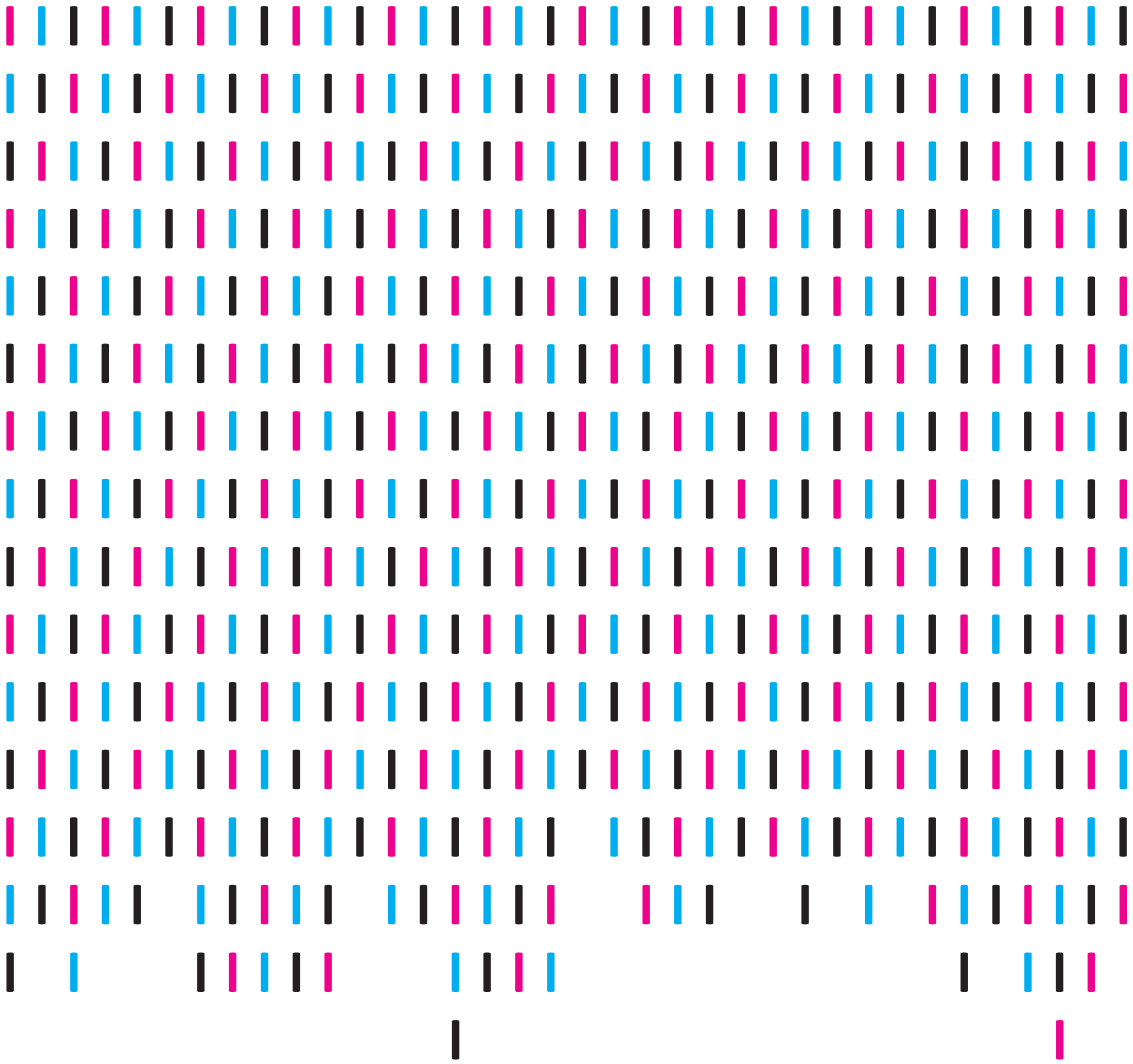

NEPS WORKING PAPERS

Stefan Zimmermann

ENTWICKLUNG EINER COMPUTER-BASIERTEN SCHWIERIGKEITSPRÄDIKTION VON LESEVERSTEHENSAUFGABEN

NEPS Working Paper No. 64
Bamberg, Januar 2016

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
https://www.neps-data.de/projektübersicht/publikationen/nepsworkingpapers

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

Entwicklung einer computerbasierten Schwierigkeitsprädiktion von Leseverstehensaufgaben

Stefan Zimmermann
Universität Bamberg, Universitätsklinikum Hamburg-Eppendorf

E-Mail-Adresse des Erstautors:

stefan.zimmermann@stud.uni-bamberg.de

Bibliographische Angabe:

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

Für die hilfreichen Rückmeldungen geht herzlicher Dank an Ingrid Koller (Alpen-Adria-Universität Klagenfurt), Manfred Hofer (Universität Mannheim) und Cordula Artelt (Otto-Friedrich-Universität Bamberg).

Entwicklung einer computerbasierten Schwierigkeitsabschätzung von Leseverstehensaufgaben

Zusammenfassung

Die Vorhersagbarkeit von Aufgabenschwierigkeiten im Bereich des Leseverstehens hat sowohl praktische Konsequenzen für die Ökonomie der Testkonstruktion als auch theoretische Implikationen. Aus theoretischer Perspektive besteht ein Forschungsdesiderat in der Untersuchung derjenigen kognitiven Prozesse, die beim Lösen von Leseverstehensaufgaben im Multiple-Choice-Format zum Tragen kommen. Zudem werden in der praktischen Testentwicklung beim Schreiben von neuen Testaufgaben Informationen darüber benötigt, was eine Aufgabe schwierig respektive leicht macht, um sukzessive in zukünftigen Studien die Passung zwischen Testschwierigkeit und Personenfähigkeit zu verbessern. Da die Erprobung von Testmaterial in Pilotstudien relativ zeitaufwändig und damit kostenintensiv ist, kann hier eine modellbasierte Abschätzung der Aufgabenschwierigkeiten die Pilotstudien sinnvoll ergänzen. Zu einfache oder zu schwere Aufgaben können somit vorab identifiziert werden und von einer empirischen Pilotierung ausgeschlossen werden oder auf Grund der Erkenntnisse, was eine Aufgabe schwierig macht, umgeschrieben werden. Für eine solche praktische Anwendung ist es entscheidend, dass die Ergebnisse der Schwierigkeitsabschätzungen den Testentwicklern bei Bedarf zur Verfügung gestellt werden. Deshalb wird in der vorliegenden Arbeit auf manuelle Kodierungen von Aufgabenmerkmalen verzichtet und verschiedene automatisierbare Methoden der quantitativen (Computer-) Linguistik eingesetzt (u.a. Part-of-Speech-Tagging und korpusbasierte Worthäufigkeitsanalysen) und im Hinblick auf ihre Eignung zur Prädiktion von Aufgabenschwierigkeiten in der Adoleszenz untersucht. Dazu werden die Aufgabenschwierigkeiten im Rahmen eines Linear Logistischen Testmodells (LLTM) als Linearkombinationen von Aufgabenmerkmalen dargestellt. Die empirischen Ergebnisse einer Hauptstudie in der 9. Klasse zeigen, dass ein zufriedenstellendes Modell zur Vorhersagbarkeit von Aufgabenschwierigkeiten aufgestellt werden kann, das zudem die Konstruktvalidität des NEPS Leseverstehenstests stützt. Da in der Sprachverarbeitung verschiedene Prozesse altersabhängig sind und sich im Erwachsenenalter der Wortschatz verbessert und das verbale Arbeitsgedächtnis sich verschlechtert, wird in einem nächsten Schritt die Übertragbarkeit der Ergebnisse auf das Erwachsenenalter überprüft. Dazu wird die Prädiktion der zuvor ausgewählten Aufgabenmerkmale bei den gleichen Aufgaben in einer Pilotstudie an Erwachsenen untersucht. Empirisch zeigt sich bei den Erwachsenen zwar eine ähnlich große Varianzaufklärung in den Aufgabenschwierigkeiten, die Konfidenzintervalle einzelner Schwierigkeitskomponenten sind jedoch u.a. auf Grundlage der deutlich kleineren Stichprobe erheblich größer. Somit können zwar in diesem Altersbereich leichte und schwierige Aufgaben identifiziert werden; gleichzeitig lassen die Ergebnisse für einzelne Schwierigkeitskomponenten jedoch keinen Vergleich mit der Studie in der Adoleszenz zu, so dass die Generalisierbarkeit auf andere Altersbereiche nicht abschließend geklärt werden kann.

Schlagworte

Lesekompetenzmessung, Linear Logistisches Testmodell (LLTM), Worthäufigkeiten, Part-of-Speech-Tagging, Altersunterschiede

Abstract

The prediction of item difficulties in reading assessments has major implications for the test development process as item writing can be organised more efficiently as the need for costly pilot studies decreases. In addition new insights into the construct of reading competency and the processing of multiple choice items are gained. The test development process is time consuming: more items have to be written than are required by the final test form as many items are discarded because of their psychometric item properties in pilot studies. The prediction of item difficulty could help making this process more efficient, as very easy or very difficult items could be identified and revised beforehand. Furthermore, the test difficulty can be adjusted well-informed to the competency level of the population for a maximum of test information. For the practical application in test development a prompt provision of the estimated item difficulty is necessary. Previous research has shown that the rating of the item features can be cumbersome. To overcome these issues a new methodological approach is suggested that draws on quantitative linguistics: word frequencies are analysed to assess the vocabulary and part-of-speech-tagging is used to assess the propositional density. These item features are then used in a linear logistic test model (LLTM) to predict item difficulties in a main study with 9^{th} graders. The data analysis shows that the resulting model fits the data reasonable well and that the construct validity of the NEPS reading tests is supported. As age differences in language components such as vocabulary and the verbal working memory are well known in adulthood, in a further pilot study it was investigated if the results could be generalized. The results showed a similar good prediction of the item difficulties. However, in the LLTM the effects of the item components possessed much larger confidence intervals. Thus, the generalization of the results in the adolescence remains vague.

Keywords

Reading assessment, age differences, Linear Logistic Test Model (LLTM), Part-of-Speech-Tagging

1. Einleitung

In der Testentwicklung muss zunächst ein größerer Pool von Aufgaben (Items) entwickelt werden: “In actual test development practice, the number of test items that must be developed and pretested is typically greater, and sometimes much greater, than the number that is eventually judged suitable for use in operational test forms.“ (Chalifour & Powers, 1989, S. 120). In einer oder mehreren darauf folgenden Entwicklungsstudien werden diese Aufgaben an der Zielpopulation überprüft und dann auf Grundlage von psychometrischen Kennwerten eine Itemselektion vorgenommen. Abschließend wird aus den besten Aufgaben ein Testinstrument zusammengestellt, so dass einerseits das zu erfassende Konstrukt in seiner Bandbreite abgedeckt ist und andererseits ein reliables Testverfahren entsteht. Eine Abschätzung der Aufgabenschwierigkeiten bereits während der Itementwicklung kann helfen den Anteil von Aufgaben zu reduzieren, die später auf Grundlage von empirischen Kennwerten in den Entwicklungsstudien ausgeschlossen werden müssen. Neben einer zu niedrigen Trennschärfe kann auch die Aufgabenschwierigkeit zu einem Ausschluss einer Aufgabe führen, wenn diese für die zu testenden Personen zu leicht oder zu schwer zu lösen ist. Wird an den Einsatz von adaptiven Tests gedacht, die eine ungleich höhere Anzahl von geeigneten Aufgaben in allen Schwierigkeitsbereichen voraussetzen (Thompson & Weiss, 2011; Wagner-Menghin & Masters, 2013), kann es sinnvoll sein den Testkonstruktionsprozess mit Vorhersagen der Aufgabenschwierigkeiten zu begleiten. Gleichzeitig lässt sich durch eine bessere Passung zwischen Personenfähigkeit und Aufgabenschwierigkeit eine höhere Präzision der psychometrischen Messung erzielen; bzw. man kann, unter Beibehaltung der Messgenauigkeit, die Anzahl der zu administrierenden Aufgaben und damit die Testzeit reduzieren (Kröhne & Martens, 2011). Entgegen früherer Erwartungen lässt sich jedoch durch den Einsatz leistungsangepasster Testmaterialien – und dadurch eine über- oder unterfordernde Testsituation vermeidend – die Testmotivation der Studienteilnehmer vermutlich nicht positiv beeinflussen (vgl. Frey, Hartig & Moosbrugger, 2009; Asseburg, 2011).

Zudem besteht die Möglichkeit, Schwierigkeit erzeugende Merkmale zu nutzen, um diagnostisch mehr über die Fertigkeiten der Studienteilnehmenden zu erfahren und Testwerte inhaltsbezogen interpretieren zu können. Bei der in Large-Scale-Assessment üblichen Konstruktion von Kompetenzstufen (Hartig, 2007) steht die qualitative Beschreibung quantitativer Leistungsunterschiede im Vordergrund, so dass Aussagen darüber getroffen werden können, über welche Fähigkeiten Studienteilnehmer verfügen. Während bei der Verwendung von Kompetenzstufen diese Aussagen zunächst auf Gruppenebene getroffen und dann auf die Individuen verallgemeinert werden (ebd.), ermöglichen neue Skalierungsverfahren, wie die *cognitive diagnostic models*, Aussagen über die Beherrschung von Teilfertigkeiten auf Individualebene (Huebner, 2010). Mit diesen Ansätzen lassen sich Testwerte besser hinsichtlich inhaltlicher Kriterien interpretieren oder sogar weiterführende Schlussfolgerungen für die zu Grunde liegenden Lernprozesse ziehen (Wilson, 2008).

Neben diesen praktischen Folgen für die Testkonstruktion und -interpretation hat die Untersuchung schwierigkeitsgenerierender Merkmale wesentliche theoretische Implikationen, da sie einen Beitrag zur Konstruktvalidierung leistet (vgl. Freedle & Kostin, 1993; Hartig & Frey, 2012). Lässt sich der Zusammenhang zwischen Aufgabenschwierigkeiten und Merkmalen nachweisen, die aus der Theorie über den Gegenstandsbereich abgeleitet

worden sind, lassen sich Aussagen über die Validität des Testverfahrens und über die bei der Testbearbeitung relevanten kognitiven Prozesse treffen. Im Bereich des Leseverstehens gibt es im Bereich der Kriteriumsvalidität besonderen Forschungsbedarf: „[…] some of the crucial steps in validating the comprehension process underlying multiple-choice tests of reading have still not been taken“ (Freedle & Kostin, 1994, S. 107). Während der Prozess des Lesenlernens und der Leseprozess im Allgemeinen bereits gut untersucht sind (z. B. Perfetti, Landi & Oakhill, 2005; Kendeou, Papadopoulos & Spanoudis, 2012), fehlt es hier noch an test- und altersübergreifenden Modellen zu den kognitiven Prozessen, die beim Lösen von Leseverstehensaufgaben im Multiple-Choice-Format zum Tragen kommen. In der Literatur wird oftmals sogar ein Lesekompetenzbegriff verwendet, bei dem Testleistungen „[…] zumeist auf einem sehr allgemeinen Leseverständnisbegriff [beruhen], der sich auf das Produkt des Leseverständnisses bezieht und nicht im Hinblick auf Merkmale zugrunde liegender Prozesse expliziert wird.“ (Richter & Christmann, 2002, S. 26). Die Gefahr besteht darin, dass bei einem solchen Vorgehen lediglich eine operationalisierte Definition des zu messenden Konstrukts resultiert, die für sich genommen bedeutungslos ist (Adler, 1947), weshalb die zu Grunde liegenden Prozesse weiter theoretisch ausgearbeitet und empirisch überprüft werden müssen.

2. Theoretischer Hintergrund

2.1 Leseverstehen

Lesen ist ein konstruktiver Prozess, bei dem ein Verständnis von der Bedeutung eines Textes vom Leser aktiv konstruiert wird. Die zu Grunde liegenden kognitiven Prozesse können in hierarchisch niedrige Prozesse, d. h. Wahrnehmungsprozesse, die spezifisch fürs Lesen sind, und hierarchisch höhere Prozesse, die allgemein das Verstehen betreffen, eingeteilt werden. Richter und Christmann (2002) unterscheiden grob zwischen drei kognitiven Prozessebenen: 1. Wort- und Satzidentifikation, 2. Herstellung der lokalen Kohärenz, und 3. Herstellung der globalen Kohärenz, wobei hier zusätzlich die Verarbeitung von Superstrukturen und die Identifikation von rhetorischen Strategien gehört. In den hierarchischen Modellen des Leseverstehens (Richter & Christmann, 2002) wird davon ausgegangen, dass die Prozesse auf den verschiedenen Ebenen miteinander in Wechselwirkung stehen und nicht kaskadenförmig ablaufen, d.h. der höhere Prozess erst dann beginnt, wenn der niedrigere Prozess abgeschlossen wurde. In diesem Sinne wird der Leseprozess gleichermaßen von Wahrnehmungsprozessen (bottom up) und der Lesererwartung, -motivation sowie dem Vorwissen (top down) beeinflusst (Kintsch, 2005).

Die Dekodierung einzelner Wörter erfolgt gemäß der dual route theory entweder über den Abruf aus dem orthographischen Lexikon oder alternativ über die Umwandlung von Graphemen zu Morphemen (Coltheart, Rastle, Perry, Langdon & Ziegler, 2001). Für eine diesbezügliche doppelte Dissoziation existiert empirische Evidenz (u. a. Ziegler, Castel, Pech-Georgel, George, Alario & Perry, 2008), so dass hier von zwei unabhängigen kognitiven Systemen auszugehen ist (Fodor, 1983). Nur wenn die Dekodierung der Wörter weitgehend automatisiert abläuft, stehen für die auf den höheren Ebenen stattfindenden Verstehensprozesse genügend kognitive Verarbeitungskapazitäten zur Verfügung (Rosebrock & Nix, 2006).

Die gelesenen Wörter und Sätze werden mental in eine Prädikat-Argument Struktur überführt (Kintsch, 2007). Für den zu Grunde liegenden parsing-Prozess spielt grammatikalisches Verständnis eine entscheidende Rolle, da Referenzen, wie bspw. Pronomen, richtig zugeordnet werden müssen. Sowohl hierbei als auch bei den höheren Verstehensprozessen ist davon auszugehen, dass das (verbale) Arbeitsgedächtnis einen limitierenden Faktor der Leseleistung darstellt (Perfetti, 2001; Just & Carpenter, 1992; Just, Carpenter & Keller, 1996). Können kleinere Textstellen schließlich widerspruchsfrei mental repräsentiert werden, entsteht die sogenannte lokale Kohärenz.

Im weiteren Verlauf wird zur Erzeugung der globalen Kohärenz das Verständnis immer größerer Textstellen rekursiv hergestellt und in einem sogenannten Situationsmodell mental repräsentiert (Kintsch & Van Dijk, 1978; Kintsch, 2004). Dabei werden die Präpositionen des Textes verdichtet und restrukturiert, so dass Präpositionen, die häufig und im Zentrum des semantischen Netzwerkes auftauchen, im Sinne der Verarbeitungstiefe besser verarbeitet werden. Dabei werden z.B. textbasierte oder wissensbasierte Inferenzen (Graesser & Kreuz, 1993; Graesser, Singer & Trabasso, 1994) gezogen, die automatisch ablaufen können oder einer aktiven Verarbeitungsstrategie bedürfen, um Zusammenhänge zwischen Textteilen herzustellen (Kintsch, 2004). Auch auf der Ebene der höheren kognitiven Prozesse laufen viele Verstehensprozesse automatisiert ab. Jedenfalls ist bei der Herstellung der Kohärenz davon auszugehen, dass sich zum Teil deutliche interindividuelle Unterschiede zwischen Lesern im Hinblick auf ihre Überwachung des Verständnisses zeigen lassen (Perfetti, Landi & Oakhill, 2005): „[...] a standard for coherence broadly determines the extent to which a reader will read for understanding, make inferences, and monitor his or her comprehension“ (S. 233).

2.2 Schwierigkeitskomponenten und Kognitive Prozesse

Inzwischen liegen für verschiedene Testinstrumente im Bereich des Leseverstehens verschiedene Analysen zur Prädiktion von Aufgabenschwierigkeiten vor (Tabelle 1). Insgesamt sind die Studien eher heterogen: Die Anzahl der Beobachtungen variiert zwischen 22 und 213 Aufgaben in Abhängigkeit davon wie viele Testformen betrachtet werden. Die Modellierungskomplexität ist sehr unterschiedlich, da zwischen 2 und über 100 Aufgaben- respektive Textmerkmale in den Analysen herangezogen werden. Methodisch werden die Aufgabenschwierigkeiten entweder in einer Regressionsanalyse durch die Aufgabenmerkmale vorhergesagt oder in einem Modell der probabilistischen Testtheorie werden zusätzliche Restriktionen eingeführt, so dass die Schwierigkeitsparameter der Aufgaben durch eine Linearkombination der Schwierigkeitskomponenten (d.h. Aufgabenmerkmale) abgebildet werden. Bei einigen Arbeiten ist das Vorgehen bei der Auswahl der untersuchten schwierigkeitsgenerierenden Merkmale kritisch zu hinterfragen. Teilweise werden die Prädiktoren nicht vor der eigentlichen Analyse auf Grundlage theoretischer Überlegungen ausgewählt, sondern auf Basis ihres empirischen Zusammenhangs mit der Aufgabenschwierigkeit (so werden bei Freedle & Kostin, 1993 zuerst beinahe hundert Prädiktoren aufgestellt, die dann ausschließlich empirisch reduziert werden) oder eine zunächst theoretische Modellierung wird schlussendlich an die empirischen Daten angepasst (u. a. Embretson & Wetzel, 1987; Sonnleitner, 2008; Poinstingl, 2009), so dass das resultierende Modell möglicherweise eine Überanpassung darstellt und Befunde nicht über den Test und die Stichprobe hinaus verallgemeinert werden können (vgl.

Hartig, 2007). Somit hat die Replikation und Überprüfung von bisherigen Ergebnissen an neuen Daten in diesem Forschungsfeld einen besonderen wissenschaftlichen Wert.

Tabelle 1: Literaturreview zu Text- und Aufgabenmerkmalen, die prädiktiv für die Aufgabenschwierigkeit von Leseverstehensaufgaben sind.

Studie	**Leseverstehens-Aufgaben**	**Modellierung**
Drum, Calfee & Cook, 1981	Leseskalen im CAT (California Achievement Test), CTBS (Comprehensive Test of Basic Skills) und STEP (Sequential Test of Educational Progress)	jeweils 20 - 36 Items Auswahl der 16 besten Prädiktoren aus 52 Aufgabenmerkmale 2 verschiedene Modelle Methode: Regression, 65%-95% Varianzaufklärung
Embretson & Wetzel, 1987	Leseskala im ASVAB (Armed Vocational Aptitude Battery)	46 Items 15 Aufgabenmerkmale 6 verschiedene Modelle Methode: Linear Logistisches Testmodell, Varianzaufklärung, bis zu 61% Varianzaufklärung
Freedle & Kostin, 1993	TOEFL (Test of English as Foreign Language)	213 Items Auswahl der 11 besten Prädiktoren aus über 100 Aufgabenmerkmale Methode: Regression, bis zu 58% Varianzaufklärung
Gorin, 2006	Leseskala im GRE (Graduate Record Examination)	200 Items 15 Aufgabenmerkmale 7 verschiedene Modelle Methode: Regression, bis zu 33% Varianzaufklärung
Hartig et al., 2011	Englischtest DESI (Deutsch Englisch International)	46 Items 2 Aufgabenmerkmale Methode: Regression und Linear Logistisches Test Modell, bis zu 42% Varianzaufklärung
Hartig & Frey, 2012	Englischtest DESI (Deutsch Englisch International)	46 Items 5 Aufgabenmerkmale

		Methode: Linear Logistisches Testmodell, 49% Varianzaufklärung
Kirsch, 2001	IALS (International Adult Literacy Survey)	34 Items 4 Variablen, in die additiv eine Vielzahl von Merkmalen einfließen Methode: Regression, 87% Varianzaufklärung
Nold & Rossa, 2007	Englischtest DESI (Deutsch Englisch International)	46 Items 6 Aufgabenmerkmale Methode: Regression, 45% Varianzaufklärung
Ozuru, Rowe, O'Reilly & McNamara, 2008	GMRT (Gates MacGinitie Reading Test), Testform 7.-9. Klasse & Testform 10.-12. Klasse	Jeweils 96 Items 15 Aufgabenmerkmale (inkl. modellierten Interaktionseffekten) Methode: Hierarchical Linear Modeling, Varianzaufklärung wird nicht berichtet
Sonnleitner, 2008	LEVE-E (Lese-verstehenstest für Erwachsene)	22 Items 11-12 Aufgabenmerkmale Methode Linear Logistisches Testmodell, 96% Varianzaufklärung

Über alle betrachteten Studien hinweg haben sich folgende Aufgabenmerkmale als besonders bedeutsam erwiesen: die präpositionale Dichte, der verwendete Wortschatz im Stimulustext und in der Aufgabe, verschiedenartig definierte kognitive Anforderungen, sowie die Plausibilität der Distraktoren, weshalb im Folgenden diese Merkmale kurz dargestellt werden.

Eine besondere Bedeutung kommt der präpositionalen Dichte (d.h. Relation zwischen der Anzahl von Präpositionen zur Textlänge) zu, da ihr Einfluss auf die Aufgabenschwierigkeit auch in der Grundlagenforschung bereits gut belegt ist (Kintsch, 2007). Eine exakte Auszählung der bedeutungstragenden Wörter und damit der Anzahl von Präpositionen ist sehr zeitaufwändig und kann nicht ohne weiteres automatisiert werden: „the propositional representation of the meaning of a text […] pose a major unsolved Problem. No fully automatic parser has yet been constructed that is capable of deriving a propositional microstructure from arbitrary English text. […] Hand coding is slow and cumbersome" (ebda. S. 54). Deshalb behilft man sich z. T. über relative Häufigkeiten von Wortarten. So betrachten Drum, Calfee und Cook (1981) in ihrer Arbeit den Anteil an Inhaltswörtern, also Nomen, Verben, Adjektive oder Adverben, da diese (fast) immer mit einer Präposition einhergehen und im semantischen Netzwerk integriert werden müssen. Jedenfalls ist davon auszugehen,

dass eine hohe präpositionale Dichte und damit verbundenen syntaktische Komplexität erhöhte Anforderungen an das verbale Arbeitsgedächtnis stellen (Caplan & Waters, 1999).

Ozuru, Rowe, O'Reilly und McNamara (2008) konnten zudem zeigen, dass der verwendete Wortschatz einen Einfluss auf die Aufgabenschwierigkeit hat. In Abhängigkeit davon, ob eher gebräuchliche oder seltenere Worte im Text verwendet wurden, waren Aufgaben leichter respektive schwieriger. Hierzu wurden dann die Wörter mit Wortlisten abgeglichen und der Anteil derjenigen Wörter bestimmt, die sehr selten sind. Diese Befunde konnten auch mit einfacheren Indikatoren repliziert werden, z. B. mit dichotomen Einschätzungen (siehe Hartig, Fry, Nold & Klieme, 2011; Hartig & Frey, 2012; Sonnleitner, 2008) oder mit einer dreistufigen Einschätzung des Wortschatzes, die von Experten vorgenommen werden (Nold & Rossa, 2007).

In der Arbeit von Kirsch (2001), in der er bei den in IALS (International Adult Literacy Survey) eingesetzten Leseverstehensaufgaben ein Modell zur Vorhersage von Aufgabenschwierigkeiten entwickelte, wiesen neben der Aufgaben-Text-Relation („*type-of-match*“) die Plausibilität der Distraktoren den deutlichsten Zusammenhang mit der Aufgabenschwierigkeit auf. Gemeinsam mit der Abstraktheit der gesuchten Information und der Komplexität insbesondere der diskontinuierlichen Texte konnten diese vier Merkmale 87% der Varianz in der Aufgabenschwierigkeit der IALS-Aufgaben erklären. Dabei wurde die Aufgaben-Text-Relation als mehrdimensionales Konstrukt operationalisiert, in dem verschiedene Aspekte additiv zur Aufgabenschwierigkeit beitragen: so wird u.a. kodiert, wie umfangreich die für die Lösung notwendige Textstelle ist und ob dabei ein oder mehrere Absätze mit einbezogen werden müssen. Zudem wird mit aufgenommen, ob bei den offenen Aufgabenformaten in der Aufgabenstellung nach mehreren Informationen gleichzeitig gefragt wird und bspw. mehrere Ursachen gleichzeitig für eine richtige Lösung angegeben werden müssen. Weiterhin wird zwischen den Anforderungen Lokalisieren von Informationen, dem wiederholten Abgleichen mit dem Text, dem Integrieren von Informationen und dem Generieren von Informationen unterschieden. Ein solches Modell kann zwar die Aufgabenschwierigkeiten empirisch gut vorhersagen, es ergeben sich dennoch einige Herausforderungen, da in diesem mehrdimensionalen Merkmal sehr unterschiedliche Aspekte integriert werden, die als einzelne Indikatoren in der Anwendung auf andere Daten dann plötzlich keine oder nur eine sehr geringe Bedeutung für die Aufgabenschwierigkeit haben. So wird im Nationalen Bildungspanel auf Seiten der kognitiven Anforderungen in Anlehnung an PISA zwischen dem Lokalisieren von Detailinformationen, dem Ziehen von Inferenzen und den Bewertungs- bzw. den Reflektionsaufgaben theoretisch unterschieden und diese Unterscheidung als Konstruktionsheuristik in der Aufgabenkonstruktion verwendet; es werden damit aber keine Schwierigkeitsunterschiede impliziert, da bspw. sowohl leichte als auch schwierige Informationsentnahme Aufgaben denkbar sind. Empirisch zeigen sich in einem in dieser Art konstruierten Test keine Unterschiede in den Aufgabenschwierigkeiten zwischen diesen Aufgabentypen, weshalb bei der Prädiktion von Aufgabenschwierigkeiten beim NEPS-Leseverstehenstest die Berücksichtigung dieses Merkmals nicht sinnvoll ist (Zimmermann, Gehrer & Artelt, 2013). Eine Aufnahme einer solchen Unterscheidung als Aufgabenmerkmal hat zudem einen gravierenden Nachteil: es liegt dann nahe, unterschiedliche Merkmale in Abhängigkeit vom jeweiligen Aufgabentyp zusätzlich mit ins Modell aufzunehmen; beispielsweise kann der Umfang und die Position einer Textstelle nur dann angegeben werden, wenn es sich um eine Aufgabe handelt, die sich nicht auf den Text als Ganzes bezieht. In der Folge ergeben sich recht aufwändige

Modellierungen mit geringer statistischer Power, da dann viele Aufgabenmerkmale nur für einen Teil der Aufgaben vorliegen.

In Bezug auf die Plausiblität der Distraktoren geht Kirsch (2001) davon aus, dass Aufgaben dann besonders schwer zu lösen sind, wenn Distraktoren mit Textstellenbezug vorhanden sind, oder auf Grundlage von Hintergrundwissen plausibel formuliert worden sind und sich die ablenkende oder widersprüchliche Information im selben Abschnitt befindet wie die richtige Lösung. Für Drum, Calfee und Cook (1981) ist die Entfernung zwischen der richtigen und einer widersprüchlichen Information jedoch nicht relevant. Aus ihrer Sicht ist es dagegen wichtig, wenn plausible Distraktoren nicht direkt auf Grund von Informationen im Text verworfen werden können. Dies deckt sich mit den Überlegungen von Embretson und Wetzel (1987), die ein Modell der Aufgabenbearbeitung entwerfen. Sie gehen davon aus, dass in einem ersten Schritt die dargebotenen Antwortoptionen darauf gelesen werden, ob man sie im Hinblick auf den zuvor gelesenen Text eineindeutig als falsch beurteilen kann und somit die Komplexität des Auswahlprozesses vereinfachen kann. Lassen sich durch dieses Vorgehen die Antwortoptionen auf nur noch eine Möglichkeit reduzieren, wählt man diese als Antwort aus. Verbleiben jedoch mehrere Antwortoptionen, die man nicht falsifizieren kann, wird in einem zweiten Schritt versucht, Informationen zu finden oder Inferenzen zu generieren, die den Wahrheitsgehalt einer der verbliebenen Antwortoptionen untermauert, so dass auf Grundlage dieses Prozesses dann eine Entscheidung getroffen werden kann. Maßgeblich für die Vergleiche zwischen Antwortoptionen und Stimulustext sind dabei sowohl semantische Überschneidungen, die entweder wortwörtlich oder im Sinne der Typologie von Anderson (1972) lexikalisch paraphrasiert oder syntaktisch restrukturiert worden sind, als auch Überschneidungen in den Propositionen. Embretson und Wetzel (1987) gehen davon aus, dass die Anzahl und Komplexität der notwendigen Vergleiche Anforderungen an das verbale Arbeitsgedächtnis stellen: „A large number of comparisons probably places an overall demand on working memory […]“ (S. 190).

Schließlich sollten die im Test verwendeten Testaufgaben betrachtet werden und für die Aufgabenschwierigkeit relevante Aufgabenmerkmale abgeleitet werden. Die Verwendung von Negationen im Aufgabenstamm wird kontrovers diskutiert, da zum einen eine Steigerung der Aufgabenschwierigkeit und zum anderen eine Reduzierung der Diskrimination vermutet wird (Haladyna, Downing & Rodriguez, 2002). Empirisch belegt sind eine deutlich höhere Fehlerhäufigkeit und eine längere Bearbeitungszeit in der Beurteilung des Wahrheitsgehalts von Aussagen, die eine Negation beinhalten (Wason, 1961). Bei Multiple-Choice Aufgaben führte die Verwendung von Negationen im Aufgabenstamm sowohl in den Untersuchungen von Dudycha und Carpenter (1973) als auch der Studie von Rachor und Gray (1996) zu einer erhöhten Aufgabenschwierigkeit. Hier liefert das Modell von Embretson und Wetzel (1987) möglicherweise eine Erklärung. Lässt sich die Antwortbearbeitung in einen Falsifikationsprozess und einen Verifikationsprozess unterteilen, so verändern sich hier die kognitiven Anforderungen durch eine Negation, da dann neben drei richtigen Antwortoptionen eine falsche identifiziert werden muss. Der Falsifikationsprozess zeichnet sich dadurch aus, dass Antwortoptionen ausgeschlossen werden, die im Widerspruch zu den bisher verarbeiteten Informationen stehen, während im Verifikationsprozess nach zusätzlichen Informationen gesucht wird, die eine der Antwortoptionen stützt. Kann die richtige Lösung bei einem Item mit Negation nicht direkt im Falsifikationsprozess ermittelt werden, folgt ein aufwändiger Verifikationsprozess, so dass damit die Aufgabenschwierigkeit steigt.

3. Fragestellung und Modellbildung

Eine vertiefte Analyse der Aufgabenmerkmale greift die bestehende Kritik konstruktiv auf, dass die Prozesse beim Leseverstehen in der Forschung oftmals ausgeblendet werden (Freedle & Kostin, 1994; Richter & Christmann, 2002). Gleichzeitig lässt sich bei einer Betrachtung prozessnaher Aufgabenmerkmale die Konstruktvalidierung des NEPS-Leseverstehenstests erreichen, indem nachgewiesen wird, dass Aufgaben dann schwieriger sind, wenn sie bspw. einen komplexeren Wortschatz voraussetzen oder sich auf einen propositional dichteren Text beziehen und sich bereits gut dokumentierte Befunde (siehe 2.2) mit NEPS Daten replizieren lassen. Bisherige Studien zeigen zudem, dass die Operationalisierungen und vorzunehmenden Kodierungen der Aufgabenmerkmale nicht ganz trivial sind (Kintsch, 2007). Erstens ist das exakte Auszählen der Präpositionen sehr zeitaufwändig. Zweitens müssen verschiedene Regeln beachtet werden, so dass bei einem komplexen Kodierschema notwendigerweise die Interkoderreliabilität überprüft werden muss. Somit ist ein solches Vorgehen für die Überprüfung größerer Itempools, wie sie im Kontext von Large-Scale-Assessments anfallen können, nur eingeschränkt geeignet. Im Rahmen einer die Testkonstruktion begleitenden Forschung, in der der Bedarf an Pilotstudien dadurch reduziert wird, dass man bspw. Aufgabenschwierigkeiten modellbasiert vorhersagt, müssen diese Abschätzungen zeitnahe vorliegen, so dass Testmaterialien und –aufgaben auf dieser Grundlage überarbeitet werden können. Hier bietet es sich an im Rückgriff auf Methoden der quantitativen Linguistik und der automatisierten Sprachverarbeitung Verfahren zu entwickeln, die eine zeitökonomische Einschätzung der Aufgabenschwierigkeiten zulassen. Ein solches Vorgehen soll im Folgenden entwickelt und hinsichtlich seiner Eignung überprüft werden. In der Studie 1 wird die Prädiktion von Aufgabenschwierigkeiten in einer größeren Studie mit Schülerinnen und Schülern der 9. Klasse untersucht, bevor anschließend dieselben Testaufgaben und dasselbe statistische Modell auf eine kleineren Pilotstudie mit Erwachsenen angewandt wird, um die Generalisierbarkeit der Befunde aus Studie 1 zu untersuchen.

4. Methode

4.1 Datengrundlage

Im Folgenden werden Daten des Nationalen Bildungspanels (NEPS) ausgewertet (z. B. Blossfeld, Roßbach & Maurice, 2011), in dessen Kontext ein neuer Lesekompetenztest konstruiert wurde. Grundlage bildeten dabei quasi-authentische Texte (Gehrer, Zimmermann, Artelt & Weinert, 2013; bzw. vertieft zu den Textsorten und –funktionen Gehrer & Artelt, 2013). Fünf kontinuierliche Texte, die sich jeweils auf die in der Rahmenkonzeption zentralen Textsorten (Sachtext, Kommentare, Gebrauchstexte, Werbungen und Literarische Texte) verteilen, wurden alters- und schwierigkeitsangemessen für die jeweilige Zielpopulation ausgewählt. In der Aufgabenkonstruktion wurden zusätzlich drei verschiedene Arten von kognitiven Anforderungen berücksichtigt: es handelt sich hierbei erstens um die Entnahme von Detailinformationen aus dem Text, zweitens um die Verknüpfung von kürzeren Textstellen im Sinne einer lokalen Kohärenzbildung, und drittens um Aufgaben, die eine Bewertung des gesamten Textes erfordern. Umgesetzt wurden diese

verschiedenen Anforderungen in drei unterschiedlichen Aufgabenformaten[1]: Mehrfachwahlaufgaben, bei denen der Wahrheitsgehalt verschiedener Aussagen bewertet werden muss, Zuordnungsaufgaben, die erfordern, dass mehreren Textabschnitten jeweils eine passende Überschrift zugeordnet wird und schließlich Multiple-Choice-Aufgaben (MC-Aufgaben), bei denen die Testteilnehmer unter vier Antwortoptionen die richtige Antwort heraussuchen müssen (Beispielaufgaben finden sich bei Gehrer, Zimmermann, Artelt & Weinert, 2012). Da die MC-Aufgaben den überwiegenden Teil der Aufgaben darstellen und unterschiedliche Aufgabenformate in der Aufgabenbearbeitung unterschiedliche kognitive Anforderungen stellen, werden in den folgenden Analysen lediglich die MC-Aufgaben betrachtet. Zudem ist die Aufgabenschwierigkeit bei mehrkategorialen Items nicht klar definiert; da man hier nicht nur den Anteil der richtigen Lösungen betrachten muss, sondern auch teilrichtige Lösungen vorliegen. Die Einschränkung auf MC-Aufgaben ermöglicht ferner erst die Anwendung des Linear Logistischen Testmodells (LLTM), da es dichotom gescorte Aufgaben voraussetzt.

4.2 Methodischer Ansatz: IRT & LLTM

Im dichotomen Rasch-Modell (Rasch, 1960; Rost, 2004), dem einfachsten Modell der Item-Response-Theorie, ist die Wahrscheinlichkeit eine Aufgabe *x* richtig zu lösen lediglich abhängig von der Differenz zwischen Personenfähigkeit ξ und Aufgabenschwierigkeit σ; mit *v* seien die Personen und mit *i* die Aufgaben nummeriert (z. B. Moosbrugger, 2007, S. 224):

$$P(x_{vi} = 1) = \frac{\exp(\xi_v - \sigma_i)}{1 + \exp(\xi_v - \sigma_i)}$$

Die aus einer Skalierung resultierenden Aufgaben- und Personenparameter liegen dabei auf einer gemeinsamen Dimension; entsprechen sich Aufgabenschwierigkeit und Personenfähigkeit, ergibt sich rechnerisch eine Lösungswahrscheinlichkeit von $P(x_{vi}) = 0.50$. Übersteigt die Personenfähigkeit die Aufgabenschwierigkeit, ergibt sich eine größere Lösungswahrscheinlichkeit, respektive im umgekehrten Fall eine kleinere Lösungswahrscheinlichkeit.

Möchte man diese Aufgabenschwierigkeit als abhängige Variable erklären, so gibt es zwei verschiedene etablierte Vorgehensweisen: ein zweistufiges Vorgehen, bei dem in einem ersten Schritt die Aufgabenparameter berechnet werden, die dann in einem zweiten Schritt in einem Regressionsmodell als abhängige Variablen eingehen. Oder die Aufgabenmerkmale werden bereits während der IRT-Skalierung im Rahmen des Linearen Logistischen Testmodells (LLTM, Fischer, 1973) berücksichtigt, so dass zusätzlich zum Raschmodell (siehe Gleichung oben) die *i* Aufgabenschwierigkeiten als Linearkomposition der Aufgabenmerkmale *q*, die mit *j* durchgezählt werden, dargestellt werden (Wilson & Moore, 2011). Dann gilt: $\sigma_i = \sum_{j=1}^{k} q_{ij}$. Bei diesem Modell werden die Aufgabenschwierigkeiten im Unterschied zur Regressionsanalyse qua Modell vollständig durch die berücksichtigten Aufgabenmerkmale erklärt, da keine Residuen zugelassen werden, was zu Lasten der Modellpassung geht (Hartig, 2007).

[1] Beispielaufgaben, die es auf Grundlage der psychometrischen Eigenschaften nicht in die Haupterhebung geschafft haben, inhaltlich aber unter Berücksichtigung dieser Testrahmenkonzeption konstruiert wurden, lassen sich bei Gehrer, Zimmermann, Artelt & Weinert (2012) finden.

In einer methodenvergleichenden Studie kommen Hartig, Frey, Nold und Klieme (2011) zu dem Schluss, dass man zu ähnlichen Ergebnissen gelangt, unabhängig davon, ob man Aufgabenschwierigkeiten regressionsanalytisch untersucht oder diese versucht im Linear Logistischen Testmodell zu erklären. Eine Erweiterung des Linear Logistischen Testmodells, das dann Residuen zulässt, hat nicht zu grundlegend anderen Ergebnissen geführt, zugleich aber spezielle Anforderungen an Software und Hardware gestellt: „A disadvantage is that models with cross-random effects can only be estimated by relatively few of the programs currently available and not within standard IRT modeling software. Estimation is computationally intensive; it poses high demands on computer hardware and is very time consuming. Given the need for specialized software and the relative novelty of cross-random effect models, results are more difficult to communicate“ (Hartig et al., 2011, S. 8-9). Somit fiel in der vorliegenden Forschungsarbeit schließlich auch aus Praktikabilitätserwägungen die Entscheidung für die Anwendung des klassischen LLTM.

4.3 Operationalisierung der Text- und Aufgabenmerkmale

Auf Grund der bisherigen Erkenntnisse über schwierigkeitsgenerierende Merkmale und die alterstypisch unterschiedlich stattfindende Sprachverarbeitung wurden folgenden Merkmale in der weiteren Analyse berücksichtigt: die Gebräuchlichkeit des verwendeten Vokabulars sowohl im Stimulustext als auch in den einzelnen Aufgaben, die propositionale Dichte im Text, das Vorhandensein von Negationen im Aufgabenstamm und der Bezug zwischen Lösung bzw. Distraktoren und Text bzw. Aufgabenstamm.

Wortschatz: die Schwierigkeit des verwendeten Wortschatzes wird ermittelt, indem die Worte im Stimulustext hinsichtlich ihrer Häufigkeit in der Leipzig Corpora Collection (Quasthoff, Richter & Biemann, 2006) analysiert werden. Da der Leipziger Korpus überwiegend auf Zeitungsartikeln im Internet basiert und somit nicht Alltagsschriftsprache beinhaltet, werden insgesamt vermutlich die Schwierigkeiten von Wörtern überschätzt, die überwiegend in literarischen Texten vorkommen. Dennoch sollte dieses Vorgehen eine grobe Beurteilung der Gebräuchlichkeit des verwendeten Wortschatzes ermöglichen[2]. Für die Berechnung der Gebräuchlichkeit des Wortschatzes wurde das Paket koRpus (Michalke, 2012) genutzt, das es ermöglicht Textkorpora für Worthäufigkeitsanalysen einzulesen. Für jedes Wort liegt zunächst ein Wert vor, wie häufig dieses auf eine Million Wörter im Korpus vorkommt. Da einige wenige Worte sehr häufig und sehr viele Wort sehr selten auftreten, wurde in der folgenden Analyse nicht der Rohwert der relativen Häufigkeit verwendet, sondern der Logarithmus zur Basis 10. Damit die Einschätzung des Wortschatzes zudem nicht von Extremwerten beeinflusst wird, wurde der Median über alle Worte im Text genommen, wobei bestimmte Wortarten nicht berücksichtigt wurden. Auf diesem Weg wird ausgeschlossen, dass die z.T. großen Häufigkeitsunterschiede in hochfrequenten Worten (z. B. bei *der* oder *die*) sich in einer unterschiedlichen Einschätzung des Wortschatzes niederschlagen. So wurden zuerst Artikel, Zahlwörter, Pronomen und Partikel mit Hilfe des Part-of-Speech Tagging (siehe unten) identifiziert und für die Berechnung der Gebräuchlichkeit des Wortschatzes ausgeschlossen. Somit werden Worte, die besonders häufig in allen Texten vorkommen und somit nicht gut zwischen einfachen und schwierigen

[2] Andere Wortlisten, die man alternativ nützen könnte, basieren auch auf einem reinen Sachtextkorpus (z.B. TIGER, Bonner Zeitungskorpus, European Corpus Initiative oder der Huge German Corpus).

Texten diskriminieren ausgeschlossen. Zur Beurteilung des verwendeten Wortschatzes in einem Item, wurden der Aufgabenstamm und die vier Antwortoptionen gemeinsam als ein eigener Text betrachtet.

Aus kognitionspsychologischer Sicht ist die *propositionale Dichte* von besonderem Interesse, um einzuschätzen wie komprimiert in einem Text Aussagen dargestellt sind. Die manuelle Auszählung der einzelnen Propositionen ist zeitaufwändig, so dass in der vorliegenden Analyse softwaregestützt mit Methoden des maschinellen Lernens vorgegangen wurde, um zu einer ökonomischen und reliablen Erfassung zu gelangen. Ausgangspunkt war die Forschungsarbeit von Brown, Snodgrass, Kemper, Herman und Covington (2008), in der gezeigt wird, wie man mit Hilfe eines Part-Of-Speech- (POS) Tagging zu einer Einschätzung der propositionalen Dichte gelangt. Beim POS-Tagging wird jedes einzelne Wort in einem Text seiner Wortart zugeordnet. Die Kategorien für die zu klassifizierenden Wortarten entstammen dem Stuttgarter-Tübinger-Tagset (Schiller, Teufel, Stöckert & Thielen, 1999). Der verwendete TreeTagger (Schmid, 1995) weist dabei für die deutsche Sprache eine Genauigkeit von 96-97,5% auf, wie Vergleiche zwischen manuellen und computerbasierten Kodierungen zeigen. Analog zum Vorgehen bei einer Vorstudie, die Brown, Snodgrass, Kemper, Herman und Covington (2008) durchführten, kann eine Annäherung an die propositionale Dichte erzielt werden, indem der Anteil von Verben, Adjektiven und Adverbien sowie Nomen an allen Wörtern im Text berechnet wird, da diese Wortarten immer mit einer Proposition einhergehen[3].

Negation im Aufgabenstamm: Es wurde dichotom kodiert, ob eine Verneinung im Aufgabenstamm (= 1) enthalten ist oder nicht (= 0): Eine Verneinung im Aufgabenstamm führt im Rückgriff auf das Aufgabenbearbeitungsmodell von Embretson und Wetzel (1987 zu einer qualitativ anderen kognitiven Anforderung, da dann in der Aufgabenbearbeitung nicht eine richtige Antwort aus vier Antwortoptionen herausgesucht werden soll, sondern eine falsche Antwortoption neben drei richtigen Antworten.

Für die Aufgabenschwierigkeit sollte zudem der Bezug zwischen Lösung, Text und Aufgabenstamm zentral sein. Ist hier ein besondere lexikalische Nähe gegeben, d. h. treten hier viele gemeinsame Worte auf, sollte für den Testteilnehmer die Lösung einfacher zu finden sein (vgl. Embretson & Wetzel, 1987); wenn hingegen die Distraktoren besonders nahe am Text und dem Aufgabenstamm formuliert sind, sollte dies die Aufgaben schwieriger machen. Um die Plausibilität der Distraktoren unabhängig vom Typus der Aufgabenstellung quantifizieren und operationalisieren zu können, wurde das Programmpaket *lsa* (Wild, 2007) verwendet, um einen korrelativen Zusammenhang zwischen den Antwortoptionen und dem Stimulustext bzw. der Aufgabenstellung zu berechnen. Wenn man hingegen die Plausibilität der Distraktoren in Anlehnung an Kirsch (2001) als Entfernung im Stimulustext zwischen ablenkenden Informationen und richtiger Textstelle operationalisiert, ergibt sich das Problem, dass es insbesondere bei Aufgaben, die eine Beurteilung des ganzen Textes erfordern, schwierig ist die Textstelle und deren Umfang eineindeutig zu benennen, die relevant für die Aufgabenlösung sind. In einem ersten Schritt werden in den zu

[3] Im Unterschied zur Arbeit von Brown et. al. (2008) wurden in der vorliegenden Arbeit keine weiteren Regeln beachtet, um die Genauigkeit zu verbessern, wie bspw. die Berücksichtigung der Kombination von Hilfs- und Vollverb als eine Proposition.

untersuchenden Texten sehr häufig verwendete Wörter entfernt, die keine textspezifische Information beinhalten wie Artikel und Partikel, bevor die gekürzten Texte dann in eine Worthäufigkeitsmatrix, die sogenannte *document-term matri*x, überführt werden. Diese erlaubt es das Auftreten gemeinsamer Wörter korrelativ zu berechnen. *Lösung-Text, Distraktoren-Text, Lösung-Aufgabenstamm* und *Distraktoren-Aufgabenstamm* bezeichnen jeweils den korrelativen Zusammenhang der auf Grundlage einer document-term matrix berechnet wird. Anschließend werden die Korrelationskoeffizienten in die Fischer-Z Verteilung überführt (vgl. Hotelling, 1953). Dies ist notwendig, da Korrelationskoeffizienten nicht normal verteilt sind und nicht über Intervallskalenniveau verfügen. Bei jeder Aufgabe wurden der Aufgabenstamm und jede der vier Antwortoptionen als einzelner Text betrachtet. Die Fischer-Z-Transformation erlaubt es zudem einen Mittelwert über die drei Korrelationskoeffizienten zu bilden, die sich auf die Distraktoren beziehen. Der Zusammenhang jedes einzelnen Distraktors zum Text und zum Aufgabenstamm wurde einzeln ermittelt. Anschließend wurden für die Distraktoren die drei Korrelationskoeffizienten gemittelt, indem diese zunächst Fischer-Z-transformiert wurden.

4.4 Deskriptive Statistik zu den Text- und Aufgabenmerkmalen

Es zeigt sich, dass der Indikator zur Einschätzung der propositionalen Dichte über eine geringe Varianz verfügt, während das aufgabenbezogene Vokabular stärker streut (Tabelle 2). Die vollständige dem LLTM zu Grunde liegende Item-Merkmal-Matrix ist im Anhang abgebildet (Tabelle 8).

Tabelle 2: Deskriptive Statistik zu den Text- und Aufgabenmerkmalen.

Merkmal	Mittelwert (SD)	Min	Max
Gebräuchlichkeit des Wortschatzes im Stimulustext	1.37 (0.25)	0.95	1.75
Propositionale Dichte im Text	0.59 (0.03)	0.56	0.62
Gebräuchlichkeit des Wortschatzes in der Aufgabe	1.39 (0.40)	0.00	1.86
Verneinung im Aufgabenstamm	0.07 (0.27)	0	1
Semantische Überschneidung von Lösung und Stimulustext[1]	0.09 (0.16)	-0.18	0.45
Semantische Überschneidung von Distraktoren und Stimulustext[1]	0.05 (0.21)	-0.16	0.61
Semantische Überschneidung von Lösung und Aufgabenstamm[1]	0.20 (0.20)	-0.03	0.79
Semantische Überschneidung von Distraktoren und Aufgabenstamm[1]	0.22 (0.26)	-0.04	0.79

Anmerkungen: [1] z-standardisierter Korrelationskoeffizient.

5. Studie 1: Entwicklung einer computerbasierten Schwierigkeitsabschätzung von Leseverstehensaufgaben in einer Haupterhebung

5.1 Stichprobe

Der Leseverstehenstest wurde von Schülerinnen und Schülern der 9. Klasse im Rahmen einer NEPS-Haupterhebung bearbeitet (*N*=13 898). Aus diesen Daten resultieren die empirischen Aufgabenschwierigkeiten, die in der Studie 1 vorhergesagt werden sollen. Der Altersbereich der Stichprobe war relativ homogen (*M*=15.7 Jahre, *SD* = 0.6 Jahre). Die besuchten Schulformen verteilen sich auf Hauptschulen (*N*=3380), Schulen mit mehreren Bildungsgängen (*N*=1068), Realschulen (*N*=3018), Gesamtschulen (*N*=1544) und Gymnasien (*N*=4887). Das Geschlechterverhältnis war zwischen Mädchen (*N*=6916) und Jungen (*N*=6980) ausgeglichen.

5.2 Raschskalierung

Eine wichtige Voraussetzung für die Anwendung des Linear Logistischen Testmodells besteht in der Raschhomogenität der Aufgaben, d.h. dass diese die Modellannahmen des Raschmodells nicht verletzen dürfen. Sowohl die Skalierung des dichotomen Raschmodells als auch die spätere Skalierung im Linear Logistischen Testmodell wurden mit dem Paket

eRm (Mair & Hatzinger, 2007), Version 0.15-4 (Mair, Hatzinger & Maier, 2014), durchgeführt, das entsprechende Routinen in der Statistiksoftware R, Version 3.1.1, (R Development Core Team, 2012) zur Verfügung stellt.

Die Summe der Schwierigkeitsparameter wurde für die Skalierung auf Null gesetzt. Der im Rahmen des Nationalen Bildungspanels skalierte Lesekompetenztest enthält zusätzlich vier komplexe Multiple-Choice-Aufgaben bzw. Zuordnungsaufgaben, bei denen im Rahmen eines Partial-Credit-Modells auch teilrichtige Lösungen berücksichtigt werden. Durch die Beschränkung auf die Multiple-Choice Aufgaben, die Verwendung eines einfachen dichotomen Raschmodells und die Verwendung des Programmpakets eRm, das im Gegensatz zu ConQuest mit CML (Conditional Maximum Likelihood) rechnet, ergeben sich Item- und Personenparameter, die nicht deckungsgleich mit den Skalierungen im Scientific-Use-File sind (vgl. die Parameter bei Haberkorn, Pohl, Hardt & Wiegand, 2012).

Zur Überprüfung der Passung der Aufgaben zum Raschmodell wurde u. a. der Infit und Outfit der einzelnen Aufgaben näher betrachtet (siehe Tab. 3). Der Infit basiert dabei auf den quadrierten Residuen zwischen dem durch das Modell implizierten Zusammenhang zwischen Personenfähigkeit und Lösungswahrscheinlichkeit und dem empirisch beobachtbaren Zusammenhang. Aufgaben, die einen Infit von eins aufweisen, passen perfekt zum Modell, während Werte größer eins im Sinne der klassischen Testtheorie eine problematische Trennschärfe aufweisen: beispielsweise. wenn fähigere Personen bei einer einfachen Aufgabe eine geringere Lösungswahrscheinlichkeit aufweisen als Personen mit einer durchschnittlichen Fähigkeit. Einen Infit von weniger als eins weisen hingegen Aufgaben mit einem „overfit“ auf, d. h. wenn sie zu stark diskriminieren. Während beim Infit Aufgabenschwierigkeiten stärker gewichtet werden, die in etwa der Personenfähigkeit entsprechen, berücksichtigt der Outfit stärker Abweichungen bei Aufgaben, die eher leicht oder schwer zu lösen sind. Im Allgemeinen werden für Infit MSQ (mean-square) und Outfit MSQ zwischen 0.5 und 1.5 noch als akzeptabel bewertet (de Ayala, 2008). Während der Infit für die skalierten 27 Lesekompetenzaufgaben auf eine gute Passung zum Raschmodell hinweist, da er im Wertebereich zwischen 0.81 und 1.20 liegt, zeigen sich etwas auffälligere Werte beim Outfit. So weist ein relativ niedriger Outfit bei den Aufgaben 7 und 10 darauf hin, dass diese im Vergleich zu den Modellannahmen empirisch zu stark zwischen fähigeren und weniger fähigeren Personen diskriminieren, während die Aufgaben 6, 8, 13 und 25 eher etwas zu schwach diskriminieren.

Tabelle 3: Aufgabenparameter, Infit und Outfit der 27 Aufgaben in der 9. Klasse.

Item	Aufgaben-schwierigkeit	Infit MSQ	Infit T	Outfit MSQ	Outfit T
1 reg90110_c	-2.14 (0.05)	0.912	-2.49	0.844	-2.13
2 reg90120_c	-2.96 (0.06)	0.875	-2.37	0.726	-2.60
3 reg90150_c	-0.10 (0.02)	0.976	-1.87	0.935	-2.58
4 reg90210_c	-0.74 (0.03)	0.935	-3.83	0.936	-1.78
5 reg90220_c	+0.93 (0.02)	1.075	8.22	1.123	8.12
6 reg90230_c	-0.82 (0.03)	1.017	0.92	1.360	8.51
7 reg90240_c	-0.82 (0.03)	0.826	-10.23	0.573	-13.29
8 reg90250_c	+2.17 (0.02)	1.205	22.13	1.344	20.96
9 reg90310_c	-0.71 (0.03)	0.896	-6.31	0.727	-8.41
10 reg90320_c	-1.53 (0.03)	0.814	-7.63	0.557	-9.51
11 reg90340_c	-0.95 (0.03)	0.886	-6.09	0.804	-5.08
12 reg90350_c	-1.17 (0.03)	0.861	-6.73	0.647	-8.73
13 reg90360_c	+0.10 (0.02)	1.156	12.52	1.348	13.70
14 reg90370_c	+0.77 (0.02)	1.082	8.65	1.108	6.65
15 reg90410_c	-0.72 (0.03)	0.980	-1.16	0.898	-2.83
16 reg90420_c	+0.21 (0.02)	0.939	-5.38	0.875	-5.87
17 reg90430_c	+0.65 (0.02)	0.927	-7.41	0.872	-7.52
18 reg90440_c	-0.09 (0.02)	0.909	-6.93	0.825	-6.88
19 reg90450_c	-0.46 (0.02)	0.878	-7.98	0.709	-9.79
20 reg90460_c	+0.87 (0.02)	1.028	2.94	1.071	4.33
21 reg90510_c	+1.46 (0.02)	0.901	-11.37	0.869	-9.90
22 reg90520_c	+0.98 (0.02)	0.989	-1.15	0.954	-2.94
23 reg90530_c	+1.14 (0.02)	1.095	9.28	1.133	8.04
24 reg90540_c	+1.77 (0.02)	0.865	-14.72	0.853	-10.24
25 reg90550_c	+1.59 (0.02)	1.190	18.18	1.340	20.21
26 reg90560_c	+0.89 (0.02)	0.931	-6.45	0.867	-7.66
27 reg90570_c	-0.32 (0.03)	0.974	-1.61	0.868	-4.00

5.3 LLTM und Modellvergleich mit dem Raschmodell

Die gute Passung der Aufgaben zum Raschmodell insgesamt erlaubt es nun, die zusätzlichen Annahmen des Linear Logistischen Testmodell in der Skalierung zu berücksichtigen und empirisch zu überprüfen. Im LLTM entfallen die 27 bzw. 26 Aufgabenparameter[4] und werden durch die untersuchten 8 Aufgabenmerkmale ersetzt, so dass das Modell deutlich sparsamer mit Modellparametern auskommt und die Passung zu den empirischen Daten allgemein

[4] Der 27. Parameter kann nicht frei geschätzt werden und ergibt sich aus der Randbedingung, dass die Summe aller Aufgabenparameter Null entspricht.

schlechter wird. Im Modellvergleich (Tab. 4) zeigt sich nun, dass die Einsparung von jeweils 18 Parametern sowohl mit einer Verschlechterung der LogLikelihood in der 9. Klasse, $\chi^2(18) = 9421.70$, $p < 0.001$ einhergeht und somit festgestellt werden muss, dass das Linear Logistische Testmodell deutlich schlechter zu den empirischen Daten passt. Es ist aber Konsens, dass dieser statistische Modelltest sehr konservativ ist und daher auch brauchbare Modelle verwirft (vgl. u. a. Gorin, 2005; Poinstingl, 2009; Hartig & Frey, 2012; Reif, 2012): „However, the associated LR difference in chi-square test rather frequently turns out significant, thus leading to rejection of the LLTM even when the graphical test for goodness-of-fit indicates a good match between the estimates of the Rasch item difficulties and their LLTM predicted values“ (Dimitrov & Raykov, 2003, S. 18).

Tabelle 4: Modellvergleich zwischen Raschmodell und LLTM.

	Log Likelihood	**Parameter**
Rasch	-112403.40	26
LLTM	-121825.10	8
Differenz	9421.70	18

Eine andere Methode die Modellpassung im linear logistischen Testmodell zu beurteilen, besteht darin zu untersuchen, wie gut die Aufgabenschwierigkeiten, die sich im Raschmodell ergeben, durch die Text- und Aufgabenmerkmale vorhergesagt werden können. Dazu können in einem graphischen Modelltest die Parameter des Raschmodells gegen die Parameter des Linearen Logistischen Testmodells abgetragen werden. Der Abstand der Aufgaben von der Winkelhalbierenden weist dabei auf die Residuen hin, d. h. die nicht durch die Aufgabenmerkmale erklärte Varianz in den Aufgabenschwierigkeiten.

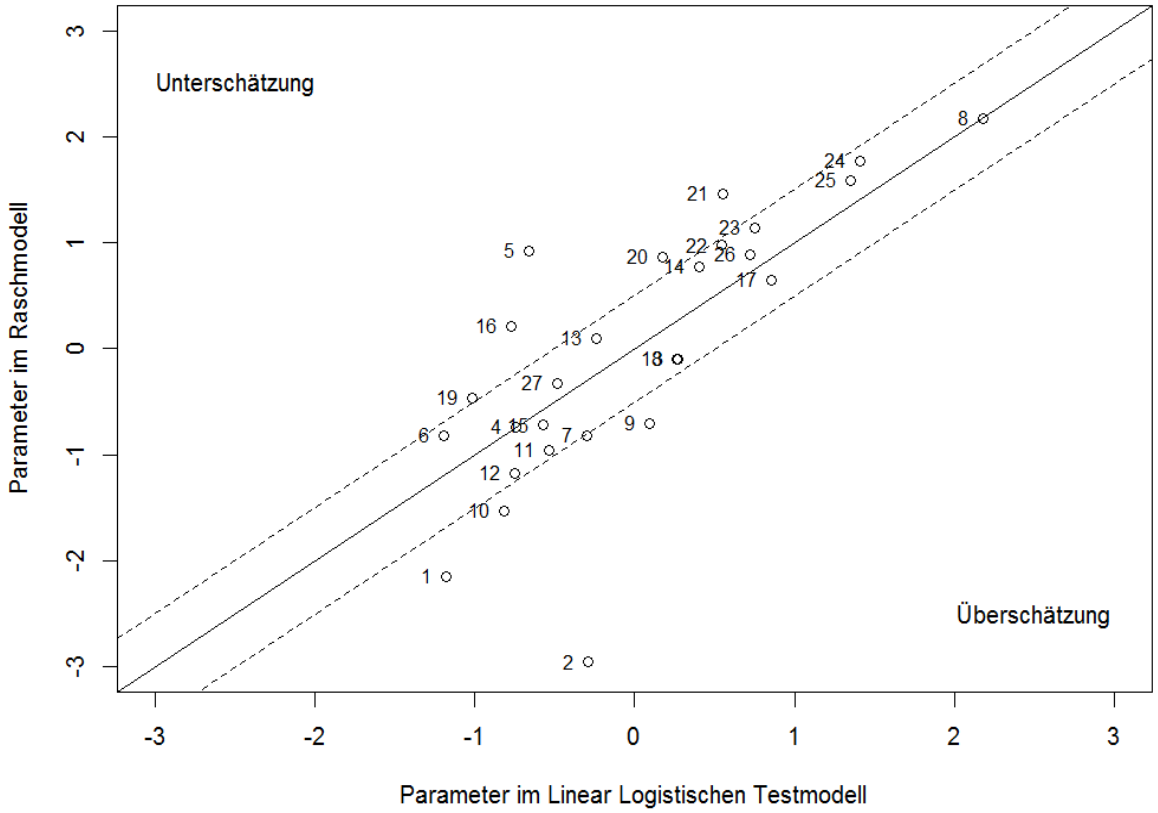

Abbildung 1: Graphischer Modelltest in der 9. Klasse.

Der graphische Modelltest in der 9. Klasse (siehe Abb. 3) zeigt, dass insbesondere die Aufgabenschwierigkeiten der Aufgaben 2 und 5 weit entfernt von der Winkelhalbierenden liegen und nicht korrekt vorhergesagt werden können. Zudem werden die Aufgabenschwierigkeiten der Aufgaben 5, 16, 19, 20 und 21 im LLTM deutlich unterschätzt, während die Schwierigkeiten der Aufgaben 1, 2, 9 und 10 deutlich überschätzt werden. Innerhalb der gestrichelten Linien liegen diejenigen Aufgaben, deren Schwierigkeit auf ± 0,75 Logits genau geschätzt werden können. Die Parameter aus der Raschskalierung und dem LLTM korrelieren deutlich miteinander, $r(25) = .77$, $p < 0.001$, so dass das resultierende Modell sinnvoll zu interpretieren ist.

5.4 Schwierigkeitsgenerierende Merkmale

Welche Merkmale haben nun auch empirisch einen Einfluss auf die Aufgabenschwierigkeiten?

Tabelle 5: Einfluss der Text- und Aufgabenmerkmale auf die Itemschwierigkeiten in der 9. Klasse..

	Parameter	**Std. Err**
Gebräuchlichkeit des Wortschatzes im Stimulustext	-0.92	0.04
Propositionale Dichte im Text	+11.12	0.30
Gebräuchlichkeit des Wortschatzes in der Aufgabe	-0.76	0.01
Verneinung im Aufgabenstamm	+1.88	0.02
Semantische Überschneidung von Lösung und Stimulustext[1]	-0.63	0.04
Semantische Überschneidung von Distraktoren und Stimulustext[1]	+0.75	0.03
Semantische Überschneidung von Lösung und Aufgabenstamm[1]	+1.21	0.03
Semantische Überschneidung von Distraktoren und Aufgabenstamm[1]	+1.90	0.02

Anmerkungen: [1] z-standardisierter Korrelationskoeffizient.

In der 9. Klasse zeigt sich ein Zusammenhang zwischen Aufgabenmerkmalen und der Aufgabenschwierigkeit, wobei die Richtung der Effekte mit den theoretischen Annahmen größtenteils übereinstimmen (Tabelle 5): wird im Stimulustext oder in einer Leseverstehensaufgabe eher gebräuchliches Vokabular verwendet, sind die entsprechenden Aufgaben einfacher zu lösen; mit zunehmender propositionalen Dichte im Text hingegen steigt die Schwierigkeit der Aufgaben an. Eine Verneinung im Aufgabenstamm führt hypothesenkonform zu einer erhöhten Aufgabenschwierigkeit, genauso wie ein engerer semantischer Bezug zwischen Distraktoren und Text bzw. Aufgabenstamm. Ist die Lösung hingegen nahe am Stimulustext formuliert, ist sie erwartungsgemäß einfacher zu lösen. Überraschenderweise führt eine semantische Überschneidung zwischen der richtigen Antwort und dem Aufgabenstamm zu einer erhöhten Aufgabenschwierigkeit; möglicherweise gehen die Testanden in der Aufgabenbearbeitung davon aus, dass eine parallele Formulierung zwischen Antwortoption und Aufgabenstamm auf eine Falschantwort hinweist.

5.5 Diskussion

In der Hauptstudie in der 9. Klasse entspricht der Einfluss der Schwierigkeitskomponenten auf die Aufgabenschwierigkeit weitgehend den theoretischen Annahmen. So zeigen sich deutliche Effekte der propositionalen Dichte, des Wortschatzes und der Relation zwischen Aufgabenstellung und Stimulustext auf die Aufgabenschwierigkeit. Somit konnten in der

vorliegenden Arbeit relevante Aufgabenmerkmale identifiziert werden, die die Aufgabenschwierigkeit maßgeblich beeinflussen. Die Prädiktion des resultierenden Modells ist etwas weniger gut als in den Vergleichsstudien (Tabelle 1). Vor dem Hintergrund, dass eine computergestützte Kodierung der Aufgabenmerkmale vorgenommen wurde und keine Kodierung von Aufgabenmerkmalen durch die Itementwickler benötigt wird, ist das Modell aber durchaus als gut zu bewerten. Die theoretische Begründung der Aufgabenmerkmale und die hypothesenkonformen Effekte der untersuchten Aufgabenmerkmale erbringen zudem einen Nachweis für die Konstruktvalidität des NEPS-Leseverstehenstest.

6. Studie 2: Eine Voruntersuchung zu alterstypischen Unterschieden in den Schwierigkeitskomponenten

Die bisherigen Studien, die sich der Prädiktion von Aufgabenschwierigkeiten im Bereich des Leseverstehens gewidmet haben, betrachten oftmals lediglich die Schwierigkeiten von Leseverstehensaufgaben innerhalb einer Altersgruppe. In diese Kategorie fallen bspw. die Ergebnisse zum Leseverstehen im Englischen in der DESI-Studie am Ende der Sekundarstufe I (Nold & Rossa; Hartig, Fry, Nold & Klieme, 2011; Hartig & Frey, 2012) und die Analyse von Aufgabenmerkmalen bei einem Leseverstehenstest für Erwachsene (Sonnleitner, 2008). Teilweise werden ausschließlich die Testaufgaben und deren empirische Schwierigkeiten betrachtet und nicht näher auf das Alter der Leserinnen und Leser eingegangen (z. B. Embretson & Wetzel, 1987; Freedle & Kostin, 1993), weshalb diese Studien keine Aussagen darüber zulassen, ob die Testbearbeitung altersspezifisch verschieden erfolgt. Betrachtet man funktionales Leseverstehen in Abhängigkeit vom Lebensalter, so fällt auf, dass in Querschnittsvergleichen eine moderate Verschlechterung der Leseleistung im hohen Erwachsenenalter festzustellen ist (van der Kamp & Boudard, 2003). Eine Erklärung für diese Entwicklung steht noch aus. Möglicherweise spielt hier die altersbedingte Verschlechterung des verbalen Arbeitsgedächtnisses (DeDe, Caplan, Kemtes & Waters, 2004) eine Rolle. Gleichzeitig verfügen jedoch ältere Erwachsene über einen größeren Wortschatz (Hartley, 1986), so dass in einem Altersvergleich verschiedene relative Stärken und Schwächen bestehen. Während alterstypische Unterschiede im mittleren und hohen Erwachsenenalter für verschiedene Sprachkomponenten bereits gut untersucht sind, besteht hier noch ein Forschungsdesiderat inwieweit diese Faktoren die Testbearbeitung in der Adoleszenz und im jungen Erwachsenenalter beeinflussen.

Im Folgenden soll daher exemplarisch an den Schwierigkeitskomponenten aus Studie 1 untersucht werden, ob diese auch im Erwachsenenalter prädiktiv für die Aufgabenschwierigkeit sind und ob das relative Gewicht der einzelnen Komponenten vergleichbar ist. Dazu wird zunächst der aktuelle Forschungsstand zu alterstypischen Unterschieden in der Sprachverarbeitung im Erwachsenenalter referiert, um anschließend abzuleiten in welchen Bereichen sich möglicherweise bereits Jugendliche von Erwachsenen unterscheiden.

In der Sprachproduktion weisen die Äußerungen von jüngeren Erwachsenen eine größere grammatikalische Komplexität und eine höhere propositionale Dichte auf (Kemper & Sumner, 2001). Eine Analyse der Dimensionalität verschiedener Sprachmaße zeigt zudem, dass im jungen Erwachsenenalter Leseverstehen und das verbale Arbeitsgedächtnis assoziiert sind, während im höheren Erwachsenenalter Leseverstehen stärker mit dem Wortschatz zusammenhängt. Dieser Befund legt aus Sicht von Kemper und Sumner (2001)

nahe, dass ältere Erwachsene aus einem umfassenderen Welt- und Wortschatzwissen bei der Bearbeitung von Leseverstehensaufgaben schöpfen können, während bei jüngeren Erwachsenen entscheidend ist, wie viele Informationen sie beim Lesen im Arbeitsgedächtnis halten können.

Hamm und Hasher (1992) untersuchten die Altersunterschiede im Ziehen von Inferenzen. Während keine Unterschiede in der Verarbeitung von richtigen Inferenzen gefunden werden konnten, zeigten sich Unterschiede in der Konstruktion von alternativen Inferenzen, die nur zu Teilen des gelesenen Textes passten, aber nicht zum Gesamttext. Die älteren Leser stimmten diesen Aussagen häufiger zu, so dass Hamm und Hasher (1992) davon ausgehen, dass es altersbedingt schwieriger wird, Inhalte im Arbeitsgedächtnis zu hemmen, wenn diese sich als irrelevant oder sogar als falsch herausgestellt haben. Diese Befunde wurden zudem von McGinnis und Zelinski (2000; 2003) repliziert. Es zeigte sich, dass die älteren Versuchsteilnehmer häufiger eine zu generalisierte Interpretation der Geschichte auswählten. Beim Verständnis von Pseudowörtern, deren Bedeutung aus dem Kontext erschlossen werden musste, konnten die älteren Versuchsteilnehmer weniger Facetten der Wortbedeutung produzieren und generalisierten die Wortbedeutung stärker.

DeDe, Caplan, Kemtes und Waters (2004) erforschten den Zusammenhang zwischen Alter, dem verbalen Arbeitsgedächtnis und verschiedenen Maßen der Sprachverarbeitung, um zu untersuchen, inwieweit ein altersbedingter Rückgang der Arbeitsgedächtnisleistung ursächlich für alterstypische Unterschiede in der Sprachverarbeitung ist. Theoretisch stellte sich hierbei die Frage, ob das verbale Arbeitsgedächtnis als eine zentrale Ressource allen Sprachverarbeitungsprozessen gleichermaßen zu Grunde liegt (vgl. Caplan & Waters, 1999) oder ob es einen besonderen Teil im Arbeitsgedächtnis zur Interpretation von Sätzen gibt. Ihre empirischen Befunde zeigen, dass für einen Leseverstehenstest und eine Satzverifikationsaufgabe die Alterseffekte über die Unterschiede im verbalen Arbeitsgedächtnis vermittelt werden, während bei einem auditiven Verstehenstest in Echtzeit kein Mediationseffekt vorliegt. Das bedeutet, dass lediglich für lesenahe Prozesse ein altersbedingter Rückgang der Leistung auf Unterschiede im verbalen Arbeitsgedächtnis zurückgeführt werden kann.

Miles und Stine-Morrow (2004) untersuchten Altersunterschiede in der Verarbeitung von Sätzen. Dazu wurden jüngeren und älteren Erwachsenen Sätze vorgelegt, die zwar gleich lang waren, aber unterschiedliche viele Propositionen enthielten. Beim Lesen wurde zugleich die Lesegeschwindigkeit erfasst. Anschließend wurden die Teilnehmer gebeten einzuschätzen, wie viel sie von dem gelesenen Satz erinnern würden. Daran anschließend folgte eine verbale Wiedergabe des Inhalts durch die Probanden. Es zeigte sich, dass die Lesedauer mit Zunahme der Propositionen zunahm und die älteren Erwachsenen insgesamt langsamer lasen als die jüngeren Erwachsenen. Zusätzlich konnte man einen Interaktionseffekt beobachten: so wiesen die älteren Erwachsenen bei Sätzen mit sehr vielen Propositionen eine minimal geringere Lesedauer auf als bei mittelschweren Sätzen. Die Autoren argumentieren, dass dies ein Hinweis darauf sein könnte, dass ältere Erwachsenen im Anbetracht von schwindenden kognitiven Ressourcen ihren Fokus auf die Bewältigung der mittelschweren Sätze richten.

Zusammenfassend lässt sich sagen, dass sich auch im Erwachsenenalter noch bedeutsame Unterschiede in den Sprachkomponenten zeigen. Ältere Erwachsene verfügen über einen

größeren Wortschatz und zeigen z. T. einen schnelleren lexikalischen Zugriff, verfügen aber über eine schlechtere Gedächtnisleistung, so dass die Auswirkungen dieser Veränderungen auf die Konstruktion des Leseverständnisses noch zu untersuchen sind.

In der Studie 2 soll der Forschungsfrage nachgegangen werden, ob sich potentielle Veränderungen in den Sprachkomponenten empirisch in einer unterschiedlichen Testbearbeitung niederschlägt.

Tabelle 6: Zuordnung der Aufgabenmerkmale zu den Sprachkomponenten.

Sprachkomponente	**Befund**	**Zusammenhang mit Aufgabenmerkmal (Hypothese)**
Wortschatz	relative Stärke im Erwachsenenalter	Gebräuchlichkeit des Wortschatzes im Stimulustext
		Gebräuchlichkeit des Wortschatzes in der Aufgabe
Verbales Arbeitsgedächtnis	relative Schwäche im Erwachsenenalter	Propositionale Dichte im Text
		Verneinung im Aufgabenstamm
Sonstiges		Semantische Überschneidung von Lösung und Stimulustext
		Semantische Überschneidung von Distraktoren und Stimulustext
		Semantische Überschneidung von Lösung und Aufgabenstamm
		Semantische Überschneidung von Distraktoren und Aufgabenstamm

Während offensichtlich ist, dass der in den Aufgaben und den Texten verwendete Wortschatz Anforderungen an den Wortschatz der untersuchten Personen stellt (Tabelle 6), ist der Zusammenhang zwischen Aufgabenmerkmalen und dem verbalen Arbeitsgedächtnis der Testanden weniger klar.

Gut belegt ist der Zusammenhang zwischen propositionaler Dichte und den Anforderungen an das verbale Arbeitsgedächtnis (Ericsson & Kintsch, 1995). Da Textinhalte, wenn sie nicht explizit auswendig gelernt werden sollen, mental in einem propositionalen Situationsmodell repräsentiert werden, gibt die Anzahl der Propositionen Aufschluss über die Anzahl der Informationen, die im Gedächtnis behalten werden müssen. Zudem spielt das Arbeitsgedächtnis auch bei der Verarbeitung von komplexer Syntax eine entscheidende Rolle (Caplan & Waters, 1999), die ihrerseits die Anzahl der Propositionen vergrößert. Eine Verneinung im Aufgabenstamm führt im Rückgriff auf das Aufgabenbearbeitungsmodell von

Embretson und Wetzel (1987 zu einer qualitativ anderen kognitiven Anforderung, da dann in der Aufgabenbearbeitung nicht eine richtige Antwort aus vier Antwortoptionen herausgesucht werden soll, sondern eine falsche Antwortoption neben drei richtigen Antworten. Für den dann stattfindenden Verifikationsprozess müssen dann bis zu drei Sätze anstatt einem im Gedächtnis behalten werden. Zudem weist die längere Bearbeitungszeit in der Beurteilung des Wahrheitsgehalts von Aussagen, die eine Negation beinhalten (Wason, 1961), darauf hin, dass das Arbeitsgedächtnis in diesem Fall stärker gefordert wird.

Ob jedoch im Altersvergleich Unterschiede bzgl. der Relation zwischen Stimulutstext und Aufgabe zu erwarten sind, ist unklar. Auf Grund der referierten Arbeiten wäre hier allenfalls zu spekulieren, dass im höheren Erwachsenenalter eine Tendenz zum (vor-) schnellen Generalisieren von Bedeutungen (McGinnis & Zelinski, 2000; 2003) möglicherweise dazu führen könnte, dass diese Personen dann eher auf Distraktoren hereinfallen, die semantisch sehr nahe am Stimulustext formuliert sind.

Aus dem Forschungsstand bzgl. der Unterschiede in der Sprachverarbeitung im Altersvergleich lassen sich zusammenfassend folgende Hypothesen ableiten: Es ist davon auszugehen, dass bei älteren Leserinnen und Lesern – im Vergleich zu jüngeren Leserinnen und Lesern – die propositionale Dichte in einem Text einen stärkeren Einfluss auf die Aufgabenschwierigkeit ausübt, da bei ihnen das verbale Arbeitsgedächtnis eher einen limitierenden Faktor des Textverstehens darstellt. Umgekehrt sollte bei jüngeren Leserinnen und Lesern der Wortschatz einen limitierenden Faktor im Textverständnis darstellen und somit der im Text und den Aufgaben vorkommende Wortschatz einen stärkeren Einfluss auf die Aufgabenschwierigkeit haben.

6.1 Stichprobe

Dieselben Aufgaben wie in der Studie 1 wurden in einer zusätzlichen Studie von Erwachsenen (*N*=504) bearbeitet, wobei der Stichprobe ein Quotenstichprobenplan hinsichtlich Alter und Bildung zu Grunde lag. Der Altersbereich der Stichprobe erstreckte sich zwischen 18 und 77 Jahren (*M*=45.9 Jahre, *SD* = 12.7 Jahre) und war zusätzlich im Hinblick auf die Bildung der Personen sehr heterogen[5]. Insgesamt waren die Frauen leicht über- (*N*=281) bzw. die Männer unterrepräsentiert (*N*=214).

6.2 Raschskalierung & Modellvergleich zwischen Raschmodell und LLTM

In der Raschskalierungen zeigt sich im Vergleich zu der Studie in der 9. Klasse ein schlechterer Itemfit (Tabelle 9 im Anhang). Während der Infit im Wertebereich zwischen 0,77 und 1,25 liegt und damit einen guten Itemfit anzeigt, sind zumindest zwei Outfitwerte als kritisch zu bewerten: Die Aufgaben 2 und 10 weisen einen Outfit von 0,46 respektive von 0,38 auf und sind damit empirisch trennschärfer als die Annahmen im zu Grunde liegenden Messmodell. Auffällig ist zudem die Aufgabe 14 mit einem Outfit von 1,40, was auf eine zu

[5] Höchster Schulabschluss: allgemeine oder fachgebundene Hochschulreife (*N*=142), Fachhochschulreife (*N*=39), Mittlere Reife (*N*=167), Hauptschulabschluss (*N*=139), Sonstiger oder kein Schulabschluss (*N*=17); Höchster beruflicher Ausbildungsabschluss: Abgeschlossenes universitäres Studium (*N*=73), Abschluss Fachhochschule (*N*=58), Beamtenausbildung (*N*=4), Meister oder Technikerabschluss (*N*=22), Abschluss einer Fachschule / Berufsfachschule / Handelsschule / Schule des Gesundheitswesens (*N*=49), Abgeschlossene Lehre (*N*=218), Sonstiger Ausbildungsabschluss (*N*=20), kein beruflicher Ausbildungsabschluss (*N*=58).

niedrige Trennschärfe hinweist, im Rückgriff auf die Regeln bei de Ayala (2008) aber toleriert werden kann.

Zumindest für 25 der 27 Aufgaben lässt sich somit die Raschkonformität nachweisen. Aus Gründen einer besseren Vergleichbarkeit mit den Ergebnissen der Studie 1 werden im Folgenden alle 27 untersucht und die Aufgaben 2 und 10 nicht aus der Analyse ausgeschlossen. Erneut zeigt sich im direkten Vergleich zwischen Raschskalierung und LLTM, dass Letzteres schlechter zu den Daten passt und zu einer Verschlechterung der LogLikelihood, $\chi^2(18) = 252.97$, $p < 0.001$, führt. Der graphische Modelltest in der Erwachsenenstudie (siehe Abb. 4) zeigt, dass die Aufgabenschwierigkeiten der Aufgaben 5, 16, 19 und 25 im LLTM deutlich unterschätzt werden, während die Schwierigkeiten der Aufgaben 1, 2, 9, 10, 11 und 18 deutlich überschätzt werden. Insbesondere die Aufgabenschwierigkeiten der Aufgaben 2 und 5 liegen weit entfernt von der Winkelhalbierenden und können nicht korrekt vorhergesagt werden. Die Parameter aus der Raschskalierung und dem LLTM korrelieren deutlich miteinander, $r(25) = .74$, $p < 0.001$, so dass das resultierende Modell dennoch sinnvoll zu interpretieren ist. Insgesamt zeigt sich in der Varianzaufklärung eine geringfügig schlechtere Prädiktion der Aufgabenschwierigkeiten im Vergleich zu der Studie im Jugendalter.

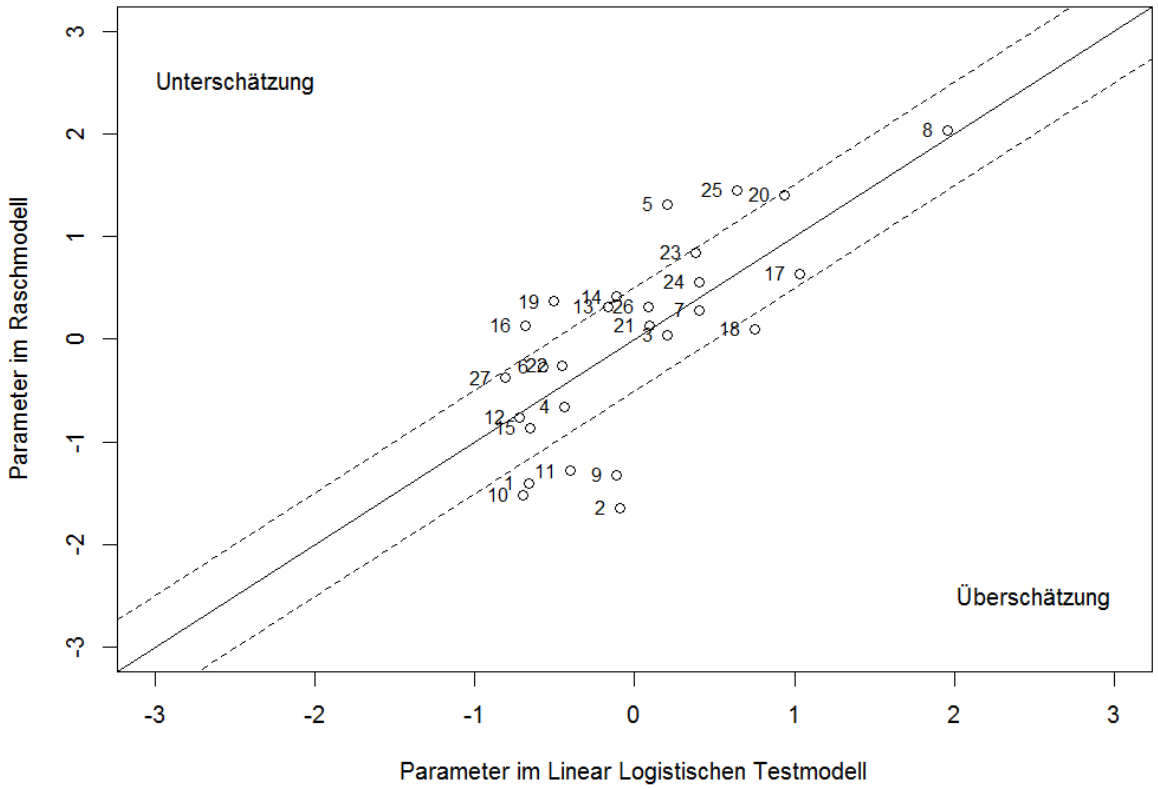

Abbildung 2: Graphischer Modelltest in der Erwachsenenstudie.

6.3 Ergebnisse & Diskussion

In Bezug auf die zuvor aufgeworfene Fragestellung, ob die untersuchten schwierigkeitsgenerierenden Merkmale bei Jugendlichen dieselben Effekte wie bei Erwachsenen haben oder nicht, werden beide Studien getrennt im LLTM skaliert und anschließend die Werte der Parameter zwischen den Studien verglichen. Auf Grundlage der theoretischen Überlegungen wurde erwartet, dass die Parameter, die sich auf den Wortschatz beziehen, bei den Erwachsenen kleiner ausfallen als bei den Jugendlichen, da die Erwachsenen insgesamt über einen größeren Wortschatz verfügen und somit die Bedeutungskraft des Wortschatzes zur Prädiktion der Aufgabenschwierigkeit abnehmen sollte. Umgekehrt sollten diejenigen Parameter, die in Zusammenhang mit dem verbalen Arbeitsgedächtnis gebracht werden können, bei den Erwachsenen größer ausfallen als bei den Jugendlichen, da diese bei den älteren Studienteilnehmern vermehrt einen limitierenden Faktor des Textverständnisses darstellen sollten. (Tabelle 7)

Tabelle 7: Einfluss der Text- und Aufgabenmerkmale auf die Itemschwierigkeiten bei Erwachsenen.

	Parameter	**Std. Err**
Gebräuchlichkeit des Wortschatzes im Stimulustext	+0.53	0.20
Propositionale Dichte im Text	-3.40	1.72
Gebräuchlichkeit des Wortschatzes in der Aufgabe	-0.04	0.09
Verneinung im Aufgabenstamm	+0.77	0.11
Semantische Überschneidung von Lösung und Stimulustext[1]	-1.12	0.21
Semantische Überschneidung von Distraktoren und Stimulustext[1]	+1.79	0.19
Semantische Überschneidung von Lösung und Aufgabenstamm[1]	+0.87	0.15
Semantische Überschneidung von Distraktoren und Aufgabenstamm[1]	+1.03	0.12

Anmerkungen: [1] z-standardisierter Korrelationskoeffizient.

In der Erwachsenenstudie scheinen auf den ersten Blick einige Zusammenhänge kontra intuitiv zu sein: so ist ein einfacher Wortschatz im Stimulustext mit einer höheren Aufgabenschwierigkeit oder eine höhere propositionale Dichte mit einer niedrigeren Aufgabenschwierigkeit assoziiert, während die Richtung der anderen Effekte vergleichbar mit den Befunden in der 9. Klasse ist. Eine genauere Betrachtung zeigt allerdings, dass insbesondere die Standardfehler, die sich auf den Wortschatz im Stimulustext, die propositionale Dichte und den Wortschatz in der Aufgaben beziehen, besonders groß sind. Das liegt zum einen an der Verteilung der Aufgabenmerkmale. Schließlich liegen den beobachtbaren 27 Aufgaben nur 5 unterschiedliche Texte zu Grunde, so dass die Variation in denjenigen Merkmalen eingeschränkt ist, die sich auf den Stimulustext beziehen (siehe Tab. 5). Zum anderen ist die Erwachsenenstudie stichprobenmäßig deutlich kleiner (N=504) im Vergleich zur Erwachsenenstudie (N=13 898), so dass zunächst die Aufgabenparameter mit einer größeren Ungenauigkeit geschätzt werden (siehe Tab. 2 die zu den Aufgabenparametern gehörenden Fehler), was sich dann auch auf das Konfidenzintervall der Effekte der Schwierigkeitskomponenten auswirkt. Unter Berücksichtigung der Konfidenzintervalle lässt sich die Richtung der beiden scheinbar widersprüchlichen Befunde in der Erwachsenenstudie nicht eindeutig bestimmen und damit nicht weiter interpretieren. Dies gilt auch für den aufgabenbezogenen Wortschatz in der Erwachsenenstudie, bei dem die Richtung des Effektes zwar den theoretischen Erwartungen entspricht aber das Konfidenzintervall so groß ist, dass nicht mit hinreichender Sicherheit ausgeschlossen werden kann, dass das Vorzeichen korrekt ist.

Diese Gemengelage erlaubt es leider nicht alle zuvor theoretisch aufgestellten Hypothesen empirisch zu überprüfen; die anderen Prädiktoren verhalten sich hingegen in beiden Studien relativ ähnlich: je üblicher der verwendete Wortschatz in der Aufgabe ist, desto leichter ist die beobachtbare Aufgabenschwierigkeit. Eine Verneinung im Aufgabenstamm führt zu einem deutlichen Anstieg in der Aufgabenschwierigkeit. Je ähnlicher die Aufgabenlösung am Text formuliert ist, desto niedriger ist die Aufgabenschwierigkeit. Sind hingegen die Distraktoren nahe am Stimulustext formuliert, steigt die Aufgabenschwierigkeit. Bei Formulierungen in den Antwortoptionen, die Wörter aus dem Aufgabenstamm aufgreifen, steigt die Aufgabenschwierigkeit unabhängig davon ob es sich um eine richtige oder falsche Antwortoption handelt. Dies lässt den Schluss zu, dass dann der Prozess eine Antwort auszuwählen insgesamt schwieriger wird.

Im Vergleich zwischen den Studien zeigt sich lediglich, dass eine Verneinung im Aufgabenstamm in der 9. Klasse mehr Schwierigkeit erzeugt als im Erwachsenenalter. Die eingangs aufgestellte Hypothese, dass im Erwachsenenalter auf Grundlage einer schlechteren Leistung des verbalen Arbeitsgedächtnisses Negationen schlechter verarbeitet werden und somit Negationen dann einen stärkeren Einfluss auf die Aufgabenschwierigkeit haben, muss daher verworfen werden.

7. Schlussfolgerungen

In der vorliegenden Arbeit konnten relevante Aufgabenmerkmale identifiziert werden, die sich computerbasiert automatisiert kodieren lassen und die Aufgabenschwierigkeit zumindest näherungsweise vorhersagen. Die dabei aufgedeckten Zusammenhänge zwischen Aufgabenmerkmalen und Aufgabenschwierigkeiten decken sich mit den eingangs aufgestellten Hypothesen. Insgesamt kann mit diesen Befunden die Konstruktvalidität des NEPS-Leseverstehenstest empirisch erbracht werden.

Ein wesentlicher Befund ist, dass moderne Verfahren der automatischen Sprachverarbeitung und der quantitativen Linguistik dazu genutzt werden können, um ein Modell zur Vorhersage von Aufgabenschwierigkeiten im Bereich des Leseverstehens aufzustellen. Ungefähr die Hälfte der Varianz in der Aufgabenschwierigkeit kann dabei vom Modell erklärt werden. Die Modellpassung zeigt aber auch, dass noch weitere Komponenten schwierigkeitsbestimmend sind, die bei der Modellierung nicht berücksichtigt wurden. Das ist nicht weiter verwunderlich, da die untersuchten schwierigkeitsgenerierenden Merkmale *post hoc* auf einen Leseverstehenstest angewendet werden, der mit einer anderen Zielsetzung entwickelt wurde: verschiedene Textsorten, kognitive Anforderungen und Aufgabenformate altersangemessen umzusetzen. Dabei ist bekannt, dass Textsorten unterschiedliche kognitive Anforderungen stellen (Voss, Carstensen & Bos, 2005). Während es bei Sachtexten nur ein korrektes Situationsmodell gibt, wird bei literarischen Texten spielerisch mit Mehrdeutigkeit umgegangen, u.a. durch die Verwendung von Metaphern und Allegorien, so dass es in der Regel mehr als ein angemessenes Situationsmodell gibt (Roick, Stanat, Dickhäuser, Frederking, Meier & Steinhauer, 2010). In der vorliegenden Analyse, bei der zunächst nur eine kleine Anzahl von Aufgaben untersucht wurde, musste zunächst ein parametersparsames Modell entwickelt werden, dass keine Überanpassung an die Daten darstellt und überhaupt eine Parameterschätzung erlaubt. Vor diesem Hintergrund, dass die Aufgabenmerkmale computergestützt erhoben wurden und keine Expertenurteile in das Modell zusätzlich eingegangen sind, ist die Modellpassung als gut zu bewerten. Gleichzeitig ist das resultierende Modell durch die Automatisierbarkeit besonders geeignet für

verschiedene Einsatzzwecke. Die geschätzten Aufgabenschwierigkeiten können genutzt werden, um potentielle sehr leichte oder sehr schwere Aufgaben im Vorfeld einer empirischen Untersuchung zu identifizieren. Diese können dann von Domänenexperten gesichtet und überarbeitet oder aus dem Test ausgeschlossen werden. Ein solches Modell ist zudem auch für andere Testinhalte interessant; bspw. um die Leseanforderungen gering zu halten, wenn andere Inhalte (z.B. naturwissenschaftliche Kompetenzen) im Fokus der Testkonstruktion stehen.

Zudem muss man die Modellpassung auch vor dem Hintergrund der im LLTM zu Grunde liegenden Annahmen bewerten. So wird im LLTM davon ausgegangen, dass die verschiedenen Schwierigkeitskomponenten additiv wirken und etwaige Defizite nicht durch andere Teilfertigkeiten kompensiert werden können. Das ist eine sehr starke Annahme. Es gibt zwar auch IRT-Modelle, die kompensatorische Effekte zulassen (siehe für einen Überlick Roussos, Templin & Henson, 2007): Bspw. ist im Dino-Modell (deterministic input noisy or-gate) die Lösungswahrscheinlichkeit einer Aufgabe lediglich davon abhängig ob *eine* der benötigten Teilfertigkeiten vorliegt. Somit wird von einer Gleichwertigkeit der verschiedenen Teilfertigkeiten ausgegangen. Daher liegt diesem Modell ihrerseits eine starke Annahme zu Grunde, die überprüft werden müsste.

Ob die vermuteten Unterschiede in der Testbearbeitung im Altersvergleich existieren, kann in der vorliegenden Arbeit nicht abschließend geklärt werden. Erst in der empirischen Analyse der Erwachsenendaten zeigte sich, dass relevante Hypothesen nicht zufriedenstellend statistisch getestet werden konnten, da der Einfluss der Schwierigkeitskomponenten in der Erwachsenenstudie - insbesondere der textbezogenen Merkmale – nicht genügend genau abgeschätzt werden kann. Somit gibt die vorliegende Arbeit keinen Hinweis darauf, dass es im Jugend- und Erwachsenenalter grundsätzliche Unterschiede in der Testbearbeitung gibt; eine altersbedingte veränderte Testbearbeitung kann aber auch nicht eindeutig ausgeschlossen werden. Eine Differential Item Functioning (DIF) Analyse aller 31 Lesekompetenzaufgaben (Pohl & Carstensen, 2013) weist durchaus daraufhin, dass fast bei der Hälfte der Aufgaben zwischen den Studien Unterschiede in den Aufgabenschwierigkeiten bestehen, wobei eine theoretische Erklärung hierfür dann immer noch aussteht. Dennoch können verschiedene Aspekte aus der Erwachsenenstudie gelernt werden: Für eine zukünftige Betrachtung von Schwierigkeitskomponenten im Altersvergleich ist es notwendig größere Itempools oder sehr große Stichproben zu untersuchen, wie der Vergleich mit den Ergebnissen in der 9. Klasse mit den vergleichsweise kleineren Konfidenzintervalle zeigen.

Für eine Untersuchung von Altersunterschieden wäre zudem eine Gruppe von altershomogenen älteren Erwachsenen wünschenswert. In der vorliegenden Arbeit war es wegen der Stichprobengröße nicht möglich innerhalb der Erwachsenengruppe altershomogene Untergruppen zu untersuchen. Experimentelle allgemeinpsychologische Studien setzen zur Untersuchung von Altersunterschieden im verbalen Arbeitsgedächtnis nicht vor dem fünften oder sechsten Lebensjahrzehnt an. Möglicherweise haben die Limitationen im Arbeitsgedächtnis auch erst dann einen Einfluss auf die Sprachverarbeitung, wenn die Bearbeitungszeit der Aufgaben limitiert ist und keine kompensatorischen Verarbeitungsstrategien angewendet werden können. Vor diesem Hintergrund, wäre es interessant zusätzlich die Bearbeitungszeiten der einzelnen Aufgaben mitzuerheben und im Altersvergleich zu betrachten. Auf Grundlage dieser Limitationen ist die vorliegende Studie zu den Altersunterschieden lediglich als eine erste Pilotstudie in diesem Bereich anzusehen.

Ein Rückgriff auf Methoden der quantitativen Linguistik ermöglicht perspektivisch weitere Ergänzungen in den psychometrischen Modellen zur Prädiktion von Aufgabenschwierigkeiten, die möglicherweise sensitiver in einem Altersvergleiche sind: 1. POS-Tagging erlaubt es die grammatikalische Komplexität von Texten abzubilden, indem der Einsatz von Konnektoren und die durchschnittliche Satzlänge mit betrachtet werden. Längere Sätze sind mit schwierigeren Aufgaben verbunden, wobei Konnektoren zu einem besseren Textverständnis beitragen, da semantische Zusammenhänge expliziert werden und weniger überbrückende Inferenzen gezogen werden müssen (Johnston & Pearson, 1982). 2. ist es möglich die Diversität im verwendeten Wortschatz mit zu berücksichtigen – hier gibt es verschiedene Indikatoren, die sich im Gegensatz zur einfachen Type-Token-Relation, auch zum Vergleich kurzer oder unterschiedlich langer Texte heranziehen lassen (McCarthy & Jarvis, 2010). 3. eine Überführung von Texten in eine *document-term-matrix* ermöglicht es einfache semantische Analysen automatisiert durchzuführen. So lässt sich die interne Kohärenz eines Textes dadurch ermitteln, indem man anhand der Verteilung von Schlüsselwörtern in verschiedenen Abschnitten des Textes untersucht, ob über alle Absätze hinweg dasselbe Thema behandelt wird oder wie viele thematische Wechsel vorkommen (Foltz, Kintsch & Landauer, 1998).

Acknowledgment

Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS) SC4, doi:10.5157/NEPS:SC4:1.1.0. 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.

Literatur

Adler, F. (1947). Operational defintions in sociology, *American Journal of Sociology, 52 (5),* 438-444.

Alexander, P. (2006). The path to competence: a lifespan developmental perspective on reading, *Journal of Literacy Research, 37 (4),* 413-436.

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension, *Review of Educational Research, 42,* 145-169.

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests*. Dissertation Christian-Albrechts-Universität zu Kiel.

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*.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R. & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging, *Behavior Research Methods, 40 (2)*, 540-545.

Caplan, D. & Waters, G. (1999). Verbal working memory and sentence comprehension, *Behavioral and Brain Sciences, 22 (1)*, 77-126.

Chalifour, C. L. & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discrimination, *Journal of Educational Measurement, 26 (2)*, 120-132.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108 (1)*, 204–256.

de Ayala, R. J. (2008). *The theory and practice of item response theory.* New York: Guilford.

DeDe, G., Caplan, D., Kemtes, K. & Waters, G. (2004). The relationship between age, verbal working memory, and language comprehension. *Psychology and Aging, 19 (4)*, 601–616.

Dimitrov, D. M. & Raykov, T. (2003). Validation of cognitive structures: a structural equation modeling approach, *Multivariate Behavioral Research, 38 (1),* 1-23.

Drum, P.A., Calfee, R.C. & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16 (4),* 486–514.

Dudycha, A. L. & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology, 58*, 116-121.

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

Ericsson, K. A. & Kintsch, W. (1995). Long-Term Working Memory. *Psychological Review, 102* (2), 211-245.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37 (6),* 359–374.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge: MIT Press.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis, *Discourse Processes, 25 (2 & 3)*, 285-307.

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

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

Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests, *Diagnostica, 55 (1)*, 20-28.

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.

Gehrer, K., Zimmermann, S., Artelt, C. & Weinert, S. (2012). *The assessment of reading competence (including sample items from grade 5 and 9)*, online verfügbar unter: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_re_2012_en.pdf [letzter Zugriff: 29.11.2013], Bamberg: Otto-Friedrich Universität Bamberg, Nationales Bildungspanel.

Gehrer, K., Zimmermann, S., Artelt, C. & Weinert, S. (2013). Framework for assessing reading competence and results from an adult pilot study, *Journal for Educational Research Online, 5 (2),* 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. & Kreuz, R. J. (1993). A theory of inference generation during text comprehension, *Discourse Processes*, 16, 145-160.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101 (3)*, 371-395.

Haberkorn, K., Pohl, S., Hardt, K. & Wiegand, E. (2012). *NEPS technical report for reading - scaling results of starting cohort 4 in ninth Grade*, online verfügbar unter https://www.neps-data.de/Portals/0/Working%20Papers/WP_XVI.pdf [letzter Zugriff am 29.11.2013], Working Paper No. 16, Bamberg: Otto-Friedrich Universität Bamberg, Nationales Bildungspanel.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment, *Applied Measurement in Education, 15 (3),* 309-334.

Hamm, V. P. & Hasher, L. (1992). Age and the availability of inferences, *Psychology and Aging, 7 (1),* 56-64.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In E. Klieme & B. Beck (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.

Hartig, J., Frey, A., Nold, G. & Klieme, E. (2011). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*.

Hartley, J. T. (1986). Reader and text variables as determinants of discourse memory in adulthood, *Psychology and Aging, 1 (2),* 150-158.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms, *Journal of the Royal Statistical Society, Series B (Methodological), 15 (2)*, 193-232.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments, *Practical Assessment, Research & Evaluation, 15 (3),* 1–7.

Johnston, P. & Pearson, P. D. (1982). *Prior knowledge, connectivity, and the assessment of reading comprehension*, Technical Report No. 245, Illinois University.

Just, M. A. & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory, *Psychological Review, 99 (1)*, 122-149.

Just, M. A., Carpenter, P. A. & Keller, T. A. (1996). The capacity theory of comprehension: new frontiers of evidence and arguments. *Psychological Review, 103, (4)*, 773-780.

Kemper, S. & Sumner, A. (2001). The structure of verbal abilities in young and older adults, *Psychology and Aging, 16 (2),* 312-322.

Kendeou, P., Papadopoulos, T. & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers, *Learning and Instruction, 22 (5)*, 354-367.

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. In R. B. Ruddell & N. J. Unrau (Hrsg.), *Theoretical models and processes of reading*, S. 1270–1328, Newark: International Reading Association.

Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: the CI perspective. *Discourse Processes, 39 (2 & 3),* 125–128.

Kintsch, W. (2007). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

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

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

Kröhne, U. & Martens, T. (2011). Computer-based competence tests in the National Educational Panel Study: the challenge of mode effects, *Zeitschrift für Erziehungswissenschaft, 14,* 169-196.

Mair, P. & Hatzinger, R. (2007). Extended rasch modeling: the eRm package for the application of IRT models in R, *Journal of Statistical Software, 20 (9),* 1-20.

Mair, P., Hatzinger, R. & Maier, M. J. (2014). *eRm: Extended Rasch Modeling*. R package version 0.15-4.

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment, *Behavior Research Methods, 42 (2)*, 381-392.

McGinnis, D. & Zelinski, E. M. (2000). Understanding unfamiliar words: the influence of processing resources, vocabulary knowledge, and age. *Psychology and Aging, 15 (2),* 335–350.

McGinnis, D. & Zelinski, E. M. (2003). Understanding unfamiliar words in young, young-old, and old-old adults: inferential processing and the abstraction-deficit hypothesis. *Psychology and Aging, 18 (3),* 497–509.

Michalke, M. (2012). *koRpus: ein R-paket zur Textanalyse*. Paper presented at the Tagung experimentell arbeitender Psychologen (TeaP), Mannheim.

Miles, J. R. & Stine-Morrow, E. A. L. (2004). Adult age differences in self-regulated learning from reading sentences, *Psychology and Aging, 19 (4),* 626-636.

Moosbrugger (2007). Item-Response-Theorie (IRT) In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion*, S. 215-260, Heidelberg: Springer.

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

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.

Perfetti, C. A. (2001). Reading Skills. In N. J. Smelser & P. B. Baltes (Hrsg.), *International Encyclopedia of the Social & Behavioral Sciences*, S. 12800-12805, Oxford: Pergamon.

Perfetti, C. A., Landi, N. & Oakhill, J. (2005). The aquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Hrsg.), *The science of reading: a handbook*, S. 227–253, Oxford: Blackwell.

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

Poinstingl, H. (2009). The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test, *Psychological Science Quarterly, 51 (2),* 123-134.

Quasthoff, U., Richter, M., & Biemann, C. (2006, May). Corpus portal for search in monolingual corpora. In *Proceedings of the fifth international conference on language resources and evaluation* (pp. 1799-1802).

R Development Core Team (2012). *R: a language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

Rachor, R. E. & Gray, G. T. (8.-12.4.1996). *Must all stems be green? A study of two guidelines for writing multiple choice stems*, Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Kopenhagen: The Danish Institute for Educational Research.

Roick, T., Stanat, P., Dickhäuser, O., Frederking, V., Meier, C. & Steinhauer, L. (2010). Strukturelle und kriteriale Validität der literarästhetischen Urteilskompetenz. Projekt literarästhetische Urteilskompetenz. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung: Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes*, S. 165-174. Weinheim: Beltz.

Roussos, L. A., Templin, J. L. & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models, *Journal of Educational Measurement, 44 (4),* 293-311.

Reif, M. (2012). Applying a construction rational to a rule based designed questionnaire using the Rasch model and LLTM, *Psychological Test and Assessment Modeling, 54 (1),* 73-89.

Richter, T. & Christmann, U. (2002). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (S. 25-28). Weinheim: Juventa.

Rosebrock, C. & Nix, D. (2006). Forschungsüberblick: Leseflüssigkeit (Fluency) in der amerikanischen Leseforschung und –didaktik. *Didaktik Deutsch, 20*, 90-112.

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

Schiller, A., Teufel, S., Stöckert, C. & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS.* Technischer Bericht. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

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

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests, *Practical Assessment, Research & Evaluation, 16 (1),* 1-9.

van der Kamp, M. & Boudard, E. (2003). Literacy proficiency of older adults, *International Journal of Educational Research, 39 (3)*, 253-263.

Voss, A., C. H. Carstensen & W. Bos (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.

Wagner-Menghin, M. M. & Masters, G. N. (2013). Adaptive testing for psychological assessment: how many items are enough to run an adaptive testing algorithm?, *Journal of Applied Measurement, 14 (2),* 106-117.

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology, 52*, 133-142.

Wild, F. (2007). *An LSA package for R.* In Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL'07), S. 11-12. Heerlen.

Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift für Psychologie, 216 (2),* 74–88.

Wilson, M. & Moore, S. (2011). Building out a measurement model to incorporate complexities of testing in the language domain. *Language Testing, 28 (4),* 441–462.

Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F.-X. & Perry, C. (2008). Developmental dyslexia and the dual route model of reading: simulating individual differences and subtypes. *Cognition, 107*, 151–178.

Zimmermann, S., Gehrer, K. & Artelt, C. (11.3.2013). *Schwierigkeitsgenerierende Merkmale bei Leseverstehensaufgaben im Nationalen Bildungspanel (NEPS): Ein Vergleich von erwachsenen und jugendlichen Lesern*, Vortrag auf der 1. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Universität Kiel.

Anhang

Tabelle 8: Die Ausprägungen der 27 Items bei den 8 untersuchten Aufgabenmerkmalen.

	Aufgabenmerkmal							
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Item 1	1.75	0.62	1.67	0	0.12	-0.07	-0.02	-0.03
Item 2	1.75	0.62	1.40	0	0.24	0.12	0.15	0.19
Item 3	1.75	0.62	1.27	0	0.13	0.06	0.29	0.33
Item 4	1.48	0.58	1.66	0	0.34	-0.05	0.12	0.28
Item 5	1.48	0.58	1.79	0	0.37	0.36	0.16	0.20
Item 6	1.48	0.58	1.49	0	-0.08	-0.14	-0.03	-0.04
Item 7	1.48	0.58	1.26	0	0.03	0.32	0.11	0.11
Item 8	1.48	0.58	1.38	1	-0.03	0.61	-0.02	0.42
Item 9	0.95	0.56	1.53	0	-0.18	-0.16	0.27	0.31
Item 10	0.95	0.56	1.38	0	0.12	0.03	-0.02	-0.02
Item 11	0.95	0.56	1.69	0	-0.07	-0.12	0.40	-0.02
Item 12	0.95	0.56	1.76	0	0.08	-0.16	0.13	0.13
Item 13	0.95	0.56	1.85	0	0.12	-0.06	0.79	-0.02
Item 14	0.95	0.56	1.52	0	0.02	-0.12	-0.03	0.70
Item 15	1.41	0.56	0.60	0	0.45	-0.01	0.21	-0.02
Item 16	1.41	0.56	0.85	0	0.27	-0.11	0.17	-0.02
Item 17	1.41	0.56	1.38	0	0.21	0.30	0.29	0.79
Item 18	1.41	0.56	1.40	0	0.37	0.40	0.42	0.42
Item 19	1.41	0.56	1.14	0	-0.01	-0.10	-0.01	-0.01
Item 20	1.41	0.56	1.54	0	-0.01	0.31	0.43	0.34
Item 21	1.43	0.62	1.53	0	-0.15	-0.10	0.12	0.50
Item 22	1.43	0.62	0.00	0	0.06	0.06	-0.02	-0.02
Item 23	1.43	0.62	1.45	0	-0.03	0.10	0.28	0.44
Item 24	1.43	0.62	1.86	1	-0.11	-0.13	0.36	-0.03
Item 25	1.43	0.62	1.43	0	0.01	0.07	0.40	0.70
Item 26	1.43	0.62	1.46	0	0.10	-0.06	0.43	0.43
Item 27	1.43	0.62	1.18	0	0.09	-0.10	-0.02	-0.02

Anmerkungen. A: Gebräuchlichkeit des Wortschatzes im Stimulustext, B: Propositionale Dichte im Text, C: Gebräuchlichkeit des Wortschatzes in der Aufgabe, D: Verneinung im Aufgabenstamm, E: Zusammenhang zwischen Lösung und Text, F: Mittlerer Zusammenhang zwischen den Distraktoren und Text, G: Zusammenhang zwischen Lösung und Aufgabenstamm, H: Mittlerer Zusammenhang zwischen den Distraktoren und Aufgabenstamm.

Tabelle 9: Aufgabenparameter, Infit und Outfit der 27 Aufgaben bei Erwachsenen.

Item	Aufgaben-schwierigkeit	Infit MSQ	Infit T	Outfit MSQ	Outfit T
1 reg90110_c	-1.40 (0.20)	0.942	-0.36	0.986	0.06
2 reg90120_c	-1.64 (0.21)	0.810	-1.24	0.464	-1.92
3 reg90150_c	+0.04 (0.13)	0.957	-0.58	0.839	-1.19
4 reg90210_c	-0.66 (0.15)	0.935	-0.63	1.146	0.76
5 reg90220_c	+1.32 (0.11)	1.030	0.64	1.090	1.23
6 reg90230_c	-0.26 (0.14)	1.075	0.91	1.016	0.15
7 reg90240_c	+0.28 (0.12)	0.888	-1.75	0.731	-2.44
8 reg90250_c	+2.04 (0.11)	1.065	1.43	1.117	1.50
9 reg90310_c	-1.32 (0.19)	0.930	-0.47	0.817	-0.59
10 reg90320_c	-1.51 (0.20)	0.770	-1.59	0.378	-2.50
11 reg90340_c	-1.28 (0.19)	0.858	-1.06	0.977	0.01
12 reg90350_c	-0.76 (0.16)	0.855	-1.40	0.568	-2.32
13 reg90360_c	+0.32 (0.12)	1.137	2.03	01.222	1.80
14 reg90370_c	+0.41 (0.12)	1.255	3.77	1.395	3.20
15 reg90410_c	-0.86 (0.17)	1.103	0.90	0.965	-0.07
16 reg90420_c	+0.14 (0.13)	0.819	-2.57	0.614	-3.21
17 reg90430_c	+0.64 (0.13)	0.959	-0.62	0.937	-0.56
18 reg90440_c	+0.10 (0.14)	0.942	-0.74	1.115	0.82
19 reg90450_c	+0.37 (0.13)	1.008	0.14	0.988	-0.06
20 reg90460_c	+1.41 (0.12)	1.073	1.39	1.184	2.26
21 reg90510_c	+0.13 (0.14)	0.941	-0.72	0.837	-1.12
22 reg90520_c	-0.26 (0.15)	1.058	0.63	0.901	-0.50
23 reg90530_c	+0.84 (0.13)	0.956	-0.63	0.903	-0.89
24 reg90540_c	+0.56 (0.14)	0.971	-0.36	0.999	0.03
25 reg90550_c	+1.45 (0.13)	1.155	2.56	1.190	2.04
26 reg90560_c	+0.32 (0.15)	0.921	-0.94	0.783	-1.55
27 reg90570_c	-0.38 (0.17)	0.849	-1.42	0.815	-0.85