität einer Entwicklungsstudie zur Konstruktion eines validen und reliablen Lesekompetenztests für die erste Welle der Haupterhebung für Erwachsene werden diskutiert. Spezieller Fokus dieses empirischen Beitrags ist dabei die Fragestellung, ob Textsorten und kognitive Anforderungen als strukturelle Elemente der Rahmenkonzeption insofern angemessen sind, als dass beide (auch) im Erwachsenenalter erlauben, hinreichend zwischen guten und schwachen LeserInnen zu

differenzieren. Basierend auf einem Vergleich eines eindimensionalen und zwei unterschiedlichen mehrdimensionalen Modellen wird der Frage nachgegangen, ob Textsorten und/oder kognitive Anforderungen der Items separate Dimensionen der Lesekompetenz ausmachen oder die im NEPS gemessene Lesekompetenz – wie intendiert – als eindimensionales Fähigkeitskonstrukt aufgefasst werden kann.

Im dritten Artikel stehen veränderte Kontextbedingungen von Lesekompetenztests im Mittelpunkt. Es wird in einer weiteren Entwicklungsstudie der NEPS-Forschergruppe Bamberg unter anderem eine experimentelle Bedingung eingesetzt, in welcher den Zielpersonen nicht erlaubt wird, wie in der Kontrollbedingung und sonstigen Lesekompetenztests, zu den bereits gelesenen Texten zurückzublättern (Gehrer, Wolter, Koller & Artelt, in Vorbereitung). Innerhalb dieser Studie wird der Fragestellung nachgegangen, ob und wenn ja, welche Aufgaben- und Textmerkmale auf eine Schwierigkeitsveränderung unter der beschriebenen Kontextbedingung Einfluss nehmen.

Es kann das Aufgabenformat als beeinflussender Faktor identifiziert werden. Insbesondere das kognitiv anspruchsvolle Format der Zuordnungsaufgabe, bei der Überschriften zu Passagen des Textes zugeordnet werden müssen, erweist sich als Einflussfaktor für Schwierigkeit (Artikel 3).

Im vierten Artikel im Rahmen der Forschergruppe um den schweizerischen EVAMAR II-Erstsprachetest für MaturandInnen (Eberle et al., 2008) wird die Fragestellung untersucht, ob die Sozialisation in unterschiedlichen deutschsprachigen Familiensprachen (Dialekt versus Standarddeutsch) einen Einfluss auf das spätere Abschneiden in einem sprachlichen Kompetenztest hat. Es erweist sich, dass die Performanz in unterschiedlichen Sprachgruppen nicht signifikant voneinander abweicht. Der muttersprachliche Dialekt im Schweizerdeutschen scheint sich unter Kontrolle des sozioökonomischen Status weder hinderlich noch förderlich auf die späteren Sprachleistungen im voruniversitären Bereich auszuwirken. Der Ansatz wird diskutiert

vor dem Hintergrund der sprachwissenschaftlichen Plurizentrik-Debatte für den deutschsprachigen Sprachraum (z. B. Ammon, 1995).

Verzeichnis der Dissertationsschriften

1. Gehrer, K. & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In C. Rosebrock & A. Bertschi-Kaufmann (Hrsg.), *Literalität erfassen: bildungspolitisch, kulturell, individuell* (S. 168–187). Weinheim: Beltz Juventa.

2. Gehrer, K., Zimmermann, S., Artelt, C. & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online, 5* (2), 50–79.

3. Gehrer, K. (2017). *Der Einfluss von Aufgaben- und Textmerkmalen auf Itemschwierigkeit – Eine experimentelle Studie zur Einschränkung der wiederholten Texteinsicht bei der Bearbeitung von Lesekompetenztestaufgaben* (NEPS Working Paper No. 67). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

4. Gehrer, K., Oepke, M. & Eberle, F. (in press). Der EVAMAR II-Deutschtest für GymnasiastInnen – Implikationen für die Plurizentrik-Debatte? In W. Davies, A. Häcki Buhofer, R. Schmidlin, M. Wagner & E. Wyss (Hrsg.), *Standardsprache zwischen Norm und Praxis. Theoretische Betrachtungen, empirische Studien und sprachdidaktische Ausblicke.* (Basler Studien zur deutschen Sprache und Literatur, Band 99). Tübingen: Francke Verlag.

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich auf dem Weg zur Promotion begleitet und unterstützt haben.

Mein besonderer Dank gilt Frau Prof. Cordula Artelt für die langen Jahre der Begleitung und fachlichen Betreuung meiner Arbeit im Gesamten sowie einiger meiner Artikel im Speziellen. Ohne ihre Geduld und interdisziplinäre Offenheit für eine Germanistin in ihrem Lesenteam sowie ihrem Promotionsfach „Empirische Bildungsforschung“ wäre ich heute nicht hier, wo ich bin.

Frau Prof. Dr. Sabine Weinert gilt mein besonderer Dank für ihr Vertrauen in meinen Beitrag zur Lesetestentwicklung der NEPS-Kompetenzsäule (damals noch unter ihrer wissenschaftlichen Leitung), das angenehme Zusammenarbeiten für den gemeinsamen JERO-Artikel und ihr langjähriges Interesse an meiner Person und Arbeit sowie für die Erstellung des Zweitgutachtens in einem äußerst knappen Zeitrahmen.

Herrn Prof. Claus H. Carstensen danke ich für seine unterstützende Begleitung, freundlichen Ermutigungen und sein stets offenes Ohr in der NEPS-Forschergruppe. Die Arbeit auch unter seiner methodischen und fachlichen Leitung bot mir eine bereichernde und jederzeit spannende Lernumwelt.

Danke all meinen vielen Kolleginnen und Kollegen im Bamberger NEPS-Arbeitsbereich Kompetenzentwicklung für ihre Unterstützung und die langjährige angenehme Zusammenarbeit, auch denen, die inzwischen nicht mehr da sind. Es war und ist eine bereichernde intensive Zeit, die ich nicht missen möchte, und welche die Grundlage bildet für die wissenschaftliche Arbeit, die ich hier vorlege. Das gemeinsame Arbeiten und Nachdenken über Kompetenzmessung und all ihre Bedingungen, Instrumente, Forschungsfragen, Methoden und Veröffentlichungen innerhalb unseres Arbeitsbereiches war stets bereichernd, manchmal herausfordernd, und stetig fördernd.

Innerhalb des kleinen Lesenteams möchte ich insbesondere Dr. Ilka Wolter und M. Sc. Stefan Zimmermann für den wertvollen Austausch und das engagierte gemeinsame Entwickeln der Lesekompetenztests „vom Text zu den Items bis ins Feld“ danken. Dr. habil. Kathrin Lockl, Dr. Ilka Wolter und Dipl.-Päd. Lena Nusser gaben mir in der abschließenden Phase wertvolle Hinweise. Dr. Shally Novita möchte ich für ihre kollegiale Unterstützung danken, die mir einen intensiven Endspurt in der Dissertationsphase ermöglichte. Danke dafür, dass ich in der letzten Phase zeitweise Aufgaben abgeben durfte, um mir zwischendurch kleine Freiräume des Schreibens einrichten zu können.

Herrn Prof. Franz Eberle und Frau Dr. Maren Oepke, Universität Zürich, möchte ich danken, dass sie mich vor über zehn Jahren für die Lesen-Testentwicklung und die wissenschaftliche Projektarbeit gewinnen und begeistern konnten. Dem Rat von Franz Eberle, dass Gymnasialunterricht immer wieder möglich sei, es aber ein Projekt wie EVAMAR II nur einmal gebe, bin ich gefolgt und habe dem Weg folgend entdeckt, dass sich daran anschließend weitere interessante (Projekt-)Türen öffnen. Danke dir, Maren, dass wir auch Jahre später nochmals gut zusammen schreiben und publizieren konnten.

Mein persönlichster und nichtwissenschaftlicher Dank gebührt meinen Eltern Marlis und Roland, deren Wertschätzung und Toleranz meinen akademischen Weg in Deutschland befördert haben. Ohne sie wäre ich heute nicht, wo und auch nicht, wie ich bin. Meiner Lebensgefährtin danke ich für ihre Begleitung und liebevolle Unterstützung in der stressigen Zeit sowie ihre stetige Ruhe und Zuversicht. Meinen Eltern danke ich zudem für die Nachsicht, dass sie mich in der anstrengendsten Phase ihrer Geschäftsauflösung mehr entbehren mussten, als ihnen lieb war. Die Freude über den späten Doktorinnentitel für ein beinahe klassisches „katholisches Mädchen vom Land“ teilen sie mit mir.

Inhalt

1 Einleitung

Die Erhebung der Daten, die für die Schriften 1, 2 und 3 die Grundlage bildeten bzw. verwendet wurden, erfolgte im Rahmen des Nationalen Bildungspanels (NEPS), welches seit 2009 repräsentativ für Deutschland individuelle Bildungsverläufe über die Lebensspanne verfolgt, in der Zeit, als die Autorin als wissenschaftliche Mitarbeiterin in dem Drittmittelprojekt (Leitung: Herr Prof. Dr. H.-P. Blossfeld), bzw. in der Phase nach der Institutionalisierung am Leibniz Institut für Bildungsverläufe e.V. in Bamberg (Leitung: Herr Prof. Dr. Roßbach), beschäftigt war. Die Autorin war an der Planung, Koordination und Durchführung der Studien beteiligt und wirkte in diesem Rahmen aktiv an der Konzeption, Entwicklung und Erprobung der eingesetzten Instrumente zur längsschnittlichen Erfassung der Lesekompetenz über die Lebensspanne sowie bei der Durchführung mehrerer Pilotierungs- und mehreren Hauptstudien in unterschiedlichen Alters- und Startkohorten mit.

Die Erhebung der Daten für die Schrift Nr. 4 erfolgte im Rahmen des schweizerischen nationalen Projektes Evamar II (Leitung: Herr Prof. Dr. Franz Eberle), welches an der Schnittstelle Gymnasium – Universität von 2005 bis 2008 die Schweizerische Maturitätsreform (MAR 95) evaluierte. Die Autorin war als wissenschaftliche Mitarbeiterin des Kernteams des Projektes an der Universität Zürich beschäftigt und maßgeblich an der Konzeption, Entwicklung und Konstruktion des Erstsprachetests und seiner Rahmenkonzeption beteiligt.

In der vorliegenden Dissertationsschrift werden zunächst die theoretische und definitorische Einbettung von Lesekompetenz sowie der Forschungsstand zu Leseprozessen eingeführt (Abschnitt 2). Anschließend werden gewisse Aspekte einer Kompetenzmessung im Querschnitt den Vorteilen einer Kompetenzmessung im Längsschnitt gegenübergestellt und auf längsschnittliche Kompetenzmessung in internationalen und nationalen Forschungskontexten eingegangen. Es schließen sich

Überlegungen an zu den besonderen Herausforderungen von Lesekompetenzmessung über die Lebensspanne (Abschnitt 2.1). Der Testung von (hoch)kompetenten Personen im akademisierten Kontext (Abschnitt 2.2) wird im Zusammenhang mit der Hochbegabtenforschung nachgegangen. Damit verknüpft wird die Möglichkeit der Generierung von Aufgabenschwierigkeit in der Testkonstruktion diskutiert, insbesondere über die Itemmerkmale Aufgabenformat (Schriften 3 und 4) und Textsorte (Schriften 1, 2 und 4).

Anschließend erfolgt ein kurzer Überblick über die zentralen Fragestellungen der vier zur Dissertationsschrift eingereichten Artikel (Abschnitt 3). Es wird die Bedeutung der Textsortenauswahl im Zusammenhang mit der NEPS-Lesekompetenzmessung über die Lebensspanne im ersten Artikel diskutiert und die Merkmale unterschiedlicher Textsorten beschrieben. Die gewählten Textsorten als auch die kognitiven Anforderungen der Testitems werden im zweiten Artikel daraufhin überprüft, ob sie separate Dimensionen der erfassten Lesekompetenz darstellen oder nicht. Die Aufgabenformate und ihr Einfluss auf die Itemschwierigkeit unter experimenteller Bedingung ohne wiederholte Texteinsicht werden im dritten Beitrag erforscht. Der Einfluss unterschiedlicher Familiensprachen bzw. unterschiedlicher Sozialisation in die deutsche Standardsprache auf die Leseverstehensleistung von MaturandInnen wird in der vierten Studie untersucht, in welcher der EVAMAR-II-Sprachtest eingesetzt wurde. Abschließend werden die dargestellten Forschungsthemen und die Herausforderungen von Lesekompetenzmessung in Large-Scale-Assessments über die Lebensspanne hinweg zusammenfassend diskutiert und Besonderheiten für schwierigkeitsangemessene reliable und valide Lesekompetenzmessung von (hoch)kompetenten Personen angesprochen (Abschnitt 4).

2 Lesen und Lesekompetenzmessung

Lesekompetenzmessung ist ein weites Feld. Seit den großen internationalen und nationalen Schulleistungsstudien und nicht zuletzt seit den PISA-Resultaten 2000 ist die Lesekompetenz als wichtiger Baustein zur Ermöglichung der Partizipation an Gesellschaft, Kultur, Politik und Bildung verstärkt in die Aufmerksamkeit der wissenschaftlichen, schulischen und bildungspolitischen Öffentlichkeit gerückt. Lesefähigkeit gilt als „universelles Kulturwerkzeug“ (Artelt et al., 2005, S. 5) und als eine der „Schlüsselqualifikationen“ beim Meistern der täglichen Anforderungen moderner Gesellschaften (Weinert, 2001). Sie ist in postindustriellen Systemen beinahe unabdingbar zum Erwerb von Wissen und Kenntnissen in vielfältigsten Bereichen und zum Erreichen vieler Bildungsziele (McElvany, 2008).

Lesekompetenz wird definiert als die über die basalen Lesefertigkeiten des Erkennens von Buchstaben und Buchstabenkombinationen (Graphemen), Silben, Worten und Wortreihenfolgen hinausgehende Kompetenz, Schrift als visuelles Abbild von Sprache inhaltlich zu entziffern mit dem Ziel der „effizienten Sinnentnahme aus unterschiedlichen Textsorten“ (Landerl, 2008, S. 582), bei der die inhaltliche Interpretation des Gelesenen im Sinne eines kompetenten Umgangs mit Texten gelingt (vgl. Artelt et al., 2005; McElvany, 2008; Rost, 2001).

Ein Text wird in der Textlinguistik als sprachliche Information definiert, welche mehr als einen Satz umfasst, doch ist die Frage, welche sprachliche Einheit als Text definiert wird, eher umstritten. Die komplexe Einheit eines Textes wird in mehrere systemische Ebenen differenziert, wovon die wichtigsten die Ebene des Wortes, des Satzes und des Diskurses, also des mündlichen oder schriftlichen Textes, sind. Kleinere und kleinste Einheiten wie Phonem, Morphem, Silbe, Phrase, Satzgefüge, aber auch der Abschnitt, sind weitere Ebenen, welche in ein übergreifendes Betrachten einer linguistischen Textverstehenstheorie eingehen. Eine Verbindung von mindestens zwei Sätzen, welche

semantisch aufeinander bezogen und durch die Intention der Produzentin oder des Produzenten zu einer systemischen Einheit verknüpft sind, wird Diskurs bzw. Text genannt. Die Funktion des Diskurses und seine Einbettung in die kommunikative Situation beschreiben ihn näher und führen zu verschiedenen Formen des Diskurses wie Beschreibung, Bericht oder Erzählung (Strohner, 1990, 79–83).

Theoretische Modellierungen des Textverstehens und der Textbeschreibung entwickelten sich seit den 1930er-Jahren im Rahmen der Lesbarkeitsforschung, der psycholinguistischen Syntaxforschung der 1950er-Jahre und der empirischen Textpsychologie der 1970er-Jahre. Das aus der Linguistik stammende Konzept der propositionalen Einheit als einer aus einem Prädikat (Ereignis, Eigenschaft) und einem Argument (Objekt, Person) bestehenden Bedeutungseinheit, wurde von Kintsch (1974) im propositionalen Textverstehensmodell, das von hierarchiehohen und hierarchieniedrigen Propositionen ausgeht, verankert. Daran anschließende Erweiterungen mit van Dijk betonten die zyklische Verarbeitung (Kintsch & van Dijk, 1978) sowie die Bildung von Makrostrukturen insbesondere bei längeren Texten (van Dijk, 1980). Spätere Forschungsströmungen der Kognitionspsychologie gehen nebst der symbolischen Repräsentation des Textes in Form von Propositionen zusätzlich von einer Ebene der analogen Repräsentation in mentalen Modellen (van Dijk & Kintsch, 1983; Johnson-Laird 1983) aus, welche unabhängig sind von sprachlichen Strukturen (vgl. den historisch-systematisierenden Überblick von Christmann & Groeben, 1996; ausführlich ebenfalls McElvany, 2008). Empirische Evidenz für das Situationsmodell konnte über eine Vielzahl von Experimentalstudien nachgewiesen werden, die sich mit dessen räumlichen, zeitlichen und kausalen Relationen auseinandersetzen (für einen Überblick Dutke, 1998). So konnte bspw. das Updating des Situationsmodelles nicht nur bei räumlichen Veränderungen der ProtagonistInnen experimentell nachgewiesen werden, sondern auch bei Veränderungen der räumlichen Umgebung selber (Wolf, Hasebrook & Rinck, 1999). Dass auch detaillierte zeitliche Informationen in ein Situationsmodell

integriert werden, zeigten experimentell Hähnel und Rinck (1999), dasselbe gilt für mentale Zustände (Barquero, 1999) und emotionale Informationen (Wentura & Nüsing, 1999).

Differenzielle Ergebnisse in Bezug auf die Verwendung des Situationsmodells im Unterschied zu der propositionalen Textrepräsentation fand Dutke (1999) für (hoch)kompetente Personen im Vergleich zu Personen mit durchschnittlicher räumlich-visueller Vorstellungsfähigkeit: Bei Texten mit mehrdeutiger Beschreibung blieb bei durchschnittlich fähigen Personen die propositionale Textrepräsentation leichter verfügbar als bei Personen mit hoher räumlich-visueller Vorstellungsfähigkeit. Letztere bildeten sowohl bei eindeutigen Szenen als auch bei modellerschwerenden Textszenen eher als durchschnittlich begabte Personen ein Situationsmodell und verfügten seltener weiterhin über die propositionale Textbasis. Dies mit dem Risiko, dass keine Modellrekonstruktion mehr vorgenommen werden kann, wenn die propositionale Basis nicht mehr verfügbar bleibt (S. 174).

Insgesamt ist festzustellen, dass die große Relevanz der Fähigkeit, sinnentnehmend zu lesen, und sich so zusätzliche Lerninhalte anzueignen, dazu geführt hat, dass die Erfassung der Lesekompetenz zentraler Bestandteil von nationalen und internationalen Vergleichsstudien (z. B. Bildungsstandards, PISA) geworden ist. Bislang sind längsschnittliche Studien zur Entwicklung und zur Messung von Lesekompetenz über das Schulalter hinaus jedoch rar bzw. umfassen nur relativ kleine Zeiträume. Das Ziel der jeweiligen Studien als auch die zu untersuchende Population hat jeweils großen Einfluss auf die zugrunde liegende Definition der erfassten Lesekompetenz (Jude, Hartig, Schipolowski, Böhme & Stanat, 2013). Für die längsschnittliche Forschung, speziell im Sinne einer Lebensverlaufsforschung, ist dieser Aspekt eine besondere Herausforderung.

2.1 Lesekompetenzmessung über die Lebensspanne

> Auf der Suche nach kausal bedeutsamen Einflussgrößen gelangen selbst raffinierte statistische Auswertungsverfahren an ihre Grenzen, wenn – wie normalerweise bei PISA – die zu erklärenden Bildungsergebnisse nur zu einem Zeitpunkt (in einer Querschnittsstudie also) gemessen wurden. Wenn man empirisch abgesichert wissen möchte, welche Rolle bestimmte Bedingungen für die Kompetenzentwicklung spielen, muss man das internationale Design von PISA (…) um eine Längsschnittkomponente mit zwei Messzeitpunkten (…) erweitern. (Prenzel, 2006, S. 16)

Angesichts prekärer Ergebnisse im Kompetenzbereich – wie die PISA-Resultate 2000 und die weiterhin bestehende Schere zwischen bildungsfernen und bildungsnahen Personen in ihren schulischen Leistungen – wird Erklärungswissen gewünscht und benötigt, denn Beschreibungen allein genügen nicht, um Verbesserungen initiieren zu können (Prenzel, 2006, 15–16).

Die beachtlichen Limitationen einer Querschnittserfassung von Kompetenzdaten lassen sich exemplarisch in Anlehnung an andere bereichsspezifische Kompetenzmessungen wie Mathematik und Naturwissenschaften, deren Relevanz für die postmoderne Gesellschaft analog zur Lesekompetenz ebenfalls international anerkannt ist, wie folgt aufzeigen.

Prenzel (2006) diskutiert für Mathematik und Naturwissenschaften anhand der Pisa 2003 ergänzenden Teilstudie Pisa-I-Plus, welche in Deutschland von 2003 bis 2004 als kurzer Längsschnitt mit zwei Messzeitpunkten vom 9. zum 10. Schuljahr hinweg durchgeführt werden konnte, anschaulich, welche inhaltlichen Vorteile eine, wenn auch begrenzte, Längsschnitterhebung gegenüber einer querschnittlichen Erfassung hat und welche Risiken und Nachteile bei letzterem in Kauf zu nehmen sind. So konnte durch die

Messung am Ende jedes Schuljahres mit konzeptionell gleichen bzw. ähnlichen Tests die erfolgte Kompetenzveränderung nach einem Jahr Schulunterricht in dem jeweiligen Fach belegt werden. Es zeigte sich nahezu dramatisch, dass die erhofften Kompetenzzuwächse im Fach Mathematik aufgrund eines Jahres Fachunterrichtes nur für rund 60 Prozent der SchülerInnen eintrafen. Bei rund 30 Prozent der Schülerinnen und Schüler stagnierten jedoch die Leistungen und bei acht Prozent gingen die Leistungen im Fach Mathematik sogar zurück. Im Fach Naturwissenschaften sah es ähnlich bzw. noch schlimmer aus, indem Leistungszuwächse nur bei 44 Prozent der getesteten SchülerInnen, Leistungsabfall jedoch bei 19 Prozent festgestellt wurden. Zur kausalen Erklärung dieser Befunde in längsschnittlichen Pfadmodellen konnten neben den kognitiven Grundfähigkeiten die fächerübergreifende Problemlösekompetenz als eigenständiger Einflussfaktor für die mathematische und naturwissenschaftliche Kompetenzentwicklung identifiziert werden (Leutner, Fleischer & Wirth, 2006).

Auf die Limitationen und Risiken der Interpretation von Querschnittsdaten gehen auch Senkbeil und Wittwer (2006) im Zusammenhang mit der Erfahrung der SchülerInnen am Computer ein. Nachdem Analysen von Daten aus PISA 2000 beeinträchtigende Effekte von häuslicher Computernutzung auf Kompetenzleistungen nachwiesen, wurde im Widerspruch dazu mit PISA 2003-Daten für Jugendliche mit langer Computererfahrung eine signifikant höhere mathematische Kompetenz berichtet. Diese widersprüchlichen Ergebnisse aus Querschnittdaten und die damit verbundenen Vermutungen über den Einflussfaktor Computer konnten dank der Auswertung der Längsschnittdaten aus PISA-I-Plus empirisch überprüft werden. Es konnte über verschiedene multiple Regressionsmodelle belegt werden, dass es keine Zusammenhänge zwischen der mathematischen Kompetenz und der häuslichen Computernutzung sowie der Computererfahrung der Jugendlichen gibt. Solche vermuteten Zusammenhänge verschwinden, sobald die mathematischen Kompetenzen der früheren Schulstufe ins Analysemodell mitaufgenommen werden können. Wie Prenzel (2006) betont,

verdeutlichen „entsprechende Kontroversen (…) die Grenzen des internationalen PISA-Designs mit einem Messzeitpunkt, Behauptungen über Einflussfaktoren ernsthaft empirisch zu prüfen“ (S. 22). Dies gilt selbstverständlich nicht nur für das PISA-Design, sondern für querschnittliche Designs von Kompetenzmessung im Allgemeinen.

Wie exemplarisch im Rahmen von PISA-Kompetenzmessungen gezeigt werden konnte, gelangen querschnittliche Messungen an die Grenzen ihrer Aussagekraft. Analysen zu unterschiedlichen Messzeitpunkten an unterschiedlichen Stichproben können zu widersprüchlichen Ergebnissen führen. Größter Mangel einer querschnittlichen Messung bleibt jedoch, dass keine Kausalanalysen durchgeführt werden und keine empirisch bestätigten Aussagen zu vermuteten Wirkfaktoren gemacht werden können. Erst eine Längsschnittausrichtung mit mindestens zwei oder drei Messzeitpunkten in sinnvollen Abständen kann kausale Begründungen zu Kompetenzentwicklung und -veränderung anbieten. Auch in der Schulforschung zeigen sich Ergebnisse bezüglich verschiedener Merkmale der Prozessqualität von Schulentwicklung erst in der Längsschnittausrichtung kausal richtig interpretierbar: So zeigen sich erst bei längsschnittlicher Betrachtung teilnehmender Schulen von PISA-E 2000 bis PISA-E 2003 die Faktoren Arbeitsmoral innerhalb des Lehrkörpers und Konsens im Lehrerkollegium als prädiktive Einflussfaktoren für die Testleistung der Schülerschaft, während bei querschnittlicher Betrachtung einzig die durch die Schulleitung festgestellte Überforderung der Schülerschaft in Zusammenhang steht mit der schlechten Testleistung (Klieme & Steinert, 2008).

Während in Deutschland die Stimmen für bildungspolitische Längsschnittforschung um die Jahrtausendwende deutlicher werden und mittels eines BMBF-Gutachtens ein „ernüchterndes Bild“ von der rückständigen Datenlage im Bereich der Bildungslaufbahnen von Kindern, Jugendlichen und jungen Erwachsenen in Deutschland aufgezeigt wird (Kristen et al., 2005), waren in anderen europäischen und internationalen Ländern vielfältige und langjährige Längsschnittstudien im Bildungsbereich etabliert.

Schon Ende der 1950er-Jahre wurde in Großbritannien eine als Vollerhebung angelegte Geburtskohorten-Panelstudie sowohl zur gesundheitlichen Entwicklung als auch zu Bildungsthemen, teilweise mit Leistungsmessung, begonnen und inzwischen mit der siebten Erhebung in 1999/2000 durchgeführt. In der zweiten britischen Geburtskohortenstudie von 1970 wurden u.a. Leistungsmessungen im Bereich Lesen, Sprache, Mathematik und logisches Denken eingesetzt; dieses Panel erstreckt sich inzwischen bereits über 40 Jahre. Methodischer Nachteil sind bei beiden Studien die teilweise großen Zeitabstände der Wiederholungserhebungen (im Alter von 5, 10, 16, 26 und 30 Jahren; Kristen et al., 2005, 7–12). Die britische Millenium Cohort Study (Connelly & Platt, 2014) erweitert die älteren Studien um wichtige soziale Aspekte im Bildungsverlauf der Individuen. Daneben werden vom britischen Bildungsministerium noch weitere Untersuchungen zu den Übergängen vom Sekundarbereich aus in den Arbeitsmarkt durchgeführt, so dass für Großbritannien von einer vielfältigen und gut aufgestellten Forschungslandschaft im Bildungsbereich und einem großen „Erfahrungsschatz“ gesprochen werden kann (Kristen et al., 2005). Mengenmäßig kann nur die USA mit sieben, den Bildungsbereich ansprechenden Längsschnittstudien mithalten, dies seit den 1970er-Jahren. Daneben weisen noch Schweden (seit 1961), Frankreich (seit 1973), die Niederlande (seit 1988) und Kanada (seit 1994) nationale Längsschnittstudien auf, welche sich primär auf den Bildungsbereich ausgerichtet haben (Kristen et al., 2005).

Das erwähnte Gutachten von Kristen et al. (2005) zu den Längsschnittstudien für die Bildungsberichterstattung verweist auf drei generelle Vorzüge von Panelstudien: Neben der möglich werdenden Beschreibung von Veränderungen und Entwicklungen im Zeitverlauf sind die „multivariate Untersuchung von individuellen und gesellschaftlichen Entwicklungs- und Wandlungsprozessen“ (S. 75) und die bereits dargestellte kausale Rekonstruktion von Ereignissen im Sinne von Wirkfaktoren ihre hauptsächlichen Stärken. Dabei soll nicht nur eine Dokumentations- und Erklärungsleistung zu

Unterschieden in den Bildungsverläufen, in Leistungsentwicklungen und weiteren Entwicklungsbereichen emotionaler, sozialer, physischer Art erbracht werden, sondern darüber hinaus stellen Panelstudien auch die Möglichkeit zur Evaluation bildungspolitischer Maßnahmen und zur Evaluation von Rahmenbedingungen und Reformen des Schulsystems dar. Deswegen ist die Empfehlung naheliegend und folgerichtig, dass eine Längsschnittsstudie zur Verfolgung individueller Entwicklungsverläufe alle wichtigen Etappen der Bildungslaufbahn berücksichtigen muss, von der frühen häuslichen Lernumwelt über die Kindergärten und die Grund- und Sekundarschule zu den weiterführenden (Aus-) Bildungswegen bis zum Eintritt in den Arbeitsmarkt, wobei möglichst alle Akteure des Bildungsprozesses befragt und deren Daten auf der Individualebene der SchülerInnen miteinander verknüpfbar sein müssen. Das systematische Zusammenführen von Befragungs- und Registerdaten, wie z. B. in Schweden, oder die adaptive Kombination verschiedener Erhebungen wie in den USA, wird dabei ausdrücklich empfohlen (76–77).

In Deutschland finden sich bis zum Beginn des 21. Jahrhunderts nur zwei große national repräsentative Erhebungen, welche Kompetenzmessung in einem längsschnittlichen Design realisiert haben: die bereits beschriebene Ergänzungsstudie PISA-I-Plus im Rahmen der PISA 2003-Erhebungen und TIMMS 1995, welche beide die Entwicklung von mathematischen und naturwissenschaftlichen Kompetenzen prüfen, wenn auch nur über ein Schuljahr hinweg (von Klasse 9 zu Klasse 10 bzw. Klasse 7 bis Klasse 8). Für den Bereich Lesen, Textverständnis und Sprache finden sich bis 2009 keine national repräsentativen Erhebungen in Deutschland bzw. im deutschsprachig europäischen Raum.

Regional auf deutsche Bundesländer beschränkte, aber dennoch bedeutsame Längsschnittstudien mit Kompetenzmessungen unter anderem auch im sprachlichen Bereich finden sich seit den 1980er-Jahren.

Mit der in den Achtzigerjahren startenden Longitudinalstudie zur Genese individueller Kompetenzen (LOGIK; Weinert, 1998) und der damit verzahnten (mit 120 Kindern überlappenden) Münchner Grundschulstudie „Schulorgansierte Lernangebote und Sozialisation von Talenten, Interessen und Kompetenzen“ (SCHOLASTIK, Weinert & Helmke, 1997), welche ab 1987 die klassenbezogenen Erfahrungen und Leistungen von über 1200 SchülerInnen im ländlichen und städtischen Großraum München untersuchte, sowie der Münchner Hauptschulstudie (vgl. Helmke, Schneider & Weinert, 1986) wurden vom Max-Planck-Institut für psychologische Forschung drei bedeutende Längsschnittstudien betrieben, welche in ihren sehr breit angelegten motivationalen und kognitiven Fragestellungen unter anderem auch den Schriftspracherwerb als Kulturtechnik untersuchten. Als eine der größten Herausforderungen beschreiben Helmke und Weinert (1997) die Entwicklung von Instrumenten, die dem (Grundschul-) Alter und dem Klassenkontext angemessen sind (S. 8). Auch wenn Panelpflege, Rückmeldung an die Lehrkräfte, die Akzeptanz der Unterrichtsbeobachtung und vieles mehr sehr komplex und aufwändig erschienen, so notieren die beteiligten WissenschaftlerInnen abschließend doch, dass „der theoretische Ertrag den großen Aufwand dieser Längsschnittstudie rechtfertigt“ (S. 11). Obwohl sie einschränken, dass bei so breit gewählten Leitfragen der Preis bleibe, dass jede davon nur relativ oberflächlich und explorativ beantwortet werden könne und die Erkenntnisse eher „begründete Hypothesen als gesicherte Erkenntnisse“ enthalten würden (S. 11), bleibt die Studie aufgrund ihrer großen, wenn auch regionalen, Stichprobe und den vielfältigen Fragestellungen sehr bedeutsam.

Im Rahmen der Untersuchung des Schriftspracherwerbs wurde in der SCHOLASTIK-Studie in der Grundschule neben den Wort- und späteren Satzdiktaten in den Klassen 1 bis 2, bzw. 3 bis 4, auch ein selbstentwickelter Leseverständnistest (Näslund, 1990) am Anfang und Ende des zweiten Schuljahres eingesetzt. Dieser besteht aus mehreren Kurzgeschichten und sich darauf beziehenden Multiple-Choice-Aufgaben. Der Test

erwies sich als relativ leicht, zeigte jedoch große interindividuelle Unterschiede auf, welche im zweiten Schuljahr vergleichsweise stabil blieben (Schneider, Stefanek & Dotzler, 1997, 117–118). In der Frage differentieller Leistungsentwicklungen wurden für die Lernfreude und die Intelligenz Haupteffekte gefunden, nicht jedoch für das Geschlecht (dies im Unterschied zu anderen Befunden wie z. B. Klipcera & Klipcera-Gasteiger,1993).

Auch in der erwähnten Münchner Längsschnittstudie LOGIK (Weinert, 1998; Schneider, 2008a), welche bemerkenswerterweise rund 25 Jahre dauerte, wurde der von Näslund (1990) entwickelte Test für Leseverständnis und Lesegeschwindigkeit zu Anfang und Ende der zweiten Klasse an einer Stichprobe von 121 Kindern eingesetzt. Im Rahmen der zweiten Nacherhebung 2004 erhielten die verbleibenden 23-jährigen Erwachsenen wiederum Lesetests (n = 112), einerseits den Lesegeschwindigkeits- und -verständnistest für die Klassen 6-12 (LGVT 6-12; Schneider, Schlagmüller & Ennemoser, 2007) andererseits auch Subtests der PISA-2000 Vorstudien (Schneider, 2008a, S. 173). Über die theoretische Vergleichbarkeit der Instrumente finden sich keine Hinweise, die Stabilität über die Zeit wird als mittelhoch (r = .40) ausgewiesen (Schneider, 2008a, S. 174). Es wurde bei der LOGIK-Studie jedoch zugunsten der Rechtschreibkompetenzen leider die Chance vertan, Lesekompetenz über einen längeren Zeitraum hinweg wiederholt zu erfassen, so dass keine Wachstumskurven und keine Erklärungsmodelle für die Entwicklung unterschiedlicher Lesekompetenzwerte generiert werden konnten (Schneider & Bullock, 2008). Bedauerlicherweise wurde die Entwicklung der Lesekompetenz in dieser beachtenswerten Längsschnittstudie somit nur lückenhaft weiterverfolgt, während die Rechtschreibleistung aufgrund ihrer wesentlichen Rolle beim Übergang in die weiterführenden Schulen über die langen Jahre der Studie bis hinein ins junge Erwachsenenalter besondere Aufmerksamkeit genoss (Schneider, 2008a, S. 171). So konnte die LOGIK-Studie beachtliche empirische Ergebnisse liefern für eine bedeutsame Stabilität der Rechtschreibleistung von der Grundschule bis ins

junge Erwachsenenalter (Schneider & Stefanek, 2007; Schneider, 2008b), sowie für eine Verbesserung der Lesegeschwindigkeit vom ersten zum zweiten Schuljahr (Schneider & Stefanek, 2004).

Die in 2005 startende bedeutsame Längsschnittstudie zu Bildungsprozessen, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (BiKS; von Maurice et al., 2007) umfasst zwei repräsentative Startkohorten (n = 547, n = 2395) aus zwei deutschen Bundesländern (Hessen und Bayern); eine interdisziplinäre Forschergruppe erarbeitete die Rahmenkonzeption der Erhebungen. Dieser Umstand schaffte neue Möglichkeiten und eine umfassende Datengrundlage zur Entwicklung von Kindern ab 3 Jahren. Besondere Berücksichtigung fand in dieser Studie die Entwicklung der (frühen) sprachlichen Kompetenzen und ihre Prädiktion auf die spätere Lesekompetenz. Ebert und Weinert (2013) konnten darlegen, wie die frühen Fähigkeiten des phonologischen Arbeitsgedächtnisses im Alter von vier Jahren die spätere basale Lesefähigkeit (Lesegeschwindigkeit) in der zweiten Klasse positiv beeinflussen, während das Leseverständnis zum gleichen Zeitpunkt von frühen sprachlichen Fähigkeiten wie Wortschatz und Grammatik vorhergesagt wird. Weitere längsschnittliche Analysen zeigten auf, dass über den Übergang von der Grund- zur weiterführenden Schule hinaus freizeitliche Leseaktivitäten einen positiven Effekt auf die Lesekompetenz hatten. Dieser Wirkzusammenhang ist vor allem für Kinder aus Familien mit höherem Bildungshintergrund bedeutsam (Pfost, Dörfler & Artelt, 2010). Die BiKS-Studie war Ausgangspunkt für neue, große Vorhaben im Feld der Bildungsstudien in Deutschland und stellt den inhaltlichen, konzeptionellen und teilweise auch personellen Anknüpfungspunkt für die größte deutschlandweite Bildungsstudie Nationales Bildungspanel (NEPS, Blossfeld, Roßbach & Maurice, 2011) dar, die inzwischen vom Leibniz Institut für Bildungsverläufe e.V. in Bamberg in Kooperation mit einem Exzellenz-Netzwerk durchgeführt wird. Neben der Erfassung relevanter Aspekte zu Bildungsentscheidungen, -renditen, sowie der Lernumwelten liegt ein besonderer Fokus

auf der Kompetenzentwicklung der insgesamt rund 60.000 Teilnehmenden in sechs verschiedenen Alterskohorten. Ab 2010 fanden im NEPS die ersten nationalen Lesekompetenzmessungen in den Haupterhebungen von vier parallelen Startkohorten, repräsentativ für Deutschland, statt (Weinert et al., 2011). So wird nun umgesetzt, was Kristen et al. (2005) in ihrem Gutachten für Deutschland empfohlen und gefordert haben.

Anforderungen an die Lesekompetenzmessung über die Lebensspanne

Nach den beschriebenen Risiken und Nachteilen von Querschnittsmessungen sowie den Vorzügen von Längsschnittmessungen soll der Blick noch weiter gespannt werden zu den Anforderungen einer längsschnittlichen Kompetenzmessung über die Lebensspanne. Für eine kohärente Lesekompetenzmessung über die Lebensspanne ist der Zeitpunkt der ersten Testung bereits eine wichtige Entscheidung und beeinflusst inhaltlich auch das zu messende Konstrukt und damit die zu entwickelnden Instrumente. Vorläuferfertigkeiten wie phonologisches Bewusstsein, Dekodierfähigkeit und -geschwindigkeit aber auch lautes (Vor-)Lesen des ersten Lesealters, welches auch ohne gleichzeitiges Verstehen erfolgen kann, sind von der als textbasierten Sinnentnahme definierten Lesekompetenz abzugrenzen. Auch Wort- und Satzverständnis sind in diesem Sinne von einer Lesekompetenz als Textverständnis zu unterscheiden, wenn sie auch die notwendige Voraussetzung für eine gelingende Sprachrezeption bilden (Lenhard & Artelt, 2009).

Die Herausforderung, Lesekompetenz über die Lebensspanne hinweg zu messen, hat mehrere Aspekte, die bedacht werden müssen.

a) Über mehrere Messzeitpunkte und verschiedene Alterskohorten hinweg muss das theoretische Konstrukt der zu erfassenden Lesekompetenz gleich gehalten werden. Das bedingt eine weitgespannte Rahmenkonzeption, welche die unterschiedlichen Leseanlässe der verschiedenen Alterspunkte im Lebenslauf und deren kontinuierliche Erweiterung, aber auch Transformation,

berücksichtigen und miteinander verknüpfen kann. Querschnittlich gängige Situationsfelder von Leseanlässen und Sprachgelegenheiten bieten keine Kohärenz über verschiedene Kohorten und die Lebensspanne. Somit sind schul- und curriculumsnahe Leseverständnisaufgaben für die Operationalisierung ebenso wenig geeignet, wie berufsnahe oder studienspezifische Leseanforderungen. Für die Gewährleistung eines einheitlichen theoretischen Konstruktes müssen infolgedessen als Grundlage Textsorten gewählt werden, welche sowohl im Kinder- und Jugendlichen- als auch im frühen und späteren Erwachsenenalter gelesen und rezipiert werden (können). Dies bedingt beispielsweise, dass ironische oder satirische Texte, welche im Schulkindalter noch nicht verstanden werden, aus einer übergreifenden Rahmenkonzeption ausgeschlossen werden. Bei der Wahl der geeigneten Textsorten für Lesekompetenztests über die Lebensspanne setzt Schrift 1 an.

b) Jedoch sollte die Erfassung des theoretischen Konstruktes der Lesekompetenz über die Lebensspanne die unglaubliche Heterogenität von Leseanforderungen und Lesegelegenheiten in gewisser Weise mitabbilden können. Es ist jedoch unmöglich, den vielfältigsten Leseanforderungen und Sprachgelegenheiten im Laufe des Lebens gerecht zu werden, nicht zuletzt, weil nicht von unbegrenzter Testzeit und Belastbarkeit der ProbandInnen ausgegangen werden kann. So beschränken sich einige Lesestudien auf die „klassischen“ zwei Textgattungen literarische Texte versus Sachtexte im weiteren Sinne (wie IGLU, z.B. Bos et al., 2007; und DESI; Beck & Klieme, 2003; 2007), andere wie PISA realisieren mit einem Multi-Matrix-Design, bei dem nicht jede Zielperson jeden Text erhält, eine größere Textvielfalt (PISA 2000 mit sechs Texttypen, von argumentativ, beschreibend bis narrativ; ab PISA 2009 ergänzt mit elektronischer Textsorte Hypertext, vgl. Naumann, Artelt, Schneider & Stanat, 2010). Neben den Textsorten in Fließtext wurde auch die Klasse der diskontinuierlichen Texte

bedeutungstragend in die PISA-Konzeption mitaufgenommen (vgl. Schaffner, Schiefele, Drechsel & Artelt, 2004), was dem Heterogenitäts- und Literacy-Gedanken entspricht und zu einer erweiterten Auffassung von Lesekompetenz im Sinne von notwendiger Text-Bild-Integration (bzw. Text-Diagramm-Integration, Text-Tabellen-Integration etc.) führt. Andere Large-Scale-Assessments, wie das Nationale Bildungspanel, verzichten auf diese erweiterte Auffassung von Lesekompetenz und konzentrieren sich auf rein kontinuierliche Textsorten, sie erhalten dadurch einen Zugewinn an Präzision in der Messung. Dem Thema Heterogenität versus Homogenität, sowie unterschiedliche Leseanlässe in der Operationalisierung von Lesekompetenz ist in Schrift 2 Raum gegeben.

c) Die geforderte Messäquivalenz über mehrere Messzeitpunkte bedingt in der validen Operationalisierung des theoretischen Konstruktes auch, dass die kognitiven Anforderungen, welche mit den Leseverstehensaufgaben gestellt werden, über die Lebenspanne und die verschiedenen Kohorten hinweg gleich gehalten werden. Auch hier, bei den Lesenverstehensanforderungen, werden von den bisherigen Lesekompetenzstudien leicht unterschiedliche Konzeptionierungen verwendet. So unterscheidet PISA theoretisch fünf kognitive Verstehensanforderungen und weist letztendlich empirisch drei aus (Artelt et al., 2001). DESI modelliert empirisch vier kognitive Anforderungen der Lesetests, nachdem theoretisch sechs Anforderungen angenommen wurden (Willenberg, 2007). PIRLS/IGLU unterscheiden zwar ebenfalls vier, jedoch qualitativ leicht andere Verstehensanforderungen (Bos et al., 2007). Somit muss für eine Messung über die Lebensspanne überlegt und theoretisch begründet werden, an welche Konzeption sich angelehnt wird. Diesem Aspekt widmet sich ebenfalls Schrift 2.

d) Weitere Aspekte, welche bei einer Messung über die Lebensspanne in den verschiedenen Testheften und Instrumenten konstant gehalten werden sollten,

sind alle weiteren Merkmale, welche das theoretische Konstrukt beeinflussen oder davon abweichen könnten. Dabei ist insbesondere auch bei unterschiedlichen Aufgabenformaten methodisch zu überprüfen, inwiefern sie das zu erfassende latente Konstrukt eindimensional abbilden. Bei der Einführung von neuen, innovativen Aufgabenformaten muss in diesem Zusammenhang beachtet werden, ob sie ermöglichen, das Konstrukt weiterhin eindimensional zu erfassen.

e) Bei einer Testung über die Lebensspanne müssen zu den verschiedenen Messzeitpunkten für jede zu messende Altersgruppe neue schwierigkeitsangemessene Testinstrumente entwickelt werden, welche miteinander (über Linkstudien oder Ankeritems) verlinkt werden müssen. Um die heterogenen Altersgruppen im Fähigkeitsbereich sowohl nach unten als auch nach oben besser erfassen zu können, empfehlen sich jeweils zwei oder drei Versionen, welche schwierigkeitsgestuft und mit verbindenden Ankerunits für die gemeinsame Skalierung eingesetzt werden. Dies bedingt einen größeren Itempool. Auch adaptives Testen ist denkbar, ist jedoch für Lesekompetenzaufgaben mit einer Nestung in dargebotenen Stimulustexten äußerst schwierig zu realisieren. Zudem bedingt adaptives Testen einen nochmals größeren Itempool (Dörfler, Golke & Artelt, 2010). Inbesondere um die (hoch)kompetenten Personen beispielsweise in akademischen Kontexten auch noch differenziert erfassen zu können, bedingt es Aufgaben mit erhöhter Schwierigkeit, was in der Umsetzung meist recht anspruchsvoll ist (Schrift 3 und 4).

f) Veränderte Kontextbedingungen können für das zugrunde gelegte theoretische Konstrukt eine Herausforderung bzw. eine empirische Unschärfe bedeuten. Für Gruppentestungen in den institutionellen und Schulzusammenhängen im Kindes- und Jugendlichenalter versus Einzeltestung in Privathaushalten im

Erwachsenenalter kann nur unter Kontrolle sämtlicher standardisierbarer Bedingungen wie Gleichhaltung der Einführung in die Testung, Einzelsitzplatz, Ruhe während der Testung, Umgang mit Pausen, automatisierte Zeitnahme etc. von vergleichbaren Testleistungen ausgegangen werden. Verschärft gilt dies auch für die Umstellung von klassischen Paper & Pencil-Testheften auf elektronische Testung am Computer. Inwiefern das Lesen von Texten und das Bearbeiten von Leseverstehensaufgaben auf Papier oder am Bildschirm gleiche oder leicht unterschiedliche Anforderungen stellen und gegebenenfalls gewisse Personengruppen mit wenig Computererfahrung leicht benachteiligen, müssen ausführliche Mode-Studien zu dieser Fragestellung noch erweisen. Erste Ergebnisse im Nationalen Bildungspanel in Klasse sieben und neun deuten darauf hin, dass bezogen auf die Itemschwierigkeiten sich insbesondere durch die Itemmerkmale Aufgabenformat und Anzahl der Textseiten am Computer schwierigkeitserhöhende Effekte ergeben, jedoch zusätzlich weitere unsystematische aufgabenspezifische Effekte solcher Kontextveränderungen existieren. Eine ausreichende Anzahl messinvarianter Items sollte dennoch eine gemeinsame Verankerung der Metrik auch bei verschiedenen Darbietungsformen ermöglichen (Bürger, Kroehne & Goldhammer, 2016).

g) Um die Qualität der Daten einer längsschnittlichen (Lese-)Kompetenzmessung über die Lebensspanne hinweg standardisiert garantieren zu können, sind neben den oben ausgeführten theoretischen Besonderheiten des Testkonstrukts und dessen Operationalisierung in Form von standardisierten Testinstrumenten auch den operativen pragmatischen Ebenen des Feldeinsatzes, der Datenerhebung und der Datenauswertung besondere Beachtung zu zollen. So ist Panelpflege ein wichtiges Thema, um der gefürchteten Panelmortalität entgegenzuwirken, welche bei systematischen Abbrüchen in speziellen Risikogruppen zu einer Verzerrung der Stichprobe führt und somit die Aussagekraft der Daten einschränkt. In diesem

Zusammenhang ist der zeitlichen Belastung und Motiviertheit der Zielperson bei wiederholten Besuchen Rechnung zu tragen, beispielsweise über eine attraktive Incentivierung. Der Testleiter- und Interviewerschulung ist ausdrückliche Aufmerksamkeit zu widmen, damit die Testsituationen sowohl in den Gruppenerhebungen im Schulkontext als auch in den außerinstitutionellen Erhebungen im Privathaushalt standardisiert und vergleichbar erfolgen. Diese und weitere Faktoren, wie personelle und finanzielle Kontinuität sowohl im Erhebungs-, Organisations- als auch im Entwicklerteam, tragen wesentlich zu einem Gelingen von Erhebungen über die Lebenspanne bei (vgl. z. B. Kristen et al., 2005). Im Rahmen dieser Dissertation wird nicht vertiefend auf diese Aspekte eingegangen.

2.2 Lesekompetenzmessung bei (hoch)kompetenten Personen

Bei der zielgruppenspezifischen Erfassung von (hoch)kompetenten Personengruppen mittels reliabler und valider Testung ist mit besonderen Herausforderungen zu rechnen. So ist dafür ein Test notwendig, der nicht nur im mittleren Bereich, sondern darüber hinaus auch im oberen Bereich der Personenfähigkeiten trennscharf und damit differenziert misst. Ein solcher Test darf also keine oder nur wenige Items mit Deckeneffekten aufweisen. Hier gerät die Testkonstruktion manchmal an die Grenzen des im Voraus Einschätzbaren und Machbaren. Es ist nicht zu leugnen, dass in einem sehr großen Gremium von internationalen ExpertInnen innerhalb des wichtigsten Itempools der letzten Jahre die adäquate Vorhersage von Aufgabenschwierigkeiten nicht annähernd gelungen ist (vgl. PISA; Artelt et al., 2001, 98–101) und dass auch qualifizierte Lehrpersonen, bei welchen Bewerten, Einschätzen und Leistungenvergleichen zu den täglichen Berufsqualifikationen zählen, kaum darin reüssieren, die Schwierigkeiten von Aufgabenstellungen richtig einzuschätzen (vgl. Rausch, Matthäi & Artelt, 2015). Die Frage, auf welche Art und anhand welcher Merkmale Schwierigkeit in der Domäne Lesen vorausgesagt werden kann, beschäftigt

schon mehr als eine Generation von ForscherInnen (vgl. kurzer Überblick bei Zimmermann, 2016).

Mit dieser versuchten Vorhersage verbunden ist die Frage nach gelungener Generierung von angemessener Schwierigkeit für bestimmte Zielgruppen. Beispielsweise kann ein international zu eichender Test in gewissen Ländern für eine bestimmte Altersstufe passend sein, in anderen nicht. So war Deutschland vor der PISA-Misere 2000/2001 bereits 1991 bei der International Study of Reading Literacy, der Lesestudie der International Association for the Evaluation of Educational Achievement (IEA, Elley, 1994), querschnittlich mitbeteiligt. Die Leseaufgaben aus diesem internationalen Pool für die dritte Jahrgangsstufe (75 Minuten Testzeit) und für die achte Jahrgangsstufe (85 Minuten Testzeit) erwiesen sich in den deutschen Pilotstichproben als zu wenig anspruchsvoll für die deutsche Schülerschaft, weshalb für die deutsche Teilstudie (auch bekannt als „Hamburger Lesestudie“) 45 Minuten Testzeit mit anspruchsvolleren Testaufgaben ergänzt wurden. Die Lesetests basierten auf Erzähl-, Sach- und Gebrauchstexten sowie diskontinuierlichen Textteilen und waren meist in Multiple-Choice-Format mit 4 Antwortalternativen, teilweise im offenen Format, vorgegeben. Die Verlinkung über die beiden Jahrgangsstufen war über vier identische Texte mit gemeinsamen Aufgaben gegeben (Lehmann, Peek, Pieper & Stritzky, 1995). Dies ist ein Beispiel, wie für Tests zielgruppenspezifisch passende Items mit höherer Schwierigkeit nachentwickelt bzw. angepasst werden mussten, um insgesamt schwierigkeitspassend zu messen.

Auch in der Hochbegabtenforschung stellen sich die ähnlichen methodischen Herausforderungen, noch verschärft durch die Thematik, dass Personen mit überdurchschnittlich hoher Intelligenz (IQ-Bereich von 130 bzw. 135 bis 200) oder mit überdurchschnittlich hohen Fähigkeiten und Hochleistungen erfasst werden sollen. Um diese hochselektive Stichprobe präzise auswählen zu können, wurden beispielsweise im Marburger Hochbegabtenprojekt (Rost, 2009) mehrere verschiedene

Intelligenztestverfahren und eine gewichtete Kombination aller Testwerte angewandt (Preckel & Vock, 2013; Rost, 2009). Um bei der Kompetenzmessung Hochbegabter Deckeneffekte zu vermeiden, ist es auch üblich, das Akzelerationsmodell anzuwenden, bei dem Aufgabensets von zwei oder drei Jahren älteren SchülerInnen vorgelegt werden (Heller, 1991, S. 285). Das Verfahren wird deshalb auch als off-level-testing bzw. above-level-testing bezeichnet. Nach Heller hat sich dies bspw. für den Kognitiven Fähigkeitstest KFT 4-13+ vergleichsweise gut bewährt. In der amerikanischen Study of Mathematically Precocious Youth (SMPY, Lubinski & Persson Benbow, 2006) wurden bei hochbegabten Jugendlichen ebenfalls niveau- statt altersangepasst die College Board Scholastic Aptitude Tests (SAT)-Tests eingesetzt, d.h. die Hochbegabten lösten Tests, die für eine vier bis fünf Jahre ältere Zielgruppe normiert waren; diese Maßnahme sorgte dafür, dass bei der Messung ihrer Leistung Deckeneffekte vermieden werden konnten (Preckel & Vock, 2013).

So wie sich diese methodischen Herausforderungen in der Hochbegabtenforschung stellen, so ist die Generierung von passender Aufgabenschwierigkeit für alle Zielgruppen wichtig, für die Erfassung von (hoch)kompetenten Personen jedoch in der Testkonstruktion besonders herausfordernd. Um Items besonders schwierig zu gestalten, sind der Literatur unterschiedliche Hinweise zu entnehmen. Im Zusammenhang mit der Lesbarkeitsforschung seit den 1920er-Jahren wurden insbesondere die Schwierigkeitsprädiktoren für Texte erforscht (Mrazek, 1979). Der Frage der Schwierigkeitseinflüsse spezifischer Merkmale von Aufgaben und ihrer Interaktion mit dem Text widmen sich seit Jahrzehnten amerikanische, englische aber auch deutschsprachige ForscherInnen (Gorin & Embretson, 2006; Zimmermann, 2016).

Schrift 3 geht der Frage von Schwierigkeitsgenerierung für Lesekompetenzitems für (hoch)kompetente Erwachsene und Studierende in einer Experimentalbedingung einer NEPS-Teilstudie nach. Schrift 4 beschreibt den nationalen Erstsprachetest von EVAMAR II, welcher unter anderem auch über lange und kurze offene Aufgabenformate die Lese-

und Sprachkompetenzen einer (hoch)kompetenten Personengruppe (MaturandInnen und Studierende) erfasst.

Generierung von passender Aufgabenschwierigkeit

Bei der Nutzung von schwierigkeitsgenerierenden Merkmalen für die Testkonstruktion können wir grundsätzlich die Ebene a) der Aufgabenstellung (Items) und b) des Stimulustextes selber und c) der Interaktion zwischen Item und Text als relevant in Betracht ziehen (Freedle & Kostin, 1994, S. 107). Während in Anlehnung an Freedle und Kostin (1994) sowie Kirsch (2001) in den folgenden Abschnitten die erwähnten drei Kategorien von Prädiktorvariablen ausgeführt werden, gehen andere Studien lediglich von einer Zweiteilung der für Schwierigkeit entscheidenden Merkmale aus, unterscheiden also theoretisch nur zwischen einer textbezogenen und einer item-, antwort- oder lösungsbezogenen Gruppe von Merkmalen (z.B. Embretson & Wetzel, 1987; Schweitzer, 2007; Sonnleitner, 2008).

Es kommen mehrere Faktoren als schwierigkeitsbeeinflussend in Frage: Freedle und Kostin (1993) beziehen beispielsweise 13 Itemvariablen und 34 Textvariablen in die Analysen mit ein. Für die dritte Kategorie von Prädiktorvariablen, welche insbesondere die Verknüpfung von Items und Text, also deren Interaktion erfassen, werden zusätzliche 28 inhaltliche Text-by-Item-Interactions-Variablen definiert, welche hauptsächlich auf semantischen Überschneidungen zwischen der Frage, dem Aufgabenstamm, den Antwortoptionen und den Passagen des Textes beruhen, welche lokal, qualitativ und quantitativ ausgezählt werden. Als signifikant für die Schwierigkeitssteigerung erweisen sich davon acht Merkmale, wovon ein einziges ein reines Item-Merkmal (Verneinung in der Lösung) ist (S. 162), weshalb Freedle und Kostin die Bedeutung der schwierigkeitsbestimmenden Text- sowie Item-Text-Interaktions-Merkmale hervorheben.

Für das Englische erwiesen sich über mehrere Studien hinweg die folgenden vier Prädikatoren innerhalb von bis zu 100 Variablen als die ausschlaggebenden schwierigkeitsbestimmenden Merkmale: a) Wortschatz, b) propositionale Dichte, c) Plausibilität der Distraktoren und d) kognitive Anforderung der Aufgabenstellung (u.a. (Embretson & Wetzel, 1987; Kirsch, 2001, 2001; Nold & Rossa, 2007; zusammenfassend zitiert nach Zimmermann, 2016, 6–10), wobei sich für Studierende über a) Wortschatz und b) propositionale Dichte keine Schwierigkeit generieren ließ (Gorin, 2005). Für das Deutsche gelang Zimmermann (2016) eine Replikation mit Multiple-Choice-Aufgaben (MC) des NEPS-Lesekompetenztests für die neunte Klasse.

Schrift 3 geht vor diesem Hintergrund dem Einfluss von unterschiedlichen Aufgabenformaten (geschlossenen Zuordnungsaufgaben und Entscheidungstabellen nebst MC) und unterschiedlichen Kontextbedingungen auf die Generierung von Itemschwierigkeit für kompetente Personen nach und kann deren Effekt teilweise bestätigen.

Für den MaturandInnen-Test von EVAMAR II (Schrift 4) wurden auch offene Aufgabenformate konstruiert, weshalb im Folgenden auf deren Besonderheiten eingegangen wird. Auch in der Hochbegabtenforschung, welche die Stanford Achievement-Tests einsetzen, haben offene Formate, als Kurzantworten oder erweiterte Antworten, ihre eigene Tradition (Greiten, 2012, S. 38).

Offenes Aufgabenformat

Aus theoretischer Sicht kann von langen offenen Formaten wie beispielsweise einer Zusammenfassung angenommen werden, dass sie eine den ganzen Text umfassende Fähigkeit auf einem „Super-Macro-Level“ (Bensoussan & Kreindler, 1990, S. 57) repräsentiert. Kintsch und Yarbrough (1982) gingen davon aus, dass lange offene Antworten bzw. „open ended questions“ Textverständnis im Sinne der Hauptideen eines Textes umfassender prüfen können als ein kurzes offenes Format wie der Cloze-

Lückentext, der ihrer Ansicht nach lediglich ein lokales Satzverständnis zu prüfen vermag. Auch Kobayashi (2002) findet bei japanischen Studierenden, dass unterschiedliche offene Testformate (Zusammenfassung schreiben, Cloze-Test, Fragen mit offenem Ende) unterschiedliche Aspekte von Textverständnis testen und sie unterschiedlich mit verschiedenen Textsorten korrelieren. Es sind jedoch neben auswertungstechnischen Gründen auch theoretische Überlegungen, welche einige Large-Scale-Assessments dazu bewegen, in ihren Testinstrumenten keine offenen Formate einzusetzen (z.B. NEPS-Lesekompetenztest; Gehrer, Zimmermann, Artelt & Weinert, 2013, S. 64).

Für das Deutsche finden Prenzel, Häußler, Rost und Senkbeil (2002) in der Schwierigkeitsprädiktionsanalyse der internationalen und nationalen PISA-Naturwissenschaftstestaufgaben, die auch das Lesen von Aufgabentexten und nebst geschlossenen Auswahlantworten auch Antworten im offenen Format beinhalten, dass innerhalb von 15 Schwierigkeitsprädiktoren (sowohl formaler, kognitiver als auch wissensbasierter Art) „die aktive Verbalisierung im Rahmen der Aufgabenbeantwortung sehr wohl einen Schwierigkeitsfaktor für die Schülerinnen und Schüler darstellt“ (2002, S. 132). Das offene Format der „langen Antworten“ erhält in ihrer Regressionsanalyse hinter der kognitiven Anforderung „Etwas ausrechnen“ und dem terminologischen Wissen das drittstärkste Regressionsgewicht (B = 0.85) in der Schwierigkeitsvorhersage, die offenen Ein-Wort-Antworten (B = 0.62) rutschen zwar auf den sechsten Platz aller Schwierigkeitsmerkmale, tragen gegenüber den geschlossenen Auswahlantworten jedoch noch deutlich zur Schwierigkeitsvorhersage bei (2002, Tabelle S. 130).

Für den Bereich „Sprache und Sprache untersuchen“ berichten Isaac und Hochweber (2011) für die dritten und vierten Klassen der Vergleichsarbeiten in der Grundschule (VERA 3) bei 107 Aufgaben aus den Subdomänen Semantik, Morphologie und Syntax innerhalb von 9 schwierigkeitsbestimmenden Aufgabenmerkmalen, dass die drei

Merkmale mit den höchsten Gewichten „meist im Rahmen offener Aufgabenstellungen bearbeitet werden“ (S. 194).

Für den Lesetest der vierten Klassenstufen bei IGLU 2006 befinden sich innerhalb der für die höchste Kompetenzstufe V veröffentlichten Items ausschließlich Fragen im offenen Format (Bos et al., 2007, 103–104), daneben werden aber auch auf der Kompetenzstufe III sowie der niedrigen Kompetenzstufe II neben Mehrfachwahlaufgaben teilweise offene Formate ausgewiesen (100–102). Für IGLU 2001 lässt sich anhand der veröffentlichten Aufgaben und der Wrightmap des Tests (Blatt & Voss, 2005, Abbildung S. 241) ablesen, dass sowohl die schwierigste als auch vier der sechs schwierigsten Aufgaben im offenen Format gehalten sind.

Für den Lesetest in Deutsch gibt Willenberg (2007) in DESI ebenfalls für neunte Klassen auf dem höchsten Kompetenzniveau D ein Beispielitem in offenem Format an, welches der Abfrage eines mentalen Modells dient (2007, 114–117). Alternativ stellt er eine zweite Variante der Frage nach dem mentalen Modell vor, welche im geschlossenen Format von richtig oder falsch anzukreuzen ist. Er führt an, dass diese zweite Variante leichter war und begründet dies unter anderem damit, „dass Mehrfachwahlaufgaben mit ihren Vorlage [*sic*] die Antwort leichter machen“ (S. 117).

Für die Oberstufe, Klassen 12 und 13, vermuten Watermann und Klieme (2006) in der Analyse zehn ausgewählter TIMSS-Geometrieaufgaben (fünf mit offenem Antwortformat, sowohl kurze (*short answer*) als auch lange Antworten (*extended response*)), dass „MC-Aufgaben andere kognitive Teilprozesse erfordern als offene Aufgaben, die eine explizite Darlegung des Lösungsweges verlangen“ (S. 331). Eine unerwartet hohe Lösungswahrscheinlichkeit einer Aufgabe der schwierigsten Kompetenzstufe 4 auch in der mittleren Gruppe der Personenfähigkeiten erklären sie damit, dass diese Aufgabe als einzige der Items dieser Kompetenzstufe nicht im offenen Format gehalten ist, sondern dass sie „ein MC-Item darstellt, die selbstständiges Argumentieren in geringerem Maße erfordert“ (S. 332). Auffallend bleibt auch, dass vier der offenen Aufgaben sich auf den

höchsten beiden Kompetenzstufen befinden (neben der erwähnten MC), und die unteren beiden Kompetenzstufen meist durch MC-Items (neben einer offenen Aufgabe) abgedeckt werden (Tabelle S. 332).

Aus den beobachteten Befunden, dass schwierige Aufgaben häufig im offenen Format konstruiert sind (z.B. VERA, IGLU, DESI, TIMMS) und dass u.a. Analysen bei PISA ergaben, dass das Format aufgrund anderer kognitiver Anforderungen ein ernst zu nehmender Schwierigkeitsprädiktor sein kann (Prenzel et al., 2002), können Effekte dieses Aufgabenmerkmales vermutet werden.

Textsorte

Nebst den literarischen und kommentierenden Textsorten, für welche von literaturwissenschaftlicher und linguistischer Seite, aber auch von Seiten der Kognitionspsychologie aufgrund ihrer Mehrdeutigkeit (Kintsch, 1994) Schwierigkeit vermutet wird (vgl. Schrift 1 und Schrift 3), wird auch für das Verwenden von Fachsprache Schwierigkeit vermutet (Prenzel et al., 2002). Für die Klassen 5 bis 10 konnte exemplarisch mit Testaufgaben zu Biologie gezeigt werden, dass Testaufgaben mit Fachbegriffen gegenüber von Testaufgaben mit Alltagssprache signifikant schwerer sind. Es konnte mittels einer Variation dieses Aufgabenmerkmales die Leistungsanforderungen der Testaufgaben hin zu erhöhter Schwierigkeit beeinflusst werden (Schmiemann, 2011). Fach- und Wissenschaftssprache ist einerseits durch einen spezifischen Wortschatz der jeweiligen Fachrichtung geprägt; Experten verwenden häufiger als Laien Begrifflichkeiten mit einem hohen Fachlichkeitsindex. Zudem sind in Fach- und Wissenschaftstexten meist anspruchsvolle Satzbaumuster zu finden, in denen bspw. häufig Nominalisierungen und Passivkonstruktionen vorkommen (Klein, 2003). Beim Eintritt in ein Universitätsstudium bringen aus Sicht der Dozierenden gewisse Erstsemestrige ungenügende sprachliche Kompetenzen mit, die es ihnen erschweren, die geforderten Eingangsanforderungen zu erfüllen (vgl. für die Schweiz: EVAMAR II, Eberle et al., 2008, 45–62). Deshalb ist es für diesen akademisierten Kontext besonders

wichtig, einen zielgruppenspezifischen Test zu entwickeln, der diese Voraussetzungen adäquat abprüft. Der EVAMAR II-Erstsprachetest, der von der Autorin maßgeblich mitentwickelt wurde, erfüllt diese Bedingung, indem als Stimulustexte für die Textverstehensaufgaben ausschließlich wissenschaftliche Fachtexte ausgewählt wurden. Um die Breite der Fachsprachlichkeiten zu gewährleisten, wurden aus den meist besuchten Fachgebieten aller schweizerischen Universitäten mehrere repräsentative Fachrichtungen berücksichtigt und deren originale Vorlesungsmaterialien für Erstsemestrige auf notwendiges Vorwissen hin abgecheckt und feinmaschig kodiert. Innerhalb dieser Kodierungsnetze wurden Textstellen identifiziert, welche gehäufte Fachsprachlichkeit aufwiesen, ohne dass diese explizit erläutert wurden. Solche Textpassagen wurden als wissenschaftliche Stimulustexte dem Erstsprachetest zugrunde gelegt (Eberle et al., 2008, 123–127).

Schrift 4 verwendet Daten aus dem Zusammenhang der schweizerischen Testung an der Schnittstelle von Gymnasium und Universität, um weiterführende Fragen bezüglich der Herkunftssprache und ihrer Auswirkungen auf die späteren Sprachleistungen im (vor)akademischen Kontext zu beantworten.

Adaptives Testen

Zur Messeffizienzsteigerung und Kostenverringerung, aber auch mit dem Ziel einer genaueren Präzision der Personenparameterschätzung, werden seit einigen Jahren standardisierte computerbasierte adaptive Testverfahren vorgeschlagen und deren Vor- und Nachteile diskutiert (z.B. Wainer & Kiely, 1987; Frey & Ehmke, 2007). Es handelt sich dabei um ein Testverfahren, bei dem der Zielperson kein vorgefertigtes Testheft vorgegeben wird, sondern ad hoc während der Testsituation entsprechend den Antworten zu den ersten Testaufgaben die folgenden Aufgaben gemäß dem festgestellten Leistungsniveau der Zielperson durch den Computer schwierigkeitsangemessen zugespielt werden. Die Aufgaben müssen somit in Formaten gehalten sein, die bereits während des Testens direkt von Computern ausgewertet

werden können; offene Formate sind hierzu weniger geeignet. Beantwortet eine Zielperson die ersten Items eines Tests falsch (wobei diese eher auf einem mittleren Schwierigkeitsniveau gehalten werden sollten), werden ihr leichtere Items zugespielt, et vice versa. Je nachdem, ob sie diese richtig oder falsch beantwortet, erhält sie in den folgenden Schritten schwierige oder leichte Items. Jede Zielperson erhält somit die für ihre Leistung adäquaten Aufgabenstellungen, und löst somit wenig bis keine Aufgaben, welche für sie viel zu leicht bzw. zu schwer sind.

Frey und Ehmke (2007) fanden bei einem experimentellen Vergleich innerhalb einer Simulationsstudie, dass die Messeffizienz und die Differenzierungsfähigkeit in den extremen Kompetenzbereichen der adaptiven Test-Version gegenüber der nicht-adaptiven Version stark überlegen sind. Dies gilt für Daten zur Überprüfung der Bildungsstandards in Mathematik für den Mittleren Schulabschluss. Während mit konventionellem Testen hauptsächlich die Personen im mittleren Kompetenzbereich differenziert und mit hoher Präzision, die Personen in den unteren und oberen Extrembereichen jedoch weniger präzise erfasst werden, kann mit Hilfe des computergestützten adaptiven Testens über alle Personen hinweg eine adäquate Messung erreicht werden. Die auf das Individuum hin optimierte Aufgabenauswahl und dadurch erreichte Messpräzisierung könnte es im Bereich Mathematik erlauben, die Testinstrumente bis auf die Hälfte von Items zu reduzieren (Frey & Seitz, 2010).

Einschränkend muss für eine solche Befundlage vermerkt werden, dass die Daten mit optimalen Itempools simuliert wurden, und nicht in Betracht gezogen wurde, dass Items meist in sog. Units vorgegeben werden, also in Einheiten, bei welchen sich Fragen oder Aufgaben auf einen Stimulustext beziehen. Dies gilt teilweise für Mathematik (Frey & Ehmke, 2007; Frey & Seitz, 2010), jedoch besonders für die Domäne Lesen (Gehrer, Zimmermann, Artelt & Weinert, 2012). Darüber hinaus muss, um eine adaptive Messung zu ermöglichen, ein überaus großer Itempool mit optimalen Schwierigkeiten vorliegen. Die Entwicklung und Pilotierung eines solchen Pools ist sehr aufwändig und

kostenintensiv. Dörfler, Golke und Artelt (2010) berechnen für dynamisches Testen im Bereich Lesekompetenz mit Feedback an die Zielperson einen dreifach so großen Itempool gegenüber konventionellem Testen. Für ein adaptives Modell darf mit ähnlichen vergrößerten Pools gerechnet werden.

3 Zentrale Fragestellungen und Befunde der vier Studien

Vor dem Hintergrund der beschriebenen Anforderungen der Lesekompetenzmessung im Längsschnitt einerseits und der zielgruppenspezifischen Passung bei (hoch)kompetenten LeserInnen andererseits sind die vier Studien- und Artikelbeiträge verortet.

Für die zentrale Fragestellung, welche Anforderungen an ein eindimensionales Konstrukt von Lesekompetenzmessung über die Lebenspanne und über heterogene Zielgruppen hinweg sowie auch für die Berücksichtigung von zielgruppenspezifischen Anforderungen an (hoch)kompetente LeserInnen erfüllt werden müssen, lassen sich einige wesentliche Aspekte beschreiben.

Ein erster zentraler Aspekt in der Fragestellung, wie man Lesekompetenz angemessen erfassen kann, sind die verwendeten Textsorten, auf denen das Testinstrument beruht. Je nachdem, ob alltagsnahe Lesekompetenz, curriculums- oder wissenschaftsnahe Lesekompetenz geprüft werden soll, sind andere Konzepte und andere Textsorten notwendig und verwendbar. In Beitrag 1 werden für die Fragestellung eines alltagnahen Lesekompetenz-Konstruktes, wie es im NEPS verwendet wird, die dafür sinnvollerweise einzusetzenden Textsorten (informierender Sachtext, Anleitung, Werbung, kommentierender und literarischer Text) dargestellt. Dabei wurden die den Textsorten immanenten Besonderheiten und Anforderungen herausgearbeitet. Für die Frage, ob trotz heterogener Textsorten ein eindimensionaler Test für Lesekompetenz konstruiert werden konnte, wurden Dimensionalitätsprüfungen vorgenommen. Es zeigte sich, dass alle verwendeten Textsorten in ein eindimensionales Lesekompetenzmodell eingehen können, dass jedoch der literarische Text mit den anderen Textsorten weniger hoch korreliert. Dies verhält sich hypothesenkonform, in Übereinstimmung mit literaturdidaktischen und kognitionspsychologischen Theorien (Kintsch, 1994), welche für

einen literarischen Text andere Rezeptionsanforderungen beschreiben als für einen rein informierenden Text.

Für die Fragestellung im Unterschied zu einer nicht alltagsbezogenen, sondern einer wissenschaftsnahen Überprüfung von Lesekompetenz wird in Artikel 4 der EVAMAR II-Test dargestellt, der für die Schnittstelle von Gymnasium zur Universität auf der Basis der Textsorte wissenschaftlicher Fachtext entwickelt wurde. Innerhalb dieser wissenschaftlichen Textsorte wurde die Breite des Lesekompetenzkonstruktes durch die Heterogenität der Texte unterschiedlicher, repräsentativer Studien- und Fachrichtungen und damit verbundener Varianz von gewählten Themen und Fachsprachlichkeit gewährleistet. Das Lesekompetenzleistungen beeinflussende Vorwissen wurde durch eine kleinmaschige Erfassung und Kodierung des notwendigen gymnasialen Eingangswissens zur Sinnentnahme aus dem authentischen Studien- und Testmaterial kontrolliert.

Ein zweiter zentraler Aspekt der Lesetestkonstruktion ist das für einen Lesetest verwendete Aufgabenformat und die damit verbundenen kognitiven Anforderungen. In Artikel 2 wird für die NEPS-Lesekompetenztests die Herleitung der drei kognitiven Anforderungen Informationenentnehmen, Schlussfolgern sowie Reflektieren und Bewerten mit den drei Aufgabenformaten Multiple-Choice, Entscheidungstabellen und Zuordnungsaufgaben dargestellt. Die Auswertungen eines breit angelegten Itempools zur Testentwicklung ergaben hoch zufriedenstellende Korrelationen der kognitiven Anforderungen von $r = .94$, .96 und .99. Gepaart mit einer modellbasierten Dimensionalitätsprüfung dieses Merkmales kann somit von einem eindeutig eindimensionalen NEPS-Lesekompetenzkonstrukt ausgegangen werden. Im EVAMAR-II-Erstsprachtest für AbiturientInnen bzw. Studierende wurden neben geschlossenen Aufgabenformaten auch offene kurze und lange Antwortformate eingesetzt. Damit wurde auf sprachdidaktische Vorbehalte gegenüber der Messung von Sprachkompetenzen mit rein geschlossenen Antwortformaten eingegangen. Dieser Aspekt geht in Schrift 4 ein.

Ein dritter zentraler Aspekt einer angemessenen Lesekompetenzmessung ist die zielgruppenspezifische Passung der Itemschwierigkeiten des Aufgabenpools. Es muss sich eine gute Übereinstimmung der Item-Schwierigkeiten mit den Personenfähigkeiten ergeben, welche im Rahmen des Rasch-Modells (Rost, 2004) in einer Wright-Map abgebildet werden kann. In der Entwicklungsstudie des Artikels 2 erwies sich die Übereinstimmung der Item-Schwierigkeiten mit den Personenfähigkeiten der Wright-Map als zufriedenstellend, jedoch im oberen Bereich als nur knapp abgedeckt. Der obere Bereich von Personenschwierigkeiten erwies sich auch in NEPS-Lesetests der erstmessenden Haupterhebungen in den fünften und neunten Klasse 5 und 9, sowie bei Studierenden und Erwachsenen als nicht allzu dicht besetzt (vgl. die Wright-Maps in Haberkorn, Pohl, Hardt & Wiegand, 2012; Hardt, Pohl, Haberkorn & Wiegand, 2013; Pohl, Haberkorn, Hardt & Wiegand, 2012; Pohl, Haberkorn & Hardt, 2014). Der Aspekt von Generierung von Items mit erhöhter Schwierigkeit war demzufolge eine besondere Aufgabe für die Entwicklung von Tests für die Wiederholungsmessung, um auch den oberen Bereich von Personenfähigkeiten der jeweiligen Altersgruppe differenziert und adäquat zu erfassen.

In Artikel 3 wurde in einer Experimentalstudie der Generierung von erhöhter Schwierigkeit besondere Aufmerksamkeit geschenkt. Obwohl durch eine Bedingung des Nichtzurückblätterns kein systematischer Schwierigkeits-Haupteffekt gefunden wurde (Gehrer et al., in Vorbereitung), erwiesen sich gewisse Items als schwieriger als in der Kontrollbedingung mit wiederholter Texteinsicht. Durch Regressionsanalysen konnte gezeigt werden, dass das Aufgabenformat Zuordnungsaufgabe ohne wiederholte Texteinsicht sowohl für schlechte als auch gute LeserInnen signifikant schwieriger war als mit Zurückblättern. Darüber hinaus erwies sich die kognitive Anforderung des Informationenentnehmens differenziell nur für die schlechten LeserInnen als Schwierigkeitsprädiktor.

In Artikel 4 wurde mit dem für sprachlich (hoch)kompetente Personen entwickelten EVAMAR-II-Erstsprachetest eine weiterführende Fragestellung innerhalb der sprachwissenschaftlichen Debatte um die deutsche Plurizentrik untersucht. Das sprachwissenschaftliche Plurizentrik-Paradigma geht davon aus, dass für viele Sprachen und so auch für das Deutsche unterschiedliche Sprachzentren mit eigenen Sprachnormen bestehen, diese sich relativ unabhängig voneinander entfalten und für ihre regional beschränkten Sprachmitglieder prägend und verbindlich sind. Keine der so regional definierten Sprachnormen ist jedoch grundsätzlich der anderen überlegen, sondern nur im eigenen Sprachraum durch die Normsetzung führend bzw. bindend. In der Deutschschweizer Sprachregion wachsen die Kinder mit der dialektalen Familiensprache Schweizerdeutsch auf und werden vor dem Eintritt in die Schule nur über Vorlesen oder Fernsehen etwas ans Hochdeutsche herangeführt. Hauptsächlich der Institution Grundschule obliegt die Aufgabe, die dialektal geprägten Kinder an den Gebrauch der Standardsprache im Schriftlichen und teilweise auch im Mündlichen heranzuführen. Im Alltag wird im Mündlichen außerhalb der Schule und der Universität kein Gebrauch von der Standardsprache/Hochsprache gemacht, höchstens in Ausnahmefällen bei Reden, gewissen Weiterbildungskursen oder mit deutschen VerhandlungspartnerInnen.

Vor diesem sprachlichen Hintergrund lag der zentrale Fokus des Beitrags 4, basierend auf Analysen von Daten mit dem EVAMAR-II-Erstsprachetest, auf der familiären Herkunftssprache und deren vermuteten Auswirkungen auf die sprachlichen Kompetenzleistungen von MaturandInnen. Beim Vergleich der Kompetenzleistungen von MaturandInnen mit dialektal schweizerdeutscher Familiensprache versus MaturandInnen mit hochdeutscher Familiensprache versus gemischter Familiensprache konnte kein Unterschied in ihren Testleistungen gefunden werden.

4 Diskussion

In den vier Schriften der vorliegenden Dissertationsschrift werden zentrale Fragestellungen 1) zur angemessenen Testkonstruktion einer Lesekompetenzmessung über die Lebensspanne und den damit verbundenen Anforderungen und 2) zur zielgruppenspezifischen Lesekompetenzmessung bei (hoch)kompetenten Personen im akademischen Umfeld und den damit verbundenen Besonderheiten bearbeitet. Die Ergebnisse der vier Schriften zu diesen beiden Bereichen werden im Folgenden zusammenfassend diskutiert.

Lesekompetenzmessung über die Lebensspanne

Vor dem Hintergrund der Bedeutung von Lesekompetenz als wichtige Kulturtechnik und Schlüsselqualifikation für die gesellschaftliche, kulturelle und politische Partizipation in modernen, postindustriellen Gesellschaften und der prekären Kompetenzmessungs-ergebnisse um die Jahrtausendwende („PISA-Schock“) wurden die Nachteile und Risiken von Kompetenzmessung in reinen Querschnitten exemplarisch an aktuellen Beispielen der Bildungsforschung dargestellt. Es wurde gezeigt, wie die Interpretation von querschnittlichen Daten teilweise zu keinem vernünftigen Erklärungswissen, sondern zu widersprüchlichen Aussagen führen kann (siehe das Beispiel zur unterschiedlichen Computernutzung bei PISA 2000 versus PISA 2003 und ihren widersprüchlichen Effekten auf die Mathematikleistungen bei 15-Jährigen: Senkbeil & Wittwer, 2006). Individuelle und gruppenspezifische Entwicklungsverläufe können bei ausschließlich querschnittlicher Kompetenzmessung nicht beschrieben werden. Kausale Erklärungen sind ebenso kaum möglich wie empirisch basierte Vorhersagen zu weiteren Entwicklungen und Empfehlungen für langfristig sinnvolle Interventionen und Maßnahmen zuhanden der Bildungspolitik. Längsschnittlich angelegte Bildungsstudien sind notwendig, um wichtiges Erklärungswissen zu generieren (Prenzel, 2006). Das Gutachten von Kristen et al. (2005) zeigte, wie in mehreren Staaten Europas und der USA seit den 1950er-Jahren diverse großangelegte und langjährige Längsschnittstudien

im Bildungsbereich durchgeführt werden, in Deutschland jedoch um die Jahrtausendwende noch nichts Derartiges vorzuweisen war. Weiter wurde dargelegt, wie in Deutschland zwar regional beschränkte, aber dennoch bedeutsame Längsschnittstudien speziell auch im sprachlichen Bereich seit Jahrzehnten wichtige Erkenntnisse brachten, so die Münchner Längsschnittstudien LOGIK und SCHOLASTIK seit den 1980er-Jahren insbesondere zur Rechtschreibleistung und ihrer Vorhersagekraft für spätere Leistungen. Die sich über die zwei Bundesländer Bayern und Hessen erstreckende Längsschnittstudie BiKS kann ab 2005 wichtige Erkenntnisse zur sprachlichen Entwicklung ab dem Grundschulalter und zu Prädiktoren für spätere Lesekompetenz beitragen (Ebert &Weinert, 2013; Pfost, Artelt & Weinert, 2013).

Für die geforderte Lesekompetenzmessung über die Lebensspanne hinweg werden verschiedene Anforderungen benannt, so a) die Konstanz des theoretischen Konstruktes über die verschiedenen Messzeitpunkte und Alterskohorten hinweg und die damit verbundene Breite der theoretischen Rahmenkonzeption, was b) eine passende Auswahl an geeigneten Textsorten (Schrift 1) sowie eine valide und reliable Umsetzung in Testaufgaben mit c) kognitiven Anforderungen (Schrift 2) und d) Aufgabenformaten (Schrift 3) bedingt. e) Die Schwierigkeiten müssen den heterogenen Alterskohorten angepasst sein, aber auch die (hoch)kompetenten Personen in akademischen Kontexten erfassen können (Schrift 3 und 4). f) Weitere Anforderungen bei einer Testung über die Lebensspanne sind eine einzuhaltende Standardisierung der Kontextbedingungen (Einzel- vs. Gruppentestung; Paper-Pencil-Testung vs. Computertestung etc.).

Kritisch anzumerken ist zum letzten Punkt, dass die Anforderung eines konstanten Kontextes und Medieneinsatzes über die Lebensspanne kaum eingehalten werden kann. Köller (2016a) berichtet leicht ernüchtert vom Modernisierungsdruck auf die Large-Scale-Assessments und bilanziert für die computerbasierten Assessments (CBA) in PISA 2015, dass sich die großen Hoffnungen vorerst nicht erfüllt haben. Die Risiken von Mode-Effekten bei der Umstellung auf CBA werden zwar diskutiert, sind aber noch nicht

final abschätzbar. Erste Hinweise deuten teilweise unsystematische Effekte für Lesekompetenzaufgaben am Computer an, andererseits wurden auch schwierigkeitserhöhende Effekte nach Aufgabenformat und Textseiten am Computer berichtet (Bürger et al., 2016).

Wie wichtig die Konstanz einer Messung über die Lebensspanne ist, zeigen die Befunde zu PISA 2015, welche als Trendbefunde nach 12 Jahren stetiger Verbesserungen der Kompetenzleistungen in den Bereichen Lesen, Mathematik und Naturwissenschaften nun erstmals wieder einen leichten Rückgang (für Mathematik) bzw. ein Stagnieren (für Lesen) bzw. ein frappantes Absinken (Naturwissenschaften) bringen. Die empirisch evidente Erklärung für diesen Einschnitt liegt in den „vielen Veränderungen, die in PISA 2015 gegenüber 2012 vorgenommen wurden. Anstelle von 13 Testversionen gab es jetzt 66, alte Aufgaben wurden von Papier- und Bleistift- auf Computerformat umgestellt, neue dynamische naturwissenschaftliche Aufgaben wurden konstruiert und schließlich wurde auch noch das Skalierungsmodell (…) geändert. Dass diese vielen Änderungen negative Effekte hatten, scheint abgesichert zu sein.“ (Köller, 2016b).

Lesekompetenzmessung bei (hoch)kompetenten Personen

Für die Anforderungen der Lesekompetenzmessung bei (hoch)kompetenten Personen wurde die Problematik der Generierung von erhöhter Itemschwierigkeit in der Testkonstruktion diskutiert (Schrift 3) sowie anhand der Textsorte wissenschaftlicher Fachtext auf die Besonderheit der Testung im akademischen Kontext eingegangen (Schrift 4).

Die Bedeutung von Aufgabenformaten für die Itemschwierigkeiten sind Ergebnisse der Studie 3, welche für das komplexe Format der Zuordnungsaufgabe unter der Experimentalbedingung des Nichtzurückblätterns schwierigkeitserhöhende Effekte berichten konnte. Die Anforderung, für jeden Textabschnitt eine zusammenfassende Kernaussage zu finden, welche die Auswahl eines passenden Untertitels aus mehreren

Optionen ermöglicht, beruht auf der erfolgreich erfolgten mentalen Repräsentation sowie einer erfolgreichen lokalen und globalen Kohärenzbildung über die einzelnen Passagen des Textes bereits beim Lesen des Textes vor dem Bearbeiten der Aufgaben. Wenn durch computergestützte Navigationsrestriktion das Zurückblättern in den Text unterbunden wird, können in dieser Experimentalbedingung allfällig vorhandene Lücken im gebildeten Situationsmodell nicht mehr nachträglich geschlossen und Fehlinterpretationen nicht mehr korrigiert werden (Schaffner et al., 2004; Kintsch, 1994). Dies macht die Anforderung dieses Aufgabenformates unter der beschriebenen Experimentalbedingung insgesamt hypothesenkonform schwieriger. Dieses Ergebnis von schwierigkeitsgenerierenden Effekten eines Aufgabenformates wird gestützt durch erste Befunde von Mode-link-Studien des NEPS, bei denen für NEPS-Lesetestaufgaben bei der Umsetzung am Computer für gewisse Aufgabenformate erhöhte Schwierigkeiten gefunden wurden gegenüber der ursprünglichen Papier-Bleistift-Version (Bürger et al., 2016).

Für (hoch)kompetente Personen (gute LeserInnen) fand sich unter der Experimentalbedingung eine größere Schwierigkeitszunahme der entsprechenden Items im Zuordnungsformat als für weniger kompetente Personen. Dies entsprach hypothesenkonform der vermuteten größeren Einschränkung der kompetenten Personen durch die Bedingung des Nichtblätterns. Wie bereits Davey (1987) und Garner und Reis (1981) experimentell für Kinder zeigen konnten, nutzen gute LeserInnen bei der Möglichkeit des Blätterns und der wiederholten Texteinsicht viel effektiver ein bestimmtes Repertoire an Aufgaben- und Textbearbeitungsstrategien. Wenig kompetente Personen nutzen demgegenüber diese viel seltener, wodurch ihr Testverhalten durch die Experimentalbedingung kaum eingeschränkt wird. Grundsätzlich muss für die Studie 3 angemerkt werden, dass aufgrund der beschränkten Itemzahl (n = 72) und beschränkten Texten (n = 6; pro Textsorte n = 2) unter

Experimentalbedingung gegebenenfalls eine Konfundierung mit dem Einzelexemplar des Textes vorliegt.

Für die Studie 4 wird die Bedeutung des offenen Aufgabenformates als weiterer möglicher Schwierigkeitsfaktor diskutiert. Es gibt theoretische und empirische Hinweise, dass die kognitiven Anforderungen insbesondere eines langen offenen Aufgabenformates schwierigkeitsgenerierende Effekte haben (Kintsch & Yarbrough, 1982; Kobayashi, 2002; Prenzel et al., 2002). Der in Schrift 4 eingesetzte EVAMAR-II-Test für Erstsprache verwendet neben geschlossenen Items (n = 50) auch viele offene Formate (n = 74, z. B. Titel und Untertitel für Abschnitte setzen; Eberle et al., 2008, S. 129), um den erhöhten Anforderungen für MaturandInnen bzw. Studierende gerecht zu werden.

Die Verwendung von offenen Formaten wird jedoch durchaus auch kritisch und kontrovers diskutiert. So wird darauf hingewiesen, dass bei der Messung von Lesekompetenz mittels offene Formate durchaus auch eine zusätzliche und andere Kompetenz, nämlich die des Schreibens, miteingeht. Semantische Rezeptionsleistungen werden durch die Schreibleistung überdeckt und Schwierigkeiten beim Niederschreiben von Antworten können als Störfaktor gesehen werden. Daraus kann gefolgert werden, dass vermutlich Personen mit besserem schriftlichem Ausdrucksvermögen systematisch bevorzugt würden. Bei offenen Aufgaben von frühen IEA-Testversionen, der 1991er-Lesestudie der International Association for the Evaluation of Educational Achievement (Elley, 1994), wurden zusätzlich erhebliche übergreifende Bewertungsprobleme im internationalen Bezugsrahmen berichtet, so dass diese für die endgültige Fassung nicht berücksichtigt werden konnten. Auswertungsprobleme und finanzielle sowie personelle Ressourcen sind vermutlich mögliche Gründe, weshalb umfangreichere Studien auf offene Aufgaben eher verzichten wollen. In der zwar national repräsentativen, aber dennoch zahlenmäßig bewältigbaren Schweizer Studie EVAMAR II beliefen sich die

Kodierarbeiten der 4-6 RaterInnen für die offenen Angaben des Erstsprachetests bspw. auf rund acht Wochen.

Entgegen den oben erwähnten Hinweisen auf mögliche Schwierigkeitsgenerierung durch offene Aufgabenformate gibt es auch andere Befunde: In den Versuchen der deutschen Teilstudie der IEA-Studie mit alternativen Aufgabenformaten (bis zu zehn Zeilen lange offene Formate) konnten nur wenige Anhaltspunkte für eine vermutete Überlegenheit solcher Formate gefunden werden (Lehmann et al., 1995, 33–34). Weitere Vergleichsstudien könnten Evidenz bringen.

Eine weitere Besonderheit, um der zielgruppenspezifischen Testung von Personen im akademischen Kontext gerecht zu werden, ist das Nutzen der Textsorte fachwissenschaftlicher Fachtext. Diese generiert Schwierigkeit durch die Verwendung expertennaher Fachbegrifflichkeit und akademisierten Satzbaus (Klein, 2003). In Schrift 4 ist der Erstsprachetest auf einer begründeten Auswahl von Fachtexten verschiedener wissenschaftlicher Richtungen aufgebaut (Eberle et al., 2008).

Anlehnend an die Bewältigung der methodischen Herausforderungen in der Hochbegabtenforschung kann für die Testung von (hoch)kompetenten Personen das Akzelerationsmodell (Heller, 1991) angewandt werden. Diese above-level-Testung, bei welcher jüngeren Zielpersonen ein Test für Ältere vorgelegt wird, macht jedoch nur im Schulalter Sinn, wo sich noch eine fortschreitende Entfaltung der Kompetenzen beobachten lässt. Stattdessen empfiehlt sich theoretisch im Studierenden- und Erwachsenenalter ein adaptives Testen, bei dem zielpersonengerecht direkt während der Testung am Computer ein passendes Testinstrument aufgrund der sofortigen Fehlerrückmeldung ad hoc Schritt für Schritt zusammengestellt wird. Dieses dynamische Testen kann jedoch gerade im Bereich Lesen mit der Vernestung der Aufgabenstruktur in den zugrundeliegenden Texten sowie dem überhöhten Bedarf an dazu erforderlichen Items (Dörfler et al., 2010) derzeit noch kritisch gesehen werden.

Literatur

Ammon, U. (1995). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten*. Berlin, New York: Walter de Gruyter.

Artelt, C., McElvany, N., Christmann N., Richter, T., Groeben, N., Köster, J. et al. (2005). *Expertise – Förderung von Lesekompetenz* (Bildungsforschung Nr. 17). Berlin/Bonn: Bildungsministerium für Bildung und Forschung.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). Opladen: Leske + Budrich.

Augst, G. & Pohl, T. (2007). Zusammenfassende Darstellung. In G. Augst, K. Disselhoff, A. Henrich, T. Pohl & P.-L. Völzing (Hrsg.), *Text-Sorten-Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter* (Bd. 48, S. 345–358). Frankfurt am Main: Peter Lang.

Barquero, B. (1999). Mentale Modelle von mentalen Zuständen und Handlungen von Textprotagonisten. *Zeitschrift für Experimentelle Psychologie, 46* (4), 243–248.

Beck, B. & Klieme, E. (Hrsg.). (2003). *DESI - Eine Längsschnittstudie zur Untersuchung des Sprachunterrichts in deutschen Schulen. Empirische Pädagogik*. Weinheim: Beltz.

Beck, B. & Klieme, E. (Hrsg.). (2007). *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. Weinheim: Beltz.

Bensoussan, M. & Kreindler, I. (1990). Improving advanced reading comprehension in a foreign language: summaries vs. short-answer questions. *Journal of Research in Reading, 13,* 55–68.

Blatt, I. & Voss, A. (2005). Leseverständnis und Leseprozess. Didaktische Überlegungen zu ausgewählten Befunden der IGLU-/ IGLU-E-Studien. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien* (S. 239–281). Münster: Waxmann.

Blossfeld, H.-P., Roßbach, H.-G. & Maurice, J. von (Eds.). (2011). Education as a longlife process. The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft* (Sonderheft 14). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bos, W., Valtin, R., Voss, A., Hornberg, S. & Lankes, E.-M. (2007). Konzepte der Lesekompetenz in IGLU 2006. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes et al. (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 81–107). Münster [u.a.]: Waxmann.

Bürger, S., Kroehne, U. & Goldhammer, F. (März, 2016). *Eine Analyse von Mode-Effekten im Nationalen Bildungspanel (NEPS) am Beispiel von Lesetests in der Sekundarstufe I*. Vortrag an der 4. Jahrestagung der Gesellschaft für empirische Bildungsforschung (GEBF), Berlin.

Christmann, U. & Groeben, N. (1996). Textverstehen, Textverständlichkeit – ein Forschungsüberblick unter Anwendungsperspekive. In H. P. Krings (Hrsg.), *Wissenschaftliche Grundlagen der Technischen Kommunikation* (S. 129–189). Tübingen: Gunter Narr.

Connelly, R. & Platt, L. (2014). Cohort profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology, 43* (6), 1719–1725.

Dörfler, T., Golke, S. & Artelt, C. (2010). Dynamisches Testen der Lesekompetenz. Theoretische Grundlagen, Konzeption und Testentwicklung. Projekt Dynamisches Testen. *Zeitschrift für Pädagogik* (Beiheft 56), 154–164.

Drechsel, B. (2010). Die Lesekompetenz in Deutschland im internationalen Vergleich – Testkonzeption und Befunde aus PISA. In M. Lutjeharms & C. Schmidt (Hrsg.), *Lesekompetenz in Erst-, Zweit- und Fremdsprache* (S. 75–90). Tübingen: Gunter Narr.

Dutke, S. (1998). Zur Konstruktion von Sachverhaltsrepräsentation beim Verstehen von Texten. Fünfzehn Jahre nach Johnson-Lairds Mental Models. *Zeitschrift für Experimentelle Psychologie, 45,* 43–59.

Dutke, S. (1999). Der Crossover-Effekt von propositionaler Textrepräsentation und mentalem Modell: Zur Rolle interindividueller Fähigkeitsunterschiede. *Zeitschrift für Experimentelle Psychologie, 46* (4), 164–176.

Eberle, F., Gehrer, K., Jaggi, B., Kottonau, J., Oepke, M. & Pflüger, M. (2008). *Evaluation der Maturitätsreform 1995 (EVAMAR). Phase II*. Bern: Staatsekretariat für Bildung und Forschung.

Ebert, S. & Weinert, S. (2013). Predicting reading literacy in primary school: The contribution of various language indicators in preschool. In M. Pfost, C. Artelt & S. Weinert (Hrsg.), *The development of reading literacy from early childhood to adolescence* (S. 93–149). Bamberg: University of Bamberg Press.

Elley, W. B. (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*. The Hague: Pergamon.

Embretson, S. E. & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11* (2), 175–193.

Freedle, R. & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing* (10), 133–170.

Freedle, R. & Kostin, I. (1994). Can multiple-choice reading tests be construct-valid? A Reply to Katz, Lautenschlager, Blackburn, and Harris. *Psychological Science, 5* (2), 107–110.

Frey, A. & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft, 10* (Sonderheft 8), 169–184.

Frey, A. & Seitz, N.-N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz. Projekt MAT. *Zeitschrift für Pädagogik* (Beiheft 56), 40–51.

Gehrer, K., Wolter, I., Koller, I. & Artelt, C. (in Vorbereitung). *Lesekompetenztestung mit und ohne Texteinsicht. Gibt es Effekte auf Itemparameter?,* Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Gehrer, K., Zimmermann, S., Artelt, C. & Weinert, S. (2012). *The assessment of reading competence (Including sample items for grade 5 and 9),* National Educational Panel Study. Zugriff am 29.04.2013. Verfügbar unter https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_re_2012_en.pdf

Gehrer, K., Zimmermann, S., Artelt, C. & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online,* (5), 50–79.

Gorin, J. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement,* (4), 351–373.

Gorin, J. S. & Embretson, S. E. (2006). Item Difficulty Modeling of Paragraph Comprehension Items. *Applied Psychological Measurement, 30,* 394–411.

Greiten, S. (2012). *Einblicke in Schulwelten intelligenter Grenzgänger - Fallstudien über hochbegabte Underachiever.* Universität Siegen. Zugriff am 01.01.2017. Verfügbar unter http://d-nb.info/1034425846/34

Groeben, N. & Hurrelmann, B. (Hrsg.). (2006). *Lesekompetenz. Bedingungen, Dimensionen, Funktionen* (2. Aufl.). Weinheim: Juventa.

Haberkorn, K., Pohl, S., Hardt, K. & Wiegand, E. (2012). *NEPS technical report for reading – Scaling results of starting cohort 4 in ninth grade* (NEPS Working Paper No. 16). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Hähnel, A. & Rinck, M. (1999). Strategische Fokussierung der Aufmerksamkeit beim Lesen narrativer Texte. *Zeitschrift für Experimentelle Psychologie, 46* (3), 177–192.

Hardt, K., Pohl, S., Haberkorn, K. & Wiegand, E. (2013). *NEPS technical report for reading – Scaling results of starting cohort 6 for adults in main study 2010/11* (NEPS Working Paper No. 25). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Heller, K. A. (1991). Hochbegabungsdiagnostik. In K. A. Heller (Hrsg.), *Begabungsdiagnostik in der Schul- und Erziehungsberatung* (S. 277–291). Bern: Huber.

Helmke, A., Schneider, W. & Weinert, F. E. (1986). Quality of instruction and class room learning outcomes: The German contribution to the IEA Classroom Environment Study. *Teaching and Teacher Education*, 2, 1–18.

Helmke, A. & Weinert, F. E. (1997). Die Münchner Grundschulstudie SCHOLASTIK: Wissenschaftliche Grundlagen, Zielsetzungen, Realisierungsbedingungen und Ergebnisperspektiven. In F. E. Weinert & A. Helmke (Eds.), *Entwicklung im Grundschulalter* (S. 3–12). Weinheim: Psychologie Verlags Union.

Isaac, K. & Hochweber, J. (2011). Modellierung von Kompetenzen im Bereich „Sprache und Sprachgebrauch untersuchen“ mit schwierigkeitsbestimmenden Aufgabenmerkmalen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43* (4), 186–199.

Johnson-Laird, P.N. (1983). *Mental models*. Cambridge, UK: Cambridge University Press.

Jude, N., Hartig, J., Schipolowski, S., Böhme, K. & Stanat, P. (2013). Definition und Messung von Lesekompetenz. PISA und die Bildungsstandards. In N. Jude & E. Klieme (Hrsg.), *PISA 2009 - Impulse für die Schul- und Unterrichtsforschung* (Zeitschrift für Pädagogik, Beiheft, Bd. 59, S. 200–228). Weinheim: Beltz.

Kintsch, E. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

Kintsch, W. (1994). Kognitionspsychologische Modelle des Textverstehens: Literarische Texte. In K. Reusser & M. Reusser-Weyeneth (Hrsg.), *Verstehen. Psychologischer Prozess und didaktische Aufgabe* (S. 39–54). Bern: Hans Huber.

Kintsch, W. & van Dijk, X (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394.

Kintsch, W. & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology,* 1-6, S. 828–834.

Kirsch, I. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured,* Educational Testing Service.

Klein, W. P. (2003). Die Spannung zwischen Fach-und Gemeinsprache als Anlass für Sprachreflexion. Beispiele aus der Computer- und Internetsprache. 11.-13. Jahrgangsstufe. *Deutschunterricht, 56* (2), 28–32.

Klieme, E. & Steinert, B. (2008). Schulentwicklung im Längsschnitt. Ein Forschungsprogramm und erste explorative Analysen. In M. Prenzel & J. Baumert (Hrsg.), *Vertiefende Analysen zu PISA 2006* (Zeitschrift für Erziehungswissenschaft, Sonderheft, S. 221–238). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klipcera, C. & Klipcera-Gasteiger, B. (1993). *Lesen und Schreiben. Entwicklung und Schwierigkeiten*. Bern: Hans Huber.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing, 19,* 193–220.

Köller, O. (2016a). Editorial. *Diagnostica*, 62 (1), 1–2.

Köller, O. (2016b). Editorial. TIMSS und PISA 2015. Was lernen wir aus den internationalen Schulleistungsstudien? *Psychologie in Erziehung und Unterricht, 84* (1), 1.

Kristen, C., Römmer, A., Müller, W. & Kalter, F. (2005). *Längsschnittstudien für die Bildungsberichterstattung - Beispiele aus Europa und Nordamerika. Gutachten im Auftrag des Bundesministeriums für Bildung und Forschung* (Bildungsreform, Bd. 10). Berlin: BMBF, Referat Publikationen.

Kubinger, K. D. (2009). *Psychologische Diagnostik. Theorie und Praxis psychologischen Diagnostizierens*. Göttingen u.a.: Hogrefe.

Landerl, K. (2008). Schriftspracherwerb. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 577–586). Göttingen: Hogrefe Verlag.

Lehmann, R. H., Peek, R., Pieper, I. & von Stritzky, R. (1995). *Leseverständnis und Lesegewohnheiten deutscher Schüler und Schülerinnen*. Weinheim, Basel: Beltz.

Lenhard, W., & Artelt, C. (2009). Komponenten des Leseverständnisses. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses. Tests und Trends* (Vol. 17, S. 1–17). Göttingen: Hogrefe.

Leutner, D., Fleischer, J. & Wirth, J. (2006). Problemlösekompetenz als Prädiktor für zukünftige Kompetenz in Mathematik und in den Naturwissenschaften. In M. Prenzel, J. Baumert, W. Blum, R. H. Lehmann, D. Leutner, M. Neubrand et al. (Hrsg.), *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 119–137). Münster [u.a.]: Waxmann.

Lubinski, D. & Persson Benbow, C. (2006). Study of mathematically precocious youth after 35 years. Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science, 1,* 316–345.

McElvany, N. (2008). *Förderung von Lesekompetenz im Kontext der Familie* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 64). Münster: Waxmann.

Mrazek, J. (1979). *Verständnis und Verständlichkeit von Lesetexten*. Frankfurt am Main: Peter D. Lang.

Näslund, J. C. (1990). The interrelationships among preschool predictors of reading acquisition for German children. *Reading and Writing*, 2, 327–360.

Naumann, J., Artelt, C., Schneider, W. & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel et al. (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 23–72). Münster: Waxmann.

Nold, G. & Rossa, H. (2007). Leseverstehen. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 197–211). Weinheim: Beltz.

Nold, G. & Willenberg, H. (2007). Lesefähigkeit. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 23–41). Weinheim: Beltz.

OECD (1999). *Measuring student knowledge and skills. A new framework for assessment.* Paris.

Pfost, M., Dörfler, T. & Artelt, C. (2010). Der Zusammenhang zwischen außerschulischem Lesen und Lesekompetenz. Ergebnisse einer Längsschnittstudie am Übergang von der Grund- in die weiterführende Schule. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42* (3), 167–176.

Pohl, S., Haberkorn, K. & Hardt, K. (2014). *NEPS technical report for reading – Scaling results of starting cohort 5 for first-year students* (NEPS Working Paper No. 34). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Pohl, S., Haberkorn, K., Hardt, K. & Wiegand, E. (2012). *NEPS technical report for reading – Scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Preckel, F. & Vock, M. (2013). *Hochbegabung. Ein Lehrbuch zu Grundlagen, Diagnostik und Fördermöglichkeiten*. Göttingen [u.a.]: Hogrefe.

Prenzel, M. (2006). Untersuchungen zur Kompetenzentwickung im Verlauf eines Schuljahres: Die Ergebnisse von PISA-I-Plus im Überblick. In M. Prenzel, J. Baumert, W. Blum, R. H. Lehmann, D. Leutner, M. Neubrand et al. (Hrsg.), *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 14–28). Münster [u.a.]: Waxmann.

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30* (2), 120–135.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J. & Schiefele, U. (Hrsg.). (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*. Münster: Waxmann.

Rausch, T., Matthäi, J., & Artelt, C. (2015). Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*, 147–158.

Rost, D. H. (2001). *Handwörterbuch Pädagogische Psychologie*. Weinheim: Psychologie Verlags Union.

Rost, D. H. (Hrsg.). (2009). *Hochbegabte und hochleistende Jugendliche. Befunde aus dem Marburger Hochbegabtenprojekt* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 72). Münster: Waxmann.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Hans Huber.

Schaffner, E., Schiefele, U., Drechsel, B. & Artelt, C. (2004). Lesekompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand et al. (Hrsg.), *PISA 2003 - Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 93–110). Münster: Waxmann.

Schmiemann, P. (2011). Fachsprache in biologischen Testaufgaben. *Zeitschrift für Didaktik der Naturwissenschaften, 17,* 115–136.

Schneider, W. (2008a). Entwicklung der Schriftsprachkompetenz vom frühen Kindes- bis zum frühen Erwachsenenalter. In W. Schneider (Hrsg.), *Entwicklung von der Kindheit bis zum Erwachsenenalter. Befunde der Münchener Längsschnittstudie LOGIK* (S. 167–186). Weinheim: Beltz.

Schneider, W. (Hrsg.). (2008b). *Entwicklung von der Kindheit bis zum Erwachsenenalter: Befunde der Münchner Längsschnittstudie LOGIK.* Weinheim Beltz.

Schneider, W. & Bullock, M. (2008). Die Längsschnittstudie LOGIK: Versuch einer zusammenfassenden Würdigung. In W. Schneider (Hrsg.), *Entwicklung von der Kindheit bis zum Erwachsenenalter. Befunde der Münchener Längsschnittstudie LOGIK* (S. 203–218). Weinheim: Beltz.

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2007). *LGVT 6-12 - Lesegeschwindigkeits- und -verständnistest für die Klassen 6-12*. Göttingen: Hogrefe Verlag.

Schneider, W. & Stefanek, J. (2004). Entwicklungsveränderungen allgemeiner kognitiver Fähigkeiten und schulbezogener Fertigkeiten im Kindes- und Jugendalter. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 36* (3), 147–159.

Schneider, W. & Stefanek, J. (2007). Entwicklung der Rechtschreibleistung vom frühen Schul- bis zum frühen Erwachsenenalter. *Zeitschrift für Pädagogische Psychologie, 21* (1), 77–82.

Schneider, W., Stefanek, J. & Dotzler, H. (1997). Erwerb des Lesens und des Rechtschreibens: Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert & A. Helmke (Hrsg.) *Entwicklung im Grundschulalter* (S. 113–129). Weinheim: Beltz.

Schweitzer, K. (2007). *Der Schwierigkeitsgrad von Textverstehensaufgaben. Ein Beitrag zur Differenzierung und Präzisierung von Aufgabenbeschreibungen*. Frankfurt am Main: Peter Lang.

Senkbeil, M. & Wittwer, J. (2006). Beeinflusst der Computer die Entwicklung mathematischer Kompetenz? In M. Prenzel, J. Baumert, W. Blum, R. H. Lehmann, D. Leutner, M. Neubrand et al. (Hrsg.), *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 139–160). Münster [u.a.]: Waxmann.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly,* 50 (3), 345–362.

Strohner, H. (1990). *Textverstehen. Kognitive und kommunikative Grundlagen der Sprachverarbeitung* (Psycholinguistische Studien). Opladen: Westdeutscher Verlag.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24* (3), 185–201.

van Dijk, T. A. (1980). *Textwissenschaft: Eine interdisziplinäre Einführung*. Tübingen: Niemeyer.

van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

von Maurice, J., Artelt, C., Blossfeld, H.-P., Faust, G., Roßbach, H.-G. & Weinert, S. (2007). *Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor- und Grundschulalter: Überblick über die Erhebungen in den Längsschnitten BiKS-3-8 und BiKS-8-12 in den ersten beiden Projektjahren*. PsyDok, 1008. Zugriff am 01.01.2017. Verfügbar unter: http://psydok.psycharchives.de/jspui/bitstream/20.500.11780/440/1/online_version.pdf

Watermann, R. & Klieme, E. (2006). Modellierung von Kompetenzstufen mit Hilfe der latenten Klassenanalyse. *Empirische Pädagogik, 20* (3), 321–336.

Weinert, F. E. (Hrsg.). (1998). *Entwicklung im Kindesalter*. Weinheim: Beltz.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and Selecting Key Competencies* (S. 45–66). Seattle: Hogrefe.

Weinert, F. E. & Helmke, A. (Hrsg.). (1997). *Entwicklung im Grundschulalter*. Weinheim: Psychologie Verlags Union.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T. & Carstensen, C. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft* (Sonderheft 14), (S. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.

Wentura, D. & Nüsing, J. (1999). Situationsmdodelle in der Textverarbeitung: Werden emotional entlastende Informationen automatisch aktiviert? *Zeitschrift für Experimentelle Psychologie, 46* (4), 193–204.

Willenberg, H. (2007). Lesen. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 107–117). Weinheim: Beltz.

Wolf, K., Hasebrook, J. & Rinck, M. (1999). Wand oder keine Wand? Die Repräsentation räumlicher Veränderungen in Situationsmodellen. *Zeitschrift für Experimentelle Psychologie, 46* (3), 152–163.

Zimmermann, S. (2016). *Entwicklung einer computerbasierten Schwierigkeitsprädiktion von Leseverstehensaufgaben* (NEPS Working Paper: No. 64). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Anhang

1. Gehrer, K. & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In C. Rosebrock & A. Bertschi-Kaufmann (Hrsg.), *Literalität erfassen: bildungspolitisch, kulturell, individuell* (S. 168–187). Weinheim: Beltz Juventa.

2. Gehrer, K., Zimmermann, S., Artelt, C. & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online, 5* (2), 50–79.

3. Gehrer, K. (2017). *Der Einfluss von Aufgaben- und Textmerkmalen auf Itemschwierigkeit – Eine experimentelle Studie zur Einschränkung der wiederholten Texteinsicht bei der Bearbeitung von Lesekompetenztestaufgaben* (NEPS Working Paper No. 67). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

4. Gehrer, K., Oepke, M. & Eberle, F. (in press). Der EVAMAR II-Deutschtest für GymnasiastInnen – Implikationen für die Plurizentrik-Debatte? In W. Davies, A. Häcki Buhofer, R. Schmidlin, M. Wagner & E. Wyss (Hrsg.), *Standardsprache zwischen Norm und Praxis. Theoretische Betrachtungen, empirische Studien und sprachdidaktische Ausblicke.* (Basler Studien zur deutschen Sprache und Literatur, Band 99). Tübingen: Francke Verlag.

Schrift 1

Gehrer, K. & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In C. Rosebrock & A. Bertschi-Kaufmann (Hrsg.), *Literalität erfassen: bildungspolitisch, kulturell, individuell* (S. 168–187). Weinheim: Beltz Juventa.

Karin Gehrer und Cordula Artelt

Literalität und Bildungslaufbahn: Das Bildungspanel NEPS

Das Konzept der Literalität(en) umfasst ein weites Feld von Aspekten, die sich auf den Erwerb und die Bedingungen eines befähigten Umgangs mit Schriftsprache beziehen (u.a. Feilke 2007). Die konzeptionelle Schriftlichkeit, welche historisch gesehen die Oralität ablöst respektive ergänzt, hat in der Aneignung von Kulturalität eine bewusstseinsbildende Komponente (u.a. Enders 2007). Teilhabe an Kultur bedeutet daher in schriftbasierten Kulturen auch den befähigten Umgang mit Schriftsprache. Ein wichtiger Eckpfeiler des so verstandenen Literalitätskonzepts stellt die Fähigkeit dar, geschriebene Texte zu verstehen und zu nutzen.

Das zentrale Ziel des nationalen Bildungspanels (NEPS) (vgl. Blossfeld/Roßbach/von Maurice 2011) besteht darin, längsschnittliche Daten für die wissenschaftliche Öffentlichkeit zu generieren, die Antworten auf gesellschaftlich und wissenschaftlich relevante Fragen zu Bedingungen und Auswirkungen von Bildung geben können. Untersuchbar werden sollen beispielsweise Fragen dazu, wie sich Kompetenzen über den Lebenslauf entwickeln, wie sie sich durch Einflüsse der Lernumwelten in der Familie, durch Bildungseinrichtungen, am Arbeitsplatz und im Privatleben verändern und welche Erträge von schulischen Bildungsprozessen im Lebensverlauf nachweisbar sind (s.a. Weinert/Artelt/Prenzel/Senkbeil/Ehmke/Carstensen 2011). Hierzu werden im NEPS Stichproben von Personen aus insgesamt sechs (Start-)Kohorten (Neugeborene, Kindergartenkinder, Fünftklässler, Neuntklässler, Studierende und Erwachsene) längsschnittlich untersucht und in regelmäßigen Abständen mit eigens entwickelten Testverfahren in verschiedenen Kompetenzbereichen untersucht. Die ausgewählten Kompetenzbereiche sollen dabei sowohl für ein erfolgreiches, verantwortungsvolles individuelles Leben bedeutsam sein, als auch für eine moderne demokratische Gesellschaft insgesamt. Die Lesekompetenz ist eine der erhobenen Kompetenzen.

1. Konzeption der Lesekompetenz im Bildungspanel NEPS[1]

Anders als bei querschnittlich angelegten international vergleichenden Large-Scale Studien wie PISA, IALS oder PIAAC besteht die besondere Herausforderung für die Rahmenkonzeption zur Lesekompetenz im NEPS darin, eine über die untersuchten Lebensabschnitte hinweg kohärente Messung von Lesekompetenz zu ermöglichen (s.a. Gehrer/Zimmermann/Artelt/Weinert eingereicht; Weinert et al. 2011). Beginnend mit Schülerinnen und Schülern am Ende der Grundschule bis hin zu älteren Erwachsenen (70 Jahre und älter) soll ein vergleichbares Konzept von Lesekompetenz zugrunde gelegt werden, das auch zu aufeinander beziehbaren individuellen Kompetenzwerten führt (s.a. Carstensen/Pohl, eingereicht). Zusammen mit dem bereits genannten Ziel, sowohl die Erträge von Bildungsprozessen für erfolgreiches individuelles Leben als auch für eine moderne demokratische Gesellschaft insgesamt als auch die Entwicklung der Lesekompetenz jenseits der Zeit formeller Beschulung abzubilden, ergeben sich für die Konzeption der Lesekompetenz wenig Alternativen zu einer Ausrichtung auf ein funktionales Verständnis der Lesekompetenz (Weinert/Artelt/Ehmke/Prenzel/Senkbeil/Walter 2008). Ein zentraler Ausgangspunkt der Rahmenkonzeption ist daher in der Orientierung an der Funktionalität und Alltagsrelevanz der Lesekompetenz zu sehen. Die Ausrichtung der Rahmenkonzeption lehnt sich damit an der *Literacy*-Konzeption internationaler Vergleichsstudien an, in der auch der Gedanke der Befähigung zur gesellschaftlichen Teilhabe zentral ist (vgl. OECD 1999). Um zudem auch die notwendige Kohärenz der Konzeption über die Lebensspanne zu erreichen, müssen die zur Konstruktion der Tests maßgeblichen Merkmale so konzipiert sein, dass sie sowohl im Kindes- und Schulalter, als auch bei Erwachsenen und im Rentenalter relevant sind. Leseanlässe, wie sie etwa in PISA als maßgebliche Kategorie der Rahmenkonzeption verwendet werden, scheinen im Anbetracht der Ziele des Bildungspanels keine geeignete Kategorie zu sein. NEPS setzt daher auf eine Konzeption mit zwei konstituierenden Merkmalen, von denen angenommen wird, dass sie in verschiedenen Lebensphasen bedeutungstragend sind: 1) Textfunktionen und damit verbundene Textsorten und 2) kognitive (Verstehens-)anforderungen. Bevor diese konstituierenden Merkmale der Rahmenkonzeption in den folgenden Abschnitten genauer dargestellt werden, werden zunächst noch einige weitere Merkmale der Rahmenkonzeption erläutert (s.a. Gehrer/Zimmermann/Artelt/Weinert eingereicht).

1 Beteiligt an der Entwicklung der Konzeption der Lesekompetenztests des Bildungspanels NEPS sind oder waren neben den Autorinnen Sabine Weinert, Franziska Fellenberg, Jana Heydrich und Stefan Zimmermann. Für die Skalierung der Kompetenzdaten zeichnen Steffi Pohl und Claus Carstensen verantwortlich.

Aufgrund der zeitlichen Restriktionen von je 30 Minuten Erhebungszeit pro Erhebungszeitpunkt werden pro Erhebung jeweils nur fünf Texte, die die fünf im NEPS differenzierten Textfunktionen abbilden, administriert. Pro Text werden dabei jeweils vier bis acht Items dargeboten, welche in ihren Frage- und Aufgabenstellungen ein lokales, schlussfolgerndes und globales Verständnis der fünf Texte erfordern (vgl. Abschnitt zu kognitiven Anforderungen). Die Mehrzahl der Aufgaben entspricht dem Multiple-Choice-Format. Aufgaben dieses Typs bestehen jeweils aus einer Frage/Aufgabe zu einem Text, zu der je vier Antwortoptionen angeboten werden, von denen eine die richtige Antwort ist. Als weiteres Aufgabenformat werden Entscheidungsaufgaben verwendet, bei denen einzelne Aussagen danach bewertet werden müssen, ob sie nach dem Text als richtig oder falsch gelten. Ein drittes Format repräsentieren die sogenannten Zuordnungsaufgaben, bei denen beispielsweise zu jedem Abschnitt eines Textes eine passende Teilüberschrift ausgewählt und zugeordnet werden muss. Bei Aufgaben des zweiten und dritten Formats werden im Zuge der Auswertung Zusammenfassungen vorgenommen, so dass Antworten mit teilrichtigen Lösungen (partial credit items) entstehen (vgl. Carstensen/Pohl eingereicht). Alle drei Aufgabenformate werden in jedem Testheft jeder Kohorte verwendet und sollen möglichst bei allen kognitiven Anforderungen und allen Textsorten eingesetzt werden.

Gegenstand der weiteren Ausführungen ist die detaillierte Darstellung und Begründung von zwei zentralen inhaltlichen Merkmalen der Rahmenkonzeption zur Lesekompetenz im Nationalen Bildungspanel: Kognitive Anforderungen und Textfunktionen, respektive Textsorten. Diese werden jeweils anhand des theoretischen Hintergrunds bzw. des Forschungsstandes und anschließend in Bezug auf ihre konkrete Fassung in der Rahmenkonzeption zur Lesekompetenz erläutert.

2. Kognitive Anforderungen im Leseprozess

Beschreibt Literalität auch die affektiv-emotionalen, reflexiv-ästhetischen und bewusstseinsbildenden Dimensionen des Lesens (Bertschi-Kaufmann/Rosebrock 2009), so befasst sich die Lesekompetenzmessung primär mit der kognitiven Dimension des Lesens und Textverstehens. Nach heute anerkanntem Paradigma ist das Lesen von Texten ein komplexer, (nicht ausschließlich) kognitiver Prozess. Die kognitiven Teilschritte des hierarchisch-aufbauenden Textverstehens sind dank prozessorientierter Methoden der Kognitionspsychologie, hauptsächlich in der Nachfolge von Kintsch und van Dijk (1978), insbesondere für pragmatische Texte gut erforscht (für einen Überblick über die Anfänge der 60er-Jahre bspw. Groeben 1982; zusammenfassend bspw. Christmann/Schreier 2003; Artelt et al. 2005.) Das Grundlagenwissen um die basalen kognitiven Teilprozesse des Lesens wird

von Fachkreisen heutzutage zum „Allgemeingut“ gerechnet (Christmann 2009, S. 181). Im Bereich der kognitiven Teilschritte des Lesens von literarisch-ästhetischen Texten bestehen teilweise noch Forschungslücken (Christmann/Schreier 2003, S. 265 ff.).

Die Lesekompetenzmessung ist am „Endprodukt“ dieses Leseprozesses interessiert: Leistungsvergleichsstudien messen den Output, sie kümmern sich um die empirische Erfassung des latenten Fähigkeitskonstruktes des Textverständnisses der Lesenden. Bei der Operationalisierung der Lesekompetenz in konkreten Aufgabenstellungen muss von der überaus komplexen Struktur der kleinschrittigen kognitiven Teilprozesse des Lesens abstrahiert werden. Das Ziel muss jedoch weiterhin eine möglichst große Übereinstimmung zwischen den Modellen der Testentwicklung und den Erkenntnissen der Grundlagenforschung sein. Bisherige Studien zur Lesekompetenz systematisieren die Verstehensanforderungen unterschiedlich, insgesamt auf abstrahierende Weise aber ähnlich (vgl. Gehrer/Zimmermann/Artelt/Weinert eingereicht). Das Nationale Bildungspanel orientiert sich in der Modellbildung der kognitiven Anforderungen an diesen theoretischen Arbeiten und berücksichtigt zudem Konzeption und Empirie im Rahmen von internationalen Vergleichsstudien. So werden etwa auch in der Rahmenkonzeption zur Lesekompetenz in der PISA-Studie drei Subdimensionen unterschieden, die sich auch als separierbare Fähigkeitsdimensionen empirisch nachweisen lassen (s. a. Artelt/Stanat/Schneider/Schiefele 2001). Unterschieden werden a) das Entnehmen von Informationen, b) das textbezogene Interpretieren und das c) Reflektieren und Bewerten (OECD 2010; s. a. Naumann/Artelt/Schneider/Stanat 2010).

2.1 Kognitive Anforderungen im Lesekompetenztest des Bildungspanels

Die kognitiven Anforderungen von Lesetexten und Leseaufgaben stellen in der Modellbildung der Rahmenkonzeption und der darauf basierenden Aufgabenkonstruktion der NEPS-Lesekompetenztests ein zentrales Merkmal dar. Aus der Literatur zur Lesekompetenz und zum Textverstehen (z. B. Kintsch 1998; Richter/Christmann 2002) und der Rahmenkonzeptionen internationaler Vergleichsstudien zur Lesekompetenz (z. B. OECD 2010) lassen sich verschiedene Arten von Verstehensanforderungen ableiten, die sich in der NEPS-Konzeption zur Lesekompetenz in drei spezifischen kognitiven Anforderungstypen der Aufgaben (Aufgabentypen) widerspiegeln. Die Varianten werden als Typen bezeichnet, da keine explizite Annahme zugrunde liegt, dass Aufgaben eines Typs notwendigerweise schwerer oder leichter sind als Aufgaben eines anderen Typs. Jeder Aufgabentyp repräsentiert jedoch eine andere Art von kognitiven Anforderungen im Verstehensprozess. In der ersten Merkmalsart handelt es sich um Aufgabenstellungen

vom Typ „Informationen ermitteln“ (Typ 1); hier müssen Detail-Informationen auf der Satzebene ermittelt werden, also Aussagen, Propositionen entschlüsselt und wiedererkannt werden. Eine erste Ausprägung dieses Aufgabentypen ist so gehalten, dass die Formulierung im Stimulustext und der Aufgabenstellung identisch ist (Typ 1.1), bei der zweiten Ausprägung weichen die Formulierungen in der Aufgabe und im Text voneinander ab (Typ 1.2).

Bei den so genannten Typ 2-Aufgaben müssen satzübergreifende Schlussfolgerungen im Sinne lokaler und globaler Kohärenzbildung gezogen werden; in der ersten Ausprägung erfolgt dies aus nahe beieinander liegenden Sätzen, in der zweiten Ausformung aus mehreren Sätzen über ganze Abschnitte hinweg und in der dritten Ausprägung (Typ 2.3) geht es darum, wichtige Gedanken im Text nachzuvollziehen, was das Verständnis größerer bzw. komplexerer relevanter Textteile voraussetzt.

Beim dritten Typ von Aufgabenstellungen werden die kognitiven Anforderungen des „Reflektierens und Bewertens“ einbezogen. In einer ersten Ausprägung (Typ 3.1) geht es hier darum, den zentralen Sachverhalt, das zentrale Geschehen oder die zentrale Aussage eines Textes zu verstehen, die zweite Ausprägung (Typ 3.2) verlangt, Adressaten und Intention eines Textes erkennen und die Glaubwürdigkeit eines Textes beurteilen zu können und die dritte Ausprägung des Merkmals Reflektieren und Bewerten bezieht sich auf Transferleistungen oder integriert die Notwendigkeit von Hintergrundwissen (Typ 3.3). Diese Aufgabenstellungen spiegeln unter anderem Anforderungen der kognitiven Repräsentation des Textes in Form eines Situationsmodells bzw. mentalen Modells wider.

Obwohl es auf den ersten Blick vielleicht den Eindruck erweckt, die Typen kognitiver Anforderungen seien schwierigkeitsgestuft, ist es doch vielmehr so, dass es sich um eine qualitative Differenzierung handelt, die primär der Abbildung eines breiten Anforderungsspektrums im Rahmen des Lesekompetenztests dient. Es kann bei jedem Typen sowohl leichte, mittlere als auch schwere Items geben; somit also sowohl leichte Items auf der Ebene des Reflektierens, als auch schwierige Items des Typs Informationen ermitteln oder Schlussfolgern. Die verschiedenen Verstehensanforderungen werden in den jeweiligen Testversionen in einem ausgewogenen Verhältnis berücksichtigt.

3. Unterschiedliche Anforderungen von Textsorten

In der Deutschdidaktik und den Bildungsstandards ist unumstritten, dass unterschiedliche Textgenres und Textsorten unterschiedliche Zugänge der Rezipienten erfordern. Die Linguistik und Textwissenschaften beschäftigen sich seit den 70er-Jahren verstärkt mit den Merkmalen und Anforderungen unterschiedlicher Textsorten und fordern in diesem Zusammenhang die

„Beschreibung einer Kompetenz zum Bilden und Verstehen von Textsorten" (Sandig 1975, S. 113).

Seit dem PISA-Schock 2000/1 ist das Bewusstsein dafür geschärft, dass das verstehende Umgehen mit alltagsnahen Sachtexten ein bis dato vernachlässigter Faktor der schulischen Förderung ist (s. a. Spinner 2004). In diesem Sinne wurde von der Didaktik bereits kurz nach PISA gefordert, dem Umgang mit heterogenem Textsortenmaterial im Klassenzimmer (wieder) mehr Raum zu geben (u. a. Hummelsberger 2003; Paule 2003).

Der traditionelle Literaturunterricht mit Lektüre und Besprechung von klassischer oder moderner Literatur wurde seit dem späten 19. Jahrhundert begleitet vom „Sachbuch" im Unterricht (Doderer 1961, zitiert nach Baurmann 2009, S. 7), und zunehmend ergänzt durch die Einübung eines kritischen Umgangs mit Gebrauchstextsorten wie Zeitungsartikeln, Leserbriefen, Arbeitsberichten, Protokollen u. ä. Insbesondere in den 70er-Jahren erhielt die „pragmatische Literatur im Unterricht" (Kügler 1970, zitiert nach Melenk 2005) mit verschiedenen Textsorten vermehrt Beachtung (Gerth, 1974); so finden sich in damaligen Heften für die Schulpraxis (Praxis Deutsch 1973, 1974) Unterrichtsvorschläge vom zweiten Schuljahr bis zu den zwölften Klassen mit journalistischen Texten, Werbetexten und Gebrauchstexten wie Anleitungen und Formularen. Daneben war aber hauptsächlich die Auseinandersetzung mit (gesellschafts-)politischer Sachliteratur und persuasiven Texten die Folge einer kritischen Neuorientierung des Deutschunterrichts (vgl. Melenk 2005). Knapp zehn Jahre nach PISA sind viele Forderungen von damals erfüllt, in Bildungsstandards und Lehrplänen umgesetzt (Baurmann 2009, S. 7).

Innerhalb der kontinuierlichen Textformen besteht von alters her eine von der Literaturwissenschaft begründete und in der Texttypologieforschung weitergeführte Dichotomisierung zwischen Literatur (Gattung Epik, d. h. Prosa) einerseits und Sachprosa andererseits. Auch die Unterscheidung zwischen fiktionalen und nichtfiktionalen Texten (Werlich 1975; Scheffel 2010), an anderer Stelle auch als faktische oder faktuale Texte auftretend, zielt letztendlich auf die Differenz zwischen literarischen Texten einerseits und pragmatischen Texten (bspw. Abraham 2003) oder Gebrauchstexten (Brinker 1985; Rolf 1993) andererseits.[2] Unterhalb dieser basalen Dichotomie zeigt jedoch weder die Gattungsforschung der germanistischen oder vergleichenden Literaturwissenschaften noch die Texttypenforschung der Textlinguistik eine übergreifende Homogenität (zum Überblick Zymner 2010; insbesondere Adamzik 2010). Es liegen mehrere Vorschläge bezüg-

2 Dass es an den Grenzen von Textklassen Übergänge im Sinne eines Kontinuums gibt, soll uns nicht daran hindern, die grundsätzlichen Kategorien als Pole aufrechtzuerhalten (vgl. Abraham 1998). Auf die zwischen den beiden Polen Literatur – Sachprosa liegenden Mischformen wie Autobiografie, Reisebericht o. ä. kann hier nicht näher eingegangen werden.

lich der Klassifikation von Textsorten vor (für einen Überblick vgl. Rolf 1993); gegenwärtig werden hauptsächlich noch fünf disziplinübergreifende Modelle[3] rezipiert (Fix 2008, S. 27). Zur Entwicklung von Lesekompetenztests kann somit nicht auf eine allgemein akzeptierte Klassifikation von Textsorten oder Textklassen zurückgegriffen werden.

Immerhin kann inzwischen die handlungstheoretisch geprägte linguistische Definition von Brinker aus den 1980er-Jahren als „Standarddefinition" von Textsorte gelten (Adamzik 2010, S. 296): „Textsorten sind konventionell geltende Muster für komplexe sprachliche Handlungen und lassen sich als jeweils typische Verbindungen von kontextuellen (situativen), kommunikativ-funktionalen und strukturellen (grammatischen und thematischen) Merkmalen beschreiben" (Brinker 1985, S. 124; Brinker 2010, S. 135). Dabei ist die Textfunktion als Basiskriterium zur Differenzierung von Textsorten zu betrachten (Brinker 1983, S. 144 ff.); die Textsorte ist somit definitionsgemäß immer an eine sie bestimmende dominierende kommunikative Funktion geknüpft (Brinker 1985, S. 128).

In dem Versuch, in die unglaubliche Textsortenvielfalt[4] eine strukturierende Ordnung zu bringen, lehnen sich frühe textlinguistische Klassifikationsansätze, welche sich zur Unterscheidung an den Textfunktionen orientieren, meist an das Organon-Kommunikationsmodell von Bühler (1934) an, und nehmen die drei Grundformen von a) Ausdrucksfunktion, b) Darstellungsfunktion und c) Appellfunktion von Texten als gegeben an. In den 1980-Jahren geht Brinker in Erweiterung davon von fünf textuellen Grundfunktionen innerhalb der Gebrauchstexte aus: a) von der Informationsfunktion (bei Textsorten wie „Nachrichten", „Beschreibung", „Sachbuch"), b) von der Appellfunktion, c) von der Obligationsfunktion, d) von der Deklarationsfunktion und e) von der Kontaktfunktion. Die letztere wird in neueren Arbeiten wieder inkludiert (vgl. Baurmann 2009, S. 12). Rolf (1993) unterscheidet in seinem an Searles angelegten hierarchischen Klassifikationsvorschlag für Gebrauchstextsorten a) assertive Textsorten, b) direktive Textsorten c) kommissive Textsorten, d) expressive Textsorten und e) deklarative Textsorten, welche in weitere Unterarten unterteilt sind. Im Rah-

3 Die Text(sorten)linguistin Ulla Fix (2008 S. 27) verweist in ihrer Forderung nach einem transdisziplinären Textsortenmodell, das eine Textwissenschaft als Querschnittsdisziplin zu leisten habe, auf die Bedeutung des Mehrebenenmodells von Heinemann/Viehweger (mit Adamzik 2010, S. 297), die Beschreibungen von Sandig/Gobyn sowie die funktionalen Modelle von Rolf (1993) und auf Brinker. Es steht uns nicht zu, hier abschließend zu entscheiden, weshalb wir uns nur grob daran orientieren können.

4 Anlehnend an alltagssprachliche Textsortenkonzepte finden sich bereits im Rechtschreibduden der 1970er-Jahre mehr als 1600 Textsorten-Begriffe, davon rund 500 grundlegende Textsorten wie „Bericht", der Rest als untergeordnete Komposita wie „Reisebericht", „Ergebnisbericht, „Wetterbericht" auftretend (Dimter 1973, zitiert nach Brinker 1985, S. 121). Rolf spricht 1993 bereits von 2100 Textsorten innerhalb der (nicht-literarischen) Gebrauchstexte (S. 132).

men von NEPS übernehmen wir gemäß Brinker (1983, 1985/2010) die Bestimmung der Textsorte durch ihre Textfunktion, lehnen uns jedoch nicht an seine oder eine andere Textklassen-Typologie der Gebrauchstexte[5] an (s. u.).

3.1 Textfunktionen und Textsorten in der Rahmenkonzeption zur Lesekompetenz im NEPS

Das zweite bestimmende Merkmal der Lesekompetenztests im Bildungspanel ist die Dimension der Textsorten und Textfunktionen. In die Rahmenkonzeption des Lesekompetenztests des NEPS wurden fünf kontinuierliche Textsortengruppen aufgenommen, welche in ihrer Heterogenität unterschiedliche Anforderungen an Leseprozesse und das Textverständnis stellen und die breite Vielfalt von Lesegelegenheiten und Textfunktionen bestmöglich innerhalb einer beschränkten Testzeit abzubilden vermögen und zudem sowohl im Alltag von (älteren) Grundschülern, als auch von älteren Erwachsenen (und den Altersgruppen dazwischen) bedeutsam sind. Auf diskontinuierliche Textsorten musste hauptsächlich aufgrund der beschränkten Testzeit zugunsten eines innerhalb der kontinuierlichen Textsorten heterogenen Lesekonstruktes verzichtet werden (s. a. Gehrer/Zimmermann/Artelt/Weinert eingereicht). Im Folgenden wird auf die fünf bei NEPS verwendeten Textsorten (Sachtexte, Anleitungen, Werbung, kommentierende Texte, literarische Texte) und die damit verbundenen Textfunktionen näher eingegangen.

3.1.1 Sachtexte

Die Bedeutung von Sachtexten als zu testende Textsortenklasse ist bei sämtlichen Vergleichsstudien der Lesekompetenz unumstritten (siehe PISA, DESI, PIRLS/IGLU, EVAMAR u. a. m.). Als alltagsnahe Textsorten finden sie sich sowohl in bildungsbezogenen, öffentlichen als auch privaten Situationsfeldern, und über die gesamte Lebensspanne hinweg. Wie bereits beschrieben, kann der Begriff Sachtexte als Oberbegriff verwendet werden und bildet somit als Sachprosa den Gegenpol zur Belletristik (Literatur). Sachtexte in diesem weiteren Sinne umfassen vielfältigste Textsortenklassen. In der Deutschdidaktik wird diese Heterogenität beispielsweise anleh-

5 Typologie von Brinker: Er leitet aus den beschriebenen Grundfunktionen fünf Grund-Textklassen ab, die mit der jeweiligen Textfunktion verbunden sind: 1) Informationstexte, 2) Appelltexte („Werbeanzeige“, „Zeitungskommentar“, „Arbeitsanleitung“, „Gebrauchsanweisung“, „Gesetz“, „Predigt“ usw.), 3) Obligationstexte („Vertrag“, „Gelöbnis“, „Angebot“ usw.), 4) Kontakttexte (mit Textsorten wie „Danksagung“, „Liebesbrief“, „Ansichtskarte“ usw.) und 5) Deklarationstexte („Vollmachten“, „Urkunden“, „Testament“, u. ä.) (Brinker 1985, S. 125).

nend an die beschriebenen kommunikativen Funktionen von Brinker (1985/2010) weiter unterteilt in a) informierende Sachtexte, b) appellierend-instruierende Sachtexte (Anleitung, Aufruf, Anzeige), c) verpflichtende Sachtexte (Garantieschein, Schulordnung, Vertrag) und d) bewirkende Sachtexte (Gutachten, Vollmacht, Zeugnis)[6] (Baurmann 2009, S. 13). Die psychologische Lesekompetenzforschung setzt die Unterkategorien a) Lehrtexte b) Persuasionstexte c) Instruktionstexte i.e.S. (Christmann/Groeben 2002).

Uns erscheint es angemessen, die unterschiedlichen Funktionen der Textsortenklassen ernst zu nehmen und stark zu gewichten, da sie unterschiedliche Merkmale in der Textstruktur, der Syntax, dem Vokabular und der Stilistik bewirken. Von der Rezeptionsseite her sehen wir uns als Lesende einem ganz anderen Text gegenüber, wenn wir uns in einen Sachtext aus einem Lehrbuch oder in einen gesetzesnahen Vertrag vertiefen. Wir müssen andere Teilfähigkeiten und Leseprozesse aktivieren, wenn wir die möglicherweise sehr implizite persuasive Struktur eines Aufrufes entdecken und die dahinter verdeckte Autorintention begreifen wollen. Mit der Sicht auf die kognitiven Anforderungen dieser unterschiedlichen Textsorten, liegt es unseres Erachtens somit nahe, der Heterogenität der Textsortenklassen im Bereich der Sachtexte i.w.S. zu entsprechen und die Untertypen gerade in der (Lesekompetenz-)Testung als eigene Kategorien zu begreifen.

Wir grenzen uns in unserem Textsortenverständnis mit dem Begriff Sachtexte somit auf die informierenden Sachtexte im engeren Sinne ein und verwenden ihn ausschließlich für erklärende Textsorten, die über einen Sachverhalt berichten und informieren. Die anderen Typen oder Subtypen von Sachtexten im weiteren Sinn ordnen wir eigenen, davon unterschiedenen Textsortengruppen innerhalb der Gebrauchstexte zu (bspw. Anzeigen, Aufrufe).

Der Typ Sachtext i.e.S. repräsentiert die Funktion des Informationenvermittelns und -entnehmens, der Weitergabe von Fakten und Erkenntnissen, und ist daher eine der wesentlichen auch mit behaltensorientiertem Lernen und Weiterbildung verknüpften Textformen. Sachtexte vermitteln Sachverhalte der alltäglichen Wirklichkeit von Experten an Laien, und werden als fachexterne Kommunikation auf einem einfacheren Sprachniveau geschrieben als (wissenschaftliche) Fachtexte, in welchen einer interfachlichen oder fachinternen Kommunikation unter Experten entsprechend, ein gehobeneres Niveau von (technischer) Fachsprache verwendet wird (Baurmann 2009, S. 11). Nebst den Fachtexten können auch die Lehrtexte von den Sachtexten i.e.S. unterschieden werden (Baurmann 2009, S. 14): Lehr-

6 Jürgen Baurmann verzichtet auf Brinkers fünfte kommunikative Funktion (Kontaktfunktion) unter dem Hinweis, dass die Kontaktfunktion meist in die anderen kommunikativen Formen miteingebaut auftritt, beispielsweise eine Danksagung, die im Rahmen einer Würdigung auftritt (Baurmann 2009, S. 12).

texte sind meist didaktisch und fachcurricular ausgerichtet und werden ausschließlich in einem didaktischen Rahmen eingesetzt.

Informierende Sachtexte umfassen als Gruppe resp. Textsortenklasse wiederum mehrere Texttypen[7], die unter anderem nach Subtypen, Mischformen und Medium gegliedert werden können. Im Bereich Zeitungssachtexte[8] wird die Textklasse der Sachtexte i.e.S. beispielsweise auch „Mitteilungstexte" genannt (Schröder 2003, S. 241), respektive „Informationstexte" (objektive Nachrichten) im Unterschied zu auffordernden und appellierenden „Meinungstexten" (Hackl-Rößler 2006, S. 33).

Zu den sprachlichen Mitteln, die in informierenden Sachtexten verwendet werden, kann gesagt werden, dass für sie „ein erhebliches Bemühen um Präzision im Ausdruck (durch sachangemessene, doch verständliche Bezeichnungen) bezeichnend" ist (Baurmann 2009, S. 21). Informierende Sachtexte weisen meist einen knappen Satzbau auf, in dem auch Ellipsen (Auslassungen) vorkommen, jedoch seltener Redundanz; verglichen mit anderen Texten finden sich häufiger Passivgebrauch, Nominalisierungen und komplexe Attribute (Baurmann 2009, S. 21). Weiter weisen insbesondere didaktische Sach- und Informationstexte mit dem Anspruch auf Verständlichkeit unter der „Dimension der kognitiven Gliederung/Ordnung" eine geordnete (Vor-)Strukturierung, einen sequenziellen Textaufbau und eine Anreicherung mit Erläuterungen, Beispielen und Analogien sowie eine konsolidierende Zusammenfassung auf (zu empirischen Ergebnissen im Rahmen der kognitionspsychologischen Textverständlichkeitsforschung siehe Christmann/Groeben 2002, S. 152 ff.). Bei den Sachtexten in Zeitungen unterscheidet man zwischen harten und weichen Nachrichten, welche sich im Aufbau, den Themen und Sprachmitteln unterscheiden: Weiche Nachrichten beschreiben Themen aus dem Bereich des Human Interest (Tiere, Prominente, Tragödien, Kurioses in Forschung, Medizin), harte Nachrichten stellen nüchterne Ereignisse aus Politik, Sport und Wirtschaft dar und folgen einer strengen Textstruktur gemäß dem Prinzip der abnehmenden Wichtigkeit; auch der wechselnde Tempusgebrauch folgt einem bestimmten Aufbauschema. Beiden Nachrichtensorten ist eigen, dass sie einfache Wörter und kurze Sätze verwenden, um für ein breites Publikum verständlich zu sein, dass Passivkonstruktionen möglichst vermieden werden, um die Distanz zu den Leserinnen und Lesern nicht zu sehr zu vergrößern, und dass sie meist einen Vorspann, auf verbalen Makrostruktureinheiten beruhende

7 Beispielsweise unterscheidet Thomas Schröder (2003) in seiner qualitativen Analyse, basierend auf 320 informierenden Zeitungstexten, prototypisch zehn verschiedene Texttypen (u.a. faktizierende Meldung, wiedergebende Meldung, erweiterte Meldung, thematischer Bericht, schildernder Bericht, exemplarischer Bericht), die sich bezüglich ihrer Gestaltungsmittel in der Strukturierung der Texthandlung (Wichtigkeitsabfolge, chronologische Abfolge) voneinander unterscheiden lassen.

8 Zu weiteren Einteilungsvorschlägen im Bereich „Sachtexte in Zeitungen" siehe Fußnote von Hackl-Rößler 2006, S. 33

Absätze und Zwischentitel aufweisen (vgl. Hackl-Rößler 2006; Weischenberg 2001).

Die Relevanz von informierenden, erklärenden und beschreibenden Sachtexten für die Lesekompetenzmessung im Bildungspanel NEPS begründet sich einerseits durch die beschriebene wichtige Stellung des Gegenstandes der Sachtexte in der Linguistik und Textforschung als auch in den internationalen Large-Scale-Assessments, andererseits beruht die Auswahl von erklärenden Sachtexten im NEPS auch auf ihrer Verbreitung im Alltag und ihrem Vorkommen und ihrer Nutzung über die ganze Lebensspanne hinweg. In der Kohorte der Fünftklässler/innen würde somit im NEPS beispielsweise ein relativ schulnaher, historisch beschreibender kurzer Sachtext über die erste Fahrt zur Arktis Platz finden können, für die Studierenden würde ein längerer und anspruchsvollerer Sachtext mit mehr Fremdwörtern und schwierigerem Satzbau bspw. im Bereich Entdeckungen und Erfindungen gewählt, der möglichst keine Fachrichtung bevorzugt, und für das heterogene Feld der Erwachsenen könnte beispielsweise eine mittellange weiche Zeitungsnachricht über neue Methoden und Erkenntnisse der Krebsforschung, in etwas größerer Schrift, eingesetzt werden. Die Operationalisierung der Lesekompetenz wird analog zu den Beispielen anhand der Textsorte Sachtexte in allen anderen gewählten Textsorten ebenso über die ganze Lebensspanne hinweg altersangemessen umgesetzt.

3.1.2 Werbung, Anzeigen, Aufrufe

Produkte- oder Dienstleistungswerbungen, Stellenanzeigen, Kursangebote und ähnliche kundenorientierte Textsorten sind Gebrauchstexte mit primär persuasiver Funktion, wobei sich die Werbeabsicht mit einer sehr wohl auch informierenden Funktion des Textes mischt: „Werbeanzeigen haben die Funktion, über das Angebot von Waren und Dienstleistungen zu informieren mit dem Ziel, den Empfänger zum Kauf derselben zu bewegen“ (Bendel 1998, S. 15). Über die Produkte- und Leistungswerbung hinaus, kann auch im weiteren Sinn von Werbung gesprochen werden bei Texten mit beeinflussender Absicht in kulturellen, gesellschaftlichen, religiösen oder politischen Bereichen (Janich 2010). Über diese linguistische Funktionsbestimmung hinaus wird die Bestimmung dieser Textgruppe beispielsweise nach publizistischen Kriterien (Unterscheidung zwischen kommerzieller und privater Werbung, Art des Werbeträgers, kostenpflichtig etc.) für NEPS nicht weiter eingegrenzt.[9] Werbetexte und Anzeigen haben heutzutage eine beinahe unentrinnbare Präsenz erreicht und werden bei steigendem Massen-

9 Diese Nichteingrenzung nach publizistischen Kriterien hat zur Folge, dass in das NEPS-Testmaterial beispielsweise auch private Einladungen zu einem Kindergeburtstag eingehen könnten, da diese auch als Anzeigen verstanden werden.

medienkonsum beinahe ständig mitrezipiert (Temath 2011). Diese Textsortengruppe weist eine zweckorientierte Werbesprache auf, welche häufig artifiziell und prinzipiell stark intentional, konstruiert und inszeniert ist (Janich 2010, S. 45, S. 129). Von der Syntax her ist eine Tendenz zu unvollständigem Satzbau gegeben (Römer 1968/1971; Janich 2010); beim Wortschatz ist grundsätzlich eine häufige Verwendung von Substantiven, insbesondere Zusammensetzungen und Neologismen (Römer 1971, S. 35f.), sowie von Adjektiven, Hochwertwörtern und sich wiederholenden Schlüsselwörtern (Römer 1968, S. 45f., S. 99, S. 131f.) auffällig, welche wie Komparative und Superlative insgesamt das Angebot aufwerten sollen (Römer 1968, S. 105f.). Neuere Forschungen weisen zusätzlich auf den hohen Anteil von Fremdwörtern, insbesondere Anglizismen hin (zur Übersicht vgl. Janich 2010).

Die Häufigkeit dieser persuasiven Gebrauchstexte bei steigender Mediennutzung im Alltag sowie die zu den Sachtexten i.e.S. unterschiedliche Funktion und damit verbundenen Textsortenmerkmalsunterschiede rechtfertigen die Aufnahme dieser Textsortengruppe in die Konzeption der Lesekompetenzmessung im NEPS. Als Beispiele der Konkretisierung dieser Textsortengruppe in den verschiedenen Alterstufen und Lebensabschnitten sind Einladungen zu Kindergeburtstagen für die Fünftklässler/innen zu nennen, Stellenanzeigen für Berufsschüler/innen und eine Werbung für eine Gruppenwanderreise für Erwachsene.

3.1.3 Anleitungen, Anweisungen

Gebrauchstexte mit primär instruktiver Funktion wie Gebrauchsanweisungen, Betriebsanleitungen (für Laien), Rezepte, Pflegehinweise für Textilien, Beipackzettel von Medikamenten und weitere anweisungsorientierte Gebrauchstexte sind in ihrer handlungsorientierten Ausrichtung als eher einfachere pragmatische Textsorten zu betrachten, deren Verstehen im Alltag über das konkrete Außenkriterium unmittelbarer Handlungskonsequenzen nachvollzogen werden kann: Wenn es gelingt, eine instruierte Handlungsanweisung mit dem erwarteten Ergebnis auszuführen, kann man von „richtigem Verstehen“ einer faktisch richtigen Anleitung ausgehen (Foppa 1994, S. 57). Sprachliche Merkmale, die Anleitungen und sonstige Texte mit primär instruktiver Funktion aufweisen, sind unter anderem hohe Anteile von Ellipsen (das Auslassen von Satzteilen), eine Tendenz zu kurzen, wenig komplexen Sätzen und eine starke Verwendung von Indikativ, Imperativ und imperativischem Infinitiv (Nickl 2001, S. 33). Seit den 50er-Jahren des 20. Jahrhunderts wurden schriftliche Gebrauchsanleitungen insbesondere

von Geräten zunehmend üblich[10] und sie gelten inzwischen als Massenkommunikationstextsorten (Nickl 2001, S. 61).

Die pragmatischen Gebrauchstextsorten Anleitungen, Anweisungen und Instruktionen sind für die Konzeption der Lesekompetenzmessung im Bildungspanel NEPS insofern relevant, als für sie davon ausgegangen werden kann, dass auch bildungsfernere Schichten mit ihnen im Alltag in Berührung kommen, beispielsweise als Beigabe zu Geräten oder Medikamenten. Mit ihrer speziellen Funktion des Erläuterns einer Handlung im Sinn einer Handlungsanweisung sind sie eher als vereinfachte Textsorten anzusehen und weisen sie eine eindeutig richtige „Lösung" des Textverstehens auf (dies im Unterschied zu literarischen Texten, siehe unten). Für die Testkonstruktion können mit ihnen vermutlich tendenziell einfache bis mittlere Aufgabenschwierigkeiten konstruiert werden. Altersangemessen über die Lebensspanne realisiert, werden beispielsweise für 10-Jährige auch Spiele- oder Bastelanleitungen eingesetzt. Für Studierende können beispielsweise Bedienungsanleitungen ungewohnter technischer Geräte eingesetzt werden, für Erwachsene alltagsnahe die erwähnten Medikamentenbeipackzettel.

3.1.4 Kommentierende Texte

Zu den kommentierenden oder auch argumentativen Texten kann man Kommentare, begründete Stellungnahmen, Rezensionen und Filmkritiken, Pro-Kontra-Diskussionen zu Sachthemen, Erörterungen, Glossen, Essays und ähnliche Textformen mit argumentativer Funktion rechnen. Einfache bis komplexe Formen davon finden sich in Leserbriefen und Zeitungskommentaren, sehr anspruchsvolle Ausprägungen sind wissenschaftliche Fachtexte und Diskurse, mit literarischen Sprachmitteln spielende Arten sind beispielsweise die Satire, das historische Narrenspiel und politisches Kabarett. Empirische Studien konnten zeigen, dass Kinder ungefähr mit 10 Jahren eine minimale argumentative Struktur in Texten erkennen, bis etwa 16 Jahren hat sich diese Fähigkeit erweitert auf komplexere Argumentationsmuster (Golder/Coirier 1994, 1996). Bei der Konzeptionierung der Lesetests gehen wir basierend auf der Argumentationsforschung (Toulmin 1958; Azar 1999; zur Argumentation im Mündlichen vgl. für einen Überblick u.a. Willenberg/Gailberger/Krelle, 2007) davon aus, dass diese Textformen mit ihrer komplexeren Struktur größere und andere Anforderungen stellen als rein informative Sachtexte, anleitende Gebrauchstexte oder Werbung.

10 Beispielweise wurde das grundsätzliche Bedienen des Telefons seit 1880 im Telefonbuch erklärt. Das Beilegen von Bedienungsanleitungen zum privaten Telefon entstand erst ab 1961 durch eine zunehmend größere Funktionsvielfalt der Apparate (vgl. Neckermann 2001, S. 92).

Bei argumentativen Texten kann zwischen einer minimal argumentativen Textstruktur und einer elaborierten Argumentationsstruktur unterschieden werden. Bei einer einfachen Argumentationsform wird eine Behauptung (claim) aufgestellt und mit einem meist persönlich gehaltenen Argument unterstützt, bei elaborativen argumentierenden Texten wird die These meist durch ein generalisiertes Argument gestützt, welches anschließend mit mehreren Restriktionen relativiert oder durch persönliche Einschätzungen gewichtet wird (Golder/Coirier 1994, S. 193). Bei komplexeren Textformen kann zwischen ein- und mehrsträngigem Argumentieren und ihren Unterformen differenziert werden (vgl. Eggs 1996); einsträngiges Argumentieren bezieht sich in all ihren Argumentationsblöcken, sei es chronologisch oder nicht-hierarchisch, mit These(n), zu entkräftender Gegenthese(n) und ihren Begründung(en), Illustratio und Konklusion auf einen Hauptargumentationsstrang, bei mehrsträngigem Argumentieren wird ein Nebenargument dermaßen gewichtig, dass es sich zu einem selbstständigen Nebenargumentationsstrang entwickelt (Eggs 1996, S. 192).

Bei einem argumentativen oder kommentierenden Text muss die Leserin oder der Leser zum richtigen Verständnis sowohl das dargestellte Thema, die beiden Positionen Pro- und Kontra, als auch die Argumentationsstruktur erkennen. Analog zum mehrschichtigen Situationsmodell bei literarischen Texten (Kintsch 1994) gehen wir davon aus, dass auch bei solchen komplexen Textformen die oder der Lesende gefordert ist, auf mehreren Ebenen eine komplexe mentale Repräsentation des Textes aufzubauen.

Für die Konzeption der Lesekompetenz im Bildungspanel NEPS sind kommentierende und argumentative Texte insofern wichtige Textsorten, als dass sich ergänzend zu den von der Theorie eher als einfach beschriebenen Gebrauchstexten hier anscheinend auch sehr anspruchsvolle Textsortenexemplare finden lassen, welche mit ihrer beschriebenen komplexen Struktur von teilweise mehrsträngigen Argumentationsmustern erhöhte Anforderungen an die Lesenden stellen. Auf der Basis solcher Texte kann es vermutlich gelingen, beispielsweise für Studierende und eine sehr gebildete Erwachsenenschicht, textbasierte Verstehensaufgaben zu konstruieren, welche die Fähigkeiten auch im oberen Bereich der Skala trennscharf abbilden können. Die Konstruktion von schwierigen Items in geschlossenem Format stellt eine große Herausforderung dar, der nur mittels komplexer und damit insgesamt schwierigen Texten zu begegnen ist.

Der altersgemäßen Entwicklung des Erkennens von argumentativen Strukturen in Texten (vgl. Golder/Coirier 1994, 1996) wird im Längsschnitt entsprochen, indem beispielsweise für 10-Jährige ein Text vorgelegt wird, in dem es um einfache Begründungen von Handlungen geht (Beispiel, warum ein Ausflug mit dem Zug geplant werden sollte), bei 16-Jährigen ein mehrsträngiger Text beispielsweise die individuellen und gesellschaftlichen Vor- und Nachteile des alpinen Skisports diskutieren könnte, und Studierende könnten Aufgaben lösen zu einem geistreichen philosophischen Essay

über den (Aus-)Verkauf von CO_2-Emissionswerten, um im Bereich Ökologie zu bleiben.

3.1.5 Literarische Texte

Literarische Texte sind im Vergleich zu den beschriebenen Textsortengruppen als komplexere „Textsorten" anzusehen; ihre Deutung und Interpretation unterliegt nicht eineindeutigen Regeln und Außenkriterien, sondern ist in hohem Maß vom Lesegenuss, der emotionalen Involviertheit und den ästhetischen Vorstellungen der Leserin und des Lesers geprägt und in einen historischen Kontext ästhetischer Subsysteme eingebettet. Literarische Texte werden durch ihre Fiktionalität, Polyfunktionalität und ästhetische Wirkung charakterisiert (Schmidt 1975, S. 68; Pérennec 2002); sie enthalten unter anderem sprachliche Mittel der Mehrdeutigkeit, Metaphern, explizite und implizite Andeutungen, Zitate und Verweise, innere Monologe, Rückblenden, Montagetechniken, Fiktionalitätssignale und verfremdende Stilmittel, wodurch ihre Interpretation anspruchsvoll und die Anforderungen an das Verstehen vielschichtig werden. Gute (moderne) literarische Texte zeichnen sich unter anderem dadurch aus, dass in ihnen nicht alles gesagt wird, was es zu sagen gäbe, dass sie verschweigen und aussparen, so dass die Lesenden gefordert sind, die „Leerstellen" selber zu füllen (Iser 1970, zitiert nach Schmidt 1975, S. 67; Andreotti 2009, S. 407). Aus literaturwissenschaftlicher Sicht kann die sogenannt textseitige Anforderung literarischer Texte nur in Annäherung bestimmt werden (Eggert 2002, S. 187). Das Rezipieren und Verstehen literarischer Texte geschieht gemäß traditioneller literaturwissenschaftlicher Sicht in einem hermeneutischen Zirkel, also einer spiraligen vertiefenden Annäherung an den Sinn des Textes. Diese vertiefende und erweiternde Mehrfachbegegnung mit einem Text widerspiegeln auch gängige Phasenmodelle der Literaturdidaktik, welche der Erstbegegnung mit dem Text und darauf beruhenden Interpretationsentwürfen meist mehrere Phasen von Analyse, Reflektion, Korrektur und Vertiefung folgen lassen (für einen Kurzüberblick vgl. bspw. Leubner/Saupe 2006, S. 32 ff.). Aus kognitionspsychologischer Perspektive geht man davon aus, dass literarisch-ästhetische Texte dazu auffordern, mehrschichtige Situationsmodelle auf verschiedenen Ebenen abzubilden, vom Handlungsablauf, den Milieubeschreibungen bis hin zur „Moral der Geschichte" (vgl. Kintsch 1994).

Literarische Texte werden in den meisten Studien zur Erfassung von Lesekompetenz miteingebunden (PISA, DESI, PIRLS/IGLU). Die Mehrdeutigkeit, Komplexität und Offenheit von literarischen Texten stellt an die Konstruktion von Testfragen hohe Anforderungen. Bei der Formulierung der Aufgaben muss auf die Mehrschichtigkeit des Textes und auf unterschiedliche Interpretationsmöglichkeiten Rücksicht genommen werden.

Dennoch darf eine alltagsnahe Testung von Lesekompetenz die Aufnahme von literarischen Texten nicht verweigern. Für eine breite und heterogene Auffassung von Lesekompetenz, wie sie das Bildungspanel NEPS beabsichtigt, sind literarische Texte aufgrund ihrer traditionell geprägten Bildungsrelevanz unverzichtbar. Aufgrund ihrer komplexen, mehrschichtigen und mehrdeutigen Anlage enthalten sie zudem ein Potenzial von Verstehensanforderungen, mit welchem es gelingen könnte, Aufgabenstellungen auch mit hohen Schwierigkeiten zu konstruieren. Um schultypenspezifischen Bildungsungleichheiten im Vornherein zu begegnen, wurden Gedichte und Theaterstücke, sowie die Stilmittel Ironie und Satire bewusst aus der Konzeption ausgeschlossen. Ein literarischer Text für Fünftklässler (oder auch für Erwachsene) kann beispielsweise ein kurzes Märchen sein (wobei klassische Grimm-Märchen aufgrund ihrer vermuteten Migranten-DIF generell zurückgestellt werden), für sämtliche Kohorten kommen als literarische Texte Auszüge aus Romanen, Kurzgeschichten oder kurze Erzählungen in Frage.

4. Fazit

Der NEPS Lesekompetenztest vereinigt fünf alltagsnahe Textsorten, die aufgrund ihrer unterschiedlichen Funktionen eine angemessene Vielfalt an Leseanforderungen abzubilden vermögen. Die verwendeten Textsorten weisen bezüglich Syntax, Wortschatz, Textaufbau, Komplexität und Verwendung von Stilmitteln teilweise stark unterschiedliche Merkmale auf. Leserseitig bedeutet dies, dass die Rezeption der Texte im Sinne eines Konstruierens des Textzusammenhanges und Textsinns von unterschiedlichen Anforderungen und Konventionen geprägt wird. Auf dem Kontinuum zwischen pragmatischen und literarischen Texten bewegt sich die Leserin und der Leser im Alltag zwischen rein informativen Sachtexten, welche die Wirklichkeit abbilden, deklaratives Wissen vermitteln und einen Anspruch auf Eindeutigkeit transportieren hin zu sehr komplexen literarischen Gebilden, welche der Ästhetik- und Polyvalenz-Konvention unterliegen (nach Schmidt 1975; vgl. Christmann/Schreier 2003) und deren Bedeutungskonstituierung sich dadurch anders gestaltet. Dazwischen stehen Mischformen, welche durch bestenfalls mehrsträngig verflochtenes Argumentieren eine weitaus höhere Komplexität erreichen als einfache Gebrauchstexte, oder solche persuasive Texte, die wie sprachspielerische Werbung, Zitate aus anderen Textarten vereinnahmen und die Lesenden zu einem Handeln veranlassen wollen. Je besser ein Leser oder eine Leserin sich auf die unterschiedlichen Anforderungen eines Textes, einer Textsorte, einlassen kann, je flexibler er oder sie die unterschiedlich geforderten Konventionen erkennen und damit umgehen kann, desto besser wird er oder sie vermutlich ein angemessenes Textverständnis entwickeln und vertiefen können.

Im NEPS-Lesekompetenztests werden zu allen beschriebenen fünf Textsorten und Textfunktionen Aufgabenstellungen entwickelt, deren kognitive Anforderungen darin bestehen, a) Informationen zu entnehmen, b) textbasierte Schlussfolgerungen zu ziehen und c) auf dem Hintergrund des Situationsmodells Aussagen zu reflektieren und zu bewerten.

Durch die systematische Berücksichtigung verschiedener Textfunktionen, die in unterschiedlichen Altersstufen in jeweils lebensnahen und altersangemessenen Texten, Textthemen und unterschiedlichen Verstehensanforderungen der darauf bezogenen Aufgaben umgesetzt werden, ist es möglich, Lesekompetenz als ein relativ breit angelegtes Fähigkeitskonstrukt zu operationalisieren und damit den an die Messung von Lesekompetenz über die Lebensspanne gestellten Zielen nahe zu kommen.

Literatur

Abraham, U. (1998): Übergänge. Literatur, Sozialisation und Literarisches Lernen. Opladen/Wiesbaden: Westdeutscher Verlag.

Abraham, U. (2003): Lese- und Schreibstrategien im themazentrierten Deutschunterricht. Zu einer Didaktik selbstgesteuerten und zielbewussten Umgangs mit Texten. In: U. Abraham/A. Bremerich-Vos/V. Frederking/P. Wieler (Hrsg.), Deutschdidaktik und Deutschunterricht nach PISA (S. 204–219). Breisgau: Filibach.

Adamzik, K. (2010): Sprachwissenschaftliche Gattungsforschung. In: R. Zymner (Hrsg.), Handbuch Gattungstheorie (S. 295–298). Stuttgart: Metzler.

Andreotti, M. (2009): Die Struktur der modernen Literatur (4., vollständig neu bearbeitete Auflage, Original 1983). Bern: Haupt.

Artelt, C./McElvany, N./Christmann N./Richter, T./Groeben, N./Köster, J. et al. (2005): Expertise – Förderung von Lesekompetenz (Bildungsforschung Nr. 17). Berlin/Bonn.

Artelt, C./Stanat, P./Schneider, W./Schiefele, U. (2001): Lesekompetenz: Testkonzeption und Ergebnisse. In: J. Baumert/E. Klieme/M. Neubrand/M. Prenzel/U. Schiefele/W. Schneider et al. (Hrsg.), PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich (S. 69–137). Opladen: Leske + Budrich.

Azar, M. (1999): Argumentative Text as Rhetorical Structure: An Application of Rhetorical Structure Theory. Argumentation, 13, S. 97–114.

Baurmann, J. (2009): Sachtexte lesen und verstehen: Grundlagen, Ergebnisse, Vorschläge für einen kompetenzfördernden Unterricht. Reihe Praxis Deutsch. Seelze: Kallmeyer/Klett.

Bendel, S. (1998): Werbeanzeigen von 1622–1798. Entstehung und Entwicklung einer Textsorte. Reihe Germanistische Linguistik: Bd. 193. Tübingen: Max Niemeyer.

Bertschi-Kaufmann, A./Rosebrock, C. (Hrsg.) (2009): Literalität: Bildungsaufgabe und Forschungsfeld. Weinheim/München: Juventa.

Blossfeld, H.-P./Roßbach, H.-G./Maurice, J. von (Hrsg.) (2011): The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft (Sonderausgabe: Bd. 14). Education as a longlife Process: The German National Educational Panel Study (NEPS). Wiesbaden: VS Verlag.

Brinker, K. (1983): Textfunktionen. Ansätze zu ihrer Beschreibung. Zeitschrift für germanistische Linguistik (ZGL), 11, S. 127–148.

Brinker, K. (1985): Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden. Berlin: Filibach.

Brinker, K. (2010). Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden. (7. durchgesehene Auflage). Berlin: ESV.
Bühler, K. (1934). Sprachtheorie: Die Darstellungsfunktion der Sprache. Jena.
Fischer, B./Carstensen, C./Pohl, S. (eingereicht): Scaling of the competence tests in the National Educational Panel Study – Practice and challenges. In: C. Artelt/S. Weinert/C. Carstensen (Hrsg.), Competence Assessment within the NEPS.
Christmann, U. (2009): Methoden zur Erfassung von Literalität. In: A. Bertschi-Kaufmann/C. Rosebrock (Hrsg.), Literalität. Bildungsaufgabe und Forschungsfeld (S. 181–200). Weinheim/München: Juventa.
Christmann, U./Groeben, N. (2002): Anforderungen und Einflussfaktoren bei Sach- und Informationstexten. In: N. Groeben/B. Hurrelmann (Hrsg.), Lesekompetenz: Bedingungen, Dimensionen, Funktionen (S. 150–173). Weinheim/München: Juventa.
Christmann, U./Schreier, M. (2003): Kognitionspsychologie der Textverarbeitung und Konsequenzen für die Bedeutungskonstitution literarischer Texte. In: F. Jannidis/G. Lauer, M. Martinez/S. Winko (Hrsg.), Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte (S. 246–285). Berlin: Walter de Gruyter.
Eggert, H. (2002): Literarische Texte und ihre Anforderungen an die Lesekompetenz. In: N. Groeben/B. Hurrelmann (Hrsg.), Lesekompetenz: Bedingungen, Dimensionen, Funktionen (S. 186–194). Weinheim/München: Juventa.
Eggs, E. (1996): Formen des Argumentierens in Zeitungskommentaren: Manipulation durch mehrsträngig assoziatives Argumentieren? In: E. Hess-Lüttich (Hrsg.), Textstrukturen im Medienwandel (S. 179–209). Frankfurt a. M.: Lang.
Enders, A. (2007): Der Verlust von Schriftlichkeit: Erziehungswissenschaftliche und kulturtheoretische Dimensionen des Schriftspracherwerbs. Berlin: LIT.
Feilke, H. (2007): Textwelten der Literalität. In: S. Schmölzer-Eibinger/G. Weidacher (Hrsg.), Textkompetenz. Eine Schlüsselkompetenz und ihre Vermittlung (S. 25–38). Tübingen: Narr Francke
Fix, U. (2008): Texte und Textsorten – sprachliche, kommunikative und kulturelle Phänomene. Berlin: Frank und Timme.
Foppa, K. (1994): ‚Verstehen im Dialog' und ‚Textverstehen': Zwei Seiten einer Medaille? Überlegungen zu einem vernachlässigten Problem. In: K. Reusser/M. Reusser-Weyeneth (Hrsg.), Verstehen. Psychologischer Prozess und didaktische Aufgabe (S. 55–68). Bern: Hans Huber.
Gehrer, K./Zimmermann, S./Artelt, C./Weinert, S. (eingereicht): NEPS Framework for Assessing Reading Competence *and* Results From an Adult Pilot Study. In: C. Artelt/S. Weinert/C. Carstensen (Hrsg.), Competence Assessment within the NEPS.
Gerth, K. (1974): Gebrauchstexte im Unterricht. In: Praxis Deutsch (Hrsg.), Zeitschrift für den Deutschunterricht. Nachdruck der Unterrichtsteile aus den Heften 1–6.
Golder, C./Coirier, P. (1994): Argumentative Text Writing: Developmental Trends. Discourse Processes, 18, S. 187–210.
Golder, C./Coirier, P. (1996): The production and recognition of typological argumentative text markers. Argumentation, Bd. 10(2), S. 271–282.
Groeben, N. (1982): Leserpsychologie: Textverständnis – Textverständlichkeit. Münster: Aschendorff.
Hackl-Rößler, S. (2006): Textstruktur und Textdesign: Textlinguistische Untersuchungen zur sprachlichen und optischen Gestaltung weicher Zeitungsnachrichten. Tübingen: Gunter Narr Verlag.
Hummelsberger, S. (2003): ‚Sachtext-Leser' oder ‚Sach-Bearbeiter'? Wie sachdienlich ist die Rede von Sachtexten in eigener Sache? In: U. Abraham/A. Bremerich-Vos/V. Frederking/P. Wieler (Hrsg.), Deutschdidaktik und Deutschunterricht nach PISA (S. 330–346). Breisgau: Filibach.

Janich, N. (2010): Werbesprache: Ein Arbeitsbuch (5. Aufl.). Tübingen: Narr Francke Attempto Verlag.
Kintsch, W. (1994): Kognitionspsychologische Modelle des Textverstehens: Literarische Texte. In: K. Reusser/M. Reusser-Weyeneth (Hrsg.), Verstehen. Psychologischer Prozess und didaktische Aufgabe (S. 39–54). Bern: Hans Huber.
Kintsch, W. (1998): Comprehension: A paradigm for cognition (4. Aufl. 2007). Cambridge: University Press.
Kintsch, W./van Dijk, T. A. (1978): Toward a model of text comprehension and production. Psychological Review, 85(5), S. 363–394.
Leubner, M./Saupe, A. (2006): Erzählungen in Literatur und Medien und ihre Didaktik. Baltmannsweiler: Schneider.
Melenk, H./Metz, K. (2005): Begriffliche Strukturierung von Fachtexten im Deutschunterricht. In: M. Fix/R. Jost (Hrsg.), Sachtexte im Deutschunterricht (S. 83–93). Hohengehren: Schneider.
Naumann, J./Artelt, C./Schneider, W./Stanat, P. (2010): Lesekompetenz von PISA 2000 bis PISA 2009. In: E. Klieme/C. Artelt/J. Hartig/N. Jude/O. Köller/M. Prenzel et al. (Hrsg.), PISA 2009. Bilanz nach einem Jahrzehnt (S. 23–72). Münster: Waxmann.
Neckermann, N. (2001): Instruktionstexte. Berlin: Weißensee Verlag.
Nickl, M. (2001): Gebrauchsanleitungen: Ein Beitrag zur Textsortengeschichte seit 1950. Tübingen: Gunter Narr Verlag.
OECD (1999): Measuring student knowledge and skills: A new framework for assessment. Paris.
OECD (2010): PISA 2009 Ergebnisse: Zusammenfassung. Zugriff zuletzt am 31. 01. 12 unter http://www.oecd.org/dataoecd/34/19/46619755.pdf
Paule, G. (2003): Sachtexte lesen und schreiben. In: U. Abraham/A. Bremerich-Vos/V. Frederking/P. Wieler (Hrsg.), Deutschdidaktik und Deutschunterricht nach PISA (S. 347–360). Breisgau: Filibach.
Pérennec, M.-H. (2002): Von der notwendigen Unterscheidung von Fiktion und Nicht-Fiktion bei einer Text-Typologie. In: U. Fix/K. Adamzik/G. Antos/M. Klemm (Hrsg.), Brauchen wir einen neuen Textbegriff? Antworten auf eine Preisfrage (S. 97–106). Frankfurt am Main: Europäischer Verlag der Wissenschaft.
Praxis Deutsch (Ed.). (1973, 1974): Zeitschrift für den Deutschunterricht: Nachdruck der Unterrichtsteile aus den Heften 1–6.
Richter, T./Christmann, U. (2002): Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In: N. Groeben/B. Hurrelmann (Hrsg.), Lesekompetenz: Bedingungen, Dimensionen, Funktionen (S. 25–58). Weinheim: Juventa.
Rolf, E. (1993): Die Funktion der Gebrauchstextsorten. Grundlagen der Kommunikation und Kognition. Berlin/New York: Walter de Gruyter.
Römer, R. (1971): Die Sprache der Anzeigenwerbung (2. unver. Aufl., Original 1968). Düsseldorf: Pädagogischer Verlag Schwann.
Sandig, B. (1975): Zur Differenzierung gebrauchssprachlicher Textsorten im Deutschen. In: E. Gülich/W. Raible (Hrsg.), Textsorten. Differenzierungskriterien aus linguistischer Sicht (2. Aufl., S. 113–124). Wiesbaden: Akademische Verlagsgesellschaft Athenaion.
Scheffel, M. (2010): Faktualität/Fiktionalität als Bestimmungskriterium. In: R. Zymner (Hrsg.), Handbuch Gattungstheorie (S. 29–31). Stuttgart: Metzler.
Schmidt, S. J. (1975): Ist ‚Fiktionalität' eine linguistische oder eine texttheoretische Kategorie? In: E. Gülich/W. Raible (Hrsg.), Textsorten. Differenzierungskriterien aus linguistischer Sicht (2. Aufl., S. 59–71). Wiesbaden: Akademische Verlagsgesellschaft Athenaion.
Schröder, T. (2003): Die Handlungsstruktur von Texten. Tübingen: Narr.

Spinner, K. H. (2004): Lesekompetenz in der Schule. In: U. Schiefele/C. Artelt/W. Schneider/P. Stanat (Hrsg.), Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000. Wiesbaden: VS Verlag

Temath, B. (2011): Kulturelle Parameter in der Werbung: Deutsche und US-amerikanische Automobilanzeigen im Vergleich. Wiesbaden: VS Verlag.

Toulmin, S. (1958): The Uses of Argument. Cambridge: University Press, 1958. (deutsch (1996): Der Gebrauch von Argumenten.Weinheim: Beltz.)

Weinert, S./Artelt, C./Prenzel, M./Senkbeil, M./Ehmke, T./Carstensen, C. (2011): Development of competencies across the life span. In: H.-P. Blossfeld/H.-G. Roßbach/J. von Maurice (Hrsg.), The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft (Sonderausgabe: Bd. 14). Education as a Lifelong Process. The German National Educational Panel Study (NEPS) (S. 67–86). Wiesbaden: VS Verlag.

Weinert, S./Artelt, C./Ehmke, T./Prenzel, M./Senkbeil, M./Walter, O. (2008): Pillar 1: Development of Competencies Over the Life Course. In: H.-P. Blossfeld (Hrsg), 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 (S. 3–61). Bamberg: Universität Bamberg.

Weischenberg, S. (2001): Nachrichten-Journalismus: Anleitungen und Qualitäts-Standards für die Medienpraxis. Wiesbaden: Westdeutscher Verlag.

Werlich, E. (1975): Typologie der Texte: Entwurf eines textlinguistischen Modells zur Grundlegung einer Textgrammatik. Heidelberg: Quelle & Meyer.

Willenberg, H./Gailberger, S./Krelle, M. (2007): Argumentation. In: B. Beck/E. Klieme (Hrsg.), Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International) (S. 118–129). Weinheim/Basel: Beltz.

Zymner, R. (Hrsg.) (2010): Handbuch Gattungstheorie. Stuttgart: Metzler.

Schrift 2

Gehrer, K., Zimmermann, S., Artelt, C. & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online, 5* (2), 50–79.

Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 5 (2013), No. 2, 50–79

Karin Gehrer, Stefan Zimmermann, Cordula Artelt
& Sabine Weinert

NEPS framework for assessing reading competence *and* results from an adult pilot study

Abstract

This article sketches the framework for assessing reading competence across the lifespan in the German National Educational Panel Study (NEPS). It gives a detailed presentation of the two central dimensions in the framework: (a) text functions and text types and (b) the cognitive requirements of reading tasks. These are discussed against the background of relevant theoretical models and research findings. A pilot study of 447 adults is reported that analyzed the dimensionality and difficulty of the reading competence test for adults. Results indicated that the test meets the NEPS research goals. The article focuses particularly on whether text types and cognitive requirements prove to be appropriate structural elements for the framework, that is, whether each distinguishes sufficiently between good and poor readers. Results also report on the dimensionality of the reading competence test. A comparison between one unidimensional and two different multidimensional models examined whether the text types and/or cognitive requirements of the items/tasks are separate dimensions of reading competence, or whether the reading competence measured in NEPS can – as intended – be conceived as a unidimensional construct. Results are discussed against the background of the scope and limitations of the NEPS reading competence test and the relevant research literature on reading competence.

Karin Gehrer, lic.phil., Dipl. HLA (corresponding author) · Stefan Zimmermann, M.Sc., National Educational Panel Study (NEPS), University of Bamberg, Wilhelmsplatz 3, 96045 Bamberg, Germany
e-mail: karin.gehrer@uni-bamberg.de
stefan.zimmermann@uni-bamberg.de

Prof. Dr. Cordula Artelt, Chair of Educational Research, University of Bamberg, Markusstraße 3, 96045 Bamberg, Germany
e-mail: cordula.artelt@uni-bamberg.de

Prof. Dr. Sabine Weinert, Chair of Psychology I: Developmental Psychology, University of Bamberg, Markusplatz 3, 96045 Bamberg, Germany
e-mail: sabine.weinert@uni-bamberg.de

Keywords
Reading competence; Language assessment; Tests; Dimensionality; Cognitive requirements; Text functions

NEPS-Rahmenkonzeption zur Messung von Lesekompetenz *und* Resultate einer Pilotstudie mit Erwachsenen

Zusammenfassung
Im Rahmen des Artikels wird das Rahmenkonzept zur Messung von Lesekompetenz über die Lebensspanne im Nationalen Bildungspanel (NEPS) skizziert. Dabei werden zwei zentrale Dimensionen dieser Rahmenkonzeption, (a) Textfunktionen bzw. Textsorten und (b) kognitive Anforderung der Leseitems, im Detail dargestellt und vor dem Hintergrund relevanter theoretischer Modelle und Forschungsbefunde diskutiert. Zudem werden Ergebnisse einer Pilotstudie berichtet, die die Angemessenheit des Lesekompetenztests für die Forschungsintentionen des NEPS basierend auf Analysen zur Dimensionalität und Schwierigkeit der Lesekompetenztests auf Basis einer Stichprobe von 447 Erwachsenen darstellen. Spezieller Fokus des Beitrags ist dabei die Frage, ob Textsorten und kognitive Anforderungen als strukturelle Elemente der Rahmenkonzeption insofern angemessen sind, als dass beide (auch) im Erwachsenenalter erlauben, hinreichend zwischen guten und schwachen Lesern zu differenzieren. Zudem werden Ergebnisse zur Dimensionalität des Lesekompetenztests dargestellt. Basierend auf einem Vergleich eines eindimensionalen und zwei unterschiedlichen mehrdimensionalen Modellen gehen wir der Frage nach, ob Textsorten und/oder kognitive Anforderungen der Items separate Dimensionen der Lesekompetenz ausmachen oder die im NEPS gemessene Lesekompetenz – wie intendiert – als eindimensionales Fähigkeitskonstrukt aufgefasst werden kann. Die Ergebnisse werden vor dem Hintergrund der Möglichkeiten und Limitationen des Lesekompetenztests des Nationalen Bildungspanels und der relevanten Forschungsliteratur zur Lesekompetenz diskutiert.

Schlagworte
Lesekompetenz; Kompetenzmessung; Test; Dimensionalität; Kognitive Anforderungen; Textsorten

1. Introduction

Being able to read is a key competence for coping with the demands of everyday life and participating in society. It also remains crucial throughout life for the acquisition and further development of knowledge and skills in countless fields. However, longitudinal studies on the development and the role of reading compe-

tence beyond school age have been either very infrequent or have covered relatively short timespans. The German National Educational Panel Study (NEPS) aims to track reading competence coherently across long stretches of the lifespan, thereby providing empirical access to one of its central issues (see Blossfeld, Roßbach, & von Maurice, 2011). However, such a consistent and coherent longitudinal empirical assessment raises both content-related and methodological challenges for both the framework of competence measurement in NEPS and the development of appropriate instruments (see Weinert et al., 2011). One central starting point for the NEPS framework is an orientation toward the functionality and everyday relevance of the competencies studied. This orientation draws on the concept of *literacy* in international comparative studies with a focus on enabling participation in society (see OECD, 1999).

The article starts by explaining how the current literature understands reading competence and text comprehension as an active process of construction occurring on several levels. It then presents theoretical and pragmatic considerations that take account of earlier concepts and studies of reading competence within the framework of large-scale assessments (LSAs) and form the background to specifying the NEPS framework for measuring reading competence. It reports on the decision not to use discontinuous texts, the state of research on text typology, and the selection of concrete text types in other LSAs; discusses work on the cognitive requirements of text comprehension tasks; and, finally, explains the most important dimensions (text types, cognitive requirements, task formats) of the NEPS framework for measuring reading competence. It also discusses the standards of test development within NEPS and how these are applied in the development of instruments and empirical pilot studies.

Finally, a larger pilot study with adults is used to analyze the parameters for item difficulty and present analyses of the dimensionality of the reading competence test by contrasting the hypothesized unidimensional model with multidimensional models of text functions or cognitive requirements as independent dimensions.

2. State of research

It was only during the phase of (radical) constructivism that research and models of text comprehension found their way back to the old hermeneutic insight that "the 'text' is finally something that constitutes itself in the experience of the recipient" (Hess-Lüttich, 1996, p. 7, translated). Since the work of Kintsch and van Dijk (1978) and Kintsch (1998; van Dijk & Kintsch, 1983), research in cognitive psychology has viewed the reading of texts as a complex active process requiring a number of interacting subabilities (see, for overviews, Artelt et al., 2005; Christmann & Groeben, 1999; Richter & Christmann, 2002). This leads to a distinction between processes on a lower and a higher hierarchic level. The lower level of the process

hierarchy contains the necessary automatic substeps of perception, identification, analysis, and all the decoding processes that lead to word recognition. On the level of sentences, the single semantic statements, the so-called propositions, are assessed within the syntactic sentence structure and related to each other within the process of local coherence formation in order to interpret a phrase meaningfully. On higher hierarchical process levels, there are also cognitive-active processes of selection, construction, and integration in which whole sequences of propositions are linked together, consolidated, selected, and generalized so that it becomes possible to understand complete text elements on the text level. Finally, comprehensive processes of global coherence formation produce so-called macrostructures on a high level of abstraction in order to grasp the global gist of a text.

Research based on the work of Kintsch (1970/1982) and Kintsch and van Dijk (1978) basically follows what was originally Bartlett's schema theory by assuming that the reader applies a cognitive-active comprehension process and uses the representation of the text content presented to finally build up a mental model (in line with Collins, Brown, & Larkin, 1980, as cited in Quathamer, 1998, p. 16). The quality and process of the mental representation (mental model), however, depend, among others, on the reader's individual abilities and capacities. In the process of building up a coherent mental representation, the reader's structural and content-specific knowledge base as well as his or her general knowledge of the world play a special role, because, for example, knowing about the function and particular structure of a special kind of text (be it a fairy tale, an entry in a dictionary, a newspaper article, or whatever) facilitates its reception. Moreover, prior content knowledge makes it easier in general to form a cognitive text representation, because both linking together associated concepts and drawing conclusions on what one has read depend on what one already knows (Kintsch, 1998; Schnotz, 1988).

3. Theoretical and pragmatic prior considerations for the NEPS framework for assessing reading competence

3.1 General considerations on the formation of models in NEPS

In contrast to models in cognitive psychology such as that described by Kintsch and van Dijk (1978) with its focus on the internal processes of information retrieval from the text and a small-scale and multidimensional process analysis, any models focusing on output have to abstract the measurement of performance from the internal structure and process analysis (Schnotz & Dutke, 2004). However, even in an LSA, the goal has to be to achieve the best possible agreement between the findings of cognitive research and the models constructed for test development.

Previous reading competence tests in the frameworks of LSAs have chosen different conceptual focuses. These range from a strong orientation toward the idea of

literacy in international studies of reading competence – such as the International Adult Literacy Survey (IALS; e.g., OECD & Statistics Canada, 1995) in the 1990s or the multicycle comparisons of school performance in the Programme for International Student Assessment (PISA; OECD, 1999, 2001, 2004, 2007, 2009, 2010; see, for Germany, e.g., Artelt, Schiefele, Schneider, & Stanat, 2002; Baumert et al., 2001; Klieme et al., 2010) – to projects based more strongly on linguistic models such as the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) that provides an orientation for studies such as the Level-One Survey (leo) of functional illiteracy (Grotlüschen & Riekmann, 2011), the Workforce Literacy Development Study (lea) on the literacy needs of employees (Grotlüschen, Kretschmann, Quante-Brandt, & Wolf, 2011), or the national reading competence test for secondary school graduates in Switzerland (EVAMAR; Eberle et al., 2008).

When operationalizing reading competence with stimulus texts and items, one general approach is to categorize the underlying texts according to the type of situation in which they are applied. In the field of reading, this has commonly led to a focus on the reasons for reading. For example, the CEFR defines the requirements on learners in terms of communication situations in (a) education, (b) work, (c) the personal domain, and (d) the public domain (Council of Europe, 2001, p. 45). The framework for reading competence in the PISA study also uses comparable situations (Baumert, Stanat, & Demmrich, 2001, p. 24; OECD, 1999).

When designing the longitudinal assessment of reading competence from childhood to retirement in NEPS, these four situations do not offer any coherence across the lifespan – the work domain is lacking in childhood and school age, and not all adults continue to take advantage of education. Therefore, NEPS works with a concept that is oriented less toward the reasons for reading and the reading situations associated with this, but focuses predominantly on the functions of text along with the types of text associated with these functions and how these relate to the cognitive requirements of reading.

3.2 Text functions and text types

Cognitive and psycholinguistic research literature addresses the ways in which the structural features and different communicative functions of text types or genres influence text comprehension because it is assumed that readers form multiple mental representations. These include constructing not only mental representations of the surface text (lexis, syntax), the semantic content, and the situation described, but also a mental representation of the communicative intention of the author and of the text genre (Schnotz & Duthke, 2004, p. 73). The following section will explain the choice of text types used in the NEPS reading framework.

3.2.1 Continuous versus discontinuous texts

One class of text that has become of increasing interest to research in recent years is that of discontinuous (or also noncontinuous) texts. For example, this class plays a major role in the design of PISA (e.g., Artelt, Stanat, Schneider, & Schiefele, 2001; Schaffner, Schiefele, Drechsel, & Artelt, 2004).

Continuous texts are the classic "running texts" of prose (in contrast to classic verse; see Göttsche, 2010, p. 38) that transport exclusively verbal information in the form of letters. Noncontinuous or discontinuous texts extend this by linking the written verbal information to pictorial information in "logical images" (tables, graphs, diagrams) or "realistic" images (illustrations) that are applied as functional rather than decorative elements. The combination of continuous and discontinuous texts results in a broader concept of reading competence (see, for PISA, Baumert et al., 2001; Schnotz & Dutke, 2004, p. 63). However, such a broader concept of reading competence has been criticized (see, e.g., Beck & Klieme, 2007a; Wieler, 2003), because of the suspected and in part also empirically confirmed heterogeneity of the demands imposed on the reader (Artelt et al., 2001; Artelt & Schlagmüller, 2004; Artelt, Stanat, Schneider, Schiefele, & Lehmann, 2004; Baumert et al., 2002; Schnotz & Dutke, 2004).

Whereas Mayer (1997, as cited in Schnotz & Bannert, 1999, p. 3) assumes that the processes of text comprehension and image comprehension run parallel, and contrasts the propositional representation of the text with the imaginal representation of the image, Schnotz and colleagues emphasize that the combination of verbal and pictorial messages in discontinuous texts leads to a new mental model construction when reading. As a result, this type of process can be assessed only with an integrative model of text and image comprehension (e.g., Schnotz & Dutke, 2004; see, for detailed empirical work, Schnotz & Bannert, 1999).

Because of the additional requirements of longitudinal modeling, the NEPS framework and the test construction based on it concentrate on a more homogeneous understanding of reading competence and focus exclusively on the classic continuous forms of text. Therefore, pictorial elements (drawings, clip arts, photographs) are used very sparingly in the NEPS test booklets and only for decoration. The mostly humorous and realistic illustrations are only there to motivate participants. In other words, participants do not have to construct any systematic relations between the (decorative) pictorial elements and the written verbal information. So-called "logical" images such as diagrams and extensive tables from which auxiliary or even the main information has to be linked to the written continuous text have been excluded consistently from the model formation and test development.

As pointed out above, the decision to not use discontinuous texts to measure reading competence in NEPS is based on the argument that continuous and discontinuous forms of text to some extent require different types of reading and comprehension processes. This argument has also been confirmed indirectly through dimension analyses (Artelt & Schlagmüller, 2004). The focus on continuous texts

is also supported by the justified supposition that written continuous text will continue to be the main medium despite the marked growth in the use of discontinuous texts and logical image elements in modern information societies (Schnotz & Dutke, 2004, p. 63). Admittedly, the gain in the precision of measurement through this concentration on purely continuous forms of text does have its price: Limiting the concept of reading competence in this way does reduce some of the relevance to daily life, when it is considered that the functionality concept of an extended reading competence in today's and future media societies naturally also includes the comprehension of functional picture elements.

3.2.2 Continuous text types

Within the continuous forms of text chosen for NEPS, there has always been a dichotomization taken from literary studies and extended by text typology research between literature (of the epic genre, i.e., prose) and nonliterature prose. The distinction between fictional and nonfictional texts (Scheffel, 2010a; Werlich, 1975) – the latter also being called factual texts – also points to the difference between literary texts and pragmatic texts (e.g., Abraham, 2003) or so-called functional texts (Brinker, 1985). Beyond this basic dichotomy[1], neither the form criticism of German studies of comparative literary studies nor the research on text types in textual linguistics reveal any general homogeneity of text typology (see, for overviews, Adamzik, 2010; Zymner, 2010). Hence, we cannot fall back on any broadly accepted classification of text types or text classes when developing reading competence tests.

Nonetheless, Brinker's linguistic definition shaped by action theory in the 1980s can serve as a "standard definition" of text type (Adamzik, 2010, p. 96): "Text types are conventionally valid patterns for complex verbal actions and can be described as specifically typical ties between contextual (situational), communicative-functional, and structural (grammatical and thematic) features" (Brinker, 1985, p. 124, 2010, p. 135, translated). Hence, the function of a text can be seen as a basic criterion for differentiating text types (Brinker, 1983, pp. 144–147). Therefore, the text type is by its very definition always tied to the dominant communicative function that determines it (Brinker, 1985, p. 128). Often, texts fulfill more than one communicative function or speech acts; since Searle's speech act theory a text "is defined as a *complex* verbal action i.e. as a hierarchically structured composition of speech acts, of those one dominates the others" (Brinker, 1983, p. 136, translated)[2].

1 The categories originating in Aristotle of fictional ("poetry") versus factual ("historiography") cannot always be separated strictly, because they overlap in certain text types such as autobiographies or the lyrical ego of a poem (Scheffel, 2010a, p. 30).

2 Brinker points out that the pure quantity of some types of semantic sentences cannot be an indicator for baseline dominance; finally, the decision on the dominance of one function in a text is given by the context or situation (Brinker, 1983, p. 135).

When attempting to find some structure for the incredible variety of text types[3], most early classification approaches in textual linguistics designed to differentiate according to the function of a text referred to the three basic text functions assumed in Bühler's (1934) Organon communication model: (a) expressive function, (b) referential function, and (c) persuasion function. In the 1980s, Brinker extended this model to five basic textual functions within instruction texts: (a) informative (in text types such as news reports, descriptions, and textbooks), (b) persuasive, (c) obligatory, (d) contacting, and (e) declarative.

The NEPS framework follows Brinker (1983, 1985, 2010) in defining text types according to their function, but does not stick to his typology of five text classes of factual or pragmatic texts[4] (see framework section).

3.2.3 Text types in other LSAs

Analog to the inconsistent state of linguistic theory, the major studies on reading competence also apply a variety of operationalizations. For example, the reading test in the longitudinal Assessment of Student Achievements in German and English as a Foreign Language (DESI) used the traditional dichotomy of text types and combined literary texts with information texts, applying two of each at every measurement time (DESI Konsortium, 2006, p. 4). PISA draws on a greater variety of texts, and its original framework is based on the text type model of Mosenthal and Kirsch (1991) that distinguishes between descriptive, narrative, expositional, and argumentative types within continuous texts. The actual design of the PISA reading test reveals six text types: descriptive, expositional, argumentative, directive (or also instructional), documenting (or records), and "narrative" – based on the five-part text typology developed by Werlich (1975, 1976).[5] The fourth cycle of PISA (2009) supplemented paper-based reading competence tests with an optional computer-based test on reading electronic texts. The framework of reading competence was correspondingly extended to include electronic texts, and, among others, the subscales assessing cognitive requirements were adapted to the requirements of electronic text types. This also led to the replacement of the original text

3 Referring to the text type concepts in daily language, one can already find more than 1,600 terms in the German Duden Spelling Dictionaries of the 1970s. About 500 of these are basic text types such as Bericht [report]; the others are secondary compounds such as *Reisebericht* [travel report], *Ergebnisbericht* [outcome report], or *Wetterbericht* [weather report] (Dimter, 1973, as cited in Brinker, 1985, p. 121).

4 Brinker's typology derives five basic text classes from the basic functions reported above that are each tied to a corresponding text function. These are information texts, persuasion texts (advertisement, newspaper commentary, instructions, directions for use, law, sermon, etc.), obligatory texts (contract, pledge, tender, etc.), contact texts (with text types such as thank you letter, love letter, postcard, etc.), and declarative texts (mandates, deeds, wills, etc.) (Brinker, 1985, p. 125).

5 PISA uses "narrative" in a very general sense that also applies to instruction or information prose compared to the narrative as genre in literary studies (see, on the theory of the narrative, Scheffel, 2010b, p. 329).

type "documentary" with texts serving the function "transactional (exchange of information)" (Naumann, Artelt, Schneider, & Stanat, 2010).

3.3 Cognitive requirements of reading tasks in other LSAs

In line with the given paradigm of viewing reading competence as an active comprehension process in which many cognitive steps impose different cognitive demands, LSAs of reading competence tap the various facets of reading competence or the cognitive requirements in the process of text comprehension with different kinds of tasks.

The most comprehensive international study of adult reading competence to date, the International Adult Literacy Survey (IALS/ALL; e.g., OECD & Statistics Canada, 1995) based its measurement of reading competence in the 1990s on the literacy model of Kirsch and Mosenthal (Kirsch, Jungeblut, & Mosenthal, 1998; Mosenthal & Kirsch, 1991; Mosenthal, 1996) and studied several facets of literacy: (a) prose literacy: the comprehension of flow texts with or without pictorial information; (b) document literacy: the reading and comprehension of documents such as forms, timetables and other tables, diagrams, and illustrations; and (c) quantitative literacy: the gathering of numerical information from forms, tables, and other texts that also involves drawing mathematical conclusions (OECD & Statistics Canada, 1995, p. 14). In the IALS/ALL, the cognitive requirements of items were distinguished according to three main aspects of information processing: locating, integrating, and generating. Locating deals with taking information from text, which partly involves drawing conclusions. Integrating requires the reader to piece together information from two or more locations in the text; these can either lie within one section or be distributed across several sections. Generating requires the reader to further process information in the text (e.g., to deliver a written answer) and to draw text-based conclusions, at times, also on the basis of background knowledge (Kirsch, 1995, p. 30; see also Kirsch, 2001). Kirsch (2001) and Mosenthal and Kirsch (1998) developed an additive item rating scheme for this that considered both features of the task or item and the interaction with the text needed to solve the text comprehension task. Depending on which different cognitive processes are necessary for the task processing (difficulty-generating features), the rating scheme assumes the item to be easier or more difficult. In various studies, the authors were able to show that estimates of task difficulty based on their rating scheme were powerful predictors of empirically ascertained task difficulties. They distinguished between the following three factors: (a) type of match, (b) type of information requested, and (c) distracting information. The processes making up the first factor, type of match, are locating, circulating, integrating, and generating. According to the second factor, type of information requested, questions on abstract information are more difficult to answer than questions on concrete things. The third factor, distracting information, addresses the plausibility of incorrect options in the text. An item is easier when there is no distracting information

in the text and more difficult when distractors are located in the same paragraph as the answer to be sought.[6] The systematization of cognitive requirements developed within the framework of the IALS study also forms the basis for the differentiation into three subscales (finding information in the text, drawing text-related conclusions, and reflecting and assessing) in the PISA framework (see below; see also OECD, 1999, 2009) that are each represented by tasks of varying difficulty (assessed empirically or in part with reference to the above schema).

The international elementary school reading study, Progress in International Reading Literacy Study (PIRLS; e.g., Mullis, Martin, González, & Kennedy, 2003; for the German Internationale Grundschul-Lese-Untersuchung, IGLU; e.g., Bos et al., 2003, 2004) used a framework distinguishing between four comprehension processes that were also reflected in the item construction: (a) recognizing and reporting explicitly given information; (b) drawing simple conclusions; (c) drawing complex conclusions, justifying them, and interpreting what one has read; and (d) testing and assessing language, content, and text elements. The first two tasks require the use of information inherent in the text; conclusions are reached by forging relations between given parts and sections. In the more complex comprehension processes (c) and (d), in which respondents have to reflect on content or structures, it is also necessary to draw on external knowledge (Bos et al., 2003, pp. 76–77). The DESI study (Beck & Klieme, 2003, 2007b; Nold & Willenberg, 2007; Willenberg, 2007) oriented its measurement of reading competence in 9th-grade classes toward text research within cognitive psychology by Kintsch[7] (1994) or van Dijk and Kintsch (1983). It used a process model of reading competence with six requirements on the theoretical levels of (a) information, (b) inferences, (c) focusing, (d) knowledge, (e) links, and (f) mental model (Nold & Willenberg, 2007; Willenberg, 2007; see, for a criticism, e.g., Bremerich-Vos & Grotjahn, 2007). Their empirical findings revealed that these could be aggregated to form the following four cognitive requirements: (a) finding information on the level of words in the context of the sentence; (b) local reading: drawing conclusions (inferences) and abstracting the topics in single sections; (c) forging links between different paragraphs and text passages – also with recourse to one's own prior knowledge; and (d) mental model: integrating the central aspects of the text (see Garbe, Holle, & Jesch, 2009, p. 27).

The international student assessment program PISA uses the theoretical structure of reading competence to differentiate five cognitive requirements: (a) developing a general understanding of the text; (b) gathering information; (c) developing a text-related interpretation; (d) reflecting on the content of the text; and (e) reflecting on the form of the text. Three subdimensions could be confirmed empir-

6 According to the schema, other factors influencing task difficulty are the number of sentences within which the correct answer is to be found, the number of answers sought and whether this number is reported, how far the information given in the question matches the information in the text, and how far the answer can be taken from the text or has to be constructed by the respondent (see Kirsch, 2001).

7 Whereas Kintsch talks about the "situation model" of the text, DESI adopts the "mental model" concept from Christmann and Groeben (1999) for the highest level.

ically: gathering information (b), text-related interpreting (a and c), and reflecting and assessing (d and e). These were added to the total scale of reading competence as report scales (Artelt et al., 2001).

NEPS is oriented toward the aforementioned theoretical analyses and empirical findings from various LSAs, and applies the distinction between cognitive requirements and difficulty-generating features for the framework and operationalization of reading competence.

4. The NEPS reading competence framework

Reading makes it possible to access and acquire a variety of life and knowledge domains. The range of reasons for reading is very broad, and reading simultaneously fulfills a multitude of different functions (see, e.g., Groeben & Hurrelmann, 2004). These range from the reading that is essential for further training and lifelong learning, across broadening one's general knowledge, up to literary and aesthetic reading. Texts convey not only information and facts but also ideas, values, and the contents of culture. The concept of reading competence in NEPS is based accordingly on a functional understanding of reading competence as also reflected in the literacy concept (see also OECD, 2009). The focus is on handling texts competently in various characteristic everyday situations.

The concept of reading competence in the NEPS framework concentrates on reading competence as text comprehension: It assesses the comprehension performance shown in replies to questions referring to the specific underlying text. Research focuses on the abilities to read a text and understand it appropriately – both as a whole and in its single statements. The emphasis is on understanding what is in the text and not primarily on memory performance for text material that has been read but is no longer available.

To represent the concept of reading competence as coherently as possible over the entire lifespan, the framework for the NEPS reading competence test specifies three features that have to be taken into account in each age- or stage-specific[8] form of the test: (a) text functions and text types, (b) cognitive requirements, and (c) item formats. However, in contrast to other LSAs that use a booklet design and thus are able to administer far more test items in the same time period, the items within each NEPS reading competence test are the same for each student within the particular age group/stage. As a result, the set of items administered is rather limited (ca. 30 items referring to five texts per measurement time). NEPS aims to assess reading competence as a comparatively homogeneous unidimensional construct. The features differentiated in the framework and test construction serve pri-

8 NEPS divides education trajectories into eight educational stages such as "From Kindergarten to Elementary School (Stage 2)". Some stages apply specific additional tests that are not continued longitudinally across the total lifespan. For example, Stage 2 is assessing phonological awareness (see Berendes, Weinert, Zimmermann, & Artelt, 2013, this issue).

marily to account systematically for the breadth of text types, cognitive requirements, and item formats.

4.1 Text functions and text types

NEPS distinguishes five text functions and their associated text types that are taken into account in each form of the test: (a) information texts, (b) commenting or argumenting texts, (c) literary texts, (d) instruction texts, and (e) advertising texts. This selection was based on the assumption that these five text functions are relevant for the everyday lives of participants of all different ages.

The continuous forms of text applied in NEPS can be characterized in terms of their functions or types (see, for more detail, Gehrer & Artelt, 2013): *Texts imparting information* represent basic texts for learning, the fundamental acquisition of knowledge, and finding information. Examples are articles, dispatches, reports, and announcements. Texts with a *commenting or argumenting function* take a particular stance or they query something, balance out arguments for and against, or include a reflective view (reader's letters, discussions, essays, academic papers). Texts with a *literary-aesthetic function* are short stories, excerpts from novels, or narratives. Special literary text types such as theater plays, satires, or poems are excluded because reading them probably depends strongly on types of education and curricula. The fourth category is made up of text types that convey *user guidance* such as assembly instructions, operating manuals, package insets for medication, work instructions, cooking recipes, and so forth. The fifth category *advertising* (*appeals, advertisements, announcements*) contains texts for advertising, job announcements, leisure activities, and so forth.[9]

The five selected text functions and their associated text types are operationalized in each test booklet as a longitudinal concept across the lifespan; that is, each test booklet measuring reading competence contains a total of five texts corresponding to these five text functions.

In contrast to PISA, NEPS is not applying any discontinuous texts with pictorial information from diagrams, tables, or graphical illustrations. Discontinuous texts are not included in the NEPS framework, as mentioned before, because they impose special requirements on readers.

9 For the reading test, preference was given to selecting texts that are as prototypical as possible for each text type. Nonetheless, the borders between some text types are not so fixed, for example, many instruction texts only report the actual operating instructions after an introductory section containing general information. There are also mixed types, for example, when advertising takes quotes from literature and integrates them humorously. According to Brinker (1983) we orientated ourselves on the all-dominant communicative function of the given text and tried to avoid such extremely mixed types when selecting texts for the NEPS reading test. For the type "information text" we set limits in the narrower sense to avoid an intermixture with other subtypes e.g. compulsory information texts (see Gehrer & Artelt, 2013, pp. 175–178).

In order to measure differences in text comprehension rather than differences in prior knowledge, the demands on the specific prior knowledge of the test persons should be kept as low as possible in a reading competence test (see, for an overview, e.g., Köster, 2003). Therefore, NEPS fundamentally excludes texts requiring specific prior knowledge from both its framework and the test construction. Such texts can be poems whose reception builds on prior knowledge of types of rhyme and verse, metric, and their place in literary history, but also specialized texts and those types of information texts with a discipline-specific vocabulary requiring special prior knowledge. Texts are primarily selected to have topics reflecting a general knowledge of the world. In addition, several different approaches were used to keep the demands on the reader's prior knowledge as low as possible, for example, by formulating the items so that they are closely related to the text (e.g., by asking which of several correct statements can be found in the text and which not) or by telling the reader to refer to the text when answering questions both in the instructions, after each text before answering the items, and, at times, even within the question stem.

4.2 Cognitive requirements

The second feature of the framework and thus of the task construction for measuring reading competence in NEPS is cognitive requirements. As mentioned above, various types of cognitive requirements can be derived from the cognitive psychological literature on reading competence and text comprehension on both hierarchically low (decoding, word recognition) and hierarchically high levels (local and global text comprehension) (e.g., Kintsch, 1998; Richter & Christmann, 2002). As also mentioned above, the latter have been operationalized slightly differently in various LSA studies of reading competence. The NEPS concept of reading competence reflects the higher cognitive requirements in three specific types of item. These variants are labeled types, because they are not based on any explicit assumptions that one type of item is necessarily more or less difficult than another type of item. However, each type of item taps another kind of cognitive requirement in the comprehension process. The first type is items on "finding information in text" (Type 1). These are items in which readers have to find detailed information on the sentence level, that is, to decode and recognize statements and propositions. A first version of this type of item is designed so that the formulation in the stimulus text and the item is identical (Type 1.1); in the second version, the two formulations differ (Type 1.2).

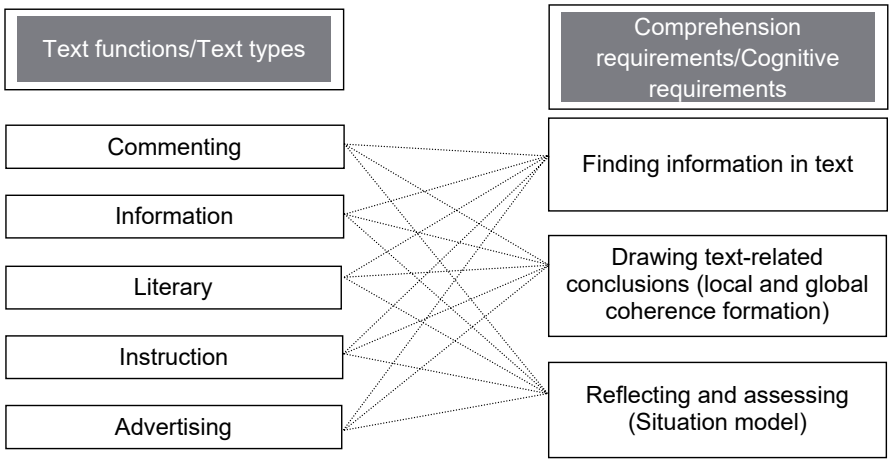
In Type 2 items, on "drawing text-related conclusions", inferences have to be drawn from several sentences in order to construct local or global coherence. In the first version, this has to be done with neighboring sentences; in the second version, with several sentences spread across whole sections of text; and in the third version (Type 2.3), it is necessary to understand the major ideas in the text, which requires the comprehension of larger or more complex sections of relevant text.

The third type of item includes the cognitive requirements of "reflecting and assessing". In the first version (Type 3.1), readers have to comprehend the central idea, event, or message in a text; in the second version (Type 3.2), they have to be able to recognize the purpose and intention of a text and judge its credibility. The third version integrates the need for background knowledge (Type 3.3). The requirements reflect, among others, the need to represent the text in the form of a situation model or a mental model.

Although, at first glance, the types of cognitive requirements may seem to be ranked according to difficulty, they differ far more qualitatively, and primarily reflect a broad spectrum of requirements on the text level within the framework of a reading competence test. In contrast to rather hierarchical models of cognitive processes on the word-, sentence- and text-level the NEPS framework and reading competence test focus exclusively on the cognitive requirements of the higher hierarchical levels in the construction processes of text comprehension. A text comprehension test like the NEPS reading competence test takes the lower processes for granted but does not assess them. The various types of tasks in the NEPS reading competence test manifest different kinds of complex cognitive requirements (finding information in text, drawing text-related conclusions, reflecting and assessing). By way of example, least of all in task type 1 a very complex process occurs while words, clause constituents, subordinate clauses and compound sentences must be combined, weighted and compared among themselves and against the stimulus text for verification or falsification.

For each type of requirement, items can be easy, intermediate, or difficult; hence, there are not only easy items on the level of reflecting but also difficult items on the level of finding information or drawing conclusions. More detail can be found in the empirical part of this article. The various cognitive requirements are to be found in all text functions and are balanced across each age appropriate test booklet (see Figure 1).

Figure 1: Text functions and cognitive requirements



4.3 Item formats

The majority of items are presented in a multiple-choice format. This type of item consists of a question related to a text with a choice of four possible answers of which one is correct. A further item format is decision-making items in which single statements have to be evaluated according to whether they are correct or incorrect in relation to the text. A third format is matching items in which a suitable heading has to be selected and assigned to each section of a text.[10] Second- and third-format items are summarized during the course of data analysis to produce items with partially correct solutions (for the scoring of the partial credit items; see Pohl & Carstensen, 2013, this issue). The different item formats should be applied across all text types, to as many cognitive requirements as possible, and across all age levels.

4.4 Test length and coverage of all features in the single tests

The reading competence test should take approximately 30 minutes to complete[11] per measurement time. In line with the framework, the final instruments for all age levels in the main national surveys each have five texts with the aforementioned five text functions and four to eight items for each text tapping local, deductive, and global comprehension (see, for concrete item examples in Grade 5 and 9, Gehrer, Zimmermann, Artelt, & Weinert, 2012). The three aforementioned cognitive requirements of the reading process in their eight different versions are balanced as far as possible in the test booklet for each starting cohort. All three item formats are applied in each test booklet.

By systematically considering the various text functions and operationalizing them in close-to-life and age-appropriate texts and text topics, and items tapping different cognitive requirements, it is possible to operationalize reading competence as a broadly based ability construct.

5. NEPS test development standards

The general challenge when implementing any measurement of competence lies in carefully transforming the theoretical foundations and framework model into empirically valid and reliable test instruments. The particular challenge when measuring reading competence longitudinally is to generate not only age-appropriate tests

10 Open unstructured formats are not used for both classificational and theoretical reasons. Measuring receptive language competencies (reading, listening) with productive procedures (e.g., writing or recounting a summary) is controversial in research, although short forms of open formats (e.g., composing titles) are in common use (e.g., PISA, EVAMAR).

11 Testing time also includes 2 minutes for self-ratings on the number of items solved (see Händel, Artelt, & Weinert, 2013, this issue, p. 170).

but also to test consistently and coherently across the lifespan. NEPS meets the demand for age-appropriate testing through an age-appropriate selection of stimulus texts and items; it meets the second demand for consistent and coherent modeling across the lifespan by consistently implementing the longitudinally designed framework.

To ensure that the text comprehension requirements correspond to the ability spectrum of the given age group, tests for all age groups are initially subjected to a careful preselection of test materials based on expert appraisals, difficulty indices, and readability indices. This should ensure that the stimulus texts are age-appropriate in terms of their length or shortness, style, syntax, vocabulary, and topics. This is followed by a multistage development and pretest process in which the items for each stimulus text are optimized successively in terms of their validity and model fit. One goal is for the items to provide an adequate spectrum of difficulty within each age group. How far this has been achieved for the first reading competence tests is one of the topics addressed in more detail in the empirical part of this article.

5.1 Empirical pilot studies to develop the test

Before they are applied in the field for the main survey, the instruments developed in NEPS have to go through several phases of cognitive interviews (see, for the method[12], Prüfer & Rexroth, 2005), smaller and larger preliminary studies, and large pilots (feasibility studies). Basically, the pool of texts and accompanying items developed for each starting cohort is four times as large as the final selection. Each development pool contains at least 20 texts and at least four examples of each text type – literary, commenting, advertising, instruction, and information. The complete pool of test material is piloted on the target population. One-half of each set of test materials is also tested on the next younger age cohort in NEPS[13] (e.g., items for 9th grade on a sample in 7th grade); the other half is also tested on the next higher age cohort (e.g., 12th grade). The samples for the pilots are representative samples from four German federal states (see, for a more detailed description of sampling procedures, Aßmann et al., 2011). After the pretests, suitable items are selected for each starting cohort in a two- to three-stage procedure. Items with less favorable parameters are optimized, and some new items are developed. Single units (i.e., texts and their attendant items) that prove to fit another age cohort better than that originally intended are reallocated. The resulting test materials are then given to the corresponding cohorts in pilot studies in four feder-

12 In the first phase of test development, cognitive interviews are suitable tools for obtaining early indications of problems that may arise (e.g., a question is not formulated clearly, certain words are not understood, text is found to be strenuous or boring, etc.).

13 In the school-age stages, reading competence is being measured every 2 years. Hence, in secondary school, it is being measured in the 5th, 7th, 9th, and, for some students, also in the 11th grade.

al states. The resulting data are, and will be, used to select the best items or units on the basis of content-related and statistical criteria. The statistical criteria[14] include fitting the items to the underlying unidimensional Rasch model, the absence of outliers for differential item functions (DIF) for gender and type of school, as well as item coverage of a broad spectrum of difficulty for one age group. The main emphasis is to cover all the criteria in the framework in a balanced way.

6. Study of the dimensionality of the reading competence test for adults

6.1 Research question

The main purpose of the features included in the framework for assessing reading competence is to cover the entire breadth of uses of reading and cognitive requirements. Due, among others, to the time constraints on testing within NEPS, the goal of test construction is to tap an intrinsically relatively homogeneous construct of reading competence in participants after psychometrically confirming the unidimensionality of the assessed construct. Items are also selected with reference to a unidimensional Rasch model. The assumption of unidimensionality is tested in this empirical section. Analyses of the dimensionality of reading comprehension tests are applied to data from a development study to ascertain whether the assumed unidimensionality can be confirmed, or whether and to what degree the text functions and cognitive requirements tap empirically distinguishable dimensions. This analysis is based on data from a pilot study with adults. It also tries to test the appropriateness of the framework using the text functions and cognitive requirements as structural features to assess reading competence: By analyzing the difficulty distribution of the items tapping the various text functions and cognitive requirements, we can analyze indirectly whether the items tapping the various levels of these features are suitable for assessing differences in ability between persons in this adult sample.[15] This should test empirically whether the test instrument resulting from the development study is appropriate for the main survey.

6.2 Sample

Participants were 447 adults (258 women and 186 men, 3 participants did not report gender). Our plan had been to stratify the sample to produce equal distributions of the variables age (birth cohorts Group 1: 1975–1989; Group 2: 1960–1974;

14 We thank Steffi Pohl for carrying out the basic scaling of the competence data for reading literacy (see, for a description of the respective analyses, Pohl & Carstensen, 2012).

15 This can already be assumed for 15-year-olds (PISA; see OECD, 2002).

Group 3: 1943–1959) and education (low, intermediate, high level of education[16]). However, we managed to only approach an equal distribution in the final sample: The youngest Group 1 contained 135 participants (30.2%); the intermediate Group 2, contained 146 (32.7%), and the oldest Group 3 contained 151 (33.8%) over 51-year-olds (18 participants did not report their age). The mean age was 43.3 years (*SD* = 2.4, range: 21.0–66.3 years). The lowest education group contained 107 persons (23.9%); 2 of these persons had left school without graduating, 2 had graduated from a special needs school [*Förderschulabschluss*], 58 had left school with basic school-leaving certificates [*einfacher Hauptschulabschluss*], and 45 with more qualified basic school-leaving certificates [*qualifizierender Hauptschulabschluss*]. The intermediate education group contained 159 persons (35.6%) with an intermediate school-leaving certificate [*Realschulabschluss*]. Group 3 with the highest education level contained 179 persons (39.8%), of whom 43 (9.6%) had a university of applied science entry certificate [*Fachhochschulreife*] and 135 (30.2%) had a university entrance certificate [*Abitur*]. Three participants gave no reports on their education.

6.3 Study implementation

The study was carried out by the *infas* survey institute. Participants were tested individually under standardized conditions in their own private homes. After completing a short questionnaire tapping the socio demographic origin, migration background and reading related behavior, they worked on four 30-minutes blocks, each containing five units consisting of one text and the accompanying items. They were paid € 15 (about $ 20) for their participation.

6.4 Instruments

A total of 26 units from the reading competence test were administered in the pilot study. By applying a multimatrix design, each participant received only 20 of the 26 units. The 26 units were distributed across 10 test booklets in which the sequence of units was also varied to balance out any effects of fatigue across all participants and tests.

The various text functions had an unequal distribution across the 26 units: There were 10 information texts, 5 commenting texts, 5 literary texts, 5 instruction texts, and 5 advertising texts. Because the greater number of information texts made the items in this category more heterogeneous, 5 of the 10 information texts were chosen at random for further analysis.

16 Level of education was determined by combining indicators on school education and the highest vocational training qualification.

We then performed an item selection. Exclusion criteria were a discriminative power of less than 0.20, an item misfit (MNSQ) greater than 1.25, and irregularities on the level of distractors (e.g., positive relation to the total test outcome). In addition, deviations in the item-characteristic curve (ICC) implied by the model from the data-based ICC were assessed qualitatively by the research team, which led to further items being dropped. A further criterion was the differential item functions (DIF) for gender and type of school. Items with extreme DIF values were dropped, whereas certain items with intermediate DIF values were retained when they were relevant for the content of a unit. This left 152 of the original 219 items. However, this set of items showed an overrepresentation of conclusion-drawing items, so a further randomly chosen 43 items were dropped from the dimension testing. Hence, the analyses presented here were based on 109 items that were roughly balanced in terms of the number of text types and the cognitive requirements (see Table 1).

Table 1: Distribution of text types and cognitive requirements in the 109-item analysis pool for testing multidimensionality

Text type/Function	Cognitive requirement			
	Finding information in text	Drawing text-related conclusions	Reflecting and assessing	∑ Items
Information	8	8	5	21
Commenting	6	6	9	21
Literary	3	9	8	20
Advertising	6	9	7	22
Instruction	10	6	9	25
∑ Items	33	38	38	109

6.5 Analysis strategy

The subsequent data analysis was performed with ACER Conquest 2.0 software (Wu, Adams, Wilson, & Haldane, 2008). The reading items were Rasch scaled (see Pohl & Carstensen, 2013, this issue). Complex items (decision-making and matching items) that were not answered completely correctly received partial credit scores. Three different models were computed with the item pool: a unidimensional model, a three-dimensional model (taking the three different cognitive requirements of the items into account), and a five-dimensional model (taking the five text types into account). We used two criteria from information theory (see Rost, 2004) as indicators for the comparative testing of the model fit: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). A further criterion was the significance of the deviance change.

7. Results on the dimensionality and difficulty of the adult reading competence test

7.1 Model fit

The results on whether the conceptual unidimensionality of reading competence assessment or the theoretical facets of the framework can be found in the empirical data structure present a differentiated picture (see Table 2). The dimension analyses showed that the text functions in particular showed clear empirical differences in the features, whereas the cognitive requirements are not verifiable as own empirical dimensions.

Table 2: Model comparison in terms of multidimensionality (N = 447, 109 Items)

Model	AIC	BIC	Deviance	Model parameters	Test versus unidimensional model
Unidimensional	35,306.33	35,970.94	34,982.33	162	–
Three-dimensional (cognitive requirements)	35,302.8	35,987.93	34,968.80	167	$\Delta\chi^2(5) = 13.53$, $p = .02$
Five-dimensional (functions of text)	35,166.03	35,888.08	34,814.03	176	$\Delta\chi^2(14) = 168.30$, $p = .00$

Note. AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

7.2 Cognitive requirements

The dimension analysis of the cognitive requirements showed that the three-dimensional model had a better numerical fit than the unidimensional model according to the AIC criterion and deviance change which attained statistical significance $\chi^2(5, N = 447) = 13.53$, $p = .02$ (see Table 2). However, the sample-weighted and therefore stricter BIC criterion had a higher value for the three-dimensional model (35,987.93) than for the unidimensional model (35,970.94), indicating a better fit with the data for the unidimensional model. Taken together, when stricter criteria were applied, the expected unidimensional construct of reading competence had a better fit than the three-dimensional model based on cognitive requirements.

All intercorrelations between the three dimensions of cognitive requirements adjusted for measurement error (see Table 3) were conspicuously high (r = .94, .96, and .99). The highest intercorrelation at r = .99 was between "reflecting and assessing" and the cognitive requirement "drawing text-related conclusions (during local and global coherence formation)". These two dimensions seem to form almost identical subfacets of reading competence in the NEPS test for adults.

Table 3: Intercorrelations between dimensions of cognitive requirements (N = 447, 109 Items)

	Finding information in text	Drawing text-related conclusions	Reflecting and assessing
Finding information in text	–		
Drawing text-related conclusions	.96	–	
Reflecting and assessing	.94	.99	–

Note. Intercorrelations are to be seen as correlations between the latent dimensions (as estimated in ACER Conquest 2.0).

7.3 Text functions and text types

The results of the dimension analyses showed that the five-dimensional model differentiating according to the text functions fitted the empirical data from the reading competence test better than the unidimensional model. The criteria were quite unequivocal: The deviance, the AIC, and sample-weighted BIC had the lowest values in the five-dimensional model compared to the unidimensional model. The deviance change criterion also attained statistical significance, $\chi^2(14, N = 447) = 168.30$, $p = .00$. Hence, the five-dimensional model based on text types was actually the best model. However, the intercorrelations corrected for measurement error between the five text functions were also very high (r = .78 to .91), although lower than those for cognitive requirements (see Table 4). Yet, there were some conspicuous findings here as well: The information texts correlated particularly strongly with the advertising (r = .91), the instruction (r = .89), and commenting texts (r = .87). There were also strong correlations of the instructions with advertising texts (r = .87) – a group of simple information texts containing appeals, calls, and announcements in which it is not necessary to follow complicated lines of argumentation. The literary texts correlated less strongly with the other texts, which is not surprising in light of the specific features of this category. With intercorrelations of r = .78 with information texts, r = .80 with instruction texts, r = .82 with commenting texts, and r = .83 with advertising texts, the literary texts showed a comparatively intermediate intercorrelation with the other text types.

Table 4: Intercorrelations between dimensions of text functions/text types (N = 447, 109 Items)

Text Functions/ Types	Information	Commenting	Literary	Instruction	Advertising
Information	–				
Commenting	.87	–			
Literary	.78	.82	–		
Instruction	.89	.85	.80	–	
Advertising	.91	.84	.83	.87	–

Note. Intercorrelations are to be seen as correlations between the latent dimensions (as estimated in ACER Conquest 2.0).

Apart from the intercorrelations, the analysis pool revealed that the most variance between readers came from the literary texts (σ^2 = 1.81). Hence, the differences between good and poor readers were markedly stronger here than in information (σ^2 = 1.5), advertising (σ^2 = 1.41), instruction (σ^2 = 1.63), or commenting texts (σ^2 = 1.41).

7.4 Selection of items for the main study and analysis of distribution of difficulty for the text functions and cognitive requirements

The 26 items selected from the preliminary study pool for the main study address five texts spread across all text functions and containing all three cognitive requirements. Basically, results showed that the selected items differentiated particularly well in the lower ability range, but not in the higher. Therefore, six further difficult items were modified or specially constructed for the main survey. However, because these were piloted after the present preliminary study pool, they cannot yet be presented with the same psychometrics.

An inspection of the distribution of item difficulties (ranging in total from -2.93 to 0.35 on the logit scale) as a function of text type revealed that the items in each text type – with the exception of commenting texts in which the items were closer together – had a difficulty spectrum ranging from 1 to 2.4 units on the logit scale. The difficulty spectrum covered by the items in each of the three cognitive requirements was higher than 1.8 units on the logit scale. For the requirement to find single pieces of information in text, the items even ranged over the entire difficulty spectrum of almost three logits for the items analyzed in this study.

8. Discussion

The specially developed NEPS reading competence test presented in this article is based on a framework oriented toward the standards of international LSAs of reading competence, and permit a longitudinal modeling of the development of reading competence over the lifespan. The NEPS framework for measuring reading competence unites the features of the cognitive requirements of the reading process with the conditions that different text functions and text types impose on readers. The tests hold both features constant across the lifespan. The presentation of the process of developing test instruments based on the NEPS framework showed how the quality of the future test instrument could be ensured by constructing a preliminary pool that was four to six times larger than needed, applying several phases of optimization and selection, and performing tests on several age cohorts. A randomly selected analysis pool[17] (109 items) from the pilot study for adults (N = 447) was used to examine whether the two features within the framework would also prove to be empirically distinguishable subdimensions of reading competence, or whether the expected unidimensionality of the assessment of reading competence could be confirmed. Results showed that it was particularly the cognitive requirements that had exceptionally high intercorrelations (r = .94, .96, and .99). This corresponded to the fact that a model comparison based on the stricter BIC criterion indicated that reading competence, as measured following the NEPS framework, is as hypothesized, a unidimensional rather than a three-dimensional construct and therefore the cognitive requirements are not verifiable as own empirical dimensions.

It seems worthwhile to compare the relatively high intercorrelations between the cognitive requirements with the relations between cognitive requirements found for the PISA reading literacy test (15-year-olds). Given that the intercorrelations of cognitive requirements in the PISA study were reported using uncorrected intercorrelations based on WLE (weighted likelihood) estimates, it is necessary to compare them to intercorrelations of the cognitive requirements in NEPS based on the same algorithm. The resulting uncorrected intercorrelations based on WLE estimators for the NEPS data for adults were markedly lower (r = .72, .77, and .71) and corresponded roughly with the uncorrected correlations in PISA (r = .74, .71, and .64). The comparison confirmed, among others, the smallest correlation found in the NEPS data between the subdimensions “finding information” and “reflecting and assessing”. However, whereas in PISA, the closest relation was found between the subdimensions “finding information” and “text-related interpreting” (Artelt & Schlagmüller, 2004), NEPS finds the closest relation between the cognitive requirements “drawing text-related conclusions” and “reflecting and assessing”.

17 Because only five texts each accompanied by 4–6 items can be applied in the main surveys due to time restrictions, the multidimensionality testing was performed with a larger pool from the development study. The analysis pool was refined by removing items that failed to reach test validity criteria and reducing the overhang of information texts. The resulting dimension analyses could be performed on the basis of 109 aggregated items.

When interpreting these different subfindings, it is necessary to recall that although NEPS applies a similar framework for cognitive requirements to that in PISA, it is not identical. For example, in contrast to PISA, the NEPS framework places more emphasis on the local and global coherence formation in the subdimension "drawing text-related conclusions", and less emphasis on integrating background knowledge in the different levels of "reflecting and assessing". Naturally, these theoretical differences in the framework influence the test construction and hence the results of both studies. In addition, the concentration on continuous texts in NEPS versus the addition of discontinuous texts in PISA also has an appreciable effect on the cognitive requirements.

It is still necessary to test whether the markedly high intercorrelation of r = .77/.99 between "reflecting and assessing" and "drawing text-related conclusions", is also due to occasional slight overlaps in the requirements of the subtypes of items. For example, the drawing text-related conclusions task of "being able to understand the important ideas in a text on the basis of comprehending more complex relevant text sections" (Type 2.3 items) and the reflecting and assessing task of "comprehending the central idea, event, or message of a text" (Type 3.1) also seem to have very similar contents.

The test of multidimensionality based on text type features showed major differences between the text functions. The fact, that the model based on the different text types showed a better model fit than the hypothesized unidimensional model of the framework, could provide an indication in the direction – as supposed from academics in the domain of languages and educational didactic –, that the different features of text types exert an influence on their reception and connected reading achievement.

The strong relation between information texts and all other text types apart from literary texts indicates that – greatly simplified – all text types used apart from literary texts can be conceived more generally as information or instruction texts; that is, the typical requirements they place on reading competence have much in common. These result concerning the difference between literary versus information based subcompetencies was already presented by the detailed analysis of the PISA data (Artelt & Schlagmüller, 2004). New results in the domain of literary-aesthetic research confirm this view; Roick and colleagues showed that for adolescents literary versus information based subcompetencies can be distinguished empirically (Roick, Frederking, Henschel, & Meier, 2013). Though both studies operationalized literary reading above all of the three literary genres – the lyrics, the drama and the epos or narrative literature –, meanwhile the NEPS framework and the reading test instruments focus exclusively on the narrative literature, remarkably, our first results in analyzing an adult pilot study goes in the same direction.

Our analysis of the text types additionally confirmed the constructional assumption linked to the NEPS framework that a heterogeneous selection of continuous text types can provide a balanced assessment of reading competence in the sense of text comprehension. An inspection of the distribution of item difficulties as a func-

tion of the text type or the specific cognitive requirement reveals that each text type or requirement delivers a satisfactory to very good differentiation in this age range.

The presentation of the items intended for the main survey revealed that the selected items differentiated particularly well in the lower ability range but not in the upper range. This finding can and has to be qualified by stating that the sample in this preliminary study was a positive selection due to the failure to achieve an equal distribution of educational attainments. It can be assumed that this led to a tendency to overestimate the performance of adults in the population and underestimate the difficulty of the items. Nonetheless, additional difficult items were constructed to tap the upper ability range more broadly.

Finally, it should be pointed out that the analyses carried out here are still only provisional trends until the analyses of further age cohorts (Grades 5 and 9 and college students) deliver comparative confirmation of the results obtained for the adult items. Particularly for highly correlated dimensions, Harell (2009), for example, has pointed out that the use of information theory criteria does not always suffice to identify the right model.[18]

Moreover, the multidimensionality of the two category systems should not just be interpreted in a disassociated way. A promising next step in this regard would be to perform an in-depth multidimensionality analysis assuming within-item multidimensionality (see Hartig & Höhler, 2008).

Acknowledgments

This work was supported by third-party funds from the German Federal Ministry of Education and Research (BMBF). We thank Franziska Fellenberg, NEPS, University of Bamberg, for her assistance in the conception and development of the reading test for the starting cohorts, and Jonathan Harrow for translating this article into English. The original, unpublished German-language version is available from the first author on request.

References

Abraham, U. (2003). Lese- und Schreibstrategien im themazentrierten Deutschunterricht. Zu einer Didaktik selbstgesteuerten und zielbewussten Umgangs mit Texten. In U. Abraham, A. Bremerich-Vos, V. Frederking, & P. Wieler (Eds.), *Deutschdidaktik und Deutschunterricht nach PISA* (pp. 204–219). Breisgau, Germany: Filibach.

Adamzik, K. (2010). Sprachwissenschaftliche Gattungsforschung. In R. Zymner (Ed.), *Handbuch Gattungstheorie* (pp. 295–298). Stuttgart, Germany: Metzler.

18 However, the simulation studies in which he was able to show this applied relatively short instruments (20 items and 40 items). His results indicate that a longer instrument is needed for highly correlated dimensions – longer like the ones used in the present study.

Artelt, C., McElvany, N., Christmann, N., Richter, T., Groeben, N., Köster, J., Schneider, W., Stanat, P., Ostermeier, C., Schiefele, U., Valtin, R., Ring, K., & Saalbach, H. (2005). *Expertise – Förderung von Lesekompetenz* (Bildungsreform No. 17). Berlin/Bonn, Germany: Bildungsministerium für Bildung und Forschung.

Artelt, C., Schiefele, U., Schneider, W., & Stanat, P. (2002). Leseleistungen deutscher Schülerinnen und Schüler im internationalen Vergleich (PISA): Ergebnisse und Erklärungsansätze. *Zeitschrift für Erziehungswissenschaft, 5*(1), 6–27.

Artelt, C., & Schlagmüller, M. (2004). Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In U. Schiefele, C. Artelt, W. Scheider, & P. Stanat (Eds.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 169–190). Wiesbaden, Germany: VS.

Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, J. Tillmann, & M. Weiß (Eds.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 69–137). Opladen, Germany: Leske + Budrich.

Artelt, C., Stanat, P., Schneider, W., Schiefele, U., & Lehmann, R. H. (2004). Die PISA-Studie zur Lesekompetenz. Überblick und weiterführende Analysen. In U. Schiefele, C. Artelt, W. Scheider, & P. Stanat (Eds.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 139–168). Wiesbaden, Germany: VS.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft [Special Issue 14]* (pp. 51–65). Wiesbaden, Germany: VS.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Schümer, G., Stanat, P., Tillmann, J., & Weiß, M. (Eds.). (2002). *PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich: Zusammenfassung zentraler Befunde*. Opladen, Germany: Leske + Budrich.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, J., & Weiß, M. (Eds.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen, Germany: Leske + Budrich.

Baumert, J., Stanat, P., & Demmrich, A. (2001). Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 15–68). Opladen, Germany: Leske + Budrich.

Beck, B., & Klieme, E. (Eds.). (2003). *DESI – Eine Längsschnittstudie zur Untersuchung des Sprachunterrichts in deutschen Schulen. Empirische Pädagogik*. Weinheim, Germany: Beltz.

Beck, B., & Klieme, E. (2007a). Einleitung. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (pp. 1–8). Weinheim, Germany: Beltz.

Beck, B., & Klieme, E. (Eds.). (2007b). *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. Weinheim, Germany: Beltz.

Berendes, K., Weinert, S., Zimmermann, S., & Artelt, C. (2013) Assessing language indicators across the lifespan within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online, 5*(2), 15–49.

Blossfeld, H.-P., Roßbach, H.-G., & Maurice, J. von (Eds.). (2011). *Education as a lifelong process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft [Special Issue 14]*. Wiesbaden, Germany: VS.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R., & Walther, G. (Eds.). (2004). *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich*. Münster, Germany: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G., & Valtin, R. (Eds.). (2003). *Erste Ergebnisse aus IGLU: Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster, Germany: Waxmann.

Bremerich-Vos, A., & Grotjahn, R. (2007). Lesekompetenz und Sprachbewusstheit: Anmerkungen zu zwei aktuellen Debatten. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (pp. 158–177). Weinheim, Germany: Beltz.

Brinker, K. (1983). Textfunktionen. Ansätze zu ihrer Beschreibung. *Zeitschrift für germanistische Linguistik (ZGL), 11*, 127–148.

Brinker, K. (1985). *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden*. Berlin, Germany: Filibach.

Brinker, K. (2010). *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden* (7th ed.). Berlin, Germany: ESV.

Bühler, K. (1934). *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Jena, Germany: Fischer.

Christmann, U., & Groeben, N. (1999). Psychologie des Lesens. In B. Franzmann, K. Hasemann, D. Löffler, & E. Schön (Eds.), *Handbuch Lesen. Im Auftrag der Stiftung Lesen und der Deutschen Literaturkonferenz* (pp. 145–223). München, Germany: Saur.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment* [Adobe Reader version]. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

DESI Konsortium. (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI)*. Frankfurt a. M., Germany: Deutsches Institut für Internationale Pädagogische Forschung (DIPF).

Eberle, F., Gehrer, K., Jaggi, B., Kottonau, J., Oepke, M., & Pflüger, M. (2008). *Evaluation der Maturitätsreform 1995 (EVAMAR): Phase II*. Bern, Switzerland: Staatsekretariat für Bildung und Forschung.

Garbe, C., Holle, K., & Jesch, T. (2009). *Texte lesen. Lesekompetenz – Textverstehen – Lesedidaktik – Lesesozialisation*. Paderborn, Germany: Schöningh.

Gehrer, K., & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In A. Bertschi-Kaufmann & C. Rosebrock (Eds.), *Literalität erfassen: bildungspolitisch, kulturell, individuell*. Weinheim, Germany: Juventa.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for grade 5 and 9). Scientific Use File 2012, Version 1.0.0*. Bamberg, Germany: University of Bamberg, National Educational Panel Study. Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_re_2012_en.pdf

Göttsche, D. (2010). Prosa als Bestimmungskriterium. In R. Zymner (Ed.), *Handbuch Gattungstheorie* (pp. 38–39). Stuttgart, Germany: Metzler.

Groeben, N., & Hurrelmann, B. (2004). *Lesesozialisation in der Mediengesellschaft: Ein Forschungsüberblick*. Weinheim, Germany: Juventa.

Grotlüschen, A., Kretschmann, R., Quante-Brandt, E., & Wolf, K. D. (Eds.). (2011). *Literalitätsentwicklung von Arbeitskräften* (Alphabetisierung und Grundbildung, Vol. 6). Münster, Germany: Waxmann.

Grotlüschen, A., & Riekmann, W. (2011). *leo. – Level-One Studie: Literalität von Erwachsenen auf den unteren Kompetenzniveaus (Presseheft)*. Hamburg, Germany: Universität Hamburg.

Händel, M., Artelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal for Educational Research Online, 5*(2), 162–188.

Harell, L. M. (2009). *Accuracy of global fit indices as indicators of multidimensionality in multidimensional Rasch analysis* (Doctoral dissertation). Blacksburg, VA: Virginia Polytechnic Institute and State University.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie, 216*(2), 89–101.

Hess-Lüttich, E. (Ed.). (1996). *Textstrukturen im Medienwandel*. Frankfurt a. M., Germany: Lang.

Kintsch, W. (1982). *Gedächtnis und Kognition*. Berlin, Germany: Springer (Original work published 1970).

Kintsch, W. (1994). Kognitionspsychologische Modelle des Textverstehens: Literarische Texte. In K. Reusser & M. Reusser-Weyeneth (Eds.), *Verstehen. Psychologischer Prozess und didaktische Aufgabe* (pp. 39–54). Bern, Switzerland: Hans Huber.

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394.

Kirsch, I. S. (1995). Literacy performance on three scales: Definitions and results. In OECD & Statistics Canada (Eds.), *Literacy, economy and society. Results of the first International Adult Literacy Survey* (pp. 27–53). Paris, France: OECD.

Kirsch, I. S. (2001). *The International Adult Literacy Survey (IALS). Understanding what was measured*. Princeton, NJ: Research Publications Office.

Kirsch, I. S., Jungeblut, A., & Mosenthal, P. B. (1998). The measurement of adult literacy. In T. S. Murray, I. S. Kirsch, & L. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first International Adult Literacy Survey*. Washington, DC: US Department of Education, National Center for Education Statistics.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., Schneider, W., & Stanat, P. (Eds.). (2010). *PISA 2009. Bilanz nach einem Jahrzehnt*. Münster, Germany: Waxmann.

Köster, J. (2003). Die Bedeutung des Vorwissens für die Lesekompetenz. In U. Abraham, A. Bremerich-Vos, V. Frederking, & P. Wieler (Eds.), *Deutschdidaktik und Deutschunterricht nach PISA* (pp. 90–104). Breisgau, Germany: Filibach.

Mosenthal, P. B. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88*(2), 314–332.

Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document literacy. *Discourse Processes, 14*(2), 147–180.

Mosenthal, P. B., & Kirsch, I. S. (1998). A new measure for assessing document complexity: The PMOSE/IKIRSCH document readability formula. *Journal of Adolescent and Adult Literacy, 41*(8), 638–657.

Mullis, I. V. S., Martin, M. O., González, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary school in 35 countries*. Chestnut Hill, MA: Boston College.

Naumann, J., Artelt, C., Schneider, W., & Stanat, P. (2010). Reading competence von PISA 2000 bis PISA 2009. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Eds.), *PISA 2009. Bilanz nach einem Jahrzehnt* (pp. 23–72). Münster, Germany: Waxmann.

Nold, G., & Willenberg, H. (2007). Lesefähigkeit. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (pp. 23–41). Weinheim, Germany: Beltz.

OECD – Organisation for Economic Co-Operation and Development. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris, France: OECD.

OECD – Organisation for Economic Co-Operation and Development. (2001). *Knowledge and skills for life: First results from the OECD Programme for International Student Assessment (PISA) 2000*. Retrieved from http://www.oecd.org/dataoecd/44/53/33691596.pdf

OECD – Organisation for Economic Co-Operation and Development. (2002). *PISA 2000 technical report*. Paris, France: OECD.

OECD – Organisation for Economic Co-Operation and Development. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris, France: OECD. Retrieved from www.oecd.org/dataoecd/16/52/35888923.pdf

OECD – Organisation for Economic Co-Operation and Development. (2007). *Kurzzusammenfassung PISA 2006*. Retrieved from www.oecd.org/data oecd/18/35/39715718.pdf

OECD – Organisation for Economic Co-Operation and Development. (2009). *PISA 2009 assessment framework – Key competencies in reading, mathematics, and science*. Paris, France: OECD.

OECD – Organisation for Economic Co-Operation and Development. (2010). *PISA 2009 results: Executive summary [PISA 2009 Ergebnisse: Zusammenfassung]*. Retrieved from www.oecd.org/dataoecd/34/19/46619755.pdf

OECD – Organisation for Economic Co-Operation and Development, & Statistics Canada. (1995). *Literacy, economy and society: Results of the first International Adult Literacy Survey*. Paris, France: OECD.

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

Pohl, S., & Carstensen, C. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online, 5*(2), 189–216.

Prüfer, P., & Rexroth, M. (2005). Kognitive Interviews. *ZUMA How-to-Reihe, 15*, 1–21.

Quathamer, D. (1998). *Kohärenzbildung beim Lesen von Texten – Nutzung und Funktion von Überblicksdiagrammen* (Doctoral dissertation). Duisburg, Germany: Gerhard-Mercator-Universität.

Richter, T., & Christmann, U. (2002). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Eds.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (pp. 25–58). Weinheim, Germany: Juventa.

Roick, T., Frederking, V., Henschel, S., & Meier, C. (2013). Literarische Textverstehenskompetenz bei Schülerinnen und Schülern unterschiedlicher Schulformen. In A. Bertschi-Kaufmann & C. Rosebrock (Eds.), *Literalität erfassen: bildungspolitisch, kulturell, individuell* (pp. 69–85). Weinheim, Germany: Juventa.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2nd ed.). Bern, Switzerland: Hans Huber.

Schaffner, E., Schiefele, U., Drechsel, B., & Artelt, C. (2004). Lesekompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Eds.), *PISA 2003 – Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (pp. 93–110). Münster, Germany: Waxmann.

Scheffel, M. (2010a). Faktualität/Fiktionalität als Bestimmungskriterium. In R. Zymner (Ed.), *Handbuch Gattungstheorie* (pp. 29–31). Stuttgart, Germany: Metzler.

Scheffel, M. (2010b). Theorien des Narrativen. In R. Zymner (Ed.), *Handbuch Gattungstheorie* (pp. 328–331). Stuttgart, Germany: Metzler.

Schnotz, W. (1988). Textverstehen als Aufbau mentaler Modelle. In H. Mandl & H. Spada (Eds.), *Wissenspsychologie* (pp. 299–330). München, Germany: PVU.

Schnotz, W., & Bannert, M. (1999). Einflüsse der Visualisierungsform auf die Konstruktion mentaler Modelle beim Text- und Bildverstehen. *Zeitschrift für Experimentelle Psychologie, 46*(3), 217–236.

Schnotz, W., & Dutke, S. (2004). Kognitionspsychologische Grundlagen der Lesekompetenz: Mehrebenenverarbeitung anhand multipler Informationsquellen. In U. Schiefele, C. Artelt, W. Scheider, & P. Stanat (Eds.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 61–99). Wiesbaden, Germany: VS.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. (2011). Development of competencies across the lifespan. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft [Special Issue 14]* (pp. 67–86). Wiesbaden, Germany: VS.

Werlich, E. (1975). *Typologie der Texte: Entwurf eines textlinguistischen Modells zur Grundlegung einer Textgrammatik*. Heidelberg, Germany: Quelle & Meyer.

Werlich, E. (1976). *A text grammar of English*. Heidelberg, Germany: Quelle & Meyer.

Wieler, P. (2003). Varianten des Literacy-Konzepts und ihre Bedeutung für die Deutschdidaktik. In U. Abraham, A. Bremerich-Vos, V. Frederking, & P. Wieler (Eds.), *Deutschdidaktik und Deutschunterricht nach PISA* (pp. 47–68). Breisgau, Germany: Filibach.

Willenberg, H. (2007). Lesen. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung*. *DESI-Studie (Deutsch Englisch Schülerleistungen International)* (pp. 107–117). Weinheim, Germany: Beltz.

Wu, M. L., Adams, R. J., Wilson, M., & Haldane, S. (2008). *ACER Conquest 2.0: Generalized item response modeling software* [Computer program]. Hawthorn, Australia: ACER.

Zymner, R. (Ed.). (2010). *Handbuch Gattungstheorie*. Stuttgart, Germany: Metzler.

Schrift 3

Gehrer, K. (2017). Der Einfluss von Aufgaben- und Textmerkmalen auf Itemschwierigkeit – Eine experimentelle Studie zur Einschränkung der wiederholten Texteinsicht bei der Bearbeitung von Lesekompetenztestaufgaben (NEPS Working Paper No.67). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Der Einfluss von Aufgaben- und Textmerkmalen auf Itemschwierigkeit

Karin Gehrer
Leibniz-Institut für Bildungsverläufe, Bamberg

Eine experimentelle Studie zur Einschränkung der wiederholten Texteinsicht bei der Bearbeitung von Lesekompetenztestaufgaben

E-Mail-Adresse der Erstautorin:

karin.gehrer@lifbi.de

Bibliographische Angaben:

Gehrer, K. (2017). *Der Einfluss von Aufgaben- und Textmerkmalen auf Itemschwierigkeit – Eine experimentelle Studie zur Einschränkung der wiederholten Texteinsicht bei der Bearbeitung von Lesekompetenztestaufgaben* (NEPS Working Paper No. 67). Bamberg, Deutschland: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Für wertvolle Kommentare bedanke ich mich bei Frau Dr. habil. Kathrin Lockl, Frau Dr. Ilka Wolter, Frau Lena Nusser und allen Kolleginnen und Kollegen des Lese-Kolloquiums, sowie Frau Prof. Dr. C. Artelt.

Der Einfluss von Aufgaben- und Textmerkmalen auf Itemschwierigkeit – Eine experimentelle Studie zur Einschränkung der wiederholten Texteinsicht bei der Bearbeitung von Lesekompetenztestaufgaben

Zusammenfassung

In einer Studie im Rahmen des Nationalen Bildungspanels (NEPS, Blossfeld, Roßbach & von Maurice, 2011) wurde 2014 für die Entwicklung von Lesekompetenztests für Wiederholungsmessungen bei Studierenden und Erwachsenen eine Experimentalbedingung implementiert, welche es ermöglichen sollte, Textverständnisfragen (Items) mit höheren Schwierigkeiten zu generieren (Gehrer, Wolter, Koller & Artelt, in Vorbereitung)[1]. Ausgehend von der Erkenntnis, dass beim Lesen von Texten auf hierarchieniederen und -höheren Ebenen des kognitiven Verarbeitungsprozesses durch ein interaktives Zusammenwirken verschiedener Teilprozesse eine inhaltsspezifische Repräsentation des Textes gebildet wird, die es ermöglicht, den Text als Ganzes zu erinnern bzw. dessen Inhalte zu speichern (Kintsch & van Dijk, 1978; Richter & Christmann, 2002), und unter der Annahme, dass gute Lesende sich unter anderem dadurch von Lesenden mit weniger hohen Leseleistungen unterscheiden, dass es ihnen – vermutlich unter Zuhilfenahme von adäquateren Lese- und Aneignungsstrategien – gelingt, ein brauchbareres Situationsmodell aufzubauen, wurde im computergestützten Assessment eine technische Navigationsrestriktion eingebaut, welche es der Zielperson bei gewissen Texten nur einmal erlaubte, den Text zu lesen. Ein Zurückblättern in den Text, wie normalerweise bei den NEPS-Leseaufgaben möglich, wurde in der Experimentalbedingung auf technischem Wege verhindert. Bezüglich des vermuteten Schwierigkeitszugewinns ließ sich über den Itempool hinweg kein Haupteffekt feststellen. Es waren innerhalb der sechs Texte unter der Experimentalbedingung gewisse Leseitems schwieriger, andere unerwartet leichter geworden (vgl. Gehrer et al., in Vorbereitung).

Darauf baut der folgende Artikel auf, in dem der Fragestellung nachgegangen wird, was sich hinter diesen unsystematischen Schwierigkeitsveränderungen möglicherweise verbirgt. Es wird angenommen, dass sich Textverständnisfragen, die auf Grund der geschilderten Experimentalbedingung eine veränderte Aufgabenschwierigkeit respektive Lösungswahrscheinlichkeit aufweisen, theoriebasiert durch item- oder textspezifische Merkmale systematisieren und beschreiben lassen. Des Weiteren werden differenzielle Effekte bei Gruppen mit unterschiedlichen Lesermerkmalen (z.B. Studierende (n = 372) mit vermutet gut ausgebildeten Lesestrategien oder Personen mit hohen versus niedrigen Lesefähigkeiten) erwartet.

Vertiefende Analysen der insgesamt 72 Items des Experimentalpools mit Klassifikationsbäumen und Regressionen zeigten, dass nicht wie erwartet (auch) Textsorte, Textlänge oder die spezifischen kognitiven Anforderungen der Aufgabenstellungen gemäß der längsschnittlichen NEPS-Lesen-Rahmenkonzeption (vgl. Gehrer, Zimmermann, Artelt & Weinert, 2013; deutsche Kurzfassung 2012) vermehrte Schwierigkeit generieren konnten, sondern dass bei diesen Lesetestaufgaben insbesondere das Aufgabenformat ein itemspezifisches schwierigkeitsgenerierendes Merkmal ist. Insbesondere über das Format

[1] Autorengruppe/Reihenfolge noch nicht final

„Zuordnungsaufgabe“, bei welchem aus mehreren möglichen Überschriften die Wahl eines passenden Zwischentitels zu jedem Abschnitt des gelesenen Textes verlangt wird, konnte gegenüber von klassischen standardisierten Multiple-Choice-Aufgaben (MC; eine richtige Antwort wird aus vier Optionen ausgewählt) und Entscheidungstabellen (für jede Aussage zum Text muss entschieden werden, ob sie richtig ist oder falsch ist) signifikant die Schwierigkeitserhöhung erklärt werden.

Differenzielle Effekte fanden sich für die Gruppe der schlechten Lesenden (*n* = 229), für welche über das Aufgabenformat hinaus die kognitiven Anforderungen von Aufgabenstellungen als Erklärung für die eingetretenen Schwierigkeitsveränderungen bestätigt wurden. Besonders Aufgaben mit der kognitiven Anforderung des Informationenentnehmens (Typ 1) werden für schlechte Lesende unter Navigationsrestriktion schwieriger. In gewissem Maß gilt dies auch für mittlere Lesende (*n* = 449), aber nicht für gute Lesende.

Schlagworte
Lesekompetenzmessung, Testen ohne wiederholte Texteinsicht, Experimentalbedingung, Itemschwierigkeit, Aufgabenformat, Erwachsene und Studierende

Inhaltsverzeichnis

1. Einleitung

Was macht Aufgaben (Items) schwierig, und unter Umständen sogar noch schwieriger? Dies ist nicht nur aus der Perspektive von Testentwicklerinnen und Testentwickler (z.B. Prenzel, Häußler, Rost & Senkbeil, 2002, S. 124; Zimmermann, 2016, S. 4) oder für Lehrpersonen im Zusammenhang mit ihrer aufgabenspezifischen Urteilskompetenz (z.B. Rausch, Matthäi & Artelt, 2015) bedeutsam, sondern auch aus theoretischer Sicht zur Klärung der Konstruktvalidität eines Lesekompetenztests (z.B. Freedle & Kostin, 1993, 1994; Sonnleitner, 2008, S. 346) eine wichtige Frage.

Für die Konstruktion von Instrumenten zur Erfassung von Personenfähigkeiten sind Testfragen relevant, welche valide und reliable Messungen ermöglichen und die Personenfähigkeiten in angemessener bzw. bestmöglicher Weise abbilden. Dies bedeutet für Kompetenzmessungen in heterogenen Gruppen, dass die Aufgabenstellungen bzw. Items sowohl im mittleren Bereich als auch in den unteren bzw. oberen Extrembereichen der möglichen Personenfähigkeiten differenziert und valide und reliabel messen können müssen. Mit anderen Worten, im Pool der Testitems müssen sich sowohl leichte als auch schwierige bis sehr schwierige Aufgabenstellungen befinden, und dies bei hoher Itemgüte (vgl. Rost, 2004). Im Bereich der Lesekompetenz stellt insbesondere die Konstruktion von sehr schwierigen Aufgaben eine besondere Herausforderung dar. Aus auswertungstechnischen Gründen werden häufig geschlossene Aufgabenformate wie Multiple Choice-Fragen oder Entscheidungstabellen gewählt (für die NEPS-Lesekompetenztests siehe deren längsschnittliche Rahmenkonzeption: Gehrer, Zimmermann, Artelt & Weinert, 2013; Kurzfassung mit Beispielen, 2012). Erfahrungsgemäß eignen sich geschlossene Items bestens für den unteren und mittleren Bereich, in einem (größeren) Feldeinsatz sind teilweise offene Formate jedoch hilfreich, um den obersten Fähigkeitsbereich noch gut abbilden zu können (vgl. z.B. Prenzel, 2002, S. 132; Willenberg, 2007, S. 117).

Bei besonders fähigen Personen oder Personengruppen weisen Lesekompetenztests oft Deckeneffekte auf, d.h. dass Fähigkeiten im oberen Bereich nicht differenziert gemessen werden können. Hier ist es für die Kompetenzmessung wichtig, durch bestimmte Item- oder Textmerkmale adäquate Schwierigkeit zu erreichen, um auch im oberen Bereich von Personenfähigkeiten trennscharf messen zu können. Im Theorieteil (Kap. 3) wird ein kurzer Überblick über verschiedene Möglichkeiten der Schwierigkeitsgenerierung in der Testkonstruktion gegeben.

Für eine Entwicklungsstudie im Rahmen des Nationalen Bildungspanels (NEPS, Blossfeld, Roßbach & von Maurice, 2011) wurden neue Lesetestaufgaben konstruiert, um fähigkeitsangemessene Schwierigkeiten insbesondere für die Wiederholungsmessungen bei Studierenden im Bereich der Lesekompetenz zu generieren. In einer Experimentalbedingung konnte überprüft werden, welche Effekte eine Restriktion des Zurückblätterns zum Text auf die Bearbeitung von Lesetestaufgaben hat. Im Theorieteil 2 werden bisherige Befunde zum Testen ohne wiederholte Texteinsicht dargestellt. Auf die Beschreibung der Experimentalstudie, das Studiendesign, die eingesetzten Instrumente und die Ergebnisse der Studie (vgl. Kopp et al., 2016; Gehrer et al., in Vorbereitung) wird im Teil 4 eingegangen. Die Forschungsfragen zur vorliegenden Überprüfung von schwierigkeitsbestimmenden Merkmalen im Rahmen der Experimentalstudie werden unter Punkt 3 und die verwendeten

Methoden im Kapitel 5 näher beschrieben, bevor anschließend die Ergebnisse dargestellt und diskutiert sowie gewisse Limitationen der vorliegenden Analyse erörtert werden.

2. Theorie und Befundlage

Gehen wir von der inzwischen akzeptierten Annahme aus, dass Lesen im Sinne von Textverständnis ein aktiver Prozess ist, der sowohl von Lesermerkmalen als auch von Textmerkmalen beeinflusst wird (vgl. bspw. Artelt, Stanat, Schneider & Schiefele, 2001, 70–73; Voss, Carstensen & Bos, 2005). Die damit verbundenen Aktivitäten bestehen aus teilweise automatisierten hierarchieniederen und hierarchiehöheren Prozessschritten, die sich gegenseitig beeinflussen und letztendlich in ein mentales Modell münden, das den gelesenen Text bzw. seinen Sinn und Inhalt idealerweise bestmöglich abbildet. Es geht somit abschließend darum, ein durch die Lektüre gewonnenes adäquates Textsituationsmodell zu generieren (vgl. Kintsch & van Dijk, 1978; Richter & Christmann, 2002; Schnotz & Dutke, 2004).

Gehen wir als Nächstes davon aus, dass die Lesefähigkeiten, die es ermöglichen, einen Text lesen und angemessen verstehen zu können, sowohl im Ganzen als auch in Einzelaussagen, über den Einsatz von validen und reliablen Lesekompetenztests erfasst werden können (z.B. EVAMAR für Studierende in der Schweiz: Eberle et al., 2008; IGLU für vierte Klassen: Voss et al., 2005; NEPS über die Lebensspanne: Weinert et al., 2011; PISA für 15-Jährige: Artelt, Stanat, Schneider & Schiefele, 2001; Artelt, Stanat, Schneider, Schiefele & Lehmann, 2004).

Die Lesekompetenzmessung erfolgt meistens in der Vorgabe von längeren oder kürzeren Texten, anhand derer die Testperson Fragen oder Aufgaben zu beantworten hat, womit ihr Textverständnis geprüft wird. Dabei ist es wesentlich, wie gut eine Leserin oder ein Leser sich innerlich eine kohärente Repräsentation des Textes bzw. der unterschiedlichen Textpassagen bilden konnte, um die anschließenden Testfragen zu beantworten.

Guten Testpersonen wird es gelingen, eine bestmögliche Repräsentation des Textes zu bilden, indem sie nicht nur auf der Textoberfläche gewisse Signale und Teilinformationen berücksichtigen, sondern auch in einem vertieften Lesen notwendige semantische und syntaktische Bezüge auf Satzebene und zwischen Satzfolgen auf Textebene erfolgreich integrieren und unter Verknüpfung größerer Textteile zu einer globalen Kohärenzherstellung gelangen. Im Zuge eines vertieften Rezeptionsprozesses, der in mehreren zyklischen Phasen erfolgt, werden sie erste Lesehypothesen verwerfen oder bestätigen (vgl. Richter & Christmann, 2002).

Schlechte Lesende werden im Unterschied dazu durch ein eher oberflächliches Lesen und gegebenenfalls mangelnde Wortschatz- und Syntaxkenntnisse sowie misslingender Inferenzenbildung nur eine eher lückenhafte Repräsentation des Textes erreichen. Sie zeichnen sich nach Stanat und Schneider (2004) durch Defizite in leseprozessnahen Faktoren aus, welche ihre Leseleistung beeinträchtigen. So werden erschwerte oder verlangsamte Worterkennungsprozesse, ein kleiner Wortschatz, eine geringere Arbeitsgedächtniskapazität und damit verbundene Defizite bei der Verknüpfung verschiedener Propositionen sowie eine damit einhergehende nicht gelingende Bildung lokaler Kohärenz oder von Makrostrukturen angenommen. Für dritte und vierte deutsche Klassen konnten Wortdekodierfähigkeiten und metakognitives Wissen zur Unterscheidung zwischen guten und schlechten Lesenden empirisch bestätigt werden (van Kraayenoord & Schneider, 1999). Ein geringeres

metakognitives Strategiewissen wurde bei Schülerinnen und Schülern der siebten und achten Klasse gefunden, welche seit mehreren Jahren eine Leseschwäche aufwiesen (Roeschl-Heils, Schneider & van Kraayenoord, 2003). In der Wiener Längsschnittstudie (Klicpera & Gasteiger-Klicpera, 1993) wurde nachgewiesen, dass diese Unterschiede zwischen guten und schwachen Lesenden von der dritten bis zur achten Klasse stabil bleiben.

2.1 Testen ohne wiederholte Texteinsicht

Üblicherweise erhalten Testpersonen die Möglichkeit, bei der Beantwortung der Fragen in den bereits gelesenen Text zurückzublättern. Dadurch können sie während der Aufgabenbearbeitung ihre Textrepräsentation jederzeit mit den neuen Informationen anreichern bzw. präzisieren und weitere notwendige Inferenzen ziehen, um den Anforderungen der konkreten Fragestellung zu genügen (z.B. mehrmaliger Abgleich der Attraktoren und Distraktoren mit gewissen Textstellen). So ist es unter der Bedingung des Blätterns und damit verbundener erneuter Texteinsicht uneingeschränkt möglich, die hierarchieniedrigen Prozessschritte der Worterkennung zu wiederholen und lokale Integrationsprozesse gegebenenfalls zu korrigieren. In der computerbasierten Anwendung können diese individuellen Strategien zur Lösungserarbeitung über die Logdatenanalyse teilweise rekonstruiert werden (z.B. Kopp, Gehrer, Artelt, Wolter & Koller, 2016; Kopp et al., in Vorbereitung). Die Bedingung des Nichtzurückblätterns stellt dem gegenüber andere Anforderungen dar. Die Bedingung des Nichtzurückblätterns bzw. das Testen von Textverständnis ohne wiederholte Texteinsicht ist v.a. in der kognitionspsychologischen Forschung von Interesse und wird beispielsweise bei Studien zum Lernen aus Texten und der Erforschung von Lernprozessen eingesetzt, welche ein Verstehen von Texten voraussetzen (nationaler Ergänzungstest zu PISA 2000: Artelt et al., 2001; Schaffner, Schiefele & Schneider, 2004). Daneben gibt es eine Reihe von Studien zur Überprüfung der Konstruktvalidität von (Multiple-Choice-) Leseverständnistests, bei welchen die experimentelle Bedingung des Testens ohne wiederholte Texteinsicht bzw. gänzlich ohne Textvorlage (Katz, Lautenschlager, Blackburn & Harris, 1990; Preston, 1964; Rost & Sparfeldt, 2007; Schroeder & Tiffin-Richards, 2014) genutzt wird.

Schaffner et al. (2004) berichten für narrative und Sachtexte den Befund einer Schlüsselrolle der kognitiven Grundfähigkeit innerhalb mehrerer Personenmerkmale (z.B. Bildungsabschluss Eltern, soziales Kapital, Selbstkonzept, Interesse, Lesemotivation), welche zusätzlich zum direkten Effekt ($\beta = .44$ bzw. $\beta = .54$ bei Erzähltexten) neben der bei Erzähltexten ebenfalls bedeutsamen Lesemotivation auch indirekt über andere kognitive Faktoren (Dekodierfähigkeit, thematisches Vorwissen) hinsichtlich des Testergebnisses des nationalen PISA-Lesetests 2000 ohne weitere Texteinsicht wirkt. Metakognitives Strategiewissen hat ebenfalls einen direkten, wenn auch kleineren Effekt ($\beta = .14$ bzw. $\beta = .22$ bei Erzähltexten) (228– 234). Ihre theoretische Annahme hinter dem Konzept „Testen ohne wiederholte Texteinsicht“ beschreiben sie wie folgt: Kann der Text in der Phase der Beantwortung der Testfragen nicht mehr eingesehen werden, führt dies dazu, dass eventuelle Lücken in der Textrepräsentation nicht mehr ad hoc geschlossen werden können (Schaffner et al., 2004, 197–198). Dadurch können auch Fehlinterpretationen nicht mehr korrigiert werden (vgl. Kintsch, 1994).

In amerikanischen Studien der Achtzigerjahre zur Beantwortung von Leseverständnisfragen mit oder ohne mögliche Texteinsicht (Davey, 1987; Garner & Reis, 1981) unterscheiden sich gute und schlechte jugendliche Lesende darin, wie effektiv sie unter der Bedingung des Zurückblätterns den Text nutzen, um zur richtigen Lösung zu gelangen. Während sich gute und schlechte Lesende unter der Bedingung ohne wiederholte Texteinsicht nicht unterscheiden, konnten die guten Lesenden bei der Möglichkeit des Zurückblätterns mehr Aufgaben korrekt beantworten, d.h. sie konnten das Zurückblättern effektiver nutzen (Davey, 1987). Garner und Reis (1981) zeigten, dass jüngere Kinder insgesamt das Zurückblättern eher selten und wenig angemessen nutzen. Jedoch etwas ältere Kinder (achte Klasse), die gleichzeitig gut lesen können, zeigten bereits einen besseren Umgang mit dem Zurückblättern und konnten davon beim Lösen der Leseaufgaben profitieren (Garner & Reis, 1981).

Das Testen ohne Zurückblättern und somit ohne wiederholte Texteinsicht wurde 2014 auch in einer Entwicklungsstudie des Nationalen Bildungspanels zur Erfassung von Lesekompetenz im Studierenden- und Erwachsenenalter eingesetzt. Gewisse Items wurden hier in einer Experimentalbedingung ohne wiederholte Texteinsicht dargeboten (vgl. Gehrer et al., in Vorbereitung). Die deskriptiven Ergebnisse der Studie zeigten, dass unabhängig von der veränderten Kontextbedingung die meisten der eingesetzten Items bezüglich Trennschärfen, Item Fit und den item-charakteristischen Kurven (vgl. Pohl & Carstensen, 2012) qualitativ hochwertig waren und sich bezüglich Reliabilität und Validität zwischen den Gruppen nicht unterschieden. Die Items wiesen zu großen Teilen kein Differential-Item-Functioning (DIF) auf und waren somit messinvariant über die verschiedenen Bedingungen. Auch zeigte sich bei Überprüfung der Dimensionalität über die beiden Gruppen hinweg weiterhin die konzeptionelle Eindimensionalität des Testes, so dass davon ausgegangen werden kann, dass unter beiden Bedingungen das gleiche Konstrukt gemessen wurde. Hinsichtlich des vermuteten Schwierigkeitszugewinns ließ sich über den Itempool hinweg kein Haupteffekt der Bedingung Navigationsrestriktion (Nichtzurückblättern) feststellen. Es waren innerhalb der sechs Texte unter der Experimentalbedingung des Nichtzurückblätterns gewisse Items schwieriger, andere unerwartet leichter geworden. Ein systematischer Effekt der computerbasierten Bedingung ohne wiederholte Texteinsicht konnte allerdings nicht gefunden werden (Kopp et al., 2016; Gehrer et al., in Vorbereitung).

Als differenzieller Effekt auf Personenebene konnte eine deutlich erhöhte Erstlesezeit bei fähigen Personen unter der Experimentalbedingung durch Kopp et al. (2016) anhand der Analyse von Logdaten bereits nachgewiesen werden. Auf anschließende Analysen auf Itemebene, insbesondere der eingetretenen Schwierigkeitsveränderungen, bezieht sich der vorliegende Beitrag.

In Anlehnung an die erwähnten Befunde von Schaffner et al. (2004) wird dem metakognitiven Strategiewissen in der experimentellen Bedingung des Nichtzurückblätterns eine signifikante Bedeutung zugesprochen. Vor dem Hintergrund der Befunde von Davey (1987) und Garnier und Reis (1981), welche gezeigt haben, dass gute Lesende das Zurückblättern häufiger und effektiver als schlechte Lesende nutzen und das wiederholte Lesen im Text zum Verifizieren von Antwortoptionen strategisch häufig einsetzen, kann in einer Bedingung des Nichtzurückblätterns davon ausgegangen werden, dass gute Lesende in ihrem üblicherweise gezeigten natürlichen strategischen Verhalten stark eingeschränkt sein werden. Ein adaptives Reagieren auf unterschiedliche Aufgabenanforderungen, die während

der Testbearbeitung gestellt werden, wird unter dieser eingeschränkten Kontextbedingung nicht möglich sein. Da gute Lesende ein besseres metakognitives Strategiewissen aufweisen (Roeschl-Heils et al., 2003; van Kraayenoord & Schneider, 1999), wird die veränderte Kontextbedingung einen größeren Einfluss auf sie haben, weil sie im Anwenden der bekannten Strategien eingeschränkt sind.

Schlechte Lesende haben im Unterschied dazu ein geringeres metakognitives Strategiewissen (Roeschl-Heils et al., 2003; van Kraayenoord & Schneider, 1999) und machen von der strategischen Bearbeitung von Leseverstehensaufgaben in der Bedingung des möglichen Zurückblätterns zum Text kaum Gebrauch (Davey, 1987; Garnier & Reis, 1981). Somit werden schlechte Lesende auch unter einer eingeschränkten Bedingung, sich nicht anders verhalten, als sie es auch mit Texteinsicht tun. Damit bleiben für schlechte Lesende die Anforderungen der Aufgabenstellungen unter beiden Bedingungen gleich und somit sollten sich die Itemschwierigkeiten für diese Gruppe unter der eingeschränkten Bedingung nicht verändern.

2.2 Was generiert Schwierigkeit?

Es wird theoretisch zwischen einer textbezogenen Gruppe und einer lösungs- oder aufgabenbezogenen Gruppe von Merkmalen, welche Schwierigkeit vorhersagen bzw. generieren (können), unterschieden (z.B. Embretson & Wetzel, 1987; Schweitzer, 2007; Sonnleitner, 2008); andere AutorInnen wie Freedle und Kostin (1994) nehmen zusätzlich auch eine dritte Gruppe von Interaktionsvariablen zwischen Items und Text wie die Verteilung der Distraktoren über die Passagen des Textes in das Kategoriensystem mit auf.

Im konkreten Fall von Lesekompetenztests können auf der ersten Ebene der Items mehrere Faktoren als schwierigkeitsbeeinflussend in Frage kommen (z.B. Freedle & Kostin, 1993, Tabelle S. 153): Es werden das Format, die Fragestellung, der Stamm der Aufgabenstellung, die Formulierung (Länge, Komplexität, Wortschatz) der Lösung und der Distraktoren sowie deren inhaltliche und sprachliche Nähe zueinander als schwierigkeitsbeeinflussend postuliert. Wie auch der Entscheidungsspielraum, der Präzisionsgrad und der verlangte Integrationsgrad der Aufgabenstellung für PISA-Aufgaben als schwierigkeitsgenerierend empirisch belegt sind (Artelt, Stanat, Schneider, Schiefele & Lehmann, 2004; vgl. Schweitzer, 2007).

Auf der Ebene des kontinuierlichen Textes sind es dessen Komplexität, seine Ein- oder Mehrdeutigkeit, das Anspruchsniveau des verwendeten Wortschatzes (z.B. Ozuru, Rowe, O'Reilly & McNamara, 2008) und der Syntax, die Satzlänge, die Strukturiertheit des Textes, seine Länge (z.B. OECD, 2009, S. 45; 2013, S. 69), die propositionale Dichte (z.B. Kintsch, 1994; Zimmermann, 2016) und je nach Textsorte auch Argumentations- oder Handlungsverflechtungen, welche das Lesen anspruchsvoller machen und für welche somit zusätzliche Schwierigkeit für die darauf bezogene Fragestellung angenommen wird.

Für die Ebene der Interaktion zwischen Aufgabenstellung und Text konnte beispielsweise die Verankerung der Distraktoren im Text (örtlich gesehen als Verteilung über den Text oder Verdichtung in einer Passage) als beeinflussender Schwierigkeitsprädiktor gefunden werden (z.B. Freedle & Kostin, 1993, S. 163). Auch Kirsch (2001) findet bei der Betrachtung der Plausibilität der Distraktoren für ihre Textstellenverankerung (Nähe zur Lösung) einen Schwierigkeitseffekt. Es muss insgesamt von einem mehrdimensionalen, komplexen

Einflussmodell ausgegangen werden, bei dem verschiedenste Merkmale von Item, Text sowie deren Interaktion bezüglich der Frage nach Schwierigkeit von Leseitems gemeinsam betrachtet werden sollten (Freedle & Kostin, 1993, S. 167)[2].

In der amerikanischen und englischen Forschung gibt es mehrere empirische Studien, welche für Englisch als Mutter- bzw. Fremdsprache innerhalb von bis zu 100 Prädiktoren die folgenden vier als ausschlaggebende schwierigkeitsbestimmende Merkmale nachweisen konnten: a) Wortschatz, b) propositionale Dichte, c) Plausibilität der Distraktoren und d) kognitive Anforderung der Aufgabenstellung (u.a. Embretson & Wetzel, 1987; Kirsch, 2001; Nold & Rossa, 2007; zusammenfassend zitiert nach Zimmermann, 2016, 6–10). Indessen konnten einzelne Studien mit experimenteller Variation von propositionaler Dichte verschiedener Textpassagen oder Verwendung eines schwierigeren Wortschatzes (Passivkonstruktionen und negative Formulierungen) beispielsweise für Studierende bei der Testbearbeitung am Computer keine schwierigkeitsgenerierenden Effekte replizieren (Gorin, 2005).

Eine bisher seltene Replikation für das Deutsche unternahm Zimmermann (2016) für die aus der Literatur identifizierten Hauptprädikatoren aller drei Ebenen Item, Text und deren Verknüpfung auf der Basis ausgewählter Lesetestaufgaben für die neunten Klassen des Nationalen Bildungspanels (*N* = 13 898). Dabei konnte er für die meisten der ausgewählten Merkmale die Schwierigkeitsvorhersage für den deutschsprachigen NEPS-Lesekompetenztest hypothesenkonform bestätigen (S. 21). Dies gilt für die Gebräuchlichkeit des Vokabulars (sowohl im Text als auch im Item), die propositionale Dichte (d.h. die Relation zwischen der Anzahl von Propositionen zur Textlänge), die semantischen Überschneidungen [3] von Lösung-Distraktoren-Stimulustext und die Verneinung im Aufgabenstamm als eine qualitativ unterschiedliche kognitive Anforderung insbesondere im Jugendlichenalter (S. 29). Zimmermann konnte mit diesem Modell mit acht signifikanten Prädiktorvariablen ungefähr die Hälfte der Varianz der Aufgabenschwierigkeiten erklären. Jedoch arbeitete auch er unter Ausschluss der Betrachtung von verschiedenen Aufgabenformaten (nur Multiple-Choice-Format ausgewählt).

2.2.1 Aufgabenmerkmale

Aus den oben erwähnten Merkmalen, welche auf der Ebene der eigentlichen Testaufgaben (Items) eines Leseverständnistests als schwierigkeitsgenerierende vermutet werden, interessiert uns aufgrund der bei NEPS-Studien vorhandenen Kodierung entlang der Rahmenkonzeption (Gehrer et al., 2013) das Merkmal des Aufgabenformates. Die kognitiven Anforderungen der Items (Kodierung als Typen) werden unter dem Abschnitt der Interaktion mit dem Text betrachtet.

[2] Freedle und Kostin (1993) beziehen beispielsweise 13 Itemvariablen, 34 Textvariablen und 28 Text-by-Item-Variablen in die Analysen mit ein (S. 155). Als signifikant für die Schwierigkeitssteigerung erweisen sich davon acht Merkmale, wovon nur ein einziges ein reines Item-Merkmal (Verneinung in der Lösung) ist (S. 162).

[3] Einzig das Merkmal „Semantische Überschneidung von Lösung und Aufgabenstamm“ führt paradoxerweise zu einer erhöhten Aufgabenschwierigkeit, was Zimmermann im Rahmen von Teststrategie diskutiert (S. 21).

Aufgabenformat

Theoretische Annahmen sowie empirische Befunde bei Large Scale-Assessments, dass für das Aufgabenformat ein schwierigkeitsgenerierender Einfluss vermutet werden kann, sind einige zu finden.

Für die geschlossenen oder gebundenen Aufgabenformate (Ingenkamp, 2005, 110–117; Rost, 2004, 61–64) ist basierend auf dem Aufgabenbearbeitungsmodell von Embretson & Wetzel (1987) anzunehmen, dass Mehrfachwahlaufgaben mit vielen Antwortoptionen (*eine* richtige Antwort aus z.B. *sechs* Optionen auswählen) sowie Mehrfachwahlaufgaben mit mehreren richtigen Antworten (Multiple-Choice mit z.B. *zwei* richtigen Antworten innerhalb von fünf Optionen[4]) gegenüber einfachen Multiple-Choice-Aufgaben, bei welchen *eine* richtige Antwort aus wenigen Optionen zu finden ist, eine höhere Schwierigkeitsanforderung darstellen (auch Davey, 1987; vgl. z.B. Sonnleitner, 2008, S. 349). Dies beruht auf dem Postulat, dass für jede Antwortoption davon ausgegangen werden kann, dass sie einzeln gegenüber dem Text falsifiziert oder verifiziert werden muss (Davey, 1987, S. 262; Embretson & Wetzel, 1987, 178–179; Rost, 2004, 62–63).

Empirische Befunde dafür, dass für das Aufgabenformat ein schwierigkeitsgenerierender Einfluss angenommen werden kann, finden wir bei älteren amerikanischen Studien zum Leseverstehen (z.B. Bormuth, 1967;Kendall, Mason & Hunter, 1980, 1980; Rankin & Culhane, 1969). So erweist sich das Format der Multiple-Choice-Items bei Kindern der fünften Klasse leichter gegenüber von Aufgaben, bei welchen die Kinder Lücken schriftlich ausfüllen sollen (Cloze-Tasks: Bormuth, 1967; Rankin & Culhane, 1969). Aber auch Lückenaufgaben, bei denen aus (drei) Optionen eine richtige ausgewählt werden kann (Maze-Task), erweisen sich gegenüber dem offenen Cloze-Lückenformat als leichter. Ebenso sind Aufgaben, bei denen gelesene Passagen frei mündlich wiedergegeben sollen (recall-Tasks) für Kinder schwieriger als Multiple-Choice-Items (Kendall et al., 1980).

Hinweise, dass Ähnliches auch für das Deutsche vermutet werden kann, finden wir heutzutage beispielsweise bei Prenzel und Kollegen (2002) für die deutschsprachigen nationalen und internationalen PISA-2000-Naturwissenschaftsaufgaben. Es werden hier in einer Regressionsanalyse von mehreren Schwierigkeitsprädiktoren formaler, kognitiver und wissensbasierter Art die offenen Antworten (lang, kurz) als schwierigkeitsgenerierende Merkmale auf dem dritten bzw. sechsten Rang (*kurze* offene Antworten) identifiziert.

Inwieweit sich dieser empirische Befund von schwierigkeitsgenerierenden Effekten des Aufgabenformates auf andere Domänen, insbesondere Lesen und auf andere Altersstufen übertragen lässt, ist in dieser Deutlichkeit weitgehend ungeprüft, jedoch finden sich Hinweise hinsichtlich der Übertragbarkeit für das offene Format in der Grundschule für den Bereich „Sprache und Sprache untersuchen“ bei VERA 3 (vgl. Isaac & Hochweber, 2011) und für den Lesetest bei IGLU (Blatt & Voss, 2005; Bos, Valtin, Voss, Hornberg & Lankes, 2007), sowie für die Sekundarstufe im Deutsch-Lesetest bei DESI (Willenberg, 2007). Auf der Oberstufe ist bei einer Kompetenzstufenmodellierung von TIMMS-Geometrieaufgaben durch Watermann und Klieme (2006) auffallend, dass von drei Aufgaben, welche auf der höchsten Kompetenzstufe messen, zwei im (langen) offenen Format gehalten sind (Watermann & Klieme, 2006).

[4] Auch als „Pick any-out of n“-Format spezifiziert, wenn die Anzahl der auszuwählenden Lösungen nicht angegeben ist (Rost, 2004, S. 64).

Für den vorliegenden Beitrag werden in Anlehnung an die Beobachtungen bei VERA, IGLU, DESI, TIMMS (schwierige Aufgaben sind mehrmalig im offenen Format konstruiert) und die Analysen bei PISA (offenes Format als ein ernst zu nehmender Schwierigkeitsprädiktor: Prenzel et al., 2002) sowie ältere amerikanische Forschung (z.B. Kendall et al., 1980) Effekte des Aufgabenformates ebenso vermutet. In die folgenden Analysen werden deshalb nicht nur die Multiple-Choice-Aufgaben, sondern alle drei bei NEPS (bisher) vorhandenen geschlossenen Formate aufgenommen (NEPS-Beispiele dazu siehe Abschnitt 5.1). Auf deren spezielle Prozesse im Unterschied zu der oben beschriebenen Multiple-Choice-Aufgabe wird nun im Folgenden kurz eingegangen.

Mit Rupp, Ferne und Choi (2006) wird angenommen, dass unterschiedliche Arten von geschlossenen Formaten (Multiple Choice, Entscheidungstabellen, Zuordnungsaufgaben) jeweils spezifische Prozesse und Strategien erfordern.

Innerhalb der geschlossenen Aufgabenformate stellen Entscheidungstabellen mit mehreren Zeilen gegenüber Multiple-Choice-Items andere bzw. vermehrte kognitive Anforderungen, indem sie ausführlichere Abgleichungsstrategien erfordern. Während bei den Multiple-Choice-Items für jede der (beim NEPS-Lesekompetenztests standardisierten) vier Antwortoptionen ein Verifizierungs- bzw. Falsifizierungsprozess durchlaufen wird, um daraus die *eine* wahrscheinlichste Antwort auszuwählen (vgl. Embretson & Wetzel, 1987; Rost 2004), müssen bei Entscheidungstabellen mit ebenso vielen Zeilen zwar ebenso viele Verifizierungs- bzw. Falsifizierungsprozesse durchlaufen werden, wobei jedoch im Unterschied zur Multiple-Choice-Aufgabe die Testperson sich auf *vier* wahrscheinlichste Antworten festlegen muss. Bei Entscheidungstabellen mit fünf oder sechs Zeilen erhöht sich dementsprechend die Zahl der Entscheidungsprozesse.

Dieser theoretischen Annahme einer erhöhten Komplexität des beschriebenen Aufgabenformates Entscheidungstabelle entspricht auch das gewählte Scoring-Verfahren der NEPS-Kompetenztestung: Hier werden Entscheidungstabellen als komplexe Items behandelt (Complex Multiple-Choice, CMC) und gehen als Partial-Credit-Items in das Item-Response-Modell ein. So wird eine Entscheidungstabelle mit mehreren Zeilen insgesamt höher gewichtet als eine simple Multiple-Choice-Aufgabe, indem jeder Zeile oder „Unteraufgabe“ (subtask) ein eigener halber (Gewichtungs-) Punkt zugewiesen wird (vgl. Pohl & Carstensen, 2012, 6–8).

Auch für die Zuordnungsaufgabe müssen auf der Annahme von unterschiedlichen Prozessen bei unterschiedlichen Formaten (Rupp et al., 2006) komplexe Abgleichungsprozesse vermutet werden. Während bei Entscheidungstabellen für jede Zeile (Unteraufgabe) davon ausgegangen wird, dass sie einzeln gegenüber dem Text zu falsifizieren oder verifizieren ist, wird für das Format „Zuordnungsaufgabe“ in Anlehnung an das Embretson & Wetzel-Prozessmodell angenommen, dass jeder optionale Zwischentitel mit jedem Textabschnitt abgeglichen wird. Der Unterschied zum individuellen Abgleichen der beiden anderen geschlossenen Formate besteht jedoch darin, dass nicht auf eine spezielle Passage, welche beim Screening häufig aufgrund eines ähnlichen Vokabulars als informationstragend identifiziert wird, fokussiert werden kann, sondern dass bei der komplexeren Zuordnungsaufgabe jeder Abschnitt auf Übereinstimmung mit den möglichen Überschriften geprüft werden muss. Jede Überschrift verdeutlicht eine Kernaussage der jeweiligen Textpassage; diese muss von den Lesenden in einer Art innerer Zusammenfassung erst als

mentale Repräsentation des Abschnittes auf einer Makro-Struktur-Ebene hypothetisch konstruiert werden (vgl. Kintsch 1978, 1994), um dann in weiteren Prozessschritten mit den möglichen Überschriften abgeglichen zu werden. Erst im weiteren Verlauf können dann die von Embretson & Wetzel (1987) beschriebenen Falsifizierungsentscheidungen bezüglich der unpassenden Überschriften und anschließenden Verifizierung der richtigen Überschrift erfolgen. Im Unterschied zu einer einfachen Multiple-Choice-Aufgabe oder komplexen Entscheidungstabelle ist hier also eine von vornherein erfolgreiche lokale und globale Kohärenzbildung für jede der verschiedenen Passagen des Textes notwendig, um die besonderen Anforderungen dieses komplexen Formates zu erfüllen.

2.2.2 Textmerkmale

Neben den beschriebenen Aufgabenmerkmalen wird davon ausgegangen, dass auch die für den Textverständnistest eingesetzten Texte wesentlich zur Generierung von Schwierigkeit beitragen. Zum Thema Textschwierigkeit bzw. Verständlichkeit als textimmanente Eigenschaft gibt es seit dem frühen 20. Jahrhundert viel empirische Forschung (für einen Überblick z.B. Mrazek, 1979, 36–50). Nach der „Frühzeit“ der quantitativen Wortschatzforschung in den 20er-Jahren wurden in der „Blütezeit (1938-1953)“ (Mrazek, 1979, S. 45) über dreißig Lesbarkeitsindices (u.a. FLESCH, 1948) entwickelt, von welchen die meisten neben Wortparametern über Satzlänge-Parameter die Schwierigkeit von Texten zu erfassen versuchen und meist hoch miteinander korrelieren. Aus der jüngeren Forschung konnte für den Wortschatz des Englischen (üblich/nicht-üblich) empirisch nachgewiesen werden, dass er bei Testaufgaben für jüngere Schülerinnen und Schülern (5. bis 7. Klassen) noch als Schwierigkeitsprädiktor wirkt, jedoch nicht mehr bei den älteren Jugendlichen der Oberstufe (Ozuru et al., 2008), bzw. dass das Vokabular weniger starken Effekt hat auf die Schwierigkeit der Items als die kognitiven Anforderungen (für den Englisch-Test bei DESI: Hartig & Frey, 2012).

Die Besonderheiten der deutschen Sprache (längere Wörter, längere Sätze als das Englische) wurden erst kaum (Mrazek, 1979, S. 50), dann beispielsweise durch den Lix-Index (Langwörter über 6 Buchstaben; Björnsson, 1968) berücksichtigt; diese sowie auch die Wiener Sachtextformel (Bamberger & Rabin, 1984) sind inzwischen automatisiert erhältlich.

Für das Deutsche ging Groeben (1971 bzw. 1978) von den vier Schwierigkeit bestimmenden Textdimensionen grammatisch-stilistische Einfachheit, semantische Dichte, kognitive Strukturierung und motivierender konzeptueller Konflikt (S. 150) aus, ähnlich dazu die Hamburger Forschergruppe um Langer und Kollegen (1974, zit. nach Groeben). Einiges später bezieht sich Willenberg (2010) im Zusammenhang mit dem DESI-Projekt bei literarischen und Sachtexten auf sechs relevante Textaspekte: die Satzlänge im Drei-Sekunden-Fenster, den Wortschatz auf den vier Ebenen Basiswörter/ Konkreta/ Abstrakta/ Fachwörter, die Junktoren (dabei v.a. Konjunktionen), Redundanz auf der Basis von Schlüsselwörtern, literarisierende Merkmale und „Verlebendigung“ (Texte mit Personen, Beispielen, emotionalen Aspekten), wobei er den Wortschatz insgesamt als ausschlaggebend für die Schwierigkeit identifiziert (S. 105).

Für die Kohärenz des Textes, welche bspw. bei Freedle & Kostin (1999), Just und Carpenter (1980) und Kobayashi (2002) im Englischen einen signifikanten Einfluss auf die Itemschwierigkeit nimmt (zit. nach Sonnleitner, 2008, S. 359), findet bspw. Sonnleitner bei einem deutschsprachigen Lesetest für Erwachsene (LEVE-E) keine signifikante Variable,

jedoch kann er im Deutschen für die propositionale Dichte bzw. Komplexität der Texte nach Kintsch und Keenan (1973) bzw. Graesser und Kollegen (1994) einen schwierigkeitssteigernden Effekt bestätigen (2008, Tabelle S. 358) – dies jedoch bei einem Test, der ausschließlich im Multiple-Choice-Format abgefragt wird.

Der Textsorte, welcher in vorliegender Arbeit als kodierte Prädiktorvariable aufgenommen ist, wird von theoretischer Seite viel Bedeutung zugemessen (vgl. Gehrer & Artelt, 2013). Textsorten sind definiert als „konventionell geltende Muster für sprachliche Handlungen" (Brinker, 2010, S. 135) und für ihre Rezeption werden jeweils spezifische Anforderungen beschrieben (vgl. zum Überblick: Gehrer & Artelt, 2013; für Sachtexte: Christmann & Groeben, 2002; für literarische Texte: Eggert, 2002; für kommentierende: Eggs, 1996; für Werbetexte: Janich, 2010; für Anleitungen: Nickl, 2001). Sachtexte im engeren Sinne als erklärende Textsorte, welche informiert und über Sachverhalte berichtet, so wie ihn die NEPS-Rahmenkonzeption konzeptionalisiert, vermitteln in einfacherer Sprache als Fachtexte Alltagswissen von Experten an Laien. Bei kommentierenden Texten ist mit einer minimal argumentativen Textstruktur bis zu einer elaborierten Argumentationsstruktur und somit erhöhten Anforderungen zu rechnen. Bei literarischen Texten wird aus kognitionspsychologischer Theoriebildung (Kintsch, 1994) davon ausgegangen, dass aufgrund ihrer Mehrdeutigkeit und Offenheit komplexe mehrschichtige Situationsmodelle auf mehreren Ebenen gebildet werden müssen (Gehrer & Artelt, 2013).

2.2.3 Interaktion Items mit Text

Für die dritte Kategorie von Prädiktorvariablen, welche insbesondere die Verknüpfung von Items und Text, also deren Interaktion erfassen, werden von Freedle und Kostin (1993) wesentlich mehr Variablen als signifikant gefunden als bei reinen Item- oder Text-Merkmalen, weshalb sie die Bedeutung dieser Kategorie hervorheben. Während sie sich jedoch mehr auf semantische Überlappungen zwischen Item und Text konzentrieren und beispielsweise die Anzahl von gleichen Wörtern in den Antwortoptionen und Textpassagen abgleichen, sollen in der vorliegenden Arbeit insbesondere die kognitiven Anforderungen als Prädiktorvariablen der Interaktion zwischen Aufgabenstellung und Text betrachtet werden.

Kognitive Anforderungen der Items

Die kognitiven Anforderungen der Items sind die Verstehensanforderungen, welche mit dem Lösen der Aufgaben verbunden sind. Meist werden diese entlang des angenommenen Informationsverarbeitungsprozesses beschrieben und systematisiert. So unterscheidet bspw. der International Adult Literacy Survey (IALS) nach Informationen lokalisierenden, integrierenden, generierenden und zyklischen Aufgabenstellungen (Kirsch, 2001, 15–16). Bei vielen Large Scale-Assessments werden für die kognitiven Anforderungen der Aufgabenstellungen schwierigkeitsgenerierende Einflüsse vermutet bzw. nachgewiesen, so findet z.B. Kirsch für IALS, dass komplexere Anforderungen des Matchings die Schwierigkeit der Aufgaben in einem erheblichen Maß mitbestimmt. Hartig und Frey (2012, S. 46) finden für den Englisch-Leseverstehenstest von DESI die kognitive Anforderung einer Frage gegenüber dem Vokabular des Textes (Textebene) und der Plausibilität der Distraktoren[5]

[5] Hartig und Frey (2012) kategorisieren interessanterweise sowohl die Plausibilität der Distraktoren der Multiple-Choice-Items als auch die kognitiven Anforderungen der Aufgaben auf der Itemebene.

(hier Itemebene) als stärkste Prädiktorvariable, insbesondere die kognitive Anforderung komplexes Schlussfolgern und Inferenzen ziehen.

Für das Deutsche wurden die kognitiven Anforderungen als weitere Analysekategorie von Prenzel und Kollegen (2002) neben formalen und wissensbezogenen Aufgabenmerkmalen beim Lösen von textlastigen naturwissenschaftlichen Testaufgaben erfasst. Insgesamt konnten dadurch 45% Varianz der Itemschwierigkeiten erklärt werden (S. 120). Es wurden für den nationalen Test insgesamt fünf kognitive Aspekte naturwissenschaftlicher Kompetenz erfasst (S. 121), als schwierigste kognitive Anforderung zeigte sich „Etwas ausrechnen“, während Textinformationen zu verarbeiten und logisch zu verknüpfen, sich als leichter erwies (129–132).

Für den Lesekompetenztest auf Deutsch werden bei DESI die Kompetenzniveaus auf Basis der empirischen Aufgabenschwierigkeiten auf vier Niveaus differenziert und inhaltlich über die Beschreibung ihrer kognitiven Anforderungen vorgenommen: So ist die Aufgabe des Identifizierens einfacher Lexik als „Fähigkeit, sinntragende Wörter im Text zu finden“ (Willenberg, 2007, S. 109) die einfachste Anforderung, gefolgt von der „lokalen Lektüre“, welche als Fähigkeit, Inferenzen zwischen Sätzen zu bilden oder den Fokus auf schwierigere Stellen zu richten“ (S. 110) das Kompetenzniveau 2 bildet. Die Anforderung der „verknüpfenden Lektüre“ (Verbindung auseinanderliegender Textstellen herstellen) definiert das zweithöchste Kompetenzniveau und die Fähigkeit, ein mentales Modell zu bilden und damit „eine innere Repräsentation wesentlicher Textaspekte“ (S. 110) zu haben, enthält als oberstes Kompetenzniveau die schwierigsten Aufgaben. Somit werden auch beim DESI-Deutschtest „Lesen“ die kognitiven Anforderungen als schwierigkeitsbeeinflussend bzw. sogar schwierigkeitshierarchisch beschrieben (Willenberg, 2007, 109–111). Ähnliches finden wir beim DESI-Test „Argumentation“ auf Deutsch, bei dem die kognitive Anforderung „Reflexion“ sich als schwieriger als die Anforderung „Einschätzung der Situation“ erweist und somit als oberes Kompetenzniveau definiert wurde (Willenberg, Gailberger & Krelle, 2007, 122–125).

Für die in der folgenden Arbeit zu untersuchenden NEPS-Lesekompetenztests wird gemäß Rahmenkonzeption für die definierten kognitiven Anforderungen keine hierarchische Schwierigkeitsstufung angenommen, sondern es werden für jeden Anforderungstyp sowohl schwierige als auch leichte Aufgaben eingesetzt (vgl. Gehrer et al., 2013, 62–63). Dies trifft zu für den üblichen Fall, in welchem die Lesenden nach Belieben beim Bearbeiten der Aufgaben in den Text zurückblättern können. Für den Fall der vorliegenden Studie mit einer Experimentalbedingung ohne wiederholte Texteinsicht wird vermutet, dass sich insbesondere informationsentnehmende Anforderungen als schwieriger erweisen als mit Zurückblättern in den Text, dass es aber dennoch nicht zu einer hierarchischen Schwierigkeitsstufung kommen wird. Items mit der Anforderung (nebensächliche) Detailinformationen aus dem Text zu entnehmen, wurden von vornherein nicht mit in die Experimentalbedingung mit aufgenommen.

3. Forschungsfragen der Analyse

Wenn auch Schaffner und ihre Kollegen (2004) im erwähnten PISA-Ergänzungstest nicht den Einfluss itemspezifischer Merkmale auf das Testergebnis untersuchen, so kann doch ihre hinter dem Konzept „Testen ohne wiederholte Texteinsicht“ stehende theoretische Annahme ebenso für unsere Experimentalstudie mit derselben Bedingung gelten: Lücken in der Textrepräsentation können nicht mehr geschlossen werden (Schaffner et al., 2004, 197–198) und Fehler in der Interpretation des Textes nicht mehr korrigiert werden (vgl. Kintsch, 1994).

Dies führt zu den hypothetischen Annahmen für die Experimentalstudie, dass unter der Bedingung des Nichtzurückblätterns 1) die Beantwortung von Fragen zum Text grundsätzlich schwieriger werden (nicht bestätigende Ergebnisse dazu vgl. Kopp et al., 2016; Gehrer et al., in Vorbereitung), dabei insbesondere 2) gewisse Fragen schwieriger werden, welche besondere Merkmale oder Anforderungen aufweisen, die unter der veränderten Kontextanforderung vermehrt zum Tragen kommen. Der Erforschung der zweiten Hypothese ist der folgende Beitrag gewidmet.

Somit lauten die weiterführenden Forschungsfragen für diesen Beitrag:

1) Lassen sich solche Items, die auf Grund der geschilderten Experimentalbedingung eine veränderte, gestiegene Itemschwierigkeit im Sinne einer geringeren Lösungswahrscheinlichkeit aufweisen, durch item- oder textspezifische Merkmale beschreiben und systematisieren?
2) Lassen sich differenzielle Effekte bei Gruppen mit unterschiedlichen Personenmerkmalen beobachten (z.B. Studierende mit vermutet gut ausgebildeten Lesestrategien oder Personen mit hohen versus niedrigeren Lesefähigkeiten)?

Zu den besonderen Merkmalen von Aufgaben, von denen angenommen wird, dass sie Schwierigkeit bewirken, gehört auf der Ebene der Item-Merkmale wie beschrieben das Aufgabenformat. Gegenüber der einfachen Multiple-Choice-Aufgabe wird mit dem Falsifizierungs- und Verifizierungs-Prozessmodell von Embretson und Wetzel (1987) angenommen, dass das Format der Entscheidungstabelle sowie der Zuordnungsaufgabe das Potenzial haben, schwierigkeitsgenerierende Prädiktoren zu sein. Wenn durch Navigationsrestriktion das Zurückblättern in den Text unterbunden wird, können in dieser Experimentalbedingung allfällig vorhandene Lücken im gebildeten Situationsmodell nicht mehr nachträglich geschlossen und Fehlinterpretationen nicht mehr korrigiert werden (Schaffner et al., 2004, 197–198; vgl. Kintsch, 1994). Dies macht die Anforderung aller drei Aufgabenformate unter der beschriebenen Experimentalbedingung insgesamt schwieriger. Im Unterschied zur Entscheidungstabelle, bei welcher mehrere Aussagen zu je einer bestimmten Textstelle (lokale Kohärenz) beantwortet werden können, bzw. zur Multiple-Choice, bei der nur eine Lösung entschieden werden muss, können sich bei einer Zuordnungsaufgabe alle Überschriftsoptionen auf alle Abschnitte des Textes beziehen, was einen zusätzlichen mehrfachen Falsifizierungs- und Verifizierungsprozess bedingt. Dieser ist nicht nur aufwändiger, sondern auch fehleranfälliger, gerade dann, wenn nicht zum fortlaufenden Abgleichen auf die Textpassagen zurückgegriffen werden kann.

Auch von den Textmerkmalen kann vermutet werden, dass sie unter der veränderten Kontextbedingung gegebenenfalls eine stärkere Rolle spielen bzw. unterschiedlichen Einfluss

nehmen: Während bei Sachtexten über die Themen und bei literarischen Text über Haupt- und Nebenfiguren und ihre Handlungsstränge voraussichtlich bereits beim aufmerksamen Erstlesen eine einprägsame Textrepräsentation gebildet werden kann, werden die mitunter verschlungenen Argumentationsstränge eines kommentierendes Textes schwerer in die Makrostrukturbildung einfließen können. Ohne erneutes Abgleichen mit dem Text werden somit vermutlich große Lücken bei darauf bezogenen Fragestellungen vorhanden sein. Auch für längere Texte wird angenommen, dass sie ohne wiederholte Texteinsicht Schwierigkeit bewirken, da die grundsätzliche Herausforderung der mentalen Repräsentation über die Länge und Anzahl von Passagen, Themen, Figuren, Argumente, Handlungsstränge hinweg steigt.

Für die unterschiedlichen kognitiven Anforderungen der Items wird hypothetisch angenommen, dass sie sich hinsichtlich ihrer Schwierigkeit unter variierenden Kontextbedingungen unterschiedlich verändern: Die kognitive Anforderung von „Detailinformationen entnehmen“ kann vermutlich ohne möglichen Abgleich mit dem Text eine größere Herausforderung für die Lesenden darstellen, da gestellte Fragen sich zufällig auf eine vielleicht individuelle vorhandene Lücke innerhalb der Textrepräsentation beziehen können und die somit ohne wiederholte Texteinsicht nicht mehr richtig beantwortet werden. Von der kognitiven Anforderung des „Reflektierens und Bewertens“ wird demgegenüber erwartet, dass sie unter der Bedingung des Nichtzurückblätterns eher stabil bleibt: Ein durch das (Erst-)Lesen des Textes generiertes Situationsmodell wird für die Beantwortung einer Reflektieren- und Bewerten-Frage herbeigezogen – man lehnt sich vermutlich eher mal zurück beim Nachdenken über einen Text, als dass man ihn nochmals und nochmals liest.

Somit lauten die konkreten Hypothesen für die erste Forschungsfrage auf der Item- und Textebene:

1.1 Das Aufgabenformat erweist sich unter der Bedingung ohne wiederholte Texteinsicht als wichtiger Schwierigkeitsprädiktor auf Itemebene. Insbesondere die Zuordnungsaufgabe könnte sich unter der Experimentalbedingung als komplexes Format erweisen, das vermehrte Schwierigkeit generiert.

1.2 Auf der Textebene wird vermutet, dass insbesondere die Textsorte der argumentativen kommentierenden Texte aufgrund ihrer komplexeren Argumentationsstruktur vermehrte Schwierigkeit unter der Experimentalbedingung generiert.

1.3 Bei den kognitiven Anforderungen in der Kategorie von Interaktion zwischen Items und Text sollten sich insbesondere die Typ 1-Fragen, welche Informationsentnahme erfordern, als schwierigkeitssteigernd erweisen.

Hinsichtlich der zweiten Forschungsfrage der differenziellen Effekte über verschiedene Personengruppen wird vermutet, dass durch die veränderte Kontextbedingung des Nichtzurückblätterns die Leseleistung bestimmter Personen oder Personengruppen abnimmt, dies wird beispielsweise erwartet für Personen, welche zuerst nur oberflächlich lesen und erst in einer möglichen zweiten Runde vertieft lesen, oder bei Personen, welche ein häufiges Zurückblättern in den Text oder Lesestrategien wie Textmarkieren gewohnt sind.

Als Hypothesen zu vermuteten Zwischengruppeneffekten gelten somit für die getrennten NEPS-Teilstichproben:

2.1 Von Studierenden als sehr fähiger Personengruppe mit eingeübten Lese- und aktualisierten Lernstrategien sowie vermutlich meist hoher Arbeitsgedächtniskapazität wird angenommen, dass sie weniger Mühe haben, mit dem anderen Bearbeitungsmodus des Nichtzurückblätterns umzugehen. Es wird angenommen, dass sie einen schnellen Strategiewechsel vornehmen können und z.B. über verlängertes und vertieftes Erstlesen (Kopp et al., 2016) die einschränkende Bedingung kompensieren können.

2.2 Innerhalb der heterogeneren Teil-Stichprobe der Erwachsenen zeigen sich spezifische Einflüsse von Item-Merkmalen unter veränderten Kontextbedingungen vermutlich weniger deutlich, da eine Vielzahl von unbekannten Personenmerkmalen die Ergebnisse beeinflussen können und nur ein Teil der Stichprobe, nämlich die Erwachsenen mit sehr hoher Lesekompetenz, auf die veränderten Kontextbedingungen mit einem angemessenen Strategiewechsel reagiert. Personen des vierten Quartils werden demzufolge getrennt betrachtet.

Verschiedene ältere Studien haben gezeigt, dass unterschiedlich fähige Personen sich in der Nutzung einer möglichen Texteinsicht in der Hinsicht unterscheiden, dass fähige Lesende das erneute Lesen im Text zur Verifizierung der Antwortoption vermehrt und erfolgreicher anwenden als Lesende mit einer geringer ausgeprägten Lesekompetenz (vgl. Davey, 1987; Garner & Reis, 1981).

Für Personengruppen mit hoher versus geringerer Lesekompetenz (Quartilspilt) werden auf dem Hintergrund der erläuterten Theorie (Kapitel 2.1) folgende Hypothesen formuliert:

2.3 Gute Lesende werden grundsätzlich von der Experimentalbedingung des Nichtzurückblätterns stärker eingeschränkt, da sie beim Bearbeiten von Leseunits vermehrt Strategien verwenden, die ein vertieftes und gründliches Lesen benötigt, wie beispielsweise das Abgleichen von Textstellen mit den verschiedenen Antwortoptionen, das Überprüfen der Lösung am Text, das Wiederlesen wichtiger Stellen und andere Strategien, welche auf einer wiederholten Texteinsicht basieren. Deshalb wird grundsätzlich davon ausgegangen, dass sich unter der Experimentalbedingung des Nichtzurückblätterns die Itemschwierigkeiten für diese Personengruppe erhöhen, da diese ihr übliches strategisches Abgleichverhalten mit Text nicht anwenden können. Ein Teil der Schwierigkeitszunahme kann von geübten Lesenden vermutlich ausgeglichen werden durch effektiven Strategiewechsel (z.B. durch verlängertes Erstlesen, siehe Kopp et al., 2016), mit welchem gute Lesende die Qualität („in die Tiefe lesen“) ihres Textlesens in der Experimentalbedingung erhöhen, um den erhöhten Anforderungen gerecht zu werden und sich die Informationen besser merken zu können. Damit sollte aber nicht die gesamte Schwierigkeitszunahme kompensiert werden können.

2.4 Schlechte Lesende sind von der Experimentalbedingung des Nichtzurückblätterns vermutlich kaum eingeschränkt. Wie dargestellt, verwenden schlechte Leserinnen zur Bewältigung von Textverstehensaufgaben auch mit der Möglichkeit des Zurückblätterns grundsätzlich weniger bis kaum effiziente oder gründliche Strategien des wiederholten Abgleichens der Antwortoptionen mit dem Text. Dadurch ändert sich auch unter experimenteller Einschränkung des Zurückblätterns ihr Bearbeitungsverhalten nicht. Daraus

kann gefolgert werden, dass sich bei dieser Personengruppe die Schwierigkeiten der Items nicht auffällig verändern sollten.

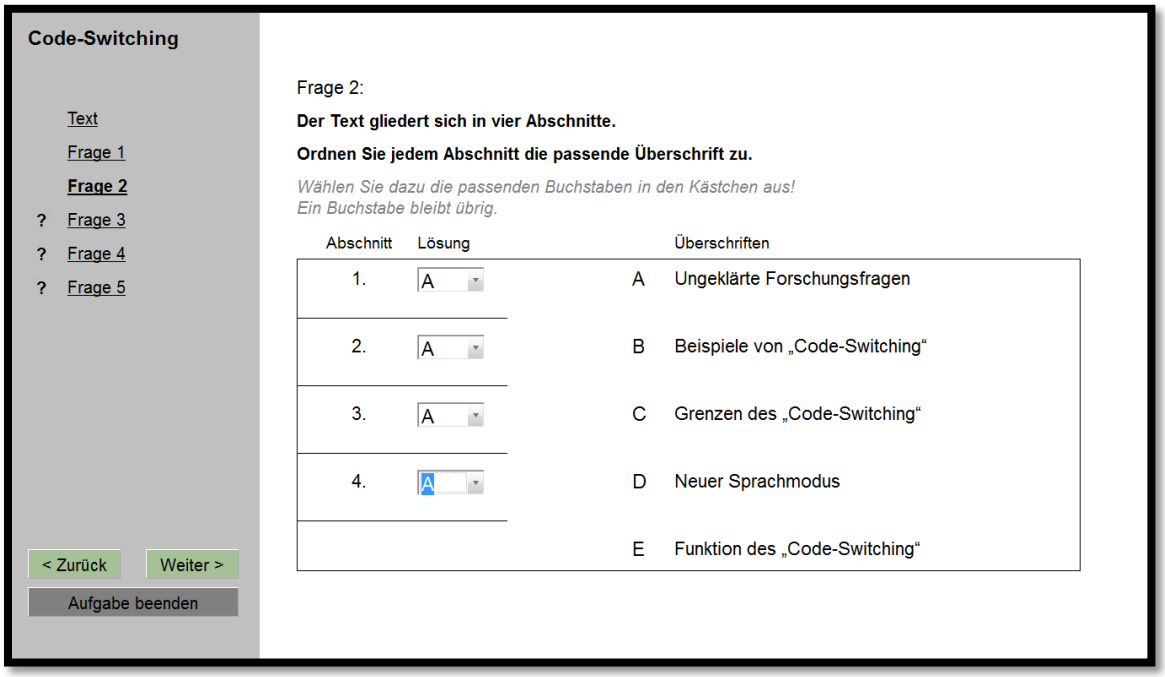
4. Beschreibung der Studie

Die für die vorliegenden Forschungsfragen genutzte Studie war eine im Herbst 2014 in vier ausgewählten Bundesländern durchgeführte Entwicklungsstudie zur Überprüfung neu konstruierter Testaufgaben für den NEPS-Lesekompetenztest für Erwachsene. Die Erhebung erfolgte durch ein externes Erhebungsinstitut und deren geschulte Interviewerinnen und Interviewer in den privaten Haushalten der Zielpersonen oder sonstigen ruhigen nicht-öffentlichen Räumen. Ein Teil der Stichprobe konnte für eine Experimentalbedingung genutzt werden (vgl. Gehrer et al., in Vorbereitung).

4.1 Instrument

In der Studie 2014 wurden mehrere neu entwickelte Testeinheiten für den späteren NEPS-Lesekompetenztest für Studierende und Erwachsene 2016 eingesetzt. Jeder NEPS-Lesekompetenztest beruht auf der NEPS-Rahmenkonzeption zur Messung der Lesekompetenz (Gehrer et al., 2013) und umfasst fünf verschiedene Textsorten unterschiedlicher Länge (vgl. Gehrer & Artelt, 2013), Komplexität und altersangemessener Themenbereiche mit dazugehörenden Testfragen. Zusammen mit den dazugehörenden Testfragen bildet ein Stimulustext (siehe Abbildung 1) einer bestimmten Textsorte eine sogenannte Leseeinheit (Unit). Während für die MC- und Tabellenaufgaben die Distraktoren und Attraktoren mit gewissen Stellen aus dem Text abgeglichen werden müssen, um sich für eine richtige Antwort zu entscheiden (pro Zeile bzw. pro Item), müssen umgekehrt bei einer Zuordnungsaufgabe alle nummerierten Abschnitte des Textes mit den Zuordnungsoptionen abgeglichen werden, um diese in der richtigen Abfolge des Textes zu ordnen bzw. unpassende Überschriften zu eliminieren (ein Beispiel, das später nicht für den Einsatz in der Haupterhebung ausgewählt wurde, siehe Abbildung 2).

Abbildung 1: Beispiel eines später nicht weiterverwendeten Textes aus dem Entwicklungspool 2014



der computerisierten NEPS-Leseitems für Erwachsene und Studierende

Abbildung 2: Beispiel des Antwortformates “Zuordnungsaufgabe“ aus dem Entwicklungspool 2014 der computerisierten NEPS-Leseitems für Erwachsene und Studierende

Auf die unterschiedlichen Anforderungen der Entscheidungstabellen (nicht weiterverwendetes Beispiel siehe Abbildung 3) im Unterschied zu den Multiple-Choice-Aufgaben wurde bereits in Abschnitt 3.1.1 eingegangen.

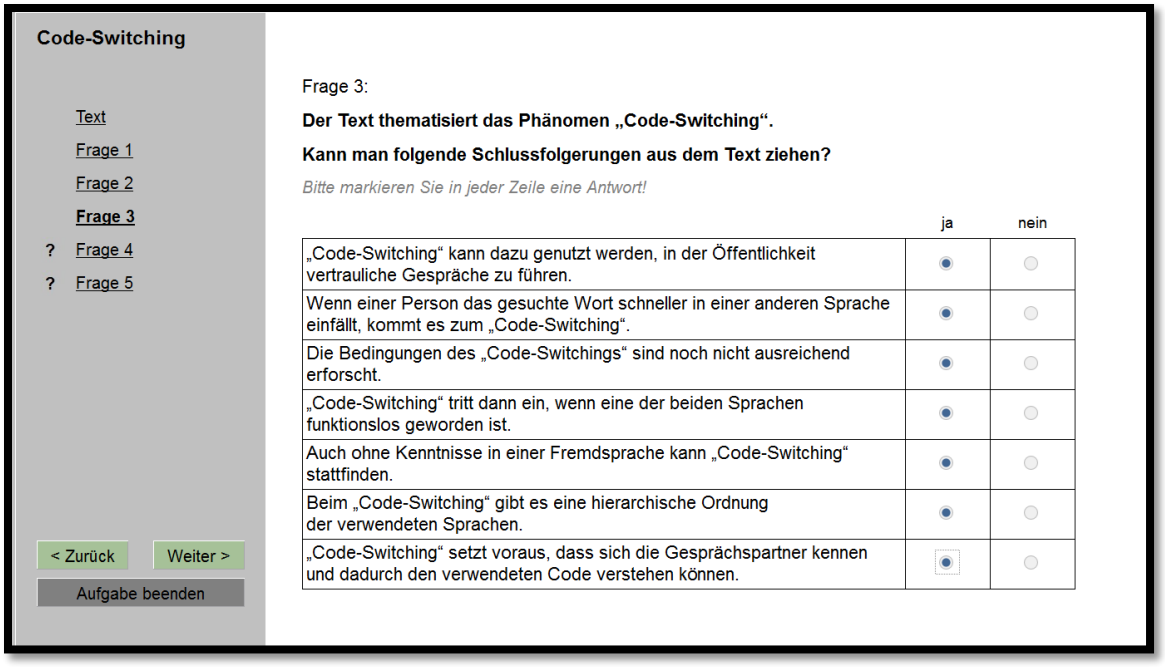


Abbildung 3: Beispiel des Antwortformates “Entscheidungstabelle“ aus dem Entwicklungspool 2014 der computerisierten NEPS-Leseitems für Erwachsene und Studierende

4.2 Design der Studie mit Experimentalbedingung

Abweichend von der sowohl bei papierbasierter Testung (PP) als auch bei computerbasiertem Assessment (CBA) in NEPS-Hauptstudien üblichen Testbedingung von Aufgabenbearbeitung mit jederzeit möglichem Zurückblättern zum anfänglich gelesenen Text bzw. Stimulus (siehe Vor- bzw. Zurückpfeile Abbildungen 2, 3), wurden in dieser Studie zur Generierung von erhöhter Schwierigkeit bei geschlossenen Aufgabenformaten einem Teil der Stichprobe (n = 450) einige Texte vorgelegt, die nur einmal gelesen werden konnten (Experimentalbedingung). Bei der Bearbeitung der darauf folgenden fünf bis acht Aufgaben und Fragen zum eben gelesenen Text konnten diese Zielpersonen nicht nochmals zum Text zurückblättern, d.h. sie erhielten keine weitere Texteinsicht (Gehrer et al., in Vorbereitung). Dies wurde im computerbasierten Assessment technisch dadurch realisiert, dass eine Navigationsrestriktion eingebaut wurde, durch welche das sonst übliche Zurückgehen über das Anklicken eines Zurückpfeils unterbunden bzw. dieser nicht angezeigt wurde. Insgesamt wurden den Zielpersonen dieser Studie 18 Texte unterschiedlicher Textsorten (Gehrer & Artelt, 2013) mit insgesamt 227 dazugehörenden Leseverständnisitems vorgelegt. Die Bedingung des Nichtzurückblätterns durch Navigationsrestriktion wurde in der Experimentalbedingung bei sechs ausgewählten Texten mit insgesamt 72 Items eingesetzt. Von den sechs Experimentaleinheiten waren zwei Texte mit 20 dazugehörenden Items literarisch, zwei Texte mit 19 Items kommentierend sowie zwei Sachtexte mit 33 Items. Die der Rahmenkonzeption entsprechenden Textsorten Anleitung und Werbung waren ausgewogen im Gesamtpool der Entwicklungsstudie enthalten, wurden aber aus inhaltlichen Gründen von der Experimentalbedingung ausgeschlossen. Die Zielpersonen der Experimentalbedingung wurden bei der Einweisung in das Testverfahren über die Navigationsrestriktion aufgeklärt. Zusätzlich wurde über ein Bildschirmfenster unmittelbar

vor dem betreffenden Text als auch über ein Dialogfenster bei Verlassen des Textes die Einschränkung des Blätterns nochmals deutlich kommuniziert[6].

Jede Zielperson bearbeitete in drei Blöcken à 28 Minuten Testzeit je sechs Texte mit dazugehörigen Aufgaben; davon wurden bei der Experimentalbedingung „ohne Texteinsicht" die ausgewählten sechs Texte mit der beschriebenen Navigationsrestriktion vorgegeben. Die sechs navigationsrestringierten Texte wurden in einem Multimatrixdesign über alle drei Blöcke hinweg sowohl an vorderster als auch mittlerer als auch hinterer Stelle platziert, unter anderem um Positionseffekte auszugleichen. Die Kontrollgruppe (*n* = 446) erhielt alle 18 Texte mit jederzeit möglicher Texteinsicht. Auch hier waren die Units im selben Multimatrixdesign über die Positionen hinweg, sowie die Blöcke vorwärts und rückwärts rotiert. Insgesamt wurden 12 Rotationen bzw. Testheftvarianten eingesetzt; die Zuweisung zu den Testpersonen erfolgte zufällig.

Bei der Entwicklung des Lesekomptenztest-Experimentes mit Navigationsrestriktion (ohne wiederholte Texteinsicht) wurde berücksichtigt, dass die unter der Experimentalbedingung eingesetzten Testaufgaben der sechs ausgewählten Texteinheiten stärker als unter normalen Bedingungen die zentralen Aspekte des Textes fokussieren und zur Lösung der Aufgaben keine nebensächlichen Detailinformationen abgerufen werden mussten. Auf diese Weise sollte ein allzu großer Gedächtniseffekt, der insgesamt vermutlich nicht restlos ausgeschlossen werden kann, eingeschränkt werden (vgl. Gehrer et al., in Vorbereitung).

4.3 Stichprobe

Die Gesamtstichprobe der Entwicklungsstudie (*N* = 896) umfasste zwei Teilstichproben, bei denen die Gruppe der Studierenden (*n* = 372) über fünf Universitäten (*n* = 283) und fünf Fachhochschulen (*n* = 89) rekrutiert wurden. Der Altersmittelwert der Studierenden lag bei 25,12 Jahren (*SD* = 3,40), der Anteil Frauen bei 63%. Die zweite Teilstichprobe umfasste 524 Erwachsene (weiblich 59,1%) mit einem Altersmittelwert von 39,35 Jahren (*SD* = 15,51). Die zufällige Rekrutierung der Teilnehmerinnen und Teilnehmer erfolgte über das beauftragte Erhebungsinstitut; es wurde eine Stratifizierung nach Altersgruppen (20-25 Jahre /26-45 Jahre /46-70 Jahre) und nach Bildungsabschlüssen (niedrig/mittel/hoch) vorgegeben (vgl. Gehrer et al., in Vorbereitung).

5. Methode

Als unabhängige Variablen (UV) wurden theoriebasiert wie beschrieben die Merkmale gemäß der Rahmenkonzeption der NEPS-Lesekompetenztests (Gehrer et al., 2012, 2013) betrachtet: die Textsorten als komplexitäts-bestimmende Merkmale auf der Textebene, die Aufgabenformate als formale Merkmale auf der Item-Ebene sowie die kognitiven Anforderungen der Aufgabenstellungen (die sogenannten drei „Typen")[7] auf der Ebene der

[6] Vor jeder der betreffenden restringierten Units wurde ein Deckblatt mit folgendem schriftlichen Hinweis eingeblendet: „Nachdem Sie den folgenden Text gelesen haben und auf „Weiter" geklickt haben, können Sie bei der Beantwortung der [bspw.] 9 Fragen nicht noch einmal zum Text zurückkehren." Beim Übergang zu den Aufgaben erschien ein Dialogfenster, bei dem die Testperson gefragt wurde, ob sie sicher ist, dass sie den Text nun verlassen will und erneut darauf hingewiesen wurde, dass sie danach nicht wieder zum Text zurückkehren kann. Die Zielperson konnte über das Anklicken von „Nein, zurück zum Text" die Erstlesezeit des Textes verlängern bzw. über „Ja, weiter" zu den Fragen gelangen.

[7] Die acht kognitiven Subtypen (Gehrer et al., 2013, 62-63) wurden als UV aufgrund ihrer hierarchischen Abhängigkeit zu den kognitiven Typen aus der Analyse ausgeschlossen. Analysen unter Einbezug der Subtypen mit Ausschluss der hierarchiehöheren Typen-Variablen ergaben keine anderen Resultate.

Item-Text-Interaktion. Als zusätzliche erklärende Variable auf Textebene wurde aufgrund von Hinweisen aus der Literatur (OECD, 2009, S. 45; 2013, S. 69) die Textlänge hinzugenommen. Abhängige Variable (AV) ist die Schwierigkeitsveränderung zwischen den Bedingungen mit und ohne wiederholte Texteinsicht auf der Basis der Lösungswahrscheinlichkeiten (Anteil richtige Lösungen aller validen Antworten) der eingesetzten Items[8]. Die Kodierungen der Items wurden vom Entwicklerteam vor Kenntnis der empirischen Itemschwierigkeiten vorgenommen. Die interne Beurteilerübereinstimmung lag zwischen .982 für die kognitiven Anforderungen (Typen[9]) und 1.00 für Aufgabenformat.

Als Gruppenvariablen zur Überprüfung von differenziellen Effekten wurde einerseits eine Teilstichprobenvariable aus dem Methodendatensatz übernommen, um die Gruppe der Studierenden von der Gruppe der Erwachsenen getrennt untersuchen zu können. Andererseits wurde eine weitere Variable basierend auf den erzielten Kompetenzwerten bzw. Summenscores im eingesetzten Lesekompetenztest gebildet, auf deren Grundlage ein Quartilsplit vorgenommen wurde, aufgrund dessen die Gesamtstichprobe in vier ungefähr gleich große Gruppen von wenig fähigen bis sehr fähigen Personen unterteilt werden konnte. Die beiden mittleren Quartilsgruppen wurden zu einer breiten Gruppe (50%) von Lesende mit mittleren Lesefähigkeiten zusammengezogen, da für die weiteren Analysen insbesondere die beiden Extremgruppen der wenig fähigen (schlechten) Lesenden und der sehr fähigen (guten) Lesenden interessierten.

Für die Deskriptiva der Ergebnisse wurde der T-Test für unabhängige Stichproben verwendet. Als methodischer Zugang zu den beschriebenen Forschungsfragen wurden zwei unterschiedliche Verfahren für die Analysen eingesetzt: Aus der algorithmischen, nicht-parametrischen „Modeling Culture“ (Breiman, 2001, S. 199) wurde zunächst das Klassifikationsbaumverfahren für einen ersten Überblick und zur Beschreibung sowie Systematisierung der item- und testspezifischen Merkmale bei Schwierigkeitsveränderung unter der Experimentalbedingung gewählt. Zur Validierung dieser Resultate und Erweiterung der Fragestellung unter Berücksichtigung von Personenmerkmalen (Personen mit hohen vs. niedrigen Lesefähigkeiten) wurden mit der klassisch parametrischen multiplen linearen Regression weitere Analysen ergänzt. Das in der Bildungsforschung eher unübliche Verfahren der Klassifikationsbäume wird nach der Beschreibung der gewählten Variablen in einem Unterkapitel kurz dargestellt.

5.1 Klassifikationsbäume

Mit dem in SPSS implementierten Klassifikationsbaum- bzw. Entscheidungsbaumverfahren[10] wurde ein visualisierendes Analyseverfahren gewählt, welches ursprünglich von Forscherinnen und Forschern des maschinellen Lernens als auch von Statistikerinnen und Statistikern in den 80er-Jahren entwickelt wurde und sich seit den 90er-Jahren in verschiedensten Anwendungsfeldern zunehmend größerer Beliebtheit erfreut. So hat z.B.

[8] Die vorliegenden Analysen werden auf der Ebene aller Unteraufgaben durchgeführt. Der Begriff Item wird also bei Partial-Credit-Items für die einzelnen Unteraufgaben (Zeilen) verwendet.

[9] Für die kognitiven Subtypen ist die Intercoderreliabilität .936.

[10] Die Begrifflichkeit wird in der Literatur uneinheitlich verwendet. Hier wird die Bezeichnung „Klassifikationsbäume“ als Überbegriff für non-parametrische explorative Analyseverfahren in der Baumstruktur gewählt. Dies im Unterschied zu bspw. Tutz (2000), der die Bezeichnung „Klassifikationsbäume“ nur verwendet bei kategorialen abhängigen Variablen (AV) und von „Regressionsbäumen“ spricht bei stetigen AVs (S. 318).

Säuberlich (2000) innerhalb verschiedener Anwendungsbereiche[11] die wichtigsten benutzten Methoden einer Rangreihung unterzogen und dabei dem Entscheidungsbaumverfahren (neben den neuronalen Netzen) eine „dominierende Rolle“ (S. 53), lange vor den Clusteranalysen, attestiert (S. 56). Lefering (1996) fand aufgrund seiner Simulationsstudien, dass die Qualität der Ergebnisse einer sorgfältig durchgeführten Klassifikationsbaumanalyse sich kaum von den Ergebnissen einer logistischen Regression unterscheiden lässt (103–104). Breiman, der 1984 mit Kollegen eine der heute gängigsten Formen eines binären Baumes, den CART–Classification and Regression Tree, entwickelt hatte, sieht Entscheidungsbäume als akkurate Alternative gerade auch für kleinere Datensätze und für komplexe Fragestellungen der Statistik an (2001; ebenso Tutz, 2012, S. 317; vgl. Parzen, 2001). Da Klassifikationsbäume aufgrund ihrer einfachen Handhabbarkeit, guten Visualisierbarkeit und guten Vorhersagegenauigkeit heutzutage in vielen Forschungsbereichen Anwendung finden, bemühen sich Forscherinnen und Forscher fortlaufend um weitere Verbesserungen einzelner Entscheidungsbaumverfahren, u.a. über verbesserte Split-Kriterien oder über Gewichtung der Modellunsicherheit (z.B. Potapov, 2012; Strobl, 2008).

Die Konstruktion eines Entscheidungs- oder Klassifikationsbaumes erfolgt aufgrund einer automatisierten Teilung der Stichprobe anhand eines der gewählten Merkmale und führt zur Bildung möglichst homogener Untergruppen. Durch die sukzessive Partitionierung, bei der jede erfolgte Zerlegung auf der vorhergehenden aufbaut, wird schnell eine nicht mehr weiteraufteilbare Untergruppe, der Endknoten (nodes), erreicht (z.B. Bühl & Zöfel, 2002, 13–84; Myers & Fucks, 2005; Tutz, 2000, 317–335). Die Klassifikationsbäume unterscheiden sich hinsichtlich ihrer Korrektklassifikationsrate und der Modellunsicherheit (Potapov, 2012, S. 57). Ein weiteres Kriterium ihrer Prognosegüte ist die Baumgröße: Bei jedem Verzweigungsschritt wird die Unreinheit für die Stichprobe kleiner, d.h. die Partionierung besser, doch die Endknoten beruhen nur noch auf wenigen Fällen, so dass für die zugrundeliegende Population der Informationsgehalt nicht mehr allzu groß sein dürfte. Deshalb werden kleinere Bäume mit größerer Prognosekraft bevorzugt. Komplexe Bäume werden zu diesem Zwecke mittels verschiedenen Techniken der „Beschneidung“ um unnötige Äste erleichtert oder aber es werden von vornherein feste Stopp-Regeln definiert (bspw. Anzahl Beobachtungen pro Knoten), welche den Baum nicht zu groß anwachsen lassen (Tutz[12], 2000, 330–332).

Für die folgenden Klassifikationsanalysen wurde das übliche Verfahren CHAID (Chi-Squared Automatic Interaction Detector-Algorithmus) verwendet, welche für eine kategoriale abhängige Variable (AV) eine schrittweise Entdeckung von Zusammenhängen auf der Basis von Chi-Quadrat-Tests erlaubt (Kass, 1980). Die in SPSS implementierte Aufbaumethode Exhaustive-CHAID (Bühl, 2016, 743–755) erstellt inzwischen für beliebige Zielvariablen optimalere Trennungen und erweist sich als präziser (Bühl & Zöfel, 2002, S. 83). Bei einer metrischen Erfassung der AV (Differenzmaß der Lösungshäufigkeiten zwischen den beiden Bedingungen) wurde somit das Verfahren Exhaustive-CHAID verwendet. Die unabhängigen Variablen (UV) konnten nominal in den Klassifikationsbaumverfahren eingehen, die Textlänge wurde ordinal erfasst (0 = kurz, 1 = mittel, 2 = lang).

[11] Säuberlich (2000) untersuchte insgesamt 110 Berichte ab 1994 aus zehn Bereichen von Astronomie, Chemie, Medizin und Gesundheitswesen, Ökonomie, Informatik bis Text-Language-Analysen (50-52).

[12] Beispiele von Klassifikationsbäumen u.a. mit Daten des sozioökonomischen Panels zeigt Tutz (2000) in seinem Statistik-Lehrbuch (S. 5, 324, 331–334, 413–415).

5.2 Multiple lineare Regression

Anlehnend an Hartigs Vorgehen (2007) zur Einschätzung der Einflüsse der einzelnen Aufgabenmerkmale auf die Schwierigkeit wird im nächsten Schritt eine klassisch parametrische multiple lineare Regression (Methode Einschluss; z.B. Fahrmeir, Kneib & Lang, 2009) gerechnet.

Für die Vorhersage von Aufgabenschwierigkeit bei Leistungstests diskutiert Hartig (2007) ein einfaches linear-additives Modell positiv und als ausreichend (S. 96). Er modelliert für die DESI-Tests mittels einer linearen Regressionsanalyse „die Aufgabenschwierigkeit als eine gewichtete Summe ihrer einzelnen [kodierten] Merkmale“, „die Regressionsgewichte Beta η_m drücken hierbei den Einfluss eines Aufgabenmerkmales auf die Aufgabenschwierigkeit aus“ (90–91; siehe dort auch die Regressionsgleichung). Als Kriterium für die Aufnahme zuvor theoretisch formulierter und in mehreren Bereichen kodierter Aufgabenmerkmale wurden die absolute Größe der Regressionsgewichte sowie inhaltlich sinnvolle, also positive Vorzeichen bestimmt; auf die Beschränkung der Aufnahme nur signifikanter Regressionsgewichte wurde bewusst verzichtet, da nicht die Generalisierbarkeit der Modelle im Fokus stand, sondern die Passung der Modelle auf der Ebene der einzelnen DESI-Aufgaben (Hartig, 2007, 94–95).

Nach der vertieften Analyse der ersten Forschungsfrage wird mit der Regression insbesondere die zweite Fragestellung geprüft, ob es für unterschiedliche Personengruppen (Studierende versus Erwachsene, gute versus schlechte Lesende) differenzielle Effekte gibt. Die abhängige Variable wird metrisch gehalten, indem die Schwierigkeitsveränderung zwischen beiden Bedingungen als Differenzmaß beider Lösungshäufigkeiten angegeben wird (Experimentalwerte minus Kontrollwerte). Die bisher nominalskalierten UVs werden gemäß Bühl (2016, S. 449) als auch Hartig (2007, S. 90) in dichotome Dummy-Variablen transformiert.

6. Ergebnisse

6.1 Deskriptiva

Mittels des T-Tests bei unabhängigen Stichproben zeigte sich, dass von den 72 Items, welche der beschriebenen Experimentalbedingung unterlagen, nur 16,7 % Prozent (12 Items) eine signifikante Schwierigkeitsveränderung aufgrund der experimentellen Kontextveränderung aufwiesen. Davon veränderten sich neun Items in die vermutete Richtung der geringeren Lösungswahrscheinlichkeit (d.h. sie wurden schwieriger), bei drei Items lag eine signifikant höhere Lösungswahrscheinlichkeit vor (d.h. sie wurden leichter). Auch bei einem liberaleren Konfidenzintervall von 90% statt 95% veränderte sich die Zahl der signifikant schwierigkeitsverändernden Items nicht. Bei zwei der sechs eingesetzten Texte ergaben sich keine signifikanten Veränderungen der zugehörigen Testfragen, davon war einer literarisch, der andere ein kommentierender Text.

Für getrennte T-Tests und getrennte Berechnung der Schwierigkeitsveränderung innerhalb der fähigkeitsmäßig heterogenen Personengruppe der Erwachsenen ($n = 524$) einerseits und andererseits der Gruppe der Studierenden ($n = 372$), welche eine besondere fähige, leistungsmäßig eher homogenere Personengruppe darstellt, ergibt sich im Vergleich, dass von diesen neun schwierigeren Items nur vier Items gleichermaßen in beiden Gruppen

signifikant schwieriger waren (Tabelle 1, Hervorhebungen). Die anderen fünf Items waren entweder nur in der einen oder der anderen Gruppe signifikant schwieriger. Die beiden Items, für die sich die größten Schwierigkeitsveränderungen unter der veränderten Kontextbedingung zeigen, waren sowohl in diesen Subgruppen als auch in der Gesamtpopulation hoch signifikant (Tabelle 1).

Tabelle 1: Differenz der Lösungswahrscheinlichkeiten unter veränderter Kontextbedingung Teilstichproben Studierende und Erwachsene – signifikant schwerer werdende Items

	Gesamt	**Erwachsene**	**Studierende**	**Merkmale**			
	(*N* = 896)	**(*n* = 523)**	**(*n* = 371)**	**Textsorte**	**Länge**	**Format**	**Typ**
Item 1	**0.169*** **	**0.168*** **	**0.164** **	K	m	MC	2
Item 2	0.107**	0.050 (n.s.)	0. 069 (n.s)	S	m	Tb	1
Item 3	0.080**	0.083*	0.066 (n.s.)	L	m	Tb	1
Item 4	0.076**	0.050 (n.s.)	0.074 *	L	m	Tb	1
Item 5	**0.089** **	**0.107** **	**0.126** **	S	lg	Tb	1
Item 6	0.133***	0.082 (n.s.)	0.082 (n.s.)	S	lg	Tb	1
Item 7	0.142***	0.075 (n.s.)	0.129 *	S	lg	ZO	3
Item 8	**0.210*** **	**0.171*** **	**0.143** **	S	lg	ZO	3
Item 9	**0.234*** **	**0.164*** **	**0.228*** **	S	lg	ZO	3
Gesamt	9	5	6				

Anmerkungen. n.s. = nicht signifikant; *signifikant = $p \leq .05$; ** = $p \leq .01$; *** = $p \leq .001$; K = kommentierend, L = literarisch, S = Sachtext; MC = Multiple-Choice-Format, Tb =Tabellenzeile, ZO = Format Zuordnungsaufgabe; m = Text mittlerer Länge, lg = langer Text.

Für die beiden Gruppen der Personen mit niedrigen Lesekompetenzwerten (Summenscores) bzw. mit hohen Lesekompetenzwerten (entsprechend des ersten und vierten Quartils) ergeben sich bei getrennten T-Tests und wiederum getrennt berechneten Veränderungen der Lösungswahrscheinlichkeiten innerhalb der Gruppen leicht differenzielle Befunde: Sieben Items waren für beide Gruppen gleichermaßen signifikant schwieriger (Tabelle 2, Hervorhebung), wobei drei davon insbesondere für die guten Lesenden um einiges schwieriger zu beantworten waren (Item 7–9).

Für die guten Lesenden finden sich unter der experimentellen Kontextbedingung niedrigere Lösungswahrscheinlichkeiten bei insgesamt 12 Items (davon 10 signifikant) um mehr als 10 Prozent d.h. diese Items waren deutlich schwieriger (Tabelle 2). Zusätzliche 21 Items (davon 13 signifikant) zeigten eine reduzierte Lösungswahrscheinlichkeit um rund 5 Prozent und waren somit etwas schwieriger (ohne Tabelle).

Tabelle 2: Differenz der Lösungswahrscheinlichkeiten unter veränderter Kontextbedingung – schwerer werdende Items bei guten vs. schlechten Lesenden

	Gute Lesende	**Schlechte Lesende**	**Textmerkmale**		**Itemmerkmale**	
	(*n* = 218)	**(*n* = 229)**	**Text**	**Länge**	**Format**	**Typ**
Item 1	**0.19****	**0.16****	K	m	MC	1
Item 2	0.20**	0.04 n.s.	S	m	Tb	1
Item 4	**0.10*****	**0.10****	L	m	Tb	1
Item 5	**0.11****	**0.20*****	S	lg	Tb	1
Item 6	**0.20****	**0.18****	S	lg	Tb	1
Item 7	**0.35*****	**0.16****	S	lg	ZO	3
Item 8	**0.34*****	**0.17****	S	lg	ZO	3
Item 9	**0.23*****	**0.19*****	S	lg	ZO	3
Item 10	0.16*	0.01 n.s.	S	m	MC	2
Item 11	0.13**	0.00 n.s.	K	m	MC	2
Item 12	0.13 n.s.	0.14*	L	k	MC	1
Item 13	0.11 n.s.	0.21**	K	m	MC	3
Item 14	-.03 n.s.	0.20**	S	m	MC	1
Item 15	-.07 n.s.	0.14*	L	k	MC	3
Item 16	-.09 n.s.	0.13*	K	m	MC	2
Item 17	0.04 n.s.	0.13*	S	lg	Tb	1
Gesamt	10*(*)	14*(*)				

Anmerkungen. n.s. = nicht signifikant; *signifikant = $p \leq .05$; ** = $p \leq .01$; *** = $p \leq .001$; K = kommentierend, L = literarisch, S = Sachtext; MC = Multiple-Choice-Format, Tb =Tabellenzeile, ZO = Format Zuordnungsaufgabe; m = Text mittlerer Länge, lg = langer Text, k = kurzer Text.

Insgesamt reduzierten sich in der Gruppe der schlechten Lesenden die Lösungswahrscheinlichkeiten von 17 Items um mehr als 10 Prozent (d.h. die Items wurden deutlich schwieriger), davon erwiesen sich 14 Veränderungen als signifikant (Tabelle 2).

Zusätzlich konnten bei weiteren 23 Items (14 signifikant) mehr als zusätzliche fünf Prozent der schlechten Lesenden sie ohne Texteinsicht nicht lösen, d.h. diese Items wurden somit etwas schwieriger (ohne Tabelle).

Es liegt eine nur leicht ungleiche Verteilung der Personenmerkmale in den getrennten Gruppen vor: Rund 27% der Erwachsenen versus 24 % der Studierenden sind schlechte Lesende, während 23 % der Erwachsenen versus 26 % der Studierenden gute Lesende sind.

6.2 Schwierigkeitsgenerierende Merkmale

6.2.1 Mit dem Klassifikationsbaum

In einem ersten explorativen Schritt werden aus den vier theoretisch angenommenen unabhängigen Variablen Aufgabenformat, kognitive Anforderungen [13], Textsorte und Textlänge mittels des Klassifikationsbaumverfahrens eindeutige Hinweise auf den Einfluss des Aufgabenformates auf die Itemschwierigkeitsveränderung identifiziert. Die abhängige Variable [14] ist wie beschrieben operationalisiert als die Differenz der Lösungswahrscheinlichkeiten zwischen den Bedingungen des Testens mit und ohne wiederholte Texteinsicht. Sowohl in der Gesamtstichprobe als auch bei getrennten Analysen für die Erwachsenen versus die Studierenden werden die Endknoten des Klassifikationsbaumes durch das Aufgabenformat gebildet (siehe Abbildung 2). Keine der anderen unabhängigen Variablen wurde in das Baummodell aufgenommen.

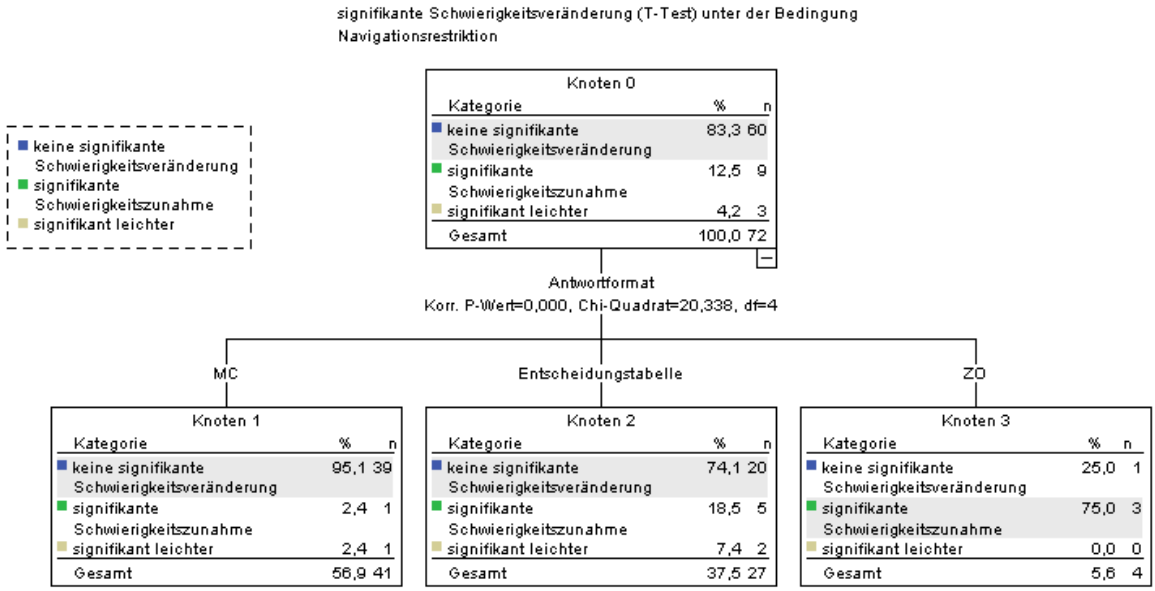


Abbildung 2: Baummodell über alle Personen (N = 896) – Zuweisung der 12 schwierigkeitsveränderten Items über die erklärende Variable „Antwortformat“

[13] Analysen unter Einbezug der Subtypen mit Ausschluss der hierarchiehöheren Variablen für die Typen ergaben keine anderen Resultate.

[14] Die abhängige Variable (AV) wurde zuerst nominal kodiert (0= keine Schwierigkeitsveränderung, 1= Schwierigkeitsveränderung) unter Verwendung der binären Aufbaumethode Quest (Bühl, 2016,770–789). Für CHAID wurde die AV in einem zweiten Schritt ordinal umkodiert (0= leichter, 1= keine Veränderung, 2= Schwierigkeitszunahme). Die Ergebnisse bleiben gleich wie beim letztlich gewählten Klassifikationsbaumverfahren Exhaustive-CHAID und metrischer Kodierung der AV.

Die Multiple-Choice-Aufgaben blieben in 95,1% der Items ohne Schwierigkeitsveränderung (n = 39), beim Format der Entscheidungstabelle wurden 18,5% der Fälle schwieriger (n = 5) und 7,4% der Subaufgaben leichter (n = 2), die Zuordnungsaufgabe erhielt in 75% der Fälle (n = 3) eine signifikante Schwierigkeitszunahme (Abbildung 2).

6.2.2 Mit Regression

Die multiple lineare Regression in der Gesamtstichprobe präzisiert diese ersten Resultate der Klassifikationsbaumanalyse. Mit einer Varianzaufklärung des Modells von 34.7% wird das Aufgabenformat „Zuordnungsaufgabe“ mit einem standardisierten Beta-Koeffizienten (Regressionsgewicht) von .597 als einziger signifikanter Prädiktor für die Veränderung von Itemschwierigkeit unter der Experimentalbedingung ohne wiederholte Texteinsicht ausgewiesen (siehe Tabelle 3).

Tabelle 3: Regression – Veränderung der Aufgabenschwierigkeit (Differenz der Lösungswahrscheinlichkeiten) über alle Personen (N = 896).

Prädiktoren	**β**	**T-Wert**	**Sig. (*p*)**
Konstante		-.579	.565
Kognitive Anforderung Typ 1	.230	1.907	.061
Kognitive Anforderung Typ 3	.070	.538	.592
Textsorte kommentierend	.126	.896	.373
Textsorte literarisch	-.092	-.602	.549
Format Tabelle	.020	.142	.888
Format Zuordnungsaufgabe	.597	4.542	.000
Textlänge kurz	.175	1.425	.159
Textlänge lang	-.099	-.629	.531
R^2	0.347		

Anmerkungen. Referenz: Kognitive Anforderung Typ 2 (Schlussfolgern); Sachtextsorte; Format Multiple Choice; mittlere Textlänge. Positive Regressionsgewichte ergeben in der Regressionsgleichung eine Zunahme von Itemschwierigkeit.

Die bei den verschiedenen Teilstichproben der Erwachsenen (58%) versus Studierenden (42%) als auch für die drei Gruppen schlechte, mittlere und gute Lesende getrennt durchgeführten Regressionen zeigen darüber hinaus ein leicht differenziertes Bild.

So erwies sich für die schlechten Lesenden (*N* = 229) wiederum das Aufgabenformat „Zuordnungsaufgabe“ als ein starker signifikanter Prädiktor für die Veränderung der Aufgabenschwierigkeit in Richtung Schwierigkeitszunahme. Als weiterer Prädiktor zeigte sich in dieser Personengruppe jedoch zusätzlich die kognitive Anforderung der Aufgabenstellung:

Während die kognitive Anforderung des Reflektierens und Bewertens (Typ 3) keinen signifikanten Effekt hatte, konnte die kognitive Anforderung des Informationen-Entnehmens (Typ 1) zusätzlich als signifikante Einflussgröße (β = .291, *p* = .032) für die Schwierigkeitszunahme identifiziert werden (Tabelle 4).

Tabelle 4: Regression – Vorhersage der Veränderung der Aufgabenschwierigkeit (Differenz der Lösungswahrscheinlichkeiten) für schlechte Lesende (N = 229).

Prädiktoren	**β**	**T-Wert**	**Sig. (*p*)**
Konstante		-.855	.396
Kognitive Anforderung Typ 1	.291	2.192	.032
Kognitive Anforderung Typ 3	-.096	-.671	.505
Textsorte kommentierend	.257	1.654	.103
Textsorte literarisch	.047	.278	.782
Format Tabelle	-.132	-.859	.394
Format Zuordnungsaufgabe	.390	2.696	.009
Textlänge kurz	.129	.955	.343
Textlänge lang	.139	.806	.423
R^2	.21		

Anmerkungen. Referenz: Kognitive Anforderung Typ 2 (Schlussfolgern); Sachtextsorte; Format Multiple Choice; mittlere Textlänge.

Für die aus zwei Quartilen gebildete Gruppe der mittleren Lesenden (*N* = 449) ergibt sich ein ähnliches Bild wie bei den schlechten Lesenden, wobei neben dem signifikanten Aufgabenformat „Zuordnungsaufgabe“ (β = .39, *p* = .009) die kognitive Anforderung Typ 1 des Informationen-Entnehmens (β = .291, *p* = .055) sich hier nur als marginal signifikante Einflussgröße in Richtung Schwierigkeitszunahme erweist.

Für die Gruppe der guten Lesenden (*N* = 218) zeigt sich im Unterschied zu den schlechten und mittleren Lesenden der letztere Befund nicht. Für die guten Lesenden wirkt die kognitive Anforderung, einem Text auch bei nur einmaligem Lesen Informationen zu entnehmen, nicht schwierigkeitsverändernd. Hingegen zeigt sich auch bei dieser Personengruppe mit erhöhten Fähigkeiten wie für alle anderen Personengruppen bei dem spezifischen Aufgabenformat der Zuordnungsaufgabe eine signifikante Schwierigkeitszunahme (β = .569, *p* ≤ .001) unter der Restriktionsbedingung (siehe Tabelle 5).

Tabelle 5: Regression – Veränderung der Aufgabenschwierigkeit (Differenz der Lösungswahrscheinlichkeiten) für gute Lesende (N = 218).

Prädiktoren	β	T-Wert	Sig. (*p*)
Konstante		.014	.989
Kognitive Anforderung Typ 1	-.014	-.115	.909
Kognitive Anforderung Typ 3	.014	.107	.916
Textsorte kommentierend	.056	.395	.694
Textsorte literarisch	-.147	-.958	.342
Format Tabelle	.139	.982	.330
Format Zuordnungsaufgabe	.569	4.302	.000
Textlänge kurz	.136	1.098	.276
Textlänge lang	-.016	-.102	.919
R^2	.338		

Anmerkungen. Referenz: Kognitive Anforderung Typ 2 (Schlussfolgern); Sachtextsorte; Format Multiple Choice; mittlere Textlänge.

7. Diskussion

Im Unterschied zu vielen anderen Studien (vgl. zusammenfassend z.B. Zimmermann, 2016) wurden deutschsprachige Leseverständnisaufgaben im Erwachsenenalter untersucht, welche nicht nur Multiple-Choice-Aufgaben enthalten, sondern in drei verschiedenen geschlossenen Formaten gehalten sind: Multiple-Choice, Entscheidungstabellen (true-false) und Zuordnungsaufgaben (matching). Zusätzlich beziehen sich die Textaufgaben auf unterschiedliche Textsorten verschiedener Länge. Es wurden die der Experimentalstudie des NEPS-Lesekompetenztests für Studierende und Erwachsene (N = 896, Kopp et al., 2016; Gehrer et al., in Vorbereitung) anschließenden Forschungsfragen beantwortet, welche Prädiktoren unter der Bedingung ohne wiederholte Texteinsicht für die Schwierigkeitsveränderungen gewisser Items verantwortlich sind und für welche Gruppen sich differenzielle Effekte ergeben. Von den 16,7 % Prozent der Items (*N* = 72), welche unter Navigationsrestriktion eine veränderte Lösungswahrscheinlichkeit aufwiesen, wurden nur sehr wenige Items (*n* = 4) sowohl in der Gesamtgruppe als auch in den Gruppen der Studierenden versus Erwachsene, bzw. schlechte versus gute Lesende signifikant schwieriger.

Sowohl mit einer Klassifikationsbaumanalyse (z.B. Tutz, 2000) als auch in Vertiefung und Erweiterung mit einer multiplen Regression (z.B. Fahrmeir et al., 2009) konnten Effekte des Aufgabenformates gefunden werden. Insbesondere das Format der Zuordnungsaufgabe, welche aus einer Auswahl von möglichen Überschriften eine Zuweisung eines passenden Zwischentitels zu jedem Abschnitt des gelesenen Textes verlangt (Beispiel siehe

Abschnitt 5.1), erwies sich in der Regression unter der besonderen Bedingung ohne wiederholte Texteinsicht hypothesenkonform als signifikant ($p \leq .001$) schwieriger, dies sowohl in der Gesamtstichprobe als auch in den getrennten Gruppen der schlechten, mittleren und guten Lesenden. In Anlehnung an Rupp, Ferne und Choi (2006) wurde für vorliegende Analyse mit NEPS-Leseitems angenommen, dass deren unterschiedliche Arten von geschlossenen Formaten (Multiple Choice, Entscheidungstabellen, Zuordnungsaufgaben) jeweils spezifische Prozesse und Strategien erfordern. In Anlehnung an das Aufgabenbearbeitungsmodell von Embretson und Wetzel für Multiple-Choice-Aufgaben (1987; auch Davey, 1987; Rost 2004) wird für jede Antwortoption, bei Entscheidungstabellen für jede Zeile (Unteraufgabe) davon ausgegangen, dass sie einzeln gegenüber dem Text falsifiziert oder verifiziert wird. Für das Format „Zuordnungsaufgabe" bedeutet dieser Lösungsprozess, dass jeder optionale Zwischentitel mit jedem Textabschnitt abgeglichen, falsifiziert oder verifiziert werden muss. Da jede Überschrift eine zusammenfassende Kernaussage der jeweiligen Textpassage ausdrückt, ist zusätzlich jedoch eine von vornherein erfolgreiche lokale und globale Kohärenzbildung und mentale Repräsentation über die einzelnen Passagen des Textes notwendig, um den besonderen Anforderungen dieses Formates gerecht werden zu können. Wenn durch Navigationsrestriktion das Zurückblättern in den Text unterbunden wird, können in dieser Experimentalbedingung allfällig vorhandene Lücken im gebildeten Situationsmodell nicht mehr nachträglich geschlossen und Fehlinterpretationen nicht mehr korrigiert werden (Schaffner et al., 2004, 197–198; vgl. Kintsch, 1994). Dies macht die Anforderung dieses Aufgabenformates unter der beschriebenen Experimentalbedingung insgesamt schwieriger. Der Befund der Analyse zu den Aufgabenformaten ist somit theorie- und hypothesenkonform.

Auf der Ebene der Textmerkmale erwiesen sich die Textsorte kommentierend-argumentativer Text sowie die Länge des Textes entgegen der Hypothese nicht als schwierigkeitsgenerierend. Dies kann auch der geringen Zahl an Textexemplaren geschuldet sein, welche in die Experimentalbedingung eingehen konnte (siehe Abschnitt 4.2). Bei der Kategorie der Merkmale Text-Item-Interaktion zeigten sich wie vermutet die kognitiven Anforderungen des Reflektierens und Bewertens sowie des Schlussfolgerns nicht als schwierigkeitssteigernde Prädiktoren. Für die kognitive Anforderung des Informationenentnehmens zeigten sich nur differenzielle Effekte.

Bezüglich differenzieller Effekte wurde für die Gruppe der Personen mit hohen Lesefähigkeiten in Anlehnung an Davey (1987) sowie Garner und Reis (1981) vermutet, dass sie aufgrund ihrer effektiveren Nutzung von Textbearbeitungs- und Aufgabenbearbeitungs-Strategien stärker als die schlechten Lesenden von einer experimentellen Einschränkung der Texteinsicht betroffen sind. Andererseits konnte auch davon ausgegangen werden, dass fähige Lesende ihre Bearbeitungsstrategie auch unter veränderten Kontextbedingungen besser als schlechte Lesende anpassen können (z.B. durch Verlängerung ihrer Erstlesezeit des Stimulustextes, siehe Kopp et al., 2016). Insgesamt wurde dennoch eine größere Schwierigkeitszunahme der Items unter der Bedingung ohne wiederholte Texteinsicht für die guten Lesenden erwartet als bei mittleren oder schlechten Lesenden.

Die Resultate der Regressionen bei getrennten Personengruppen anhand eines Quartilsplits nach Lesefähigkeiten konnte diese Hypothese bestätigen. Es zeigte sich unter der Experimentalbedingung, dass gerade bei guten Lesenden die Lösungswahrscheinlichkeit von

Zuordnungs-Items signifikant und stärker als bei mittleren und schlechten Lesenden abnahm, d.h. die Schwierigkeit von Zuordnungs-Items für gute Lesende größer wurde als für andere Personengruppen. Dies lässt die Schlussfolgerung zu, dass gute Lesende ihre erstgebildeten Texthypothesen zu einzelnen Textpassagen stärker und öfters als schlechte Lesende während eines zyklischen Rezeptionsprozesses bestätigen bzw. revidieren. Ohne wiederholte Texteinsicht können sie ihre gewohnte effektive Strategie des fortwährenden Abgleichens mit dem Text nicht nutzen, wodurch sich die Aufgabenstellung für sie erschwert.

Für die Gruppe der Personen mit geringeren Lesefähigkeiten fanden sich über das Aufgabenformat hinaus zusätzliche differenzielle Effekte für die kognitiven Anforderungen von Aufgabenstellungen, welche teilweise als Erklärung für die eingetretenen Schwierigkeitsveränderungen bestätigt wurden: Aufgaben mit der kognitiven Anforderung des Informationenentnehmens (Typ 1) wurden für schlechte Lesende unter der Navigationsrestriktion signifikant schwieriger. In gewissem Maße gilt dies auch für mittlere Lesende, aber nicht für gute Lesende. Für schlechte Lesende scheinen Aufgaben unter Navigationsrestriktion, d.h. ohne wiederholte Texteinsicht, insbesondere also dann schwieriger zu werden, wenn Fragen zum detaillierten und lokalen Textverständnis beantworten werden müssen.

Insbesondere bei diesen weniger komplexen kognitiven Anforderungen kann für gute Lesende vermuten werden, dass sie dank des Strategiewechsels hin zu einem verlängerten Erstlesen (vgl. Kopp et al, 2016) potenzielle Schwierigkeitssteigerungen der experimentellen Navigationsrestriktion besser kompensieren konnten: Sie verbesserten in der Experimentalbedingung vermutlich die Qualität ihres Textlesens vor der Aufgabenbearbeitung im Sinne eines intensivierten „in die Tiefe Lesens“, um den erhöhten Anforderungen des Nichtblätterns gerecht zu werden und sich die Informationen besser merken zu können.

Als Einschränkungen der Analyse sind folgende zu benennen: Obwohl die NEPS-Entwicklungsstudie eine große Gesamtzahl an Leseaufgaben (N = 227) aufwies, konnte für den Teil der Experimentalbedingung lediglich eine beschränkte Itemzahl (n = 72) eingesetzt werden. Dies hatte den Nachteil, dass für die Vielfalt an einzelnen Prädiktoren nur eine beschränkte Anzahl von Items vorlag. So konnten insgesamt nicht mehr als sechs Texte in die Experimentalbedingung aufgenommen werden, pro verwendeter Textsorte nur je zwei Stück; so standen nicht mehrere kürzere Texte und auch nicht mehrere Zuordnungsaufgaben zu Verfügung. Insgesamt kann somit eine Konfundierung mit dem Einzelexemplar des Textes aufgrund der kleinen Itemzahl sowie der test-immanenten Nestung von Fragen und Text nicht ausgeschlossen werden. Da für die Experimentalbedingung aus Kapazitätsgründen nicht alle fünf Textsorten der NEPS-Rahmenkonzeption eingesetzt werden konnten, bleiben die Aussagen auf die hier verwendeten Textsorten Sachtext, literarischer Text und kommentierender Text beschränkt.

Methodisch führte die Beschränkung der Items und damit geringe Zellenbesetzung dazu, dass in der Regression keine Interaktionsanalysen (bspw. Interaktionsterm kognitive Anforderungen mit Textsorte oder -länge) durchgeführt werden konnten, wodurch die dritte Ebene der schwierigkeitsbestimmenden Merkmale nur über eine qualitative Kodierung gewährleistet werden konnte, in der die spezifische Interaktion der kognitiven Anforderung der Items mit dem jeweiligen Textexemplar erfasst wurde. Im Verfahren des

Klassifikationsbaums hatte diese Beschränkung der Items zwar den methodischen Vorteil, dass keine Stopp-Regeln bezüglich der Anzahl Beobachtungen pro Knoten eingebaut werden mussten, um den Baum in seiner allfälligen Größe artifiziell zu beschneiden (zum Verfahren vgl. Tutz, 2000, 330–332), der Nachteil zeigt sich jedoch darin, dass sich die Anzahl Beobachtungen pro Knoten im untersten Bereich befindet, insbesondere bei den Endknoten (Knoten 3: $n = 4$).

Um Positions- und Reihenfolgeeffekte für die Items auszugleichen und unerwartet großen not-reached-Missings entgegen zu wirken, werden in NEPS-Entwicklungsstudien für große Aufgabenpools sorgfältige Multi-Matrix-Designs eingesetzt. Ob und wie die unterschiedlichen Reihenfolgen und Positionen der Items aber gegebenenfalls auch Lerneffekte innerhalb der Testsituation, besonders bei anspruchsvollen Formaten wie der beschriebenen Zuordnungsaufgabe und auch bei der Bedingung des Nichtzurückblätterns, bewirken, konnte noch nicht abschließend geklärt werden. Dieser nächsten Forschungsfrage sollte in zukünftigen Analysen noch nachgegangen werden.

Methodisch kann bei der Beantwortung der zweiten Forschungsfrage der differenziellen Effekte für verschiedene Personengruppen in Bezug auf den Vergleich der Studierenden versus Erwachsene eine gewisse Ungenauigkeit der Zuweisung über die Methodendatensatzvariable nicht ausgeschlossen werden. Diese Variable wurde über die Erstellung der Teilstichproben definiert (Rekrutierung der Studierenden über Universitäten und Fachschulen versus Auffrischung eines Pilotpanels Erwachsene nach Quotenmerkmalen). In der Teilstichprobe der Erwachsenen befinden sich somit auch einige Studierende, welche nicht über die Hochschulen gewonnen wurden. Es wäre in der Erwachsenenteilstichprobe eine Umkodierung des Personenstatus allenfalls denkbar aufgrund der Angaben zu den Fragen zur berufliche Tätigkeit („Welche berufliche Tätigkeit üben Sie derzeit aus bzw. haben Sie zuletzt ausgeübt?“ [offene Angabe] und „Wenn Sie noch nie eine hauptberufliche Tätigkeit ausgeübt haben, klicken Sie bitte „trifft nicht zu“ an“). Beide Angaben scheinen aber nicht eindeutig einen Studiumsstatus zu verneinen. Da die spezifische Frage „Befinden Sie sich zur Zeit in einem Studium?“ im Erwachsenensample nicht vorhanden ist, wurde auf eine Umkodierung verzichtet.

Eine weitere Limitation der Studie liegt darin, dass Personenmerkmale nur schwierig in die Analysen eingehen konnten. Da hier Regressionen auf Itemebene (Items als „Fälle“) gerechnet werden, konnten Merkmale von Personen nur über die Analyse für getrennte Stichproben nach bestimmten Gruppen, wie hier Studierende versus Erwachsene und gute versus schlechte Lesende, vorgenommen werden. Für die wünschenswerte Kontrolle weiterer Lesermerkmale wie Muttersprache, Bücherbesitz, Lesehäufigkeiten, Computergeübtheit und andere in der Experimentalstudie erfragten Variablen müsste ein anderes Verfahren gewählt werden. Inhaltlich wertvoll wären für diese Studie auch die Erfassung weiterer Personenmerkmale als Kontrollmaße gewesen: Maße zur Erfassung der selbstberichteten Strategienutzung oder zu Problemlösefähigkeiten, der Lesemotivation oder zum thematischen Interesse und dem Vorwissen wie der verbalen Intelligenz wären wünschenswert, um den vorliegenden Aussagen zu differenziellen Effekten für bestimmte Lesergruppen weiter nachgehen zu können.

Insbesondere für die Kontrolle der Effekte der Kapazität des Kurzzeitgedächtnisses und der Leistungsfähigkeit des Arbeitsgedächtnisses sowie speziell der bereichsspezifischen

Textgedächtnisleistung wäre für eine solche Studie mit Experimentalbedingung ohne wiederholte Texteinsicht, bei welcher eine spezifische Rolle von Teilen des expliziten Gedächtnisses vermutet werden kann, die Erhebung eines solchen Zusatzmaßes angebracht gewesen. Für die Münchner Längsschnittstudie LOGIK (Weinert, 1998) konnte nachgewiesen werden, dass mehrere Gedächtnisaspekte, so das bereichsspezifische Gedächtnis für Geschichten und Texte sowie das Arbeitsgedächtnis, als Prädiktoren für individuelle Leistungsunterschiede bereits sehr früh deutlich werden und bis ins frühe Erwachsenenalter relativ stabil bleiben (Knopf, Schneider, Sodian & Kolling, 2008).

Für weiterführende Untersuchungen differenzieller Effekte wäre es wünschenswert, solche zusätzlichen Personenmerkmale erfassen und in weiteren Analysen berücksichtigen zu können. Isaac & Hochweber (2011, S. 193) zeigen beispielsweise für Sprachherkunft und Bücherbesitz, dass ihre Interaktion mit den Aufgabenmerkmalen (bei Jugendlichen) Unterschiede in den schwierigkeitsgenerierenden Effekten bewirken.

Als letzter Ausblick ist anzustreben, dem geschlossenen Aufgabenformat „Zuordnungsaufgabe“ in nächsten Entwicklungsstudien mit höheren Itempools und wenn möglich unter Experimentalbedingung geflissentlich weitere analytische Aufmerksamkeit zu schenken, nicht nur in Bezug zu den bisherigen geschlossenen Formaten der NEPS-Lesekompetenztests und in Bezug zu weiteren Textsorten, sondern auch in Unterscheidung zu weiteren innovativen geschlossenen und halboffenen Leseformaten künftiger Lesekompetenztests des Bildungspanels.

Insgesamt bleiben abschließend die limitierten Aussagen bestehen, dass das Aufgabenformat einen nicht zu unterschätzenden Einfluss auf Itemschwierigkeiten unter veränderten Kontextbedingungen zu haben scheint, und dass sich differenzielle Effekte für Aufgaben insbesondere mit der kognitiven Anforderung des Informationen-Entnehmens zeigen, in dem Sinne, dass diese bei Lesenden im Erwachsenenalter eher für weniger fähige Lesende eine gewisse Herausforderung darstellen, nicht aber für gute Lesende.

Literatur

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). Opladen: Leske + Budrich.

Artelt, C., Stanat, P., Schneider, W., Schiefele, U. & Lehmann, R. H. (2004). Die PISA-Studie zur Lesekompetenz. Überblick und weiterführende Analysen. In U. Schiefele, C. Artelt, W. Scheider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 139–168). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bamberger, R. & Rabin, A. T. (1984). New Approaches to Readability: Austrian Research. *The Reading Teacher* 37 (6), 512–519.

Björnsson, C.-H. (1968). *Lesbarkeit durch Lix*. Pedagogiskt centrum, Stockholms skolförvaltn.

Blatt, I. & Voss, A. (2005). Leseverständnis und Leseprozess. Didaktische Überlegungen zu ausgewählten Befunden der IGLU-/ IGLU-E-Studien. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien* (S. 239–281). Münster: Waxmann.

Blossfeld, H.-P., Roßbach, H.-G. & von Maurice, J. (Hrsg.) (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Sonderheft 14.*

Bos, W., Valtin, R., Voss, A., Hornberg, S. & Lankes, E.-M. (2007). Konzepte der Lesekompetenz in IGLU 2006. In W. Bos, S. Hornberg & K.-H. Arnold (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 81–107). Münster [u.a.]: Waxmann.

Bormuth, J. R. (1967). Comparable Cloze and Multiple-Choice Comprehension Test Scores. *Journal of Reading, 10* (5), 291–299.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16,* 199–231.

Brinker, K. (2010). *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden.* (7. durchgesehene Auflage). Berlin: ESV.

Bühl, A. (2016). *SPSS 23. Einführung in die moderne Datenanalyse* (15., aktualisierte Auflage). Hallbergmoos: Pearson.

Bühl, A. & Zöfel, P. (2002). *Erweiterte Datenanalyse mit SPSS. Statistik und Data Mining.* Wiesbaden: Westdeutscher Verlag.

Christmann, U. & Groeben, N. (2002). Anforderungen und Einflussfaktoren bei Sach- und Informationstexten. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (S. 150–173). Weinheim, München: Juventa.

Davey, B. (1987). Postpassage Questions: Task and Reader Effects on Comprehension and Metacomprehension Processes. *Journal of Reading Behavior, 19* (3), 261–283.

Eggert, H. (2002). Literarische Texte und ihre Anforderungen an die Lesekompetenz. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (S. 186–194). Weinheim, München: Juventa.

Eggs, E. (1996). Formen des Argumentierens in Zeitungskommentaren: Manipulation durch mehrsträngig assoziatives Argumentieren? In E. Hess-Lüttich (Hrsg.), *Textstrukturen im Medienwandel* (S. 179–209). Frankfurt a. M.: Lang.

Embretson, S. E. & Wetzel, C. D. (1987). Component Latent Trait Models for Paragraph Comprehension Tests. *Applied Psychological Measurement, 11* (2), 175–193.

Fahrmeir, L., Kneib, Th. & Lang, S. (2009). *Regression. Modelle, Methoden und Anwendungen.* Berlin: Springer.

Freedle, R. & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing* (10), 133–170.

Freedle, R. & Kostin, I. (1994). Can multiple-choice reading tests be construct-valid? A Reply to Katz, Lautenschlager, Blackburn, and Harris. *Psychological Science, 5* (2), 107–110.

Garner, R. & Reis, R. (1981). Monitoring and Resolving Comprehension Obstacles: An Investigation of Spontaneous Text Lookbacks among Upper-Grade Good and Poor Comprehenders. *Reading Research Quarterly, 16* (4), 569–582.

Gehrer, K. & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In C. Rosebrock & A. Bertschi-Kaufmann (Hrsg.), *Literalität erfassen: bildungspolitisch, kulturell, individuell* (S. 168–187). Weinheim: Beltz Juventa.

Gehrer, K., Wolter, I., Koller, I. & Artelt, C. (in Vorbereitung)[15]. *Lesekompetenztestung mit und ohne Texteinsicht. Gibt es Effekte auf Itemparameter?* Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Gehrer, K., Zimmermann, S., Artelt, C. & Weinert, S. (2013). NEPS Framework for Assessing Reading Competence and Results From an Adult Pilot Study. In C. Artelt, S. Weinert & C. H. Carstensen (Hrsg.), *Competence Assessment within the NEPS.* Journal for Educational Research Online, 50–79.

Gorin, J. S. (2005). Manipulating Processing Difficulty of Reading Comprehension Questions. The Feasibility of Verbal Item Generation. *Journal of Educational Measurement, 42* (4), 351–373.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing Inferences During Narrative Text Comprehension. *Psychological Review, 101* (3), 371–395.

Groeben, Norbert (1978). *Die Verständlichkeit von Unterrichtstexten*. *Dimensionen und Kriterien rezeptiver Lernstadien.* Münster: Aschendorff (2. Aufl.).

[15] Autorengruppe/Reihenfolge noch nicht final

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 83–99). Weinheim: Beltz.

Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau, 63* (1), 43–49.

Ingenkamp, K. (2005). *Lehrbuch der Pädagogischen Diagnostik* (5. überarb. Aufl.). Weinheim, Basel: Beltz.

Isaac, K. & Hochweber, J. (2011). Modellierung von Kompetenzen im Bereich „Sprache und Sprachgebrauch untersuchen" mit schwierigkeitsbestimmenden Aufgabenmerkmalen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 43(4)*, 186–199.

Janich, N. (2010). *Werbesprache: Ein Arbeitsbuch* (5. Aufl.). Tübingen: Narr Francke Attempto Verlag.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistic, 29,* 119–127.

Katz, S., Lautenschlager, G. J., Blackburn, A. B. & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science, 1* (2), 122-127.

Kendall, J. R., Mason, J. M. & Hunter, W. (1980). Which Comprehension? Artifacts in the Measurement of Reading Comprehension. *The Journal of Educational Research, 73* (4), 233–236.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394.

Kintsch, W. & Keenan, J. (1973). Reading Rate and Retention as a Function of the Number of Propositions in the Base Structure of Sentences. *Cognitive Psychology 5*, 257–274.

Kintsch, W. & Yarbrough, J.C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology 74*, 828–34.

Kintsch, W. (1994). Text Comprehension, Memory, and Learning. *American Psychological Association 49 (4),* 294–303.

Kirsch, I. S. (2001). *The International Adult Literacy Survey (IALS). Understanding What Was Measured.* Princeton: Research Publications Office.

Klicpera, C. & Gasteiger-Klicpera, B. (1993). *Lesen und Schreiben. Entwicklung und Schwierigkeiten*. Bern: Huber Verlag.

Knopf, M., Schneider, W., Sodian, B. & Kolling, T. (2008). Die Entwicklung des Gedächtnisses vom Kindergartenalter bis ins frühe Erwachsenenalter - Neue Erkenntnisse aus der LOGIK-Studie. In W. Schneider (Hrsg.), *Entwicklung von der Kindheit bis zum Erwachsenenalter. Befunde der Münchner Längsschnittstudie LOGIK* (S. 85–102). Weinheim, Basel: Beltz PVU.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*, 193–220.

Kopp, F., N.N. (in Vorbereitung*). Context, Competence and Strategy use: Reading behavior under systematically varied contexts as indicator for adaptive use of reading strategies.*

Kopp, F., Gehrer, K., Artelt, C., Wolter, I. & Koller, I. (11.03.2016). *Sind gute Lesende unter widrigen Bedingungen flexible Strategienutzende?* Vortrag an der 4. Jahrestagung der Gesellschaft für empirische Bildungsforschung (GEBF), Berlin.

Lefering, R. (1996). *Klassifikationsbäume - Ein multivariates Prognosemodell in der klinischen Anwendung und im Vergleich zur logistischen Regression* (Dissertation). Köln: Universität.

Mrazek, J. (1979). *Verständnis und Verständlichkeit von Lesetexten*. Frankfurt am Main: Lang.

Myers, C. & Fucks, S. (2005). Klassifikations- und Regressionsbäume. In H. Moosbrugger, J. Hartig & D. Frank (Hrsg.), Studierendenauswahl (Riezlern-Reader XIV). Frankfurt am Main: Institut für Psychologie der J. W. Goethe-Universität. Zugriff am 27.06.2016 unter http://publikationen.ub.uni-frankfurt.de/oai/container/index/docId/2409

Nickl, M. (2001). *Gebrauchsanleitungen: Ein Beitrag zur Textsortengeschichte seit 1950*. Tübingen: Gunter Narr Verlag.

OECD (2009), *PISA 2009 Assessment Framework: Key competencies in Reading, Mathematics and Science,* OECD Publishing.

OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy,* OECD Publishing.

Ozuru, Y., Rowe, M., O'Reilly, T. & McNamara, D. S. (2008). Where is the difficulty in standardized reading tests: the passage or the question? *Behavior Research Methods, 40 (4)*, 1001–1015.

Parzen, E. (2001). Comment, *Statistical Science 16,* 224–226.

Pohl, S. & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Potapov, S. (2012). *Zur Verbesserung der Splitkriterien bei Klassifikationsbäumen und Ensemble-Methoden* (Dissertation, Universität Erlangen-Nürnberg).

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30* (2), 120–135.

Preston, R. C. (1964). Ability of Students to Identify Correct Responses Before Reading. *The Journal of Educational Research, 58* (4), 181–183.

Rankin, E. F. & Culhane, J. W. (1969). Comparable Cloze and Multiple-Choice Comprehension Test Scores. *Journal of Reading, 13* (3), 193–198.

Rausch, T., Matthäi, J. & Artelt, C. (2015). Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47* (3), 147–158.

Richter, T. & Christmann, U. (2002). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (S. 25–58). Weinheim, Germany: Juventa.

Roeschl-Heils, A., Schneider, W. & van Kraayenoord, C. E. (2003). Reading, metacognition and motivation: A follow-up study of German students in Grades 7 and 8. European *Journal of Psychology of Education*, 18 (1), 75–86.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. überarb. und erw. Aufl.). Bern: Hans Huber.

Rost, D. H. & Sparfeldt, J. R. (2007). Leseverständnis ohne Lesen? Zur Konstruktvalidität von multiple-choice-Leseverständnistestaufgaben. *Zeitschrift für Pädagogische Psychologie, 21* (3/4), 305–314.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*, 441–474.

Säuberlich, F. (2000). *KDD und Data Mining als Hilfsmittel zur Entscheidungsunterstützung*. Frankfurt am Main: Peter Lang.

Schaffner, E., Schiefele, U. & Schneider, W. (2004). Ein erweitertes Verständnis der Lesekompetenz: Die Ergebnisse des nationalen Ergänzungstests. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 197–242). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schnotz, W. & Dutke, S. (2004). Kognitionspsychologische Grundlagen der Lesekompetenz: Mehrebenenverarbeitung anhand multipler Informationsquellen. In U. Schiefele, C. Artelt, W. Scheider, & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 61–99). Wiesbaden: VS.

Schroeder, S. & Tiffin-Richards, S. P. (2014). Kognitive Verarbeitung von Leseverständnisitems mit und ohne Text. *Zeitschrift für Pädagogische Psychologie, 28* (1-2), 21–30.

Schweitzer, K. (2007). *Der Schwierigkeitsgrad von Textverstehensaufgaben. Ein Beitrag zur Differenzierung und Präzisierung von Aufgabenbeschreibungen*. Frankfurt am Main: Peter Lang.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly, 50* (3), 345–362.

Stanat, P. & Schneider, W. (2004). Schwache Leser unter 15-jährigen Schülerinnen und Schülern in Deutschland: Beschreibung einer Risikogruppe. In U. Schiefele, C. Artelt, W.

Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000*. (S. 243-273). Wiesbaden: VS Verlag für Sozialwissenschaften.

Strobl, Caroline (2008). *Statistical issues in machine learning - Towards Reliable Split Selection and Variable Importance Measures.* Zugriff am 27.06.2016 unter https://edoc.ub.uni-muenchen.de/8904/1/Strobl_Carolin.pdf

Tutz, Gerhard (2000). *Die Analyse kategorialer Daten: Anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression* (Lehr- und Handbücher der Statistik). München, Wien: Oldenbourg.

Tutz, Gerhard (2012). *Regression for categorical data.* Cambridge: University Press.

van Kraayenoord, C. E. & Schneider, W. (1999). Reading achievement, metacognition, reading self-concept and interest: A study of German students in grades 3 and 4. *European Journal of Psychology of Education*, 14 (3), 305–324.

Voss, A., Carstensen, C. H. & Bos, W. (2005). Textgattungen und Verstehensaspekte: Analyse von Leseverständnis aus den Daten der IGLU-Studie. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien* (S. 1–36). Münster: Waxmann.

Watermann, R. & Klieme, E. (2006). Modellierung von Kompetenzstufen mit Hilfe der latenten Klassenanalyse. *Empirische Pädagogik*, 20 (3), 321–336.

Weinert, F. E. (Hrsg.). (1998). *Entwicklung im Kindesalter*. Weinheim: Beltz Psychologie-Verlags-Union.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T. & Carstensen, C. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong Process. The German National Educational Panel Study (NEPS)*. Zeitschrift für Erziehungswissenschaft [Special Issue], vol. 14, pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.

Willenberg, H. (2007). Lesen. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 107–117). Weinheim: Beltz.

Willenberg, H. (2010). Ein handhabbares System, um Textschwierigkeiten einzuschätzen. Vorschläge für eine Textdatenbank von Sachtexten. In M. Fix & R. Jost (Hrsg.), *Sachtexte im Deutschunterricht. Für Karlheinz Fingerhut zum 65. Geburtstag* (Diskussionsforum Deutsch, Bd. 19, 2. unver. Auflage). Baltmannsweiler: Schneider-Verlag Hohengehren.

Willenberg, H., Gailberger, S. & Krelle, M. (2007). Argumentation. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 118–129). Weinheim: Beltz.

Zimmermann, S. (2016). *Entwicklung einer computerbasierten Schwierigkeitsprädiktion von Leseverstehensaufgaben* (NEPS Working Paper No. 64). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS): Entwicklungsstudie B98, Erwachsene und Studierende 2014. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom Bundesministerium für Bildung und Forschung (BMBF) finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) an der Otto-Friedrich-Universität Bamberg in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt

Schrift 4

Gehrer, K., Oepke, M. & Eberle, F. (in press). Der EVAMAR II-Deutschtest für GymnasiastInnen – Implikationen für die Plurizentrik-Debatte? In W. Davies, A. Häcki Buhofer, R. Schmidlin, M. Wagner & E. Wyss (Hrsg.), *Standardsprache zwischen Norm und Praxis. Theoretische Betrachtungen, empirische Studien und sprachdidaktische Ausblicke.* (Basler Studien zur deutschen Sprache und Literatur, Band 99). Tübingen: Francke Verlag.

Gehrer, Karin, Oepke, Maren & Franz Eberle:

Der EVAMAR II-Deutschtest für GymnasiastInnen – Implikationen für die Plurizentrik-Debatte?

Abstract

Mit dem folgenden Beitrag wollen wir, angeregt durch die Basler Podiumsdiskussion zum Spannungsfeld der verschiedenen nationalen Varietäten der deutschen Sprache (Ammon 1995; zur Gegenposition Koller 1999), der Frage nachgehen, ob die für die Schweiz repräsentativen empirischen Daten der Studie EVAMAR II (Eberle et al. 2008) für die sprachwissenschaftliche Plurizentrik-Debatte innerhalb des Deutschen ein gewisses Analysepotenzial bieten und überprüfen, ob wir für sprachliche Leistungsunterschiede auf universitärem Niveau empirische Hinweise auf den Einfluss der Familiensprache finden.

Seit der Maturitätsevaluation EVAMAR II wissen wir empirisch gesichert, dass die Schweizer Maturandinnen und Maturanden an der Schnittstelle Gymnasium – Universität eine auffallend breite Streuung der getesteten Leistungen und Fähigkeiten aufweisen, mit teilweise beachtlichen Defiziten (Eberle et al. 2008: 16). Die Studie EVAMAR II belegte für die Deutschschweiz und die Romandie [1] an für den Matura-Abschlussjahrgang 2007 repräsentativen Zielpersonen, dass diese eine ausgesprochen breite Varianz von Kompetenzen in den getesteten Bereichen Mathematik, Biologie und Deutsch bzw. Französisch aufwiesen. Diese stark unterschiedlichen Leistungen spiegeln die heterogenen Oberstufencurricula und unterschiedlichen gymnasialen Prüfungsanforderungen, gegebenenfalls auch die Stadt-Land- oder Kantonsgrößenunterschiede innerhalb der Bildungslandschaft Schweiz wider.

Auch in dem uns hier in diesem Sammelband ausschließlich interessierenden Sprachbereich des Deutschen zeigten die Ergebnisse des EVAMAR II-Tests, obwohl die Resultate etwas besser waren als in Mathematik und Biologie, sowohl zwischen den Personen als auch zwischen den Klassen eine auffallend breite Streuung der Leistungen und Fähigkeiten. Daraus lässt sich folgern, dass ein gewisser Teil der möglichen StudienanfängerInnen für ein künftiges Studium an den Universitäten nicht die notwendigen sprachlichen Eingangskompetenzen mitbringt (vgl. Eberle et al 2008: 374, 383-384). Wenn man dann sieht, dass die Mehrheit der längsschnittlich befragten Studierenden das Wissen und Können aus dem gymnasialen Fach Deutsch für ihr im Studium derzeit gewähltes Hauptfach als „eher wichtig“ einstufen (vgl. das Nachfolgeprojekt von EVAMAR II, Oepke/ Eberle 2014), könnten uns diese sprachlichen Lücken eher beunruhigen.

Vor dem Hintergrund der Plurizentrik-Debatte möchten wir diese sprachlichen Leistungsunterschiede nun im Folgenden daraufhin prüfen, ob sie gegebenenfalls von der Zugehörigkeit zu der einen oder anderen „Sprachgruppe“ abhängig oder beeinflusst sind. Mittels im Fragebogen erfragter Variablen zur Familiensprache und dem Geburtsland der Eltern hofften wir, dass es möglich sei, mehrere vergleichbare Gruppen innerhalb der für die Schweiz großen Gesamtstichprobe (N = 3800) zu identifizieren, die bezüglich ihrer Sprachverwendung bzw. Sprachherkunft unterschiedlich sind.

[1] Das italienischsprachige Tessin wies leider aus verschiedenen Gründen eine niedrig ausgefallene Beteiligungsquote der Schulen an der Kompetenzmessung auf. Die Ergebnisse des Tessins konnten somit nicht in die Auswertungen dieses Teilprojektes einbezogen werden (Eberle et al. 2008: 374).

Die Fragestellung, ob zwischen der Gruppe der MaturandInnen mit Schweizerdeutsch bzw. Schweizerdeutscher Standardsprache (Gruppe 1) und den Gruppen der MaturandInnen mit Bundesdeutscher Standardsprache eines Elternteils (Gruppe 2) respektive beider Eltern (Gruppe 3) beim Ablegen einer Kompetenztestung in Deutsch auf dem Niveau von universitären Anforderungen starke Leistungsunterschiede auftreten oder nicht, bildete den Fokus unserer Untersuchung. Die Analysen ergeben das Bild, dass die vermuteten Gruppen relativ klein und (leider) nur unter gewissen Annahmen interpretierbar sind. Unter diesem Vorbehalt zeigen die Ergebnisse, dass die Leistungen der drei gebildeten sprachlichen Gruppen weder im Gesamttest noch in den einzelnen Subskalen (Grammatik/Orthographie, Leseverstehen, Wortschatz) signifikant voneinander abweichen. Somit existieren auf den ersten Blick innerhalb der Schweizer Bevölkerung anscheinend keine auffälligen Sprachstandsunterschiede bei MaturandInnen aus einem Elternhaus mit Sozialisation in Schweizerdeutsch und Schweizerdeutscher Standardsprache gegenüber MaturandInnen aus einem Elternhaus mit Bundesdeutscher Standardsprache. Inwiefern diese Ergebnisse unter welchen Annahmen belastbar sind oder nicht, diskutieren wir im Folgenden.

1 Einleitung – EVAMAR und Plurizentrik

Die Frage, die im Rahmen der Tagung zur Plurizentrik in Basel an uns herangetragen wurde, ist folgende: Was können wir mit der Studie EVAMAR II und deren Ergebnissen zu den Erstsprachekompetenzen von Maturandinnen und Maturanden bezüglich des Forschungsgedanken der Plurizentrik des Deutschen beitragen?

Breiter gefasst meint dies vielleicht auch den grundsätzlicheren Gedanken, was und wie Large-Scale-Assessments (LSA), bzw. empirische Erhebungen von Sprachfähigkeiten, zur Thematik der Plurizentrik beitragen könnten, auch wenn sie nicht eigens für die Untersuchung dieser speziellen Fragen ausgerichtet wurden. Es ist[2] im Rahmen von wissenschaftlicher Öffentlichkeit und vernetzter Ressourcennutzung einige Gedanken wert, sich zu überlegen, wie breitangelegte, qualitativ hochwertige vorhandene Daten der Schweizerischen Eidgenossenschaft genutzt werden könnten, sich der Beantwortung von Plurizentrikfragen auf empirischem Wege anzunähern. Gerade LSA-Daten bieten mit ihren für die Grundgesamtheit repräsentativen großen Fallzahlen eine mächtige Aussagekraft, welche kleiner angelegte Studien in einem engeren Forschungsrahmen so nicht aufbringen können. Empirische Kompetenzmessungen zum Bereich Deutsch beziehen sich auf die normierten Standardvorgaben von Deutsch als erster Schulsprache (L1), erhoben in standardisierter Weise, meist im Gruppenkontext. An der Pyramidenspitze der Sprachnorm[3] der deutschen Standardsprache, wie sie insbesondere an Gymnasien und Universitäten gelehrt und verwendet wird, werden die ermittelten Personenfähigkeiten gemessen und die erfassten Sprach- oder Kompetenzniveaus entlang einer vertikalen Messlatte unterteilt. Ob innerhalb einer solchen normierten Erhebung bezüglich der Standardsprache Deutsch empirisch begründete und verlässliche Aussagen zum Einfluss von plurizentrischen Faktoren auf den Kompetenzerwerb gemacht werden könnten, wollen wir im Folgenden anhand von einigen Überlegungen und ausgewählten Analysen versuchsweise nachgehen.

---

[2] (– dankenswerterweise nicht zuletzt auf löblichen Wunsch der Herausgeberinnen dieses Bandes –)

[3] Neuland verwendet das eingängige Bild einer „Stilpyramide“ des Deutschen, mit der „Hochsprache als selbstevidenter Zielnorm an der Spitze und den Dialekten als historische Sprachformen an der Basis“ (Neuland 2004: 3)

Wir nähern uns der Thematik an, indem wir zuerst zur Verständlichkeit und Einbettung der Studie einen allgemeinen Überblick über die EVAMAR-Projekte geben, gefolgt von einer Beschreibung der eingesetzten Testinstrumente und einem kurzen Einblick in die querschnittlichen Gesamtergebnisse des Bereichs Deutsch, kurz vor dem Abschluss der Schweizerischen Maturitätsprüfung und dem damit verbundenen Erreichen der Studierfähigkeit (ausführlich dazu der Schlussbericht: Eberle et al. 2008). In dem anschließenden empirischen Teil berichten wir von einer Analyse, die wir basierend auf den vorhandenen EVAMAR II-Daten zur Thematik Plurizentrik vorgenommen haben.

Da wir unseren Artikel als einen interdisziplinären Beitrag verstehen, verzichten wir darauf, uns vertieft in die sprachwissenschaftlichen Richtungen und Begriffsklärungen einzubringen. Wir sind uns jedoch bewusst, dass das Forschungsgebiet der Deutschschweizer Sprachlandschaft herausfordernd ist mit seinem „Nebeneinander von Mundarten und Standardsprache" (Sieber/ Sitta 1986: 16; für einen einfachen Überblick Siebenhaar/ Wyler 1997). Die Besonderheit der deutschsprachigen Schweiz ist unseres Wissens konzeptionell noch nicht einheitlich gefasst (Christen 2004: 13-14). So wird dieses Phänomen von den einen als mediale Diglossie gesehen – i.S.v. „man *schreibt* Standardsprache, man *spricht* Mundart" (Sieber/ Sitta 1986: 20). Von anderen wird es weiterhin als „kanonische" Diglossie von genetisch verwandten Sprachformen bzw. zwei Varietäten der gleichen Sprache klassisch nach Ferguson verteidigt (Haas 2004: 83), und in Gegenpositionen u.a. diskutiert als ein „besondere[r] Fall der Zweisprachigkeit" (Berthele 2004: 131), bei der die Steuerungsfaktoren von Nähe und Distanz, von Informalität und Formalität entscheidend sind für die Wahl der jeweiligen sprachlichen Varietät (Berthele 2004: 85). Im Unterschied zu Deutschland mit seinen ebenfalls vielen Dialekten kann in der Schweiz ein Zuwachs an Prestige nicht einfach über die Verwendung des Standarddeutschen erreicht werden[4], weil die Mundarten in der Schweiz *die* Umgangssprache sind und meist eine hohe Wertschätzung besitzen (Sieber/ Sitta 1986: 169), und somit ist „jemand, der in einer Alltagssituation (Hoch-)Deutsch *spricht*, gerade kein (Deutsch-)Schweizer" (Koller 1999: 139). Für den Spracherwerb der Schweizerischen Standardsprache, des Hochdeutschen[5], sprechen wichtige empirische Studien von einem „erweiterten Erstspracherwerb"[6] (Häcki Buhofer/ Burger 1998: 137), wobei beim vorschulischen Lernprozess auch einige Züge von Zweitspracherwerb[7] zu beobachten sind (Häcki Buhofer/ Burger 1998: 89; qualitativ u.a. auch Schneider 1998). Für die muttersprachlich schweizerdeutschen Kinder fängt vor- und außerschulisch meist über den Kontakt mit dem Medium Fernsehen ein frühes Annähern an die Standardsprache an (Häcki Buhofer/ Burger 1998: 42-51). Die professionelle Sozialisierung in die Standardsprache erfolgt für die meisten schweizerdeutschen Kinder erst mit dem Eintritt in die Grundschule. Es gibt inzwischen jedoch empirische Hinweise für positive Effekte zugunsten des Standarddeutschen, wenn die sprachliche Sozialisation früher, also bereits in der

[4] „Während man die H-Varietät [=Standardsprache, Anm.d.V.] in FERGUSONs Paradigma braucht, um sozial **auf**zusteigen, braucht man in der deutschen Schweiz L [= Dialekt], um überhaupt sozial **ein**zusteigen." (Berthele 2004: 119)

[5] Da in der Schweiz die Begrifflichkeit „Hochdeutsch" für den Gebrauch der schweizerdeutschen Standardsprache in Abgrenzung zu den schweizerdeutschen Dialekten/ Mundarten gebräuchlich ist, und teilweise auch von SprachwissenschaftlerInnen durchaus weiterhin verwendet wird (bspw. auch in Häcki Buhofer 2000; Landert Born 2011), möchten wir uns vorbehalten, deutsche Standardsprache und Hochdeutsch in diesem Bericht teilweise oszillierend zu verwenden. Da, wo die konkrete nationale Varietät der deutschen Sprache angesprochen wird, sprechen wir innerhalb der deutschen Standardsprachen von schweizerdeutscher Standard- bzw. bundesdeutscher Standardsprache.

[6] Formen aus der Mundart werden im ungesteuerten Spracherwerb in einem direkten Transfer oder über Transferregeln ins Hochdeutsche der Lernersprache aufgenommen (Häcki Buhofer/ Burger 1998: 88-89).

[7] Zweitsprachliche Differenzierungsstrategien insbesondere bei der Verwendung des Präteritums (Häcki Buhofer/ Burger 1998: 88-89).

Kindergartenstufe, beginnt (Landert Born 2011). Nicht zuletzt die schlechten PISA-Ergebnisse der Schweiz in Deutsch/Lesen im Jahr 2000 veranlassten einige Kantone dazu, zugunsten eines verbesserten frühen Spracherwerbs der Standardsprache bereits im Kindergarten das „Hochdeutsche" als gesprochene Sprache einzuführen.

Für diesen Sammelband zur Plurizentrik sind jedoch nicht die Fragen nach dem Spannungsfeld zwischen Standarddeutsch und Dialekt(en) wesentlich, sondern vielmehr das Spannungsfeld zwischen den nationalen Standard-Varietäten des Deutschen im Sinne Ammons (1995), also hier zwischen bundesdeutschem Standarddeutsch (mit seinen Teutonismen) und Schweizerischem Standarddeutsch (mit seinen Helvetismen)[8]. In diesem Zusammenhang wollen wir hier bereits kritisch darauf hinweisen, dass die bei EVAMAR erhobenen Daten nicht auf sprachwissenschaftliche Forschungsfragen, sondern auf Kompetenzmessungen und bildungspolitisch relevante Informationen bezüglich Schnittstellenpassung ausgerichtet waren; nichtsdestotrotz soll ein Versuch gewagt werden, zu prüfen, ob die EVAMAR II-Daten darüber hinaus gewisse erste empirische Hinweise innerhalb der sprachwissenschaftlichen Plurizentrik-Debatte anbieten können.

2 Die EVAMAR-Studien

2.1 Allgemeines

Im Dezember 2004 wurden in der Schweiz die Ergebnisse der ersten Phase der Evaluation der Maturitäts-Reform von 1995 (EVAMAR I; vgl. Ramseier et al. 2004) vorgestellt. Ziel dieser ersten Untersuchung war, das Gelingen der schweizerischen Gymnasialreform von 1995, welche die Wahlmöglichkeiten erweitert und die Maturarbeit eingeführt hatte, mittels Befragungen der Schülerschaft, von Lehrpersonen und Schulleitungsmitgliedern zu begutachten. Von März bis August 2003 wurden über 21000 GymnasiastInnen, 2300 Lehrpersonen sowie Schulleitungen von 148 Gymnasien in drei Landessprachen der Schweiz befragt (Ramseier et al. 2004: 5). Die Bilanz dieser Untersuchung ist insgesamt positiv ausgefallen; unter anderem fühlten sich 76% der MaturandInnen generell (eher) gut auf ein Hochschulstudium vorbereitet (Ramseier et al. 2004: 122, 145).

Die daraufhin folgende EVAMAR II-Studie (vgl. Eberle et al. 2008) hatte die Überprüfung des gymnasialen Bildungsziels der allgemeinen Studierfähigkeit zum Ziel und damit eine Einschätzung der Passgenauigkeit der Schnittstelle zwischen Gymnasium und Universitäten bzw. Hochschulen. Allgemeine Studierfähigkeit im Sinne der Ziele des Maturitätsreglements von 1995 meint dabei, dass das Schweizer Gymnasium für jedes Studium zu qualifizieren hat.

Im ersten Teil der verschiedenen Teilprojekte wurden einerseits die Anforderungen von Dozierenden erfragt und andererseits die universitären Lehrmittel sowie Prüfungen für Erst- und Zweitsemestrige der 16 gemessen an den Studierendenzahlen größten universitären Studienfächer der Schweiz analysiert, um damit die Anforderungen an den Eintritt in ein Studium bzw. an das Konstrukt allgemeine Studierfähigkeit empirisch zu bestimmen. Die Entwicklung von Tests aufgrund dieser Ergebnisse wird im folgenden Abschnitt näher beschrieben.

Den zweiten zentralen Teil bildeten die schweizweit bei einer Stichprobe von rund 3800 MaturandInnen im Jahr 2007 durchgeführten Tests, welche zum Ziel hatten, den Ausbildungsstand der Maturandinnen und Maturanden am Ende des schweizerischen

[8] Koller spricht in diesem Zusammenhang etwas ironisch von einer Verschärfung der Diglossie hin zu einer „*Triglossie Schweizerdeutsch / schweizerische nationale Varietät des Deutschen / (bundes-)deutsche Standardsprache*" (1999: 152).

Gymnasiums im Hinblick auf die Anforderungen verschiedener Studien zu erheben, und damit deren allgemeine Studierfähigkeit zu validieren. Untersucht wurde der Stand in der jeweiligen Landessprache, in Mathematik und Biologie, sowie die überfachlichen Fähigkeiten anhand von Aufgaben aus der (naturwissenschaftlich orientierten) Eignungsprüfung für das Fach Medizin. Der Sprachtest war territorial definiert, d.h. es wurde der Sprachstand Deutsch getestet in den deutschsprachigen resp. Französisch in den französischsprachigen und Italienisch in den italienisch- oder rätoromanisch-sprachigen Gebieten. Jeder Test dauerte 45 Minuten. Das Hauptgewicht des Sprachtests lag auf übergreifenden, für alle Studienrichtungen wesentlichen Sprachkompetenzen, und er beinhaltete sowohl rezeptive Teile (Textverstehen, Wortschatz) als auch produktive bzw. reflexive Anteile (Grammatik, Orthographie). Das Ergebnis war insgesamt zufriedenstellend (Eberle et al. 2008: 383). Aber es zeigten sich auch große Leistungsunterschiede sowohl zwischen Einzelnen als auch zwischen Klassen und Profilen und zwischen Kantonen mit unterschiedlichen Maturitätsquoten (Eberle et al. 2008: 372-379). So konstatierte abschließend der Projektleiter Prof. Franz Eberle: „Die Auswertung der Testergebnisse offenbart in allen Bereichen eine erstaunlich breite Streuung, vor allem angesichts der Tatsache, dass die Schülerinnen und Schüler kurz vor der Verleihung der für alle Studienfächer geltenden, universalen Qualifikation ‚Hochschulreife' standen. Es kann deshalb davon ausgegangen werden, dass nicht alle Maturandinnen und Maturanden in allen drei getesteten Fachbereichen über Kompetenzen verfügen, die den universitären Anforderungen aller Studienfächer entsprechen." (Eberle et al. 2008: 374). „Die gefundene breite Streuung der Testresultate bedeutet gleichzeitig auch, dass die Gymnasien nicht alle ihre Maturandinnen und Maturanden mit Kompetenzen entlassen, die in der ganzen Breite als mindestens genügend eingeschätzt werden können" (Eberle et al. 2008: 383).

Zusätzliche Analysen der Notendaten zeigten, dass im Bereich Sprache 4,7 % der Maturandinnen und Maturanden als insgesamt ungenügend qualifiziert wurden (Gesamt-Maturanote 3.9 oder tiefer[9]) (Eberle et al. 2008: 170). In der schriftlichen Abschlussprüfung Sprache (meist in Form eines Aufsatzes) waren sogar 19,6 % der Maturandinnen und Maturanden ungenügend (Eberle et al. 2008: 170, 375).

Wie wichtig gute Deutschkompetenzen für die allgemeine Studierfähigkeit und für den Studienerfolg sind, wurde unter anderem in einer vom Schweizerischen Nationalfonds geförderten Längsschnitt-Nachfolgestudie untersucht, mit welcher von Januar 2011 bis März 2013 die deutschsprachige Teilstichprobe der EVAMAR II-Stichprobe weiter begleitet wurde. Neben den weiteren Wegen der ehemaligen GymnasiastInnen wurde auch die Bedeutung der bei der Matura erhobenen Kompetenzen auf den späteren Studienerfolg und damit die Passung des hinter dem heterogenen Konstrukt der allgemeinen Studierfähigkeit steckenden Wissens und Könnens mit den Anforderungen der verschiedenen Studienrichtungen analysiert (vgl. Oepke/ Eberle 2010; 2014: 189). Dabei zeigte sich unter anderem, dass insbesondere zum Zeitpunkt des Abiturs gute Deutschkompetenzen zum späteren Studienerfolg beitragen können und dies auch in Studienfachgruppen, die als eher „mathematiklastig" gelten (Oepke/ Eberle 2016).

[9] In der Schweiz ist die Note 6 die beste, Note 1 die schlechteste. Noten unter 4.0 gelten als ungenügende Qualifikation.

2.2 Der Sprachtest

Im Projekt EVAMAR II wurde zur Kompetenzmessung der schriftsprachlichen Fähigkeiten und Fertigkeiten von Studierenden auf der Grundlage eines theoretisch fundierten Kompetenzrasters ein umfassender Sprachtest entwickelt[10]. Das EVAMAR-Kompetenzraster, auf dem der Test beruht, wurde in Anlehnung an den Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER) im Kernteam von einer Germanistin konzipiert und versteht sich als eine für die Textsorte „universitäre Fachtexte“ auf dem Niveau von kompetenter Sprachverwendung eigenständige Erweiterung (Eberle et al. 2008: 81-85). GER bietet in Europa einen umfassenden, transparenten und kohärenten Referenzrahmen für das Erlernen und Lehren von Sprachen, seine Kompetenzniveaus sind von A1 bis C2 für lebenslanges Lernen definiert und empirisch kalibriert (Europarat 2001).

Die Bedeutung spezifischer Textsorten für die Testkonstruktion ist im internationalen Rahmen von Kompetenzmessung in Large-Scale-Assessments (LSA) anerkannt; es werden literarische Texte, Sachtexte, aber auch Anweisungen, argumentative Texte und diskontinuierliche Texte eingesetzt, um verschiedene Anforderungen an die sprachlichen Kompetenzen erfassen zu können (vgl. bspw. Gehrer/ Zimmermann/ Artelt/ Weinert 2013: 57-58). Da jede Textsorte andere spezifische Anforderungen stellt (vgl. bspw. Gehrer/ Artelt 2013: 172-173, 175-183), ist es wesentlich, für die valide Erfassung von Kompetenzen der Ziel- und Altersgruppe angemessene Textsorten zu wählen. Für die Überprüfung der Studierfähigkeit im sprachlichen Bereich bei EVAMAR II stellt die gewählte Textsorte „universitäre Fachtexte“ den passenden validen Texttyp dar.

Der EVAMAR II-Sprachtest umfasst die beiden ausgewählten großen Bereiche Verstehen und Sprachreflexion (siehe Abbildung 1), wobei im rezeptiven Bereich Verstehen die Subskala ‚Allgemeines Leseverstehen’ im Sinne eines Lesens zur Orientierung sowie die Subskala ‚Detailliertes Leseverstehen’ im Sinne von Informationen und Argumentation verstehen eingesetzt wurden. Im Bereich Sprachreflexion wurden in der Subskala ‚Wortschatz’ einerseits rezeptive Testaufgaben zum Wortschatz-Spektrum und zur Wortschatz-Beherrschung konstruiert andererseits auch sprachproduktionsnahe Aufgabenstellungen entwickelt (siehe Abbildung 2). In der Subskala ‚Grammatische Kompetenz’ und ‚Orthographie’ wurden mittels überwiegend offenen Formaten ebenfalls produktionsorientierte Aufgabenstellungen generiert (siehe Abbildung 3).

Die Studie und die damit verknüpften Testentwicklungen konzentrierten sich wie beschrieben auf die grundlegenden Kompetenzen, welche wesentlich sind, um die Anforderungen der ersten Semester an einer Universität zu erfüllen. Damit wurde innerhalb des Bereiches der Bildungs- bzw. Wissenschafts- (oder Fach-)sprache neben Teilaspekten von produktionsorientierter reflexiver Sprachkompetenz wie Grammatik und Orthographie großes Gewicht auf Teil-Kompetenzen gelegt, welche das Lesen und Textverstehen betreffen, davon ausgehend, dass Lesen sowohl „die Basiskompetenz [bildet], mithilfe derer neues Wissen angeeignet wird“ (Eberle et al 2008: 83) als auch dass mit dieser grundlegenden rezeptiven Tätigkeit viel Zeit innerhalb eines universitären Studiums verbracht wird. Die im Hinblick auf das spätere Verfassen wissenschaftlicher Arbeiten – und für die Forschungsfragen der Plurizentrik-Debatte – bedeutungsvolle Kompetenz der schriftlichen Sprachproduktion als auch die Teilkompetenz „Sprechen“ konnte aus Ressourcengründen leider nicht in das

[10] Die Entwicklung des Sprachtests erfolgte in Deutsch; die Übersetzung in die zwei anderen Landessprachen erfolgte durch je zwei Personen und wurde durch die zweisprachigen Sprachexpertinnen der Projektpartner Romandie und Tessin feinübersetzt und gemeinsam mit der Testentwicklerin justiert (Eberle et al. 2008: 128).

Untersuchungsfeld einbezogen werden.[11] Teile produktiver Sprachkompetenzen lassen sich hingegen in den offenen Formaten finden, die vorwiegend in den Grammatik- und Orthographie-Subskalen Verwendung fanden (Eberle et al 2008: 82-83).

Abb. 1: Kompetenzraster für die EVAMAR II-Sprachsubkalen (mit Kann-Beschreibungen)

			Textsorten-Orientierung
			Universitäre Textsorte
VERSTEHEN	**Lesen**	**Allgemeines Leseverstehen (Zur Orientierung lesen)**	Kann alle Arten geschriebener Texte verstehen und kritisch interpretieren (einschliesslich abstrakter, strukturell komplexer, nicht-literarischer Texte). Kann ein breites Spektrum langer und komplexer Texte, *auch wissenschaftliche* Texte, verstehen und dabei feine stilistische Unterschiede und implizite Bedeutungen erfassen. Kann lange und komplexe Texte, *auch wissenschaftliche Texte,* rasch durchsuchen. Kann rasch den Inhalt und die Bedeutung von Artikeln und Texten zu einem breiten Spektrum *wissenschaftlicher* Themen erfassen und entscheiden, ob sich ein genaueres Lesen lohnt.
		Detailliertes Leseverstehen: Information & Argumenta-tion verstehen	Kann ein weites Spektrum langer, komplexer Texte, *denen Studierende im ersten Studienjahr an der Universität begegnen*, verstehen und dabei feinere Nuancen auch von explizit oder implizit angesprochenen Einstellungen und Meinungen erfassen. Kann wichtige Einzelinformationen auffinden.[12]
SPRACH-REFLEXION		**Wortschatz**	Beherrscht einen sehr reichen Wortschatz und ist sich der jeweiligen Konnotation bewusst. Durchgängig korrekte und angemessene Verwendung des Wortschatzes.
		Grammatische Kompetenz	Zeigt auch bei der Verwendung komplexer Sprachmittel eine durchgehende Beherrschung der Grammatik.
		Orthographie	Die schriftlichen Texte sind frei von orthographischen Fehlern.

(Aus: Eberle et al. 2008: 85)

2.3 Das Instrument

Die Entwicklung der konkreten Testaufgaben (Items) zur Überprüfung des Textverstehens, des Wortschatzes, der Grammatik und Orthographie erfolgte in einem aufwändigen

[11] Ebenso musste die rezeptive Hörkompetenz aus dem Design entfallen (Eberle et al 2008: 83).

[12] „Die [GER-]Teilkompetenz „wichtige Einzelinformationen auffinden“ wurde für EVAMAR II aus inhaltlichen Gründen aus dem Bereich des Allgemeinen zum Detaillierten Leseverstehen umgeordnet.“ (Eberle et al 2008: 85)

mehrstufigen Verfahren anhand authentischen Textmaterials (Vorlesungen, Skripte, universitäre Lehrbücher) der 16 gemessen an den Studierendenzahlen größten universitären Studienfächer der Schweiz. Durch die Inhaltsanalysen der von den Universitäten zur Verfügung gestellten Studienmaterialien wurden Sinneinheiten[13] identifiziert, die nicht im entsprechenden Lehrmaterial vermittelt, sondern aufgrund des Besuchs gymnasialen Unterrichts als vorhanden vorausgesetzt wurden und somit als ‚Eingangswissen' gelten (Eberle et al 2008: ausführlich zur Methode des Eingangswissen Identifizierens 36-44). Jedes Test-Item bezieht sich auf eine Textstelle / Sinneinheit der authentischen Studienmaterialien und wurde „konstruiert mit dem Ziel zu messen, ob eine Maturandin oder ein Maturand eine Sinneinheit in einer gewissen Tiefe verstanden hat" (Eberle et al 2008: 120) bzw. damit auch produktiv umzugehen weiß.
Zusätzlich zu den allgemein gültigen Leitlinien für die Konstruktion von Testaufgaben wurde die Vorgabe umgesetzt, dass die Items „so realistisch wie möglich die konkreten kognitiven Anforderungen widerspiegeln, mit denen die Studentin beziehungsweise der Student im ersten Semester konfrontiert ist. Alle Items beziehen sich [deshalb] auf die Situation des Lesens und Verstehens [und Sich-Aneignens] von Studienunterlagen (Skripte und Bücher)" (Eberle et al. 2008: 120). Für die Sprachitems wurde ein Verfahren gewählt, welches innerhalb der zugrunde gelegten Fachbücher die jeweils ersten Kapitel favorisierte, welche chronologisch entlang des Vorlesungsverlaufes von den Erstsemester-Studierenden zu bearbeiten waren (Eberle et al. 2008: 123).

Für die Items im Kompetenzbereich ‚Wortschatz' wurden häufig kodierte Sinneinheiten ausgewählt; diese repräsentieren vorausgesetztes Wortschatzwissen, welches in Bezug zum engen Lesekontext des Studienmaterials abgefragt werden konnte. Um erhöhten Korrekturaufwand zu vermeiden, wurden die Wortschatzitems meist im Multiple-Choice-Format entwickelt (Beispiel siehe Abbildung 2).
Für die Entwicklung von Testaufgaben zur Überprüfung der grammatikalischen und orthographischen Fähigkeiten wurden ebenfalls mittels der beschriebenen Inhaltsanalysen identifizierte Stellen authentischen Studienmaterials ausgewählt und vorwiegend über offene Formate beispielsweise die Produktion von Satzbauveränderungen eingefordert. Als Beispiel einer Aufgabenstellung in Grammatik gilt folgendes: „Verkürzen Sie die folgenden Sätze so, dass sie formal keinen Nebensatz mehr aufweisen, aber weiterhin alle Informationen vorhanden sind." Und eine weitere Aufgabenstellung im Grammatikteil: „Formen Sie in den folgenden Sätzen die unterstrichenen Satzteile zu Nebensätzen um." Nebst den offenen Formaten wurden auch produktionsnahe Grammatikitems im MC-Format konstruiert (siehe Abbildung 3).
Offene Formate wurden auch im Bereich der Orthographie eingesetzt, wie beispielsweise die Aufgabenstellung „Unterstreichen Sie alle Fehler im folgenden Text und schreiben Sie die Verbesserung des betreffenden Wortes oder des Satzteiles daneben. Korrigieren Sie auch stilistische Fehler."
Die Grundlage der Items für das allgemeine und detaillierte Leseverstehen bildeten ebenfalls häufig codierte Sinneinheiten bzw. Sinneinheiten-Häufungen; teilweise musste hier (insbesondere um Schwierigkeit zu generieren) der Bezug erweitert werden auf einen größeren Textzusammenhang (Eberle et al. 2008: 123; zur Veranschaulichung des Verfahrens ‚vom Lehrbuch zur konkreten Aufgabe' anhand der Sinneinheiten siehe insbesondere: 124-127). Die Items in den Bereichen der Lesenverstehensaufgaben wurden sowohl im geschlossenen als auch im offenen Format (siehe Abbildung 4) konstruiert.

[13] „Sinneinheiten können Fachausdrücke, Kategorien, Klassifikationen, Prinzipien bis hin zu einer ganzen Theorie beziehungsweise einem ganzen Modell sein." (Eberle et al 2008: 120)

Abb. 2: Beispiel einer sprachproduktionsnahen Wortschatz-Aufgabe

Aufgabe 1.2 **Welches Substantiv ersetzt in folgendem Kontext inhaltlich am besten das Fremdwort „...repertoire“?** **Vernachlässigen Sie die dadurch erforderlichen Beugungen.** „Wir können sagen, dass diese Reaktion zu den ursprünglichen und natürlichen Verhaltensmöglichkeiten gehört; sie ist Bestandteil eines zumindest in seinem Grundbestand angeborenen Verhaltensrepertoires.“ (Steiner [2001], S. 16) 1 ❑ ...mechanismus 2 ❑ ...vorrat 3 ❑ ...instinkt 4 ❑ ...weise	D_W_12

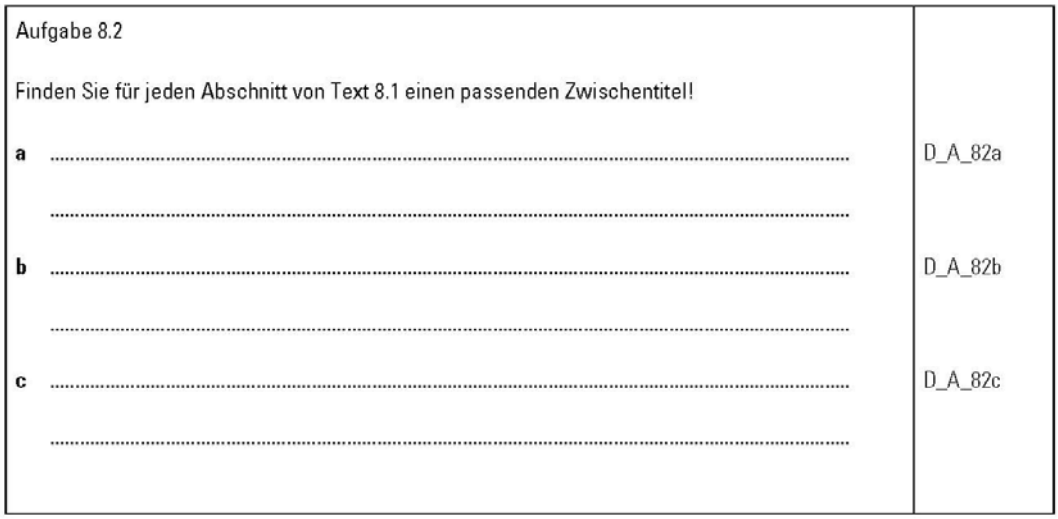
(Aus: Eberle et al. 2008: 125)

Abb. 3: Beispiel einer leichten produktionsnahen Grammatikaufgabe

Textauszug 8.1 (…) Ein weiteres Beispiel für eine Theorie ist die Rezeptortheorie. Die Zuckerkrankheit, der Diabetes mellitus, ist eine schon seit langem bekannte Erkrankung und wurde früher als Ausscheidung honigsüssen Urins bezeichnet. Frerichs, ein Berliner Pathologe, beobachtete im vorletzten Jahrhundert bei Patienten mit einem Diabetes mellitus, dass die Bauchspeicheldrüse dieser Patienten unter dem Mikroskop anders aussah als bei Gesunden und damit der Diabetes eine Erkrankung der Bauchspeicheldrüse ist. Claude Bernard, der berühmte französische Physiologe und Gründervater der experimentellen Medizin, wies nach, dass Zucker in der Leber produziert wird, und so zählte man den Diabetes zu den Lebererkrankungen. (…) (Steurer [2005]. *Wissenschaftstheoretische Grundlagen der Medizin*, Skript zur Vorlesung, OLAT-pdf, S. 7–9)	
Aufgabe 8.6 Sie finden in oben stehendem Text 8.1 mehrere Sätze, in denen „Diabetes“ vorkommt. Welches grammatikalische Geschlecht hat dort dieses Wort? Kreuzen Sie die richtige Antwort an. 1 ❑ femininum 2 ❑ neutrum 3 ❑ maskulinum 4 ❑ keines	D_G_86

(Aus: Eberle et al. 2008: 149)

Abb. 4: Beispiel einer Aufgabe im offenen Format[14]



(Aus: Eberle et al. 2008: 153)

2.4 Die Items

Nach der qualitativen Beurteilung der Sprachaufgaben durch einen professoralen Experten wurden in zwei Runden Vortests in Abschlussklassen an ausgewählten Gymnasien (n = 65; bzw. n = 180) durchgeführt, beide Pretests dienten der Optimierung der Testitems. Nach Pretests in allen drei Sprachregionen erfolgte eine weitere, dieses Mal gemeinsame Optimierung. Die Überprüfung und allfällige Revision der Items bzw. Auswahl für die Haupterhebung erfolgte aufgrund der Kriterien Rasch-Modellpassung „Infit" (zwischen 0.8 und 1.2), Trennschärfekoeffizient der Einzel-Items (>0.2), ausgewogener Distraktorenverteilung und Rasch-konformem Verlauf der „Item Characteristic Curve" (ICC) (Eberle et al. 2008: 121-122).

Insgesamt wurden 77 ‚übergreifende' Items in die Hauptauswertung aufgenommen, diese waren sowohl für Deutsch als auch für Französisch valide und reliabel. Für Deutsch konnten zusätzlich 47 Items aus dem Grammatikteil ergänzt werden. In die Hauptauswertung gingen für Deutsch somit 124 Test-Items mit zufriedenstellenden Testcharakteristika ein. Sie verteilen sich wie folgt auf die Kategorien des oben beschriebenen Kompetenzrasters:

- Allgemeines Leseverstehen (Zur Orientierung lesen): 19 Items
- Detailliertes Leseverstehen (Information und Argumentation verstehen): 40 Items
- Wortschatz: 19 Items
- Grammatische Kompetenz: 17 Items
- Orthografie: 29 Items

In die beiden Oberformate ‚offen' und ‚geschlossen' lassen sich die Aufgaben wie folgt einteilen: Knapp 60 Prozent aller Sprachitems wurden im offenen Format konstruiert (insgesamt 74 Items), währendem rund 40 Prozent der Aufgaben (50 Items) im geschlossenen Format gehalten wurden; davon gut zwei Drittel im klassischen Multiple-Choice-Format

[14] Weitere exemplarische Beispiele des Sprachtests siehe Eberle et al. 2008: 146-154.

(MC; Ankreuzen der richtigen Antwort aus mehreren Optionen). Etwa ein Drittel der weiteren geschlossenen Aufgabenformate waren Variationen wie etwa Zuordnungsaufgaben (z. B. Zuordnung von vorgegebenen Titeln zu Textabschnitten) oder die Setzung von Marginalien.

‚Offen lang' waren 14 anspruchsvolle Aufgabenstellungen (18.6% aller offenen Aufgaben), bei denen beispielsweise mehrere Zwischentitel für längere Abschnitte wissenschaftlicher Texte formuliert werden mussten. Im Bereich Grammatik mussten im offenen Format bspw. vorgegebene komplexe Satzbauten in grammatikalisch richtige bzw. einfachere umgeschrieben werden.

Im Format ‚offen kurz' musste bspw. in einen Lückentext ein fehlendes Wort hineingeschrieben oder eine fehlende Endung eingepasst werden (vgl. Eberle et al. 2008: 129).

Die Kodierung der offenen Antworten in Deutsch erfolgte jeweils durch zwei durch die Testentwicklerin geschulte Korrektorinnen. Die Intercoder-Reliabilitäts-Prüfung mit dem IRC-Koeffizienten von Früh (2004: 179) führte bei lediglich einem Item zu einem nur genügenden Wert und sonst zu guten bis hervorragenden Ergebnissen (Eberle et al. 2008: 142).

2.5 Die Methode

Der Sprachtest wurde in Papierform vorgegeben. In der ursprünglichen Langversion für EVAMAR II standen 45 Minuten für jeden Test zur Verfügung. Es wurde ein anerkanntes Multi-Matrix-Design verwendet, bei dem die Testhefte über gleiche Aufgabenblöcke miteinander verknüpft werden. Dadurch werden die Fähigkeitsparameter der Personen auf der Basis der Gesamtzahl des eingesetzten Itempools geschätzt. Ermittelt wurden Rasch-skalierte Personenfähigkeiten, die über die Gesamtstichprobe von EVAMAR auf jeweils eine Skala mit einem Mittelwert von $M = 500$ und einer Standardabweichung $SD = 100$ transformiert wurden. Berichtet wurden sowohl die Ergebnisse des Gesamttests Sprache als auch die Mittelwerte der Subskalen Allgemeines Leseverstehen, Detailliertes Leseverstehen und Wortschatz. Die sprachreflexiven Teilbereiche Grammatik und Orthografie konnten aus konstruktionstechnischen Gründen nur in der Deutschschweiz eingesetzt werden und wurden zu einer gemeinsamen Subskala zusammengezogen (Eberle et al 2008: 145).

Für Deutsch wurden 168 Test-Items in der Haupterhebung eingesetzt und fünf verschiedene Testhefte erstellt, wovon jede Person zwei bearbeitete. Positions- und Reihenfolgeeffekte wurden ausgeglichen (vgl. Eberle et al. 2008: 128). Die unterschiedlichen Testhefte wurden zufällig auf die Maturandinnen und Maturanden verteilt, so dass jede Aufgabe jeweils von ähnlich vielen Zielpersonen bearbeitet wurde (Eberle et al. 2008: 120). Die Erhebungen wurden in einem für alle Schulen vergleichbaren Zeitraum von Ende April bis Anfang Juli 2007 durchgeführt (maximal drei Wochen vor Ende des regulären Unterrichts vor den Maturaprüfungen). Die Testdurchführung fand durch die Schulen selbst auf der Grundlage einer genauen Anleitung statt.

2.6 Die Stichprobe

Die GymnasiastInnen der Schweiz, die im Sommer 2007 die Maturitätsprüfungen ablegten, bildeten die Zielpopulation[15], auf deren Grundlage die Auswahl der Stichprobe erfolgte (Eberle et al. 2008: 373). Mittels eines einstufigen Verfahrens wurden proportional zu ihren Größen 260 Klassen ausgewählt, aus denen jeweils sämtliche SchülerInnen zur Teilnahme eingeladen wurden (vgl. Cochran 1977: 150; Kish, 1965: 182; Lehtonen & Pahkinen, 1995: 7; [single-stage cluster sampling with probability proportional to size] zitiert nach Eberle et al. 2008: 141). Erfreulich hoch lag die Rücklaufquote bei 91% auf Klassenebene bzw. 85% auf Personenebene (siehe Tabelle 1), so dass von rund 3800 MaturandInnen auswertbare Daten vorhanden sind (Eberle et al. 2008: 143).

Tab. 1: Anzahl teilnehmende Klassen sowie MaturandInnen nach Regionen bzw. Straten; mit Rücklaufquoten

Stratum	Stichprobe Klassen	Teilnahme	Rücklauf-quote Klassen	Stichprobe Personen	Teilnahme	Rücklauf-quote Personen
1 Zürich	75	67	89%	1439	1204	84%
2 Deutsch-schweiz MD3[16]	30	27	90%	587	496	84%
3 Deutsch-schweiz klein[17]	30	25	83%	578	459	79%
4 Deutsch-schweiz groß[18]	30	29	97%	590	560	95%
5 Romandie I (MD3)	30	27	90%	636	509	80%
6 Romandie II	30	30	100%	612	545	89%
Total	225	205	91%	4442	3773	85%

(Eberle et al. 2008: 142, hier ohne Tessin, ‚Total' angepasst)

Die Vergleiche erfolgten nach Regionen bzw. Straten, wobei ‚kleine' Kantone mit wenigen Maturaklassen bzw. ‚große' Kantone mit mehr als 15 Maturaklassen zusammengefasst wurden, andererseits zwei Straten diejenigen Gymnasialklassen zweier Regionen umfassen, in welchen nur 3 Mindest-Jahre am Gymnasium besucht wurden (Eberle et al. 2008: 376).

2.7 Ergebnisse Deutsch

Für die Aufgaben des Sprachtests konnte gezeigt werden, dass die Schweizer MaturandInnen sie im Mittel zu etwas mehr als die Hälfte richtig lösen konnten. Im Bereich Grammatik und Orthographie konnten im Mittel deutlich mehr Aufgaben als die Hälfte richtig bewältigt werden, wobei hier die maximalen Punktzahlen nicht erreicht wurden. Damit lagen die ermittelten Personenfähigkeiten im Durchschnitt auf einem Anforderungsniveau von Aufgaben oberhalb des mittleren Schwierigkeitsgrades, wobei die Streuung sowohl auf der Klassenebene als auch der Personen „beachtlich" war (Eberle et al. 2008: 374).

[15] Ausnahme: Der Kanton Basel-Landschaft war aufgrund von späteren Prüfungen nicht in der Grundgesamtheit enthalten. Der Kanton Genf verzichtete auf eine Teilnahme an der Studie (Eberle et al. 2008: 373).
[16] MD3: nur 3 Mindest-Jahre am Gymnasium nötig (erstes Oberstufen-Jahr an Sek II-Schule möglich): BE
[17] Kleine Kantone (Klassenzahl <= 15): AI, NW, OW, GL, UR, AR, SH, VD, ZG
[18] Große Kantone (Klassenzahl > 15): SZ, SO, TG, GR, BS, AG, SG, LU

3 Empirischer Teil

3.1 Analysen und Forschungsfrage

Vor dem Hintergrund einer im Sinne dieses Sammelbandes als plurizentrisch begriffenen Sprachlandschaft des Deutschen möchten wir mittels des beschriebenen Tests und dem Datensatz einer der größten schweizerischen Erhebungen an der Schnittstelle von Gymnasium – Universität folgende Frage empirisch prüfen: Zeigen sich Unterschiede in den Leistungstestergebnissen im Bereich Sprache in Abhängigkeit davon, ob Schülerinnen und Schüler in einer Deutschschweizer Familie mit einem Schweizer Sprachhintergrund mit den Sprachformen (dialektales) Schweizerdeutsch und Schweizer Standarddeutsch oder einem bundesdeutschen familiären Sprachkontext aufgewachsen sind? Diese Forschungsfrage begreifen wir als Annäherung an die Thematik der Plurizentrik des Deutschen, im Bewusstsein, dass wir aufgrund des Hineinspielens von dialektalen Ausprägungen sowohl des Schweizerdeutschen als auch des Bundesdeutschen sprachwissenschaftlich gesehen keine trennscharfen Aussagen bezüglich des Effektes des Schweizer Standarddeutschen versus des Bundesdeutschen Standarddeutschen machen können.

Die Hypothese zur genannten Forschungsfrage könnte lauten, dass SchülerInnen, welche im deutschsprachigen Teil der Schweiz meist über die Familiensprache in einem der schweizerdeutschen Dialekte (dem Schweizerdeutschen) und danach in der Schule[19] in der Schweizer Standardsprache sozialisiert werden, gegenüber SchülerInnen, welche in Deutschland oder mit einem oder zwei bundesdeutschen Elternteilen sozialisiert und damit vermutlich näher der bundesdeutschen Standardsprache/ der deutschen Hochsprache sozialisiert sind, benachteiligt sind, weil sie durch ihre mündliche dialektale Familiensprache in einer Form sozialisiert werden, die nicht der universitären Sprachnorm entspricht. Die Vermutung wäre also, dass bei einer Kompetenzmessung mittels eines Sprachtests Deutsch, der wie beschrieben bei universitären Fachtextsorten[20] sowohl das allgemeine Textverständnis als auch das detaillierte Leseverständnis, sowie Wortschatz, Grammatik und Orthografie prüft, die Leistungsergebnisse von im Schweizerdeutschen sozialisierten Gymnasiastinnen und Gymnasiasten niedriger ausfallen als die ihrer Hochdeutsch nahen Schulkameradinnen und -kameraden. Da sowohl der Wortschatz- als auch der Grammatik- und Orthographieteil im EVAMAR-Sprachtest sehr produktionsnahe konstruiert sind, könnten sich vielleicht in diesen Subskalen vermutete Auffälligkeiten in Richtung der Plurizentrik-Debatte zeigen.

3.2 Methode

In den Befragungsdaten von EVAMAR stehen keine direkten Variablen zur Fragestellung der Sozialisierung in Schweizer Standardsprache oder in bundesdeutscher Standardsprache zur Verfügung. Deshalb müssen, um die Fragestellung annähernd operationalisieren zu können, Hilfsvariablen herangezogen werden, welche sich auf die vorhandenen Befragungsdaten stützen. Es können Angaben zur Familiensprache kombiniert werden mit Auskünften zu den Geburtsländern der Familienmitglieder, um möglichst homogene

[19] Versuche mit „Hochdeutsch im Kindergarten“ sind in der Zwischenzeit erfolgreich (vgl. Landert Born 2011), jedoch für unsere untersuchten Maturajahrgänge noch nicht relevant.

[20] Es wird nicht davon ausgegangen, dass die Fachtexte, die an schweizerischen Universitäten verwendet und für EVAMAR II untersucht wurden, sich von Fachtexten an deutschen Universitäten stark unterscheiden. Bei deren Inhaltsanalyse (Teilprojekt A1) wurden zwar weder bei der Feinrasterung noch bei der Grobkodierung auffällige Teutonismen bzw. Helvetismen *explizit* erfasst, jedoch wären solche Sinneinheiten über die Wissenskategorie AA „Wissen über Terminologien“ als vorausgesetztes Eingangswissen kodiert worden (Eberle et al. 2008: 37-41).

Subgruppen zu bilden, welche sich in ihrem Sprachgebrauch unterscheiden. Das Verfahren wird kurz beschrieben und anschließend kritisch diskutiert.

Die Familiensprache muss mittels zwei Variablen erschlossen werden aus Selbstauskünften zu a) der Sprache mit der Mutter und b) der Sprache mit dem Vater. In der Erhebung wurde im Fragebogen gefragt, „welche Sprache [man] überwiegend … mit der Mutter“ bzw. „… mit dem Vater“ spreche (siehe Abbildung 5). Es standen auf einer Likert-Skala von 1-4 folgende vier Antwortkategorien zur Verfügung: (1) „nur Deutsch/ Schweizerdeutsch“; (2) „meistens Deutsch/ Schweizerdeutsch, aber manchmal auch eine andere Sprache“; (3) „meistens eine andere Sprache, aber manchmal auch Deutsch/ Schweizerdeutsch“; (4) „nur eine andere Sprache“.

Abb. 5 Fragebogenitems zur Sprachverwendung

4. Welche Sprache sprechen Sie überwiegend...?

Falls Sie nur noch ein Elternteil haben, beantworten Sie die Frage nur für diese Person.

		... mit Ihrer Mutter	... mit Ihrem Vater	... mit Ihren Freunden
a)	nur Deutsch/Schweizerdeutsch	❑	❑	❑
b)	meistens Deutsch/Schweizerdeutsch, aber manchmal auch eine andere Sprache	❑	❑	❑
c)	meistens in einer anderen Sprache, aber manchmal auch Deutsch/Schweizerdeutsch	❑	❑	❑
d)	nur eine andere Sprache	❑	❑	❑

Bei diesen Sprachverwendungsitems muss in Bezug auf ihre Verwendung in der sprachwissenschaftlichen Plurizentrik-Debatte an dieser Stelle kritisch angemerkt werden, dass in der Befragung (leider) nicht fein differenziert wurde zwischen Deutsch im Sinne eines bundesdeutschen Standarddeutsch oder schweizerischen Standarddeutsch versus Schweizerdeutsch im Sinne von regionalen alemannischen Dialekten innerhalb des Gebietes der Deutschschweiz.

Dadurch wird bei einer gewünschten Stichprobenaufteilung in unterschiedliche Sprachgruppen die Trennung allein aufgrund dieser Items bzw. Variablen nicht möglich.

Für die Gruppeneinteilung in Schweizerdeutsch vs. Standarddeutsch sprechende bzw. gemischtsprachige Familien wurde somit zusätzlich die Angabe der Geburtsländer der Familienmitglieder hinzugezogen: Auf die Fragen „Wo wurden Sie geboren? Wo wurden Ihre Eltern geboren?“ konnte mit den beiden Antwortkategorien „in der Schweiz“ oder „in einem anderen Land als in der Schweiz“ geantwortet werden (siehe Abbildung 6).

Abb. 6 Fragebogenitems zum Geburtsland

2. Wo wurden Sie geboren? Wo wurden Ihre Eltern geboren?

Bitte machen Sie in jeder Spalte ein Kreuz!

		Sie	Mutter	Vater
a)	in der Schweiz	❑	❑	❑
b)	in einem anderen Land als in der Schweiz	❑	❑	❑

Mit diesem Umweg über die Abfrage nach den Herkunftsländern der Familienmitglieder kann ermittelt werden, welche Personen mit welchen Eltern in der Schweiz geboren und

vermutlich in der Schweiz aufgewachsen sind, und somit vermutlich Schweizerdeutsch (im Sinne von regionalen mündlichen Dialekten) sprechen, und andererseits welche Personen mit welchen Eltern nicht in der Schweiz geboren/aufgewachsen sind, die dann vermutlich eher Deutsch im Sinne von Hochdeutsch (bundesdeutsche Standardsprache) verwenden.

Aufgrund der Kombination der beschriebenen Variablen Sprache und Geburtsland wurden drei Gruppen mit unterschiedlichen Eltern-Kind-Konstellationen im Hinblick auf den familiären Sprachgebrauch Schweizerdeutsch vs. Hochdeutsch gebildet.

Die erste Gruppe umfasst die „nur Schweizerdeutsch sprechenden" Familien, die zweite Gruppe Maturandinnen und Maturanden mit einem schweizerdeutschen und einem hochdeutschsprachigen Elternteil, die dritte Gruppe jene mit zwei hochdeutschsprachigen Elternteilen. Betrachtet wurden nur Maturandinnen und Maturanden aus den deutschsprachigen Kantonen (N = 2719 Versuchspersonen), für die Angaben zu beiden Elternteilen und gleichzeitig Testergebnisse vorliegen (N = 2351).

Zur ersten Gruppe der „Schweizerdeutschen Familien" gehören demnach Maturandinnen und Maturanden, die selbst in der Schweiz geboren sind, deren Elternteile beide in der Schweiz geboren sind, und die mit beiden Elternteilen „nur Deutsch/ Schweizerdeutsch" reden, unter der begründeten Annahme, dass es sich um alemannisches Schweizerdeutsch (Dialekt) handelt. Dieser deutlich größten Gruppe gehören insgesamt n = 1527 Maturandinnen und Maturanden an. In der zweiten Gruppe befinden sich Maturandinnen und Maturanden, deren eines der beiden Elternteile im Ausland geboren ist und die sich mit beiden Elternteilen „nur auf Deutsch/ Schweizerdeutsch" unterhalten (n = 188). Die im Vergleich mit den beiden anderen Gruppen sehr kleine dritte Gruppe beinhaltet Schülerinnen und Schüler, deren beide Elternteile im Ausland geboren sind und bei denen die Kommunikation mit Mutter und Vater „nur auf Deutsch/ Schweizerdeutsch" erfolgt (n = 44). Hier bewegt sich die Annahme in der Richtung, dass in der Familie Hochdeutsch/ Bundesdeutsche Standardsprache gesprochen wird. Maturandinnen und Maturanden, die mit ihren Eltern nicht „nur auf Deutsch/ Schweizerdeutsch" (Item 4a, siehe Abbildung5) sprechen, sondern „manchmal" bis „nur" eine andere Sprache (Items 4b-d) verwenden, wurden in den Analysen nicht berücksichtigt (n = 592), da es nur um den Vergleich des Schweizerdeutschen mit einem hochdeutschen Sprachhintergrund der Schülerinnen und Schüler gehen soll, aber nicht um den Vergleich mit Deutsch als Fremdsprache.

Kritisch anzumerken ist, dass dieser Strategie des Auseinanderdividierens der Sprachgruppen einige Annahmen zugrunde liegen und dass Ausreißer[21] vermutlich nicht vermieden werden können. So müsste beispielsweise die Annahme, dass in der Schweiz geborene Personen, welche mit ihren ebenfalls in der Schweiz geborenen Eltern „nur Deutsch/ Schweizerdeutsch" sprechen, vermutlich „Einheimische" und nicht MigrantInnen sind und somit vermutlich einen Schweizerdeutschen Dialekt sprechen, bis mindestens in die dritte Generation geprüft werden (für die Sprachvariable im Zusammenhang mit der Bestimmung von Migrationsgruppen vgl. bspw. NEPS, Kristen, Olczyk & Will 2016). Speziell für das Thema der verschiedenen Plurizentren des Deutschen kann hinsichtlich der Kombination von „nur Deutsch/ Schweizerdeutsch in der Familie" mit dem Geburtsland Nicht-Schweiz (leider) nicht eindeutig bestimmt werden, ob es sich in diesem Fall um bundesdeutsches Standarddeutsch oder um österreichisches Standarddeutsch handelt.

[21] Beispielsweise könnte die Mutter aus einem asiatischen Land stammen, der Vater aber vielleicht aus einem osteuropäischen Land und die nun in der Schweiz lebende Familie hätte sich auf Deutsch als (einzige) Familiensprache geeinigt. Leider konnte die Geburtslandfrage aus Ressourcengründen nicht offen abgefragt werden, so dass über das Herkunftsland nicht Eineindeutigkeit besteht.

Zur Prüfung der oben beschriebenen inhaltlichen Hypothese wurden univariate Varianzanalysen mit dem Faktor „Gruppenzugehörigkeit“ und den abhängigen Variablen der Leistungstestergebnisse durchgeführt. Anschließend wurde zur Kontrolle von Drittvariablen die „berufliche Stellung der Eltern“ als zusätzlicher Faktor in die Varianzanalysen mitaufgenommen.

3.3 Ergebnisse

In Tabelle 2 sind die Mittelwerte, Standardabweichungen und Standardfehler der einzelnen Testergebnisse (in Form der Rasch-skalierten Personenfähigkeiten) für die verschiedenen Gruppen aufgeführt. Der Mittelwert der Personenfähigkeiten des Deutsch-Gesamttests der vorliegenden Stichprobe entspricht mit 501 Punkten gerade dem Mittelwert der gesamten EVAMAR II-Stichprobe für Erstsprache. Des Weiteren ist auch in der vorliegenden EVAMAR II-Substichprobe die bereits im theoretischen Teil beschriebene breite Streuung der Personenfähigkeiten erkennbar, die für den Gesamttest hier von 187 bis 780 Punkten reichen (siehe Tabelle 2).

Die nahe beieinanderliegenden Mittelwerte der Personenfähigkeiten der einzelnen Gruppen deuten bereits an, dass sich die Maturandinnen und Maturanden mit unterschiedlichen Sprachhintergründen in ihren Ergebnissen des Deutschtests nicht unterscheiden. Über alle drei Gruppen hinweg gesehen ergeben sich für die Ergebnisse des Gesamtsprachtests keine signifikanten Effekte der Gruppenzugehörigkeit (*F*-Wert (2, 1756) = .515, Irrtumswahrscheinlichkeit p = .598, Effektstärke partielles Eta-Quadrat η_p^2 = .001). Dies gilt ebenso für die Teilergebnisse der rezeptiven Fähigkeiten ‚Detail-Leseverständnis’ (F(2, 1756) = .465, p = .628, η_p^2 = .001) und ‚Allgemeines Leseverständnis’ (F(2,1756) = 1.165, p = .312, η_p^2 = .001) als auch für die produktionsnäheren Subskalen ‚Wortschatz’ (F(2, 1756) = .266, p = .767, η_p^2 = .000) und ‚Grammatik’ (F(2, 1756) = .028, p = .973, η_p^2 = .000). Für alle diese sprachlichen Subskalen zeigen sich ebenso wie beim Gesamttest keine signifikanten Effekte der Gruppenzugehörigkeit aufgrund der unterschiedlichen Familiensprachen (dialektales) Schweizerdeutsch versus Hochdeutsch bzw. gemischtsprachig. Darüber hinaus zeigen auch die einzeln durchgeführten Gruppenvergleiche (Post hoc-Tests nach Bonferroni zur Korrektur der Potenzierung des statistischen Alpha-Fehlers), dass sich die MaturandInnen der sprachlichen drei Gruppen untereinander statistisch nicht signifikant in ihren Testleistungen unterscheiden.

Lediglich für die Gruppe der MaturandInnen mit zwei bundesdeutschen Elternteilen erweist sich beim ‚Allgemeinen Leseverständnis’ (M = 518) der Unterschied zu den beiden anderen Gruppen (M = 501 bzw. 499) auf der Basis von zwei T-Tests zwischen Gruppe 1 bzw. 2 und 3 als praktisch bedeutsam (Effektgröße Cohens d = -0.23 bzw. -0.25) bei ansonsten aufgrund der Gruppengröße statistisch nicht signifikanten Werten (t(1569) = -1.489, p = .137 im Vergleich mit der Gruppe der Schweizerdeutschen; bzw. im Vergleich mit der gemischten Gruppe 2: t(230) = -1.475, p = .142).

Weitere Varianzanalysen zeigen jedoch, dass diese beiden Effekte kleiner Größenordnung auf Bildungsunterschiede zwischen den drei Gruppen zurückzuführen sind: Da vor allem in der Gruppe der Familien mit zwei bundesdeutschen Elternteilen überzufällig häufig Mütter und Väter mit Hochschulabschluss anzutreffen sind, wurde die berufliche Stellung der Eltern kontrolliert; unter Kontrolle dieser Drittvariablen gehen die Effektstärken des Faktors sprachliche Gruppenzugehörigkeit beider Gruppenunterschiede auf nicht mehr bedeutsame Niveaus zurück (von $\boldsymbol{F}$(1, 1569) = 2.218, p = .137, η_p^2 = .001 auf F(1, 1323) = .597, p = .440, η_p^2 = .000 im Vergleich mit der Gruppe der Schweizerdeutschen; bzw. im Vergleich mit Gruppe 2: von F(1, 230) = 2.175, p = .142, η_p^2 = .009 auf F(1, 164) = .457, p = .500, η_p^2 =

.003). Maturandinnen und Maturanden mit zwei vermuteten bundesdeutschen Elternteilen weisen demzufolge ein schwach bedeutsam besseres allgemeines Leseverständnis als ihre schweizerdeutschen SchulkollegInnen der anderen beiden Gruppen auf, da ihre Eltern über ein besonders hohes Bildungsniveau verfügen.

Tab. 2: Leistungstestergebnisse über die drei Sprach-Gruppen

		N	Mittel-wert	Standard-abweichung	Mini-mum	Maxi-mum	Standard fehler
Gesamtergebnis Erstsprache	beide Eltern Schweizer-deutsch	1527	501	83	187	780	2.13
	ein Elternteil Hochdeutsch	188	495	81	246	663	5.93
	beide Eltern Hochdeutsch	44	503	94	224	676	14.10
	Gesamt	1759	501	83	187	780	1.99
Ergebnisse Wortschatz	beide Eltern Schweizer-deutsch	1527	497	72	237	735	1.84
	ein Elternteil Hochdeutsch	188	493	70	312	713	5.08
	beide Eltern Hochdeutsch	44	496	83	349	715	12.51
	Gesamt	1759	496	72	237	735	1.72
Ergebnisse Grammatik und Orthographie	beide Eltern Schweizer-deutsch	1527	503	76	192	861	1.94
	ein Elternteil Hochdeutsch	188	502	79	201	695	5.76
	beide Eltern Hochdeutsch	44	501	79	282	683	11.98
	Gesamt	1759	502	76	192	861	1.82
Ergebnisse Detail-Leseverständnis	beide Eltern Schweizer-deutsch	1527	505	82	163	729	2.11
	ein Elternteil Hochdeutsch	188	500	79	209	675	5.75
	beide Eltern Hochdeutsch	44	498	109	195	701	16.40
	Gesamt	1759	504	83	163	729	1.97
Ergebnisse Allgemeines Leseverständnis	beide Eltern Schweizer-deutsch	1527	501	76	185	700	1.94
	ein Elternteil Hochdeutsch	188	499	76	284	704	5.57
	beide Eltern Hochdeutsch	44	518	75	321	747	11.28
	Gesamt	1759	501	76	185	747	1.81

Für alle anderen Gruppenunterschiedsanalysen ergaben sich bei einer Kontrolle der beruflichen Stellung der beiden Elternteile auch weiterhin keine Effekte der Gruppenzugehörigkeiten auf die Ergebnisse des Deutschleistungstests (ohne Tabelle). Unter der Annahme, dass die Gruppenzugehörigkeit die Sprachgewohnheiten Schweizerdeutsch (Dialekt) vs. Hochdeutsch/ Standarddeutsch in den Familien adäquat widerspiegeln, spielt es unseren Ergebnissen zufolge für das Abschneiden der Maturandinnen und Maturanden beim Deutsch-Leistungstest keine Rolle, ob in den Familien überwiegend Hochdeutsch/ Standarddeutsch oder Schweizerdeutsch gesprochen wird.

4 Diskussion

Auf dem Hintergrund der bereits im Methodenteil geäußerten kritischen Anmerkungen zu unserer Differenzierung der Sprachgruppen anhand der im Datensatz von EVAMAR II vorhandenen Sprach- und Geburtslandvariablen wollen wir unsere Ergebnisse einer versuchsweisen Analyse von Sprachstandsunterschieden von Maturandinnen und Maturanden im EVAMAR-Sprachtest Deutsch und deren empirische Bedeutung für die sprachwissenschaftliche Plurizentrik-Debatte im europäischen Raum im Folgenden kurz diskutieren.

Unsere Ergebnisse zum Vergleich der drei angenommenen Schülergruppen „Schweizerdeutsch(Dialekt-)Sprechende“ vs. „Hochdeutsch(Standarddeutsch)-Sprechende“ vs. „Schülerinnen und Schüler mit gemischter Familiensprache Hochdeutsch/Schweizerdeutsch“, zeigen keine erwarteten signifikanten Unterschiede in den Leistungen im Deutschtest. Unter dem Vorbehalt von nicht eineindeutiger Identifizierung von Gruppenangehörigen[22] konnte empirisch gezeigt werden, dass innerhalb des Plurizentrums des Schweizer Standarddeutschen sozialisierte Schülerinnen und Schüler mit der Muttersprache Schweizerdeutsch gegenüber den (vermutlich) näher am Hochdeutschen sozialisierten Maturandinnen und Maturanden[23] in einem auf universitärer Sprachnorm ausgerichteten Deutsch-Kompetenztest weder schlechtere noch bessere Ergebnisse zeigten.

Darüber hinaus ist insbesondere die Gruppe der Familien mit zwei (mutmaßlich) hochdeutschen Elternteilen sehr klein, die Gruppengrößen sind sehr unterschiedlich. Dies birgt die Gefahr, dass die sehr große Vergleichsgruppe der Schweizerdeutschen sehr heterogen sein könnte und damit nicht kontrollierte Drittvariablen das Ergebnis verzerren.

Auch in den die Plurizentrik-Debatte interessierenden sprachproduktionsnahen Subskalen Grammatik und Orthografie sowie Wortschatz konnten – vorbehältlich der erwähnten Einschränkungen – keine empirisch signifikanten Leistungsunterschiede zwischen den drei Sprachgruppen nachgewiesen werden.

Offensichtlich gelingt es entweder den Gymnasien, die Norm der deutschen Standardsprache für alle gleichermaßen als Orientierungspunkt gut zu vermitteln, so dass die sprachliche familiäre (dialektale) Alltagsgewohnheit weder eine hinderliche noch eine

[22] Einschränkend zur vorliegenden Studie muss gesagt werden, dass bei der Erfragung der Sprachverwendung in den Familien nicht direkt zwischen Schweizerdeutsch und Hochdeutsch unterschieden wurde (siehe Abbildung 5). Die Unterscheidung der Gruppen erfolgte daher auf Basis der Annahme, dass MaturandInnen, deren Eltern im Ausland geboren sind und mit denen „nur Deutsch/ Schweizerdeutsch“ gesprochen wurde, tatsächlich einerseits aus Deutschland stammen und zum zweiten Hochdeutsch verwenden. In Deutschland findet sich jedoch auch heute noch eine Vielfalt an Dialekten in Gebrauch (vgl. bspw. die empirische Arbeit von Huesmann 1998), wobei vermutet werden kann, dass die Familien mit Hochschulabschluss sich eher um eine standardorientierte Familiensprache bemühen (Koller 1992).

[23] Dabei konnte der Einfluss bundesdeutscher Dialekte nicht nachverfolgt oder eindeutig ausgeschlossen werden.

förderliche Rolle zu spielen scheint, oder andererseits gelingt es den im dialektalen Schweizerdeutsch aufgewachsenen Jugendlichen insgesamt als Gruppe, ohne Leistungseinbußen gegenüber anderen Sprachgruppen, die Sprachanforderungen des universitären Standarddeutschen zu erfüllen.

Empirisch zu überprüfen wäre, ob dieses Resultat erst in den späteren Jahren der schulischen Sozialisation auftritt, also vielleicht auch eine Frage der Schuldauer oder der Schulform ist, so dass sich in unteren Klassen vielleicht noch größere Unterschiede in der Sprachleistung als zum Zeitpunkt des Übergangs an die Hochschulen zeigen. Hinweise darauf geben bspw. qualitative Ergebnisse von Koller (1992), dass die zuhause Hochdeutsch sprechenden Jugendlichen bundesdeutscher Familien nach eigener Einschätzung teilweise besser die Standardsprache verwenden als ihre schweizerdeutschen SchulkollegInnen und ihnen ihre familiäre Sprachsituation vor allem in der Primarstufe gewisse Vorteile verschafft habe (317-318).

Auch müsste vermutlich der Einfluss des Mediums Fernsehens neben der familiären Sprachsituation darüber hinaus noch empirisch miteinbezogen werden (vgl. Landert Born, 2011: 186; Häcki Buhofer/ Burger 1998); diese Daten standen uns bei EVAMAR II retrospektiv nicht zur Verfügung.

Wir vermuten insgesamt, dass die hohen Hürden in der Schweiz, an ein Gymnasium aufgenommen zu werden (schweizweit liegt die Maturitätsquote bei lediglich ca. 20%), auch im Hinblick auf das Vorliegen der sprachlichen Fähigkeiten bereits sehr selektiv wirken, so dass sich durch diese Selektion in der Schulform bereits keine Unterschiede mehr zwischen den Gruppen zeigen. Interessant wäre es daher, auch an nicht-gymnasialen Schulen ähnliche Gruppenuntersuchungen vorzunehmen, um diese Vermutung ausschließen zu können.

Das Resultat nicht vorhandener Unterschiede in den Sprachleistungen der angenommenen Sprachgruppen könnte man als weiteres Zeichen für eine gelingende Vorbereitung der Gymnasien auf die Hochschulen bezeichnen, wie sie die EVAMAR II-Analysen insgesamt zum Ergebnis haben. Dennoch bleibt der grundsätzlich auch hier gezeigte Befund der großen Unterschiede in den Leistungen zwischen den Personen und des Defizits am unteren Ende der Leistungsskala. Dieses Ergebnis sprachlicher Lücken bestätigt sich ebenfalls für die AbgängerInnen der schweizerischen Berufsmaturitätsschulen (vgl. Studie OEKOMA, u.a. Eberle/ Schumann 2013). Die Befunde der vorliegenden Studie bestätigen im Kleinen diejenige der großen EVAMAR-Studie, dass die Bemühungen zur Verbesserung der sprachlichen Fähigkeiten wie die aktuell in der Schweiz auf bildungspolitischer Ebene unternommenen Anstrengungen zur besseren Förderung der Sprachkompetenzen im Gymnasium weiter vorangetrieben werden sollten. Diese Förderung soll nicht zulasten anderer Fachinhalte gehen, sondern mittels rechtzeitiger, ergänzender und individueller Förderung der in diesen Bereichen schlechten Schülerinnen und Schüler (Eberle et al. 2015).

Bezüglich der eingangs gestellten Fragestellung, ob mithilfe der EVAMAR II-Daten es möglich ist, empirisch gesicherte Antworten für das Feld der sprachwissenschaftlichen Plurizentrik-Debatte zu liefern, müssen wir nach eingehender Beschäftigung mit dem Datenmaterial uns einerseits mit dem Verweis zufriedengeben, dass eine exakte Aufteilung nach plurizentrisch definierten Sprachgruppen innerhalb der deutschen Sprachverwendung nahezu unmöglich ist, dürfen aber andererseits darauf verweisen, dass unter bestmöglicher Verwendung des Datenmaterials sich im EVAMAR-Sprachtest auf universitärem Niveau keine signifikanten Leistungsunterschiede zwischen Gruppen von GymnasiastInnen mit unterschiedlichen Varietäten von deutscher Familiensprache zeigen, auch nicht in Subskalen, die der Sprachproduktion (und damit der Plurizentrik-Debatte) nahe sind. Mit diesem

eingeschränkten Befund können wir nur darauf hoffen, dass in nächster Zeit es mit sprachwissenschaftlich ausgerichteten empirischen Studien möglich sein wird, ein feineres Ergebnisbild für die Plurizentrik-Thematik zu zeichnen.

Literaturverzeichnis

Ammon, Ulrich (1995): *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten.* Berlin, New York: Walter de Gruyter.

Berthele, Raphael (2004): Vor lauter Linguisten die Sprache nicht mehr sehen - Diglossie und Ideologie in der deutschsprachigen Schweiz. In: Christen, Helen (Hrsg.): *Dialekt, Regiolekt und Standardsprache im sozialen und zeitlichen Raum.* Beiträge zum 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen. Marburg/Lahn. Wien: Edition Praesens, 111-136.

Christen, Helen (2004): Vorwort: Vom Wissen um die soziale Komponente arealer Sprachvariation zu ihrer Erforschung. In: Christen, Helen (Hrsg.): *Dialekt, Regiolekt und Standardsprache im sozialen und zeitlichen Raum.* Beiträge zum 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen. Marburg/Lahn. Wien: Edition Praesens, 7-20.

Eberle, Franz, Karin Gehrer, Beat Jaggi, Johannes Kottonau, Maren Oepke & Michael Pflüger (2008): *Evaluation der Maturitätsreform 1995 (EVAMAR). Phase II.* Bern: Staatsekretariat für Bildung und Forschung. http://edudoc.ch/record/29677/files/Web_Evamar-Komplett.pdf (11.08.2016)

Eberle, Franz & Stephan Schumann (2013): Ökonomische und weitere Kompetenzen von Deutschschweizer Berufsmaturanden und Gymnasiasten im Vergleich. In: *Gymnasium Helveticum* 3, 18-21.

Eberle, Franz, Christel Brüggenbrock, Christian Rüede, Christof Weber & Urs Albrecht (2015): *Basale fachliche Kompetenzen für allgemeine Studierfähigkeit in Mathematik und Erstsprache*. Schlussbericht zu Handen der EDK. Universität Zürich: Institut für Erziehungswissenschaft. http://www.ife.uzh.ch/research/lehrstuhleberle/forschung/bfkfas/downloads/Schlussbericht_final_V7.pdf (11.06.2015).

Europarat - Rat für kulturelle Zusammenarbeit (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt. http://www.goethe.de/z/50/commeuro/i0.htm (01.06.2015).

Früh, Werner (2004): *Inhaltsanalyse. Theorie und Praxis.* Stuttgart: UTB.

Gehrer, Karin & Cordula Artelt (2013): Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. In: Rosebrock, Cornelia & Andrea Bertschi-Kaufmann (Hrsg.): *Literalität erfassen: bildungspolitisch, kulturell, individuell.* Weinheim: Beltz Juventa, 168-187.

Gehrer, Karin, Stefan Zimmermann, Cordula Artelt & Sabine Weinert (2013): NEPS Framework for Assessing Reading Competence and Results From an Adult Pilot Study. In: Artelt, Cordula, Sabine Weinert & Claus H. Carstensen (Hrsg.): *Competence Assessment within the NEPS.* JERO Journal for Educational Research Online/ Journal für Bildungsforschung Online, 5 (2): Waxmann, 50-79.

Haas, Walter (2004): Die Sprachsituation der deutschen Schweiz und das Konzept der Diglossie. In: Christen, Helen (Hrsg.): *Dialekt, Regiolekt und Standardsprache im sozialen und zeitlichen Raum.* Beiträge zum 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen. Marburg/Lahn. Wien: Edition Praesens, 81-110.

Häcki Buhofer, Annelies (Hrsg.) (2000): *Vom Umgang mit sprachlicher Variation*. Tübingen; Basel: A. Francke.

Häcki Buhofer, Annelies & Harald Burger (1998): *Wie Deutschschweizer Kinder Hochdeutsch lernen. Der ungesteuerte Erwerb des gesprochenen Hochdeutschen durch Deutschschweizer Kinder zwischen sechs und acht Jahren*. Stuttgart: Steiner.

Huesmann, Annette (1998): *Zwischen Dialekt und Standard: Empirische Untersuchung zur Soziolinguistik des Varietätenspektrums in Deutschland*. Tübingen: Max Niemeyer.

Kish, Leslie (1965): *Survey Sampling.* New York: John Wiley & Sons.

Koller, Werner (1992): *Deutsche in der Deutschschweiz. Eine sprachsoziologische Untersuchung*. Aarau: Sauerländer.

Koller, Werner (1999): Nationale Sprach(en)kultur der Schweiz und die Frage der "nationalen Varietäten des Deutschen" (1999) In: Gardt, Andreas, Haß-Zumkehr, Ulrike, Roelcke, Thorsten (Hrsg.): *Sprachgeschichte als Kulturgeschichte*. Berlin/New York: De Gruyter, 133-170. Auch online http://wernerkoller.com/wissenschaftliches.html (09.08.2016).

Kristen, Cornelia, Melanie Olczyk & Gisela Will (2016): Identifying Immigrants and Their Descendants in the National Educational Panel Study. In: Blossfeld, Hans-Peter, Jutta von Maurice, Michael Bayer & Jan Skopek (Hrsg): *Methodological Issues of Longitudinal Surveys. The Example of the National Educational Panel Study.* Wiesbaden: Springer Fachmedien, 195-213.

Landert Born, Karin (2011): Hochdeutsch im Kindergarten: Was meinen die Schweizer Kinder dazu? In: Vogt, Franziska, Miriam Leuchter, Annette Tettenborn, Ursula Hottinger, Marianna Jäger & Evelyne Wannack (Hrsg.): *Entwicklung und Lernen junger Kinder.* Münster: Waxmann, 185-195.

Lehtonen, Risto & Erkki Pahkinen (1994): *Practical Methods for Design and Analysis of Complex Surveys.* New York: John Wiley & Sons.

Neuland, Eva (2004): Sprachvariation im Fokus von Sprachunterricht. Zur Einführung in das Themenheft. In: *Der Deutschunterricht* 56 (1: Sprachvariation im heutigen Deutsch), 2-7.

Oepke, Maren & Franz Eberle (2010): *Studierfähigkeit von Maturandinnen und Maturanden. Eine Follow-up-Studie zur EVAMAR II-Untersuchung. Antrag an den SNF vom 01.10.2010*. Universität Zürich: Institut für Erziehungswissenschaft.

Oepke, Maren & Franz Eberle (2014): Studierfähigkeit und Studienfachwahl von Maturandinnen und Maturanden. In: Eberle, Franz, Barbara Schneider-Taylor & Dorit Bosse (Hrsg.): *Abitur und Matura zwischen Hochschulvorbereitung und Berufsorientierung*. Wiesbaden: Springer, 185-214.

Oepke, Maren & Franz Eberle (2016): Erstsprach- und Mathematikkompetenzen – wichtig für die (allgemeine) Studierfähigkeit? In Kramer, Jochen, Marko Neumann & Ulrich Trautwein (Hrsg.), *Abitur und Matura im Wandel. Historische Entwicklungslinien, aktuelle Reformen und ihre Effekte*. Wiesbaden: Springer Fachmedien, 215-252.

Ramseier, Erich, Jürgen Allraum, Ursula Stalder, François Grin, Roberta Alliata & Stephan Müller et al. (2004): *Evaluation der Maturitätsreform 1995 (EVAMAR). Neue Fächerstruktur – Pädagogische Ziele – Schulentwicklung. Schlussbericht zur Phase 1.* Bern: Schweizerische Konferenz der Erziehungsdirektoren und Bundesamt für Bildung und Wissenschaft.

Schneider, Hansjakob (1998): *„Hochdeutsch – das kann ich auch" Der Erwerb des Hochdeutschen in der deutschen Schweiz: eine Einzelfallstudie zur frühen mündlichen Sprachproduktion.* Bern: Lang.

Siebenhaar, Beat & Alfred Wyler (1997): *Dialekt und Hochsprache in der deutschsprachigen Schweiz*. Zürich: Pro Helvetia.

Sieber, Peter & Horst Sitta (1986): *Mundart und Standardsprache als Problem der Schule*. (Reihe Sprachlandschaft, Bd. 3) Aarau: Sauerländer.