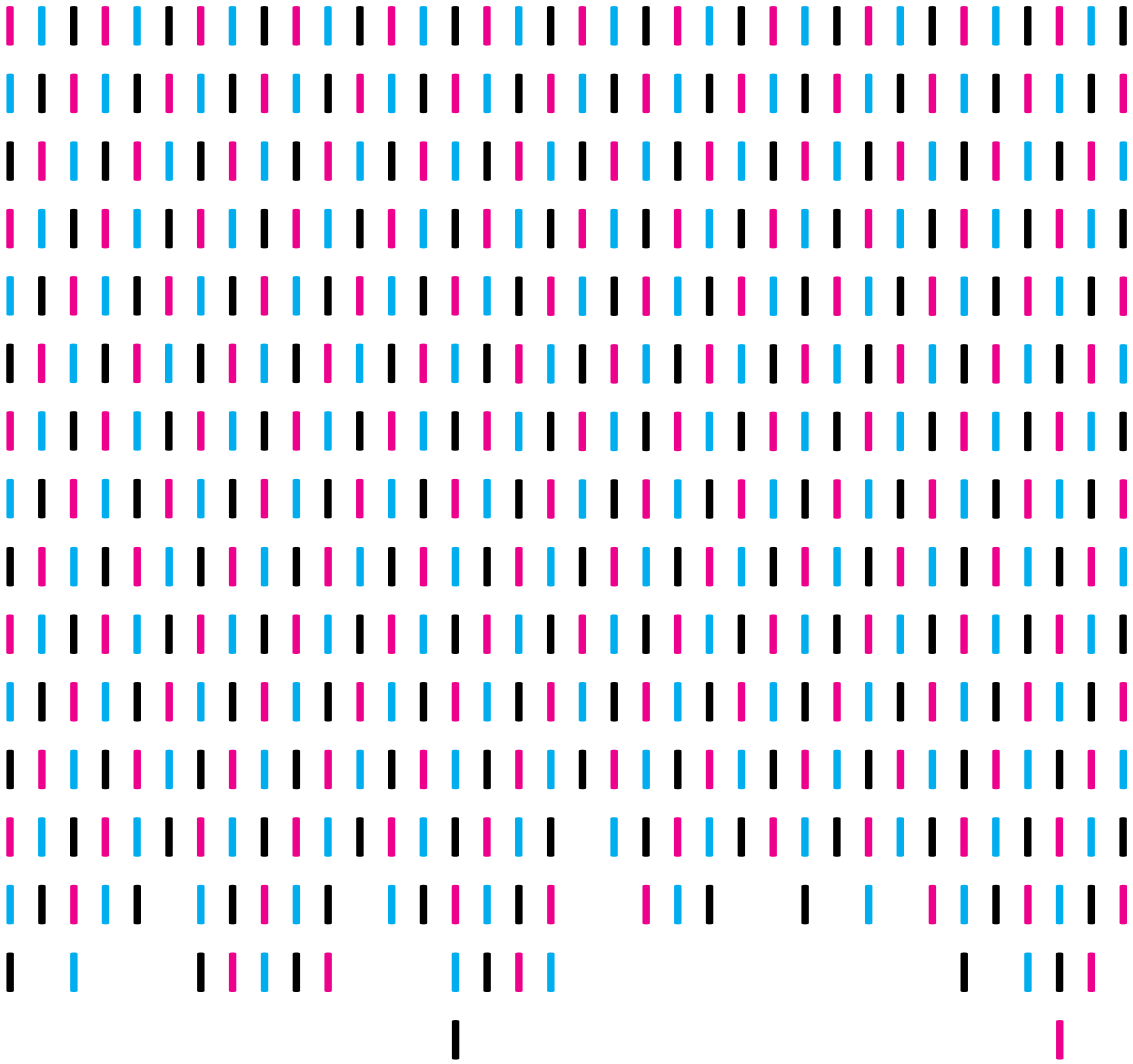


NEPS WORKING PAPERS

Karin Gehrer

DER EINFLUSS VON AUFGABEN- UND TEXTMERKMALEN AUF ITEM-SCHWIERIGKEIT

NEPS Working Paper No. 67
Bamberg, Januar 2017

LIfBi
LEIBNIZ-INSTITUT FÜR BILDUNGSVERLÄUFE e.V.

NEPS
National Educational Panel Study

Working Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).

The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at **www.neps-data.de** (see section "Publications").

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

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

Zusammenfassung ... 2
1. Einleitung ... 5
2. Theorie und Befundlage ... 6
2.1 Testen ohne wiederholte Texteinsicht ... 7
2.2 Was generiert Schwierigkeit? ... 9
2.2.1 Aufgabenmerkmale ... 10
2.2.2 Textmerkmale ... 13
2.2.3 Interaktion Items mit Text ... 14
3. Forschungsfragen der Analyse ... 16
4. Beschreibung der Studie ... 19
4.1 Instrument ... 19
4.2 Design der Studie mit Experimentalbedingung ... 21
4.3 Stichprobe ... 22
5. Methode ... 22
5.1 Klassifikationsbäume ... 23
5.2 Multiple lineare Regression ... 25
6. Ergebnisse ... 25
6.1 Deskriptiva ... 25
6.2 Schwierigkeitsgenerierende Merkmale ... 28
6.2.1 Mit dem Klassifikationsbaum ... 28
6.2.2 Mit Regression ... 29
7. Diskussion ... 31
Literatur ... 36

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.

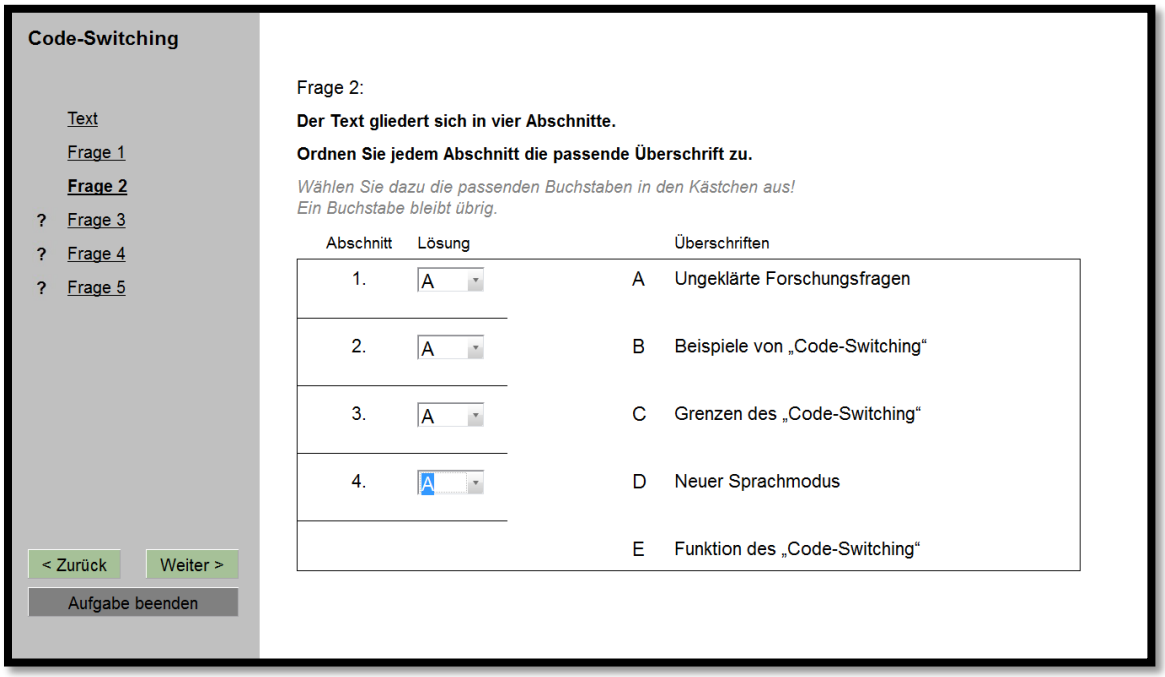
4. Beschreibung der Studie

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.

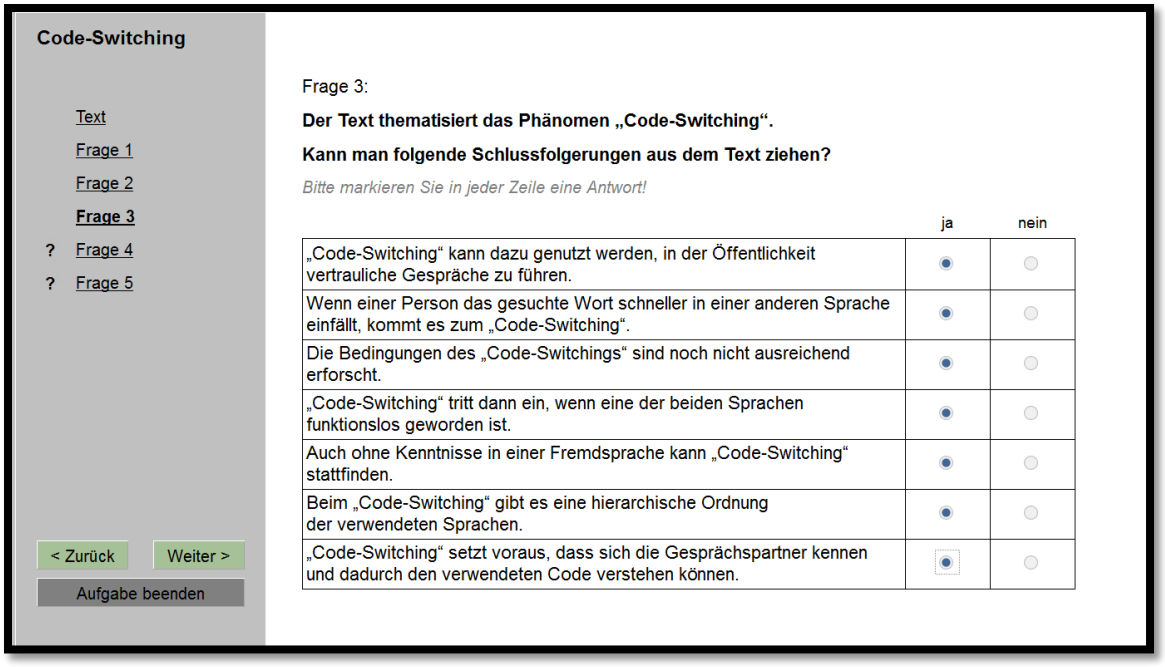


Abbildung 3: Beispiel des Antwortformates “Entscheidungstabelle“ aus dem Entwicklungspool 2014 der computerisierten NEPS-Leseitems für Erwachsene und Studierende

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.

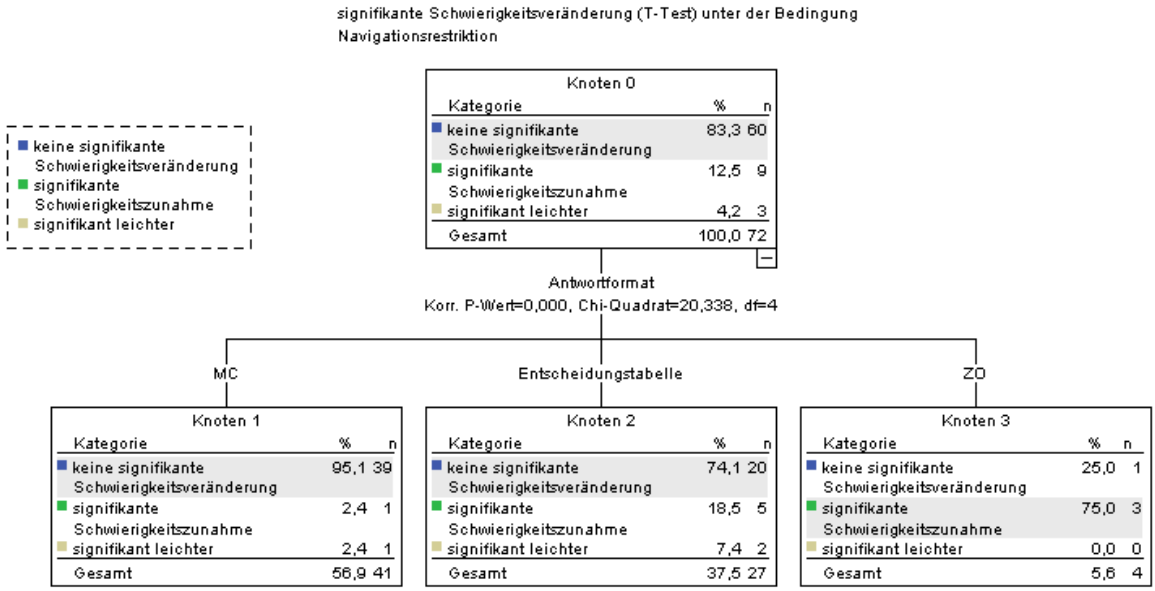


Abbildung 2: Baummodell über alle Personen (N = 896) – Zuweisung der 12 schwierigkeitsveränderten Items über die erklärende Variable „Antwortformat“

[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