Diagnostische Entscheidungen mit dem Treatment Decision Model - ein entscheidungstheoretischer Ansatz auf Basis von Evaluationsstudiendaten

Inaugural-Dissertation zur Erlangung des Doktorgrades der Philosophie an der Ludwig-Maximilians-Universität München

vorgelegt von

Elisabeth Barbara Kraus

aus

Würzburg

2022

Referent: Prof. Dr. Markus Bühner

Korreferent: Prof. Dr. Sven Hilbert

Tag der mündlichen Prüfung: 10.05.2022

Danksagung

Danken möchte ich meinen Betreuern und Gutachtern Markus Bühner und Sven Hilbert, die mich stets in meiner Entwicklung als Wissenschaftlerin gefördert und begleitet haben. Danke im Besonderen für den guten Austausch und die Unterstützung in dieser Arbeit meine Ideen zu verfolgen. Ich habe mich gut aufgehoben gefühlt. Danke auch an alle Kolleginnen und Kollegen für den fruchtbaren Austausch und die offenen Türen. Des Weiteren möchte ich meiner Familie, meiner Verlobten und meinen Freundinnen und Freunden danken, die mir bei Schwierigkeiten jeglicher Art immer fest zur Seite standen und ohne deren Geduld diese Arbeit nicht möglich gewesen wäre.

Zusammenfassung

Die vorliegende Arbeit formalisiert, parametrisiert und implementiert ein diagnostisches Treatmententscheidungsmodell. Zur Formalisierung wird die Treatmententscheidung in den Kontext statistischer Entscheidungstheorie eingebettet. Die Parametrisierung erfolgt mit psychometrischen, sowie non-parametrischen statistischen Modellen und statistischen Entscheidungsregeln. Implementiert wird das Treatmententscheidungsmodell am Beispiel der Auswahl zweier Lesetrainings mit dem Ziel der maximalen Steigerung der Lesekompetenz. Auf Basis von Evaluationsstudiendaten der FiLBY-Studie (Fachintegrierte Leseförderung Bayern) werden hierfür in Studie 1 psychometrische Modelle der Lesekompetenzmessung mit dem Bayerischen Lesetest entwickelt. In Studie 2 wird der Trainingserfolg zweier FiLBY-Lesetrainings mit den in Studie 1 gemessenenen Lesekompetenzen funktional in Beziehung gesetzt. Die funktionalen Zusammenhänge werden mit generalisierten additiven Modell geschätzt und bilden die Nutzenfunktion des entscheidungstheoretischen Problems. In Studie 3 werden eine Erwartungsnutzen- und eine Maximin-Entscheidungsregel entwickelt. Anschließend werden diese auf die Nutzenfunktionen aus Studie 2 unter Berücksichtigung der messbedingten Unsicherheit aus Studie 1 angewendet. So entsteht ein Treatmententscheidungsmodell, das in der konkreten Anwendung nach der Messung zweier Komponenten der Lesekompetenz eine Trainingswahlentscheidung zwischen dem FiLBY-2- und dem FiLBY-3-Training ermöglicht und dabei auch die Entscheidungssicherheit quantifiziert. Alle Ergebnisse werden inhaltlich mit Bezug auf die Lesekompetenzentwicklung und -förderung, sowie aus entscheidungstheoretischer Perspektive diskutiert.

Glossar

Folgende Abkürzungen und Notationen werden in dieser Arbeit verwendet.

Abkürzungen

z. B. = zum Beispiel

bzw. = beziehungsweise

etc. = et cetera

vgl. = vergleiche

IGLU = Internationale Grundschul-Lese-Untersuchung

PISA = Programme for International Student Assessment

BYLET = Bayerischer Lesetest

SLS = Salzburger Lese-Screening

Statistische Begriffe und Notationen

MIRT = Multidimensionale Item-Response-Theorie

KTT = Klassische Testtheorie

SEM = Strukturgleichungsmodell

ML = Maximum Likelihood

logL = Loglikelihood

MHRM = Metropolis-Hastings-Robinson-Monro

MAP = Maximum-a-Posteriori

CFI = Comparative fit index

RMSEA = Root mean squared error of approximation

BIC = Bayesian information criterion

$\widehat{\ }$ = geschätzter Wert

$\sim\sim$ = Kovarianz

$=\sim$ = Ladung

$\Sigma(X)$ = Kovarianzmatrix von X

NA = Not available (Wert nicht verfügbar)

Mathematische Notation

$\mathbb{E}$ = Erwartungswert

$\sum$ = Summe

$\prod$ = Produkt

$x^{\backprime}$ = x transformiert

e = Eulersche Zahl

log = logarithmiert

$|$ = gegeben

∇ = Gradient

$\int$ = Integral

X^{-1} = Inverse von X

$\subseteq$ = Teilmenge

inf = Infimum

max = Maximum

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Diese Arbeit befasst sich mit Entscheidungen. Im Alltag treffen wir rund 20 000 Entscheidungen jeden Tag (Rettig, 2012). Einige sind automatisiert und dringen kaum ins Bewusstsein vor, andere bereiten uns schlaflose Nächte. Das Gehirn verarbeitet dazu eine unzählige Anzahl an Informationen, verknüpft aktuelle Wahrnehmungen mit gespeicherten Erfahrungen und berücksichtigt das erwartete Ergebnis auf einer Vielzahl von Dimensionen (Gold & Shadlen, 2007). Was so intuitiv ganz wunderbar zu funktionieren scheint, stellt die Wissenschaft vor große Herausforderungen. Wirtschaftswissenschaftler:innen arbeiteten schon lange mit mathematisch motivierter Nutzenmaximierung (Gilboa, 2009), Machine-Learning Algorithmen drängen die Internet-Nutzer:innen zum Konsum vermeintlich auf sie zugeschnittener Inhalte (Gomez-Uribe & Hunt, 2015). In vielen Bereichen der angewandten Wissenschaften versucht man den menschlichen Entscheidungsprozess zu verstehen und zu beeinflussen. Ein weiterer Bereich, der von Entscheidungen geprägt ist, ist die Diagnostik. Mit diesem Themenfeld möchte sich diese Arbeit befassen. Der Begriff stammt aus dem Griechischen und umfasste ursprünglich die Feststellung oder Bestimmung von Krankheiten. In der Mitte des 19. Jahrhunderts wurde der Begriff "Diagnostik" jedoch ausgeweitet und beschreibt nun alle diejenigen Aktivitäten, die dazu dienen einen körperlichen oder geistigen Zustand zu beschreiben (Passow, 1852/2008).

Es soll nun einleitend eruiert werden, was Diagnostik in den Bildungswissenschaften ausmacht und wie diagnostische Entscheidungen in Forschung und Praxis getroffen werden. Anschließend soll ein statistisches Verfahren für diagnostische Treatmententscheidungen entwickelt werden, das den Prozess der Diagnostik unterstützen und den theoretischen Anforderungen an diagnostische Entscheidungen Rechnung tragen kann. Die theoretischen Überlegungen und Grundlagen, sowie die vorgestellten statistischen Methoden sind in weiten Teilen der Sozialwissenschaften anwendbar. Diese Arbeit ist jedoch im Bereich der empirischen Bildungsforschung angesiedelt. Alle Daten und Anwendungsbeispiele stammen aus den Daten der FiLBY-Studie (Fachintegrierte Leseförderung Bayern). Sie ist eine sehr umfangreiche bayernweite Evaluationsstudie, die die Wirksamkeit von Lesetrainings über einen Zeitraum von drei Jahren untersucht. Dem geneigten Leser, der geneigten Leserin sei es aber freigestellt, das Verfahren nach Belieben in andere Anwendungskontexte zu übertragen.

1.1 Diagnostik in Bildungswesen und Wissenschaft

Nicht nur im medizinischen, juristischen und psychologischen Kontext, auch im Schul- und Bildungswesen, spielt die Diagnostik eine zunehmend größere Rolle. Bis in die 1970er Jahre des vergangenen Jahrhunderts fand die Thematik der Leistungsmessung oder allgemeiner der pädagogischen Diagnostik so gut wie keine Aufmerksamkeit in der Wissenschaft. Auch in der Praxis war das Thema der Diagnostik kaum präsent. Noch in den 1970er Jahren zählte der deutsche Bildungsrat zwar die Leistungsbeurteilung, nicht aber die Leistungsdiagnostik zu den Aufgaben von Lehrkräften (Lukesch, 1998). Die Leistungsbeurteilung ist durch unstandardisierte Verfahren, wie eine subjektive Einschätzung oder die Zensur schriftlicher und mündlicher Leistungen gekennzeichnet. Im Gegensatz dazu umfasst die Leistungsdiagnostik auch standardisierte Verfahren. Zusätzlich untersucht die Leistungsdiagnostik auch Ursachen und Teilkompetenzen von schulischer Leistung. Sie ist daher grundsätzlich förderorientiert (Klauer, 1982). Die Perspektiven der pädagogischen und der pädagogisch-psychologischen Diagnostik sind Thema der folgenden Abschnitte.

In der pädagogisch-psychologischen Diagnostik nahm das Thema der Leistungsdiagnostik zu Beginn der 1970er Jahre zunehmend mehr Raum im wissenschaftlichen Diskurs ein (Eberwein, Schui, & Krampen, 2006). Es entstanden erste Theorien über pädagogisch-psychologische Diagnostik, die nun auch in eigenen Zeitschriften Gehör fanden. So wurde etwa 1981 die Reihe “Tests und Trends” gegründet (Hasselhorn, Schneider, & Marx, 2000). Sie stellt aktuelle Testverfahren für den Praxiseinsatz vor. Einen weiteren Entwicklungsschub verdankte die pädagogisch-psychologische Diagnostik schließlich der Verbreitung von internationalen Vergleichsstudien, allen voran der PISA Studie (Baumert, Stanat, & Demmrich, 2001). Durch sie wurde eine Welle von nationalen Bildungsstandserhebungen und ein öffentliches Interesse für Bildung und Leistungsmessung ausgelöst, welches bis in die heutigen 2020er Jahre anhält. So kommt es, dass Diagnostik heutzutage ca. 29% der Arbeitszeit von pädagogischen Psycholog:innen in Anspruch nimmt (Roth & Herzberg, 2008). Inhalte der Tätigkeiten umfassen dabei vor allem das Testen von Leistungsständen und der Leistungsfähigkeit (Schmidt-Atzert & Amelang, 2012), ergänzt um motivationale und Angstdiagnostik. Insgesamt liegt aber ein starker Fokus der pädagogisch-psychologischen Diagnostik auf Selektion. Zentrale Inhalte sind zum Beispiel die Intelligenzmessung zur Rechtfertigung einer Sonderbeschulung. Darüber hinaus zählt auch die Beschreibung von überdauernden Zuständen, wie die Diagnose von Lese-Rechtschreib-Störungen oder Schulangst zu den Aufgaben der pädagogisch-psychologischen Diagnostik.

Auch in der Pädagogik steht historisch die Selektion von Lernenden für höhere Bildungswege stark im Vordergrund. In den 1960er Jahren entwickelten sich ergänzend erste Einschulungstests, deren

Einsatz sich in der Schuleignungsfeststellung sich jedoch als wenig gewinnbringend erwies (Ingenkamp, 2007). Über die Einführung von Zensuren zu Selektionszwecken entwickelte sich die pädagogische Diagnostik zunehmend parallel zur pädagogisch-psychologischen Diagnostik. So wurden nun Zensuren auch zur Leistungsstandmessung herangezogen (Ingenkamp, 2007). Zurecht wurde und wird diese Praxis kritisiert und wirft die Frage auf, welche Rollen Pädagogik und pädagogische Psychologie im Themenfeld der pädagogisch(-psychologischen) Diagnostik jeweils zugeordnet werden sollen und welche Beiträge sie leisten können. Darüber hinaus, wird die pädagogische Diagnostik als Teil des pädagogischen Handelns von Lehrkräften verstanden. Nur durch ein kontinuierliches Monitoring der Lernprozesse seiner:ihrer Lernenden kann ein:e Lehrende:r die Lehrgeschwindigkeit und die Lerninhalte an die Schüler:innen anpassen (Jank & Meyer, 2002). So ist pädagogische Diagnostik im engeren Sinne automatisch Teil eines jeden Lehr-Lernprozesses und trägt zur professionellen Entwicklung von Lehrenden bei. Mit einer steigenden Anzahl an Förderangeboten und Lernprogrammen wird zunehmend auch das Handeln der Lehrkraft um die diagnostischen Entscheidungen der Auswahl der Lernmaterialien, Programme, oder allgemein Treatments ergänzt. Zusätzlich übernehmen automatisierte, computergestützte Lernprogramme mehr und mehr die diagnostischen Aufgaben der Lehrkraft und verlangen nach formalisierten diagnostischen Entscheidungen, die über eine intuitive, erfahrungsgeleitete und individuelle Einzelfalldiagnostik hinausgehen (Lee & Park, 2008). Dass diese Idee nicht neu ist, zeigt dabei exemplarisch eine Studie von Nitko und Hsu (1974), in welcher ein formalisierter Zugang zur Auswahl von Lernmaterialien gefordert wird.

1.2 Diagnostik in der empirischen Bildungsforschung

Die Bildungsforschung ist ein Bereich, der die verschiedenen bisher diskutierten diagnostischen Strömungen integriert und um eine von Forschungsinteressen geleitete Perspektive ergänzt. Im Rahmen von internationalen Vergleichsstudien, aber auch in der Evaluationsforschung werden standardisierte Verfahren zur Leistungsmessung eingesetzt. Zunehmend entsteht aus den Erkenntnissen dieser Forschungsprojekte der Anspruch, aus ihren Ergebnissen Handlungs- und Förderempfehlungen für die evaluierten Lehrkräfte abzuleiten. Dabei verschwimmen die Grenzen zwischen pädagogischer und psychologischer Diagnostik und der Fokus der Leistungsmessung und der Leistungssteigerung treten gleichberechtigt in den Vordergrund (Ingenkamp, 2007). Dass dieser Prozess nicht immer einfach ist, und oft Diskrepanzen zwischen der Evaluation von Maßnahmen und der Optimierung individuellen Lernens liegen, wird bereits bei Klauer (1982) diskutiert.

Diese Diskrepanz begründet sich unter anderem darin, dass Diagnostik in Panel- und Evaluationsstudien in der Regel kompetenzorientiert erfolgt, während zur Ableitung individueller Förderbedarfe

bisher auch eine prozessorientierte Diagnostik von Nöten ist (Lee & Park, 2008). Erkenntnisse aus Panel- und Evaluationsstudien werden dabei mehr als Basis für politische und verwaltungstechnische Entscheidungen betrachtet und nicht als Basis für eine individuelle Lernprozesssteuerung (Koeppen, Hartig, Klieme, & Leutner, 2008).

1.3 Ziele bildungswissenschaftlicher Diagnostik

Was sind also die Ziele der pädagogischen oder auch der bildungswissenschaftlichen Diagnostik? Übertragen aus der psychologischen Diagnostik ergibt sich als übergeordnetes Ziel

"Psychologisches und pädagogisches Wissen und psychologische Techniken bereit[zu]stellen, die dazu beitragen, (in Einzelfällen) praktische Probleme zu lösen" (Fisseni, 2004).

Auch ältere Definitionen der pädagogischen Diagnostik beinhalten durchaus die Zielvorstellung, von der Diagnostik als Werkzeug des Lehrenden, das dazu dient, geeignete Methoden auszuwählen. So schreiben etwa Reulecke und Rollett (1976):

"Diagnostik in schulischen Entscheidungssituationen hat den Zweck, Informationen zur Optimierung des pädagogischen Handelns zu gewinnen. Entsprechend unterscheidet man zwischen pädagogischer Diagnostik im engeren Sinn, die Planung und Kontrolle von Lehr- und Lernprozessen zum Gegenstand hat und pädagogischer Diagnostik im weiteren Sinn, die alle diagnostischen Aufgaben im Rahmen der Bildungsberatung umfasst."

Gleichzeitig grenzt sich pädagogische Diagnostik von der Bildungsforschung ab, die allgemeine Zusammenhänge aufzeigen will. Sie versteht sich dementgegen als Wissenschaft mit Fokus auf dem Einzelfall (Ingenkamp, 2007; Klauer, 1982).

Auch von anderen Autor:innen werden als konkretes Ziel pädagogischer Diagnostik die Selektion von Personen für verschiedene Einrichtungen des Bildungswesens und die Auswahl geeigneter Interventionsverfahren, oder die Steuerung des Unterrichtshandelns genannt (Schmidt-Atzert & Amelang, 2012).
Mit dem Fokus auf den letzten beiden Punkten gliedern etwa Schmidt-Atzert und Amelang (2012) die Ziele pädagogischer Diagnostik wie folgt:

- Selbst- und Fremdkorrektur falscher Lernergebnisse
- Erkennen von Lerndefiziten
- Bestätigung erfolgreicher Lernschritte
- Planung nachfolgender Lernschritte

- Motivierung durch Hinweise auf Lernerfolge und Steuerung des Schwierigkeitsgrads der nächsten Lernschritte
- Verbesserung der Lernbedingungen

Dass dabei die individuelle Lernsteuerung und die Evaluation der eingesetzten Methoden auf der Gruppenebene nicht trivial zu vereinbaren sind, wird häufig diskutiert. Nicht nur Pellegrino und Kollegen (2001) stellen fest, dass je nach Ziel der Beurteilung unterschiedliche Messinstrumente benötigt werden. Diese sind auch an die Beantwortung unterschiedlicher Forschungsfragen gekoppelt. Auch eine Studie von Huff und Goodman (2007) bestätigten, dass Lehrkräfte detaillierte, kontextbezogene Informationen über die Lernfortschritte ihrer Schüler:innen benötigen, um ihren Unterricht effizient anzupassen. So verwenden nur etwa 20-30% der Lehrkräfte in den Vereinigten Staaten Ergebnisse aus staatlich vorgeschriebenen oder kommerziellen Tests, um ihren Unterricht zu reflektieren. Als Folge wird in der Unterrichtspraxis zur Steuerung des Lernens fast ausschließlich auf unstandardisierte, subjektive Leistungsbeurteilung der Lehrkräfte vertraut (Huff & Goodman, 2007). Gleichzeitig zeigt ein Konglomerat an Studien (Anmarkrud & Bråten, 2012; Bremerich-Vos, Wendt, & Bos, 2017; Österbauer, Bachinger, Winter, Paasch, & Illetschko, 2020), dass Lehrkräfte in der Regel häufig nicht über basale diagnostische Fähigkeiten verfügen und Lehrmaterialien fast ausschließlich unabhängig von den Charakteristika der Kinder auswählen. Standardisierte Erhebungsinstrumente und in Evaluationsstudien erhobene Leistungsstände auch für den Unterrichtsalltag nutzbar zu machen, stellt also eine der aktuellen Herausforderungen der Bildungswissenschaften dar. Ähnlich der Forderung nach Präzisionsmedizin, anstelle von evidenzbasierter Medizin (Beckmann & Lew, 2016), sollte auch in den Bildungswissenschaften eine individuelle Gestaltung des Lernprozesses das Ziel sein und Diagnostik eines der Werkzeuge auf dem Weg dorthin.

1.4 Beschreibungsdimensionen von Diagnostik

Die pädagogische Diagnostik als gesamtes Forschungsgebiet wiederum ist komplex und nicht jedes Teilgebiet eignet sich gleichermaßen die im vorangegangenen Abschnitt formulierten Ziele zu erfüllen. Zwar definieren Klauer (1982): "Pädagogische Diagnostik ist das Insgesamt von Erkenntnisbemühungen im Dienste aktueller pädagogischer Entscheidungen." oder auch Rollett (1976)

"Unter (päd.) Diagnostik soll … zunächst die theoriegeleitete Datengewinnung und -reduktion im Rahmen eines gewichteten Entscheidungsverfahrens im Hinblick auf ein vorgegebenes Behandlungsziel verstanden werden."

Dennoch bleiben sie in Bezug auf konkrete diagnostische Fragestellungen abstrakt. Es zeigt sich

jedoch im pädagogischen Kontext ein vermehrtes Augenmerk auf dieser Zweiteilung des diagnostischen Prozess. Gleichzeitig lässt sich das Gebiet der Diagnostik genauer zerlegen als eine Zweiteilung in Datengewinnung und Datennutzung mit Hinblick auf das diagnostische Ziel. Daher soll zunächst der Themenkomplex Diagnostik auf die in dieser Arbeit interessierenden Dimensionen eingeschränkt werden. Hierzu kann die Diagnostik entlang drei Dimensionen beschrieben werden. Neben den Fragen, wer Gegenstand der Diagnostik sein soll, unterscheiden sich die diagnostischen Methoden darin, wie Erkenntnisse gewonnen werden, und schließlich auch darin, welche Instrumente dafür eingesetzt werden.

So wird zunächst zwischen Individual- und Umweltdiagnostik unterschieden. Bei ersterer steht das Verhalten und Erleben einzelner Personen im Vordergrund. Die Umweltdiagnostik konzentriert sich hingegen auf Verhalten in Interaktion mit anderen oder auf Gruppenprozesse (Ziegler & Bühner, 2012). Ein klassisches Instrument für Umweltdiagnostik stellt die Soziometrie dar. Gemeinhin bekannt ist zum Beispiel der soziographischer Test von Bullis-Seelmann (Engelmayer, 1968). Die Umweltdiagnostik soll jedoch in dieser Arbeit nicht von Interesse sein. Des Weiteren wird zwischen standardisierter und unstandardisierter Diagnostik unterschieden. Von standardisierter Diagnostik spricht man, wenn Verfahren in Durchführung, Auswertung und Interpretation reglementiert sind und an einer großen Stichprobe normiert wurden. Hierunter fallen Tests und strukturierte Interview- und Beobachtungsverfahren (Schmidt-Atzert & Amelang, 2012). Zur unstandardisierten Diagnostik gehören hingegen Verfahren, die an die zu untersuchende Person oder während des diagnostischen Prozess durch die Diagnostiker:innen angepasst werden können. Beispiele sind ein freies Interview oder die Auswertung einer autobiografischen Erzählung. Inferenzstatistische Verfahren beruhen auf der Schätzung von Populationsparametern, die auf Basis von Stichprobendaten geschätzt werden und unter vergleichbaren Bedingungen erhoben wurden. Da diese Arbeit auf Daten aus empirischen Evaluationsstudien basieren soll, wird im Weiteren der diagnostische Zugang auf standardisierte Verfahren eingeschränkt. Eine weitere Antwort auf die Frage wie Diagnostik ausgeübt wird, ergibt sich in der Weise, wie eine Erkenntnis in eine diagnostische Entscheidung überführt wird. Liegt die Erkenntnis in Form eines Messwerts vor, so kann dieser Wert bezüglich einer Entscheidungsgrenze beurteilt werden. Handelt es sich bei einer Entscheidungsgrenze um einen festen Wert, so spricht man von kriteriumsorientierter Diagnostik. Wird eine Entscheidungsgrenze in Bezug zu einer Populationsverteilung festgelegt, spricht man von normorientierter Diagnostik (Ziegler & Bühner, 2012). So ist zum Beispiel die Alkoholdiagnostik bei einer Polizeikontrolle kriteriumsorientiert. Denn ein:e betrunkene:r Fahrer:in verliert den Führerschein ab einer bestimmten Promillegrenze, ganz unabhängig davon, wie viel Blutalkohol die anderen Verkehrsteilnehmer:innen haben. Bei einer Intelligenzmessung hingegen wird das Ergebnis in Bezug zu den Ergebnissen gleichaltriger Personen

mit ähnlichem Bildungsniveau gesetzt. Bei der Leistungsmessung in den Bildungswissenschaften werden fast ausschließlich normorientierte Interpretationen angewandt. So ist es auch in dieser Arbeit der Fall. Damit ist nun der Themenbereich dieser Arbeit auf standardisierte, normorientierte Individualdiagnostik festgelegt.

1.5 Diagnostische Entscheidungen

Jede standardisierte normorientierte Individualdiagnostik endet dabei mit einer diagnostischen Entscheidung (Van Der Linden, 1987). Die diagnostische Entscheidung ist also das zentrale Ziel der Diagnostik und kann in unterschiedlichem Grad der Formalisierung getroffen werden. Dabei der werden nach Van der Linden (1987) vier Arten der diagnostischen Entscheidung unterschieden: placement decisions (Treatmententscheidungen), selection decisions (Selektionsentscheidungen), mastery decisions (Leistungsniveauentscheidungen) und classification decisions (Klassifikationsentscheidungen). Jede Art der diagnostischen Entscheidung verfügt über spezifische Eigenschaften ist durch ein Set an unterschiedlichen Prämissen gekennzeichnet. Hier erfolgt eine Einschränkung auf die Treatmententscheidungen.

1.6 Nicht formalisierte diagnostische Entscheidungen

Treatmententscheidungen können sowohl formalisiert als auch auf nicht formalisiert erfolgen. Das hängt maßgeblich davon ab, wie standardisiert der diagnostische Prozess im jeweiligen Anwendungsfall ist. Im Folgenden werden daher traditionelle, nicht oder nur wenig formalisierte diagnostische Entscheidungsverfahren formalisierten Ansätzen gegenübergestellt.

1.6.1 Diagnostische Entscheidungen als Ergebnis des diagnostischen Prozesses

Als Abschluss der diagnostischen Bemühungen wird eine diagnostische Entscheidung getroffen. Sie ist das Endergebnis eines diagnostischen Prozesses. Lehrbücher beschreiben einen prototypischen diagnostischen Prozess, der einzelne diagnostische Handlungen zusammenfasst und in einer diagnostischen Entscheidung oder einem diagnostischen Urteil mündet (z. B. Schmidt-Atzert & Amelang, 2012). Das Urteil wird oft mit der Beantwortung der ursprünglichen Fragestellung gleichgesetzt. Hierbei wird nur teilweise berücksichtigt, dass jede diagnostische Untersuchung fehlerbehaftet ist. Bei ungenügender Evidenz wird ein iteratives Verfahren von immer neuen diagnostischen Untersuchungen postuliert, das schlussendlich aber immer in der Beantwortung der Fragestellung resultiert. Dabei wird jedoch in der Regel nicht klar spezifiziert, wann eine Fragestellung als beantwortet gilt

und eine diagnostische Entscheidung als gerechtfertigt. Die Urteilsbildung selbst wird als Integration der diagnostischen Erkenntnisse beschrieben. Sie erfolgt durch den:die Diagnostiker:in und geschieht in einem von Reflexion und Erfahrung geprägten subjektiven Abwägungsprozess. Die diagnostische Entscheidung selbst wird als Antwort auf eine konkrete Fragestellung betrachtet. Die Fragestellung soll hier die Zurodnung eines Treatments zu einer Person sein (Schmidt-Atzert & Amelang, 2012). Dies impliziert, dass der:die Diagnostizierende nur in seltenen Fällen eine unbegrenzte Anzahl an Wahlmöglichkeiten für seine:ihre diagnostische Entscheidung hat. Geht es etwa darum, eine Schuleignung festzustellen, so bieten sich als diagnostische Entscheidung "schulfähig," "nicht schulfähig" und "auf Basis der Daten nicht zu beurteilen" an. Natürlich sind diese Attribute auch in kontinuierlicher Form denkbar, insofern das Ziel die Beschreibung der Schuleignung war. Während die Aussage, dass ein Kind "sehr schulfähig" oder "nur wenig schulfähig" zur Beschreibung der Schuleignung zwar ungewohnt aber nicht unplausibel erscheint, kann sie kaum handlungsweisend wirken. In der Praxis muss im gegebenen Beispiel eine dichotome Entscheidung getroffen werden. Das Kind kann entweder eingeschult oder zurückgestellt werden. Es gilt also eine Entscheidungsregel festzulegen, die auch als natürlicherweise kontinuierlich angenommene Zustände (z. B. Schuleignung) in eine dichotome Entscheidung (Einschulung: ja/nein) überführt. Eid und Petermann (2006) bemerken hierzu, dass die Entscheidungsfindung wissenschaftlichen Kriterien entsprechen müsse. Sie spezifizieren jedoch nicht weiter, was eben solche wissenschaftlichen Kriterien ausmacht und führen keine konkrete Kriterien an. Schmidt-Atzert und Amelang (2012) verweisen darauf, dass Untersuchungshypothesen explizit darzulegen sind, Verfahren auszuwählen, die möglichst spezifisch auf die Fragestellung zugeschnitten sind und die in kontrollierten Untersuchungsbedingungen zum Einsatz kommen. Des Weiteren wird die Erfassung der Prognosegenauigkeit bzw. der Validität der Diagnose als wünschenswert erachtet (Schmidt-Atzert & Amelang, 2012). Für psychometrische Verfahren führen Bühner und Ziegler (2012) Entscheidungsregeln ein, die auf Basis von Konfidenzintervallen und angenommenen Populationsverteilungen latenten Konstrukten Messwertkategorien zwischen unterdurchschnittlich und überdurchschnittlich zuordnen. Weitere auch entscheidungstheoretisch fundierte Entscheidungsregeln werden zwar als wünschenswert erachtet (Irtel, 1996), jedoch für die Praxis als bislang ohne breite Anwendung (Schmidt-Atzert & Amelang, 2012) diskutiert. Um also konkrete Stichprobendaten zur Verbesserung diagnostischer Entscheidungen nutzen zu können, muss neben den beschriebenen prozeduralen Vorgängen auch eine formalisierte Betrachtung von diagnostischen Entscheidungen erfolgen (Irtel, 1996). Die Formalisierung definiert, auf welche Art und Weise die Daten in den diagnostischen Prozess einfließen.

1.6.2 Unterrichtliches Handeln: Entscheidungen auf Basis subjektiver Expertenurteile

Die meisten diagnostischen Entscheidungen im Bildungskontext werden nach wie vor als subjektives Expertenurteil gefällt. Dabei entscheiden Lehrkräfte auf Basis von Unterrichtsbeobachtung, Schülerbeobachtungen und den Ergebnissen aus Lernzielkontrollen, welche Inhalte sie vertiefen, ob neuer Stoff eingeführt oder zusätzliche Erklärungen und Übungsphasen eingeschoben werden. Jeder der genannten Aspekte kann dabei als Treatment aufgefasst werden, das dazu dienen soll, das Lernen zu fördern oder zu ermöglichen. Die Lehrkräfte nehmen also im Unterricht die Rolle der Diagnostiker:innen ein, dessen Gestaltung als eine Abfolge von Treatmententscheidungen aufgefasst werden kann. Da Unterricht ein interaktives, analoges Geschehen ist, bietet sich eine standardisierte Entscheidungsfindung in der Praxis meist nicht an oder ist auch gar nicht möglich. Grundsätzlich schlagen zwar Grove, Zald, Lebow, Snitz und Nelson (2000) vor, immer dann auf statistische Urteilsbildung zurückzugreifen, wenn es eine empirisch ermittelte Verrechnungsvorschrift gibt. Da wie oben beschrieben in vielen Fällen eine solche empirisch ermittelte Verrechnungsvorschrift nicht vorliegt, legen die Diagnostiker:innen durch die Integration der Ergebnisse und Wahl der Entscheidungsregeln subjektiv fest, wie die diagnostischen Treatmententscheidungen gebildet werden.

1.6.3 Treatmententscheidungen auf Basis von Evaluationsforschung

Eine weitere - oft politisch initiierte - Art der diagnostischen Entscheidung ist die Auswahl von Unterrichtsinhalten oder Unterrichtsprogrammen. Dies geschieht in der Regel über die Aufnahme bestimmter Lerninhalte oder -ziele in die Lehrpläne. Auch eine solche Entscheidung ist eine diagnostische Treatmententscheidung und sie basiert nicht selten auf Daten von Evaluationsstudien (z. B. Slavin, 2002). Wird etwa ein Programm oder eine Methode übereinstimmend als dem herkömmlichen Unterricht überlegen betrachtet, so werden diese als neuer Unterrichtsstandard implementiert (Lee & Park, 2008). Da auch dies eine Form der diagnostischen Treatmententscheidung darstellt, wird in Studie 3 ausführlicher diskutiert welche Grundannahmen in ihr stecken.

1.7 Formalisierung von diagnostischen Entscheidungen

Im Kontrast zum Expertenurteil und zur politischen Treatmententscheidung gibt es auch in den Sozialwissenschaften verschiedene Ansätze nicht nur die Informationsgewinnung, sondern auch die diagnostischen Entscheidungen selbst zu formalisieren. Dabei ergibt sich der folgende Grundkonsens:

Diagnostische Messungen können eindimensional oder multidimensional sein. Für die diagnostische Entscheidung werden pro Dimension Entscheidungsregeln festgesetzt, die den Messwerten oder bestimmten Kombinationen von Messwerten dichotome Entscheidungen zuordnen. In der Regel wird ein eindimensionales, kontinuierliches Kriterium angenommen (z. B. ein Testwert), welches durch eine oder mehrere Entscheidungsgrenze(n) in zwei oder mehr Bereiche geteilt wird. Die Verrechnung der einzelnen Kriterien oder Messungen zu einem Entscheidungskriterium kann dabei reflexiv oder formativ erfolgen. Dabei bilden häufig auf ein Kriterium gefittete Regressionsanalysen die Basis des Entscheidungsmodells. Eine Bildung der Entscheidungsgrenzen erfolgt dennoch meistens durch ein Expertenurteil oder Erfahrungswerte. In den meisten Fällen ist das Urteil also als ein klinisches anzusehen, auch wenn eine Verrechnungsvorschrift verwendet wird (Ziegler & Bühner, 2012).

Eine alternative Herangehensweise bietet eine entscheidungstheoretische Betrachtung. Gleichwohl ist die Idee, den gesamten diagnostischen Prozess der Messung, Evaluation der Ergebnisse und Entscheidungsfindung auf die spezifische diagnostische Situation anzupassen, nicht neu (Irtel, 1996; Van Der Linden, 1987; Van der Linden, 1998; vgl. Van der Linden & Mellenbergh, 1978; Vrijhof, Mollenbergh, & Van den Brink, 1983). Bereits 1965 formulieren Cronbach und Gleser in ihrem viel zitierten Buch "Psychological Tests and Personnel Decisions" die Maxime, diagnostische Entscheidungen auf das Gerüst der Entscheidungstheorie zu stützen und ersuchen eine Ergänzung der Testtheorie um entscheidungstheoretische Prinzipien. Bevor in diesem Framework entwickelte Modelle vorgestellt werden, soll an dieser Stelle eine kurze allgemeinere Einführung in die Entscheidungstheorie erfolgen.

1.8 Entscheidungstheoretische Perspektive

Eine mögliche Bewertung von diagnostischen Entscheidungen bietet ein entscheidungstheoretischer Ansatz. Im Gegensatz zu einer dichotomen Entscheidungsbewertung (falsch/richtig), formuliert die Entscheidungstheorie für jede Kombination aus Kriterium und Entscheidungsalternative einen zugeordneten Nutzen (u) bzw. Verlust (l). Man nennt die Kriterien dabei Umweltzustände und bezeichnet sie mit θ. Die Entscheidungsalternativen nennt man Aktionen und bezeichnet sie mit a. Aus dieser Kombination lässt sich datenfrei für jede Kriteriumsausprägung eine oder mehrere optimale Entscheidungen ableiten. Eine optimale Entscheidung bedeutet, dass diejenige Aktion (a_i) ausgewählt wird, die ein vorgegebenes Optimalitätskriterium maximiert. Man betrachte hierfür das folgende Beispiel, in welchem entschieden werden soll, ob ein:e Schüler:in Nachhilfe erhalten soll. Die Entscheidung soll in Abhängigkeit eines Kriteriumszustands getroffen werden, der erklärt, warum der:die Schüler:in schlechte Leistungen erbringt. Die Tabelle 1, gibt an, wie die beiden

Tabelle 1: Beispielhafte Nutzentabelle

		Kriteriumszustand θ	
	Nutzen (u)	Leistungsdefizit aus Desinteresse	Leistungsdefizit wegen Krankheitstagen
Entscheidungsalternativen (a_i)	Nachhilfe	0	1
	keine Nachhilfe	0.5	0

Entscheidungsalternativen (Nachhilfe vs. keine Nachhilfe) mit einem fiktiven Gewinn (Nutzen) in Zusammenhang stehen.

In diesem Beispiel etwa, ist für desinteressierte Menschen keine Nachhilfe und für erkrankte Menschen Nachhilfe, die optimale Aktion. Gleichzeitig ist jedoch der Nutzen einer richtigen Entscheidung in den beiden Fällen unausgeglichen, so dass eine Fehlklassifikation eines:r Erkrankten als desinteressiert, schwerer wiegt (Verlust von 1), als eine Fehlklassifikation eines:r Desinteressierten als erkrankt (Verlust von 0.5). Eine reine Betrachtung der Raten verschiedener Fehlklassifikationen (Sensitivität, Spezifität, Richtig-positiv- und Richtig-negativ-Rate) reicht also unter entscheidungstheoretischer Perspektive zur Festlegung von Entscheidungsregeln nicht aus.

Die Evaluation der Entscheidung erfolgt in einer entscheidungstheoretischen Herangehensweise a priori. Sie besteht daraus, Personen mit möglichst hoher Sicherheit bestimmten Kriteriumsausprägungen zuzuordnen und die Entscheidungsregeln zu finden, die den fehlerbehafteten Messwerten diejenigen Entscheidungsalternativen zuordnen, welche den Nutzen maximieren bzw. den Verlust minimieren. Voraussetzung hierfür ist jedoch, dass sich für alle Kombinationen aus Merkmalsausprägung und Entscheidungsalternative Nutzen bzw. Verlustwerte formulieren lassen. Ist das Zielkriterium definiert und messbar, so lassen sich die Nutzen- bzw. Verlustwerte aus den Ergebnissen einer experimentellen Untersuchung ableiten (Dehejia, 2005).

Im obigen Beispiel muss also definiert sein, welchen Verlust ein falsch-negatives gegenüber einem falsch-positiven Ergebnis hat, um aus dieser Gewichtung eine optimale Entscheidungsgrenze abzuleiten. Hier könnte der Nutzen etwa einer Verrechnung aus Leistungspunkten in einem Schultest und investierter Zeit entsprechen. Im Experiment hätten desinteressierte und erkrankte Schüler:innen jeweils randomisiert Nachhilfe oder keine Nachhilfe erhalten und anschließend den Schultest absolviert. Erhält man nun für zukünftige Schüler:innen Informationen über die Ursachen ihrer schlechten Schulleistungen so kann anhand der Nutzentabelle abgeschätzt werden, ob sich Nachhilfe für diese Schüler:innen lohnt.

1.8.1 Psychologisch-pädagogische Ansätze

Erste Ansätze für die Auswahl von Unterrichtsmethoden eine entscheidungstheoretische Perspektive einzunehmen, finden sich bereits bei Crobach und Gleser (1965). Sie postulieren - dem Stand der berechenbaren Funktionen in dieser Zeit geschuldet - jedoch ausschließlich lineare Zusammenhänge zwischen Ausgangsleistung und Nutzen unter verschiedenen Treatments. Daraus ergibt sich in ihren Anwendungen maximal eine Entscheidungsgrenze zwischen den Treatmentpaaren. In Bezug zu verschiedenen Charakteristika zeigen sich diese frühen diagnostischen Entscheidungsmodelle als sehr anpassungsfähig. So können in ihnen bereits Testkosten, sequentielle Entscheidungen, Testvaliditäten und verschiedene Annahmequoten berücksichtigt werden. Auch werden Probleme der Bandbreite und Spezifität der Messungen diskutiert und bereits angenommen, dass der Zusammenhang zwischen Treatmentvoraussetzungen und dem Nutzen oder Payoff zwar vermutlich nicht immer linear ist, jedoch in der damaligen Zeit nicht anders zu schätzen war (Cronbach & Gleser, 1965).

Auch andere Autoren stellen in ihren Veröffentlichungen Verrechnungen auf entscheidungstheoretischer Ebene an, wie etwa Brogden (1949) oder Boudreau (1991). Ebenso Finney (1962), der auf Basis von einem zweistufigen Auswahlverfahren die optimalen Zulassungsquoten für eine universitäre Ausbildung bestimmte. Auch diese Modelle haben aber die Einschränkung, dass sie Linearität im Nutzen annehmen und Validitätsmatrizen benötigten. Validitätsmatrizen bestehen aus einer vollständigen Beschreibung der Kriteriumsverteilungen nach den verschiedenen Entscheidungen. Sie sind so bei Treatmentscheidungen schwer realisierbar, da per Definition nicht beobachtbar ist, wie sich Personen im nicht realisierten Treatment entwickelt hätten. Zudem wurde der Nutzen stets als finanzieller Nutzen operationalisiert, was gerade im Bildungskontext zu nur schwer zu schätzenden Nutzenfunktionen führt, da Lernfortschritte nur bedingt in finanzielle Gewinne zu überführen sind. Mit Ausnahme der gerade dargestellten Modelle, ergibt sich für Schmidt-Atzert und Amelang (2012) auch 60 Jahre nach Cronbach die Feststellung, dass entscheidungstheoretische Prinzipien in der psychologischen Praxis bislang keine breite Anwendung erfahren haben.

1.8.2 Statistisch-entscheidungstheoretische Ansätze

Statistisch und mathematisch motivierte Entscheidungsmodelle zur Treatmentwahl wurden im Rahmen der Psychometrieentwicklung (Van Der Linden, 1987) und im Bereich der Sozialplanung (Dehejia, 2005) entwickelt. Sie eröffnen Evaluationsstudien einen neuen Fokus. Anstelle der Untersuchung von mittleren Treatmenteffekten, rückt nun die Analyse des unterliegenden Entscheidungsproblems in den Vordergrund. Diese Modelle sind hoch flexibel und berücksichtigen nicht nur verschiedenste

Arten der Unsicherheit (Manski, 2004), sowie nicht lineare Zusammenhänge zwischen Kriterien und Nutzen, sondern leiten auch asymptotisch optimale Treatmententscheidungsregeln für verschiedene Nutzenfunktionen her (Hirano & Porter, 2009). Damit bieten sie eine fundierte Basis für die Schätzung und Anwendung komplexer entscheidungstheoretischer Prinzipien auf Daten aus experimentellen und quasi-experimentellen Evaluationsstudien. Gleichzeitig werden diese Modelle in den angewandten Wissenschaften noch nicht rezipiert, weswegen ihre Anwendung in der Regel auf frei zugängliche Datensätze der Sozialforschung und spezielle Fragestellungen aus dem Kontext der Kriminalstatistik begrenzt ist.

Dass eine Vernetzung durchaus zu fruchtbaren Ergebnissen führen kann, zeigt die Erprobung ähnlicher entscheidungstheoretischer Modelle in der Medizin. Hier etwa wird unter dem Stichwort der Präzisionsmedizin der Fokus bei der Treatmentwahl zunehmend auf individualisierten Entscheidungsregeln gelegt (Qian & Murphy, 2011).

1.9 Fazit

Bisher findet wenig Integration der pädagogischen, psychologischen und statistischen Entscheidungsverfahren statt. Gleichzeitig gibt es einen Bedarf, datengestützte und individuelle Treatmententscheidungen fällen zu können. Eine Möglichkeit ist daher, die Einbettung der Treatmententscheidung in den entscheidungstheoretischen Kontext auch im pädagogischen und bildungswissenschaftlichen Kontext zu entwickeln. Eine Möglichkeit hierfür stellt das in dieser Arbeit entwickelte Treatment Decision Model (TreaDeM) dar. Im folgenden Abschnitt wird es vorgestellt.

2 Das TreaDeM (Treatment Decision Model)

Eine diagnostische Entscheidung im Bildungskontext lässt sich in drei Bestandteile zerlegen. Vorab muss jedoch ein Zielkriterium festgelegt sein, zu dessen Maximierung die Gelingensbedingungen der Entscheidungsoptionen bekannt sein müssen. Ist es das Ziel, ein Treatment auszuwählen, so werden die Entscheidungsoptionen durch die möglichen Treatments definiert. Statistisch gesehen entspricht das Zielkriterium dem Nutzen, die Gelingensbedingungen den Umweltzuständen und die Treatments den Aktionen. Das heißt, es sollte gesichert sein, dass unterschiedliche Treatments bei unterschiedlichen Umweltzuständen vorteilhaft sind. Wenn es ein Treatment gibt, das in jeder möglichen Situation (d.h. zum Beispiel für alle Kinder und alle Kontexte) allen anderen Treatments überlegen ist, dann wird die diagnostische Entscheidung trivial. In der Theorie ist das zwar denkbar, in der Praxis jedoch wird im Grunde niemals ein Treatment für alle Akteure und alle Situationen optimal sein. Die Basis der Gelingensbedingungen ist die valide und reliable Messbarkeit der Bedingungen. Es ergibt aus praktischer Perspektive wenig Sinn, theoretische Gelingensbedingungen zu formulieren, die nicht messbar sind. Messbar werden die Gelingensbedingungen jedoch nur, wenn sie als operationalisierbare latente oder manifeste Konstrukte Bestand haben. Es ist also die Messung des Zielkriteriums (Nutzens) und der vermuteten Gelingensbedingungen (Umweltzustände), die die Basis jeder diagnostischen Entscheidung bilden. Deswegen befasst sich Studie 1 dieser Arbeit mit der messtheoretischen Modellierung. Erst wenn diese erfolgreich ist, kann ein Gelingensmodell gebildet werden. Entscheidungstheoretisch gesehen entspricht das Gelingensmodell der Bestimmung der Nutzenfunktion. Dabei gilt es, die verschiedenen Treatments unter den Voraussetzungen, situativer und personeller Eigenschaften mit Hinblick auf das Zielkriterium zu erforschen. Konkret gilt es die Fragestellung zu beantworten, welche der Treatments gegeben der individuellen Voraussetzungen und Eigenschaften einer Person zu welchem Erfolg (Nutzen) führen. Mit einer empirisch fundierten Herangehensweise an die Modellierung eines Gelingensmodells beschäftigt sich die zweite Studie dieser Arbeit. Erst im dritten Schritt kann das Vorgehen einer informierten Entscheidung diskutiert werden. Welche Annahmen über menschliches Entscheiden getroffen werden und wie diese sich auch auf diagnostische Entscheidungen auswirken, wird in der dritten Studie dieser Arbeit thematisiert.

Die Ausgestaltung der drei Bestandteile hängt wiederum mit den Charakteristika der Entscheidungssituation zusammen. Diese lässt sich danach kategorisieren, aus wie vielen Treatments gewählt wird, ob es Quoten gibt, die zur Auswahl eingehalten werden muss, und ob Test- und Treatmentkosten berücksichtigt werden sollen. Liegt keine Quote vor, so können beliebig viele Personen demselben Treatment zugeordnet werden. Zusätzlich muss definiert werden, ob es sich um eine sequentielle oder einmalige Entscheidung handelt und ob das Treatment adaptiv ist. Diese Arbeit fokussiert

sich auf den folgenden Fall: Fixe Treatments werden selektiert, es gibt keine Quote und die Test- und Treatmentkosten sind vernachlässigbar. Die Treatmententscheidung erfolgt in einem einstufigen Verfahren. Natürlich bietet das erarbeitete Framework aber die Möglichkeit anders geartete Entscheidungssituationen zu berücksichtigen.

Zusammenfassend ergeben sich drei relevante Bestandteile einer diagnostischen Entscheidung:

1. Eine Messung (s. Studie 1)
2. Eine Nutzenfunktion (s. Studie 2)
3. Eine Entscheidungsfunktion (s. Studie 3)

In den weiteren Teilen der Arbeit wird am Beispiel der Auswahl von einem aus zwei Lesetrainings gezeigt, wie sich Evaluationsstudiendaten nutzen lassen, um individuelle Treatmentempfehlungen abzuleiten.

3 Studie 1: Umsetzung und Anwendung des TreaDeMs - die Messung

In diesem Teil der Arbeit wird die Messung thematisiert. Dafür werden die psychometrischen Eigenschaften, insbesondere die Skalierbarkeit im Sinne einer testtheoretischen Modellierung des Bayerischen Lesetests (BYLET) untersucht. Hierzu wird per Kreuzvalidierung aus drei theoretisch plausiblen Modellen das passendste ausgewählt. Die Modellierung erfolgt sowohl mit Methoden der Item-Response-Theorie (IRT) als auch mit Methoden der Klassischen Testtheorie (KTT). Darüber hinaus werden die Ergebnisse der beiden Schätzverfahren miteinander verglichen und die Standardgütekriterien der Validität und der Reliabilität des BYLETs bestimmt.

3.1 Messung als Basis der diagnostischen Treatmententscheidung

Mit Blick auf die zentrale Fragestellung dieser Arbeit gilt: wenn man nichts über einen Gegenstand weiß, kann man auch keine sinnvollen Entscheidungen ihm bezüglich treffen. Entscheidungstheoretisch betrachtet, entspricht die Messung dem Informationsgewinn über die Umweltzustände. Gibt es keine strikt dominierende Aktion, kann die Entscheidung dadurch verbessert werden, etwas über die Umweltzustände zu lernen. Eine strikt dominierende Aktion bedeutet, dass sie unter allen Umweltzuständen besser ist als alle anderen Aktionen. Dass es bei der Leseförderung keine dominierenden Aktionen gibt, haben zahlreiche Studien gezeigt. Es ist bekannt, dass für unterschiedliche Stadien des Leselernprozess unterschiedliche Fördermaßnahmen unterschiedlich gut wirken (Harris & Hodges, 1995; Indrisano & Chall, 1995; Kuhn & Stahl, 2003). Es hilft also eine passende Leseförderma̧ßnahme auszuwählen, wenn man etwas über die Lesekompetenz des Kindes weiß, das unterrichtet werden soll.

Daher muss der erste Schritt zur Beantwortung der diagnostischen Fragestellung welches Treatment für das Kind ausgewählt werden soll, die Messung der Lesekompetenz sein. Im Folgenden wird zunächst das Thema Messen behandelt, bevor auf den Messgegenstand Leseverstehen eingegangen wird.

3.1.1 Die psychometrische Messung

Gemeinhin wird unter Messen das Zuordnen von einem numerischen Relativ zu einem empirischen Relativ verstanden. Dabei ist es das erklärte Ziel, das empirische Relativ möglichst exakt numerisch abzubilden (Bühner, 2011; Maranell, 2017). In der klassischen Herangehensweise unter pädagogisch-

psychologischer Perspektive werden Messungen hinsichtlich der Hauptgütekriterien Validität, Reliabilität und Objektivität beurteilt. Unter der in dieser Arbeit gewählten entscheidungstheoretischen Perspektive, treten neben Validität und Reliabilität vor allem der prädiktive Nutzen des Messwerts für die Entscheidung in den Vordergrund. Soll nämlich die Entscheidung auf den Messwert bedingt werden, so ist gerade die Interaktion aus Messwert und Aktion in Bezug auf das Optimalitätskriterium wünschenswert. Ist diese stark ausgeprägt und lässt sich die Optimalität von Treatments gut anhand der gewählten Messwerte unterscheiden, so spricht man von einem hohen differential Payoff (Cronbach & Gleser, 1965).

3.1.2 Gütekriterien der Messung unter entscheidungstheoretischer Perspektive

Wie im vorangegangenen Abschnitt beschrieben, sollen aus entscheidungstheoretischer Sicht die gemessenen Konstrukte prädiktiv für den Nutzen der zur Auswahl stehenden Treatments sein. Das heißt, es ist ein Messinstrument gesucht, das valide mit einem Zielkriterium, das den Nutzen abbildet, assoziiert ist. Das wiederum bedeutet, wenn die Veränderung eines Konstrukts selbst das Ziel des Treatments ist (z. B. die Steigerung einer Kompetenz) und diese Veränderung einer Kompetenz gemessen werden soll, so ist ein Messinstrument von Nöten, das im klassischen Sinne (inhalts)valide ist. Entscheidungen über Treatments auf Basis eines inhaltsleeren Konstrukts zu treffen, wäre uninteressant. Zudem muss das Messinstrument veränderungssensitiv sein, damit etwaige Veränderungen im Konstrukt sich auch in einer Veränderung in den Messwerten zeigen können. Darüber hinaus sollte die Veränderung auf dem Zielkonstrukt (zum Beispiel dem Trainingserfolg) wegweisend sein für ein weiter in der Zukunft liegendes Zielkriterium. So befähigt zum Beispiel eine hohe Lesekompetenz neben anderen Kriterien zu höherer Bildung, die wiederum zu gut bezahlten Berufen führt (Preston, 2006). Zusammenfassend bedeutet dies, die Messung soll ein Konstrukt valide abbilden, das in Zusammenhang mit einem gewünschten Zielkriterium steht oder das aus anderen Gründen für wertvoll erachtet wird (Cronbach & Gleser, 1965).

Es folgt auch, dass das Konstrukt so reliabel gemessen werden muss, dass Veränderungen messbar werden. Die Reliabilität kann dabei aus verschiedenen Gründen nicht im gesamten Leistungsbereich gleich groß sein. Zum einen ist zum Beispiel die Trennschärfe im Bereich der Schwierigkeit eines Items am höchsten (Baker & Kim, 2004). Die Schwierigkeiten der Items sind jedoch nur in Ausnahmefällen über den Itempool hinweg gleich verteilt. In der Regel werden immer die meisten Items eine mittlere Schwierigkeit aufweisen. Dies hängt schon damit zusammen, dass die meisten Tests so konstruiert werden, dass sie die Populationsleistung trennscharf abbilden. Da die Populationsleistung aber normalverteilt ist, ist die Leistungsdichte im mittleren Segment am höchsten, was

wiederum viele Items mit einer hohen Trennschärfe für den mittleren Leistungsbereich benötigt. Weiterhin gilt für das Framework der Item-Response-Theorie, dass die Schätzungen der Schwierigkeitsparameter genauer werden, je mehr Probanden sich in diesem Leistungsbereich befinden. Es gilt also bei der Konstruktion des Messinstruments zu berücksichtigen, in welchem Leistungsbereich es mit welcher Genauigkeit messen soll.

Bildet das Messinstrument in der Treatmententscheidung sowohl das Zielkonstrukt, als auch das Auswahlkriterium, so muss das Messinstrument veränderungssensitiv sein. Veränderungssensitiv bedeutet, dass das Messinstrument so genau messen muss, dass Fortschritte nicht innerhalb des Messfehlerbereichs liegen. Für das hier konstruierte Messinstrument reicht jedoch zunächst eine Veränderungssensitivität auf Gruppenebene. Das bedeutet, dass nicht für jede einzelne Person ein Fortschritt messbar sein muss. Tatsächlich wird an späterer Stelle gezeigt, dass eine hohe Reliabilität für die Einzelfalldiagnostik und damit auch für die Einzelfallentscheidung nur dann von hohem Interesse ist, wenn die Messwerte in der Nähe von Schnittstellen verschiedener Nutzenfunktionen liegen. Wenn es aber darum geht, Leistungsniveaus von ganzen Treatmentgruppen zu bestimmen, kann davon ausgegangen werden, dass sich die Messfehler einzelner Personen bei ausreichend großer Stichprobe ausmitteln (Hilbert et al., 2020).

Als Fazit der eben dargestellten Anforderungsanalyse ergibt sich, dass ein spezifisches Messinstrument mit nicht zu hoher Bandweite benötigt wird. Darüber hinaus ist eine Modellklasse von Nöten, die unterschiedliche Reliabilitäten schätzen kann. Dies kann über die Modellierung der Testergebnisse mit einem testtheoretischen Modell bewältigt werden (Bühner, 2011; Rost, 2004).

3.1.3 Testtheoretische Modelle

Testtheoretische Modelle sind der mathematische Weg, empirische Relative in numerische Relative zu übersetzen. Dabei überprüfen sie die Datenstruktur auf Systematiken, die für eine theoretisch angenommene inhaltliche Struktur sprechen. So können zum Beispiel die Gliederung des Konstrukts in einzelne Facetten oder Konstrukthomogenität getestet werden (Bühner, 2021; Rost & Spada, 1978). Es gibt zwei testtheoretische Schulen: die klassische Testtheorie und die Item-Response-Theorie. Beide Schulen umfassen unterschiedliche Analyseverfahren, die Informationen liefern, welche sich in der psychometrischen Modellierung gut ergänzen. Daher sollen beide Theorien im Folgenden kurz dargestellt werden.

3.1.3.1 Klassische Testtheorie (KTT) In der KTT stellt die Itemantwort selbst bereits eine Messung dar, die sich aus einem wahren Wert und einem Messfehler zusammensetzt. Die KTT

wird daher auch als Messfehlertheorie bezeichnet (Gulliksen, 2009). Eine Kernannahme der KTT ist, dass wahre Werte und Messfehler sowie die Messfehler untereinander unkorreliert sind. Als Modellierung als konfirmatorische Faktorenanalyse (KFA) mit einem Strukturgleichungsmodell (SEM) wird die Zusammensetzung von wahren Werten und Fehlern identifiziert. Dabei können bei ausreichender Modellanpassung Messeigenschaften angenommen werden - allerdings nur, wenn keine Korrelationen zwischen Fehlern und latenten Variablen oder zwischen den Fehlern selbst spezifiziert werden. Ein klassisches SEM stützt sich auf multiple Korrelationen und basiert daher auf starken Annahmen, wie Intervallskalenniveau, multivariate Normalverteilung und dem Fehlen von Multikollinearität (Byrne, 2013; Lei & Wu, 2007). Nicht viele Messinstrumente erfüllen diese Kriterien tatsächlich. Fundamentalist:innen wie Bond und Fox (2013), Linacre (1996), Panayides, Robinson und Tymms (2010) oder Michell (1997) warnen daher eindringlich davor, davon auszugehen, dass vordefinierte Skalenformate automatisch Daten produzieren, die den Anforderungen der beabsichtigten Skalenniveaus entsprechen, wie sie von Stevens (1946) eingeführt wurden. So kann etwa nicht automatisch davon ausgegangen werden, dass fünfstufige Likert-Skalen Daten auf Intervallskalenniveau produzieren. Erst durch Erweiterung der Schätzmethoden unter anderem durch Muthén (1998), konnten SEM-Modelle auch auf dichotome oder allgemein kategoriale Daten angewendet werden. Ein Vorteil von KTT-Analysen ist, dass mögliche Fehlspezifikationen der latenten Variablen durch Analysen mit Modifikationsindizes erkannt werden können (Byrne, 2013).

3.1.3.2 Item-Response-Theorie (IRT) In der Item-Response-Theorie (IRT) wird die Messung als eine logistische, probabilistische Beziehung zwischen der Schwierigkeit des Items und der Fähigkeit der Person verstanden. Die zugrunde liegende Idee besagt, dass eine steigende Fähigkeit einer Person zu einer höheren Wahrscheinlichkeit führt, ein einzelnes Item zu lösen. In den strengsten Messmodellen sind dabei die Personenfähigkeit und die Itemschwierigkeit die einzigen Prädiktoren der Wahrscheinlichkeit, ein Item zu lösen (Bond & Fox, 2013; Michell, 1997). Komplexere IRT-Modelle können auch unterschiedliche Trennschärfen von Items, Raten oder eine geclusterte Personenstruktur berücksichtigen. Als Erweiterung der eindimensionalen IRT-Modelle wurden in den 1970er und 1980er Jahren multidimensionale IRT, die sogenannten MIRT-Modelle entwickelt. Sie umfassen ähnlich den aus der KTT bekannten konfirmatorischen Faktorenanalysen mehrere latente Dimensionen und erweitern damit die Möglichkeit der IRT, auch komplexere Messmodelle zu schätzen (Mulaik, 1972; Reckase, 1972; Sympson, 1978; Whitely, 1980). Die Schätzung von MIRT-Modellen erfolgt in der Regel likelihoodbasiert, so dass verschiedene Modelle anhand von globalen Fit Indizes wie dem BIC verglichen werden können (Reckase, 2009).

IRT-Modelle unterscheiden sich in zwei wesentlichen Punkten von Modellen der klassischen Test-

theorie. Erstens wird zwischen den Personenparametern und den Itemantworten kein linearer, sondern ein logistischer Zusammenhang angenommen. Zweitens wird der Einfluss der Items auf das Lösungsverhalten vom Einfluss der Personenfähigkeit in der Schätzung getrennt. Dieser letzte Punkt ist für die strukturellen Eigenschaften der Testmodelle von großer Bedeutung (Irtel, 1996). In eindimensionalen IRT-Modellen können so die Itemparameter getrennt von den Personenparametern geschätzt werden. In den MIRT-Modellen entfällt diese Eigenschaft, da hier die Schätzung der Itemparameter und der Personenparameter abwechselnd aber in jeder Iteration der Schätzung hintereinander erfolgt (Cai, 2010; Reckase, 2009). Nach dieser methodischen Einführung in das Messen widmet sich der folgende Abschnitt dem Messgegenstand des Leseverstehens.

3.2 Der Messgegenstand Leseverstehen

Das Leseverstehen entsteht aus der Interaktion eines Menschen mit Text (Groeben, 1982). Da beide Elemente, Mensch und Text, zum Konstrukt Leseverstehen beitragen, sollen auch aus theoretischer Sicht beide Seiten beleuchtet werden. Einmal wird also die Perspektive der Lesenden als Kompetenzansatz dargestellt. Anschließend wird das Leseverstehen aus der Perspektive der Textschwierigkeitsforschung präsentiert.

Um das große Forschungsgebiet des Lesens einzugrenzen, sollen hier nur Prozess- und Teilkompetenztheorien diskutiert werden, die sich auf solche Prozesse beziehen, die dem Lesen inherent sind. Hierzu zählt etwa das Dekodieren, die Automatisierung des Dekodierens und Inferenzleistungen beim Leseverstehen. Es sollen keine Prozesse oder Fähigkeiten betrachtet werden, die dem Lesen inherente Prozesse ermöglichen. Konkret ist der Autorin bewusst, dass Intelligenz (Cain, Oakhill, Barnes, & Bryant, 2001; Catts, Adlof, & Weismer, 2006), (verbales) Gedächtnis (Perfetti, Marron, & Foltz, 1996; Swanson, Cochran, & Ewers, 1989) und Aufmerksamkeit (Ghelani, Sidhu, Jain, & Tannock, 2004) auch und zum Teil maßgeblich Einfluss auf die Lesekompetenz nehmen (Cutting & Scarborough, 2006). Sie tragen jedoch zur Beantwortung der Frage, was Leseverstehen ist und wie man es messen kann, wenig bei. Es soll hier also keine Störungsperspektive eingenommen werden und von gesunden Lesenden ausgegangen werden. Das heißt, es soll beim Lesen vorausgesetzt gesetzt werden, dass ein ausreichendes Maß an Intelligenz, Merkfähigkeit und Aufmerksamkeit gegeben ist.

3.2.1 Lesen als komplexes Zusammenspiel von Teilkompetenzen

In den folgenden zwei Abschnitten werden verschiedene Theorien zur Lesekompetenz vorgestellt, welche das theoretische Fundament des in dieser Studie modellierten Lesetests bilden. Im Prinzip

unterscheiden die Theorien sich darin, in wie viele Teilkompetenzen sie das Leseverstehen unterteilen und ob sie einen statischen kompetenzorientierten oder einen dynamischen, strukturellen Ansatz wählen. Lesen wird in vielen Theorien als eine komplexe Kulturfähigkeit verstanden und unter anderem neurowissenschaftlich untersucht. Hier soll allerdings keine technische Betrachtung der Leseprozesse, die vor allem im Bereich der kognitiven Neurowissenschaften und der Linguistik untersucht werden, dargestellt werden (Kintsch, 1988; Richter & Christmann, 2002). Dort spielt das Dekodieren von Schriftzeichen eine wesentliche Rolle. Dass für das Leseverstehen ein Zusammenspiel dieser primär visuellen bottom-up und erfahrungsbasierter top-down Prozesse essentiell ist, darüber besteht viel Einigkeit. Weniger Einigkeit besteht in der kompetenzbasierten Betrachtung des Lesens. Daher sollen hier verschiedene Theorien vom Lesen als Zusammenspiel einzelner Kompetenzen im Sinne von Klieme, Leutner und Weinert (2006) und (2001) diskutiert werden.

3.2.1.1 Einfache Theorien zur Lesekompetenz Als erstes wird der Ansatz der sogenannten "Simple View of Reading" (Hoover & Gough, 1990) vorgestellt. Nach der Simple View of Reading besteht die Lesekompetenz aus zwei interagierenden Komponenten: Dekodieren und Sprachverständnis. Das Sprachverständnis wiederum besteht aus Gliedern, Inferieren und der Bildung eines Diskurses über das Gelesene. In der Simple View of Reading wird das Sprachverständnis nicht als Teilkompetenz des Lesens aufgefasst. Vielmehr ermöglicht die Fähigkeit des Dekodierens, das Sprachverständnis von geschriebenem Text. Dass das Dekodieren und das Sprachverständnis interagieren, zeigt sich laut der Simple View of Reading im Leselernprozess (Catts et al., 2006). Dieser sei am Anfang durch das Dekodieren dominiert. Das bedeutet, dass in einem frühen Leselernstadium das Leseverstehen maßgeblich von der Dekodierfähigkeit abhängt, jedoch nur wenig vom Sprachverständnis. Mit zunehmender Lesefähigkeit dreht sich der Einfluss der beiden Teilfähigkeiten auf die Leseleistung um. Das Leseverstehen wird nun maßgeblich durch das Sprachverständnis bestimmt. Weitere Stufen oder Teilfähigkeiten werden in dieser Theorie abgelehnt, weil sie als Teilkompetenzen des Sprachverständnisses nicht mehr als Teilkompetenz des Leseverstehens selbst aufgefasst werden. Begründet wird dieser Standpunkt damit, dass Menschen auch Texte verstehen können, ohne Lesen zu können. So können etwa Personen mit Dyslexia Texte verstehen, wenn sie diese anhören anstelle sie zu lesen. Auch empirisch lassen sich dabei die beiden Teilkompetenzen differenzieren. Durch verschiedene Messungen des Textverständnis, Sprachverständnis, der phonologischen Verarbeitung und der Dekodierfähigkeit konnten etwa bei Catts und Kollegen (2006) Subpopulationen von Lesenden identifiziert werden, welche entweder nur in den Dekodierfähigkeiten oder nur im sprachlichen Textverständnis eingeschränkt waren.

3.2.1.2 Mehrstufige Modelle der Lesekompetenz Neben der Simple View of Reading entstanden auch Theorien zur Lesekompetenz, welche der Lesekompetenz Fähigkeiten zuordnen, welche über das reine Dekodieren von Wörtern hinausgehen. Solche mehrstufigen Modelle der Lesekompetenz zerlegen das Lesen in die aufeinander aufbauenden Teilkompetenzen der Worterkennung sowie des Leseverstehens auf Satz- und Textebene (Van Dijk & Kintsch, 1983). Kompetenzen werden dabei als komplexe, trainierbare Fähigkeitskonstrukte verstanden, die es Individuen in spezifischen Kontexten ermöglichen, lebensnahe Anforderungen zu bewältigen (Klieme & Leutner, 2006; Connell, Sheridan, & Gardner, 2003; Koeppen et al., 2008). Einige Autor:innen beschreiben das Lesen als Fähigkeit, die durch aufeinander aufbauende Teilkompetenzen charakterisiert wird. Diese ermöglichen das Erfassen von Text auf unterschiedlichen Niveaus (Chall, 1983; Indrisano & Chall, 1995). In den internationalen Vergleichsstudien wie IGLU oder PISA setzt sich zur Lesekompetenzmessung dabei zunehmend ein fünfstufiges Kompetenzmodell durch. Laut diesem Kompetenzmodell entwickelt sich das Leseverstehen auf fünf von einander abgrenzbaren Kompetenzstufen. Die erste Stufe umfasst die Kompetenz der Worterkennung. Lesende, die sich auf Kompetenzstufe I befinden, können also einzelne Worte lesen und diesen eine semantische Bedeutung zuordnen. Stufe I wird in der Regel spätestens mit der Beendigung des Erstleselehrgangs abgeschlossen und stellt eine absolute Mindestvoraussetzung für jedes weitere Leseverstehen dar. In Stufe II entwickeln Lesende die Fähigkeit, Sätze und deren Inhalt zu erlesen und können einzelne Informationen wortgetreu oder auch sinngemäß entnehmen. Stufe III beschreibt darüber hinaus Leseverstehenskompetenzen, die auf Inferenzen beruhen. Inferenzen, die Lesende auf Kompetenzstufe III ziehen können, setzen zumeist nur einzelne Satzteile oder einzelne Sätze miteinander in Beziehung. Eine abschnittsweise Inferenz, bei der Informationen über mehrere Sätze hinweg integriert werden müssen, werden der Kompetenzstufe IV zugerechnet. In Stufe IV müssen die Lesenden zudem Vorwissen über Textstrukturen, oder Handlungswissen in den Prozess einbringen. Zum Beispiel können Lesende auf Stufe IV aus der Beschreibung typischer freundschaftlicher Verhaltensweisen erschließen, dass zwei Charaktere miteinander befreundet sind, ohne dass die Freundschaft explizit im Text benannt wird. In Stufe V gelingt es Lesenden zusätzlich, auch Weltwissen mit den gelesenen Informationen zu verknüpfen. So können sie sich ein sogenanntes mentales Modell eines ganzen Textes bilden, das auch eine Bewertung von Texten auf der Metaebene ermöglicht. Hierunter fallen sowohl eine kritische Bewertung als auch eine Einordnung des Texts innerhalb verschiedener Textsorten. Auch die Reflexion über die Funktion einzelner Textelemente oder -abschnitte wird von Stufe V abgedeckt (Artelt, Stanat, Schneider, & Schiefele, 2001; Bremerich-Vos et al., 2017).

3.2.1.3 Strukturelle Modelle von Lesekompetenz

Neben Teilkompetenzmodellen lohnt sich mit Hinblick auf die Lesekompetenzmessung auch eine Betrachtung von strukturellen Modellen der Lesekompetenz. Strukturelle Modelle von Lesekompetenz wie zum Beispiel das Direct and Inferential Mediation Model (DIME-Modell) (Cromley & Azevedo, 2007) zerlegen die Lesekompetenz nicht in Teilkompetenzen, sondern zeigen die angenommenen Verbindungen zu ähnlichen Kompetenzen auf. Das DIME-Modell wurde auf Basis sehr vieler empirisch gefundener Zusammenhänge des Leseverstehens mit anderen Konstrukten entwickelt. So zeigte sich, dass Strategien, Vokabular und Vorwissen, logisches Schlussfolgern und das Wortlesen (Leseflüssigkeit) miteinander aber auch in direkter Verbindung zum Leseverstehen stehen.

Strukturelle Modelle von Lesekompetenz, wie das in Abbildung 1 dargestellte DIME-Modell, liefern daher für die Lesekompetenzmessung wichtige Informationen zur Erfassung der konvergenten und divergenten Validität. Nimmt man etwa an, dass die Leseverstehensleistung auch von der Leseflüssigkeit abhängt, so sollten die Testleistungen aus einem Leseverstehenstest mit der eines Leseflüssigkeitstests substanziell zusammenhängen. Jedoch sollte die Varianz im Leseverstehen nicht vollständig durch die Leseflüssigkeit erklärbar sein. Zudem zeigt das DIME-Modell auf, dass auch das Vorwissen einen substanziellen Einfluss auf das Leseverstehen nimmt. Möchte man also einen Leseverstehenstest entwickeln, dessen Ergebnisse möglichst nicht vom Vorwissen der Getesteten abhängen, so sollte auf ein Thema zurückgegriffen werden, zu dem alle Lesenden entweder kein Vorwissen oder mit hoher Sicherheit sehr viel Vorwissen besitzen. Da ersteres leichter zu realisieren ist, bietet es sich bei der Testkonstruktion an, mit Science-Fiction-Texten zu arbeiten, da ausgedachte Welten für alle Lesenden gleichermaßen unbekannt sind.

3.2.2 Zusammenfassung

Betrachtet man die Seite der Lesenden, so ergeben sich aus allen theoretischen Betrachtungen verschiedene Teilkompetenzen, die zum Leseverstehen beitragen. Die Unterschiede in den vorgestellten Theorien liegen dabei hauptsächlich in der Anzahl der Teilkompetenzen und ihrer Interaktionen in Bezug auf das Leseverstehen. Während die Simple View of Reading das Leseverstehen in das Dekodieren und das Sprachverständnis zerlegt, nimmt das in den internationalen Vergleichsstudien verwendete fünfstufige Modell eine feinere Unterteilung vor, indem verschiedene Niveaus der zum Leseverstehen benötigten Inferenzen festgelegt werden. Der gemeinsame Nenner ist aber immer, dass zum Leseverstehen Wörter dekodiert werden müssen, Dekodiertes miteinander verknüpft und Dekodiertes mit Vorwissen in Bezug gesetzt werden muss. Daher sollte ein Leseverstehenstest auch erfassen, ob Lesende die genannten Teilkompetenzen beherrschen. Wie leicht dies gelingt, hängt je-

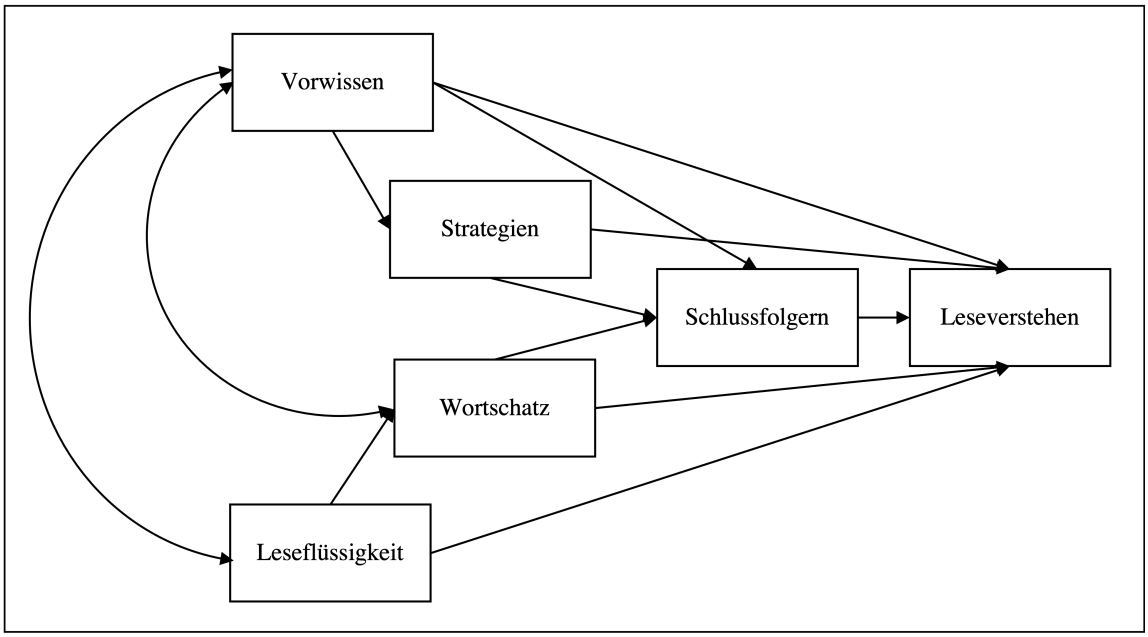


Abbildung 1: DIME-Modell (nach Cromley und Acevedo, 2007)

doch nicht nur von der Kompetenz der:s Lesenden ab, sondern auch von Merkmalen des gelesenen Textes.

3.2.3 Die Bedeutung der Textschwierigkeit

Daher soll nun die zweite Seite, die Seite des Texts betrachtet werden. Auch die Textschwierigkeit kann auf verschiedenen Ebenen betrachtet werden. Während die Literaturwissenschaften primär Antworten auf die Frage nach der Passung von Lesealter und Inhalten, Handlungskompositionen und der sprachlichen Gestaltung sucht, nimmt die Psycholinguistik stärker grammatikalische und sprachwissenschaftliche Aspekte der Textschwierigkeit in den Fokus (Groeben, 1982). In dieser Betrachtung, scheint sich die Textschwierigkeit in ähnliche Teilkomponenten zu zerlegen, wie es auch die Lesekompetenz von Lesenden tut. So zeigte sich etwa in einer Studie von McNamara, Graesser, Cai, & Kulikowich (2011), dass sich die Kodierungen verschiedener Texte mittels Hauptkomponentenanalyse in die Faktoren der Narrativität, syntaktischer Einfachheit, Wortpassung, sowie in referenzielle Kohäsion und tiefe Kohäsion zerlegen ließen. Dafür hatten verschiedene Expert:innen Texte in Bezug auf einzelne Kriterien bewertet. Diese Bewertungen wiederum dienten dann als Datengrundlage für die Hauptkomponentenanalyse. Daher erscheint es auch plausibel, dass sich etwa bei Amendum, Conradi und Hiebert (2018) ein Zusammenhang zwischen der Textschwierigkeit und dem Leseverstehen zeigte. Als dritte Betrachtungsebene ergibt sich die Operationalisierung der Textschwierigkeit anhand von Obeflächenmerkmalen von Texten, wie die Wort- oder Satzlänge. Diese Operationalisierung ist bereits in kommerziellen Angeboten wie dem Lexile Framework

(Stenner, 1996), aber auch in frei verfügbaren Angeboten wie dem LIX (Lenhard & Lenhard, 2014-2017) oder RATTE (Wild & Pissarek, 2018) implementiert. Obwohl so Textschwierigkeit schnell und systematisch erfasst werden kann, spielt die Messung und Manipulation der Textschwierigkeit für Unterrichtsmaterialien oder Lesetests bisher eher eine untergeordnete Rolle (Benjamin, 2012; Mesmer, 2008).

3.2.4 Zusammenfassung

Zusammenfassend lässt sich sagen, dass die Textschwierigkeit ein das Verständnis beeinflussender Faktor ist. Daraus lässt sich gerade für die Konstruktion von Leseverstehenstests ableiten, dass die Textschwierigkeit der verwendeten Texte geprüft und berücksichtigt werden sollte. Eine erfolgreiche Lesekompetenzmessung sollte also die verschiedenen Teilkompetenzen des Leseverstehens abbilden. Zudem sollte sie Texte einsetzen, die in Hinblick auf die Zielgruppe über eine angemessene Textschwierigkeit verfügen.

3.2.5 Messung von Leseverstehen

In den letzten Abschnitten wurde vorgestellt, wie man Lesekompetenz definieren kann und aus welchen Teilkompetenzen sie sich zusammensetzt. Im Folgenden soll es nun um die Messung von Kompetenzen gehen. Dafür soll erst dargestellt werden, wie man Kompetenzen im Allgemeinen messen kann und dann, was man daraus für die Messung der Lesekompetenz im Speziellen ableiten kann.

3.2.5.1 Allgemeines zur Messung von Kompetenzen Wenn es um die Messung von Kompetenzen im Allgemeinen geht, unterscheidet man zwischen Kompetenzstufenmodellen und Kompetenzstrukturmodellen (Hartig & Klieme, 2006; Klieme & Leutner, 2006). Laut Koeppen und Kollegen (2008) sollen valide Kompetenzmessungen auf theoretisch fundierten und empirisch getesteten Kompetenzmodellen basieren. Diese Modelle müssen die interne Struktur von Kompetenzen in Form von spezifischen Grundfertigkeiten und Fähigkeiten darstellen, verschiedene Kompetenzniveaus mit bereichsspezifischen Leistungen beschreiben und die Veränderungen in Lern- und Entwicklungsprozessen berücksichtigen (Koeppen et al., 2008). Die Empirie zeigt jedoch, dass die aktuellen Messinstrumente dies häufig noch nicht umsetzen. Dies kann man am Beispiel der Kompetenzmessung in den internationalen Vergleichsstudien zeigen. Hier spielt die Messung von Kompetenzen eine zentrale Rolle (Baumert et al., 2001). So wurden in PISA sowie in IGLU (Bremerich-Vos et al., 2017) zwar Kompetenzstufenmodelle postuliert, die Aufgaben jedoch nicht a priori zu bestimmten Kompetenz-

stufen zugeteilt oder mit Hinblick auf diese konstruiert. Stattdessen wurden Gesamtscores aus allen Items gebildet, welche dann post hoc über eine Grenzwertbildung in Kompetenzstufen unterteilt wurden. Tatsächlich wurde diese Herangehensweise bisher in allen getesteten Lernbereichen gewählt, sowohl für das Leseverstehen (Artelt et al., 2001), die mathematischen Fähigkeiten (Klieme, Neubrand, & Lüdtke, 2001), die Naturwissenschaften (Prenzel, Rost, Senkbeil, Häußler, & Klopp, 2001) als auch für das allgemeine fachunabhängige Problemlösen (Dossey, Hartig, Klieme, & Wu, 2004). Es wurde also bereichsunabhängig eine Vorgehensweise gewählt, die aus theoretischer Sicht nicht zufriedenstellend ist. Denn sind aus theoretischer Sicht Teilkompetenzen oder Kompetenzstufen identifizierbar, so sollten Aufgaben bereits a priori spezifisch für die jeweilige Teilkompetenz oder Kompetenzstufe konstruiert und ihr zugeordnet werden. Es ist also wünschenswert, bei zukünftigen Messinstrumenten die Zuteilung der Testitems zu den Kompetenzstufen vorab festzulegen. So kann diese Zuordnung von Items zu Teilkompetenzen oder Kompetenzstufen empirisch überprüft werden (Koeppen et al., 2008). Zur empirischen Überprüfung bieten sich wiederum psychometrische Modelle an. Es muss also als wissenschaftlicher Standard gelten, Kompetenzen wie das Leseverstehen mit einem standardisierten Testverfahren zu erheben.

3.2.5.2 Bisherige Lesekompetenzmessung Im Gegensatz zur anderen Bereichen des Deutschunterrichts hat sich im Bereich der Lesekompetenzmessung bereits eine formelle, standardisierte Diagnostik etabliert. Der deutsche Markt bietet Testverfahren wie ELFE, ELFE II (Lenhard, Schneider, Lenhard, & Schneider 2018), die VSL (Walter, 2013) und eine Anzahl weiterer Verfahren für den Einsatz zur Einzelfalldiagnostik des Leseverstehens an. Diese Verfahren überzeugen zwar in ihren testtheoretischen Gütekriterien, eine Orientierung an Kompetenzstufen wurde aber bisher bei ihrer Konstruktion nicht berücksichtigt. Auch eine bewusste Berücksichtigung und Manipulation der Textschwierigkeit ist in bisherigen Leseverstehenstests nicht vorhanden. Darüber hinaus weisen viele verfügbare Testverfahren ungünstige Nebengütekriterien insbesondere Probleme bei der Test-Fairness und dem Anwendungsbereich auf. Dabei wird kritisiert, dass Testtexte inhaltlich für Mädchen ansprechender sind als für Jungen (Coles & Hall, 2002). Die gemessenen Unterschiede in der Lesekompetenz können also auf Unterschieden in der konkreten Testbearbeitungsmotivation beruhen (Oakhill & Petrides, 2007). Zudem sind die Testtexte häufig Sachtexte, wodurch das Vorwissen stark zum Tragen kommt (Cromley & Azevedo, 2007; García & Cain, 2014). Schlussendlich gibt es auch aktuell nur wenige Verfahren, welche den Übergang zwischen Primarstufe und Sekundarstufe abdecken. So können Leistungsentwicklungen zwischen der vierten und siebten Klasse in der Regel nur schwer abgebildet werden (Galuschka, Rothe, & Schulte-Körne, 2015). Aus diesem Grund wurde ein neuer Leseverstehenstest - der Bayerische

Lesestest (BYLET) - entwickelt. Er wird im folgenden Abschnitt vorgestellt.

3.2.6 Der Bayerische Lesetest (BYLET)

Der Bayerische Lesetest präsentiert den Kindern eine Fantasiegeschichte, welche in vier Abschnitte unterteilt ist. Die Abenteuerwelt des Science-Fiction bietet den Vorteil, dass sie vorwissensarm ist und Jungen wie Mädchen gleichermaßen anspricht (Merisuo-Storm, 2006). Es ist bekannt, dass sowohl Vorwissen als auch thematisches Interesse sich positiv, aber auch differentiell auf die Leseverstehensleistung auswirken (Baldwin, Peleg-Bruckner, & McClintock, 1985). Die Leseleistung von Mädchen hängt weniger stark von ihrem thematischen Interesse am Text ab (Oakhill & Petrides, 2007). Deswegen wurde mit der Themenwelt des Science-Fiction versucht einer durch Interesse und Vorwissen hervorgerufenen Verzerrung in der Messung des Leseverstehens entgegenzuwirken. Die vier Textabschnitte verfügen über eine zunehmende Textlänge sowie Textschwierigkeit. Die Textschwierigkeit wurde dabei mit RATTE (Wild & Pissarek, 2018) gemessen und ist so gestaltet, dass der Gsmog Werte zwischen zwei und fünf annimmt. Der Test umfasst insgesamt 20 Multiple-Choice-Fragen. Nach der Lektüre jedes Abschnitts werden vier Fragen beantwortet, welche auf den Lesekompetenzniveaus II - IV des IGLU-Kompetenzstufenmodells angesiedelt sind. Abschließend beantworten die Kinder vier zusätzliche Multiple-Choice-Fragen, welche sich auf den ganzen Text beziehen und mit allen anderen Items gemeinsam das Kompetenzniveau V abbilden. Durch das geschlossene Fragenformat wird ausgeschlossen, dass Defizite im Schreiben bei der Messung des Leseverstehens zum Tragen kommen.

Der BYLET wurde für die Klassenstufen zwei bis sieben konstruiert. Er liegt in drei Parallelversionen (A, B und C) vor, so dass eine Veränderungsmessung möglich ist, ohne dass Behaltenseffekte befürchtet werden müssen.

3.2.7 Validierung der Messung des Leseverstehens mit dem BYLET

Das zentrale Gütekriterium einer jeden psychometrischen Messung ist die Validität (Messick, 1995). Sie lässt sich in Inhaltsvalidität, statistische Validität, konvergente und divergente Validität unterteilen. Diese vier Unterarten der Validität werden nun in Bezug zum BYLET diskutiert.

3.2.7.1 Inhaltsvalidität Die Inhaltsvalidität verlangt, dass ein Messmodell zur Theorie passt (Bühner, 2011; Messick, 1995). Verschiedene Theorien postulieren unterschiedliche Teilkompetenzen des Leseverstehens (s. Abschnitt 3.2.1). Es muss also bei der Modellierung des BYLETs die Frage beantwortet werden, ob sich die feingliedrigere Aufteilung der Kompetenzstufen des IGLU-Modells

gegenüber der einfacheren Annahme, dass es nur die Teilkompetenzen des Dekodierens und des Sprachverständnisses gibt, durchsetzt. Zudem sollte geprüft werden, ob sich die Textschwierigkeit als eigenes Element des Leseverstehens aus der Datenstruktur extrahieren lässt.

3.2.7.2 Statistische Validität Statistische Validität ist gegeben, wenn die Rückschlüsse aus den Analysen mit der in den Daten enthaltenen Information übereinstimmen. Im Falle der psychometrischen Modellierung zeigt sich statistische Validität zum Beispiel in der Robustheit der Analysen, welche in dieser Studie mittels Kreuzvalidierung geprüft wird. Die Kreuzvalidierung wird darüber hinaus für viele Anwendungsfälle als grundsätzlich bestes Modellwahlkriterium empfohlen, da sie fast ohne Modellannahmen auskommt. Im Gegensatz dazu basieren gängige Modellwahlkriterien, wie Fit Indizes auf Modellannahmen, welche mit hoher Wahrscheinlichkeit falsch sind (Arlot & Celisse, 2010). So sind zum Beispiel Likelihood-basierte Modellwahlkriterien wie das AIC oder BIC nur dann optimal, wenn die Modelle wirklich korrekt spezifiziert sind. Bei der Kreuzvalidierung hingegen werden beide Teilstichproben gleichermaßen von möglicher Fehlspezifikation beeinflusst, so dass diese bei der Modellwahl per Kreuzvalidierung nicht zum Tragen kommt. Darüber hinaus wird die statistische Validität darüber erfasst, dass verschiedene Schätzmethoden auf dieselben Modelle angewandt und ihre Ergebnisse verglichen werden. In dieser Studie wird dafür sowohl im Framework der IRT als auch der KTT eine Modellschätzung vorgenommen.

3.2.7.3 Konvergente und divergente Validität Schlussendlich sollte eine valides neu entwickeltes Messinstrument mit etablierten Messinstrumenten korrelieren, insofern diese dasselbe, ein ähnliches oder ein Teilkonstrukt des interessierenden Konstrukts erfassen (Lukesch, 1998). Ist dies gegeben, so spricht man von einer guten konvergenten Validität. Divergente Validität hingegen zeichnet sich dadurch aus, dass mit den Werten von Messinstrumenten, die etwas anderes messen, nicht oder zumindest nicht so stark korrelieren sollte, wie mit inhaltlich verwandten Konstrukten. In dieser Studie etwa wird erwartet, dass sich ein Zusammenhang der Werte des BYLETs mit dem SLS und der Deutschnote ergibt. Der Zusammenhang der BYLET-Werte mit der Mathenote sollte jedoch substanziell kleiner sein als der Zusammenhang mit der Deutschnote der Kinder.

3.2.8 Psychometrische Messung von Leseverstehen mit dem BYLET

Aus der Theorie zum Leseverstehen ergeben sich verschiedene Forschungsfragen, die im Zuge der Modellierung beantwortet werden sollen. Als erstes gilt es, die Anzahl der Abstufungen der Kompetenzstufen zu identifizieren. Zudem sollte untersucht werden, ob sich die Textschwierigkeiten aus den Antwortmustern als Faktoren extrahieren lassen. Diese angenommenen inhaltlichen Strukturen

lassen sich in einer Vielzahl von angenommenen Faktorenstrukturen abbilden, welche sich wiederum mit einer Vielzahl aus verschiedenen Schätz- und Optimierungsverfahren untersuchen lassen. Aus diesem Grund wurde im Jahr 2018 eine Pilotstudie (Kraus, Wild, Schilcher, & Hilbert, 2021) durchgeführt. In der Pilotstudie wurden acht verschiedene Modelle mit verschiedenen Schätzverfahren gegeneinander getestet. Zusätzlich wurde in der Pilotstudie auch überprüft, ob die Nebengütekriterien der Testfairness und Anwendbarkeit erfüllt waren.

3.2.8.1 Erkenntnisse aus der Pilotstudie Ausführliche Ergebnisse der Pilotstudie lassen sich in einem Research Report im Open Science Framework (osf) unter https://osf.io/bmp54/?view_only=3db22c0b3cec4383820514243e431b1b nachlesen. In Zusammenfassung ergab sich, dass die aus der Theorie abgeleiteten Modelle gut passten. Deskriptiv wurden sowohl die Items der höheren Kompetenzstufen als auch die Fragen der schwierigeren Textabschnitte weniger häufig gelöst, als die Items der niedrigeren Kompetenzstufen und der einfacheren Textabschnitte. Dabei war die Testversion BYLET-C insgesamt schwieriger als die Testversionen BYLET-A und BYLET-B. Die Anzahl der zu testenden Modelle konnte für die Hauptstudie von acht auf drei reduziert werden. Dabei konvergierte nur der bayesianische Schätzalgorithmus MHMR in angemessener Zeit, so dass für die Hauptstudie die anderen Schätzer nicht verwendet wurden. Gleichzeitig zeigte sich, dass die Wahl der Startwerte Einfluss auf die Modelloptimierung hatte, so dass für die Hauptanalysen eine Startwerteoptimierung geplant wurde. Bei der Presonenparameterschätzung erwies sich das bayesianische MAP-Verfahren als optimal, da es sowohl mit den Summenwerten assoziiert war als auch zu normalverteilten Personenparametern führte. Darüber hinaus zeigte sich der Test als genderfair und unkompliziert in Durchführung und Auswertung. Trotz multiplem Testen unterschieden sich Jungen und Mädchen in keiner der drei Testversionen des BYLETs.

3.2.8.2 Rationale In dieser Studie sollten also verschiedene testtheoretische Modelle mit unterschiedlicher inhaltlicher Bedeutung mit verschiedenen Verfahren geschätzt werden und ihre Robustheit per Kreuzvalidierung bestimmt werden. Darüber hinaus sollten die hier getesteten Modelle nicht nur auf globale Modellpassung untersucht werden, sondern im Rahmen der KTT-Analysen auch auf lokale Fehlspezifikationen geprüft werden.

Aus der Theorie zum Leseverstehen und der Pilotstudie ergaben sich drei plausible Modelle, welche die Existenz einer zweistufigen oder einer vierstufigen Lesekompetenz und den Einbezug der Textschwierigkeit modellieren. Die Abbildungen 2, 3 und 4 verdeutlichen die angenommene Struktur der getesteten Modelle.

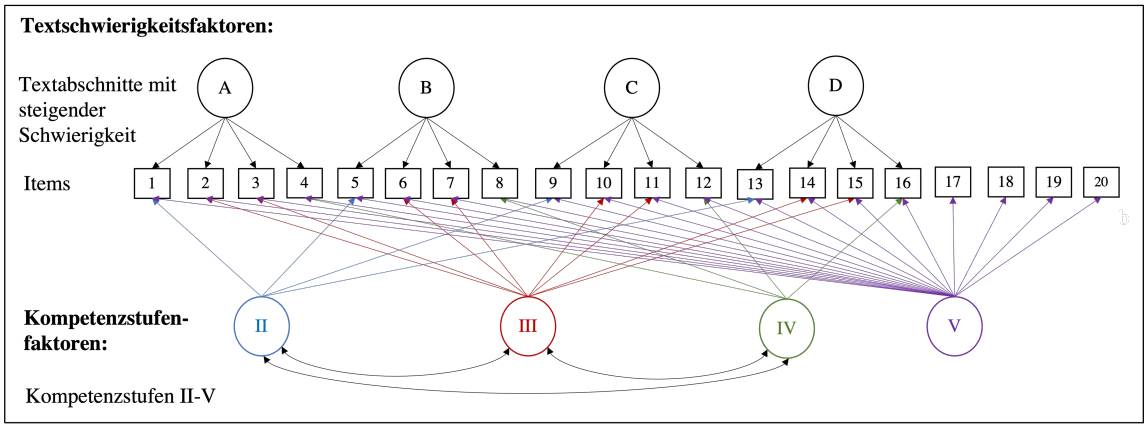


Abbildung 2: Theoriebasiertes Modell

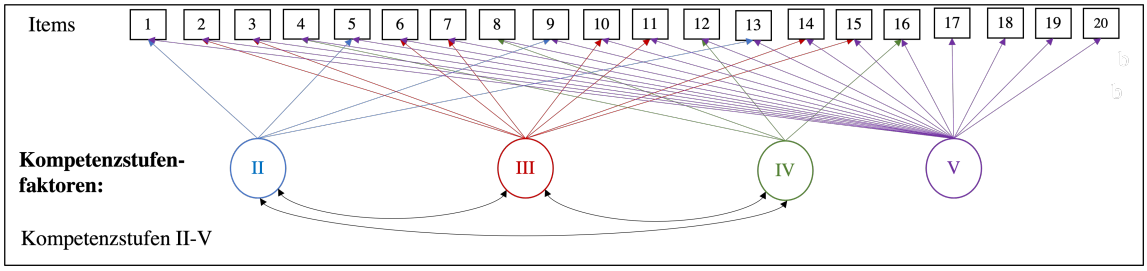


Abbildung 3: Sparsames Modell

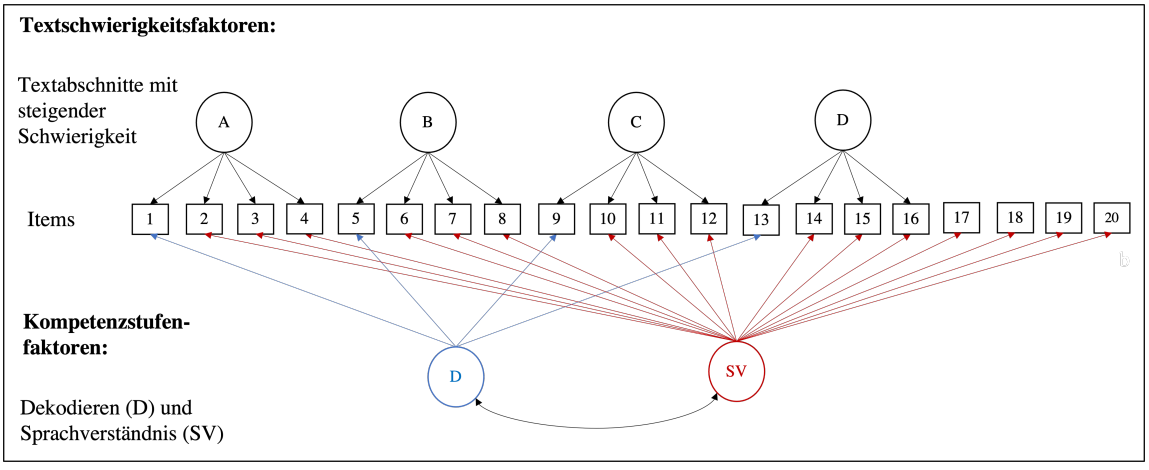


Abbildung 4: Alternatives Modell

3.2.9 Hypothesen und Forschungsfragen

Die vorliegende Studie verfolgte vier Ziele:

1. Aus der Pilotstudie vermutete Gender-Fairness bestätigen
2. Replikation der präferierten Modelle aus der Pilotstudie mit Methoden der IRT und der KTT
3. Per Kreuzvalidierung die Robustheit der globalen Passung der Modelle bestimmen
4. Bestimmung von lokaler Fehlspezifikation, Reliabilität und konvergenter, divergenter externer Validität

Dabei dient das erste Ziel zur Bestätigung der bereits vermuteten günstigen Nebengütekriterien. Das zweite Ziel stellt die psychometrische Modellierung des BYLETs und damit die Untersuchung der Skalierbarkeit dar. Ebenfalls zur Modellierung gehört die Bestimmung der Robustheit der Modellschätzungen und die Bestimmung lokaler Fehlspezifikationen. Die Robustheit stellt jedoch auch einen Indikator für die statistische Validität dar. Als weitere Validitäten sollen die konvergente Validität und die divergente Kriteriumsvalidität über Zusammenhänge mit verwandten Konstrukten und Kriterien bestimmt werden. Abschließend galt es, auch das letzte Hauptgütekriterium, die Reliabilität und damit eine Eignung des BYLETs für die Einzelfalldiagnostik, zu untersuchen.

3.3 Methode

Im folgenden Abschnitt wird zunächst die Stichprobe beschrieben. Anschließend werden die Durchführung der Testungen, der Bayerische Lesetest und die durchgeführten Analysen vorgestellt.

3.3.1 Stichprobe

Zur psychometrischen Modellierung des BYLETs lagen für die Version BYLET-A Daten von 6644 Kindern, für die Version BYLET-B Daten von 7551 Kindern und für die Version BYLET-C Daten von 6994 Kindern vor. Davon stammten mit 6483 (Version BYLET-A), 7376 (Version BYLET-B) und 6785 (Version BYLET-C) Tests die meisten Daten von Kindern der Klassenstufen zwei bis vier aus der FiLBY-Studie. Alle weiteren Daten entstammten einer zusätzlichen Erhebung an Gymnasien in den Klassenstufen sechs und sieben. Tabelle 2 gibt einen Überblick, der Stichprobenzusammensetzung, gegliedert nach Klassenstufen. Insgesamt nahmen 10 390 Kinder an der Erhebung teil. Etwa die Hälfte, nämlich 4237 Kinder der Stichprobe waren männlich, 4282 der Kinder waren weiblich und 1871 gaben ihr Geschlecht nicht an.

Tabelle 2: Teilstichprobengröße in Abhängigkeit des Alters der Getesteten

	Mitte Kl.2	Ende Kl.2	Mitte Kl.3	Ende Kl.3	Anfang Kl.4	Weihnachten Kl.4	Anfang Kl.6	Anfang Kl.7
EG	3239	4655	2085	3597	2445	208	0	0
GY	0	0	0	0	0	0	295	270
KG	1759	1855	0	460	341	0	0	0

Anmerkung. GY = Gymnasiale Stichprobe; EG = Experimentalgruppe; KG = Kontrollgruppe; Kl. = Klasse.

Die Differenz aus der Gesamtanzahl der Tests und der Gesamtanzahl der Kinder ergibt sich dadurch, dass die Kinder der FiLBY-Stichprobe den BYLET wiederholt mit wechselnden Versionen bearbeiteten und so mehrfach in die Stichprobe eingingen. Die Stichprobe war jedoch so zusammengesetzt, dass von keinem Kind mehr als ein Testergebnis pro Version in die Stichprobe einging.

Tabelle 3: Teilstichprobengröße in Abhängigkeit des Migrationshintergrunds der Getesteten

	Mitte Kl.2	Ende Kl.2	Mitte Kl.3	Ende Kl.3	Anfang Kl.4	Weihnachten Kl.4	Anfang Kl.6	Anfang Kl.7
kMG	3824	4930	1513	2785	2001	181	219	193
MG	539	685	242	382	202	10	74	50

Anmerkung. MG = Migrationshintergrund; kMG = kein Migrationshintergrund.

Darüber hinaus wurde die Auswahl so getroffen, dass Teilstichproben so gleichverteilt wie möglich über die Altersgruppen waren. Es nahmen Kinder teil, die sich zwischen der Mitte der zweiten Jahrgangsstufe und Beginn der siebten Jahrgangsstufe befanden. Die Grundschulkinder wurden im Verlauf ihrer Schullaufbahn fünf Mal mit dem BYLET getestet. Mit Beginn in der zweiten Hälfte der zweiten Jahrgangsstufe im Jahr 2019 wurden die Kinder jedes halbe Jahr mit dem BYLET getestet. Die Testungen fanden Mitte der zweiten Jahrgangsstufe, Ende der zweiten Jahrgangsstufe, Mitte der dritten Jahrgangsstufe, Ende der dritten Jahrgangsstufe und zu Beginn und Mitte der vierten Jahrgangsstufe statt. Die Gymnasialkinder besuchten die sechste und siebte Jahrgangsstufe und nahmen nur einmal im Herbst 2020 an der Testung teil. Da die Stichprobe im Verlauf der FiLBY-Studie immer kleiner wurde, und die gymnasiale Teilstichprobe im Vergleich zur Grundschulstichprobe kleiner war, ergab sich trotzdem eine Unterrepräsentation der älteren Kinder in der Stichprobe. Die meisten Kinder waren in Deutschland geboren und hatten damit keinen Migrati-

onshintergrund. Eine genaue Aufschlüsselung der Kinder mit Migrationshintergrund innerhalb der Stichprobe ist in Tabelle 3 dargestellt.

3.3.2 Durchführung

Die Daten des BYLETs und des SLS aus der Grundschule wurden im Rahmen der FiLBY-Studie erhoben. Dabei wurden diese jeweils vor und nach zur Studie gehörigen Lesetrainings von den Lehrkräften durchgeführt. Die Kinder erhielten 40 Minuten zur Testbearbeitung für den BYLET und drei Minuten für das SLS. Die Testbögen lagen in Papierform vor. Im Anschluss an die Datenerhebung wurden die Rohdaten von den Lehrkräften in Exceltabellen eingegeben und an die Forschenden der Universität Regensburg übermittelt. In der Kontrollgruppe und den Gymnasialklassen wurden die Testbögen an die Universität Regensburg versandt und die Ergebnisse dort von Projektmitarbeitenden digitalisiert.

3.3.3 Materialien

Beim BYLET handelt es sich um den oben beschriebenen, neu entwickelten Leseverstehenstest mit drei Parallelversionen, bei dem die Schülerinnen und Schüler 20 Multiple-Choice-Fragen zu vier in ihrer Komplexität ansteigenden Textabschnitten beantworten. Gegenstand des Texts ist eine Science-Fiction-Geschichte: Eine Weltraumcrew erkundet einen neuen Planeten und erlebt ein Abenteuer. Da der Test zusätzlich eine steigende Aufgabenschwierigkeit innerhalb der Textabschnitte aufweist, soll er das Leseverstehen innerhalb eines breiten Leistungsspektrums erfassen. Die Art der Konzeption vermeidet vorwissensbedingte und geschlechtsspezifische Verzerrungen und bietet deshalb eine gute Voraussetzung für eine faire Messung des Leseverstehens. Die demografischen Informationen wurden mithilfe eines Fragebogens erfasst, der von den Kindern falls nötig mit Hilfe der Lehrkraft ausgefüllt wurde. In der gymnasialen Stichprobe gaben die Lehrkräfte Informationen über den Migrationshintergrund auf einem gesonderten Informationsblatt an. Das Geschlecht wurde in dieser Teilstichprobe auf den Testbögen selbst vermerkt. Das zur konvergenten Validierung eingesetzte Salzburger Lese-Screening (SLS) (Wimmer & Mayringer, 2016) wurde genauso durchgeführt wie der BYLET und nach dem Manual ausgewertet. Das SLS wurde mit dem Raschmodell (Rasch, 1960) modelliert. Es lag jedoch ausschließlich von den FiLBY-Kindern vor, nicht aber von der gymnasialen Teilstichprobe. Als externes Kriterium wurde in der FiLBY-Stichprobe für die Messzeitpunkte am Ende des Schuljahrs zusätzlich die Deutsch- und die Mathenote erfragt.

3.3.4 Analysen

Alle Analysen wurden mit der Analysesoftware R (R Core Team, 2020) durchgeführt. Die Datenaufbereitung erfolgte mit *tidyverse* (Wickham et al., 2019) und *multilevel* (Bliese, 2016). Die psychometrischen Analysen wurden mit den packages *mirt* (Chalmers, 2012) und *lavaan* (Rosseel, 2012) durchgeführt. Für die an anderer Stelle (s. Abschnitt 4.5.1.1) näher beschriebene Raschmodellierung des Salzburger Lese-Screenings wurde das package *eRm* (Mair, Hatzinger, Maier, Rusch, & Mair, 2020) verwendet.

3.3.4.1 Deskriptive Analysen Zunächst wurden deskriptive Analysen durchgeführt. Dafür wurde die Lösungshäufigkeit der einzelnen Items bestimmt. Diese wurde anschließend für die einzelnen Klassenstufen, Jungen und Mädchen und für Kinder mit und ohne Migrationshintergrund berechnet. Anschließend wurden die polychorischen Itemkorrelationen bestimmt und ausgewertet, ob Items, die zu einem postulierten Faktor gehörten, im Median stärker miteinander korrelierten, als sie dies mit Items von anderen Faktoren taten.

3.3.4.2 Psychometrische Modellierung Zur psychometrischen Modellierung wurden multidimensionale Item-Response-Modelle (MIRT) und Strukturgleichungsmodelle (SEMs) geschätzt und deren Modellgüte bestimmt. Die Modellwahl erfolgte per Kreuzvalidierung.

3.3.4.2.1 Multidimensionale Item-Response-Theorie (MIRT) Zur psychometrischen Modellierung wurde zunächst der Zugang über die multidimensionale Item-Response-Theorie (MIRT) (Reckase, 2009) gewählt. MIRT-Modelle setzen als Erweiterung des 2-PL-Modells (Birnbaum, 1968) die dichotomen Itemantworten in einen logistischen Zusammenhang mit einer vorher spezifizierten Anzahl an latenten Variablen. Allgemein unterscheidet man zwischen kompensatorischen und nichtkompensatorischen MIRT-Modellen (Reckase, 2009). Letztere zeichnen sich dadurch aus, dass die Leistungsfähigkeit auf einem Faktor durch die Fähigkeit auf anderen Faktoren begrenzt wird. Es ist also in diesen Modellen nicht möglich ein maximales bzw. minimales Verhältnis der Personenwerte der verschiedenen Faktoren zu über- bzw. unterschreiten (Adams, Wilson, & Wang, 1997). Hier wurde ein nicht-kompensatorisches Modell gewählt, da es nicht plausibel erscheint, dass latente Faktoren höherer Kompetenzstufen die gemessene Fähigkeit der niedrigeren Kompetenzstufen begrenzen sollte. Zusätzlich zu den Itemschwierigkeiten und den Varianzen und Kovarianzen der latenten Faktoren wurden Ladungen geschätzt. Ladungen stellen in diesen Modellen Trennschärfeparameter dar. Die Modellgleichung des nicht-kompensatorischen Modells lautet im allgemeinen Fall (Van der Linden, 2016):

$$p(u_i = 1; \theta_p) = \frac{e^{a_i' \theta_p + d_i}}{1 + e^{a_i' \theta_p + d_i}}$$

mit:

- $p(u_i = 1)$ = Wahrscheinlichkeit das Item i zu lösen
- θ_p = Vektor der latenten Faktoren
- a_i' = Trennschärfenparameter bzw. Ladung
- d_i = Schwierigkeitsparameter

Insgesamt wurden die drei im Theorieteil begründeten Modelle geschätzt. Das erste geschätzte Modell war das theoriebasierte Modell. Es umfasst vier Faktoren, die die Kompetenzstufen II-V des Leseverstehens abbilden. Weitere vier Faktoren repräsentieren die Level der Textschwierigkeit. So gehört jedes Item zu jeweils einem Kompetenzstufenfaktor der Faktoren II-IV. Zusätzlich gehört jedes Item zum Generalfaktor, also zur Kompetenzstufe V und zum Textschwierigkeitsfaktor des entsprechenden Textabschnitts. Zusätzlich nimmt das theoriebasierte Modell eine volle Korrelationsstruktur zwischen den Kompetenzstufenfaktoren II-IV an. Alle anderen Faktorkorrelationen sind hingegen auf 0 fixiert und werden daher nicht geschätzt. Abbildung 5 zeigt grafisch, welche Ladungen und Korrelationen im theoriebasierten Modell geschätzt wurden.

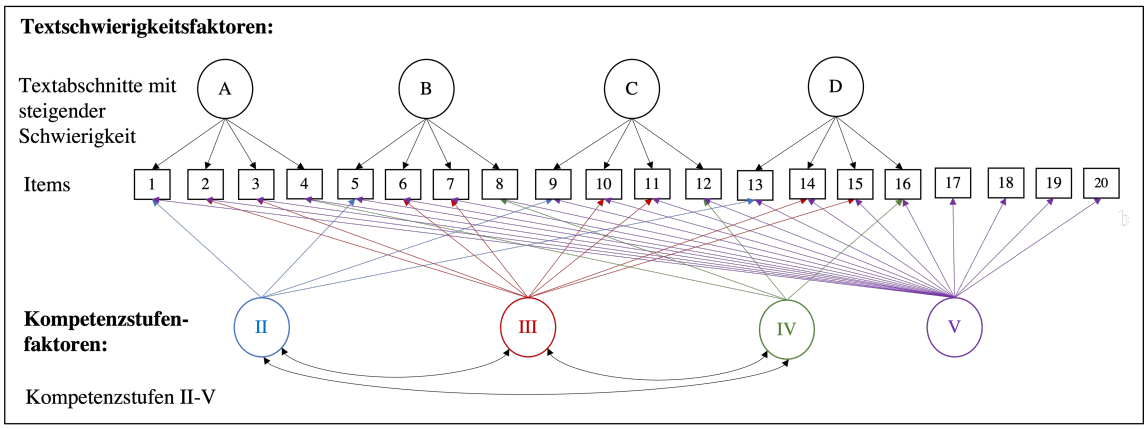


Abbildung 5: Theoriebasiertes Modell

Das zweite Modell war ein sparsameres Modell. Da sich in der Pilotierung gezeigt hatte, dass die Textschwierigkeitsfaktoren nicht immer zu identifizieren waren, wurde im zweiten Modell auf diese verzichtet. Das sparsame Modell beinhaltete also nur die vier Kompetenzstufenfaktoren. Auch in diesem Modell wurden nur die Korrelationen zwischen den Faktoren der Stufen II, III und IV zugelassen. Das sparsame Modell ist in Abbildung 6 dargestellt.

Das dritte getestete Modell beinhaltete nur zwei Faktoren, die das Leseverstehen im Sinne der Dekodierfähigkeit und des Sprachverständnisses abbildeten. Zusätzlich umfasste es die vier Faktoren für die ansteigende Textschwierigkeit. Es diente dazu, ein alternatives Verständnis von Lesekompetenz

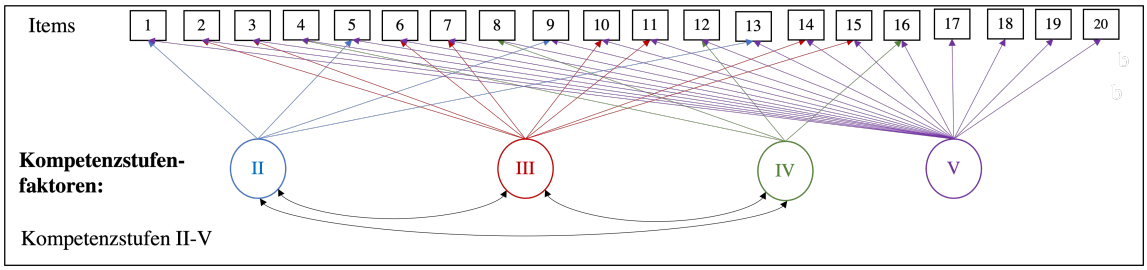


Abbildung 6: Sparsames Modell

abzubilden und damit das von den Autor:innen des Bayerischen Lesetests angenommene Konzept von Leseverstehen kritisch zu hinterfragen. In diesem alternativen Modell werden nur die Items, die eine Informationsentnahme messen, dem Faktor des Leseverstehens zugeordnet. Alle anderen Items werden als Messung der kognitiven Fähigkeiten bzw. des allgemeinen Sprachverständnisses aufgefasst. Die Ladungsstruktur des alternativen Modells ist in Abbildung 7 dargestellt.

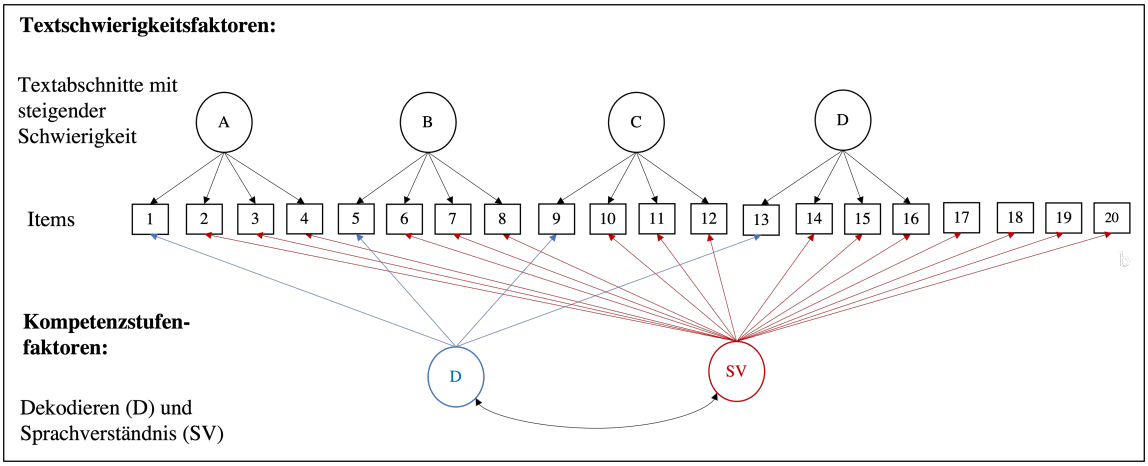


Abbildung 7: Alternatives Modell

Modellschätzung

Für die Schätzung der MIRT-Modelle stehen verschiedene Algorithmen zur Verfügung. Neben verschiedenen Versionen des Expectation-Maximization (EM)-Algorithmus (z. B. MCEM, QMCEM, Bock & Aitkin, 1981; Wei & Tanner, 1990) kann auch eine bayesianische Schätzung per Metropolis-Hastings-Robbins-Monro (MHRM)-Algorithmus (Cai, 2010) erfolgen. Da dieser für komplexe faktorielle Strukturen mit mehr als vier Faktoren empfohlen wird (Chalmers, 2012), und sich auch in der Pilotierung durch eine hohe Konvergenzrate und praktikable Rechenzeiten hervorhob, wurden alle Modelle ausschließlich mit dem MHRM-Algorithmus geschätzt. Jeder Modellschätzung wurde dabei eine Startwerteoptimierung vorgeschaltet. Der MHRM-Algorithmus fasst Elemente aus MCMC (dem Metropolis-Hastings (MH)-Algorithmus, Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) mit der stochastischen Approximation der Robbins-Monro (RM)-Methode

(Robbins & Monro, 1951) zusammen. Der MHRM-Algorithmus basiert auf drei Grundideen:

1. Die Modellparameter lassen sich in Itemparameter und Personenparameter unterteilen. Die Personenparameter werden als fehlende Daten betrachtet (Cai, 2010) und stochastisch approximiert. Dann gilt unter der Annahme, dass der Personenparametervektor ($\mathbf{x_i}$) bekannt ist, dass sich die Loglikelihood des MIRT-Modells als Summe multinomieller Loglikelihoods des ordinalen (hier des logistischen) Regressionsmodells ergibt (z. B. McCullagh 1980). Die Personenparameter bilden die Regressoren, die Itemantworten die abhängige Variablen. Die Itemparameter entsprechen den Regressionsgewichten. Es gilt dann für die Loglikelihood der Itemparameter:

$$logL(\theta|Z) = \sum_{j=1}^{n}[\sum_{i=1}^{N}\sum_{k=0}^{1} I_{y_{ij}} log\ p(y_{ij} = k|\theta, x_i)]$$

mit:

- $\mathbf{Z}$ = Matrix aus beobachteten Itemantworten y_{ij} und angenommenen Personenparametern $\mathbf{x_i}$
- θ = Vektor der Itemparameter
- j = Index der Items
- i = Index der Personen
- $I_{y_{ij}}$ = Indikatorfunktion für $y = 1$
- $p(y_{ij} = k|\theta, x_i)$ = Wahrscheinlichkeit, dass Person i Item j löst, gegeben den Itemparametervektor θ und den Personenparametervektor $\mathbf{x_i}$

Die Loglikelihood ist damit eine einfache Summe aus den logarithmierten Lösungswahrscheinlichkeiten aller im Datensatz vorhandenen Itemantworten.

2. Diese Loglikelihood lässt sich additiv in zwei unabhängige Teile zerlegen, von denen der erste Teil sowohl von den Itemantworten y_{ij} also auch von den eigentlich zunächst unbekannten Personenparametern $\mathbf{x_i}$ abhängt. Der zweite Teil hängt jedoch nur von den angenommenen Personenparametern, respektive von deren gemeinsamer Verteilung ab:

$$logL(\omega|Z) = logL(\theta|Z) + logL(\vartheta|X)$$

mit:

- ω = Itemparametervektor und Gruppenparameter (Parameter, welche die Verteilung der Personenparameter beschreiben)
- θ = Vektor der zu schätzenden Itemparameter
- $\mathbf{Z}$ = Matrix aus beobachteten Itemantworten y_{ij} und angenommenen Personenparametern $\mathbf{x_i}$
- ϑ = Element des Vektors, der die Verteilung der Personenparameter beschreibt

3. Beide Teile der Loglikelihood lassen sich getrennt voneinander, abwechselnd optimieren. Eine ausführliche Beschreibung davon findet sich bei Cai (2010). Hier wird der Prozess nur sehr komprimiert dargestellt. Beim Optimierungsprozess werden abwechselnd mithilfe des MH-Algorithmus unter Ausnutzung der Proportionalität von $\Pi(X|Y,\omega)$ und $L(\omega|Z)$ jeweils verschiedene Ziehungen der $\mathbf{x_i}$ realisiert.

Diese werden anschließend unter Ausnutzung von $\nabla_\omega l(\omega|Y) = \int_{\mathcal{E}} s(\omega|Z)\Pi(dX|Y,\omega)$ verwendet, um $\nabla_\omega l(\omega|Y)$ zu approximieren und daraus den Gradienten $s(\omega|Z)$ zu bestimmen, sowie die volle Informationsmatrix upzudaten.

Der aus den gemittelten Ziehungen gewonnene Gradient $s_{k+1}(\omega|Z)$ der k-ten Iteration dient wiederum dazu, gemeinsam mit der upgedateten Informationsmatrix die Modellparameter upzudaten. Hierfür wird der Robbinson-Monro-Algorithmus (Robbins & Monro, 1951) verwendet. Er optimiert durch Störeinflüsse verrauschte Funktionen. Das ganze Verfahren läuft iterativ, bis die Modellparameter im Rahmen einer a priori festgelegten Toleranz konvergieren.

Modellgütebewertung

Es ist bei psychometrischen Modellen sehr schwierig ihre Güte ohne einen Referenzpunkt zu bewerten. Daher ist es üblich und ratsam die Passung mehrerer plausibler Modelle miteinander zu vergleichen und das am besten passendende Modell als finales Modell auszuwählen (Browne, 2000). Die Modellgüte lässt sich in Abhängigkeit der gewählten Schätzverfahren mit unterschiedlichen Kennwerten bemessen. Ist die Modellschätzung zum Beispiel Likelihood basiert, so können diese oder Abkömmlinge wie das BIC (Schwarz, 1978) zur Modellgütebewertung herangezogen werden. Wird das Modell jedoch über die Optimierung einer anderen Diskrepanzfunktion geschätzt, so muss auf absolute oder komparative Fit Indizes wie z. B. das RMSEA oder das CFI zurückgegriffen werden. Ist die Diskrepanzfunktion oder eine Transformation davon χ^2-verteilt, so kann auch ein Hypothesentest zur Modellpassung herangezogen werden. Da diese jedoch eine erhöhte Sensitivität bei großen Stichproben aufweisen, wurde in dieser Studie die Modellpassung ausschließlich über Fit Indizes bestimmt. Für die MIRT-Modellierung wurde etwa das BIC als Modellwahlkriterium ausgewählt, da es sparsamere Modelle bevorzugt (Vandekerckhove, Matzke, & Wagenmakers, 2015).

3.3.4.2.2 Strukturgleichungsmodelle Die Schätzung eine psychometrischen Modells als Strukturgleichungsmodell, bietet neben den Parameterschätzungen und den globalen Fit Indizes, die Möglichkeit, Fehlspezifikationen des postulierten latenten Modells zu identifizieren. Deshalb wurde zusätzlich zur MIRT-Modellierung auch eine Modellierung als Strukturgleichungsmodell vorgenommen.

Modellschätzung

Die Modellschätzung erfolgte als konfirmatorische Faktorenanalyse mit dem weighted-least-squares-mean-variance-adjusted (WLSMV)-Schätzer (Asparouhov, Muthén, & Muthén, 2006). Diesen erhält man durch Minimierung der folgenden allgemeinen Diskrepanzfunktion:

$$F = (\sigma(\Theta_0) - s)' W^{-1} (\sigma(\Theta_0) - s)$$

mit:

- s = Stichprobenkennwerten des unrestringierten Modells
- $\sigma(\Theta_0)$ = modellimplizierte Kennwerte
- W = Gewichtungsmatrix

Anstelle von Pearson-Korrelationen wurden polychorische Korrelationen verwendet. Die Gewichtungsmatrix W ist dabei eine Diagnonalmatrix und damit immer positiv definit. Man erhält sie folgendermaßen: die realisierten Itemantworten werden als Resultat einer an Schwellenwerten dichotomisierten latenten, kontinuierlichen Variable angenommen und bilden die Gewichte der Diagnonalmatrix W.

Die Parameterschätzungen und die zugehörigen Standardfehler erhält man dann durch die Minimierung der gewichteten kleinsten Quadrate. Hierzu werden neben der Gewichtungsmatrix die geschätzte asymptotische Kovarianzmatrix der polychorischen Korrelationsschätzungen verwendet (Muthén, 1984). Für normalverteilte latente Variablen und große Stichproben (N $>$ 200) erweist sich die WLSMV-Schätzung als stabil (Li, 2016). Als Optimierungsverfahren der Diskrepanzfunktion wurde das Broyden–Fletcher–Goldfarb–Shanno (BFGS)-Verfahren eingesetzt (Broyden, 1970; Shanno, 1970). Es gehört zu den Quasi-Newton-Verfahren, welche die Hesse-Matrix mittels eines stochastischen Verfahrens approximieren.

Modellgütebewertung

Da die WLSMV-Schätzung kein Likelihood basiertes Verfahren ist, stehen als Modellgütekriterien keine Indizes zur Verfügung, die auf der Likelihood basieren (d.h. AIC, BIC, LR-Tests). Bei der Schätzung werden jedoch Modell implizierte Kovarianzmatrizen erzeugt, auf welchen Distanzmaße wie χ^2-Statistiken, und globale Fit Indizes wie RMSEA oder CFI berechenbar sind. Zusätzlich kann die Modellgüte, im Sinne möglicher Fehlspezifikationen über Modifikationsindizes bemessen werden (Byrne, 2013).

Bestimmung der lokalen Fehlspezifikation

Im Rahmen der SEM-Modellierung wurden die Modifikationsindizes berechnet. Sie geben an, wel-

che im Modell auf 0 fixierten Parameter bei einer freien Schätzung einen signifikant von einem festgelegten Grenzwert verschiedenen Parameterschätzwert erhalten würden. Die Betrachtung der Modifikationsindizes liefert also einen Hinweis, durch welche zusätzlichen Pfade oder Ladungen ein Modell verbessert werden könnte (Byrne, 2013). Mit Hinblick auf eine Testrevision geben Modifikationsindizes etwa durch den Vorschlag korrelierter Fehlerterme Hinweise auf überarbeitungswürdige Items.

3.3.4.2.3 Kreuzvalidierung Geplant wurde eine dreifache Kreuzvalidierung. Durchgeführt wurde jedoch eine zehnfache Kreuzvalidierung, weil ein erster Durchlauf mit dreifacher Kreuzvalidierung zu wenig Informationen für eine fundierte Modellentscheidung geliefert hatte. Die Kreuzvalidierung wurde ursprünglich für die lineare Regression entwickelt. Dabei wurde zunächst das Regressionsmodell an einem Trainingsdatensatz geschätzt, anschließend die Regressionsgewichte extrahiert und für eine neue Teststichprobe $\hat{y}$ berechnet. Diese $\hat{y}$ korrelierte man dann mit dem tatsächlich beobachteten y-Vektor und erhielt so eine Maßzahl für die Modellpassung an einer neuen Stichprobe. Kreuzvalidierung ist ein gutes Mittel, um Modellanpassung an zufällige Datenstrukturen zu vermeiden. Obwohl in sehr großen Stichproben dieser Effekt zufälliger Fehler bei der Modellwahl nicht stark zum Tragen kommt, ist der Einfluss der zufälligen Fehler schlecht abzuschätzen. Daher kann auch in großen Stichproben Kreuzvalidierung eine Abschätzung des zufälligen Fehlers ermöglichen (Browne, 2000).

3.3.4.2.4 Kreuzvalidierung MIRT Für die Kreuzvalidierung der MIRT-Modellierung gibt es zahlreiche Möglichkeiten. Hier wurde das folgende Vorgehen gewählt. Der Datensatz wurde in zehn Teile zerlegt, von denen in jeder Iteration neun als Trainingsset zusammengefasst und der zehnte Teil als Testset definiert wurde. Am Trainingsset wurde zunächst das sogenannte Trainingsmodell geschätzt. Für jedes der spezifizierten Modelle wurden im Trainingsmodell die Startwerte optimiert und anschließend das Modell mit den besten Startwerten geschätzt. Die besten Startwerte ergaben sich aus der besten Modellpassung, gemessen am BIC. Die Parameter des so geschätzten Modells wurden extrahiert und als Parameterrestriktionen für die Schätzung des Testmodells gesetzt. Technisch geschieht dies über das Setzen der Parameter als Startwerte und die Bestimmung des Modellfits ohne weitere Optimierung der Startwerte. Die Modellpassung des Testmodells mit den fixierten Startwerten diente dann als Maß für die Modellpassung an einem neuen Datensatz. Auch hier diente das BIC als Maß für die Modellpassung.

3.3.4.2.5 Kreuzvalidierung SEM Im Grunde analog wurde die Kreuzvalidierung der SEM-Modelle durchgeführt. Zunächst wurde die Gesamtstichprobe wiederholt in eine Trainings- und eine Teststichprobe geteilt und anschließend für jedes Set an Stichproben die Kreuzvalidierung durchgeführt. Bei der Kreuzvalidierung der SEMs wurde die Trainingsstichprobe verwendet, um die Momentenmatrix $S_{training}$ zu berechnen. Anschließend wurde der Itemparametervektor geschätzt, indem die Diskrepanzfunktion F in Bezug auf den Parametervektor minimiert wurde. Die empirische Momentenmatrix S_{test} der Teststichprobe wurde geschätzt und zur vom Trainingsmodell implizierten Momentenmatrix über die Diskrepanzfunktion in Verbindung gesetzt. Der Kreuzvalidierungsindex (CVI) ergab sich also als Distanz zwischen der empirischen Momentenmatrix des Testsets und der vom trainierten Modell erwarteten Momentenmatrix (Browne, 2000).

Bedingt man die Kreuzvalidierung auf eine spezifische Trainingsstichprobe, so kann man zeigen, dass der CVI ein konsistenter Schätzer für den Generalisierbarkeitsfehler ist (Browne, 2000).

In dieser Studie wurde zur besseren Interpretierbarkeit die Diskrepanzfunktion F über die χ^2-Statistik zum Root Mean Square Error of Approximation (RMSEA) (Brown & Cudeck, 1993) und zum Comparative Fit Index (CFI) (Bentler, 1990) verrechnet. Der RMSEA berechnet sich nach:

$$RMSEA = \sqrt{max(\frac{F(S,\Sigma(\hat{\Theta}))}{df} - \frac{1}{N-1}, 0)}$$

mit:

- df = Anzahl der Freiheitsgrade
- $F(S, \Sigma(\hat{\Theta}))$ = Diskrepanzfunktionswert
- N = Stichprobengröße

Je kleiner also die Diskrepanzfunktion F im Vergleich zur Stichprobengröße N wird, desto kleiner wird der RMSEA. Je besser also das Modell passt, umso kleiner wird die Abweichung der modellimplizierten Momentenmatrix $\Sigma(\hat{\Theta})$ von der empirischen Momentenmatrix S. Als Referenzwerte geben Hu und Bentler (1999) an, dass ab einem RMSEA-Wert < 0.06 von einer guten Modellpassung ausgegangen werden kann.

Der CFI berechnet sich nach:

$$\text{CFI} = 1 - \frac{\hat{F_{\text{H}}}}{\hat{F_{\text{N}}}}$$

mit:

- $\hat{F_{\text{H}}}$ = F-Wert des zu testenden Modells
- $\hat{F_{\text{B}}}$ = F-Wert des Nullmodells

(Xia & Yang, 2019)

Das Nullmodell nimmt an, dass die Daten zufällig entstanden sind und in Wahrheit alle Kovarianzen null sein müssten. Hier gilt es also, dass der Bruch $\frac{\hat{F}_H}{\hat{F}_N}$ dann besonders klein wird, wenn einfach gesprochen, die Daten besser zum postulierten Modell als zum Nullmodell passen. In diesem Fall wird eine kleine Zahl von der 1 subtrahiert und der CFI wird groß. Beim CFI gilt gemeinhin ein Wert $> .95$ als ein akzeptabler Wert. Sind jedoch grundsätzlich die Zusammenhänge im Datensatz eher schwach ausgeprägt, so wird der CFI auch bei korrekter Modellspezifikation nicht sehr groß, da die Diskrepanz zum Nullmodell von vorneherein gering ist. Deswegen gilt es als Empfehlung, immer sowohl einen absoluten als auch einen komparativen Fit Index zur Modellwahl zu verwenden (Bühner, 2021; Hu & Bentler, 1999).

Hier wurden die Fit Indizes auf die Momentenmatrix der Teststichprobe und die modellimplizierte Momentenmatrix der Trainingsstichprobe angewendet. Dies war möglich, da in allen Testsstichproben die Stichprobengröße gleich war.

3.3.4.3 Bestimmung der Personenscores, Faktorwerte und der Reliabiliät Im Anschluss an die Modellschätzung - oder genauer gesagt - im Anschluss an die Schätzung der Itemparamter bzw. Ladungen, wurden auch die Personenparameter bzw. Faktorwerte berechnet. Zur Stabilisierung der Schätzungen wurden die Personenparameter mit der MAP-Methode geschätzt (Baker & Kim, 2004; Mislevy, 1986). Im Framework der klassischen Testtheorie wurden die Faktorwerte mit der Empirical Bayes Modal (EBM)-Methode (Varriale & Vermunt, 2012) geschätzt. Im Rahmen der Bestimmung der Personenparameter und der Faktorwerte wurde auch die Reliabilität der Faktoren modellbezogen ermittelt. Hierfür wurde die interne Konsistenz nach Zumbo, Gadermann und Zeisser (2007) verwendet.

3.3.4.4 Bestimmung der Validität Anschließend wurden die Personenparameter und Faktorwerte für jeden Faktor miteinander korreliert, um so ein Maß für die Ähnlichkeit der Schätzverfahren (als MIRT-Modell bzw. KFA) zu erhalten. Zusätzlich wurde untersucht, ob sich ein Zusammenhang der Faktorwerte mit dem Alter der Kinder, beziehungsweise dem Erhebungszeitpunkt zeigte. Nachdem die Lesekompetenz im Laufe der Schullaufbahn üblicherweise steigt, wäre die Korrelation im Sinne einer konvergenten Kriteriumsvalidität der Messung wünschenswert. Zur weiteren Bestimmung der konvergenten und der Kriteriumsvalidität wurden die Faktorwerte bzw. Personenparameter mit den Schulnoten der Fächer Deutsch und Mathematik sowie mit der mit dem SLS erhobenen Leseflüssigkeit korreliert.

3.4 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse der deskriptiven Auswertung, der psychometrischen Modellierung, sowie die Kennwerte der Reliabilität und Validität des BYLETs dargestellt. Um Redundanzen zu vermeiden werden nur die Ergebnisse der Testversion A ausführlich dargestellt. Die Ergebnisse der Versionen B und C befinden sich im Anhang A.

3.4.1 Deskriptive Ergebnisse

Zunächst wurden für alle Items die Lösungshäufigkeiten in Abhängigkeit des Erhebungszeitpunkts und damit in Abhängigkeit des Alters der Kinder bestimmt. Abbildung 8 verdeutlicht, dass die Lösungshäufigkeiten der einzelnen Items stiegen, je älter die Kinder wurden. Eine Ausnahme ergibt sich für die Items 4, 7 und 8 für die Erhebungszeitpunkte zur Mitte und zum Ende der dritten Jahrgangsstufe. Hier lagen die ersten coronabedingten Schulschließungen zwischen den Messzeitpunkten. Weitere Ausnahmen ergaben sich für die Items 6, 7, 9, 10, 11, 14 und 17 für die Erhebungszeitpunkte Anfang der vierten Klasse und Weihnachten der vierten Klasse.

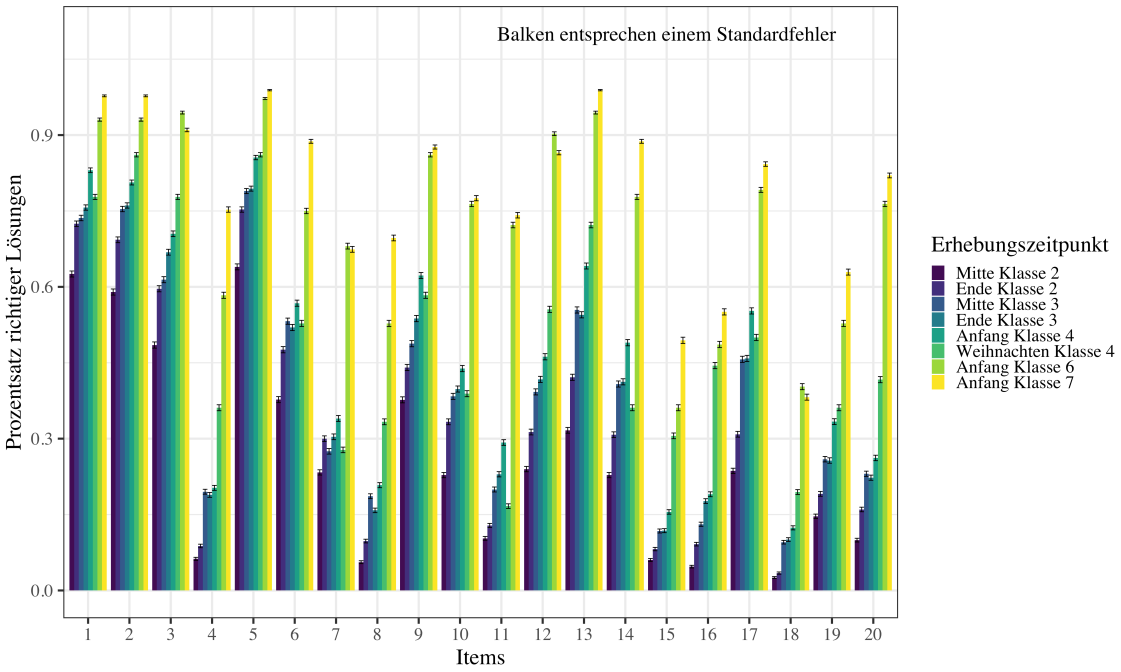


Abbildung 8: Lösungshäufigkeit in Abhängigkeit des Alters

Jeweils vier aufeinanderfolgende Aufgaben, also 1-4, 5-8, 9-12, 13-16 gehörten zu den einzelnen Textabschnitten. Innerhalb jedes Abschnitts wurde zunächst eine Aufgabe zur Kompetenzstufe II, dann zwei Aufgaben zur Kompetenzstufe III und schließlich eine Aufgabe zur Kompetenzstufe IV gestellt. Betrachtet man die Lösungshäufigkeiten innerhalb der Textabschnitte, so sieht man, dass

die Lösungshäufigkeiten mit steigender Kompetenzstufe sanken. Es wurde also meistens das erste Item jedes Abschnitts deutlich häufiger gelöst, als das zweite und dritte Item, während das vierte Item am seltensten korrekt beantwortet wurde. Zudem sieht man auch, dass die Lösungshäufigkeit über die Abschnitte hinweg sank. Die Aufgaben zum Textabschnitt A (Items 1-4), der über die geringste Textschwierigkeit verfügte, wurden also häufiger gelöst, als die Aufgaben des Textabschnitts B (Items 5-8). Diese wiederum wurden häufiger gelöst als die Aufgaben des Textabschnitts C (Items 9-12), usw. Die Standardfehler bewegten sich für alle Schätzungen der Lösungshäufigkeiten in der Größenordnung 10^{-3}.

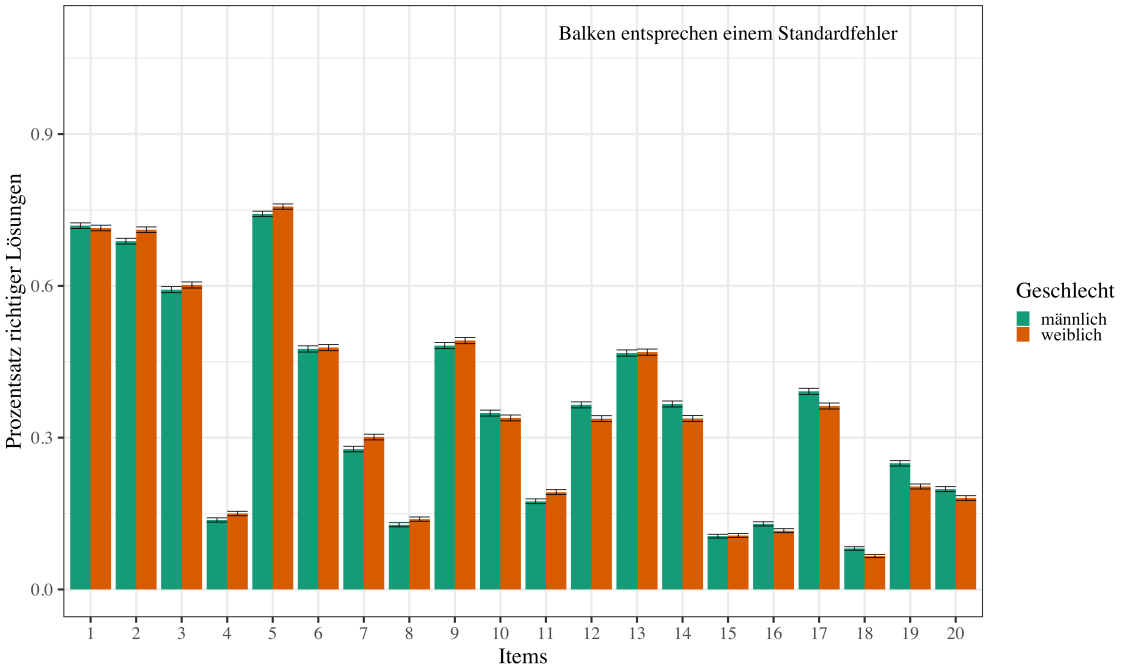


Abbildung 9: Lösungshäufigkeit in Abhängigkeit des Geschlechts

In Abbildung 9 sieht man, dass Jungen und Mädchen sich nicht wesentlich in der Lösungshäufigkeit der einzelnen Items unterschieden. Lediglich bei den Items 12, 14, 17 und 19 schienen die Jungen deskriptiv einen Vorteil gegenüber den Mädchen erhalten zu haben. Demgegenüber schnitten Kinder mit Migrationshintergrund sichtbar schlechter ab als ihre Altersgenossen ohne Migrationshintergrund. Sie beantworteten im Vergleich zu Kindern, die in Deutschland geboren waren mit Ausnahme des Items 19 alle Aufgaben mit geringerer Häufigkeit korrekt. Besonders ausgeprägt waren die Unterschiede in den Items 9, 12, und 13 (vgl. Abbildung 10).

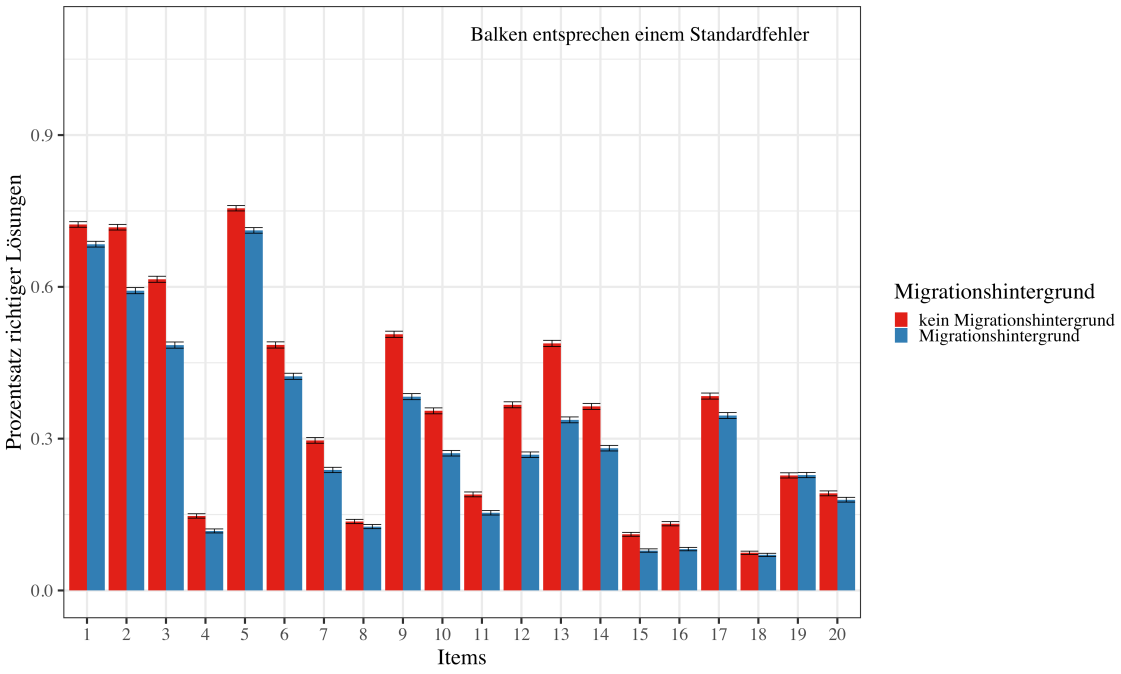


Abbildung 10: Lösungshäufigkeit in Abhängigkeit des Migrationshintergrunds

3.4.1.1 Itemkorrelationen mit Hinblick auf die angenommene Faktorstruktur Zur deskriptiven Beschreibung der psychometrischen Datenstruktur wurde die Itemkorrelationsmatrix so gruppiert, dass jeweils die Items, die gemeinsam einem Faktor zugeordnet werden sollten, eine Gruppe bildeten. Anschließend wurde der Median der Itemkorrelationen innerhalb dieser Gruppen gebildet. Dann wurden die Mediane der Korrelationen der gruppeninternen Items mit den Medianen der gruppenexternen Items verglichen. Dies wurde für die vom Simple View of Reading postulierten zwei Kompetenzstufenfaktoren, für die vom Stufenmodell postulierten vier Kompetenzstufen, sowie für die vier Textschwierigkeitsfaktoren durchführt.

Tabelle 4: Vergleich der Mediane der Itemkorrelationen der Textschwierigkeitsfaktoren

Itemgruppen	Abschnitt A	Abschnitt B	Abschnitt C	Abschnitt D
$MD_{innerhalb}$	0.36	0.27	0.40	0.52
$MD_{außerhalb}$	0.30	0.29	0.33	0.35

Anmerkung. Vergleich der Mediane (MD) der Itemkorrelationen, innerhalb und außerhalb, der durch die Faktoren bestimmten Itemgruppen.

Tabelle 5: Vergleich der Mediane der Itemkorrelationen der Kompetenzfaktoren

Itemgruppen	Stufe II	Stufe III	Stufe IV	Dekodieren	Sprachverständnis
$MD_{innerhalb}$	0.41	0.31	0.44	0.41	0.31
$MD_{außerhalb}$	0.32	0.32	0.32	0.32	0.32

Anmerkung. Vergleich der Mediane (MD) der Itemkorrelationen, innerhalb und außerhalb, der durch die Faktoren bestimmten Itemgruppen.

Es zeigte sich, dass mit Ausnahme der Kompetenzstufe III, dem Abschnitt B und den Sprachverständnisitems alle Items im Median deskriptiv stärker mit Items des eigenen Faktors korrelierten als mit Items von anderen Faktoren. Während vor allem Kompetenzstufe II und IV sowie die Abschnitte C und D substanzielle Korrelationsunterschiede zeigten, fielen in den anderen Gruppen die Unterschiede deutlich geringer aus (vgl. Tabelle 4 und 5).

3.4.2 Psychometrische Modellierung

Im Anschluss an die deskriptiven Analysen wurden die verschiedenen aus der Theorie abgeleiteten psychometrischen Modelle geschätzt. Hierfür wurden sowohl MIRT-Modelle als auch SEMs verwendet.

3.4.2.1 MIRT-Modelle Im Rahmen der MIRT-Modellierung wurden alle drei angenommenen Modelle mittels MHRM-Algorithmus geschätzt, dabei die Startwerte optimiert und die Modellpassungen per Kreuzvalidierung bestimmt. Abschließend wurde das bestpassende Modell ausgewählt und am vollen Datensatz erneut geschätzt. Die so erhaltenen Item- und Personenparameter wurden für die weiteren Berechnungen genutzt.

3.4.2.1.1 Modellwahl über Kreuzvalidierung Abbildung 11 zeigt die BICs in den Testsets der 10 Testsets der Kreuzvalidierung. In fast allen Testsets besaß das sparsame, zweite Modell den kleinsten BIC. Wie in Abbildung 11 mit einem schwarzen Quadrat gekennzeichnet, verfügte es auch im Durchschnitt über die beste Passung und wurde deswegen als finales Modell ausgewählt. Eine nur zweistufige Lesekompetenz und ein substanzieller Einfluss der Textschwierigkeit auf das Leseverstehen konnten damit empirisch nicht bestätigt werden.

Auch bei der Testversion BYLET-C verfügte das sparsame Modell 2 über die beste Modellpassung. Lediglich bei der Testversion BYLET-B erhielt das dritte, alternative Modell einen niedrigeren,

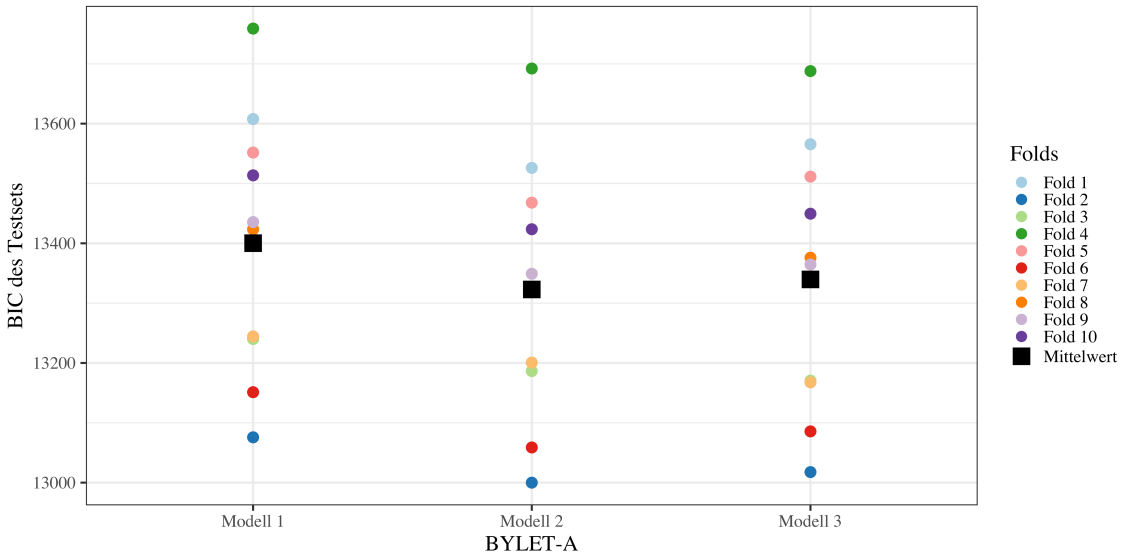


Abbildung 11: Modellwahl MIRT

durchschnittlichen BIC. Die Differenz zum mittleren BIC des sparsamen Modells betrug hier allerdings nur vier Punkte. Der Unterschied war hier also gering.

3.4.2.1.2 Kennwerte des finalen Modells

Tabelle 6: Fitstatistiken

BIC	Loglikelihood	df	p	RMSEA	CFI
130435	-64958	151	< 0.001	0.03	0.97

Anmerkung. Kennwerte des finalen MIRT-Modells für den BYLET-A; df = Freiheitsgrade; p = p-Wert.

Wie Tabelle 6 zeigt, besaß das finale Modell des BYLET-A eine sehr gute Modellpassung. So lag sowohl der RMSEA unter 0.06, als auch der CFI über 0.95.

Tabelle 7: Personenparameterkorrelationen

	II	III	IV	V
II	1	0.58	0.01	-
III	0.58	1	0.10	-
IV	0.01	0.10	1	-

Anmerkung. Römische Ziffern beziehen sich auf die Personenparameter der jeweiligen Kompetenzstufe.

Bei den Faktorkorrelationen zeigte sich ein moderater Zusammenhang zwischen der zweiten und dritten Kompetenzstufe ($r = 0.58$), sowie ein sehr schwacher Zusammenhang zwischen der dritten und vierten Kompetenzstufe ($r = 0.10$). Die Stufen II und IV waren hingegen unkorreliert ($r = 0.01$). Die weiteren Korrelationskoeffizienten sind Tabelle 7 zu entnehmen.

Tabellen 12 und 13 zeigen die Ladungsmatrix und die Kommunalitäten der Items. Auffällig sind zwei negative Ladungen der Items 13 und 14 auf den Faktoren der Kompetenzstufen II und III. Item 7 und 19 verfügten mit 0.17 und 0.16 über die geringsten Kommunalitäten. Allgemein fielen die Ladungen auf dem Faktor der Kompetenzstufe III am geringsten aus. Dies stimmte mit den deskriptiven Ergebnissen insofern überein, als auch dort die Items innerhalb des Faktors III nur unwesentlich stärker korrelierten, als sie dies mit Items von anderen Faktoren taten.

3.4.2.2 SEM Modelle Auch im Rahmen der SEM-Modellierung wurden alle drei theoretisch begründeten Modelle geschätzt und die Modellpassungen per Kreuzvalidierung bestimmt. Zur Schätzung wurde hier der WLSMV-Algorithmus verwendet. Die Modellpassung wurde über das CFI und den RMSEA bestimmt. Abschließend wurde auch in der SEM-Modellierung das bestpassende Modell ausgewählt, am vollen Datensatz erneut geschätzt und die Modifikationsindizes berechnet und ausgewertet.

3.4.2.2.1 Modellwahl über Kreuzvalidierung Zur Modellwahl konnte hier das BIC nicht verwendet werden, weil die WLSMV-Schätzung nicht likelihoodbasiert ist. Daher wurden RMSEA und CFI verwendet. In Abbildung 12 sind die RMSEAs der zehn Testsets aus der Kreuzvalidierung dargestellt. Wieder zeigte sich, dass das sparsame, zweite Modell in den meisten Folds den geringsten RMSEA und auch den durchschnittlich geringsten RMSEA hatte. Das zweite Modell war also nach dem Kriterium des RMSEA das präferierte Modell.

Wie in Abbildung 13 dargestellt, bestätigte die Betrachtung der CFIs in den Testsets die Präferenz für das sparsame, zweite Modell. Auch hier verfügte das zweite Modell mit den höchsten Werten in allen Testsets über die beste Modellpassung. In keinem Testset überschritten jedoch die Passungsunterschiede eine maximale Differenz von 0.02. Zusammenfassend lässt sich also festhalten, dass alle drei getesteten Modelle über eine angemessene Passung verfügten.

Auch bei den Versionen BYLET-B und BYLET-C ergab sich ein vergleichbares Bild, so dass schlussendlich für alle drei Testversionen das zweite Modell als finales Modell festgelegt wurde. Im folgenden Abschnitt werden deswegen nur die Kennwerte der Schätzung des zweiten Modells am vollständigen Datensatz präsentiert.

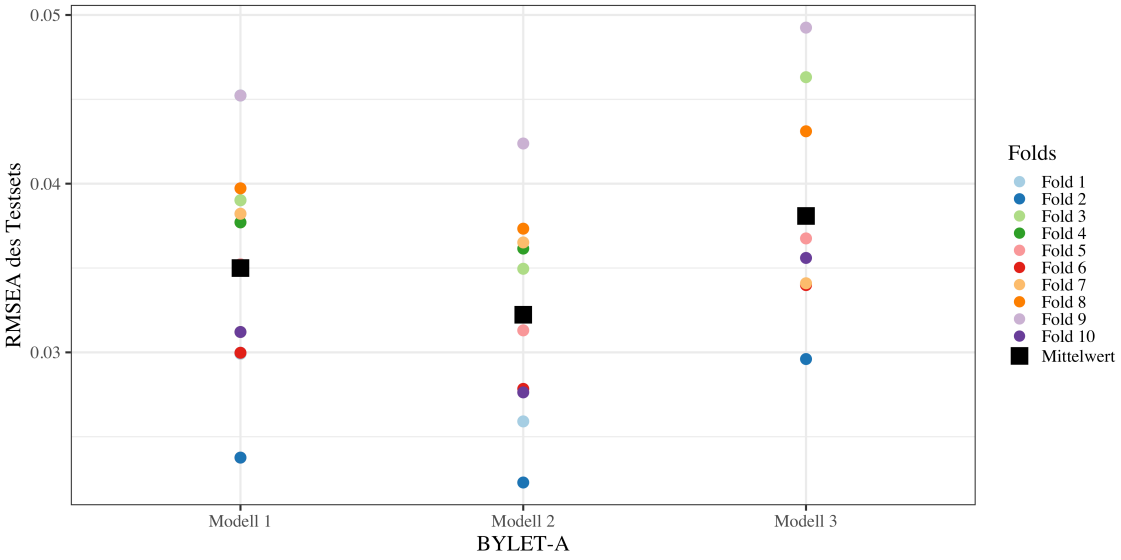


Abbildung 12: Modellwahl SEM RMSEA

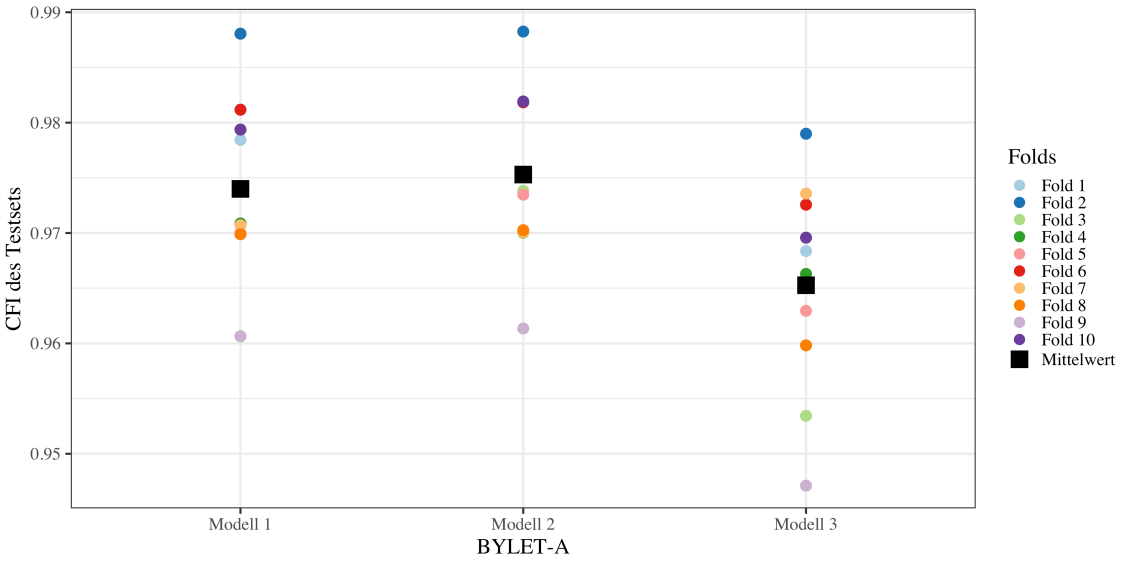


Abbildung 13: Modellwahl SEM CFI

3.4.2.2.2 Kennwerte des finalen Modells Als Kennwerte des finalen Modells ergaben sich eine Teststatistik von 600.78, bei 151 Freiheitsgraden und einem p-Wert < 0.001. Der χ^2-Test wurde also nicht signifikant.

Tabellen 12 und 13 stellen unter anderem die Ladungen der Items auf die Faktoren bei WLSMV-Schätzung dar. Ähnlich wie bei der MIRT-Modellierung zeigten sich vereinzelt sehr geringe oder leicht negative Ladungen. So waren zum Beispiel die Items 4, 11 und 15 negativ mit den Kompetenzstufenfaktoren assoziiert.

Tabelle 8: Faktorkorrelationen

	II	III	IV	V
II	1	0.74	-0.17	-
III	0.74	1	0.05	-
IV	-0.17	0.05	1	-

Anmerkung. Römische Ziffern beziehen sich auf die Faktorwerte der jeweiligen Kompetenzstufe.

Die Faktorkorrelationsmatrix zeigte einen starken Zusammenhang zwischen der zweiten und dritten Kompetenzstufe, jedoch keinen Zusammenhang zwischen der dritten und vierten Kompetenzstufe. Die zweite und vierte Kompetenzstufe waren sogar schwach negativ assoziiert. Die exakten Korrelationen können in Tabelle 8 nachgelesen werden.

3.4.2.2.3 Modifikationsindizes des finalen SEM-Modells In allen drei BYLET-Versionen zeigten sich nicht vernachlässigbare lokale Fehlspezifikationen mit Modifikatiosindizes im zwei und dreistelligen Bereich. Die Modifikationsindizes fielen dabei bei der Version BYLET-A am geringsten aus. Die vorgeschlagenen Lockerungen der Parameterrestriktionen waren jedoch in allen drei Modellen unterschiedlich. Lediglich die Zurodnung von Item 6 zu Faktor II, fand sich übereinstimmend in allen drei Modellen. Sie wurde so als Modifikation vorgenommen und der Modellfit erneut bestimmt.

Tabelle 9: Modifikationsindizes des finalen Modells

Finales Modell					Modifiziertes Modell				
Parameterset	mi	epc	power	dec	Parameterset	mi	epc	power	dec
i17~~i19	75.25	0.18	1.00	n.r.	i17~~i19	75.78	0.18	1.00	n.r.
IV=~i11	45.09	0.24	1.00	n.r.	IV=~i11	42.13	0.24	1.00	n.r.
i5~~i6	43.46	0.14	1.00	n.r.	II=~i12	35.92	0.19	1.00	n.r.
II=~i12	33.84	0.18	1.00	n.r.	i17~~i20	32.95	0.12	1.00	n.r.
i17~~i20	32.51	0.12	1.00	n.r.	i14~~i17	31.93	0.10	1.00	n.r.
III=~i12	31.56	0.16	1.00	n.r.	III=~i1	31.70	0.44	0.97	r.
i14~~i17	31.33	0.10	1.00	n.r.	IV=~i17	28.84	-0.20	1.00	n.r.
II=~i6	29.16	0.36	0.99	r.	III=~i12	24.61	0.14	1.00	n.r.
IV=~i17	28.21	-0.19	1.00	n.r.	i19~~i20	23.38	0.11	1.00	n.r.
i2~~i3	27.74	0.34	1.00	r.	i11~~i4	23.09	0.11	1.00	n.r.
i11~~i4	24.02	0.12	1.00	n.r.	i13~~i16	20.84	0.10	1.00	n.r.
i19~~i20	23.16	0.11	1.00	n.r.	i1~~i6	17.85	-0.11	1.00	n.r.
i13~~i16	22.49	0.10	1.00	n.r.	i13~~i12	17.52	-0.08	1.00	n.r.
i11~~i8	17.58	0.10	1.00	n.r.	i6~~i7	17.49	0.08	1.00	n.r.
i13~~i12	17.17	-0.08	1.00	n.r.	i5~~i6	16.91	0.11	1.00	n.r.
III=~i19	15.40	-0.10	1.00	n.r.	III=~i19	16.55	-0.11	1.00	n.r.
i9~~i12	15.31	0.07	1.00	n.r.	i11~~i8	16.42	0.10	1.00	n.r.
i6~~i7	14.75	0.08	1.00	n.r.	i9~~i12	15.17	0.07	1.00	n.r.
i11~~i17	14.44	-0.09	1.00	n.r.	i4~~i17	15.03	-0.09	1.00	n.r.
II=~i11	14.43	-0.29	0.97	n.r.	i11~~i17	14.15	-0.08	1.00	n.r.

Anmerkung. mi = Modifikationsindex; epc = erwarteter Parameterwert; delta (Grenzwert der Parameterrelevanz) = 0.3; dec = Entscheidung; n.r. = nicht relevant; r. = relevant.

Tabellen 9 und 10 zeigen, die Modifikationsindizes und die Veränderung der globalen Fit Indizes nach der zusätzlichen Schätzung der Ladung von Item 6 auf Faktor II. Dabei zeigt Tabelle 10, dass sich die globalen Fit Indizes durch die Modifikation nur sehr geringfügig veränderten. Es wurde deswegen das theoretisch begründete Modell als finales Modell für alle weiteren Analysen verwendet.

Tabelle 10: Fit Indizes nach Modifikation des finalen Modells

Modell	RMSEA	CFI
Finales Modell	0.0212	0.9915
Modifiziertes Modell	0.0205	0.9921

Anmerkung. RMSEA = Root mean squared error of approximation; CFI = Comparative Fit Index; Angabe auf vier Nachkommastellen genau, um geringe Unterschiedlichkeit zu verdeutlichen.

3.4.2.2.4 Reliabilität des finalen Modells Im Anschluss an die Schätzung wurden die Reliabilitäten für die Kompetenzstufen und für das Gesamtmodell aus der WLSMV-Schätzung berechnet. Hierfür wurde die interne Konsistenz von Zumbo und Kollegen (2007) verwendet. Wie in Tabelle 11 dargestellt, verfügte nur der Faktor der Kompetenzstufe V über eine sehr gute Reliabilität mit einem ordinalen Alpha-Wert > 0.90. Die anderen Faktoren verfügten jedoch mit Werten > 0.70 ebenfalls über akzeptable Reliabilitäten.

Tabelle 11: Reliabilitäten

	II	III	IV	V	Gesamtreliabilität
Ordinales Alpha	0.72	0.79	0.75	0.91	0.91

Anmerkung. Ordinales Alpha = Interne Konsistenz nach Zumbo und Kollegen (2007).

Es kann also davon ausgegangen werden, dass der BYLET auch zur Einzelfalldiagnostik einsetzbar ist.

3.4.3 Vergleich der testtheoretischen Theorien

Zum Abschluss der psychometrischen Modellierung wurden die Ladungen und die Korrelationen der Faktorwerte aus der MIRT-Modellierung und der SEM-Modellierung gegenübergestellt. Dies diente der Untersuchung der Robustheit der statistischen Auswertung und damit der Überprüfung der statistischen Validität.

3.4.3.1 Vergleich der Ladungen Beim Vergleich der Ladungen zeigte sich, dass diese alle bis auf eine Differenz von 0.05 übereinstimmten. Diese hohe Kongruenz spricht also für eine robuste Schätzung und für eine Äquivalenz der beiden Schätzverfahren auf interpretativer Ebene. Eine

genaue Gegenüberstellung der Ladungen, Signifikanzprüfung der Ladungen und Kommunalitäten zeigen die Tabellen 12 und 13.

Tabelle 12: Vergleich der Ladungen Stufe II und III

Items	II mirt	II sem	II sem p	III mirt	III sem	III sem p
1	0.47	0.47	< 0.001	-	-	-
2	-	-	-	0.58	0.56	< 0.001
3	-	-	-	0.58	0.56	< 0.001
4	-	-	-	-	-	-
5	0.45	0.45	< 0.001	-	-	-
6	-	-	-	0.15	0.20	< 0.001
7	-	-	-	0.10	0.11	< 0.001
8	-	-	-	-	-	-
9	0.26	0.26	< 0.001	-	-	-
10	-	-	-	0.01	0.01	0.545
11	-	-	-	0.09	0.07	0.019
12	-	-	-	-	-	-
13	-0.11	-0.09	0.001	-	-	-
14	-	-	-	-0.05	-0.03	0.309
15	-	-	-	0.00	0.00	0.951
16	-	-	-	-	-	-
17	-	-	-	-	-	-
18	-	-	-	-	-	-
19	-	-	-	-	-	-
20	-	-	-	-	-	-

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; p = p-Wert.

Tabelle 13: Vergleich der Ladungen Stufe IV und V

Items	IV mirt	IV sem	IV sem p	V mirt	V sem	V sem p	Kommunalität
1	-	-	-	0.41	0.40	< 0.001	0.40
2	-	-	-	0.56	0.54	< 0.001	0.65
3	-	-	-	0.50	0.49	< 0.001	0.59
4	0.40	0.37	< 0.001	0.55	0.54	< 0.001	0.46
5	-	-	-	0.55	0.54	< 0.001	0.51
6	-	-	-	0.46	0.47	< 0.001	0.23
7	-	-	-	0.40	0.41	< 0.001	0.17
8	0.52	0.54	< 0.001	0.55	0.55	< 0.001	0.57
9	-	-	-	0.54	0.55	< 0.001	0.37
10	-	-	-	0.58	0.59	< 0.001	0.33
11	-	-	-	0.64	0.64	< 0.001	0.41
12	0.02	-0.05	0.103	0.63	0.65	< 0.001	0.40
13	-	-	-	0.87	0.86	< 0.001	0.76
14	-	-	-	0.70	0.70	< 0.001	0.49
15	-	-	-	0.64	0.60	< 0.001	0.41
16	0.22	0.22	< 0.001	0.70	0.68	< 0.001	0.54
17	-	-	-	0.58	0.58	< 0.001	0.33
18	-	-	-	0.57	0.52	< 0.001	0.32
19	-	-	-	0.39	0.38	< 0.001	0.16
20	-	-	-	0.53	0.52	< 0.001	0.29

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; p = p-Wert.

3.4.3.2 Korrelationen der Faktorwerte Anschließend wurde überprüft, inwiefern auch die Bestimmung der Personenparameter beziehungsweise Faktorwerte der beiden Schätzverfahren vergleichbar waren. Dafür wurden sowohl die Faktorwerte der SEMs als auch die Personenparameter der MIRT-Modelle für alle Faktoren geschätzt und vollständig miteinander korreliert. Die Ergebnisse dieser Analysen sind in Tabelle 14 abgebildet.

Tabelle 14: Korrelation der Personenparameter und der Faktorwerte

	IIsem	IIIsem	IVsem	Vsem	IImirt	IIImirt	IVmirt	Vmirt	Alter
IIsem	1	0.94	-0.29	0.14	0.93	0.82	-0.11	0.21	0.08
IIIsem	-	1	-0.18	0.15	0.83	0.94	0.01	0.21	0.08
IVsem	-	-	1	0.12	-0.16	-0.05	0.94	0.08	0.15
Vsem	-	-	-	1	0.11	0.12	0.18	0.99	0.46
IImirt	-	-	-	-	1	0.79	-0.03	0.15	0.10
IIImirt	-	-	-	-	-	1	0.10	0.15	0.09
IVmirt	-	-	-	-	-	-	1	0.16	0.16
Vmirt	-	-	-	-	-	-	-	1	0.45

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen.

Die durch die unterschiedlichen testtheoretischen Ansätze bestimmten Faktorwerte derselben Kompetenzstufen korrelierten sehr hoch miteinander. Die Korrelationen zwischen den Kompetenzstufen hingen kaum von der Schätzmethode ab. Es korrelierte also beispielsweise die Kompetenzstufe II stark mit der Kompetenzstufe III, unabhängig davon, ob diese mit einem MIRT-Modell oder einem SEM geschätzt wurde. Dies ist ein gutes Indiz dafür, dass trotz der unterschiedlichen Schätzverfahren zwischen SEMs und MIRT-Modellen dieselbe Information gewonnen wurde und eine Vergleichbarkeit der Methoden gegeben ist. Die letzte Spalte der Tabelle 14 zeigt, wie die einzelnen Kompetenzstufen mit dem Alter der getesteten Kinder korrelierten. Hier zeigten sich für die Faktoren II - IV nur schwache Zusammenhänge. Lediglich die Kompetenzstufe V schien substanziell mit dem Alter der Kinder anzusteigen.

3.4.4 Konvergente und divergente Validität

Zur Bestimmung der konvergenten Validität wurden für alle Erhebungszeitpunkte der FiLBY-Stichprobe die Faktorwerte des BYLETs mit den Personenparametern des SLS korreliert. Auch hier zeigten sich nur für die Kompetenzstufe V substanzielle Zusammenhänge. Alle anderen Kompetenzstufen waren nur schwach mit den Werten des SLS assoziiert (vgl. Tabelle 15).

Tabelle 15: Korrelation der BYLET-Faktorwerte mit dem Salzburger Lese-Screening

BYLET	SLS					
Stufe	Mitte Kl.2	Ende Kl.2	Mitte Kl.3	Ende Kl.3	Anfang Kl.4	Weihnachten Kl.4
IIsem	0.14	0.13	0.12	0.13	0.12	0.33
IIIsem	0.15	0.14	0.14	0.15	0.15	0.31
IVsem	0.01	0.03	0.03	0.03	0.07	-0.03
Vsem	0.47	0.46	0.44	0.45	0.40	0.34
IImirt	0.12	0.11	0.11	0.12	0.11	0.33
IIImirt	0.14	0.13	0.12	0.14	0.13	0.30
IVmirt	0.05	0.08	0.07	0.08	0.11	0.05
Vmirt	0.47	0.46	0.44	0.45	0.40	0.35

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; SLS = Salzburger Lese-Screening; Kl. = Klasse., Stufen sind Kompetenzstufen des Bayerischen Lesetests

Tabelle 16: Korrelation der BYLET-Faktorwerte mit den Schulnoten

	Deutschnote	Mathenote
IIsem	-0.16	-0.12
IIIsem	-0.19	-0.14
IVsem	-0.03	-0.02
Vsem	-0.45	-0.36
IImirt	-0.14	-0.10
IIImirt	-0.18	-0.14
IVmirt	-0.08	-0.06
Vmirt	-0.45	-0.37
Deutschnote	1	
Mathenote	0.68	1

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; Schulnoten reichen von 1 bis 6. 1 ist die beste Note.

Zur Bestimmung der konvergenten und divergenten Kriteriumsvalidität wurden abschließend die Zusammenhänge der BYLET-Faktorwerte mit der Deutsch- und der Mathenote bestimmt. Dabei zeigte sich, wie erwartet, dass die Faktorwerte stärker mit der Deutsch- als mit der Mathenote zusammenhingen. Beide Schätzverfahren zeigten auch hier sehr ähnliche Korrelationsmuster. Generell waren jedoch die Zusammenhänge zwischen den Schulnoten stärker als die Zusammenhänge zwischen Deutschnote und Leseverstehen. Es ist also denkbar, dass hier eine Dominanz des Methodeneffekts gegenüber des Inhalteffekts vorlag (vgl. Tabelle 16).

3.5 Diskussion

Im folgenden letzten Abschnitt dieser Studie werden die Ergebnisse zusammengefasst und interpretiert. Anschließend wird diskutiert, wie die Ergebnisse im entscheidungstheoretischen Kontext zu betrachten sind und ein Ausblick auf die Verwendung des BYLETs für weitere Forschungsvorhaben und die Einzelfalldiagnostik gegeben.

3.5.1 Zusammenfassung und Interpretation der Ergebnisse

Noch immer gehört eine strikt theoriegeleitete Testkonstruktion nicht zum Standard in der Testkonstruktion von Leseverstehenstests (Galuschka et al., 2015). Auch die Überprüfung theoretisch angenommener Teilkompetenzen oder Kompetenzstufen mittels mehrdimensionaler psychometrischer Modelle findet in der Entwicklung von Leseverstehenstests kaum Anwendung. In dieser Studie wurde daher ein neuer Leseverstehenstest, der Bayerische Lesetest (BYLET) entwickelt, dessen Items die Anforderungen der Kompetenzstufen IGLU-Kompetenzstufenmodell (Bremerich-Vos et al., 2017) abbilden. Es wurde mit psychometrischen Modellen überprüft, ob sich die Kompetenzstufen in der Datenstruktur zeigten. Der BYLET erwies sich in den deskriptiven Analysen als ein gender-fairer Test, der die oft aufgefundene Überlegenheit der Mädchen im Leseverstehen (Logan & Johnston, 2009, 2010) nicht replizieren konnte.

Mit einer Durchführungsdauer von 45 Minuten und einer standardisierten Kodierung und Auswertung verfügt der BYLET damit über gute Nebengütekriterien. Bei der psychometrischen Modellierung mit MIRT-Modellen und SEMs zeigte sich übereinstimmend ein Modell als das bestpassende, welches das Leseverstehen in vier Kompetenzstufen unterteilte, jedoch keine zusätzliche Modellierung der verschiedenen Textschwierigkeiten berücksichtigte. Hierbei ist es jedoch möglich, dass eine Betrachtung der Textschwierigkeit über den Gsmog der Komplexität von Textschwierigkeit nicht genügend Rechnung getragen hat. McNamara, Graesser und Louwerse (2012) etwa vertreten die

Auffassung, dass Textschwierigkeit vor allem durch die Kohäsion des Textes bestimmt wird. Eine reine Berücksichtigung der Wortlänge und Satzlänge wie beim Gsmog reiche nicht aus, um die Textschwierigkeit umfassend abzubilden.

Dahingegen konnte die Vierstufigkeit des Leseverstehens, welche auch zur Konstruktion des BYLETs herangezogen wurde, auch empirisch gefunden werden. Trotz einer Unterrepräsentation von kompetenten Lesenden verfügten die Modelle, welche vier gegenüber nur zwei Kompetenzstufen annahmen, über die besseren Modellpassungen. Auf der einen Seite ist die empirisch gefundene Vierstufigkeit ein starkes Indiz für die konzeptionelle Vierstufigkeit des Leseverstehens. Gleichzeitig muss aber auch berücksichtigt werden, dass der BYLET mit eben diesem theoretischen Zugang konstruiert wurde. Es ist also nicht auszuschließen, dass sich noch weitere Kompetenzstufen oder Teilkompetenzen des Leseverstehens auf Itemebene abbilden ließen. Eine psychometrische Modellierung kann hier zwar theoretische Argumentationen widerlegen, aber keinesfalls ersetzen. Wie Itemformulierungen lauten, und ob diese Facetten des Leseverstehens wiederspiegeln, bleibt eine inhaltliche Fragestellung.

Die Modellschätzungen erwiesen sich als sehr robust. Beide testtheoretischen Zugänge führten zu sehr ähnlichen Ladungsmatrizen. Die Personenparameter der MIRT-Modelle und die Faktorwerte der SEMs derselben Faktoren waren hoch korreliert. Dafür zeigten sich innerhalb der Faktoren teilweise sehr niedrige oder negative Korrelationen. Sie werfen die Frage auf, ob es sich bei den modellierten Kompetenzstufen tatsächlich um Stufen oder eventuell doch eher um Teilkompetenzen handelt. Für die Stufenannahme spricht, dass alle Items mit zunehmendem Alter und zunehmender Lesekompetenz häufiger gelöst wurden. Als auffällig zeigte sich vor allem die Kompetenzstufe IV, welche kaum oder sogar negativ mit den anderen Kompetenzstufen korrelierte. Zudem verfügte diese über eine verhältnismäßig geringe Reliabilität. Ob dies der Zusammensetzung der Stichprobe oder inhaltlichen Gründen auf der Konstruktebene geschuldet ist, bleibt für zukünftige Studien zu untersuchen. Darüber hinaus zeigten sich für die Faktorwerte der Kompetenzstufen II-IV keine relevanten Zusammenhänge mit dem Alter der getesteten Kinder. Da diese Faktoren jedoch Teilkompetenzen des Leseverstehens darstellen sollten, hätten diese mit zunehmendem Alter zu höheren Werten führen sollen. Dieser erwartete Zusammenhang zeigte sich jedoch ausschließlich für die Kompetenzstufe V. Auch hier müssen weiterführende Untersuchungen zeigen, inwiefern sich diese querschnittlichen Analysen auch für längsschnittliche Entwicklungen bestätigen lassen.

Unter Berücksichtigung der Ladungsmatrix und der Item-Kommunalitäten ließen sich einzelne, problematische Items identifizieren. In der hier vorgestellten Version BYLET-A etwa, luden die Items 11-15 fast gar nicht auf die zugeordneten Faktoren. Auffällig ist, dass diese Items im Test alle auf-

einander folgten und sich auf die beiden mittleren Abschnitte des Texts bezogen. Sollte der BYLET in Zukunft überarbeitet werden, so sollte eine inhaltliche Analyse der Items, zum Beispiel mithilfe kognitiver Interviews erfolgen. Retrospektiv erscheint es möglich, dass manche Kinder bei der Bearbeitung nicht wie in der Durchführungsanleitung empfohlen die Fragen jeweils nach den einzelnen Textabschnitten, sondern nach Lesen des kompletten Textes beantworteten. So könnte neben dem Leseverstehen auch das Erinnerungsvermögen Einfluss auf die Itembeantwortung genommen haben. Im Sinne der Primacy- und Recency-Effekte (Murdock, 1962) erscheint es plausibel, dass gerade die Items zum mittleren Textabschnitt schlechter beantwortet wurden als die Items zum letzten Textabschnitt.

In Bezug zu Reliabilität und Validität verfügte der BYLET über zufriedenstellende Eigenschaften. Dass die Reliabilitäten der Kompetenzstufen II - IV eher gering ausfielen, dürfte auch der geringen Anzahl an Items von jeweils vier Items pro Stufe geschuldet sein. Es ist daher zu überlegen, ob für die Anwendung des BYLETs in der Einzelfalldiagnostik weitere Fragen für die Kompetenzstufen II - IV formuliert werden sollten, um so auch die Werte dieser Kompetenzstufen reliabler abbilden zu können. Demgegenüber wies die Kompetenzstufe V, in die Antworten aller Items eingingen, eine sehr gute Reliabilität > 0.90 auf, so dass Personenparameter dieses Faktors auf jeden Fall auch zur Einzelfalldiagnostik geeignet erscheinen.

In Bezug zur Validität zeigten sich wie erwartet mittlere Zusammenhänge mit der Leseflüssigkeit, welche mit dem SLS erfasst wurde. Da Leseflüssigkeit vor allem im Leselernprozess als Vorläuferfähigkeit oder auch als “Flaschenhals” des Leseverstehens beschrieben wird (Chall, 1983; Pikulski & Chard, 2005), entspricht ein moderater Zusammenhang mit Werten des BYLETs den theoriegeleiteten Erwartungen. Kriteriumsvalidität zeigte sich in einer moderaten Korrelation der Kompetenzstufe V beider Schätzverfahren mit der Deutschnote. Da die Deutschnote neben der Leseleistung auch die Rechtschreibung, Grammatikleistung und die Schreibfähigkeit berücksichtigt, wäre ein höherer Zusammenhang eher erwartungswidrig. Dass die Zusammenhänge der Faktorwerte mit der Mathenote nicht so stark ausgeprägt waren, spricht für eine gute divergente Validität des BYLETs.

3.5.2 Beurteilung des BYLETs aus methodischer Sicht

Die hier präsentierten Ergebnisse sprechen aus methodischer Sicht für den Einsatz des BYLETs als Messinstrument für das Leseverstehen. Die psychometrische Skalierbarkeit ist die mathematische Voraussetzung für eine Verrechnung einzelner Itemantworten zu den Faktorscores und Personenparametern und stellt die Basisanforderung an jede psychometrische Messung dar (Bond & Fox, 2013). Die Bestimmung der Validität stellt die inhaltliche Interpretation des Messwerts als Maß des Lese-

verstehens sicher und die Reliabilität gewährleistet, dass das Leseverstehen mit einer akzeptablen Genauigkeit erfasst werden kann. Die vorgestellten Analysen erlauben den Einsatz des BYLETs also sowohl für weitere Forschungsvorhaben, als auch als Instrument der Einzelfalldiagnostik. Im Unterschied dazu macht auch eine gelungene Kompetenzmodellierung keine Aussage über den Erwerbsprozess der Kompetenz. Denn es lässt sich auch im Falle eines perfekten Vorhersagemodells für die Schwierigkeit der zugrunde liegenden Items und einer eindimensional gemessenen Kompetenz nicht ableiten, dass der Erwerb der Kompetenz genau jenem Muster folgt oder folgen sollte, das über die Abfolge von Kompetenzstufen beschrieben wird (Isaac & Hochweber, 2011). Das heißt, es kann aus einzelnen Messwerten des BYLETs nicht abgeleitet werden, wie ein Kind trainieren muss, um sein Leseverstehen zu verbessern.

3.5.3 Limitationen

Einschränkungen der hier getätigten Interpretationen der Analyseergebnisse ergeben sich aus der Zusammensetzung der Stichprobe. Diese war anteilsmäßig von schwachen Lesenden dominiert. Nur etwa 3-5% der Stichprobe besuchte bereits eine weiterführende Schule. Es ist daher nicht auszuschließen, dass sich die Ladungsstruktur verändert, wenn nur oder mehr ältere Kinder in der Stichprobe sind. So ist es zum Beispiel denkbar, dass dann mehr Kinder auch die schwierigeren Items der Kompetenzstufen III und IV lösen und sich diese besser als eigene Faktoren in der Datenstruktur identifizieren lassen. Es scheint nämlich so zu sein, dass gerade die Items der Kompetenzstufen III und IV der späteren Textabschnitte C und D besonders selten gelöst wurden und über nur schwache Zusammenhänge mit den anderen Faktoren verfügten. Eine zweite Limitation bezieht sich auf die Anwendbarkeit des BYLETs in der Praxis. Durch die Modellierung mit Modellen, welche für die Items Ladungen schätzen, stellt der Summenwert keine suffiziente Statistik der Personenfähigkeit dar (Bond & Fox, 2013). Dies bedeutet, dass die Auswertung des BYLETs nicht oder nur schlecht “per Hand” erfolgen kann. Stattdessen ist eine computerisierte Auswertung nötig, welche zum Beispiel in einer Webseite oder einem Computerprogramm implementiert werden muss.

3.5.4 Diskussion unter entscheidungstheoretischer Perspektive

Diese Arbeit entwickelt ein Treatmententscheidungsmodell. Hierfür liefert die Messung die Information über die Kovariablen und in der Anwendung auf die Lesetrainingswahl auch die Informationen für das Zielkriterium. Eines der Zielkriterien stellt den Zuwachs im Leseverstehen dar. Die Treatmententscheidung wird auf die Kovariablen bedingt und bezüglich des Zielkriteriums optimiert. Die Messunsicherheit führt also im Rahmen des Treatmententscheidungsmodells zu Unsicherheit in der

bedingten Zielkriteriumsverteilung. Inwiefern also die geringeren Reliabilitäten der Kompetenzstufen II - IV zu unsichereren Entscheidungen führen, hängt von der bedingten Verteilung ab. Diese wiederum ist durch die Form der Nutzenfunktion bestimmt, welche Studie 2 dieser Arbeit behandelt. Ob eine Berücksichtigung aller Kompetenzstufen einen zusätzlichen Gewinn in der Nutzenanalyse der Treatments bringt, wird ebenfalls in Studie 2 erörtert. Hier wird sich zeigen, welche Kompetenzstufen die Differenzen der Fortschritte nach verschiedenen Trainings am besten vorhersagen - welche Kompetenzstufen also über den höchsten differential Payoff verfügen.

3.5.5 Ausblick

Als Ausblick ergibt sich also zunächst eine Nutzung der BYLET-Messwerte für entscheidungstheoretische Fragestellungen innerhalb dieser Arbeit. Außerhalb dieser Arbeit wäre eine Online-Implementation mit einer sich ständig aktualisierenden Normierung und Modellierung des BYLETs wünschenswert. Für die "einfache" Anwendung des Tests, wie sie für die Unterrichtspraxis benötigt wird, könnte nur die Kompetenzstufe V ausgewertet werden, und damit eine Normierung erstellt werden. So könnte der Test auch analog ausgewertet werden. Ob sich hier auch weitere günstige psychometrische Eigenschaften wie Messinvarianz annehmen lassen, können zukünftige Modellierungen zeigen.

4 Studie 2: Umsetzung und Anwendung des TreaDeMs - die Nutzenfunktion

Wie eingangs eingeführt, besteht das in dieser Arbeit erarbeitete Entscheidungsmodell aus drei Komponenten: einer Messung, einer Nutzenfunktion und einer Entscheidungsfunktion. In der ersten Studie wurde das Messmodell beschrieben, mit welchem das Leseverstehen auf vier Dimensionen erfasst werden kann. Die nun folgende, zweite Studie befasst sich mit der Nutzenfunktion. Man könnte sagen, sie liefert die Entscheidungsmatrix eines klassischen entscheidungstheoretischen Problems. Die Wahrscheinlichkeit für das Eintreffen der Umweltzustände wird hier noch nicht berücksichtigt. Tatsächlich lässt sich die hier geschätzte Nutzenfunktion nicht als Matrix schreiben. Denn sowohl die Umweltzustände, also die Leseleistungen vor den Trainings, als auch die Zielkriterien, nämlich die Leseleistungsfortschritte nach den Trainings, werden durch stetige Variablen abgebildet. Es wird hier am Beispiel der Entscheidung zwischen verschiedenen Lesetrainings gezeigt, wie man die Nutzendimension festlegen kann und wie man Evaluationsstudiendaten verwenden kann, um den erwarteten Nutzen der Trainings für einzelne Personen vorherzusagen. Als Nutzendimension wird der Leistungszuwachs in Leseflüssigkeit und Leseverstehen in zwei standardisierten Schulleistungstests definiert. Die Nutzenfunktion selbst beruht auf den in Studie 1 berechneten Personenwerten der Item-Response-Modelle und stellt einen funktionalen Zusammenhang zwischen Ausgangsleistung und demografischen Merkmalen und dem Lesekompetenzzuwachs nach zwei verschiedenen Lesetrainings her. Betrachtet werden Trainings aus der FiLBY-Studie: ein Leseflüssigkeitstraining mit Sachtexten (FiLBY-2) und ein Lesestrategietraining anhand von literarischen Ganzschriften (Lektüren) (FiLBY-3). Es geht also konkret um die Frage, welches Training gegeben einer bestimmten Ausgangsleistung in Leseflüssigkeit und Leseverstehen mit höheren Zuwächsen in Leseflüssigkeit und Leseverstehen assoziiert ist. Dabei wird berücksichtigt, dass sich diese Zusammenhänge in verschiedenen Teilpopulationen unterscheiden können. Als Teilpopulationen wurden dabei Mädchen und Jungen, sowie Kinder mit und ohne Migrationshintergrund betrachtet.

Im Folgenden wird zunächst wieder auf die Definition und Bedeutung des Nutzens aus entscheidungstheoretischer und diagnostischer Sicht eingegangen. Im Anschluss werden die Erkenntnisse auf das konkrete Beispiel des Nutzens von Lesetrainings im Sinne von Lesekompetenzzuwächsen übertragen.

4.1 Operationalisierungen des Nutzens

Bevor eine Nutzenfunktion aufgestellt werden kann, muss definiert werden, was genau der Nutzen sein soll. Soll ein konkretes entscheidungstheoretisches Problem gelöst werden, so reicht eine mathematische Definition des Nutzens nicht aus. Der Nutzen muss auch operationalisiert werden. Das bedeutet, dass festgelegt werden muss, welche beobachtbaren und messbaren Ereignisse den Nutzen definieren.

4.1.1 Klassische, statistische Perspektive

Im entscheidungstheoretischen Kontext ist ein Nutzen der Output einer Nutzenfunktion (Peterson, 2017). In einem Entscheidungsproblem der Form $(\mathbb{A}, \Theta, u(.))$ weist eine Nutzenfunktion jeder möglichen Kombination aus Aktionen ($\mathbb{A}$) und Umweltzuständen (Θ) einen Nutzenwert ($u(a_i, \theta_j)$) zu. Der Aktionenraum ($\mathbb{A}$) ist dabei in der Regel eine endliche Menge an diskreten Handlungsoptionen (siehe Wald (1950) und Brown (1986) für Ausnahmen). Θ hingegen beschreibt einen Raum einer kontinuierlichen oder diskreten Zufallsvariable, welche auch vektorwertig sein kann. a_i ist demnach eine der möglichen Aktionen im Entscheidungsproblem, welche vom:n der Entscheidungstragenden ausgewählt werden kann. θ_j ist eine mögliche Realisation eines zum Zeitpunkt der Entscheidung unbekannten Umweltzustands. Wäre es bekannt, welches θ_j sich im Entscheidungsproblem realisiert, so wäre $(u(a_i, \theta_j))$ ein Optimalitätskriterium bezüglich der a_i. Das bedeutet, dass bei bekanntem Umweltzustand dasjenige a_i ausgewählt werden müsste, für welches $(u(a_i, \theta_j))$ den höchsten Wert erreicht. Das Entscheidungsproblem entsteht also erst durch die Zufälligkeit der Umweltzustände. Nutzen werden üblicherweise in sogenannten Nutzentabellen dargestellt. Je nach Form des Entscheidungsproblems kann die Nutzenfunktion stetig oder diskret sein (Peterson, 2017). Stetige Nutzenfunktionen ergeben sich aus stetigen Nutzenwerten und stetigen Umweltzuständen. Diskrete Nutzenfunktionen ergeben sich aus Funktionen mit diskreten Nutzenwerten oder diskreten Umweltzuständen. Stetige Nutzenfunktionen lassen sich grafisch als Kurven darstellen. Dabei wird für jede Aktion eine eigene Kurve gezeichnet. Auf der X-Achse werden dann die Umweltzustände und auf der Y-Achse der Nutzen abgebildet. Abbildung 14 verdeutlicht dies.

Alternativ zum Nutzen kann auch ein Verlust formuliert und anstelle einer Nutzenfunktion eine Verlustfunktion aufgestellt werden. Da sich ein Verlust mathematisch als lineare Transformation eines Nutzens ergibt, soll aber auf diesen Fall nicht weiter eingegangen werden.

Der Nutzen hängt davon ab, wie sich Aktionen unter bestimmten Umweltzuständen auf Zielkriterien auswirken. An einem einfachen Beispiel: nimmt man einmal an, man müsste entscheiden, ob

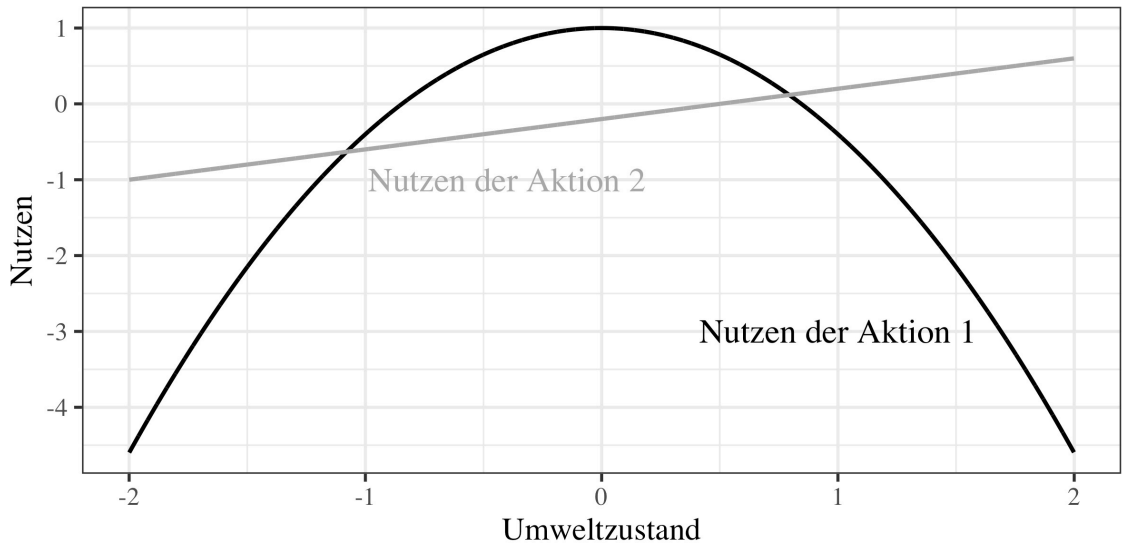


Abbildung 14: Allgemeine Darstellung einer Nutzenfunktion

man zum Campingwochenende entweder eine Regenhose oder eine Sonnencreme mitnimmt. Das Zielkriterium soll das Wohlbefinden während des Campings sein. Dann hängt der Nutzen der zwei möglichen Aktionen "Regenhose mitnehmen" oder "Sonnencreme mitnehmen" maßgeblich vom Umweltzustand "Wetter" ab. Regnet es, so wäre man mit einer Regenhose gut beraten, scheint hingegen die Sonne, so wäre eine Sonnencreme von größerem Nutzen. Konkrete Nutzenwerte zuzuordnen und eine vollständige Nutzenfunktion zu spezifizieren, ist dabei nicht nur in diesem kleinen fiktiven Beispiel ein zentrales Element entscheidungstheoretischer Modellierungen.

In den Wirtschaftswissenschaften blickt man auf entscheidungstheoretische Forschung von über 250 Jahren zurück. Sie versucht auf Basis von Experimenten und Beobachtungen eben diese Nutzenfunktionen aus dem Verhalten von Menschen abzuleiten und Regelmäßigkeiten in menschlichen Entscheidungen zu entdecken, die sich mit mathematischen Modellen beschreiben lassen (z. B. Peterson, 2017).

Als Grundannahme wird dabei stets angenommen, dass Menschen Kombinationen aus Aktionen und Umweltzuständen Nutzenwerte zuschreiben, Umweltzustände mit gewissen Wahrscheinlichkeiten eintreten und Menschen auf Basis dieser Informationen ihren persönlichen Erwartungsnutzen maximieren. Der Erwartungsnutzen ist dabei als erwarteter Nutzen bei gegebenen Wahrscheinlichkeiten für bestimmte Umweltzustände und gegebenem Nutzen für Kombinationen aus Aktionen und Umweltzuständen definiert (z. B. Peterson, 2017). Ein klassisches Entscheidungsproblem besteht also aus Aktionen, Umweltzuständen, Nutzenwerten. Hinzu kommt eine Entscheidungsfunktion, die eine Nutzenfunktion in die Entscheidung überführt, also in den Aktionenraum zurück projiziert. In einem datengestützten Entscheidungsproblem werden bei der Entscheidungsfunktion Wahrschein-

lichkeiten über Umweltzuständen berücksichtigt, in einem datenfreien Entscheidungsproblem nicht. In dieser Studie soll nur der erste Schritt betrachtet werden: die Verbindung von Aktionen und Umweltzuständen in der Form von Nutzenfunktionen.

Denkt man zurück an das oben eingeführte Beispiel zum Equipment für den Campingausflug, können nun Kombinationen aus Aktionen und Umweltzuständen Nutzenwerte zugeordnet werden. So könnte etwa der Zustand "man hat eine Regenhose dabei und es regnet" mit einem Nutzen von fünf Einheiten belegt sein, während der Zustand "man hat eine Regenhose dabei und es scheint die Sonne" mit einem Nutzen von minus fünf Einheiten belegt sein könnte. Der Nutzen ergibt sich damit immer aus einer Kombination von Aktion und Umweltzustand, da nur diese beiden gemeinsam einen Zustand definieren, dem ein subjektiver Nutzen zugeordnet werden kann. Es wäre per se weder vorteilhaft eine Regenhose noch eine Sonnencreme mitzunehmen, genauso ließe sich nicht pauschal festlegen, ob Regen oder Sonnenschein per se von Vorteil sind. In der Empirie ist es jedoch für Menschen in der Regel nicht möglich oder sehr schwierig, vollständige Nutzentabellen zu spezifizieren. Denn zur vollständigen Spezifikation von Nutzentabellen muss nicht nur jede Kombination aus Umweltzustand und Aktion bekannt und bewertbar sein, auch sollten die zugeordneten Nutzenwerte bestimmten logischen Ordnungen, wie etwa der Transitivität folgen (Gilboa, 2009). Andere Forschungsfelder blicken bereits auf eine Tradition in der Erstellung von Nutzenfunktionen zurück. Daher werden nun zunächst die dort entwickelten Ansätze vorgestellt. Anschließend wird eine Perspektive entwickelt, um die Erkenntnisse auf das Feld der Diagnostik zu übertragen.

4.1.2 Verhaltensökonomische Perspektive

In der verhaltensökonomischen Perspektive gibt es zwei Ansätze, um Nutzenfunktionen zu spezifizieren. Im Ersten Ansatz wird unter Annahme von verschiedenen Axiomen bei gegebenen Präferenzen bezüglich Zuständen auf subjektive Wahrscheinlichkeitsverteilungen der Umweltzustände geschlossen. So könnte zum Beispiel eine Regenhose dabeizuhaben, wenn es regnet, gegenüber einer Sonnencreme dabeizuhaben, wenn es nicht regnet, gleich gewichtet sein. Wenn sich dann eine Person bei ungewisser Wetterlage für die Sonnencreme entscheidet, so scheint die Person Sonne für wahrscheinlicher zu halten. Alternativ wird bei gegebenen Wahrscheinlichkeitsverteilungen auf subjektive Wertigkeiten von Ereignissen geschlossen. Manche Ansätze (z. B. Savage, 1972) versuchen auch beides gleichzeitig zu tun. Zentral sind dabei fast immer Gedankenexperimente, bei denen sich Proband:innen zwischen verschiedenen Szenarien, oder Lotterien entscheiden müssen und bei denen der Nutzen fast immer finanzieller Natur ist. Im gesamten betriebswirtschaftlichen Kontext wird der Nutzen fast ausschließlich als ein monetärer Nutzen verstanden, oder zumindest ein Nut-

zen, der sich in funktionalem Zusammenhang mit einem monetären Nutzen befindet (Bernoulli, 1738/1954; Kahneman & Tversky, 1984). Das Ziel dieses Forschungsfelds ist dabei die Generierung von deskriptiven Theorien. Es geht darum zu beschreiben und zu erklären, wie und warum sich Menschen entscheiden.

4.1.3 Diagnostische Perspektive

Die hier gewählte Perspektive unterscheidet sich in den letztgenannten Punkten von der verhaltensökonomischen Perspektive. Als erstes ist zu nennen, dass diese Arbeit keine deskriptive Theorie erarbeiten möchte. Sie soll stattdessen eine normative Perspektive einnehmen. Es soll eine Nutzenfunktion aus Daten geschätzt werden und aus der Nutzenfunktion in Studie 3 eine optimale Einzelfallentscheidung abgeleitet werden. Es soll also nicht deskriptiv untersucht werden, wie Diagnostiker:innen Entscheidungen bezüglich verschiedener Treatments treffen. Stattdessen sollen aus Evaluationsstudiendaten Empfehlungen abgeleitet werden, wie zukünftig Diagnostiker:innen Entscheidungen treffen sollten. Dafür muss im Gegenzug vorausgesetzt werden, dass Entscheidungstragende die hier verwendeten Zielkriterien auch persönlich wertschätzen. Nur so kann daraus eine Akzeptanz der Nutzenfunktion und darauf aufbauend eine Akzeptanz der Entscheidungsfunktion folgen. Während in den Wirtschaftswissenschaften große Einigkeit darin besteht, dass Menschen nach dem größtmöglichen finanziellen Nutzen streben, ist in den Bildungswissenschaften oft sehr umstritten, welche Werte als erstrebenswert zu erachten sind und ob sich diese auf einer einzigen Dimension, die man mit Bildungserfolg betiteln könnte, abbilden lassen (vgl. Archer, 2013; Marzano & Kendall, 2006).

Eine zentrale Herausforderung für entscheidungstheoretische Betrachtungen in den Bildungswissenschaften ist es also, die Einheit des Nutzens festzulegen. Eine Definition von Nutzen als finanzieller Gewinn, wie in der verhaltensökonomischen Perspektive üblich, bietet sich für eine Anwendung der Entscheidungstheorie im Gebiet der Diagnostik nur sehr eingeschränkt an. Außerhalb von Unternehmenskontexten, welche bei der Personalauswahl entscheidungstheoretisch vorgehen könnten, ist der Nutzen verschiedener diagnostischer Entscheidungen nur sehr schwer in Geldwerten auszudrücken. Gerade in den Bildungswissenschaften sind die Zusammenhänge zwischen Ausgaben und Revenues zeitlich zerdehnt und über viele weitere Umstände vermittelt und beeinflusst. Es lässt sich nicht berechnen, wie viel Euro es "wert ist," wenn ein Schulkind eine Standardabweichung in einem standardisierten Lesetest an Fortschritten macht. Im Rahmen der bildungswissenschaftlichen Diagnostik müssen also andere Zielkriterien als der finanzielle Gewinn gefunden werden.

Zudem hängt die Wahl der Nutzendimension direkt von der Art der diagnostischen Entschei-

dung ab. Van der Linden (1980) folgend kann jede diagnostische Entscheidungen als Selektions-, Klassifikations-, Mastery- oder Treatmententscheidung aufgefasst werden. Für Selektionsentscheidungen steht bei den Entscheidungstragenden der Nutzen für eine Institution oder der Nutzen für die Allgemeinheit im Vordergrund. Ein Beispiel ist die Selektion von Personen für ein Studium oder ein Stipendium. Hier sind die öffentlichen Gelder begrenzt und es gilt sie möglichst gewinnbringend für die Allgemeinheit einzusetzen und die Personen mit den höchsten Erfolgschancen auszuwählen. Für Treatmententscheidungen steht demgegenüber der Nutzen des Individuums im Vordergrund. Dabei können zum Beispiel eine Therapieart, ein Training oder das Erhalten von Förderunterricht als Treatments betrachtet werden. Hier besteht bereits vor der Entscheidungssituation der Anspruch auf oder der Bedarf für ein Treatment. Das Entscheidungsproblem beantwortet nur noch die Frage, welches der zur Verfügung stehenden Treatments am geeignetsten für ein Individuum ist.

Will man in der Diagnostik also eine entscheidungstheoretische Perspektive einnehmen, so muss zunächst das Entscheidungsproblem an sich sowie die Einheit des Nutzens definiert werden. Dies wurde jedoch bisher kaum systematisch gemacht (siehe z. B. Cronbach & Gleser, 1965; Rudner, 2009 für Gegenbeispiele).

4.1.3.1 Bisherige Definition von Nutzen im diagnostischen Setting

Bisher gibt es im diagnostischen Bereich insgesamt kaum entscheidungstheoretische Betrachtungen. Darüber hinaus herrscht im pädagogisch-psychologischen Bereich wenig Einigkeit, was überhaupt ein Zielkriterien von diagnostischen Entscheidungen sein könnten. Oft bleiben die Formulierungen sehr wage.

Obgleich es schon Überlegungen gibt, neben den Hauptgütekriterien auch Nützlichkeit in engerem Sinne beim diagnostischen Handeln zu berücksichtigen. So berichtet etwa Fisseni (2004), dass sich generell die Frage der Nützlichkeit von Gütekriterien entfernt - hin zu einsatzspezifischen Nutzenkriterien. Ein weiteres entscheidungstheoretisches Modell geht auf Edwards, 1977 zurück und nennt sich MAUT (Edwards, 1977). Diese berechnen den Nutzen mittels einer linearen Gewichtung von Attributen, während deren Nutzenbewertung durch Expertenratings erfolgt. Daher wird im Folgenden ein eigener Ansatz zur Implementation von einer Nutzenfunktionsgenerierung im diagnostischen Kontext vorgestellt, der sich an Überlegungen von Cronbach und Gleser (1965), dem Konzept der Aptitude-Treatment-Interaktion (Cronbach & Snow, 1977) und allgemeinen entscheidungstheoretischen Prinzipien orientiert.

4.1.3.2 Definition des Nutzens im diagnostischen Entscheidungsproblem

Dafür muss klar definiert werden, was das Zielkriterium sein sollte. Im Bildungskontext steht häufig der Kompetenzzuwachs an erster Stelle. Natürlich ist es auch legitim, zeitliche oder finanzielle Ressourcen

mit einzubeziehen und nicht zuletzt pädagogisch-psychologische Aspekte wie Lernmotivation, Freude am Lernen und so weiter berücksichtigen zu wollen. In der Regel werden Kosten und Nutzen eines Treatments direkt verrechnet. Im verhaltensökonomischen Kontext, wo sich Menschen zwischen verschiedenen Lotterien entscheiden müssen, werden die potenziellen Gewinne verschiedener Spiele mit den Teilnahmekosten direkt zu einem Nutzen verrechnet. Dies stellt im bildungswissenschaftlichen Kontext ein Problem dar. Hier sind die Nutzenwerte eher diffus, wie z. B. abstrakte Punkte auf Standardskalen, während die Kosten in der Regel nicht diffus sind. Testmaterialien werden in Euro bezahlt, Zeit, die für das Testen oder verschiedene Treatments aufgewendet werden muss lässt sich in Stunden angeben, etc. Diese auf einer gemeinsamen Skala zu verrechnen ist schwierig (Cronbach & Gleser, 1965). Zudem wird es in der Praxis sehr schwer sein, alle relevanten Aspekte zu berücksichtigen. Es kann aber dennoch eine Abwägung und Verrechnung verschiedener Zielkriterien zu einer Nutzendimension dem Aufstellen einer Nutzentabelle vorausgehen. Mögliche Herangehensweisen dafür finden sich in der Literatur zur multikriteriellen Entscheidungstheorie (Fishburn & Keeney, 1974; Morton & Fasolo, 2009) und sollen hier nicht weiter erörtert werden. Im Anwendungsbeispiel dieser Studie unterscheiden sich die Treatments nur in den Inhalten und nicht auf anderen Dimensionen wie der Trainingsdauer, dem Preis der Materialien oder dem Vorbereitungsaufwand. Deswegen wird ausschließlich der Lernzuwachs als Zielkriterium gewählt und der Nutzen als Zuwachs an normierten Testpunkten definiert. Da nun die Einheit des Nutzens definiert ist, befasst sich der nächste Abschnitt mit den Rahmenbedingungen der diagnostischen Treatmententscheidung.

4.1.3.3 Definition der Entscheidungssituation Grundsätzlich gilt, für die systematisch, statistische Herangehensweise ist eine saubere, mathematische Definition der Entscheidungssituation unumgänglich. Dafür muss definiert werden, was die Aktionen des Entscheidungsproblems sind. Im hier betrachteten Fall sollen Treatments für Personen ausgewählt werden. Weiterhin muss beachtet werden, ob es eine Auswahlquote gibt, ob das Treatment dosierbar oder fest ist, und ob es sich um eine einschrittige oder mehrschrittige Entscheidung handelt. Auch ob Fehlentscheidungen in einem zweiten Schritt korrigiert werden können, nimmt Einfluss auf die Formalisierung des Entscheidungsproblems. Zuletzt muss entschieden werden, ob Informationen über die Umweltzustände gewonnen werden können. Falls dies der Fall ist, muss entschieden werden, wie diese Informationen gewonnen werden und ob die Kosten der Informationsbeschaffung im Entscheidungsproblem berücksichtigt werden sollen. Eine fast vollständige Abhandlung der verschiedenen Szenarien findet sich bei Cronbach und Gleser (1965).

In dieser Arbeit wird nur der folgende Fall betrachtet:

- Treatmententscheidung
- Informationsgewinnung durch Testen mit standardisierten Leistungstests
- keine Beachtung der Testkosten
- keine Quote
- keine Beachtung der Treatmentkosten
- Entscheidung zwischen zwei Treatments
- Single stage Entscheidung
- Treatment nicht dosierbar

Sind das Zielkriterium und die Rahmenbedingungen definiert, können Überlegungen zur Schätzung der Nutzenfunktion getätigt werden.

4.2 Die Schätzung des Nutzens

Wie der Nutzen statistisch geschätzt werden kann, hängt maßgeblich vom Design der Studie ab, aus welcher die Daten zur Nutzenschätzung stammen. Grundsätzlich soll ein funktionaler Zusammenhang zwischen individuellen Merkmalen und Trainingserfolg unter verschiedenen Treatments hergestellt werden. Es muss also für jedes Treatment eine eigene Nutzenfunktion geschätzt werden. Dies muss jedoch nicht in verschiedenen Modellen geschehen. Beruht die Nutzenschätzung auf Evaluationsstudiendaten, wie in dieser Studie, so geht es darum eine kontrafaktische Realität zu schaffen (Cook, Campbell, & Shadish, 2002). Die Frage lautet konkret, wie hätte sich die Leistung eines Kindes in einer alternativen Realität mit einem anderen Treatment entwickelt? Dabei ist aus theoretischer Perspektive davon auszugehen, dass dieser Zusammenhang nicht linear ist. Möglicherweise unterscheidet sich der Zusammenhang zudem für unterschiedliche Teilpopulationen (Cummins, 2012; Logan & Johnston, 2009; O'Connor et al., 2002). Die meisten Trainings oder auch Treatments allgemein wirken besonders gut, wenn bestimmte Ausgangsbedingungen gegeben sind. Diese zu identifizieren soll Ziel der Nutzenfunktionsschätzung sein. Die Schätzung der Nutzenfunktion ist also gleichzeitig eine Bedingungsanalyse und wird in Bezug auf Interventionen auch unter dem Begriff der Aptitude-Treatment-Interaktion (ATI) geführt. Hier werden klassischerweise einen Treatmenterfolg bedingende Aptitude-Variablen in linearen Regressionen als Moderatoren eingesetzt (Cronbach & Snow, 1977; Freebody & Tirre, 1985; Fuchs et al., 2019). Alternativ und

allgemeiner lassen sich diese Zusammenhänge in der Form von Manski (2004) darstellen.

$$W(z, p, Q) \equiv \int \{\sum_{\xi \in \chi} p(x = \xi) \cdot \sum_{t \in T} z(t, \xi, \psi) \cdot E(u[y(t), t, \xi] | x = \xi)\} dQ(\psi)$$

mit:

- $W(z, p, Q)$ = Risiko der Entscheidungsregel z
- ξ = Kovariablenvektor
- $u([y(t), t, \xi] | x = \xi)$ = Nutzen von Treatment t, wenn Kovariablenvektor $\xi = x$
- ψ = alle möglichen Datensätze, die sich dem Daten generierenden Prozess folgend realisieren könnten

Allerdings gilt hier, dass es für die Wahrscheinlichkeiten, welche die Treatmentwahl beschreiben, in der Regel keine geschlossenen Ausdrücke gibt und sie numerisch schwer zu berechnen sind. Daher schlägt Manski vor, Bounds für die bedingten Treatmenteffekte zu berechnen. Eine Alternative zu Manskis Ansatz ist eine Modellierung der empirischen Zusammenhänge über statistische Modelle. Hier werden anstelle der bedingten Verteilungen, bedingte Verteilungen unter den Modellrestriktionen geschätzt. Zur Modellierung der Zusammenhänge können lineare, aber auch allgemeinere funktionale Zusammenhänge angenommen werden (Novick & Lindley, 1978). Der Ansatz der Modellierung als funktionaler Zusammenhang wurde unter anderem von Cronbach und Gleser (1965), Cronbach und Snow (1977) oder Van der Linden und Kollegen empfohlen (1998). Lineare Zusammenhänge führen dazu, dass es unter Konstanthaltung aller anderen Prädiktoren maximal einen Schnittpunkt der Treatmentgeraden pro Prädiktor gibt. An diesem Schnittpunkt übersteigt dann der Nutzen eines Trainings den eines anderen. Da dies den eingangs postulierten komplexen Wirksamkeitsbedingungen nicht gerecht werden kann, scheint es ratsam, Modelle zu wählen, die flexible funktionale Zusammenhänge modellieren können. Hier bieten sich zum Beispiel nonparametrische Regressionen mit Splines, oder allgemeiner generalisierte additive Modelle (GAM) an. Diese Modelle erhalten zudem die Stetigkeit der Funktion. Es erscheint nämlich inhaltlich fragwürdig, dass eine stetige Ausgangsbedingung (Variable) sich in einer Sprungfunktion mit den kontinuierlichen Nutzenwerten befindet.

Im vorangegangenen Abschnitt wurde nun die Nutzenfunktion allgemein definiert. Als nächstes gilt es Kriterien zu finden, die eine gute Nutzenfunktion ausmachen. Dabei sollten neben klassischen Modellgütekriterien wie Modellpassung, unverzerrte Schätzung und hohe Varianzaufklärung auch noch weitere Kriterien berücksichtigt werden. Eine Nutzenfunktion soll nicht nur die Zielvariable gut erklären, auch soll sie möglichst gut zwischen den Treatments differenzieren. So beschreiben schon Cronbach und Gleser (1965), dass es auf den differential Payoff ankommt. Entscheidend ist nicht, ob ein Messwert einen Nutzen vorhersagt, sondern ob er die Unterschiede im Nutzen zwischen

den verschiedenen Treatments vorhersagen kann, also ob er einen Benefit von einem Treatment gegenüber einem anderen vorhersagen kann. Das bedeutet, der Nutzen muss erstens möglichst genau geschätzt werden und zweitens sich zwischen den Trainings möglichst stark unterscheiden. Aber das heißt wiederum, dass vor allem Interaktionseffekte der Prädiktoren mit der Treatmentvariable von Relevanz sind, selbst wenn es die Haupteffekte nicht sind. Zum Beispiel wäre es möglich, dass Kinder mit Migrationshintergrund von einem Treatment mehr profitieren als von einer anderen, auch wenn Kinder mit Migrationshintergrund an sich nicht generell stärker von Treatments allgemein profitieren.

Es ist dabei aus inhaltlichen Gründen naheliegend, Prädiktoren auszuwählen, die eng verwandt mit den Zielkriterien sind. Aber es ist nicht zwingend notwendig. Theoretisch können sich Zielkriterien auch mit Variablen vorhersagen lassen, die keine Augenscheinvalidität besitzen. Mit Blick auf die späteren Anwenderinnen und Anwender erscheint es jedoch ratsam, zumindest solche Kriterien auszuwählen, deren Vorhersagekraft auch inhaltlich begründbar ist.

4.3 Der Gegenstand der Nutzenfunktion: Lesekompetenzentwicklung

Im vorherigen Abschnitt wurde allgemein beschrieben, wie sich Nutzenfunktionen im bildungswissenschaftlichen Kontext definieren und schätzen lassen. Nun soll konkretisiert werden, wie man eine Nutzenfunktion zur Bestimmung des Lesekompetenzzuwachses durch verschiedene Lesetrainings modellieren kann. Dazu ist es unerlässlich, bestehendes Wissen zur Lesekompetenz und Lesekompetenzentwicklung zu integrieren und auf Basis dessen passende Prädiktoren und Zielkriterien auszusuchen.

4.3.1 Lesekompetenz

Lesekompetenz im engeren Sinn setzt sich aus Dekodieren, Leseflüssigkeit und Leseverstehen zusammen. Die Leseflüssigkeit bildet dabei eine Brückenfunktion zwischen dem Dekodieren und dem Leseverstehen. Weiter gefasste Definitionen beziehen zudem soziale und motivationale Komponenten von Lesekompetenz mit ein, wie etwa eine gelungene Anschlusskommunikation oder ein positives Selbstbild als Leserin oder Leser (Rosebrock & Nix, 2017). In dieser Studie wird jedoch eine technischere Perspektive eingenommen. Lesekompetenz soll ausschließlich als Zusammenspiel zwischen Dekodieren, Leseflüssigkeit und Leseverstehen verstanden werden.

Dekodieren bedeutet, den Graphemen die passenden Phoneme zuzuordnen und ihren semantischen Inhalt zu erfassen. Dekodieren bildet die Basis des Leseverstehens (Dehaene, 2014).

Leseflüssigkeit beschreibt die Fähigkeit akkurat und automatisiert mit angemessener Betonung und Geschwindigkeit zu lesen (z. B. Hudson, Pullen, Lane, & Torgesen, 2008; Rasinski, Reutzel, Chard, & Linan-Thompson, 2011). Eine ausführlichere Diskussion zur Definition von Leseflüssigkeit findet sich zum Beispiel bei Kuhn, Schwanenflugel und Meisinger (2010).

Leseverstehen im engeren Sinn setzt sich aus lokaler und globaler Kohärenzbildung zusammen. Das Ziel ist die Bildung eines mentalen Modells, in welches neben den Textinhalten auch Wissen über Darstellungsformen und Superstrukturen integriert. Lokale Kohärenz bezieht sich auf das Verstehen einzelner Sätze. Ein Satz besitzt eine sogenannte propositionale Struktur. Das heißt, dass nicht mehr die einzelnen Wörter die grundlegenden Informationseinheiten darstellen, sondern Wörtergruppen. Diese Wörter sind über ihre Bedeutung, aufgrund von grammatikalischen Strukturen oder über eine Kombination von Bedeutung und Grammatik miteinander verbunden. Diese Verbundenheit zu verstehen wird lokale Kohärenzbildung genannt (Lenhard & Schneider, 2009). Bei der globalen Kohärenzbildung werden größere Texteinheiten erarbeitet. Kohärenzbildung wird auch als hierarchiehoher Prozess bezeichnet (vgl. Richter & Christmann, 2002). Hierarchiehohe Prozesse sind zudem durch ihren potenziell strategisch-zielorientierten Charakter gekennzeichnet. Die durch die lokale Kohärenzbildung gebildeten Propositionen werden mit Vorwissen verknüpft und zu einem sogenannten Situationsmodell integriert (Kintsch & Van Dijk, 1978).

Leseverstehen ist also ein Prozess, der sich hierarchisch aus der Fähigkeit Buchstaben und Wörter zu dekodieren, einer Automatisierung dieses Prozesses und der Integration von Einzelinformationen zusammensetzt. In der Entwicklung des Leseverstehens sind daher die hierarchieniederen Prozesse den hierarchiehöheren Prozessen zumindest teilweise vorgeschaltet (Chall, 1983; LaBerge & Samuels, 1974).

4.3.2 Lesekompetenzentwicklung

Die Lesekompetenzentwicklung wird in verschiedene Phasen eingeteilt, in denen die aufeinander aufbauenden Teilkompetenzen schwerpunktmäßig erworben werden. Die Phasen der Lesekompetenzentwicklung sind die Dekodierphase, die Phase der Leseflüssigkeitssteigerung und die Phase der zunehmenden Inferenzleistung beim Lesen. Sie gehen ineinander über, jedoch limitieren Fähigkeiten der unteren Phasen die Entwicklung der weiteren, höheren Lesekompetenzen. Theoretische Überlegungen (LaBerge & Samuels, 1974) und empirische Daten (Fuchs et al., 2019) legen nahe, dass der Zusammenhang zwischen diesen Lernvoraussetzungen und der Wirksamkeit von spezifischen Lesetrainings nicht linear ist. In der Literatur finden sich viele Hinweise auf sensible Phasen und Kompetenzbereiche, in denen einzelne Maßnahmen besonders effektiv sind (Indrisano & Chall,

1995; McDonald Connor et al., 2009). Aus der Betrachtung des Entwicklungsprozesses leitet sich ab, dass verschiedene Lesekompetenzen (Dekodieren, Flüssigkeit, Inferenzen) nacheinander und teilweise auch unabhängig voneinander trainiert werden können und trainiert werden sollten. Es ist zu erwarten, dass sich deswegen Leseflüssigkeitstrainings stärker auf die Leseflüssigkeit und Lesestrategietrainings stärker auf das Leseverstehen auswirken. Es ist außerdem zu erwarten, dass etwa ein auf Inferenzen ausgelegtes Lesestrategietraining wenig wirksam ist, wenn Lernende damit trainieren, deren Leseflüssigkeit noch unzureichend ausgebildet ist.

4.3.3 Lesekompetenz trainieren

Eine Übersicht, wie Leseflüssigkeit und Leseverstehen im Grundschulalter trainiert werden sollten, gibt Tabelle 17 (Duffy, 1993; Hudson, Lane, & Pullen, 2005; Kuhn & Stahl, 2003; Morrow, 2019; Rosebrock & Nix, 2017).

Tabelle 17: Merkmale der Effektivität von Lesetrainings

Effektivitätsmerkmale	
Leseflüssigkeitstrainings	Leseverstehenstrainings
Gelegenheit zum Lesen wird geben	Strategievermittlung
Wiederholtes Lesen von Texten	Modellierung von Strategien
Modellierung des Lesens	Zeit für das Üben wird eingeplant
Individuelle Unterstützung des Lesens	Anschlusskommunikation und Reflexion finden statt
Lautes Lesen	
Wiederholdung der Strategien zum Dekodieren	

Anmerkung. Vgl. Duffy, 1993; Hudson et al., 2005; Kuhn & Stahl, 2003; Morrow, 2019; Rosebrock & Nix, 2017.

Doch auch demografische Merkmale der Lernenden wie etwa das Geschlecht oder ein Migrationshintergrund können sich differenziell auf die Wirksamkeit verschiedener Trainingsmaßnahmen auswirken (Lervåg & Aukrust, 2010). Beim Einfluss von demografischen Merkmalen findet wahrscheinlich eine Vermittlung über soziale und motivationale Ebene statt. So postulieren Logan und Johnston (2010), dass Mädchen deswegen im Durchschnitt schneller und besser lesen lernen als ihre männlichen Altersgenossen, weil das Lesen kulturell als eine weibliche Tätigkeit aufgefasst wird, es über-

wiegend weibliche Lesevorbilder gibt und altersgerechte Texte thematisch stärker auf die Interessen von Mädchen ausgerichtet sind (Millard, 2002). Es wurde zwar versucht mit den FiLBY-Trainings ein für beide Geschlechter gleichermaßen ansprechendes Programm zu entwickeln, dennoch soll in dieser Studie untersucht werden, ob es für Jungs und Mädchen differentiell wirkt. In Bezug auf auf den Migrationshintergrund zeigte sich in vorangegangenen Studien, dass sich dieser vor allem dann negativ auf den Leselernprozess auswirkt, wenn er mit sprachlichen Defiziten oder der Zugehörigkeit zu bildungsbenachteiligten Bevölkerungsgruppen assoziiert ist (Cummins, 2012). Daher sollte auch überprüft werden, ob sich die Ausgangsleistungen bei Kindern mit und ohne Migrationshintergrund unterschiedlich auf den Trainingserfolg auswirken. Trainings, die diese wissenschaftlich belegten Wirksamkeitskomponenten berücksichtigen, sind die beiden FiLBY-Trainings (FiLBY-2 und FiLBY-3), welche im Folgenden kurz vorgestellt werden sollen.

4.3.3.1 FiLBY-2 Training Das FiLBY-2-Training basiert auf dem bereits erfolgreich evaluierten Konzept "Lesen durch Hören" (Gailberger, 2011) und gliedert sich in zwei Teile. Diese Studie bezieht sich jedoch nur auf den zweiten Teil. Ziel des FiLBY-2-Trainings ist es, dass Schüler:innen der zweiten Klassenstufe eine angemessene Leseflüssigkeit entwickeln. Um dies zu erreichen wurden 62 altersangemessene Sachtexte geschrieben. Mit diesen sollten die Lehrkräfte mit ihren Klassen täglich über mindestens sechs Wochen hinweg üben. Die Sachtexte werden dabei über eine Audioaufnahme in verschiedenen Geschwindigkeiten vorgelesen. Das Vorlesen, sowie das wiederholte Lesen erleichtert den Aufbau eines Sichtwortschatzes, da die Zuordnung von Graphemen und Phonemen wiederholt eingeprägt werden kann. Durch das Vorlesemodell erhalten die Schüler:innen zudem eine gute Zielvorstellung von der erwarteten Lesegeschwindigkeit, sowie von der korrekten Prosodie. Im Training bekommen die Schüler:innen den Text zweimal vorgelesen. Beim ersten Durchlauf deuten die Schüler:innen jeweils auf die Wörter, welche das Vorlesemodell gerade liest. Beim zweiten Durchlauf lesen sie leise flüsternd mit. Anschließend im dritten Durchlauf lesen die Schüler:innen den Text selbst. Beim selbstständigen Lesen wird der Leseprozess durch eine:n Tandempartner:in unterstützt. So können etwaige Lesefehler erkannt und korrigiert werden. Die am Projekt beteiligten Lehrkräfte erhalten eine dreitägige Präsenzfortbildung. Diese Fortbildung wird durch eine Lernplattform mit sieben Modulen zu Grundlagen der Leseförderung und deren Umsetzung in FiLBY ergänzt, von denen zwei Module (Was ist Lesekompetenz? Wie fördert man die Leseflüssigkeit?) vor der Durchführung verpflichtend absolviert werden müssen. Für die tägliche Praxis steht eine ausführliche Lehrerhandreichung mit Unterrichtsskizzen und weitere Materialien zur Verfügung (Schilcher, Wild, & Steinert (2019); genauere Einblicke in das Training können hier eingesehen werden: www.lesen.bayern.de/filby2).

4.3.3.2 FiLBY-3 Training

Während in FiLBY-2 die Leseflüssigkeit als Basis des Leseverstehens fokussiert wird, steht zu Beginn der dritten Klassenstufe (FiLBY-3; Wild, Schilcher & Steinert, 2019) die Vermittlung effektiver literarischer (Vorwissen aktivieren, Figur, Ereignis und Situation untersuchen) Strategien im Vordergrund. Darüber hinaus werden auf die Informationsentnahme und Informationsverarbeitung ausgerichtete Lesestrategien (Vorwissen aktivieren, Überfliegen, Visualisieren) vermittelt. Die Strategien werden über mindestens vier Wochen hinweg täglich geübt. Im Sinne des Scaffoldings wird der Strategieeinsatz von der Lehrkraft unterstützt. Im Unterrichtsgespräch und in Tandemarbeit werden die literarischen Strategien an einer Ganzschrift geübt. Die verwendeten Ganzschriften wurden von Projektmitarbeitenden ausgewählt. So wird sichergestellt, dass die bearbeiteten Geschichten auf altersgerechte Weise die Erzählelemente der Figur, des Ereignisses und der Situation präsentieren und die Strategien von den Kindern erfolgreich angewendet werden können. Die Lehrkräfte nehmen auch im FiLBY-3-Trainings jeweils vor dem Training an einer zweitägigen Fortbildung teil und erhalten eine Einführung in das Training. Darüber hinaus stehen ihnen Lehrerhandreichungen zur Verfügung, die die didaktischen Grundlagen des Trainings nochmals zusammenfassen sowie Stundenverläufe und weiteres Unterrichtsmaterial enthalten.

4.3.4 Zusammenfassung

Ziel von Lesetrainings soll primär die Steigerung von Lesekompetenz sein. Also ist der Fortschritt in der Lesekompetenz ein gutes Zielkriterium für eine Nutzenfunktion. Da die Lesekompetenzentwicklung sequentiell erfolgt, sollten Fähigkeiten auf Teilkompetenzen z. B. Leseflüssigkeitskompetenz und Leseverstehenskompetenz gute Indikatoren für die Wirksamkeit von auf Leseflüssigkeit oder Inferenz ausgelegten Trainings sein. Es ist bekannt, dass demografische Merkmale die Kompetenzsteigerung bedingen können. Demnach sollte ihr Einfluss auf den Nutzen ebenfalls überprüft werden. Da zusätzlich sensible Phasen angenommen werden müssen, sollte die Nutzenfunktion als flexibler funktionaler Zusammenhang geschätzt werden. Es sollten also zu Modellierung Funktionen verwendet, werden, welche auch nicht lineare Zusammenhänge abbilden können. Eine Modellklasse, die dem gerecht wird, sind generalisierte additive Modelle (GAM) (Hastie & Tibshirani, 2017; Wood, 2017). Sie können sowohl lineare Effekte als auch flexible Zusammenhänge in Form von Smoothing Splines und Tensorprodukt Splines modellieren.

4.4 Methode

Im folgenden Abschnitt wird zuerst die Stichprobe beschrieben. Anschließend werden die Durchführung der Studie, die verwendeteten Trainings- und Testmaterialien und die durchgeführten Analysen vorgestellt.

4.4.1 Stichprobe

An der vorliegenden Studie nahmen 5893 Schülerinnen und Schüler teil, die während der Studienlaufzeit die zweite und dritte Klassenstufe besuchten. 5343 der Kinder absolvierten im Frühjahr 2019 das FiLBY-2-Training. 5092 der Schülerinnen und Schüler nahmen im Herbst 2019 am FiLBY-3-Training teil. Insgesamt nahmen 4542 Schülerinnen und Schüler an beiden Trainings teil. Das Geschlechterverhältnis war ausgeglichen (weiblich: 2692, männlich: 2670, ohne Angabe: 531). Ein Großteil der Kinder hatte keinen Migrationshintergrund (kein Migrationshintergrund: 4645, Migrationshintergrund: 652, ohne Angabe: 596).

4.4.2 Durchführung

Die Lehrkräfte führten jeweils vor und nach den Trainings standardisierte Prä- und Posttests mit Leseflüssigkeits- und Leseverstehenstests durch. Dazwischen trainierten sie mit dem FiLBY-2- und mit dem FiLBY-3-Training. Konkret testeten die Lehrkräfte im März 2019, führten anschließend zwischen März und Juni 2019 das FiLBY-2-Training durch und testeten wieder im Juli 2019 Leseflüssigkeit und Leseverstehen. Nach den Sommerferien im Zeitraum Oktober bis Februar trainierten die Lehrkräfte mit dem FiLBY-3-Training und testeten anschließend im März 2020 erneut Leseflüssigkeit und Leseverstehen. Dieser Studienabschnitt war Teil einer größeren Evaluationsstudie, welche insgesamt von September 2018 bis Juli 2021 stattfand. Der hier betrachtete Studienzeitraum erstreckte sich aber nur von März 2019 bis März 2020. Die untersuchten Kinder besuchten in diesem Zeitraum die zweite und dritte Klassenstufe. Alle Kinder trainierten zuerst mit den FiLBY-2 Sachtexten und anschließend mit den FiLBY-3 Lektürematerialien.

4.4.3 Material

Die Leseflüssigkeit wurde mit dem Salzburger Lese-Screening 2-9 (SLS 2-9, Wimmer & Mayringer, 2016) erfasst. Das SLS ist ein standardisiertes Verfahren, bei dem die Kinder innerhalb von drei Minuten so viele Sätze wie möglich auf semantische Korrektheit überprüfen sollen. Der Test wurde nach den Vorgaben des Testmanuals und bezogen auf Normtabellen ausgewertet, welche sich aus einer

studieneigenen testtheoretischen Modellierungen mit Raschmodellen ergaben. Um eine Vergleichbarkeit zwischen den Messzeitpunkten zu ermöglichen, wurden alle Tests an denselben Normtabellen ausgewertet. Dabei wurden für die eingesetzten Testversionen SLS-A2, SLS-B1, SLS-B2, sowie SLS-C (Eigenkonstruktion) jeweils eigene Normtabellen erstellt und so etwaige Messunterschiede der Versionen eliminiert. Das Leseverstehen wurde mit dem Bayerischen Lesetest (BYLET) erfasst. Der BYLET ist ein neu entwickelter Leseverstehenstest mit drei Parallelversionen, in welchem die Kinder 20 Multiple-Choice-Fragen zu vier Textabschnitten beantworten. In diesen Textabschnitten wird eine Fantasiegeschichte einer Weltraumcrew beschrieben, die auf einem neuen Planeten landet und dort Abenteuer erlebt. Die geschlechtsneutrale und vorwissensarme inhaltliche Gestaltung der BYLET-Texte bietet gute Voraussetzungen für eine faire Messung des Leseverstehens. Dies konnte bereits in einer Pilotstudie bestätigt werden (Kraus et al., 2021). Zur Auswertung wurden die 20 Multiple-Choice-Fragen als richtig oder falsch kodiert und wie in Studie 1 beschrieben psychometrisch modelliert. Da hier nur eine Modellierung für alle Messzeitpunkte durchgeführt wurde, mussten die Personenparameter nicht gesondert auf einer gemeinsamen Skala normiert werden. Jedoch wurde auch für den BYLET jede Testversion einzeln modelliert.

Die demografischen Daten der Schülerinnen und Schüler wurden mit einem Fragebogen erhoben, der in Anlehnung an den Hintegrundfragebogen aus IGLU (Bos et al., 2010) gestaltet war. Dort wurden neben weiteren Variablen, die hier keine Berücksichtigung erfahren, das Geschlecht und der Migrationshintergrund abgefragt. Der Migrationshintergrund wurde als Geburt des Kindes im Ausland operationalisiert.

Als Trainingsmaterial dienten die oben vorgestellten FiLBY-2 und FiLBY-3 Materialien. Diese sind unter https://www.lesen.bayern.de/filby2/ und https://www.lesen.bayern.de/filby3/ größtenteils öffentlich und kostenfrei verfügbar.

4.4.4 Analysen

Alle Analysen wurden mit der Analysesoftware R (R Core Team, 2020) durchgeführt. Die Datenaufbereitung erfolgte mit *tidyverse* (Wickham et al., 2019). Für die psychometrischen Analysen wurden die packages *mirt* (Chalmers, 2012) und *eRm* (Mair et al., 2020) verwendet. Zur Schätzung der Nutzenfunktion wurde zusätzlich das package *mgcv* (Wood & Wood, 2015) verwendet.

4.4.4.1 Psychometrische Analysen Zur Überprüfung der Messeigenschaften der eingesetzten psychometrischen Tests wurde für das SLS eine Raschmodellierung und für den BYLET eine Modellierung mittels multidimensionalem Item-Response-Modell (MIRT) durchgeführt. Da das SLS

in vier Parallelversionen eingesetzt wurde, wurde die Raschmodellierung für jede Version einzeln durchgeführt. Zur Schätzung der Modellparameter wurde die konditionale Maximum-Likelihood-Methode verwendet. Die Modellpassung wurde mittels Andersens Likelihood-Ratio-Tests (LR, Andersen, 1973) überprüft. Die MIRT-Modellierung des BYLETs wurde in Studie 1 ausführlich behandelt.

4.4.4.2 Deskriptive Analysen Zunächst wurden die Verteilungskennwerte (Mittelwert, Range, Standardabweichung) der beiden Lesekompetenzen (Leseflüssigkeit und Leseverstehen) für alle Teilstichproben ermittelt. Teilstichproben waren jeweils Messwerte vor den Trainings und nach den Trainings, wobei der Messzeitpunkt im Juli 2019 (Ende der zweiten Jahrgangsstufe) sowohl der Messzeitpunkt nach dem FiLBY-2-Training als auch der Messzeitpunkt vor dem FiLBY-3-Training war.

4.4.4.3 Nutzenfunktionsmodellierung Die Nutzenfunktion wurde mit einem generalisierten additiven Modell (GAM) geschätzt. Als Optimalitätskriterium wurde das GCV (generalisierte Kreuzvalidierungskriterium) verwendet. Die allgemeine Modellgleichung eines GAMs lautet:

$$y_i = f_1(z_{i1}) + ... + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i$$

mit:

- y_i = Wert der abhängigen Variable y der Person i
- $f_q(z_{iq})$ = Smoothing Spline des kontinuierlichen Prädiktors z_q der Person i
- β_0 = Konstante
- β_1 = linearer Effekt des Prädiktors x_1
- ϵ_i = Residuum

Alle - auch die Basisfunktionen - wurden penalisiert. Die Basisfunktionenpenalisierung diente der Variablenselektion. Durch eine Penalisierung der Basisfunktionen können auch die linearen Effekte eines Prädiktors auf 0 geschrumpft werden. Der Prädiktor bleibt so zwar im Modell, sein Einfluss wird jedoch effektiv eliminiert (Wood, 2017). Es wurden zwei Modelle geschätzt: eines für den Zuwachs an Leseflüssigkeit und eines für den Zuwachs an Leseverstehen. Der Zuwachs wurde als individuelle Differenz zwischen den Messzeitpunkten berechnet. Die Spline-Komponenten der GAMs wurden immer pro Treatment (FiLBY-2 und FiLBY-3) berechnet. Als Schätzmethode wurde die restringierte Likelihood optimiert (REML). Interaktionseffekte der demografischen Variablen und der Treatments wurden durch eigene Variablen repräsentiert, die mit der Funktion *interaction* erstellt wurden. Da ein volles Modell mit allen möglichen Teilkompetenzen, und den demografischen

Variablen Migrationshintergrund und Geschlecht sowie deren Interaktion in der Schätzung nicht konvergierte, wurde die Modellbildung wie folgt vorgenommen:

Als Basismodell wurden alle Teilkompetenzen des BYLETs und SLS als Prädiktoren aufgenommen, sowie die Interaktion aus den beiden Gesamtskalen des Leseverstehens (BYLET-V) und der Leseflüssigkeit (SLS). Anschließend wurden die Prädiktoren entfernt, deren Koeffizienten für beide Treatments auf einen Wert kleiner eins geschrumpft wurden. In der GAM-Modellierung entsprechen die Koeffizienten den effektiven Freiheitsgraden. Diese geben den polynomiellen Grad der Modellierung der verwendeten Splines an. Bei Werten kleiner eins kann also davon ausgegangen werden, dass der Einfluss der zugehörigen Variable weniger als linear, und damit vernachlässigbar, ist. Anschließend wurden die demografischen Variablen einzeln und dann in Kombination aufgenommen. Die sich so ergebenden verschiedenen Modelle wurden anhand des BIC verglichen. Das BIC wurde mit Hinblick auf ein sparsames Wunschmodell als Modellwahlkriterium ausgesucht Im letzten Schritt wurde die geschachtelte Datenstruktur durch das Hinzufügen der random Effekte berücksichtigt. Abbildung 15 veranschaulicht den Modellbildungsprozess.

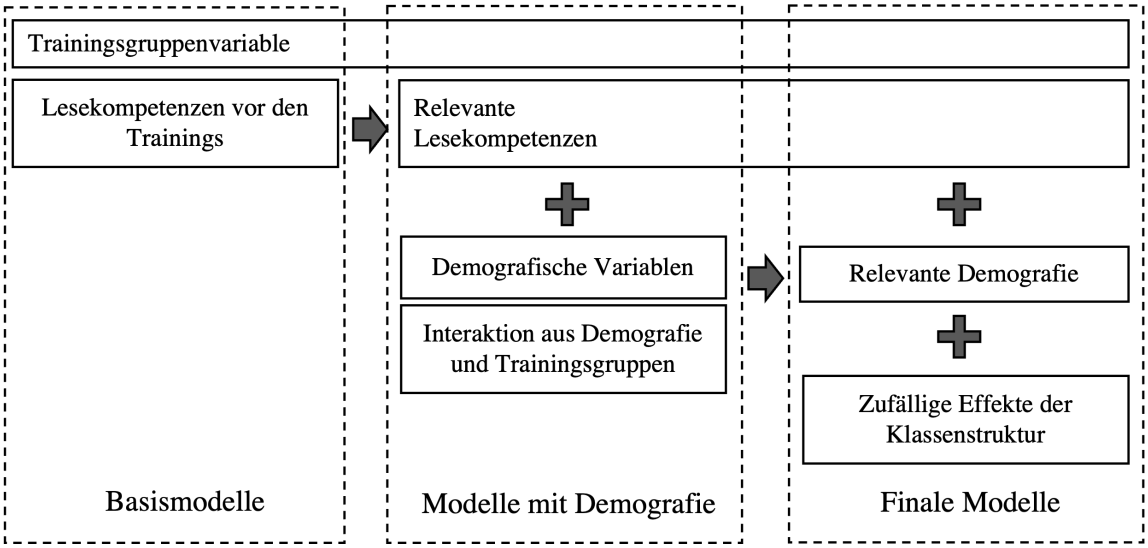


Abbildung 15: Modellbildungsprozess der Nutzenfunktionen

Die Modellwahl diente dem Ziel ein möglichst gut passendes und gleichzeitig sparsames Modell zu erhalten. Ein sparsames Modell wurde angestrebt, weil es in seiner Funktion als Nutzenfunktion und damit als Entscheidungsgrundlage für Anwender:innen praktikabel bleiben sollte. Die random Effekte wurden am Ende hinzugefügt, da sie für die zukünftige Entscheidungsfindung nicht berücksichtigbar sind. Sie benötigen Informationen über die Schulklasse. Für zukünftige Kinder, denen mithilfe des TreaDeMs ein Training zugewiesen werden soll, sind keine Informationen über ihre Klasse bekannt.

Für die grafische Darstellung und als Basis für die Entscheidungsfunktion wurde zudem pro Lesekompetenz ein Modell gebildet, das nur zwei metrische Prädiktoren enthält. Sie werden im Folgenden als "reduzierte Modelle" bezeichnet. In diesem letzten Schritt wurde erneut überprüft, ob die random Effekte tatsächlich zur Modellverbesserung beitrugen. Hierfür wurden die BICs der reduzierten Modelle mit und ohne random Effekte verglichen. Zur Modellanalyse wurden die k-Indizes berechnet und Residuenplots erstellt. k-Indizes prüfen, ob die Basisdimension für die Smoothing Splines angemessen ist (Wood, 2017). Sie basieren auf dem Verhältnis der Varianz benachbarter und zufällig gezogener Residuen. Je weiter dieser Wert unter eins liegt, desto wahrscheinlicher ist es, dass in den Residuen noch systematische Zusammenhänge mit der Kovariablen vorhanden sind. Der p-Wert wird durch eine Simulation berechnet, bei der die Nullhypothesenverteilung durch das Ziehens aus zufällig gemischten Residuen ermittelt wird.

4.4.4.4 Bestimmung des differential Payoffs Der differential Payoff beschreibt den Unterschied zwischen den beiden Lesetrainings bezüglich des erwarteten Trainingserfolgs bei gleichen Lernvoraussetzungen. Er wird hier als punktweise Differenz der vorhergesagten Nutzenwerte verstanden, der entsteht, wenn im Sinne eines kontrafaktischen Vergleichs die Variable der Trainingsgruppe verändert wird. Daher wird der differential Payoff als Grafik dargestellt. In dieser befinden sich die Leistungsvoraussetzungen in Leseflüssigkeit und Leseverstehen auf den beiden Achsen, die Farbe gibt den zugehörigen differential Payoff an. In einer weiteren Grafik wird der differential Payoff den punktweisen Standardfehlern der Nutzenfunktionsschätzungen gegenübergestellt, um so ein deskriptives Maß für die entscheidungsweisende Kraft der geschätzten Nutzenfunktion zu erlangen. Diese Grafik ist also gewissermaßen eine Veranschaulichung der Effektstärke, denn sie zeigt die punktweise Erwartungswertdifferenz in Bezug zur Varianz der Erwartungswertschätzungen. Abschließend wird der differential Payoff anhand seines Vorzeichens dichotomisiert. So entstehen Entscheidungsflächen, deren Farbe das Training mit der höheren geschätzten Wirksamkeit kennzeichnet. Die Grenzen der Entscheidungsflächen zeigen den Wechsel in der geschätzten höheren Wirksamkeit zwischen dem FiLBY-2- und dem FiLBY-3-Training an.

4.5 Ergebnisse

Im folgenden Abschnitt werden zuerst die Ergebnisse der psychometrischen Analysen präsentiert. Anschließend werden die deskriptiven Auswertungen berichtet und die Ergebnisse der Nutzenfunktionsmodellierung mit den Zielkriterien des Leseflüssigkeits- und des Leseverstehenszuwachses dargestellt. Der Abschnitt endet mit einer Betrachtung des differential Payoffs und vorläufigen Entscheidungskarten, welche gegeben eines Sets an Kovariablen die erwarteten Trainingserfolge nach dem FiLBY-2- und dem FiLBY-3-Training gegenüberstellen.

4.5.1 Psychometrische Analysen

Zur Gewinnung der Messwerte der Leseflüssigkeit und des Leseverstehens wurden psychometrische Analysen durchgeführt und die Testverfahren zur besseren Interpretierbarkeit normiert.

4.5.1.1 Raschmodellierung des SLS Das SLS (Wimmer & Mayringer, 2016) ist ein Speed-Test, so dass hier eine versionsinterne Raschmodellierung mit der konditionalen Maximum-Likelihood-Methode durchgeführt wurde. Alle vier eingesetzten Versionen waren raschskalierbar und zeigten bei einer zufälligen Teilung der Stichprobe nicht-signifikante Andersen LR-Tests (vgl. Tabelle 18).

Tabelle 18: Andersen LR-Testergebnisse für alle vier SLS-Versionen

	SLS-A	SLS-B1	SLS-B2	SLS-C
Andersen LR-Wert	45.62	100.11	91.66	83.72
Freiheitsgrade	84	99	99	94
p	1	0.450	0.687	0.767

Anmerkung. LR = Likelihood Ratio; $p = p$-Wert. Unterschiedliche Freiheitsgrade resultieren aus unterschiedlicher Anzahl maximal gelöster Items.

Die Personenhomogenität konnte auch in den grafischen Modelltests bestätigt werden. Diese sind in Abbildung 16 dargestellt. Die Itemschwierigkeiten stiegen fast kontinuierlich mit der Reihenfolge der Testitems im Test an. Eine Übersicht über die exakten Itemschwierigkeiten gibt Anhang B. Anschließend wurden die Personenparameter mit der Maximum Likelihood Methode geschätzt und z-standardisiert.

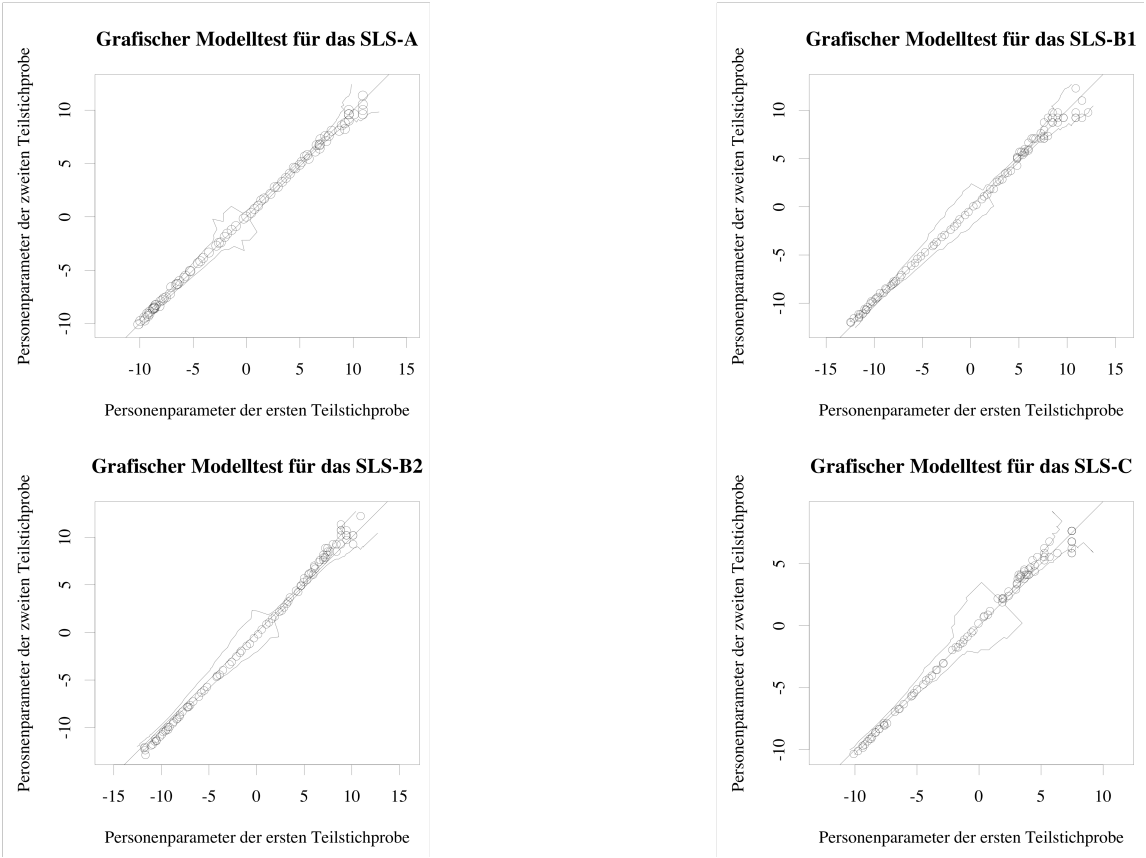


Abbildung 16: Grafische Modelltests der Raschmodellierung des SLS

4.5.1.2 MIRT-Modellierung des BYLETs Diese Analysen wurden ausführlich in Studie 1 dargestellt. Für diese Studie wurden die aus den MIRT-Modellen per MAP-Methode (Baker & Kim, 2004; Mislevy, 1986) geschätzten Personenparameter verwendet.

4.5.2 Deskriptive Analysen

Die deskriptive Auswertung in Tabelle 19 zeigt, dass die Kinder in allen Teilkompetenzen über die Zeit im Durchschnitt Fortschritte erzielten, während die Standardabweichungen fast konstant blieben. Die Werte sind in den z-standardisierten Skalenwerten der psychometrischen Modellierungen angegeben, in welche alle drei Messzeitpunkte eingingen. Die Ergebnisse bedeuten also, dass die Leistung im Querschnitt an einzelnen Messzeitpunkten nicht weniger stark streute, als die Leistung der gesamten Stichprobe über alle drei Messzeitpunkte hinweg. Ebenso zeigte sich, dass zu allen Messzeitpunkten die Leistungsskalen fast gleichermaßen abgedeckt wurden und sich die Minimal- und Maximalwerte kaum veränderten.

Tabelle 19: Kennwerte der Lesekompetenzwerteverteilungen der Teilstichproben

Messzeitpunkt	Minimum	Mittelwert	Maximum	Standardabweichung
BYLET II 2. Kl. Mitte	-3.57	-0.13	2.21	1.00
BYLET II 2. Kl. Ende	-3.22	-0.05	2.44	1.00
BYLET II 3. Kl. Mitte	-3.87	0.10	2.67	0.97
BYLET III 2. Kl. Mitte	-3.11	-0.12	2.41	0.99
BYLET III 2. Kl. Ende	-3.20	-0.03	2.90	1.00
BYLET III 3. Kl. Mitte	-3.42	0.09	3.04	0.98
BYLET IV 2. Kl. Mitte	-2.93	-0.10	4.94	0.90
BYLET IV 2. Kl. Ende	-3.05	-0.06	4.43	0.96
BYLET IV 3. Kl. Mitte	-2.64	0.17	4.36	1.15
BYLET V 2. Kl. Mitte	-2.10	-0.31	2.84	0.92
BYLET V 2. Kl. Ende	-2.20	-0.10	3.04	0.99
BYLET V 3. Kl. Mitte	-2.07	0.34	3.41	0.98
SLS 2. Kl. Mitte	-3.01	-0.34	3.34	0.83
SLS 2. Kl. Ende	-3.20	0.07	5.44	1.05
SLS 3. Kl. Mitte	-2.37	0.56	3.73	1.01

Anmerkung. Kl. = Klasse; SLS = Salzburger Lesecreening; BYLET = Bayerischer Lesetest; II - V = Kompetenzstufen; Werte sind an Gesamtstichprobe z-standardisiert.

4.5.3 Nutzenfunktionsmodellierung

Zur Nutzenfunktionsmodellierung wurden zwei Nutzenfunktionen geschätzt: eine für den Fortschritt in der Leseflüssigkeit, gemessen mit dem SLS und eine für den Fortschritt im Leseverstehen, gemessen mit dem BYLET. Beide Modellbildungen und Modellanalysen werden getrennt dargestellt.

4.5.3.1 Fortschritt der Leseflüssigkeit Zur Modellierung des Nutzens im Bezug auf die Leseflüssigkeit wurde zunächst die Modellselektion durchgeführt und anschließend die Parameter der finalen Modelle geschätzt. Abschließend wurde eine Modellanalyse mit Betrachtung der Passungskriterien und der Residuenplots durchgeführt.

4.5.3.1.1 Modellselektion Getestet wurden die folgende Modelle mit den in den Tabellen 20 bis 22 dargestellten Effekten.

Tabelle 20: Basismodell und reduziertes Modell

Unabhängige Variablen/Name	Basismodell	Reduziert
Abkürzung	*mod*0	*mod*1
Lineare Effekte		
Faktorvariable	Treatment	Treatment
Nonlineare Effekte		
Gruppierungsvariable	Treatment	Treatment
Smoothing Spline	SLS	SLS
Smoothing Spline	BYLET Stufe II	BYLET Stufe II
Smoothing Spline	BYLET Stufe III	-
Smoothing Spline	BYLET Stufe IV	BYLET Stufe IV
Smoothing Spline	BYLET Stufe V	-
Tensorprodukt Spline	SLS x BYLET Stufe V	SLS x BYLET Stufe V

Anmerkung. Getestete Effekte des Basismodells und des reduzierten Modells für den Leseflüssigkeitszuwachs.

Tabelle 21: Modell nur mit Demografie und Modell mit Geschlecht

Unabhängige Variablen/Name	Nur Demografie	Geschlecht
Abkürzung	*mod*2	*mod*3
Lineare Effekte		
Faktorvariable	Treatment x Migrationshintergrund x Geschlecht	Treatment
Nonlineare Effekte		
Gruppierungsvariable	-	Treatment x Geschlecht
Smoothing Spline	-	SLS
Smoothing Spline	-	BYLET Stufe II
Smoothing Spline	-	BYLET Stufe IV
Tensorprodukt Spline	-	SLS x BYLET Stufe V

Anmerkung. Getestete Effekte des Modells nur mit Demografie und des Modells mit Interaktion aus Geschlecht und Treatment für den Leseflüssigkeitszuwachs.

Tabelle 22: Modelle mit Geschlecht und/oder Migrationshintergrund

Unabhängige Variablen/Name	Migrationshintergrund	Geschlecht und Migrationshintergrund
Abkürzung	*mod*4	*mod*5
Lineare Effekte		
Faktorvariable	Treatment	Treatment
Nonlineare Effekte		
Gruppierungsvariable	Treatment x Migrationshintergrund	Treatment x Migrationshintergrund x Geschlecht
Smoothing Spline	SLS	SLS
Smoothing Spline	BYLET Stufe II	BYLET Stufe II
Smoothing Spline	BYLET Stufe IV	BYLET Stufe IV
Tensorprodukt Spline	SLS x BYLET Stufe V	SLS x BYLET Stufe V

Anmerkung. Getestete Effekte der Modelle mit Geschlecht und/oder Migrationshintergrund für den Leseflüssigkeitszuwachs.

Es wurden also zuerst alle Leistungsvariablen in das Modell aufgenommen und anschließend diejenigen im Modell behalten, deren Einfluss nicht auf 0 regularisiert wurde. Zu diesem reduzierten Modell (*mod1*) wurden anschließend verschiedene demografische Variablen (Geschlecht und Migrationshintergrund) hinzugefügt. So entstanden sechs verschiedene Modelle, aus denen dasjenige mit dem kleinsten BIC als finales Modell ausgesucht wurde. Dabei ergab sich die in Tabelle 23 gezeigte Modellwahl.

Es passte also das Basismodell (*mod0*) am besten. Für *mod0* betrug das REML-Kriterium einen Wert von 5331.74, der Skalierungsfaktor 0.38 und es gingen 5641 Beobachtungen in die Analyse ein. Es konnte kein bedeutsamer Einfluss der demografischen Variablen auf den Fortschritt in der Leseflüssigkeit gefunden werden. Es zeigte sich in den parametrischen Effekten, dass sich das FiLBY-2-Training insgesamt günstiger auf den Fortschritt in der Leseflüssigkeit auswirkte als das FiLBY-3-Training.

Tabelle 23: Modellwahl zum Leseflüssigkeitsfortschritt ohne random Effekte

Modellbezeichnung	*df*	BIC
mod0	33.22	10847.28
mod1	54.41	11046.09
mod2	9.00	10996.60
mod3	79.88	11211.58
mod4	45.65	10982.92
mod5	60.94	11086.84

Anmerkung. df = Freiheitsgrade; mod = Modell.

Die Effekte der Smoothing Splines zeigten sich meistens nur für das FiLBY-2-Training, die Variablen BYLET-III und BYLET-V wurden sogar in beiden Gruppen auf einen Wert kleiner eins regularisiert, so dass ihr Einfluss als vernachlässigbar anzusehen ist. Die Varianzaufklärung war mit 0.07 allerdings ziemlich gering.

4.5.3.1.2 Finale Modelle mit random Effekten

Tabelle 24: Finales Modell mit random Effekten

Unabhängige Variablen	Finales Modell
Lineare Effekte	
Faktorvariable	Treatment
Nonlineare Effekte	
Gruppierungsvariable	Treatment
Smoothing Spline	SLS
Smoothing Spline	BYLET Stufe II
Smoothing Spline	BYLET Stufe III
Smoothing Spline	BYLET Stufe IV
Smoothing Spline	BYLET Stufe V
Tensorprodukt Spline	SLS x BYLET Stufe V
Random Effekte	Klasse

Anmerkung. Getestete Effekte des finalen Modells mit random Effekten für den Leseflüssigkeitszuwachs.

Im letzten Schritt der Modellierung wurden dem Basismodell die random Effekte hinzugefügt, wie in Tabelle 24 dargestellt. Es ergab sich ein REML-Kriterium einen Wert von 4391.63, der Skalierungsfaktor betrug 0.24 und es gingen 5641 Beobachtungen in die Analyse ein.

Tabelle 25: Parametrische Effekte der finalen Modelle des Leseflüssigkeitsfortschritts

	Ohne random Effekte			Mit random Effekten		
Name	Effekt	Standardfehler	p -Wert	Effekt	Standardfehler	p -Wert
Konstante	0.42	0.01	< 0.001	0.42	0.03	< 0.001
FiLBY-3	-0.14	0.02	< 0.001	-0.09	0.02	< 0.001

Tabelle 26: Nonparametrische Effekte der finalen Modelle des Leseflüssigkeitsfortschritts

	Ohne random Effekte				Mit random Effekten			
Name	edf	ref	Testwert	WKT	edf	ref	Testwert	WKT
s(SLS):FiLBY-2	6.09	48	2.92	< 0.001	6.72	19	27.52	< 0.001
s(SLS):FiLBY-3	4.88	49	2.77	< 0.001	9.07	19	181.71	< 0.001
s(BYLET V):FiLBY-2	0.98	16	4.08	< 0.001	0.99	12	14.97	< 0.001
s(BYLET V):FiLBY-3	0.95	14	1.46	< 0.001	0.98	8	11.01	< 0.001
s(BYLET II):FiLBY-2	3.84	49	0.26	0.003	0.57	11	0.18	0.089
s(BYLET II):FiLBY-3	0	49	0	0.353	0.00	19	0.00	0.477
s(BYLET III):FiLBY-2	0.4	49	0.01	0.204	0.94	19	0.16	0.05
s(BYLET III):FiLBY-3	0.26	14	0.02	0.245	0.81	10	0.72	0.022
s(BYLET IV):FiLBY-2	1.37	49	0.1	0.024	0.00	19	0.00	0.49
s(BYLET IV):FiLBY-3	0	49	0	0.403	0.00	19	0.00	0.741
ti(SLS,BYLET V):FiLBY-2	4.7	16	1.2	< 0.001	10.72	304	0.15	< 0.001
ti(SLS,BYLET V):FiLBY-3	0	16	0	0.377	0.94	78	0.04	0.423
random Effekte					229.22	248	13.09	< 0.001

Anmerkung. edf = effektive Freiheitsgrade; ref = maximale Freiheitsgrade; WKT = p-Wert; SLS = Salzburger Lese-Screening, BYLET = Bayerischer Lesetest; römische Ziffern bezeichnen Kompetenzstufen (II - V); s() = Smoothing Spline; ti() = Tensorprodukt Spline.

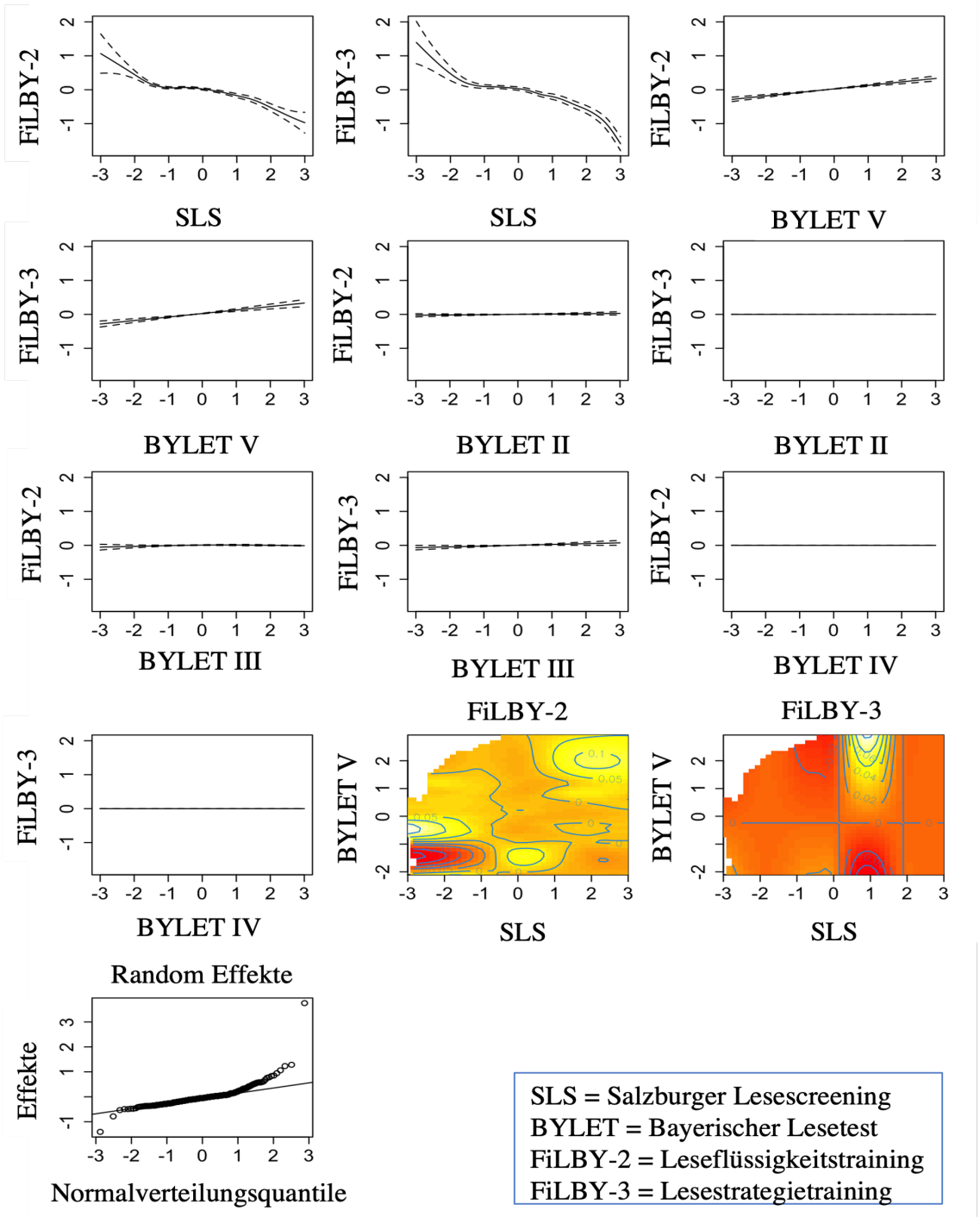


Abbildung 17: Splines des finalen Modells der Leseflüssigkeit mit random Effekten

Es zeigte sich, wie in Tabelle 25 dargestellt, in den parametrischen Effekten, dass sich das FiLBY-2-Training auch unter Berücksichtigung der Klassenstruktur insgesamt günstiger auf den Fortschritt in der Leseflüssigkeit auswirkte.

Die Effekte der Smoothing Splines zeigten sich meistens nur für das FiLBY-2-Training. Die Variablen BYLET-IV und BYLET-V wurden wieder in beiden Gruppen auf einen Wert kleiner eins regularisiert, so dass ihr Einfluss als vernachlässigbar anzusehen ist (vgl. Tabelle 26). Die Varianzaufklärung war mit 0.41 deutlich höher als im Modell ohne random Effekte. Das deutet auf einen deutlichen Einfluss der Klassenstruktur auf den Fortschritt in der Leseflüssigkeit hin. Abbildung 17 zeigt die Spline-Komponenten des finalen Modells inklusive random Effekten.

4.5.3.1.3 Reduziertes finales Modell Da eine möglichst sparsame Modellierung ein Ziel der Nutzenfunktionsmodellierung war, wurde abschließend das finale Modell um alle Terme bereinigt, deren Smoothing Splines unter einen Wert von eins regularisiert worden waren (vgl. Tabelle 27). Dieses reduzierte finale Modell verfügte über eine sehr ähnliche Passung, wie das vollständige finale Modell.

Tabelle 27: Reduziertes finales Modell mit random Effekten

Unabhängige Variablen	Reduziertes finales Modell
Lineare Effekte	
Faktorvariable	Treatment
Nonlineare Effekte	
Gruppierungsvariable	Treatment
Smoothing Spline	SLS
Tensorprodukt Spline	SLS x BYLET Stufe V
Random Effekte	Klasse

Anmerkung. Getestete Effekte des reduzierten finalen Modells mit random Effekten für den Leseflüssigkeitszuwachs.

Dies zeigte sich auch im Vergleich der BICs. Es musste nur eine geringfügig schlechtere Modellpassung für eine Reduktion der Modellvariablen in Kauf genommen werden (vgl. Tabelle 28).

Tabelle 28: Modellpassungsvergleich reduziertes mit nicht reduziertem finalen Modell mit random Effekten

Modellbezeichnung	*df*	BIC
Finales Modell	33.22	10847.28
Finales Modell mit random Effekten	270.13	10083.31
Reduziertes, finales Modell mit random Effekten	292.58	10358.01

Anmerkung. df = Freiheitsgrade.

Es zeigte sich im reduzierten finalen Modell erneut in den parametrischen Effekten, dass sich das FiLBY-2-Training insgesamt günstiger auf den Fortschritt in der Leseflüssigkeit auswirkte (vgl. Tabelle 29).

Tabelle 29: Parametrische Effekte des reduzierten finalen Modells des Leseflüssigkeitsfortschritts

Name	Effekt	Standardfehler	p -Wert
Konstante	0.41	0.03	< 0.001
FiLBY-3	-0.10	0.02	< 0.001

Die Effekte der Smoothing Splines waren nun für beide Treatments gleichermaßen ausgeprägt. Zudem wurden alle Effekte auch signifikant (vgl. Tabelle 30). Die Varianzaufklärung war mit 0.40 als vergleichbar zum vollständigen finalen Modell. Das Kriterium nahm dabei einen Wert von 4443.85, bei einem Skalierungsfaktor von 0.25 bei 5641 Beobachtungen an.

Tabelle 30: Nonparametrische Effekte des reduzierten finalen Modells des Leseflüssigkeitsfortschritts

Name	edf	ref	Testwert	WKT
s(SLS):FiLBY-2	7.19	19	17.55	< 0.001
s(SLS):FiLBY-3	8.76	19	87.04	< 0.001
ti(SLS,BYLET V):FiLBY-2	17.40	323	0.42	< 0.001
ti(SLS,BYLET V):FiLBY-3	10.56	307	0.14	< 0.001
random Effekte	229.06	248	12.92	< 0.001

Anmerkung. edf = effektive Freiheitsgrade; ref = maximale Freiheitsgrade; WKT = p-Wert; SLS = Salzburger Lese-Screening, BYLET = Bayerischer Lesetest; römische Ziffern bezeichnen Kompetenzstufen (II - V); s() = Smoothing Spline; ti() = Tensorprodukt Spline.

In der einzelnen Betrachtung von Leseflüssigkeitsvoraussetzung und zeigte sich, dass bei beiden Treatments Kinder besonders profitierten, die über eine schlechte Leseflüssigkeit verfügten. Diese Effekte wurden durch den gemeinsamen Effekt des Tensorprodukt Splines teilweise ausgeglichen. Darüber hinaus ergaben sich im FiLBY-2-Training jedoch Vorteile für solche Kinder, die in einer der beiden Lesekompetenzen bereits vor dem Training eine mittlere Leistung erbrachten, in der jeweils anderen Lesekompetenz jedoch unterdurchschnittlich abschnitten.

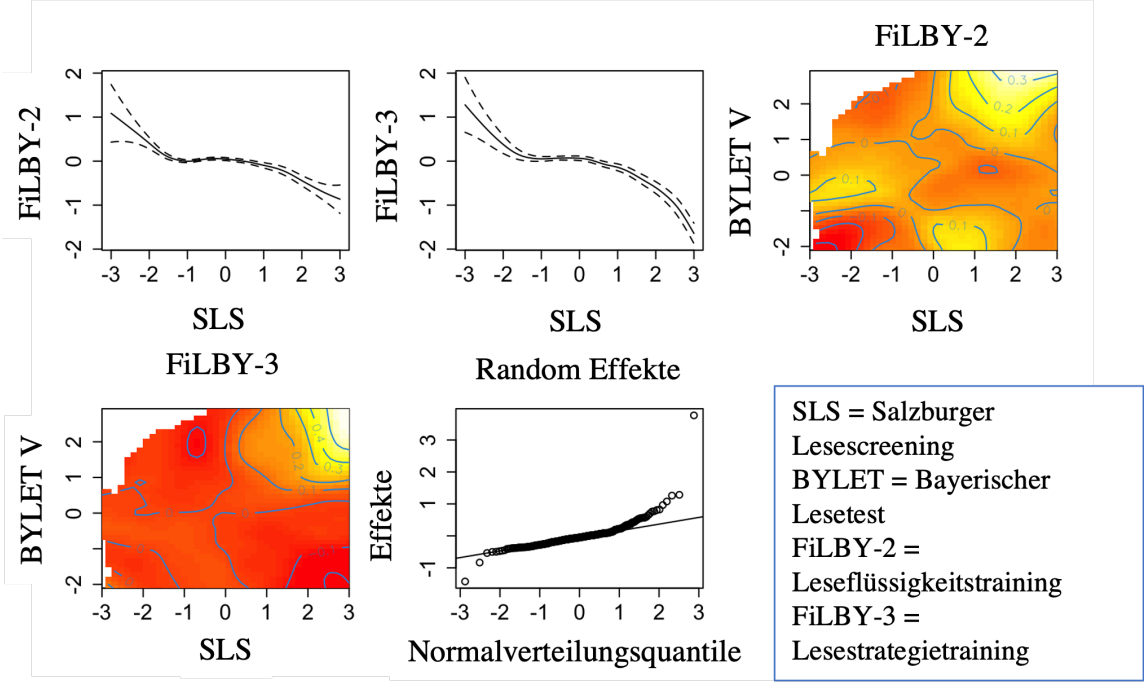


Abbildung 18: Splines des reduzierten finalen Modells der Leseflüssigkeit

Dies war beim Strategietraining nicht der Fall. Hier schien es sich in leichter Tendenz eher positiv auszuwirken, wenn Leseverstehen und Leseflüssigkeit sich auf einem ähnlichen, möglichst hohen Niveau befanden. Bei den parametrischen Effekten zeigte sich, wie erwartet, dass das Leseflüssigkeitstraining dem Lesestrategietraining in Bezug zur Leseflüssigkeitssteigerung insgesamt überlegen war (vgl. Abbildung 18).

Insgesamt scheint es möglich, den Fortschritt in der Leseflüssigkeit auch mit weniger Variablen, dafür aber komplexeren Zusammenhängen bei vergleichbarer Modellpassung zu erklären. Die komplexeren Zusammenhänge zeigen sich in den höheren effektiven Freiheitsgraden.

4.5.3.1.4 Modellanalyse Die Modellanalyse wurde für das finale Modell mit random Effekten, sowie für dessen reduzierte Variante vorgenommen. Für das finale Modell mit random Effekten ergaben sich die in Tabelle 31 dargestellten Kennwerte. Man sieht am k-Index, dass die Splines eine gute Passung hatten.

Tabelle 31: Modellcheck des finalen Modells des Leseflüssigkeitfortschritts

Effekt	k	edf	k-Index	WKT
s(SLS):FiLBY-2	19	6.71	0.98	n.s.
s(SLS):FiLBY-3	19	8.95	0.98	s.
s(BYLET V):FiLBY-2	19	0.99	1.00	n.s.
s(BYLET V):FiLBY-3	19	0.98	1.00	n.s.
s(BYLET III):FiLBY-2	19	1.31	1.02	n.s.
s(BYLET III):FiLBY-3	19	0.81	1.02	n.s.
s(BYLET IV):FiLBY-2	19	0.00	0.98	n.s.
s(BYLET IV):FiLBY-3	19	0.00	0.98	n.s.
ti(SLS,BYLET V):FiLBY-2	304	10.40	1.01	n.s.
ti(SLS,BYLET V):FiLBY-3	287	3.41	1.01	n.s.
random Effekte	249	229.00	NA	NA

Anmerkung. edf = Effektive Freiheitsgrade; WKT = p -Wert; s() = Smoothing Spline; ti() = Tensorprodukt Spline; s. = signifikant; n.s. = nicht signifikant.

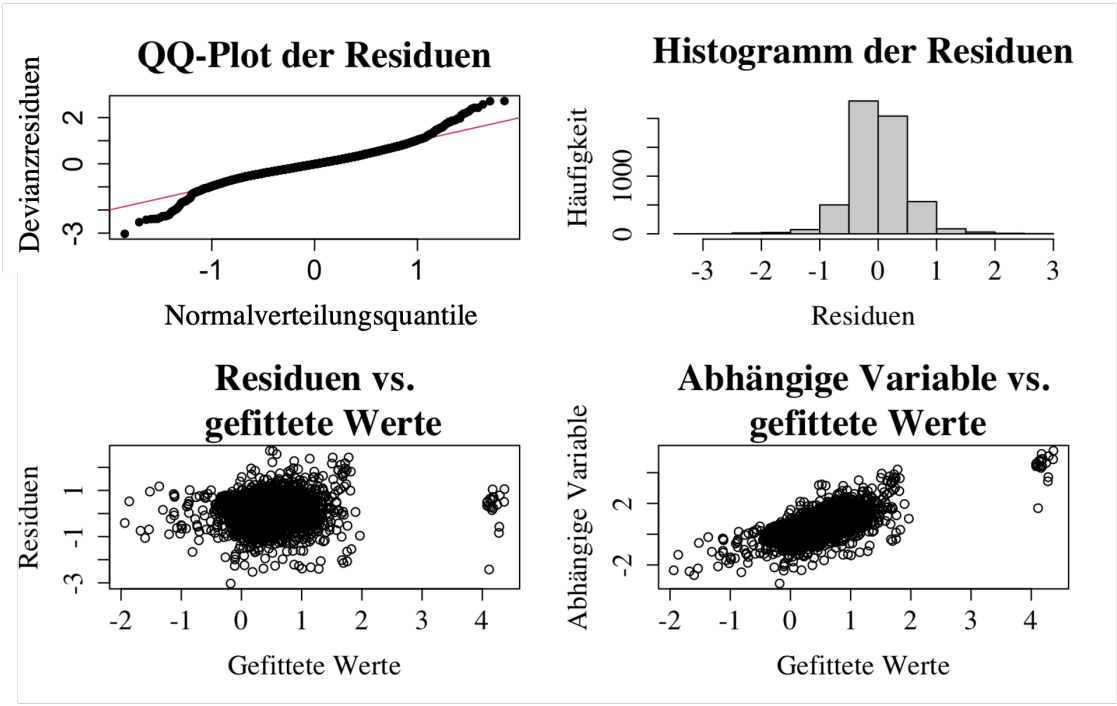


Abbildung 19: Residualanalyse des finalen Modells mit random Effekten der Leseflüssigkeit

Die Residuenanalyse bestätigte, dass die Residuen von den gefitteten Werten unabhängig waren

und annährend normalverteilt. Als auffällig erwies sich eine Gruppe von Beobachtungen, die einen extrem hohen Fortschritt von über vier Standardabweichungen zeigte (vgl. Abbildung 19).

Für das reduzierte finale Modell mit random Effekten zeigten sich die in Tabelle 32 präsentierten Kennwerte.

Tabelle 32: Modellcheck des reduzierten finalen Modells des Leseflüssigkeitfortschritts

Effekt	k	edf	k-Index	WKT
s(SLS):FiLBY-2	19	7.19	1.01	n.s.
s(SLS):FiLBY-3	19	8.76	1.01	n.s.
ti(SLS,BYLET V):FiLBY-2	323	17.40	0.98	n.s.
ti(SLS,BYLET V):FiLBY-3	307	10.56	0.99	n.s.
random Effekte	249	229.06	NA	NA

Anmerkung. edf = Effektive Freiheitsgrade; WKT = p -Wert; s() = Smoothing Spline; ti() = Tensorprodukt Spline; s. = signifikant; n.s. = nicht signifikant.

Auch dieses Modell verfügte über eine gute Passung. Der Einfluss der im Modell verbleibenden Variablen veränderte sich kaum im Vergleich zum vollständigen, nicht-reduzierten Modell.

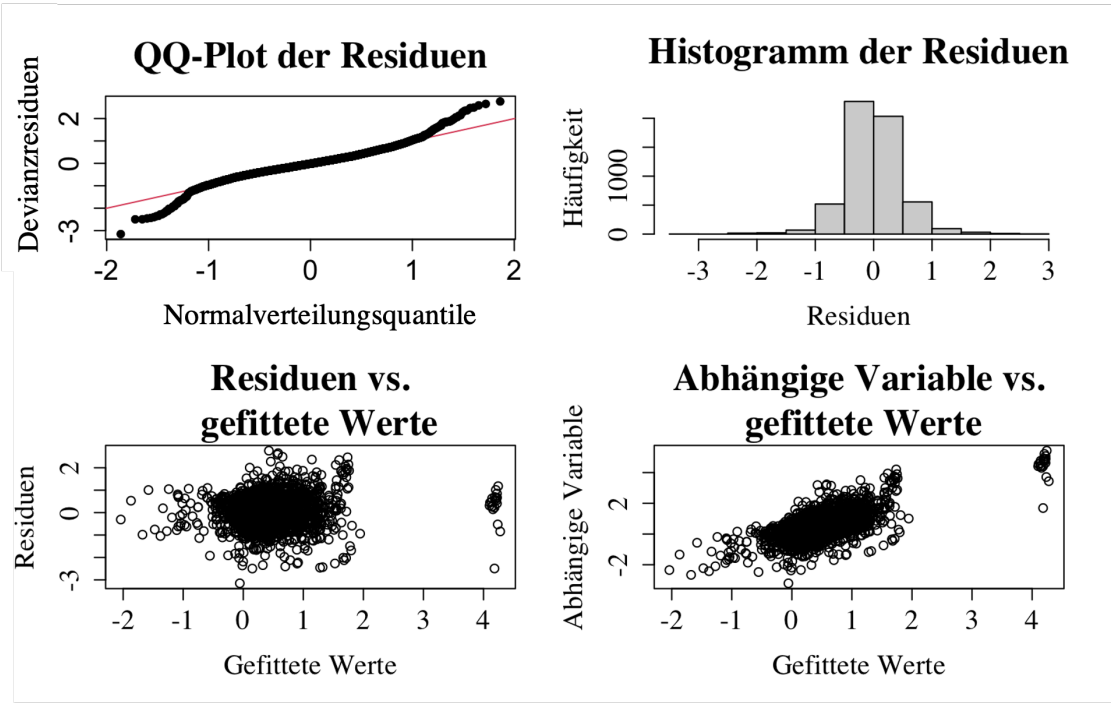


Abbildung 20: Residualanalyse des reduzierten finalen Modells der Leseflüssigkeit

Man erkennt in Abbildung 20, dass die Residuen vor allem in den Extrembereichen stark von

der Normalverteilung abwichen. Auch gilt, dass die Schätzungen außerhalb des Bereichs von -2.5 bis +2.5 relativ ungenau waren. Daher sollte die Funktion außerhalb dieses Bereichs mit Vorsicht interpretiert werden.

Die Residualplots zeigen eine Unabhängigkeit der Residuen von den gefitteten Werten, sowie einen linearen Zusammenhang zwischen den gefitteten Werten und den tatsächlichen Werten der abhängigen Variable. Dies ist ein Indikator für eine unverzerrte Schätzung. Aber es ergab sich auch im reduzierten Modell eine kleine Gruppe von Ausreißern, deren abhängige Variable ungewöhnlich hohe Werte aufwies.

4.5.3.1.5 Finales Modell ohne Ausreißer Eine Anschlussanalyse bestätigte, dass die Ausreißerbeobachtungen alle aus nur zwei verschiedenen Schulklassen stammten. Abbildung 21 verdeutlicht dies und zeigt, dass für diese beiden Ausreißerklassen die Residuen zudem sehr schief verteilt waren, da fast alle Punkte überhalb der in schwarz eingezeichneten Winkelhalbierenden lagen.

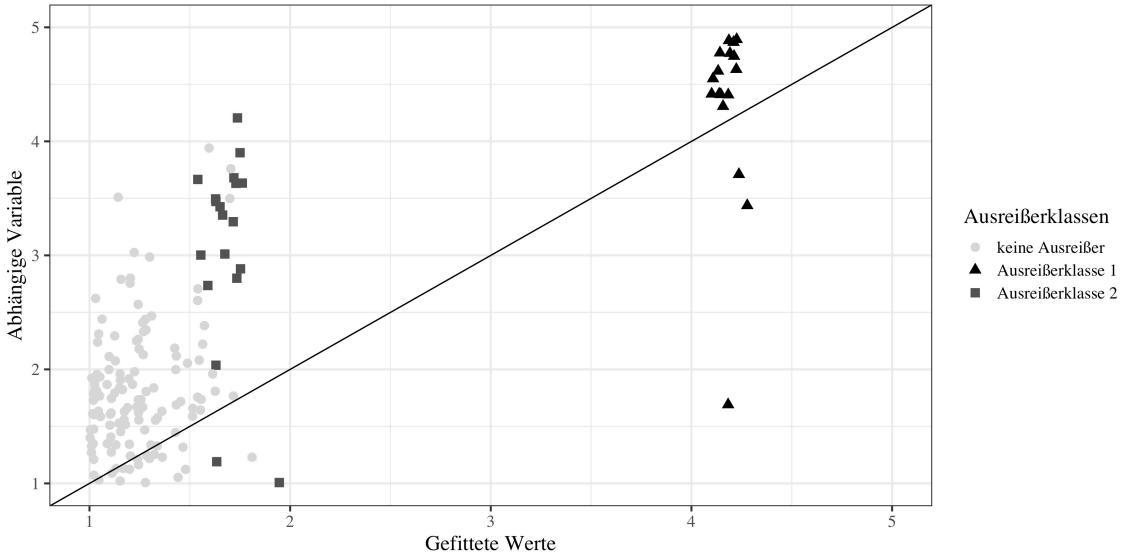


Abbildung 21: Ausreißeranalyse der Modellierung des Leseflüssigkeitsfortschritts

Die beiden Lehrkräfte der Ausreißerklassen wurden daraufhin kontaktiert, um auszuschließen, dass es sich bei ihren Werten um einen Fehler bei der Datenerhebung oder -eingabe gehandelt hatte. Eine Lehrkraft meldete sich nicht zurück. Die zweite Lehrkraft meldete zurück, dass es sich nicht um ein Versehen gehandelt hatte, sondern um eine außergewöhnlich talentierte Schulklasse. Da jedoch diese Klassen ausschließlich in der Leseflüssigkeit, nicht aber im Leseverstehen extrem überdurchschnittliche Leistungen aufwiesen, wurden als ein letzter Modellbildungsschritt die beiden Ausreißerklassen aus der Analyse entfernt.

Durch das Entfernen der Ausreißerbeobachtungen aus der Analyse veränderten sich die Modell-

parameter geringfügig, wie Tabelle 33 und 34 zu entnehmen. Es wurde wieder ein kleiner aber signifikanter negativer Effekt für das FiLBY-3-Training geschätzt. Durch das Entfernen der Ausreißerklassen wurde zudem der Tensorprodukt Spline für das FiLBY-2-Training mit 23.31 effektiven Freiheitsgraden deutlich komplexer geschätzt als im Modell mit Ausreißern. Hier betrugen die effektiven Freiheitsgrade 10.56.

Tabelle 33: Parametrische Effekte des reduzierten finalen Modells des Leseflüssigkeitsfortschritts ohne Ausreißer

Name	Effekt	Standardfehler	p -Wert
Konstante	0.39	0.02	< 0.001
FiLBY-3	-0.06	0.02	0.002

Tabelle 34: Nonparametrische Effekte des reduzierten finalen Modells ohne Ausreißer

Name	edf	ref	Testwert	WKT
s(SLS):FiLBY-2	7.95	9.75	15.28	< 0.001
s(SLS):FiLBY-3	6.28	7.90	18.64	< 0.001
ti(SLS,BYLET V):FiLBY-2	23.31	340.00	0.42	< 0.001
ti(SLS,BYLET V):FiLBY-3	8.40	316.00	0.07	0.001
random Effekte	226.11	257.00	7.24	< 0.001

Anmerkung. edf = effektive Freiheitsgrade; ref = maximale Freiheitsgrade; WKT = p-Wert; SLS = Salzburger Lese-Screening, BYLET = Bayerischer Lesetest; römische Ziffern bezeichnen Kompetenzstufen (II - V); s() = Smoothing Spline; ti() = Tensorprodukt Spline.

Auch der Modellfit fiel ohne Ausreißer etwas geringer aus. Ohne Ausreißer konnten nur noch 28% der Varianz erklärt werden. Wie in Abbildung 22 zu sehen, veränderten sich die Splineschätzungen durch das Entfernen der Ausreißerklassen nur geringfügig. Nach wie vor profitierten vor allem langsam Lesende, sowie sehr schnell und sehr gut verstehend Lesende von den beiden Trainings.

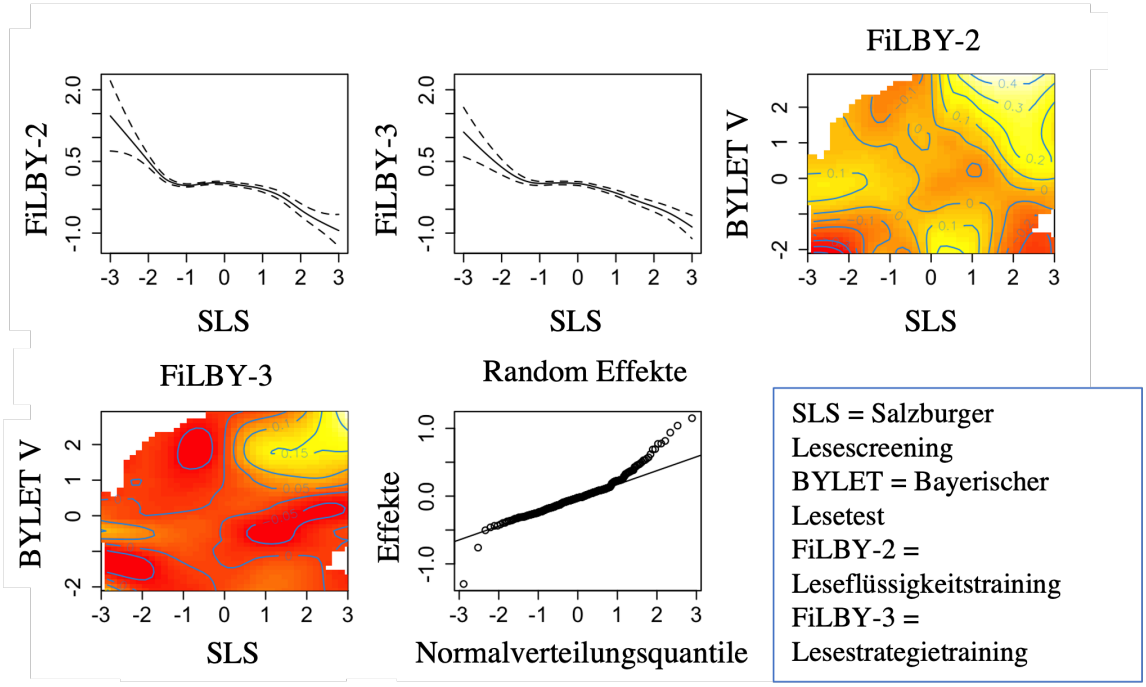


Abbildung 22: Splines des reduzierten finalen Modells ohne Ausreißer der Leseflüssigkeit

4.5.3.1.6 Modellanalyse ohne Ausreißer Nun zeigte sich in der Residuenanalyse in Abbildung 23, dass im Modell ohne Ausreißer ein linearer Zusammenhang zwischen gefitteten Werten und der abhängigen Variable vorlag und die Residuen normalverteilt waren. Die Modellanalyse mittels k-Indizes fiel ebenfalls durchweg positiv aus. Dies zeigt Tabelle 35.

Tabelle 35: Modellcheck des finalen Modells des Leseflüssigkeitfortschritts ohne Ausreißer

Effekt	k	edf	k-Index	WKT
s(SLS):FiLBY-2	19	7.95	1.02	n.s.
s(SLS):FiLBY-3	19	6.28	1.02	n.s.
ti(SLS,BYLET V):FiLBY-2	340	23.31	0.99	n.s.
ti(SLS,BYLET V):FiLBY-3	316	8.40	0.98	n.s.
random Effekte	258	226.11	NA	NA

Anmerkung. edf = Effektive Freiheitsgrade; WKT = p -Wert; s() = Smoothing Spline; ti() = Tensorprodukt Spline; s. = signifikant; n.s. = nicht signifikant.

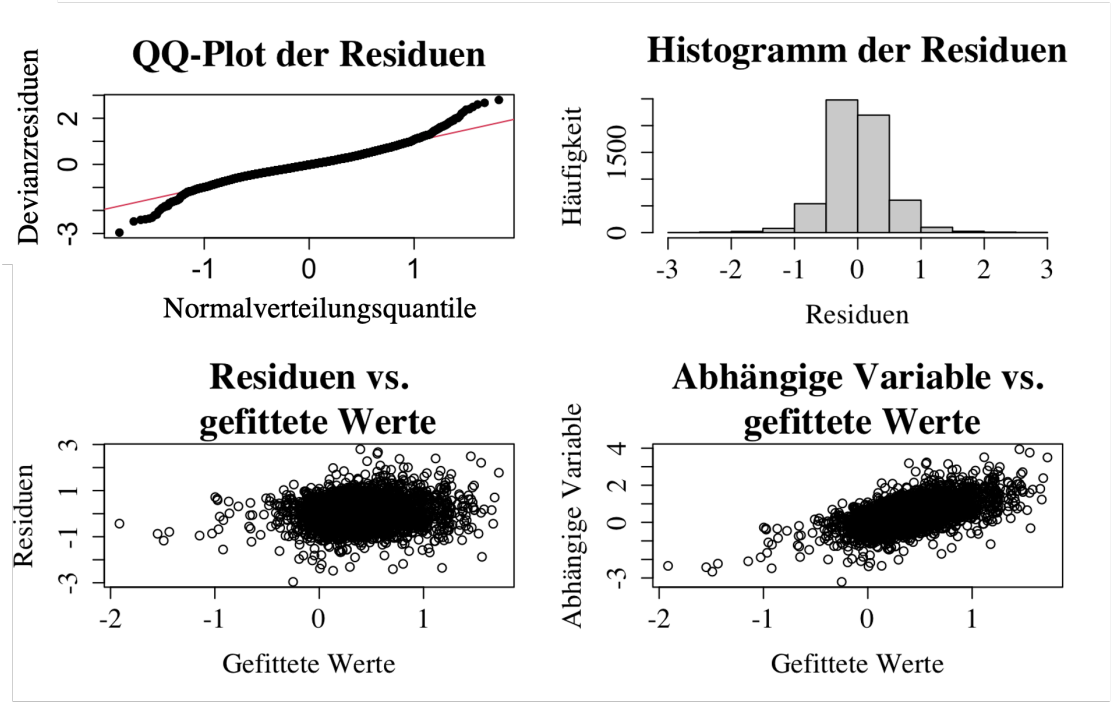


Abbildung 23: Residualanalyse des reduzierten finalen Modells ohne Ausreißer der Leseflüssigkeit

4.5.3.2 Fortschritt des Leseverstehens Zur Modellierung des Nutzens im Bezug auf das Leseverstehen wurde zunächst die Modellselektion durchgeführt. Hierfür wurden sechs verschiedene Modelle geschätzt und ihre Passung anhand ihrer BICs verglichen. Anschließend wurden die Parameter der finalen Modelle geschätzt. Abschließend wurde eine Modellanalyse mit Betrachtung der Passungskriterien und der Residuenplots durchgeführt.

4.5.3.2.1 Modellselektion Auch für die Fortschrittsmodellierung des Leseverstehens wurden verschiedene Modelle getestet. Diese sind in den Tabellen 36 bis 38 dargestellt. Es wurden also analog zur Leseflüssigkeitsmodellierung zuerst alle Leistungsvariablen in das Modell aufgenommen und anschließend diejenigen im Modell behalten, deren Einfluss nicht auf 0 regularisiert wurde. So reduzierten sich die relevanten Leistungsvariablen um die Interaktion aus BYLET V und SLS, sowie um die Stufen II und III des BYLETs. Zu diesem reduzierten Modell (*mod1*) wurden anschließend die demografische Variablen (Geschlecht und Migrationshintergrund) hinzugefügt.

Tabelle 36: Basismodell und reduziertes Modell

Unabhängige Variablen/Name	Basismodell	Reduziert
Abkürzung	*mod*0	*mod*1
Lineare Effekte		
Faktorvariable	Treatment	Treatment
Nonlineare Effekte		
Gruppierungsvariable	Treatment	Treatment
Smoothing Spline	SLS	SLS
Smoothing Spline	BYLET Stufe II	-
Smoothing Spline	BYLET Stufe III	-
Smoothing Spline	BYLET Stufe IV	BYLET Stufe IV
Smoothing Spline	BYLET Stufe V	BYLET Stufe V
Tensorprodukt Spline	SLS x BYLET Stufe V	-

Anmerkung. Getestete Effekte des Basismodells und des reduzierten Modells für den Leseverstehensszuwachs.

Tabelle 37: Modell nur mit Demografie und Modell mit Geschlecht

Unabhängige Variablen/Name	Nur Demografie	Geschlecht
Abkürzung	*mod*2	*mod*3
Lineare Effekte		
Faktorvariable	Treatment x Migrationshintergrund x Geschlecht	Treatment
Nonlineare Effekte		
Gruppierungsvariable	-	Treatment x Geschlecht
Smoothing Spline	-	SLS
Smoothing Spline	-	-
Smoothing Spline	-	BYLET Stufe IV
Tensorprodukt Spline	-	-

Anmerkung. Getestete Effekte des Modells nur mit Demografie und des Modells mit Interaktion aus Geschlecht und Treatment für den Leseverstehenszuwachs.

Tabelle 38: Modelle mit Geschlecht und/oder Migrationshintergrund

Unabhängige Variablen/Name	Migrationshintergrund	Geschlecht und Migrationshintergrund
Abkürzung	*mod*4	*mod*5
Lineare Effekte		
Faktorvariable	Treatment	Treatment
Nonlineare Effekte		
Gruppierungsvariable	Treatment x Migrationshintergrund	Treatment x Migrationshintergrund x Geschlecht
Smoothing Spline	SLS	SLS
Smoothing Spline	-	-
Smoothing Spline	BYLET Stufe IV	BYLET Stufe IV
Tensorprodukt Spline	-	-

Anmerkung. Getestete Effekte der Modelle mit Geschlecht und/oder Migrationshintergrund für den Leseverstehenszuwachs.

So entstanden sechs verschiedene Modelle, aus denen dasjenige mit dem kleinsten BIC als finales Modell ausgesucht wurde. Es ergaben sich dabei die folgenden Modellpassungen (vgl. Tabelle 39).

Tabelle 39: Modellwahl zum Leseverstehensfortschritt ohne random Effekte

Modellbezeichnung	*df*	BIC
mod0	25.14	12306.51
mod1	22.73	12321.88
mod2	9.00	14139.14
mod3	33.40	12406.93
mod4	28.11	12360.41
mod5	45.86	12490.69

Anmerkung. df = Freiheitsgrade; *mod* = Modell.

Auch hier erhielt das Basismodell den kleinsten BIC. Das Basismodell (*mod0*) passte also am besten. Es verfügte über einen REML-Wert von 6086.83. Der Skalierungsfaktor wurde auf 0.52 geschätzt

und in der Analyse wurden 5561 Kinder berücksichtigt.

Tabelle 40: Parametrische Effekte der finalen Modelle des Leseverstehensfortschritts

	Ohne random Effekte			Mit random Effekten		
Name	Effekt	Standardfehler	p -Wert	Effekt	Standardfehler	p -Wert
Konstante	0.18	0.01	< 0.001	0.18	0.02	< 0.001
FiLBY-3	0.05	0.02	0.031	0.02	0.03	0.405

Tabelle 41: Nonparametrische Effekte der finalen Modelle des Leseverstehensfortschritts

	Ohne random Effekte				Mit random Effekten			
Name	edf	ref	Testwert	WKT	edf	ref	Testwert	WKT
s(BYLET V):FiLBY-2	6.00	49	40.54	< 0.001	5.14	9	293.79	< 0.001
s(BYLET V):FiLBY-3	1.00	15	16.96	< 0.001	1.00	9	46.27	< 0.001
s(SLS):FiLBY-2	4.15	48	9.83	< 0.001	3.77	9	73.84	< 0.001
s(SLS):FiLBY-3	1.75	49	0.77	< 0.001	2.00	9	8.98	< 0.001
s(BYLET II):FiLBY-2	0.57	15	0.09	0.116	0.26	8	0.05	0.233
s(BYLET II):FiLBY-3	0.00	49	0.00	0.326	0.00	9	0.00	0.466
s(BYLET III):FiLBY-2	0.94	15	0.98	< 0.001	0.95	9	2.58	< 0.001
s(BYLET III):FiLBY-3	0.83	15	0.33	0.015	0.83	8	0.68	0.015
s(BYLET IV):FiLBY-2	1.63	49	0.06	0.163	1.78	9	0.48	0.095
s(BYLET IV):FiLBY-3	0.00	49	0.00	1.000	0.00	9	0.00	0.833
ti(BYLET V,SLS):FiLBY-2	0.75	16	0.19	0.041	0.51	16	0.07	0.148
ti(BYLET V,SLS):FiLBY-3	0.00	16	0.00	1.000	0.00	16	0.00	1.000
random Effekte					148.72	250	1.51	< 0.001

Anmerkung. edf = effektive Freiheitsgrade; ref = maximale Freiheitsgrade; WKT = p-Wert; SLS = Salzburger Lese-Screening, BYLET = Bayerischer Lesetest; römische Ziffern bezeichnen Kompetenzstufen (II - V); s() = Smoothing Spline; ti() = Tensorprodukt Spline.

Bereits die Leistungsvariablen erklärten einen guten Anteil der Varianz, mit einem R^2 von 0.30. Das FiLBY-3-Training bewirkte insgesamt mehr, als das FiLBY-2-Training. Dies kann am festen Effekt des FiLBY-3-Trainings abgelesen werden (vgl. Tabelle 40). Es zeigte sich - ähnlich wie bei der Leseflüssigkeit, dass die Funktion für das FiLBY-2-Training komplexer geschätzt wurde (vgl.

Tabelle 41). Die Splines für das FilBY-2-Training verfügten über eine höhere Anzahl an effektiven Freiheitsgraden. Die Interaktion aus SLS und BYLET-V, sowie BYLET-II und BYLET-III hingegen wurden aus dem Modell regularisiert. Sie erhielten eine effektive Anzahl an Freiheitsgraden kleiner eins.

4.5.3.2.2 Finale Modelle mit random Effekten

Tabelle 42: Finales Modell mit random Effekten

Unabhängige Variablen	Finales Modell
Lineare Effekte	
Faktorvariable	Treatment
Nonlineare Effekte	
Gruppierungsvariable	Treatment
Smoothing Spline	SLS
Smoothing Spline	BYLET Stufe II
Smoothing Spline	BYLET Stufe III
Smoothing Spline	BYLET Stufe IV
Smoothing Spline	BYLET Stufe V
Tensorprodukt Spline	SLS x BYLET Stufe V
Random Effekte	Klasse

Anmerkung. Getestete Effekte des finalen Modells mit random Effekten für den Leseverstehenszuwachs.

Im abschließenden Schritt wurden dem Basismodell die random Effekte hinzugefügt (s. Tabelle 42). Daraus ergaben sich als Modellkennwerte, ein REML-Kriterium von 6022.26 und ein Skalierungsfaktor von 0.48. Wie sich zeigte, konnte die Varianzaufklärung durch die zusätzlichen random Effekte nur geringfügig gesteigert werden. Sie betrug $R^2 = 0.34$. Das FiLBY-3-Training bewirkte insgesamt nur wenig mehr, als das FiLBY-2-Training. Dieser parametrische Effekt wurde zudem nicht signifikant. Es zeigte sich, ähnlich wie bei der Leseflüssigkeit, dass die Funktion für das FiLBY-2-Training komplexer geschätzt wurde.

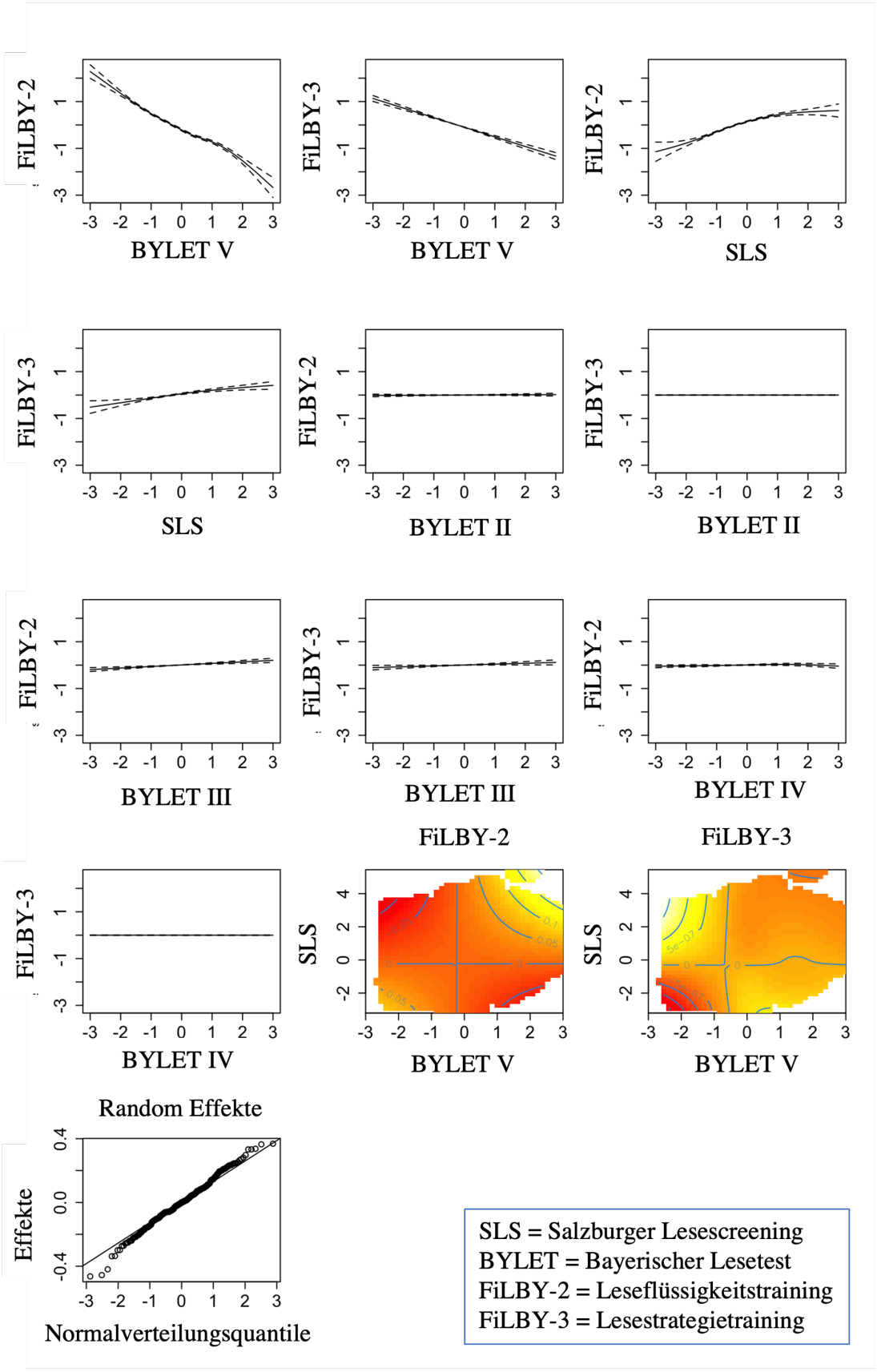


Abbildung 24: Splines des finalen Modells des Leseverstehens

Die Effekte der Interaktion aus SLS und BYLET-V, sowie die Effekte für den BYLET-II und BYLET-III wurden aus dem Modell regularisiert. Zudem fiel der zufällige Klasseneffekt deutlich geringer aus (vgl. Tabelle 41).

4.5.3.2.3 Reduzierte finale Modelle Um dem Ziel der Sparsamkeit in den Modellen gerecht zu werden, wurde ein weiteres reduziertes Modell aus dem präferierten Modell abgeleitet. Das reduzierte finale Modell verfügte über eine sehr ähnlicher Modellpassung. Es entstand durch das Entfernen derjenigen Smoothing Splines aus dem finalen Modell, welche durch die Regularisierung unter einen Wert von eins reduziert wurden und ist in Tabelle 43 dargestellt. Diese Terme wurden damit effektiv aus der Gleichung regularisiert.

Tabelle 43: Reduziertes finales Modell mit random Effekten

Unabhängige Variablen	Reduziertes finales Modell
Lineare Effekte	
Faktorvariable	Treatment
Nonlineare Effekte	
Gruppierungsvariable	Treatment
Smoothing Spline	SLS
Smoothing Spline	BYLET Stufe V
Random Effekte	Klasse

Anmerkung. Getestete Effekte des reduzierten finalen Modells mit random Effekten für den Leseverstehenszuwachs.

Der Vergleich der BICs zwischen dem reduzierten und dem finalen Modell zeigte auch hier wieder, dass nur eine geringfügig schlechtere Modellpassung für eine deutliche Reduktion der Komplexität in Kauf genommen werden musste (vgl. Tabelle 44). Es zeigte sich auch, dass das Modell ohne random Effekte einen kleineren BIC erlangte, als das reduzierte finale Modell mit random Effekten. Daher wurde für das Leseverstehen auch noch das reduzierte Modell ohne random Effekte geschätzt. Es besaß ebenfalls einen sehr kleinen BIC und wird daher ebenfalls in Tabellen 45 und 46 vorgestellt.

Tabelle 44: Modellpassungsvergleich reduziertes mit nicht reduziertem finalen Modell mit random Effekten

Modellbezeichnung	*df*	BIC
Finales Modell	25.14	12306.51
Finales Modell mit random Effekten	174.99	13083.98
Reduziertes, finales Modell mit random Effekten	169.10	13085.30
Reduziertes, finales Modell ohne random Effekte	17.33	12309.02

Anmerkung. df = Freiheitsgrade.

Dennoch veränderten sich die parametrischen und nonparametrischen Effekte im Vergleich zu den nicht reduzierten Modellen nur geringfügig (vgl. Tabellen 45 und 46).

Tabelle 45: Parametrische Effekte der reduzierten finalen Modelle des Leseverstehensfortschritts

	Ohne random Effekte			Mit random Effekten		
Name	Effekt	Standardfehler	p -Wert	Effekt	Standardfehler	p -Wert
Konstante	0.19	0.01	< 0.001	0.19	0.02	< 0.001
FiLBY-3	0.04	0.02	0.092	0.02	0.02	0.493

Tabelle 46: Nonparametrische Effekte der reduzierten finalen Modelle des Leseverstehensfortschritts

	Ohne random Effekte				Mit random Effekten			
Name	edf	ref	Testwert	WKT	edf	ref	Testwert	WKT
s(BYLET V):FiLBY-2	5.62	19	104.32	< 0.001	5.41	19	142.61	< 0.001
s(BYLET V):FiLBY-3	1.00	13	19.08	< 0.001	1.00	10	41.47	< 0.001
s(SLS):FiLBY-2	4.18	19	26.69	< 0.001	3.97	19	37.84	< 0.001
s(SLS):FiLBY-3	1.80	19	2.09	< 0.001	2.05	19	4.54	< 0.001
random Effekte	NA	-	-	-	150.95	250	1.57	< 0.001

Anmerkung. edf = effektive Freiheitsgrade; ref = maximale Freiheitsgrade; WKT = p-Wert; SLS = Salzburger Lese-Screening, BYLET = Bayerischer Lesetest; römische Ziffern bezeichnen Kompetenzstufen (II - V); s() = Smoothing Spline; ti() = Tensorprodukt Spline.

Insgesamt war weder das FiLBY-3-Training, noch das FiLBY-2-Training signifikant wirksamer. Die

Splines wurden wieder für das FiLBY-2-Training komplexer geschätzt, bildeten aber insgesamt fast lineare Zusammenhänge.

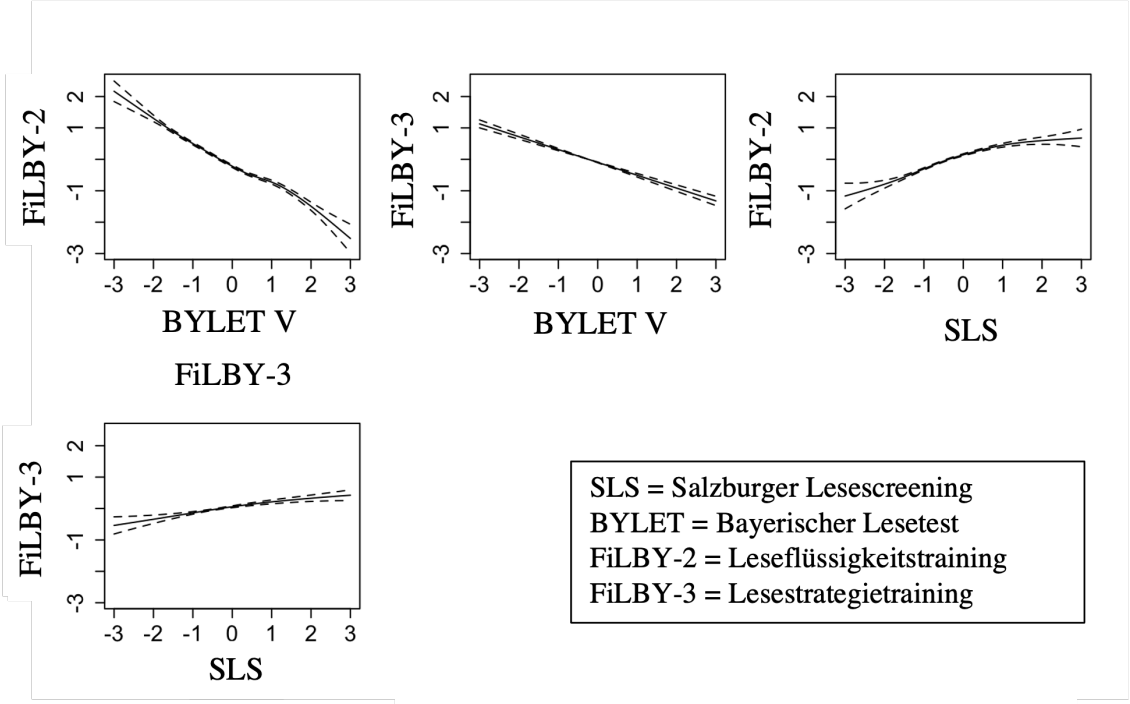


Abbildung 25: Splines des reduzierten finalen Modells des Leseverstehens

Abbildung 25 zeigt die Splines des reduzierten finalen Modells ohne random Effekte. Die Splines des reduzierten finalen Modells mit random Effekten werden nicht dargestellt, da sie sehr ähnlich waren, das Modell aber insgesamt eine schlechtere Modellpassung hatte. Man erkennt, dass Kinder umso stärker von den Trainings in beiden Lesetrainings profitierten, je flüssiger sie lasen und je schwächer ihr Leseverstehen entwickelt war. Dabei waren diese Zusammenhänge für das FiLBY-2-Training stärker ausgeprägt und weniger linear als für das FiLBY-3-Training.

4.5.3.2.4 Modellanalyse Die Modellanalyse wurde sowohl für das finale Modell mit random Effekten, sowie für dessen reduzierte Variante ohne random Effekte durchgeführt. Für das finale Modell mit random Effekten ergaben sich die in Tabelle 47 und Abbildung 26 dargestellten Kennwerte. Die Residuenplots zeigten dabei keinerlei Auffälligkeiten. So zeigten sich weder im QQ-Plot, noch im Histogramm relevante Abweichungen von der Normalverteilung. Auch konnten keine systematischen Zusammenhänge zwischen den Residuen und der unabhängigen Variable entdeckt werden. Alle k-Indizes lagen nahe an der Eins oder darüber und wurden nicht signifikant.

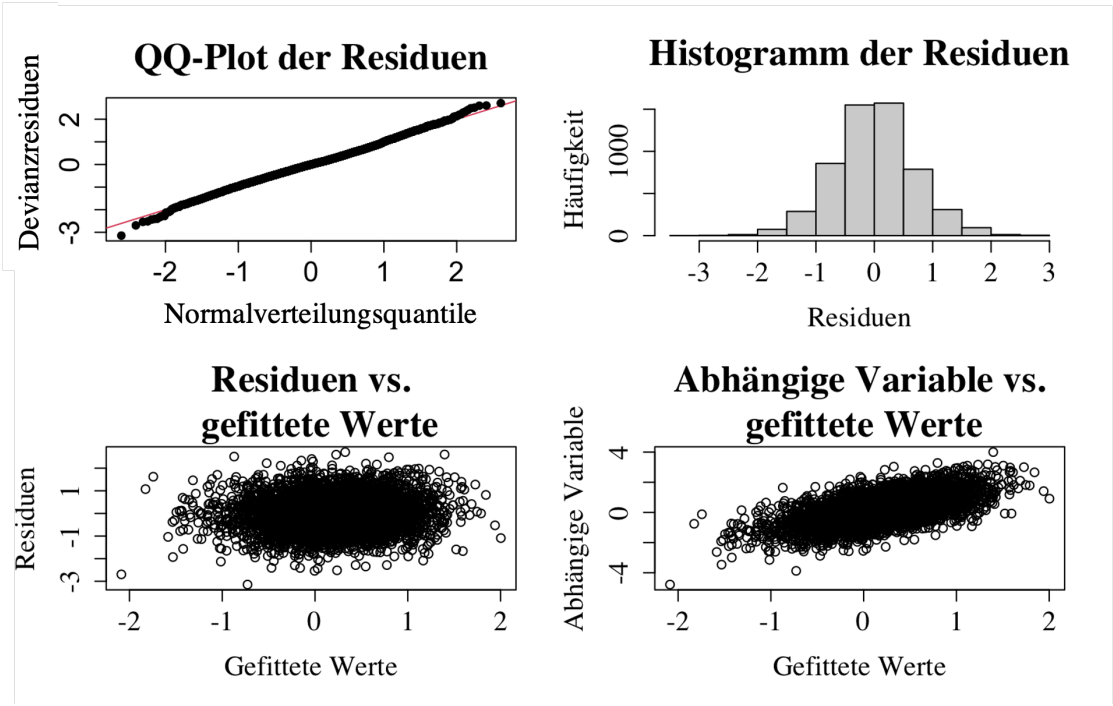


Abbildung 26: Residualanalyse des finalen Modells des Leseverstehens

Tabelle 47: Modellcheck des finalen Modells des Leseverstehensfortschritts

Effekt	k	edf	k-Index	WKT
s(BYLET V):FiLBY-2	9	5.14	1.02	n.s.
s(BYLET V):FiLBY-3	9	1.00	1.02	n.s.
s(SLS):FiLBY-2	9	3.77	0.98	n.s.
s(SLS):FiLBY-3	9	2.00	0.98	n.s.
s(BYLET II):FiLBY-2	9	0.26	1.00	n.s.
s(BYLET II):FiLBY-3	9	0.00	1.00	n.s.
s(BYLET III):FiLBY-2	9	0.95	1.00	n.s.
s(BYLET III):FiLBY-3	9	0.83	1.00	n.s.
s(BYLET IV):FiLBY-2	9	1.78	1.01	n.s.
s(BYLET IV):FiLBY-3	9	0.00	1.01	n.s.
ti(BYLET V,SLS):FiLBY-2	16	0.51	1.01	n.s.
ti(BYLET V,SLS):FiLBY-3	16	0.00	1.00	n.s.
random Effekte	251	149.00	NA	NA

Anmerkung. edf = Effektive Freiheitsgrade; WKT = p -Wert; s() = Smoothing Spline; ti() = Tensorprodukt Spline; s. = signifikant; n.s. = nicht signifikant.

Für das reduzierte finale Modell ohne random Effekte ergaben sich die in Tabelle 48 und Abbildung 27 dargestellten Modellkennwerte. Es zeigte sich, dass die Residuen fast perfekt normal verteilt waren. Im Zusammenhang von den Residuen mit den gefitteten Werten zeigten sich keine Ausreißer. Auch konnte keine Heteroskedastizität festgestellt werden. Auch die k-Indizes waren mit Werten nahe an der Eins alle unauffällig. Zudem wurden sie nicht signifikant.

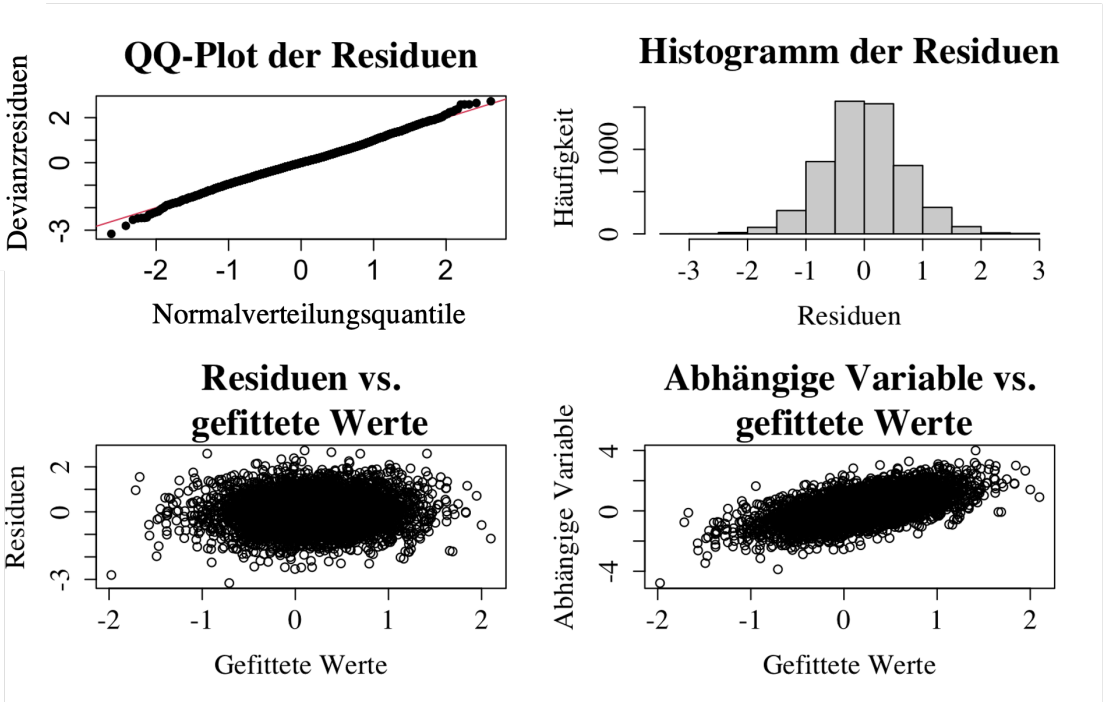


Abbildung 27: Residualanalyse des reduzierten finalen Modells des Leseverstehens

Tabelle 48: Modellcheck des reduzierten finalen Modells des Leseverstehensfortschritts

Effekt	k	edf	k-Index	WKT
s(BYLET V):FiLBY-2	19	5.62	1.00	n.s.
s(BYLET V):FiLBY-3	19	1.00	1.00	n.s.
s(SLS):FiLBY-2	19	4.18	0.98	n.s.
s(SLS):FiLBY-3	19	1.80	0.98	n.s.

Anmerkung. edf = Effektive Freiheitsgrade; WKT = p -Wert; s() = Smoothing Spline; s. = signifikant; n.s. = nicht signifikant.

4.5.3.3 Differential Payoff Zur Bewertung in diesem spezifischen entscheidungstheoretischen Kontext gilt es bei der Modellierung auch das differential Payoff zu berücksichtigen. Es geht also darum, ob die Prädiktoren für die zwei Treatments möglichst unterschiedliche Vorhersagen für

den Leistungszuwachs machen. Das Ziel ist es, damit eine möglichst eindeutige Nutzenfunktion zu generieren. Die Bestimmung des differential Payoffs wurde grafisch angegangen.

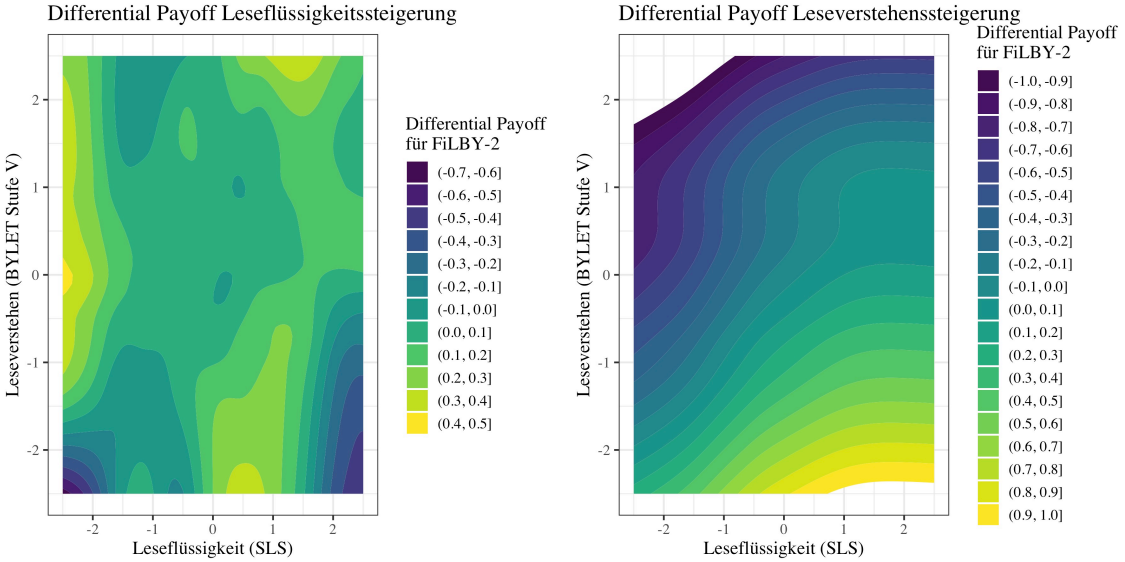


Abbildung 28: Payoffkarte für Leseflüssigkeit und Leseverstehen

Auf der Abbildung 28 ist der differential Payoff für das FiLBY-2-Training dargestellt. Je höher der Wert ist, desto stärker profitierten Kinder mit den auf der X- und Y-Achse abgebildeten Lernvoraussetzungen vom FiLBY-2-Training (im Vergleich zum FiLBY-3-Training). Es zeigte sich Folgendes. Vor allem relativ schnell, aber schlecht verstehend lesende Kinder profitierten von FiLBY-2 stärker als von FiLBY-3 für die Leseflüssigkeit. Während noch sehr langsam lesende Kinder in der dritten Klasse auch vom Strategietraining FiLBY-3 stark für die Leseflüssigkeit profitieren. Sie profitierten stärker als es vergleichbar schlecht und langsam lesende Kinder dies beim FiLBY-2-Training in der zweiten Klasse getan hatten. Die Ergebnisse sind also nicht primär kontraintuitiv, sondern als Aufholeffekt der sehr leistungsschwachen Kinder in der dritten Klasse zu verstehen.

Die rechte Grafik in Abbildung 28 zeigt, dass es für den Leseverstehenszuwachs eine Art intraindividuelle Schwelle gibt, die nötig ist, um von Lesestrategietraining stärker zu profitieren als man es von Leseflüssigkeitstraining tut. Es scheint so zu sein, dass das Strategietraining vor allem für diese Kinder effektiv ist, deren Leseverstehen mindestens so gut ist, wie ihre Leseflüssigkeit. Berücksichtigt man zusätzlich, dass die Nutzenschätzung mit Unsicherheit behaftet ist, so zeigt sich das in Abbildung 29 dargestellte Bild. Der differential Payoff in Bezug auf die Leseflüssigkeit ist also gerade in den Leistungsbereichen am größten, wo auch die Schätzunsicherheit ausgedrückt in den punktweisen Standardfehlern der Nutzenfunktionen am höchsten ist.

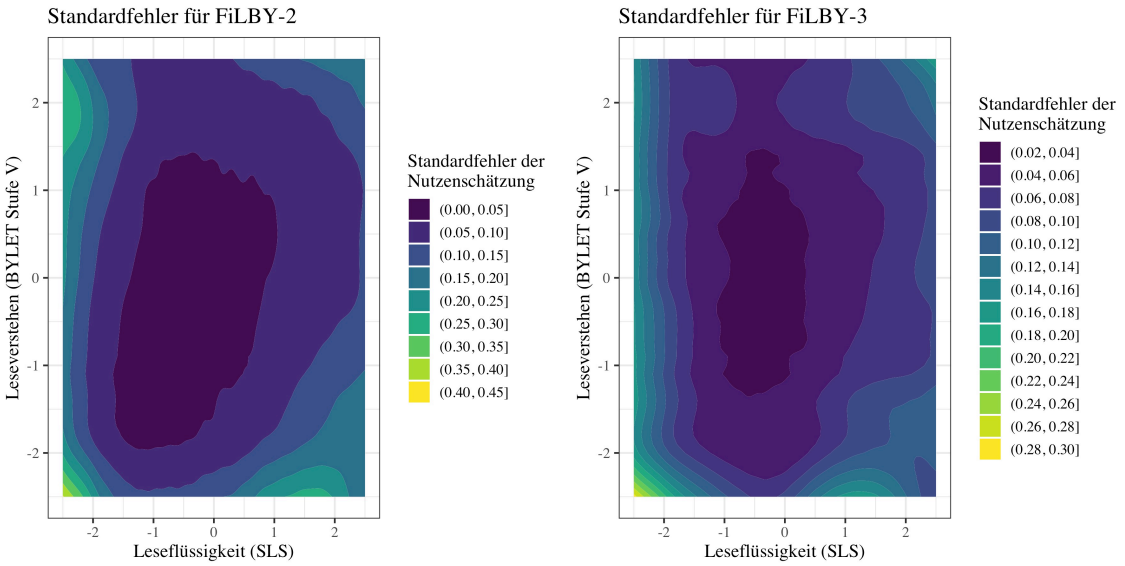


Abbildung 29: Vergleich Standardfehler

4.5.3.4 Entscheidungskarten Fasst man die Nutzenfunktion als deterministische Abbildung der Lernvoraussetzungen auf einen Nutzen unter verschiedenen Treatments auf, so lässt sich eine Entscheidungsoberfläche generieren. Sie entsteht, indem die Lernvoraussetzungen in der Leseflüssigkeit und im Leseverstehen auf den beiden Achsen abgebildet werden und der höhere erwartete Fortschritt der beiden Treatments die Farbe des Punktes bestimmt. Dichotomisiert man die Entscheidungsoberfläche, indem eine Farbe für Vorteile des FiLBY-2-Trainings und eine Farbe für die Vorteile des FiLBY-3-Trainings gewählt werden, ergibt sich eine Entscheidungskarte (vgl. Abbildung 30).

Wählt man den Fortschritt in der Leseflüssigkeit als Nutzenkriterium, so zeigt sich eine generelle Dominanz des FiLBY-2-Trainings. Lediglich für sehr schwache Leser:innen scheint das FiLBY-3-Training überlegen zu sein. Wie Abbildung 30 zeigt, wird hier das Strategietraining neben sehr schwachen Schüler:innen, auch Schüler:innen empfohlen, die über eine sehr gute Leseflüssigkeit, aber ein unterdurchschnittliches Leseverstehen verfügen. Zusätzlich sollen Schüler:innen mit dem FiLBY-2-Training trainieren, deren Leseflüssigkeit schon sehr gut ausgeprägt ist aber die im Leseverstehen noch stärkere Defizite aufweisen.

Demgegenüber zeigt Abbildung 30 für die Leseverstehenssteigerung eine fast lineare Entscheidungsgrenze. Es scheint also so zu sein, dass Kinder erst dann vom FiLBY-3-Training profitieren, wenn ihr Leseverstehen ihre Leseflüssigkeit übersteigt.

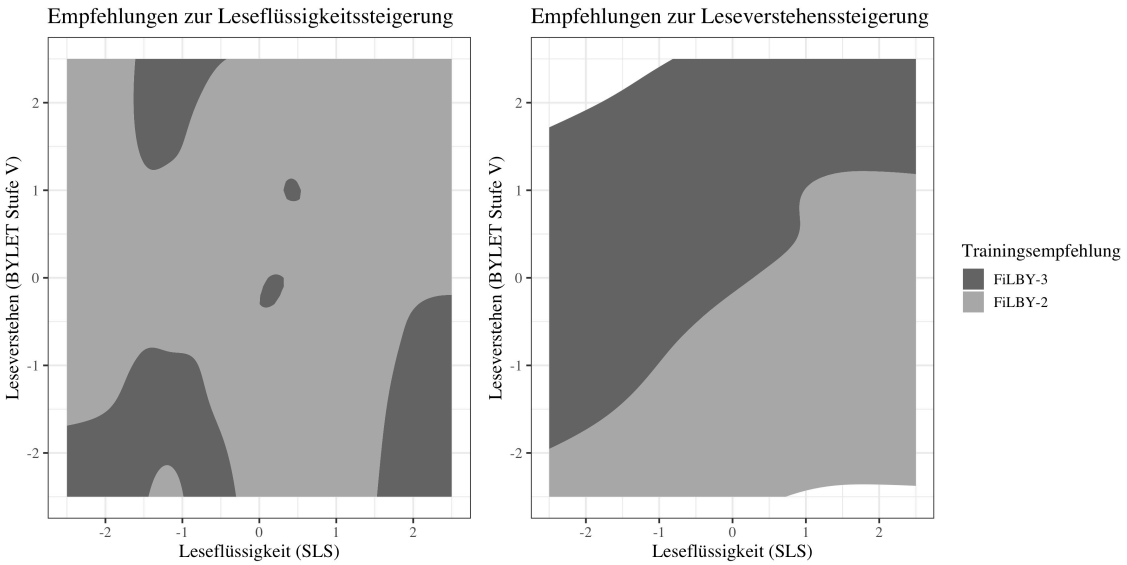


Abbildung 30: Trainingsempfehlungsvergleiche zur Leseflüssigkeits- und Leseverstehenssteigerung

4.6 Diskussion

Das Ziel dieser zweiten Studie war es, eine Nutzenfunktion zu schätzen. Hierfür wurden möglichst sparsame Modelle gesucht, welche den erwarteten Trainingserfolg in Abhängigkeit leicht zu erfassender persönlicher Merkmale abbilden. Im Folgenden sollen die Analyseergebnisse zusammengefasst und diskutiert werden.

4.6.1 Zusammenfassung und Interpretation der Ergebnisse

Der Trainingserfolg wurde über den Fortschritt zwischen den jeweiligen Prä- und Postmessungen der beiden Trainings operationalisiert. Dabei wurden die Fortschritte im Leseverstehen sowie in der Leseflüssigkeit in zwei getrennten Modellierungen betrachtet. Mittels generalisierten additiven Modellen wurde hierbei aus den Prätrainingsleistungen der Lesekompetenz und demografischen Variablen der Trainingserfolg vorhergesagt. Es zeigte sich, dass die demografische Variablen Geschlecht und Migrationshintergrund nicht zur Trainingserfolgsvorhersage beitrugen. Als prädiktiv zeigten sich dafür die Ausgangsleistungen im Leseverstehen der Kompetenzstufe V und in der Leseflüssigkeit. Die Zusammenhänge mit den Trainingserfolgen des FiLBY-2- und FiLBY-3-Trainings zeigten eine Komplexität, die zwischen sechs und 18 effektiven Freiheitsgraden lag. Es zeigte sich also insgesamt ein eher komplexer Zusammenhang zwischen den Ausgangsleistungen und den Trainingserfolgen. Betrachtet man die festen Effekte, so zeigte sich wie erwartet, dass sich zur Leseflüssigkeitssteigerung das FiLBY-2-Training insgesamt als wirksamer erwies. Demgegenüber erwies sich das FiLBY-3-Training nicht insgesamt wirksamer zur Leseverstehenssteigerung. Stellt man die

Funktionen zur Trainingserfolgsvorhersage der beiden Trainings punktweise gegenüber, so lassen sich erste Trainingsempfehlungen aus der Modellierung ableiten. Hierbei zeigte sich ein heterogenes Bild für die Leseflüssigkeitssteigerung. Es scheint so zu sein, dass hierfür immer dann das FiLBY-2-Training vorteilhaft ist, wenn beide Leistungsbereiche mindestens durchschnittlich ausgeprägt sind. Dass für Kinder, welche sehr langsam lesen und deren Leseverstehen nur sehr schwach ausgeprägt ist, das FiLBY-3-Training empfohlen wird, dürfte auf die fehlende Unabhängigkeit der Trainings zurückzuführen sein. Da die beiden Trainings sequentiell absolviert wurden, ist es aus theoretischer Überlegung heraus wahrscheinlich, dass diese Empfehlung auf einem Aufholeffekt beruht. Dieser bedeutet, unter der Annahme, dass Kinder keine Rückschritte in ihrer Entwicklung machen, dass sehr schwache Leser:innen zu Beginn der dritten Jahrgangsstufe im FiLBY-3-Training einen Leistungsschub machten, der über den durch die systematische Leseförderung im FiLBY-2-Training zum Ende der zweiten Klassenstufe hinausgeht. Dieser Leistungsschub kann ein zeitverzögerter Effekt des FiLBY-2-Trainings sein, aber auch auf externe Einflüsse oder Reifungsprozesse zurückzuführen sein. Gut interpretierbar ist hingegen die Empfehlung des Lesestrategietrainings für Kinder, deren Leseflüssigkeit schon sehr gut entwickelt ist, die aber gleichzeitig noch Defizite im Leseverstehen aufweisen. Für diese Kinder erscheint es plausibel, dass ein stärkeres Lesevestehen im Sinne einer top-down-Steuerung des Lesens auch die Leseflüssigkeit weiter erhöht, da die nächsten gelesenen Satzteile durch die innere Erwartungshaltung vorentlastet werden (Dehaene, 2014).

Auch denkbar ist, dass für sehr schwache Leser:innen beide Trainings überfordernd waren. Betrachtet man die Heatmaps (vgl. Abbildung 17) der Vorhersage der Trainingserfolge für die Leseflüssigkeit, so sieht man, dass gerade Kinder, welche sowohl ein sehr schwaches Leseverstehen als auch eine sehr schlechte Leseflüssigkeit besitzen, am wenigsten vom FiLBY-2-Training für die Leseflüssigkeit profitieren. Gleiches zeigte sich in Bezug auf das Leseverstehen für das FiLBY-3-Training. Als Fazit bleibt also festzuhalten, dass für sehr schwache Leser:innen gesonderte Fördermaßnahmen, die über FiLBY-2 und FiLBY-3 hinausgehen, eingesetzt werden sollten.

In Bezug auf die Steigerung des Leseverstehens erwies sich das FiLBY-3-Training dann als wirksamer, wenn das individuelle Leseverstehen die individuelle Leseflüssigkeit überstieg. Dieses Phänomen lässt sich gut über die im Theorieteil diskutierte “Flaschenhals”-Funktion der Leseflüssigkeit für das Leseverstehen erklären. Es hat sich bereits in Studien (Chall, 1983; Pikulski & Chard, 2005) gezeigt, dass Lesestrategietrainings erst dann ihre volle Wirksamkeit entfalten, wenn die Lernenden über eine angemessene Leseflüssigkeit verfügen. Interessant ist, dass hierbei kein kriteriumsorientierter Leseflüssigkeitswert gefunden wurde, welcher die beginnende Wirksamkeit des Strategietrainings definiert. Vielmehr scheint es so zu sein, dass Kinder auch dann im Vergleich weniger vom Strate-

gietraining profitieren, wenn zwar ihre Leseflüssigkeit schon sehr hoch ist, jedoch ihr Leseverstehen eher gering ausgeprägt ist. Ob dieses Phänomen einer durch hohe Lesegeschwindigkeit ausgelöste Illusion von Testverständnis zugrunde liegt, stellt einen Ansatz für weitere Forschungsprojekte dar. Es zeigt aber auch, dass eine reine Fokussierung auf die Steigerung der Leseflüssigkeit ohne zielgerichtete und passgenaue Leseverstehensförderung nicht zielführend ist. Offen bleibt dabei, wie das Leseverstehen gefördert werden kann, wenn es unabhängig von der Leseflüssigkeit sehr schwach ausgeprägt ist.

Weiterhin zeigte sich, dass an den Rändern der Nutzenfunktion sowohl die differential Payoffs als auch die Schätzunsicherheiten am größten waren. Letzteres war durch die geringere Datendichte an den Verteilungsrändern bedingt. Da naturgemäß nur wenige Kinder extreme Werte der Leistungsskalen erreichen, basiert hier die Nutzenfunktion nur auf wenigen Datenpunkten. Zudem werden Smoothing Splines so konstruiert, dass sie außerhalb des durch Daten bestimmten Funktionsbereichs linear extrapolieren (Fahrmeir, Kneib, Lang, & Marx, 2007). So wäre es möglich, dass bestehende differential Payoffs an der Grenze der Stichprobenverteilung durch die lineare Extrapolation überschätzt werden.

4.6.2 Limitationen

Die bedeutendste Limitation der vorgestellten Studie ist ihr Design. Fast alle Kinder haben beide diskutierten Lesetrainings sequentiell durchlaufen. Ausnahmen entstanden nur, wenn Kinder die Schule wechselten, oder eine Klasse wiederholten. Dadurch konnten die Effekte des FilBY-3-Trainings nicht wie in einem experimentellen Design von den Effekten des FiLBY-2-Trainings getrennt werden. Das bedeutet, dass in der Interpretation der Effekte des FiLBY-3-Trainings gedanklich immer “die Effekte des FiLBY-3-Trainings nach abgeschlossenem FiLBY-2-Training” hinzugefügt werden sollte. Damit reduziert sich streng genommen die entscheidungsleitende Kraft des Modells auf die Frage, ob ein Kind, das bereits das FiLBY-2-Training durchlaufen hat, bereit ist, am FiLBY-3-Training teilzunehmen oder ob es weiter mit dem FiLBY-2-Training arbeiten sollte. Das hier vorgestellte Modell vermag jedoch nicht, eine Trainingsempfehlung für das FiLBY-3-Training auszusprechen, wenn zuvor noch kein FiLBY-2-Training durchgeführt wurde. Für diese Art der Empfehlung hätte es in der Stichprobe auch Kinder geben müssen, die direkt am FiLBY-3-Training teilnehmen, ohne zuvor das FiLBY-2-Training absolviert zu haben. Denn nur durch ein experimentelles Design mit einer zufälligen Zuordnung zu den Trainingsbedingungen FiLBY-2 und FiLBY-3 ist die Gruppenvariable (Bedingungsvariable) unabhängig vom Zielkriterium und nur dann kann die Verteilung des Zielkriteriums auf die Bedingung selbst zurückgeführt werden und nicht auf eine Interaktion aus

Bedingung und dem Zuordnungsprozess zur Bedingung (Manski, 2003).

Weitere Limitationen zeigen sich im Einfluss, den Ausreißerklassen auf die Nutzenfunktion für die Leseflüssigkeit hatten. Daran anschließend bleibt zu diskutieren, inwieweit die Datenqualität für die gewünschten Trainingsempfehlungen ausreichte. Neben den Ausreißerklassen zeigte sich, dass ein Großteil der Varianzaufklärung im Zielkriterium der Leseflüssigkeitssteigerung auf die random Effekte zurückzuführen war. Die random Effekte berücksichtigen die Abhängigkeit in den Daten, die durch die Clusterung von Schulkindern in Schulklassen zustande kam. So gibt der große Einfluss der random Effekte einen Hinweis darauf, dass zusätzliche Prädiktoren auf Klassenebene das Modell verbessern könnten. Praktisch gesprochen bedeutet es, dass es einige Klassen mit großen und einige Klassen mit kleinen Fortschritten gab, welche von der Klassenstruktur und damit vermutlich auch von der Lehrkraft abhingen. Es wäre daher hilfreich in erneuten Untersuchungen, diese als Kovariablen in das Modell zu integrieren und so die Vorhersage für künftige Schulkinder zu verbessern. Auch möglich ist, dass Verzerrungen durch Fehler bei der Datenerhebung und Dateneingabe entstanden sind. Die Lehrkräfte führten die Tests selbst durch. Dass jedoch Lehrkräfte in der Regel nicht über Fachkenntnisse in der empirischen Forschung verfügen, macht eine teilweise suboptimale Durchführung von Testungen und Trainings wahrscheinlich (Stelter & Miethe, 2019). Daher kann an dieser Stelle auch empfohlen werden, Datenerhebungen möglichst von geschultem wissenschaftlichen Personal durchführen zu lassen und schwer verfälschbare standardisierte Tests einzusetzen.

Als dritte Limitation zeigte sich, dass der differential Payoff gerade an den Rändern des betrachteten Leistungsbereichs am größten war. Gerade dort war aber auch die Schätzunsicherheit am höchsten, da die Leistung normalverteilt ist und somit die Dichte der Verteilung an den Rändern geringer war, als in der Mitte. Will man also robuste Entscheidungen treffen, so darf nicht nur der Erwartungswert der prädiktiven Verteilungen der Zielkriterien gegeben einem Set an Kovariablen pro Training betrachtet werden, sondern auch die Unsicherheit in der Verteilung.

4.6.3 Diskussion unter entscheidungstheoretischer Perspektive

Unter entscheidungstheoretischer Perspektive stellt die Modellierung der Nutzenfunktion das zentrale Element des Entscheidungsmodells dar. Das Zielkriterium entscheidet maßgeblich darüber, ob ein Entscheidungsmodell vom:von der zukünftigen Anwender:in eingesetzt wird, oder nicht. Die Definition des Nutzens legt ein Wertesystem zugrunde, das von Entwickelnden und Anwender:innen geteilt werden muss (Slovic, Fischhoff, & Lichtenstein, 1977). In dieser Nutzenfunktionsmodellierung wurde der Zugewinn an Leseverstehen und Leseflüssigkeit - gemessen mit standardisierten

Leistungstests - verwendet. Das Zielkriterium wurde also sehr eng definiert. In der Unterrichtspraxis ist es wahrscheinlich, dass viele weitere Kriterien die Entscheidung für den Einsatz eines Lesetrainings beeinflussen. Hier spielen auch die Freude der Kinder am Training, der zeitliche und finanzielle Aufwand, eine Integrationsmöglichkeit in den Unterricht oder auch die persönliche Präferenz der Lehrkraft für bestimmte Unterrichtsmethoden und -formen, Erwartungshaltungen der Schülereltern und vieles mehr eine Rolle.

Die Nutzenfunktion bildet darüber hinaus zwar die Basis für die Entscheidung, sie bildet aber - anders als hier im Kapitel der Entscheidungskarten - in der Entscheidungstheorie nicht den Endpunkt des entscheidungstheoretischen Problems. Im klassischen entscheidungstheoretischen Problem wird die Nutzenfunktion nämlich als deterministisch aufgefasst (Peterson, 2017). Durch diese Definition ergibt es sich im nächsten Schritt eine - nun möglicherweise nicht mehr deterministische - Entscheidungsfunktion zu finden, welche den Aktionenraum unter Berücksichtigung der Nutzenfunktion optimal auf eine einzelne Aktion zurückführt (Peterson, 2017). In der Anwendung zur Treatmententscheidung ist jedoch bereits die Nutzenfunktion nicht deterministisch, da sie als statistisches Modell aus empirischen Daten geschätzt wird (Dehejia, 2005). Dadurch ergibt sich selbst für ein festes Set an Kovariablen eine Unsicherheit in der Nutzenvorhersage, welche durch die Standardfehler der Modellparameter quantifiziert werden kann. Ist also die Schätzung durch schlechte Modellpassung oder kleine Stichproben mit großer Unsicherheit behaftet, so schwindet die entscheidungsweisende Kraft des Nutzenmodells für die entscheidungstheoretische Verwendung. In dieser Studie wurde etwa ein großer Anteil der Varianz durch zufällige Effekte, also Effekte, die durch eine Gruppierung der Beobachtungen entstehen, erklärt. Dies ist im entscheidungstheoretischen Kontext nachteilig, da für zukünftige Beobachtungen keine Aussagen über eben diese durch die Clusterung der Daten entstandenen Effekte getroffen werden können. Einfach ausgedrückt: für zukünftige Schüler:innen ist nicht bekannt, ob sie sich in einer starken oder schwachen Schulklasse befinden. Hinzu kommt, dass auch die Messwerte, welche als Kovariablen fungieren, Messfehler behaftet sind. Zur Schätzung eines statistischen Modells auf Gruppenebene sind diese oft zu vernachlässigen, da es allgemein gilt, dass die Messfehler und Messwerte verschiedener Beobachtungen unkorreliert sind und sich so auf Gruppenebene ausmitteln (Hilbert et al., 2020). Ist das Ziel jedoch für eine einzelne Person eine optimale Entscheidung zu treffen, so gewinnen auch die einzelnen Messfehler an Bedeutung. Wie dies berücksichtigt werden kann, wird in der nächsten und letzten, dritten Studie vorgestellt.

5 Studie 3: Umsetzung und Anwendung des TreaDeMs - die Entscheidungsfunktion

Wie eingangs eingeführt, besteht das in dieser Arbeit entwickelte Entscheidungsmodell zur Treatmententscheidung aus drei Komponenten: einer Messung, einer Nutzenfunktion und einer Entscheidungsfunktion. In der ersten Studie wurde exemplarisch ein Messmodell beschrieben, mit welchem die individuellen Voraussetzungen (hier das Leseverstehen auf vier Dimensionen) erfasst werden können. Die zweite Studie befasste sich mit der Nutzenfunktion. Sie lieferte die Entscheidungsbasis eines klassischen entscheidungstheoretischen Problems, noch ohne Berücksichtigung einer Wahrscheinlichkeit für das Eintreffen der Umweltzustände. Die Nutzenfunktion bildete auf Basis von empirischen Daten einen Zusammenhang zwischen den Voraussetzungen und dem Treatmenterfolg ab. Die beiden ersten Studien dienten also dazu, eine Informationsstruktur zu schaffen, auf deren Basis nun Entscheidungen getroffen werden können. In der nun folgenden dritten Studie wird das Entscheidungsmodell komplettiert, indem zwei Entscheidungsfunktionen entwickelt werden. Entscheidungsfunktionen weisen - unter Berücksichtigung einer unsicheren Informationslage - den Voraussetzungen Aktionen oder hier eben Treatments zu. Die Entscheidungsfunktionen des TreaDeMs basieren auf in der statistischen Entscheidungstheorie etablierten Funktionen der Erwartungsnutzenmaximierung und des Maximin-Prinzips (Savage, 1951). Genau wie die Voraussetzungsmodellierung mit MIRT-Modellen und die Nutzenfunktionsmodellierung mittels GAMs sind auch die entwickelten Entscheidungsfunktionen nicht kontextspezifisch. Dennoch werden ihre Ergebnisse wieder exemplarisch auf die bereits bekannten Daten der FiLBY-Studie angewendet. Zunächst sollen beide Herangehensweisen auch inhaltlich im Sinne einer formalisierten Entscheidungstheorie eingeführt werden.

5.1 Das entscheidungstheoretische Problem

Es ist eine in der Forschung intensiv betrachtete Fragestellung wie Menschen sich in vollständig spezifizierten Entscheidungssituationen entscheiden sollten und wie sie es tatsächlich tun (z. B. Gilboa, 2009). Vollständig spezifiziert meint hierbei, dass der Aktionenraum vollständig bekannt ist, und für alle Umweltzustände Wahrscheinlichkeitsverteilungen spezifiziert sind. Zudem gilt, dass unter gewählter Aktion und eingetretenem Umweltzustand der zugeordnete Nutzen nicht stochastisch, sondern deterministisch eintritt. Dieser Forschung liegen mathematische Argumentationen zugrunde und beruht in der Regel auf Experimenten mit Lotterien oder Gedankenexperimenten. Dort ist der Aktionenraum die Auswahl an Lotterien, die ein:e Spieler:in spielen kann. Die Umweltzustände

sind die Ausgänge der assoziierten Zufallsprozesse (z. B. Münzwurf) und der Nutzen ist der Gewinn des Spiels abzüglich der Spielkosten. Der Determinismus in der Nutzenfunktion bedeutet hier, dass der Gewinn bei positivem Ausgang des Spiels immer ausgezahlt wird. Schränkt man die Entscheidungssituationen auf Treatmentwahlsituationen ein, so befassen sich nur wenige (psychologische) Diagnostiker:innen und Entscheidungstheoretiker:innen auf mathematische Art damit, wie Treatments für Personen ausgewählt werden sollen. Gegenbeispiele sind Manski, Stoye, Van der Linden und Rudner.

5.2 Die Struktur des Entscheidungsproblems

Ungeachtet dessen, dass nicht alle Autor:innen dieselbe Formulierung des hier diskutierten Entscheidungsproblems wählen - einige verzichten gar auf eine formalisierte Formulierung. Es erscheint dennoch sinnvoll, die unterschiedlichen Herangehensweisen in dieselbe Darstellungsform zu überführen. Hierfür soll als erstes festgelegt werden, dass sich eine Person vollständig über ihren Vektor an Voraussetzungsvariablen definiert. Das heißt, dass Merkmale, welche nicht beobachtet wurden, auch für die Lösung des Entscheidungsproblems nicht berücksichtigt werden sollen.

Darauf aufbauend konnten drei Forschungszweige identifiziert werden, die sich mit dem Thema der Treatmentwahl auseinandersetzen: Evaluationsforschung, diagnostische Forschung aus den Sozialwissenschaften und statistische Entscheidungstheorieforschung mit Schwerpunkt auf komplexer Unsicherheit. Alle drei Strömungen arbeiten bisher unabhängig voneinander. Es lassen sich keine Hinweise in Form von Zitationen finden, die auf die Wahrnehmung der jeweilig anderen Forschungsarbeiten hindeuten. Dennoch gibt es eine gemeinsame Struktur in der Problemstellung, nämlich die durch Stichprobendaten gestützte Treatmentwahl für Individuen. Diese lässt sich mit den folgenden sieben Punkten beschreiben:

1. Es liegen empirische Daten vor, deren Verteilung Informationen bezüglich des Erfolgs von mindestens zwei Treatments beschreiben

$$f(u|T_0) \neq f(u|T_1)$$

mit:

- u = Erfolgskriterium/Nutzenkriterium
- f = Verteilungsfunktion
- T_i = Treatment i

2. Es gibt ein Erfolgskriterium, das die Wirksamkeit der Treatments misst. Dieses Erfolgskriterium u ist eindimensional. Wenn mehrere Erfolgsdimensionen berücksichtigt werden sollen, so

sind diese bereits miteinander verrechnet.

3. Das Erfolgskriterium entspricht dem Nutzen. Wenn das in einer empirischen Untersuchung verwendete Kriterium nicht dem Nutzen entspricht, so ist dieses durch eine Transformation in einen Nutzen überführt. Denkbar wäre also, dass der Entscheidung eine Umformung aus verschiedenen Erfolgskriterien der folgenden Form vorausgegangen ist:

$$u = f(\mathbf{y})$$

mit:

- u = Erfolgskriterium/Nutzenkriterium
- f = eine Funktion, zum Beispiel eine Gewichtung
- $\mathbf{y}$ = ein Vektor messbarer Erfolgsvariablen

4. Es lässt sich für jede zukünftige Person eine prädiktive Nutzenverteilung bedingt auf die einzelnen Treatments bestimmen. Dies kann auch eine uneigentliche Verteilung sein, zum Beispiel eine Einpunktverteilung. Diese prädiktive Nutzenverteilung kann eine auf Kovariablen bedingte Verteilung sein. Die prädiktive Nutzenverteilung kann sich aus der Schätzung eines statistischen Modells ergeben.

$$f(u_j|T_0) = f(f_u(\mathbf{x_j}, T_0))$$

mit:

- $f(u_j|T_0)$ = individuelle Nutzenverteilung der Person j unter Treatment 0
- $f_u(\mathbf{x_j}, T_0)$ = Funktion, die Treatment 0 und Kovariablenvektor $\mathbf{x}$ der Person j auf u_j abbildet

5. Es gibt einen Entscheidungsmechanismus, eine Entscheidungsregel, welche die auf die Treatments bedingten prädiktiven Nutzenverteilungen in eine Treatmententscheidung überführt.

$$\{f(u_j|T_0), f(u_j|T_1), ..., f(u_j|T_n)\} \rightarrow T_{optimal}$$

mit:

- $T_{optimal}$ = optimales Treatment

6. In die Überführung von prädiktiven Nutzenverteilungen in Treatmententscheidungen fließen Optimalitätskriterien ein. Die Wahl der Optimalitätskriterien hängt von einer Reihe von Vorüberlegungen ab:

- Welche Unsicherheiten in der Entscheidungssituation sollen berücksichtigt werden?
  - Varianzen der individuellen Vorhersageverteilung
  - Schätzunsicherheiten

   - komplexe Unsicherheit (z. B. partielle Identifikation)
   - Messfehler in den Kovariablen (hier neu entwickelt)

7. Das Optimalitätskriterium wird über eine Verlustfunktion berechnet. Es muss also geklärt werden, wie der Verlust (die Loss-Funktion) spezifiziert werden soll. Dafür müssen die folgenden Fragen beantwortet werden:

- Wie soll die Unsicherheit in die Entscheidung einfließen? Zur Berücksichtigung der verschiedenen Arten der Unsicherheit im Sinne von stochastischer und komplexer Unsicherheit bieten sich bayesianische oder auch minimaxbasierte Entscheidungsregeln an.
- Handelt es sich um eine einzelne Entscheidung oder eine wiederholte Entscheidung? Eng verknüpft mit der Frage, ob es sich um einzelne oder wiederholte Entscheidungen handelt, steht das Verwenden von Entscheidungsregeln, die auf asymptotischem Verhalten von Schätzern basieren. Für einzelne Entscheidungen ist deren Verwendung aus theoretischer Sicht diskutabel.
- Soll eine Entscheidungsregel das persönliches Sicherheitsbedürfnis, die persönliche Risikofreude berücksichtigen? Hierbei gilt es eine individuelle Nutzenstruktur zu berücksichtigen. Zentral ist die Frage, ob die Loss-Funktion symmetrisch sein soll. Werden beispielsweise Hypothesentests als Entscheidungsfunktion eingesetzt, so ist dies nicht der Fall.

Diese Fragen sollen im Folgenden ausführlicher dargestellt werden. Anschließend werden bestehende Entscheidungsregeln unter der hier eingenommenen Perspektive diskutiert und ein eigener Ansatz zur Parametrisierung der Treatmententscheidung vorgestellt.

5.3 Einflussfaktoren auf Optimalität von Entscheidungsstrategien

Optimalität ist kein objektiv definierbarer Begriff. Optimalität ist immer an bestimmte Prämissen geknüpft, die nicht nur durch die Entscheidungssituation, sondern auch durch die entscheidende Person definiert werden müssen (z. B. Rosenhead, Elton, & Gupta, 1972).

5.3.1 Berücksichtigung der Arten der Unsicherheit

Die Unsicherheit im hier diskutierten entscheidungstheoretischen Problem der Treatmententscheidung besitzt drei Ursachen. Als erstes ist die Unsicherheit bei der Schätzung der Nutzenfunktion zu berücksichtigen. Sie stammt aus der Nutzenfunktion und lässt sich auf theoretischer Ebene in Bias, Varianz und den nicht irreduziblen Fehler zerlegen (Friedman, 1997). Wird die Nutzenfunktion als

ein statistisches Modell geschätzt, so wird die Modellpassung niemals absolut sein. Hier gilt es, wie bei jeder statistischen Modellierung einen Trade-Off zwischen Bias und Varianz zu finden (Friedman, 1997). Hinzu kommen unvermeidbare Fehler, die sich aus der Zufälligkeit der Stichprobe ergeben. Alle zusammen führen dazu, dass die individuelle Vorhersageverteilung für neue Beobachtungen keine Einpunktverteilung ist, sondern mit einer gewissen Unsicherheit behaftet. Diese Unsicherheit drückt sich in Abhängigkeit des Modells in der Regel durch die Varianz der individuellen Vorhersageverteilung aus. Je größer die Varianz ist, desto höher ist die Unsicherheit. Eine weitere Quelle der Unsicherheit kann sich aus dem Design der Studie ergeben, auf deren Daten eine Nutzenfunktion geschätzt wird. Erfolgt etwa die Zuordnung der Probanden zu den Treatments nicht zufällig, so führt das Problem der partiellen Identifikation der Zielvariable zu nicht identifizierbaren Einflüssen auf die Zielvariable (Manski, 2003). Nimmt ein:e Proband:in nur an einer der Treatmentoptionen teil, so ist die Treatmenterfolgsvariable stets partiell identifiziert. Wann dies zu Unsicherheit im Entscheidungsproblem führt, wird nachfolgend kurz dargestellt.

Die Idee von experimentellen Treatmentstudien ist es, eine sogenannte kontrafaktische Realität nachzuempfinden. Eine kontrafaktische Realität bezeichnet dabei eine Wirklichkeit, die nicht eingetreten ist, aber theoretisch als Folge einer anders getroffenen Entscheidung hätte eintreten können. Bezogen auf ein Experiment ist eine kontrafaktische Realität der Ausgang des Experiments unter einer nicht realisierten Treatmentbedingung. Der tatsächliche Effekt eines spezifischen Treatments ist also der Unterschied zwischen dem, was passiert ist, und dem, was in einer anderen Realität hätte passieren können (Cook et al., 2002). Da kontrafaktische Realitäten nicht realisiert werden können, bedeutet dies, dass die Ergebnisvariablen von Treatmentuntersuchungen immer unvollständig sind. Diese Unvollständigkeit nennt man partielle Identifikation.

Partielle Identifikation allgemein bedeutet, dass der Sampling-Prozess einer Zufallsvariable nur teilweise beobachtbar ist. Daraus folgt, dass die Beobachtungen interessierende Größen (Parameter) der Zufallsvariable nicht identifizieren können. Partielle Identifikation ist also kein Schätzproblem, das sich mit größeren Stichproben oder klüger gewählten Schätzstatistiken beheben lässt (Manski, 2003). Der Ansatz der partiellen Identifikation ist eine komplementäre Sichtweise der klassischen Sichtweise der Punktidentifikation. Er distanziert sich von der Annahme, dass Verteilungen vollständig spezifiziert werden können und stellt empirischen Forscher:innen daher Methoden zur Verfügung, die dabei helfen, das Spektrum an Informationen zu untersuchen, die wir über einen interessierenden Parameter unter Verwendung verschiedener Annahmen erhalten (Tamer, 2010).

In Treatmentstudien sorgt ausschließlich Randomisierung für Unabhängigkeit zwischen der Treatmentwahl und der abhängigen Variable. So können die Verteilungen der beobachteten Daten

eindeutig identifiziert werden und nicht gemessene Störeinflüsse, die neben dem Treatment Einfluss auf die Zielvariable nehmen, mitteln sich bei ausreichend großen Stichproben aus.

Eine dritte Quelle der Unsicherheit in Nutzenfunktionen entspringt den Messfehlern in den Kovariablen. Bisher wurde in allen bislang bekannten Modellen angenommen, dass die Kovariablen fehlerfrei gemessen werden können. In den Sozialwissenschaften ist es aber üblich, psychometrische Werte aus Fragebögen und Tests zur Entscheidungsfindung zu verwenden. Diese sind in der Regel mit nicht zu vernachlässigenden Messfehlern behaftet. Die Messfehler werden entweder aus den psychometrischen Modellen selbst oder unter Verwendung der globalen Reliabilitäten ermittelt und in Form von Konfidenzintervallen ausgedrückt (z. B. Rost, 2004; Thissen & Wainer, 2001). Diese Studie soll deshalb erstmals eine Entscheidungsfunktion vorstellen, welche die Unsicherheit durch die Messfehler in den Kovariablen bei der Treatmententscheidung berücksichtigt (vgl. Abschnitt 2).

5.3.2 Einzelne Entscheidungen vs. wiederholte Entscheidungen

Seit Wald (1950) gilt das asymptotische Verhalten von Schätzern als ausschlaggebend für ihre Gütebewertung und statistische Tests werden so konstruiert, dass bei unendlicher Wiederholung ihr Fehler erster Art gegen ein festgelegtes Alphaniveau konvergiert.

Demgegenüber befasst sich die Finite-Sample Statistik mit Optimalitätskriterien, die greifen, bevor die Asymptotik einsetzt (Manski, 2004). Damit können auch im Treatmententscheidungsproblem zwei Fokusse gesetzt werden. Wenn asymptotisches Verhalten von Schätzern als relevant betrachtet werden soll, so kann dies als Fokus auf dem Allgemeinwohl verstanden werden. Denn dann wird berücksichtigt, dass für sehr viele Personen einen Treatmententscheidung getroffen wird und damit zum Beispiel der mittlere Treatmenterfolg der Population maximiert werden soll. Es kann aber auch der Fokus auf die einzelne Person gelegt werden. Diese profitiert nicht von guten Treatmententscheidungen für andere Personen, sondern wird vermutlich eine Entscheidungsstrategie bevorzugen, die nicht den mittleren Treatmenterfolg in der Population maximiert, sondern bereits bei einmaliger Wiederholung oder wenigen Wiederholungen optimal ist. Dies könnte der Fall sein, wenn etwa das eigene Risiko für Misserfolg minimiert wird (vgl. Kahneman & Tversky, 1984). Ein illustrierendes Beispiel für dieses Phänomen bietet das St. Petersburg Paradoxon (Bernoulli, 1738), das wie folgt lautet. Ein Wettbüro bietet ein Glücksspiel für eine:n einzelne:n Spieler:in an, bei dem in jeder Runde eine faire Münze geworfen wird. Der anfängliche Einsatz beginnt bei zwei Dollar und wird jedes Mal verdoppelt, wenn Kopf erscheint. Wenn zum ersten Mal Zahl erscheint, endet das Spiel und die:der Spieler:in gewinnt den gesamten Betrag im Pott. So gewinnt die:der Spieler:in zwei Dollar, wenn beim ersten Wurf Zahl erscheint, vier Dollar, wenn beim ersten Wurf Kopf und

beim zweiten Wurf Zahl erscheint, acht Dollar, wenn bei den ersten beiden Würfen Kopf und beim dritten Wurf Zahl erscheint, usw. Mathematisch gesehen, gewinnt die:der Spieler:in in Erwartung $\mathbb{E}(u) = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + ... = 1 + 1 + 1 + ... = \infty$ Dollar. Es zeigt sich hier, dass der Erwartungswert der angebotenen Lotterie zwar ∞ ist, der durchschnittliche Preis, den Probanden bereit sind zu zahlen, jedoch nur bei wenigen Dollar liegt (Hayden & Platt, 2009). Damit ergibt sich, dass bei individuellen Entscheidungen asymptotisches Verhalten von Zufallsprozessen nicht zwangsläufig berücksichtigt wird. Dabei Probanden grundsätzlich die rationale Handlungsfähigkeit abzusprechen, greift zu kurz (Gilboa, 2009). Andere Erklärungsansätze vermuten, dass es neben der objektiven Nutzenstruktur eine überlagernde indviduelle Nutzenstruktur gibt, welche nicht linear in der objektiven Nutzenstruktur (hier den Geldwerten ist) (Blavatskyy, 2005). Auch ist es plausibel, dass das Greifen der Asymptotik für echte Lebenssituationen aufgrund endlicher Ressourcen an Geld und Zeit schlichtweg nicht relevant ist (Feller, 2008). Daher ist es fraglich, ob auf Asymptotik basierende Entscheidungsregeln für Individuen jemals als rational gelten können. Denn ähnlich wie im St. Petersburg Paradoxon könnte es sein, dass diese zwar im Erwartungswert also im Unendlichen optimal sind, diese Eigenschaft aber für die einzelne und auch die endlich oft wiederholte einzelne Entscheidung praktisch nicht relevant sind.

Legt man den Fokus dennoch auf wiederholte Entscheidungen, so lassen sich in der Entscheidungstheorie sogenannte randomisierte Entscheidungsregeln ableiten. Hier wird eine einzelne Entscheidung auf Basis eines Zufallsexperiments getroffen. Dieses Vorgehen kann sinnvoll sein, wenn zwei oder mehrere Entscheidungsregeln gleich optimal sind, und/oder wenn man sich im Sinne der Spieltheorie mit einem feindlichen Gegenspieler konfrontiert sieht. Wird die einzelne Entscheidung ausgelost, so wird das eigene Handeln für den feindlichen Gegenspieler unberechenbar(er). Da das hier vorgestellte Modell jedoch für die Anwendung in der Unterrichtspraxis konstruiert wird, werden solche randomisierten Entscheidungsregeln ausgeschlossen. Es erscheint aus individueller und Anwenderperspektive für Einzelfallentscheidungen nicht plausibel, dass Entscheidungen besser werden, wenn man sie auf Basis von Zufallsexperimenten trifft. Außerdem gilt, dass zu jeder randomisierten Entscheidungsregel eine nicht-randomisierte Bayes-Entscheidungsregel existiert, welche mindestens gleich gut ist (Ferguson, 1967).

5.3.3 Individuelle Risikoaversion

Neben objektiv und mathematisch begründbaren Entscheidungsgrundsätzen beinhaltet das Entscheiden auch eine psychologische Komponente. Nicht alle Menschen entscheiden sich in denselben Situationen gleich. Es gibt beim Entscheiden interindividuelle Unterschiede, die sich unter anderem

auf unterschiedliche Risikofreudigkeit zurückführen lassen (Artinger, Artinger, & Gigerenzer, 2019). Ein entscheidungstheoretisches Modell sollte daher auch die Flexibilität besitzen, solche individuellen Präferenzen abbilden zu können. Eine Möglichkeit diese individuellen Präferenzen abzubilden, ist die Annahme einer individuellen Nutzenfunktion, welche mit der objektiven Nutzenfunktion verrechnet ist beziehungsweise diese verändert oder aktualisiert. In Bezug auf Risikoaversivität hat es sich in empirischen Untersuchungen gezeigt, dass der individuelle Nutzenverlust für Geldverluste im Vergleich zum individuellen Nutzen für Geldgewinne deutlich größer ausfällt (Kahneman & Tversky, 1984). Gerade im Bereich der Sozialplanung und der politischen Entscheidungen wird zum Beispiel grundsätzlich konservativ und risikoavers entschieden (Manski & Tetenov, 2007). Auch für Treatmententscheidungen kann daher angenommen werden, dass zumindest einige Entscheidungstragende eine sehr konservative Haltung einnehmen und zum sogenannten defensiven Entscheiden neigen (Artinger et al., 2019; Manski & Tetenov, 2007). Typisch wäre für diese Personen, dass sie bei ihrer Treatmentwahl anstelle eines großen Erfolgs einen möglichst kleinen Misserfolg anstreben.

Ein möglicher Umgang mit Risikoaversie in der entscheidungstheoretischen Modellierung ist das Überstülpen einer individuellen Nutzenfunktion auf die objektive Nutzenstruktur mit anschließender Erwartungswertmaximierung (Clemen, 1996; Kirkwood, 1997, 2004). Siehe auch Keefer (1991) für eine Anwendung im Bereich des Finanzmanagements.

Als Alternative dazu kann neben dem Erwartungsnutzen auch die Varianz der erwarteten Nutzenverteilung mit in die Entscheidungsregeln einbezogen werden (Jia & Dyer, 1996; Kroll, Levy, & Markowitz, 1984; Markowitz, 2014). Schlussendlich können auch Entscheidungsregeln verwendet werden, welche grundsätzlich immer vom Worst-Case-Szenario, also vom ungünstigsten Fall ausgehen und diesen optimieren. Eine Entscheidungsregel, die diese maximal mögliche Risikoaversion umsetzt, ist das Minimax-Prinzip. Es wurde im Rahmen der Spieltheorie entwickelt (Nash, 1951) und ist dann optimal, wenn die Natur, bzw. der Zufallsprozess, welcher die Umweltzustände realisiert, als feindlicher Gegenspieler aufgefasst wird. Dieser realisiert dann in Abhängigkeit der eigenen Entscheidung immer den schlecht möglichsten Umweltzustand. Zu Recht wird das Minimax-Prinzip häufig als unrealistisch oder sehr pessimistisch kritisiert und sollte nur dann angewendet werden, wenn alle Komponenten des Entscheidungsmodells auch von praktischer Relevanz sind (Chamberlain, 2000; Good, 1952). Enthält das Entscheidungsproblem de facto unrealistische Ereignisse, so führt das Minimax-Prinzip in der Regel nicht zu als rational wahrgenommenen Entscheidungen. Denken wir etwa an das Einleitungsbeispiel mit Regenhose und Sonnencreme zurück. Das entscheidungstheoretische Problem wird im Beispiel ad Absurdum geführt, wenn extrem unwahrscheinliche Szenarien mit katastrophalen Folgen wie etwa ein Terroranschlag zu den möglichen Umweltzustän-

den des Problems gehören. Das Minimax-Prinzip würde stets von diesem Szenario ausgehen und in Folge würden Aktionen präferiert (hier zum Beispiel das Haus nie wieder verlassen), welche bei intuitiver Betrachtung als wenig rational gelten dürften. Eine häufige Anwendung des Minimax-Prinzips ergibt sich daher auf Stichprobenverteilungen (begrenzt, da Anzahl an Beobachtungen in Stichproben immer endlich sind), auf getrimmte Verteilungen (z. B. Stoye, 2012) oder für Fälle, in welchen keine praktisch irrelevanten Umweltzustände mit sehr geringen Eintretenswahrscheinlichkeiten vorkommen (z. B. Chamberlain, 2000).

5.4 Bisherige Modelle

Im vorangegangenen Absatz wurde auf theoretischer Ebene die Struktur der Treatmententscheidung analysiert und Bedingungen formuliert, die Einfluss auf die Optimalität verschiedener Entscheidungsstrategien nehmen. Im Folgenden soll nun dargestellt werden, welche Treatmententscheidungsmodelle unterschiedliche Akteure der bildungswissenschaftlichen, psychologischen und statistischen Treatmentforschung explizit und implizit verwenden. Daran anknüpfend wird dann eine eigene Variante eines TreaDeMs vorgestellt.

5.4.1 Traditioneller Ansatz der Evaluationsforschung

Treatments jeglicher Natur - seien es große staatlich organisierte Förderprogramme oder kleine für Abschlussarbeiten entwickelte Interventionen - wollen, ja müssen gar evaluiert und in ihrer Wirksamkeit bestätigt werden, wenn sie innerhalb der Community Bestand haben sollen. Fast immer bleibt dabei der eigentliche Grund der Evaluation im Ergebnisbericht implizit: die Empfehlbarkeit des Programmes gegenüber anderen Programmen oder einer Standardprozedur. So verweist auch Newcomer in ihrem Handbuch der Evaluation (Newcomer, Hatry, & Wholey, 2015) zwar darauf, dass neben Signifikanztests auch die Effektgröße bei Empfehlungen betrachtet werden müssen - wie für wen und auf welchem Algorithmus Empfehlungen ausgesprochen werden sollten, wird jedoch nicht ausgeführt. Gleiches zeigt sich in Einzelstudien, welche Fördermaßnahmen im Bildungsbereich evaluieren (vgl. Cronbach & Snow, 1977; Gold, Trenk-Hinterberger, & Souvignier, 2009; Lee & Park, 2008; Van Keer & Verhaeghe, 2005). Signifikante Unterschiede zwischen geförderten und Kontrollgruppen bilden die Basis, die Programme als erfolgreich zu betrachten und implizit Treatmentempfehlungen dafür auszusprechen. Welche Annahmen über individuelle Treatmenteffekte stecken jedoch in der Betrachtung signifikanter Gruppenunterschiede? Verwendet man ein klassisches varianzanalytisches Design mit zwei oder mehr Gruppen, so ergibt sich für alle Personen dieselbe prädiktive Treatmenteffektverteilung. Es gilt also:

$$f(u_i|T_n) = f(u_j|T_n)$$

mit:

- u_i = Nutzen der Person i
- u_j = Nutzen der Person j
- T_n = Treatment n
- f = Verteilungsfunktion der Nutzenwerte
- $f(u|T_n)$ = Verteilungsfunktionen der Nutzenwerte von Treatment n

Die Vorhersageverteilung für jede einzelne Person entspricht nämlich der Verteilung der Kriteriumswerte nach Durchführung der Treatments. Da in dieser Betrachtung die Vorhersageverteilung nicht auf weitere Kovariablen bedingt wird, wird der Treatmenteffekt also für alle Personen konstant geschätzt. Als Nutzen wird die abhängige Variable definiert. Sie beschreibt in der Regel den Trainingserfolg auf einer interessierenden Dimension. Als Entscheidungsfunktion dient in dieser Betrachtung der Hypothesentest zwischen den Verteilungen der Kriteriumswerte der Treatmentgruppen, z. B. ein t-Test für unabhängige Stichproben. Bezüglich der berücksichtigten Unsicherheiten bedeutet die Anwendung eines Hypothesentests, dass die Unsicherheit in der Verteilung in Form der Varianz der Verteilung berücksichtigt wird, sowie Schätzunsicherheiten der Mittelwerte und der Varianzen der Verteilungen. Zur Gewinnung des Optimalitätskriteriums (kritischer Testwert des statistischen Tests) wird eine nicht-symmetrische Verlustfunktion angewendet, welche den Fehler erster Art stärker gewichtet, als den Fehler zweiter Art. Hypothesentests implizieren also eine Aversion gegen Unterschiede in der Wirksamkeit. Man könnte diese Aversion auch als konservative Haltung bezeichnen.

Als Kritik ist anzumerken, dass ein konstanter Treatmenteffekt unwahrscheinlich ist. Aus theoretischer Überlegung heraus lassen sich für alle Anwendungsfälle Gelingensbedingungen und Prämissen formulieren unter denen Treatments wirksam sind (Van der Linden, 1981). Eine Ausnahme mögen Treatmentevaluationen bilden, deren Stichprobe bereits auf Basis bereits bekannter Gelingensbedingungen vorselektiert ist. Zudem erscheint ein Hypothesentest als Entscheidungskriterium eher schlecht geeignet, da Hypothesentests zum einen auf Asymptotik basieren und da die zugehörige Loss-Funktion zu recht strengen Cutoffs führt. Wie eingangs im Abschnitt 5.2 bereits ausgeführt, ist ein Entscheidungskriterium kritisch zu sehen, das asymptotisches Verhalten von Statistiken ausnutzt, wenn das Entscheidungsproblem aber nur für eine einzelne Entscheidung eingesetzt wird. Darüber hinaus zeigt Manski (2004), dass sich grundsätzlich Entscheidungen schon viel früher verbessern lassen. Früher das bedeutet, bevor ein Treatmenteffekt das in den Sozialwissenschaften gängige Signifikanzkriterium von einem p-Wert kleiner gleich 0.05 erreicht. Zudem nehmen Signi-

fikanztests zu geringen Alphaniveaus immer eine ungleiche Gewichtung der Testfehler erster und zweiter Art vor. Indem das Alphaniveau fixiert wird und sich das Betaniveau aus der Effektgröße und der Stichprobengröße ableitet, verfügen die üblichen Evaluationsstudien selbst bei angemessener Teststärke von 0.80 zu einem rund viermal größeren Betafehler im Vergleich zum Alphafehler. Entscheidungstheoretisch entspricht dies aber einer Gewichtung, welche dem gleichkommt, dass der Alphafehler viermal so schlimm ist wie der Betafehler. Gerade diese ungleiche Gewichtung der Fehlertypen erscheint bei der Treatmententscheidung als wenig sinnvoll. Unterscheiden sich nämlich die Treatments tatsächlich nicht in ihrer Wirksamkeit, so kann im entscheidungstheoretischen Sinne eigentlich keine Fehlentscheidung getroffen werden. Eine übermäßige Absicherung gegen die Gleichheit in der Wirksamkeit durch ein niedrig gewähltes Alphaniveau erscheint also unter entscheidungstheoretischer Perspektive nicht als nachvollziehbar.

5.4.2 Diagnostische Modelle mit deterministischem, linearem Nutzen

In der psychologischen Diagnostik wurden systematische Überlegungen zur Treatmentwahl zu Beginn der 60er Jahre intensiviert. Aufbauend auf statistischen Modellen, wie sie etwa von Brodgen (1949) in den 40er und 50er Jahren formuliert wurden, präsentieren Cronbach und Gleser (1965) verschiedenste diagnostische Entscheidungsmodelle. Zur Schätzung der Nutzenfunktion schlagen sie eine Regressionsanalyse vor. Die Einfachheit der Analyse begründen sie mit den Rechenkapazitäten der damaligen Zeit. In dieser Regression bildet ein Nutzenkriterium die abhängige Variable und mögliche Kovariablen für den Treatmenterfolg bilden die Prädiktoren. So ergibt sich in ihrem Modell ein individueller Treatmenteffekt, welcher linear abhängig von den Kovariaten ist. Die Schätzwerte der Regressionsgeraden werden dann als deterministische Nutzenfunktion aufgefasst. Damit wird die individuelle Vorhersageverteilung quasi als Einpunktverteilung angenommen, welche durch den gefitteten Wert des Regressionsmodells definiert wird. Diese Prozedur wird für jedes Treatment durchgeführt, so dass für jedes Treatment ein linearer Zusammenhang zwischen den Kovariaten und dem Treatmenterfolg entsteht (Cronbach & Snow, 1977). Als Entscheidungsfunktion dient die Ordnungsrelation der auf die Kovariaten bedingten Vorhersagen der jeweiligen Treatmentgeraden. Konkret wird für eine Person mit festen Kovariaten dasjenige Treatment ausgewählt, dessen Nutzenfunktion den höchsten Wert annimmt. Im Rahmen der psychologischen Diagnostik stellen diese Überlegungen, welche hier nur sehr verknappt dargestellt wurden, eine erste umfassende und systematische Beschreibung des unterliegenden Entscheidungsproblems der Treatmententscheidung dar. Dennoch bleibt anzumerken, dass in diesen Modellen die Unsicherheit, welche aus der Schätzung der Nutzenfunktion resultiert, unberücksichtigt bleibt. Einmal geschätzt, wird der Zusammenhang zwi-

schen Kovariablen und Zielkriterium als deterministisch angesehen. Damit reduziert sich zwar die Entscheidungsfunktion in ihrer Komplexität auf eine einfache Ordnungsrelation, die Modellpassung des Regressionsmodells und die tatsächliche Unsicherheit in der prädiktiven Vorhersageverteilung wird aber mit ungewissem Grad unterschätzt. Zudem ist es fraglich, ob der Zusammenhang zwischen Kovariablen und Treatmenterfolg tatsächlich immer linear ist. Gerade als Extrapolation in die Randbereiche gemessener Kovariablen erscheint ein linearer Zusammenhang fast grundsätzlich unrealistisch. Dies ist in Bezug zu Nutzenfunktionen im Besonderen der Fall. Unbegrenzte, lineare Zusammenhänge implizieren nämlich immer, dass Nutzenwerte beliebig groß und klein werden können. Eben die Restriktion der Linearität im Nutzen wird unter anderem von Van der Linden (1981) diskutiert, dessen Modelle diese Restriktion auflösen. Daher werden seine Modelle im nächsten Abschnitt vorgestellt.

5.4.3 Psychologisch-psychometrisch motivierte Modelle

Aufbauend auf Modellen, welche den Nutzen einzelner Treatments als lineare Funktionen von Kovariaten definieren und Entscheidungsgrenzen über die Schnittpunkte der so geschätzten Regressionsgeraden definieren, stellt Van der Linden (1981) einen weiteren entscheidungstheoretisch motivierten Ansatz zur Treatmententscheidung vor. In seinem Modell werden auf Basis von experimentellen Daten individuelle prädiktive Verteilungen für ein Nutzenkriterium oder eine Funktion eines Nutzenkriteriums geschätzt. Durch die Bestimmung des Risikos der Verteilungen ergeben sich so Entscheidungsgrenzen im Kovariablenraum. Das Risiko wird, der allgemeinen statistischen Theorie folgend über eine Erwartungswertbildung über die Nutzenfunktion bestimmt. Hierbei findet eine Einschränkung auf monotone Entscheidungsregeln statt. So gilt stets, dass ein jeweiliges Treatment nur auf einem geschlossenen Intervall des Kovariablenraums als optimal gelten kann. Hinzu wird als Form der bedingten Verteilungen auf klassische Verteilungsfunktionen wie zum Beispiel die Normal-Ogive-Verteilungsfunktion zurückgegriffen. Diese treten auf, wenn ein regressionsanalytisches statistisches Modell zur Schätzung der prädiktiven Verteilung eingesetzt wird.

Als Entscheidungsfunktion dienen allerdings auch in diesem Modell wieder die Ordnungsrelationen der Erwartungswerte der Nutzenverteilungen. Dabei wird immer dasjenige Treatment mit dem größten Erwartungswert empfohlen. Diese Herangehensweise stellt in gewisser Hinsicht eine Verallgemeinerung des von Cronbach und Gleser (1965) vorgestellten Entscheidungsmodells dar. Dennoch beschränkt sich das Modell von Van der Linden auf monotone Entscheidungsregeln. Somit ist die Form, die die Nutzenfunktion annehmen kann, auf monotone Funktionen beschränkt. Zudem wird zwar die Unsicherheit aus der Schätzung der Nutzenfunktion im Modell berücksichtigt, die Unsi-

cherheit in der Messung der Prädiktoren der Nutzenfunktion jedoch nicht.

Verschiedene Entscheidungsfunktionen werden im Kontext diagnostischer Entscheidungen vermutlich erstmals bei Rudner (2009) im Rahmen seiner "measurement decision theory" vorgestellt. Das zur Klassfikation und Bestimmung des Masterys entwickelte Modell versteht sich primär als ein testtheoretisches. Es schätzt für Itemantwortmuster Wahrscheinlichkeitsverteilungen für verschiedene Masteryzustände. Voraussetzung ist, dass die einzelnen Itemlösungswahrscheinlichkeiten unter den verschiedenen Masteryzuständen a priori bekannt sein müssen. Diese können zum Beispiel aus Vorstudien stammen, welche mit einem externen oder einem Globalurteil den Masteryzustand definierten. So ergeben sich zu jedem Itemantwortmuster für jeden möglichen Masteryzustand eine prädiktive Wahrscheinlichkeitsverteilung. Die so entstehenden verschiedenen Verteilungen können dann über verschiedene Entscheidungsfunktionen in eine Entscheidung für einen der Masteryzustände überführt werden. Hier werden die "maximum probability of error decision" (1), die "maximum likelihood decision" (2), das "MAP-Kriterium" (3) und das "Bayes risk Kriterium" (4) vorgestellt. Sie entsprechen dabei einem Signifikanztest (1), dem Erwartungsnutzenkriterium bei Betrachtung der Likelihood als Nutzenfunktion (2), einem Erwartungsnutzenkriterium unter Berücksichtigung einer zwingenden Quote der Masteryzustände (3) und dem klassischen Bayes-Entscheidungskriterium der statistischen Entscheidungstheorie (4) (vgl. Berger, 1985). In Abhängigkeit der Eigenschaften des vorliegenden entscheidungstheoretischen Problems bieten diese verschiedenen Entscheidungsfunktionen erstmals die Möglichkeit Präferenzen der Entscheidungsträger:innen und Besonderheiten der Entscheidungssituation angemessen zu berücksichtigen.

5.4.4 Mathematisch-statistisch motivierte Modelle

Des Weiteren gibt es ein Teilgebiet der statistischen Entscheidungstheorie, welches sich mit Treatmententscheidungen als Anwendungsbereich der partiellen Identifikation auseinandersetzt. Als Autoren sind hier z. B. Manski, Dehejia oder Stoye zu nennen. In ihren Modellen wird, ähnlich wie in den psychologisch-psychometrischen Modellen, zunächst in Abhängigkeit von Kovariaten eine auf die Treatments bedingte prädiktive Vorhersageverteilung eines Nutzens geschätzt. In dieser werden viele verschiedenen Quellen der Unsicherheit berücksichtigt. Deheja (2005) etwa verwendet zur Schätzung einen bayesianischen Ansatz. In diesem berücksichtigt er nicht nur die Unsicherheit in der Vorhersageverteilung mittels ihrer Varianz, sondern auch die Unsicherheit in ihren Parametern, welche sich in seiner Studie aus der Verwendung unterschiedlicher Priors ergibt. Als Entscheidungsfunktion stellt Deheja einen Vergleich der individuellen Erwartungswerte der prädiktiven Vorhersageverteilungen vor und analysiert diese unter verschiedenen Präferenzen der Entscheidungs-

tragenden. Als Präferenzen werden eine Aversion gegen Ungleichheit der Treatmenteffekte keiner Aversion gegen Ungleichheit gegenübergestellt. Eine Aversion gegen Ungleichheit bedeutet hier, dass Entscheidungsregeln bevorzugt werden, welche die erwarteten Zuwächse durch ein Förderprogramm möglichst gleichmäßig auf die Individuen mit ihren unterschiedlichen Voraussetzungen verteilen. So können Treatmentregeln gebildet werden, die nicht Individuen mit besonders guten oder besonders schlechten Voraussetzungen bevorzugen oder benachteiligen.

Andere Autoren wie Manski und Stoye wenden Methoden der partiellen Identifikation auf prädiktive Vorhersageverteilungen an und stellen damit erstmals Modelle vor, welche explizit für Daten aus nicht-experimentellen Settings geeignet sind (Manski & Tetenov, 2007; Stoye, 2012). In ihren Modellen ergeben sich wegen der partiell identifizierten Vorhersageverteilungen als Entscheidungsfunktion eine Betrachtung der stochastischen Dominanz. Bei getrimmten oder diskreten Verteilungen wenden sie das Minimax-Kriterium an. Weitere Entscheidungsfunktionen lassen sich aufgrund der nicht punkt-identifizierten Verteilungsparameter nicht anwenden.

5.4.5 Zusammenfassung

Es zeigen sich also zusammenfassend viele Herangehensweisen, das entscheidungstheoretische Problem der Treatmententscheidung zu formalisieren und zu lösen. Es lassen sich unter Verwendung von Kovariaten prädiktive Verteilungen von Nutzenwerten schätzen und dabei Unsicherheiten in der Schätzung, den Parametern und der Identifikation der Verteilungen berücksichtigen. Auch lassen sich dabei verschiedene Präferenzen der Entscheidungsträger:innen bei der Bildung der Entscheidungsregeln formalisieren. Eine weitere Quelle der Unsicherheit im Entscheidungsprozess bildet die Messung der Kovariaten. Im Folgenden wird deswegen ein neues Modell vorgestellt, das diese Messunsicherheit berücksichtigt.

5.4.6 Entscheidungsfunktionen für die individuelle Treatmentwahl in dieser Studie

In dieser Studie wird eine Entscheidungsfunktion erarbeitet, welche bei Kenntnis über Leseverstehens- und Leseflüssigkeitskompetenz von Grundschüler:innen entweder das FiLBY-2-, oder das FiLBY-3-Training als wirksamer kennzeichnet. Dabei ergeben sich die prädiktiven Nutzenverteilungen aus der in Studie 2 geschätzten Nutzenfunktionen. Diese Studie baut also auf einer vollständigen Spezifizierung der prädiktiven Nutzenverteilung auf. Allerdings verfügte die Nutzenfunktion aus Studie 2 über eine relativ große Unsicherheit durch die Schätzung. In Studie 2 zeigte sich nur eine mittelmäßige Varianzaufklärung. Das heißt, es müssen weitere noch unbekannte Einflüsse auf den Trainingserfolg angenommen werden. Trotzdem wird in dieser

Studie der Erwartungswert der prädiktiven Nutzenverteilung für einzelne Personen als deterministischer Nutzen angenommen. Diese Einschränkung wird vorgenommen, weil die punktweisen Standardfehler für beide Treatments im interessierenden Kovariablenbereich in etwa gleich groß sind (vgl. Abbildung 29). So wird zwar die absolute Unsicherheit in der Entscheidung unterschätzt, das Optimalitätskriterium selbst aber nicht verzerrt. Die Unsicherheit im Entscheidungsproblem stammt in diesem Modell also nicht aus der Nutzenfunktion, sondern aus den Messmodellen der Prädiktoren der Nutzenfunktion. Diese sind in dieser Studie zwei standardisierte Lesetests, deren Messfehler aus der psychometrischen Modellierung gewonnen werden. So ergibt sich für jedes Individuum eine zweidimensionale Wahrscheinlichkeitsverteilung im Kovariablenraum, welche mit der Nutzenfunktion verrechnet wird, um so ein Optimalitätskriterium zu erhalten.

Dabei soll der Fokus auf Einzelentscheidungen gelegt werden, da bei der Wahl eines Lesetrainings das Wohlergehen jedes einzelnen Kinds zählt, nicht das aller Kinder im Mittel. Bezüglich der persönlichen Risikopräferenz, ist es der Anspruch, dass diese von der:vom zukünftigen Entscheidungsträger:in selbst definiert werden soll. Daher werden zwei Entscheidungsfunktionen vorgestellt: eine Erwartungsnutzenmaximierung und eine Maximin-Regel. Die Erwartungsnutzenmaximierung entspricht dabei einer risikoneutralen Haltung. Die Anwendung der Maximin-Regel entspricht einer maximal risikoaversen Haltung.

5.5 Methode

Die hier vorgestellten Analysen bauen direkt auf den vorangegangenen Studien 1 und 2 auf. Es werden zur Bestimmung zweier Entscheidungsfunktionen sowohl die psychometrischen Modellierungen aus der Studie 1, als auch die in Studie 2 geschätzten Nutzenfunktionen verwendet. Es wird jeweils eine Entscheidungsfunktion per Erwartungsnutzenmaximierung und eine, die dem Maximin-Prinzip folgt, entwickelt. Dies wird für die beiden Zielkriterien der Steigerungen in Leseflüssigkeit und Leseverstehen durchgeführt. Im Folgenden wird daher kurz auf die Datengrundlage und anschließend auf die Implementation der beiden Entscheidungsfunktionen eingegangen. Alle Analysen wurden in R (R Core Team, 2020) durchgeführt.

5.5.1 Datengrundlage

Als Datengrundlage für die Entwicklung der Entscheidungsfunktionen wurden exemplarisch die MIRT-Modelle der Version des BYLET-A und die Raschmodelle des SLS-B1 verwendet. Die individuellen Messwertverteilungen wurden für den BYLET aus den Punktschätzern und Standardfehlern

der MAP-Schätzung gewonnen. Für das SLS wurde aus den Maximum Likelihood Schätzern und deren Standardfehlern eine Wahrscheinlichkeitsverteilung der individuellen Personenparameter gebildet. Da sowohl die bayesianische a Posterioriverteilung der Personenparameter, als auch die Likelihood (diese zumindest asymptotisch) normalverteilt sind (Reckase, 2009), lag es nahe für die intraindividuelle Personenparameterverteilung eine Normalverteilung anzunehmen. Da die Varianz der Personenparameterschätzungen zwischen den Versionen leicht variiert, wurde so sichergestellt, dass nicht alle Schlussfolgerungen zusätzlich auf die Testversionen bedingt werden mussten. Als Nutzenfunktionen wurden die finalen, reduzierten Nutzenmodelle verwendet, welche den Trainingserfolg nur auf Basis des BYLETs und des SLS vorhersagen. Die random Effekte wurden bei der Entscheidungsfunktionsbildung nicht berücksichtigt, da sie für zukünftige Schüler:innen nicht bekannt sind.

5.5.2 Erwartungsnutzen-Entscheidung

Die Erwartungsnutzen-Entscheidung beruht auf der Verrechnung einer diskreten oder stetigen Nutzenfunktion mit einer stetigen oder diskreten Wahrscheinlichkeitsverteilung für die Umweltzustände des entscheidungstheoretischen Problems. Der Erwartungsnutzen wird dabei für jede Aktion einzeln berechnet. Handelt es sich um diskrete Wahrscheinlichkeitsverteilungen, so ergibt sich der Erwartungsnutzen als

$$\mathbb{E}(u(a_i)) = \sum_{j=1}^{n} p(\theta_j)u(a_i, \theta_j)$$

mit:

- $\mathbb{E}(u(a_i))$ = Erwartungsnutzen der Aktion a_i
- $p(\theta_j)$ = Wahrscheinlichkeit für den Umweltzustand θ_j
- $u(a_i, \theta_j)$ = Nutzen der Aktion a_i, wenn θ_j eintritt
- n = Anzahl der möglichen Umweltzustände

Handelt es sich um stetige Wahrscheinlichkeitsverteilungen der Umweltzustände, so wird die Summe durch das Integral ersetzt und der Erwartungsnutzen ergibt sich zu:

$$\mathbb{E}(u(a_i)) = \int_{\theta} u(a_i, \theta)d\pi(\theta)$$

mit:

- $\mathbb{E}(u(a_i))$ = Erwartungsnutzen der Aktion a_i
- $\pi(\theta)$ = Wahrscheinlichkeitsverteilung für die Umweltzustände θ
- $u(a_i, \theta)$ = Nutzen der Aktion a_i, für θ

In dem hier vorgestellten Modell stellen die Messwerte der Lesekompetenz die Umweltzustände dar, über die aufgrund von Messfehlern Unsicherheit in Form von Wahrscheinlichkeitsverteilungen besteht. Die Aktionen bilden das FiLBY-2- und das FiLBY-3-Training. Die Nutzenfunktion wurde in Studie 2 entwickelt.

5.5.2.1 Bestimmung der Entscheidungsfunktion Wird zur Entscheidung eine Erwartungsnutzenmaximierung herangezogen, so muss zunächst der Erwartungsnutzen berechnet werden. Im hier dargestellten Kontext wurde zur Erwartungsnutzenbildung zunächst die Wahrscheinlichkeitsverteilung der Personenparameter mit den vorhergesagten Nutzenwerten aus der Nutzenfunktion gewichtet. Anschließend wurde der Erwartungswert durch eine Integralbildung bestimmt. Dabei wurde der Definitionsbereich der Nutzenfunktion und der Wahrscheinlichkeitsverteilung der Kovariablen auf einen plausiblen Bereich eingeschränkt. Da bei standardisierten Personenparametern rund 99% der Werte innerhalb zweieinhalb Standardabweichungen liegen, wurden die Werte -2.5 und $+2.5$ als Grenzen festgelegt. Für die restlichen ein Prozent der extremsten Beobachtungen, wurden als Approximation die Vorhersagen für die Werte -2.5 bzw. 2.5 verwendet. So konnte sichergestellt werden, dass keine Funktionsbereiche der Nutzenfunktion Einfluss auf die Entscheidung nahmen, in welchen die Nutzenfunktionsschätzung nur eine Extrapolation darstellte, also de facto nicht auf Daten geschätzt wurde. Gleichzeitig mussten keine Beobachtungen aus der Analyse ausgeschlossen werden. Es wurde also für die beiden verglichenen Trainings der folgende Ausdruck berechnet:

$$\mathbb{E}_{\pi(X_i)}(u(T_n)) = \int_{-2.5}^{2.5} u(T_n|X_i)d\pi(X_i)$$

mit:

- T_n = Treatment n
- X_i = Kovariablenvektor der Person i ($\theta_{sls}, \theta_{bylet}$)
- θ_{sls} = Personenparameter des Salzburger Lese-Screenings
- θ_{bylet} = Personenparameter des Bayerischen Lesetests
- $\pi(X_i)$ = Wahrscheinlichkeitsverteilung des Kovariablenvektors der Person i

Die finale Treatmententscheidung fiel dann auf das Treatment, welches bei gegebenen Personenparametern (Kovariablenvektor) den höchsten Erwartungsnutzen aufwies.

5.5.2.2 Bestimmung der Entscheidungssicherheit Die Entscheidungssicherheit wurde über die Differenz der Optimalitätskriterien für beide Treatments bestimmt. Um diese Differenz bewerten zu können, muss die Skala des Erwartungsnutzens und damit auch der Erwartungsnutzendifferenz

interpretierbar sein. Im klassischen statistischen Entscheidungsproblem wird der Erwartungsnutzen gewonnen, indem die gewichtete Nutzenfunktion über den gesamten Umweltzustandsraum integriert wird. In dieser Studie wurde der Umweltzustandsraum jedoch auf den Parameterraum zwischen -2.5 und 2.5 begrenzt. Je nach Lage des Erwartungswerts von π befand sich also nur zwischen knapp 25% und 99% der Wahrscheinlichkeitsmasse im Auswertungsbereich des Integrals. Knapp 99% wenn der Erwartungswert genau in der Mitte des plausiblen Bereichs liegt, knapp 25% wenn der Erwartungswert genau an einer der vier Ecken liegt. Damit ergab der Erwartungsnutzen sich nicht, wie im klassischen bayesianischen Entscheidungskriterium als gewichteter Nutzen. Deswegen wurden zur Betrachtung des differential Payoffs die Erwartungsnutzenwerte normalisiert. Dies geschah, indem sie durch den Erwartungsnutzen geteilt wurden, welcher sich unter der Annahme einer konstanten Nutzenfunktion von eins ergibt. Damit konnten die Erwartungsnutzenwerte auf die ursprüngliche Nutzenskala zurücktransformiert und als gewichteter Nutzen interpretiert werden. Sie bildeten ein Maß für die Entscheidungssicherheit, denn ihre Unterschiedlichkeit ging nur auf die Unterschiede in den Nutzenfunktionen der einzelnen Treatments zurück, nicht aber auf die Lage der Wahrscheinlichkeitsverteilungen der Kovariablen.

5.5.3 Maximin-Entscheidung

Das Maximin-Prinzip (im Falle einer Nutzenbetrachtung) oder Minimax-Prinzip (im Falle einer Verlustbetrachtung) trägt seinen Algorithmus schon im Namen. Es maximiert den minimalen Nutzen, beziehungsweise minimiert den maximalen Verlust. Im Gegensatz zur Erwartungsnutzen-Entscheidung basiert die Maximin-Entscheidung damit nicht auf Wahrscheinlichkeitsverteilungen. Es müssen also keine Annahmen über die Verteilungen der Umweltzustände getroffen werden. Das Maximin-Prinzip befasst sich damit direkt mit dem Problem der statistischen Schlussfolgerungen für endliche Stichproben, ohne Rückgriff auf die Annäherungen für große Stichproben der asymptotischen statistischen Theorie (Manski, 2004). Berger (1985) etwa schlägt daher das Maximin-Prinzip als Absicherung der bayesianischen Regeln vor. Das Maximin-Prinzip sichert gegen den größtmöglichen Schaden ab. Good (1952) etwa betrachtet eine Menge subjektiver Nutzenfunktionen und argumentiert, dass eine Minimax-Lösung vernünftig ist, wenn nur vernünftige subjektive Nutzenfunktionen und Umweltzustände in Betracht gezogen werden. Eine weitere Betrachtungsmöglichkeit des Maximin-Prinzips ist, es als Erwartungsnutzenmaximierung unter einer Wahrscheinlichkeitsverteilung zu sehen, unter der für jede Aktion der Umweltzustand mit dem geringsten Nutzen mit Wahrscheinlichkeit eins eintritt. Die Maximin-Entscheidungsregel lautet:

Man wähle dasjenige a^*, für das gilt:

$$\inf_{\theta \in \Theta} u(a^*, \theta) \geq \inf_{\theta \in \Theta} u(a, \theta) \text{ für alle } a \text{ in } \mathbb{A}$$

mit:

- $\mathbb{A}$ = Aktionenraum
- a^* = Maximin-Aktion
- $u(a, \theta)$ = Nutzen der Aktion a unter θ
- Θ = Raum der Umweltzustände

5.5.3.1 Bestimmung der Entscheidungsfunktion Zur Bestimmung der Maximin-Entscheidungsfunktion wurde die Schätzung der Personenparameter likelihoodbasiert durchgeführt. Die Likelihood wurde anschließend standardisiert, so dass sie am Maximum-Likelihood-Schätzer den Wert von eins annimmt (Cattaneo, 2013). Das primäre Ziel war es dann, für jeden Messwert den Bereich der möglichen Personenparameter auf einen Bereich der plausiblen Personenparameter einzugrenzen. Dafür wurde ein Cutoff von 0.5 der standardisierten Likelihood gewählt. So wurden nur solche Personenparameter berücksichtigt, deren Plausibilität mindestens halb so groß war, wie die des ML-Schätzers. An dieser Stelle sind grundsätzlich natürlich auch andere Cutoffs denkbar. Der Wert 0.5 wurde mangels besserer Erfahrungswerte als vernünftig eingeschätzt. Es wurde also für jede Person der Bereich, in welchem die Nutzenfunktion ausgewertet werden sollte, vom ursprünglichen Parameterbereich auf einen plausiblen Parameterbereich verkleinert. Mathematisch wurde also Folgendes durchgeführt:

$$\Theta \rightarrow \Theta^* \; mit \; \Theta^* \subseteq \Theta \text{ und } \Theta^* = f(\Theta, X_i)$$

mit:

- Θ = Parameterraum der Lesetests
- Θ^* = Plausibler Parameterraum für Person i mit Kovariablenvektor X_i
- X_i = Kovariablenvektor, der sich aus Antwortmustern der beiden Lesetests ergibt

Anschließend wurden in dem sich so ergebenden Bereich die Nutzenfunktionen beider Trainings ausgewertet und jeweils der minimale Nutzen der Trainings bestimmt. Im letzten Schritt wurden die beiden minimalen Nutzen verglichen und die Treatmententscheidung zugunsten des Treatments mit dem höchsten minimalen Nutzen gefällt.

5.5.3.2 Bestimmung der Entscheidungssicherheit Auch bei der Maximin-Entscheidung konnte die Entscheidungssicherheit am differential Payoff festgemacht werden. Da bei der Maximin-

Entscheidung die Nutzenfunktion nicht transformiert wird, behält sie ihre Einheit und die Differenz der beiden minimalen Nutzen (differential Payoff) drückt den Payoff im ungünstigsten Fall aus. Zu beachten ist: der differential Payoff drückt hingegen nicht den garantierten Payoff aus. Natürlich ist es möglich, dass sich die Bereiche der plausiblen Nutzenwerte der Treatments überlappen und es so keinen garantiert positiven differential Payoff für eines der beiden Treatments gibt.

5.6 Ergebnisse

Im Folgenden werden die Ergebnisse der Anwendung der Erwartungsnutzen-Entscheidung und der Maximin-Entscheidung dargestellt. Hierfür wurden auf einem Gitter von Personenparametern zwischen -2.5 und 2.5 die Entscheidungsfunktionen für die Leseflüssigkeit und das Leseverstehen angewandt. Anschließend wurden die Ergebnisse grafisch dargestellt.

5.6.1 Entscheidungen für den Leseflüssigkeitszuwachs unter Erwartungsnutzenmaximierung

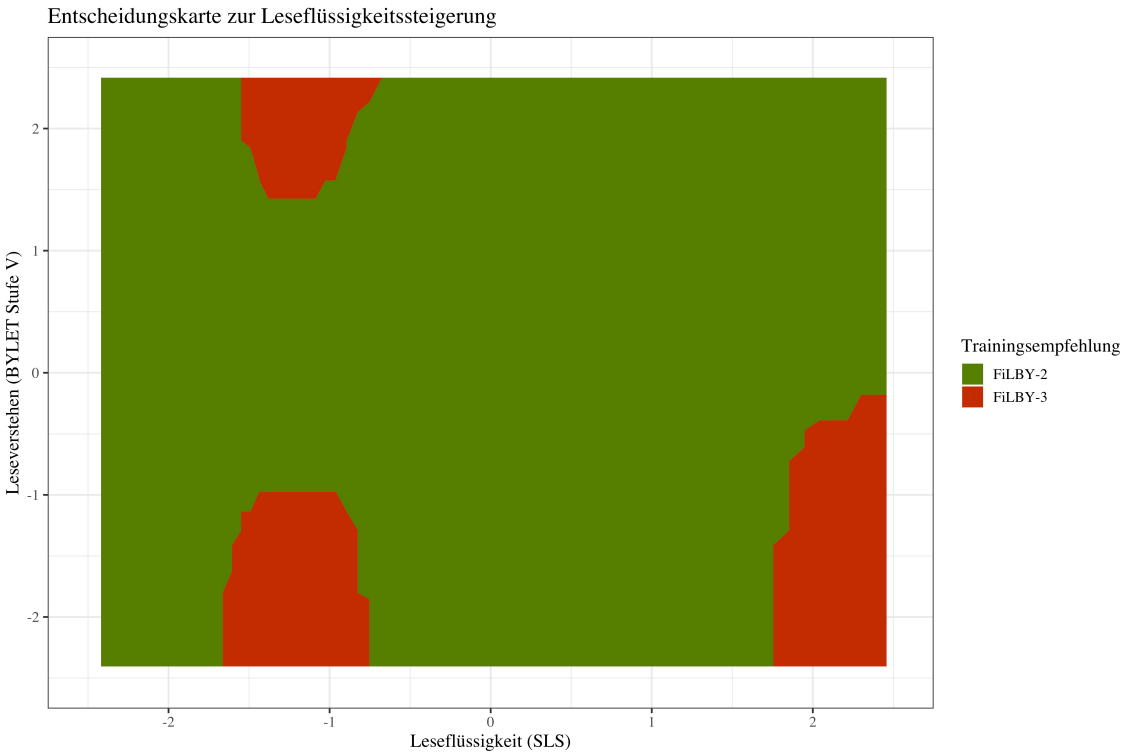


Abbildung 31: Trainingsempfehlung zur Leseflüssigkeitssteigerung

Bildet der Zuwachs der Leseflüssigkeit das Zielkriterium, so erkennt man in Abbildung 31, dass in fast allen Leistungsbereichen das FiLBY-2-Training empfohlen wurde. Lediglich bei sehr stark ausgeprägter Leseflüssigkeit, bei gleichzeitig schwach ausgeprägtem Leseverstehen wurde das FiLBY-

3-Training empfohlen. Dieser Befund ist inhaltlich sehr plausibel. Grundsätzlich gilt, dass eine adäquate Leseflüssigkeit die Voraussetzung für eine gute Leseverstehenskompetenz ist. Gleichzeitig begünstigt aber ein gutes Leseverstehen auch eine höhere Leseflüssigkeit. Wird beim Lesen ein Situationsmodell entwickelt, so entsteht eine Erwartung an die nächsten Wörter, grammatikalische Strukturen und Satzverläufe. Diese Erwartungen erleichtert eine schnelle Verarbeitung der neuen Textpassagen (Dehaene, 2014). Zusätzlich wird das FiLBY-3-Training in zwei kleineren Bereichen empfohlen, die bei eher geringer Leseflüssigkeit und extrem ausgeprägtem Leseverstehen liegen. Dies erscheint inhaltlich nicht besonders plausibel, da das FiLBY-3-Training das anspruchsvollere Training darstellt. Möglich ist jedoch, dass es sich beim Bereich, in welchem beide Kompetenzen schwach ausgeprägt waren, um einen Aufholeffekt handelt. Dies ist möglich, da die beiden Trainings nicht in einem experimentellen Design, sondern nacheinander durchgeführt wurden. Es wäre also denkbar, dass Kinder, die in der zweiten Jahrgangsstufe während des FiLBY-2-Trainings nur wenige Fortschritte erzielten, schließlich in der dritten Jahrgangsstufe einen Entwicklungsschub erlebten. Dies ist im Sinne eines Reifeprozess nicht unplausibel, aber jedoch vermutlich unabhängig vom im Unterricht durchgeführten Lesetraining. Auch denkbar ist, dass es sich um Zufallseffekte handelt.

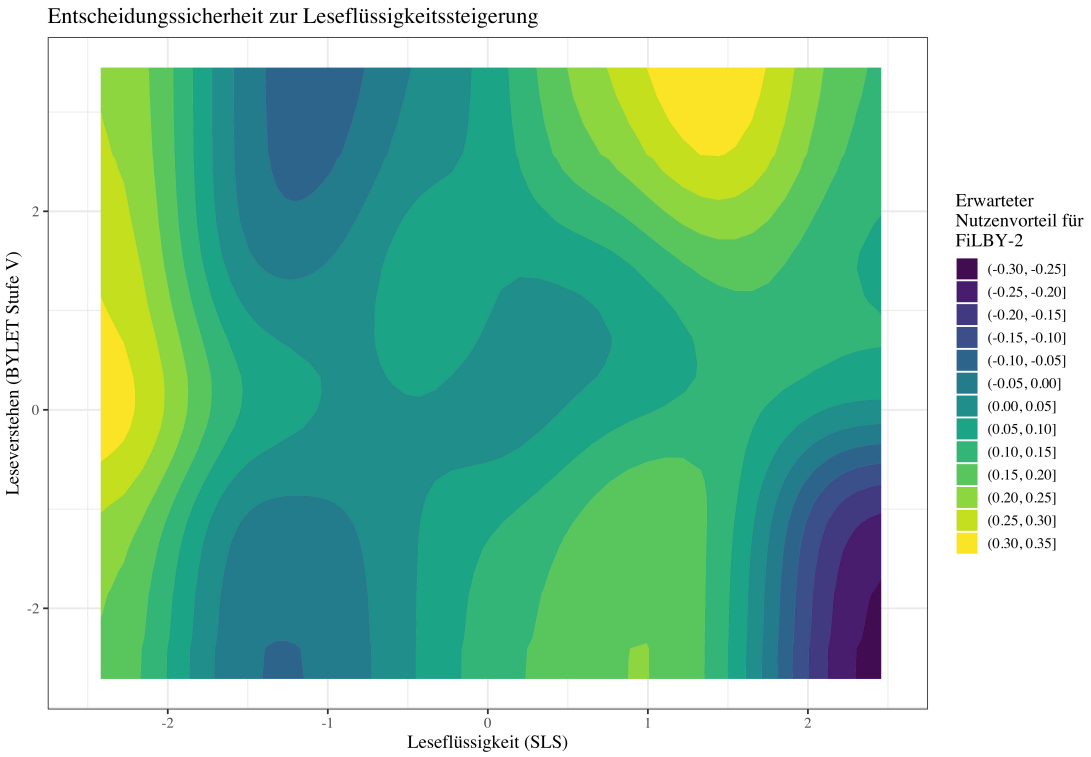


Abbildung 32: Trainingsempfehlungssicherheit zur Leseflüssigkeitssteigerung

Betrachtet man nicht nur die dichotome Entscheidungskarte, sondern zusätzlich die Entscheidungssicherheitskarte (vgl. Abbildung 32), so sieht man, dass die zwei kleinen Bereiche, in denen FiLBY-3 empfohlen wird, dies nur mit sehr geringer Sicherheit tun. Die Entscheidungssicherheitskarte zeigt

an, wie groß die erwarteten Unterschiede im Leseflüssigkeitszuwachs zwischen dem FiLBY-2- und dem FiLBY-3-Training sind. Die Skala befindet sich auf der Einheit der Nutzenfunktion, also in z-standardisierten Werten. Positive Werte bedeuten einen Vorteil für das FiLBY-2-Training, negative einen Vorteil für das FiLBY-3-Training. Liegen die Werte bei 0 oder nahe an der 0, so sind in diesem Leistungsbereich beide Trainings gleichermaßen wirksam und die gefundenen Wirksamkeitsunterschiede mit größerer Sicherheit zufällig entstanden.

5.6.2 Entscheidungen für den Leseverstehenszuwachs unter Erwartungsnutzenmaximierung

Es zeigte sich, dass sich das Leseverstehen sowohl mit dem Leseflüssigkeitstraining als auch mit dem Lesestrategietraining verbessern ließ (s. Abbildung 33). Dabei schien es jedoch so zu sein, dass erst ab einem gewissen Level an Leseverstehen das Lesestrategietraining zu voller Effektivität kommt. Für Kinder deren Leseverstehen nur sehr schwach ausgeprägt ist, erscheint es daher sinnvoller zunächst weiter die Leseflüssigkeit zu trainieren und damit das Leseverstehen zu verbessern.

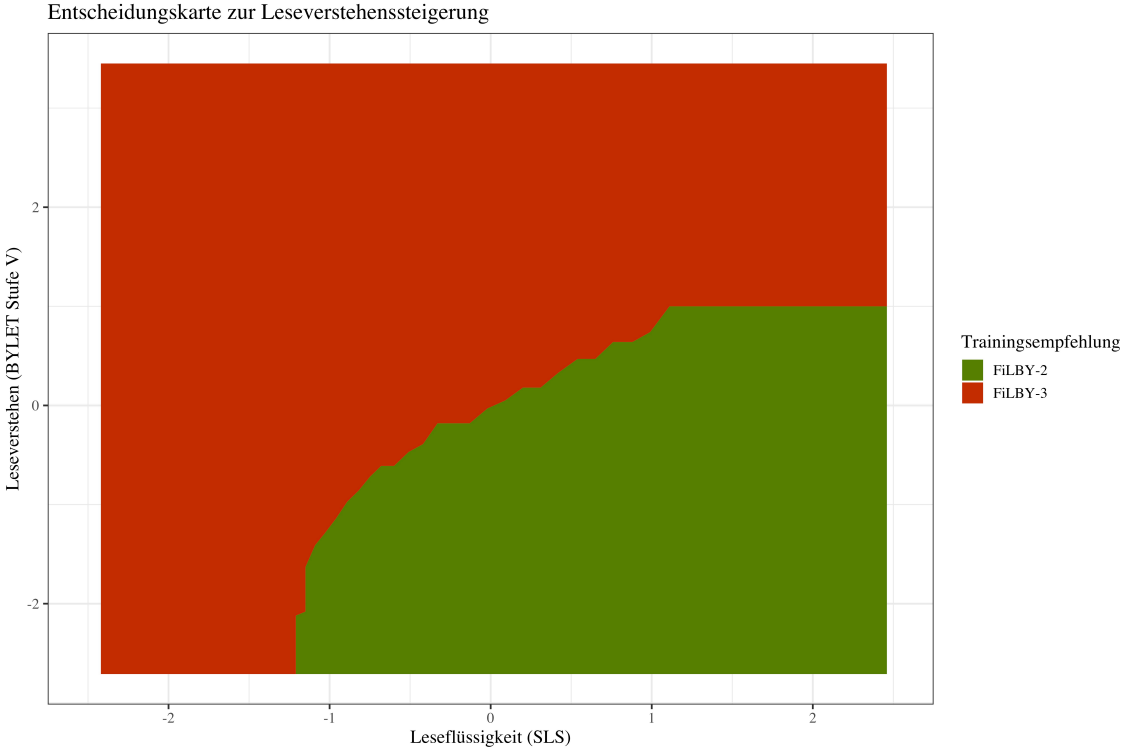


Abbildung 33: Trainingsempfehlung zur Leseverstehenssteigerung

Dieser Befund passt zu bisherigen Forschungsergebnissen, welche die Leseflüssigkeit als einen Art Flaschenhals zum Leseverstehen identifizieren (Chall, 1983; Pikulski & Chard, 2005). Dennoch zeigte sich auch hier wieder, dass für Kinder mit Defiziten in Leseflüssigkeit und Leseverstehen das

FiLBY-3-Training empfohlen wird. Auch hier könnte es sich wieder, ähnlich wie beim Leseflüssigkeitszielkriterium, um einen Aufholeffekt gehandelt haben.

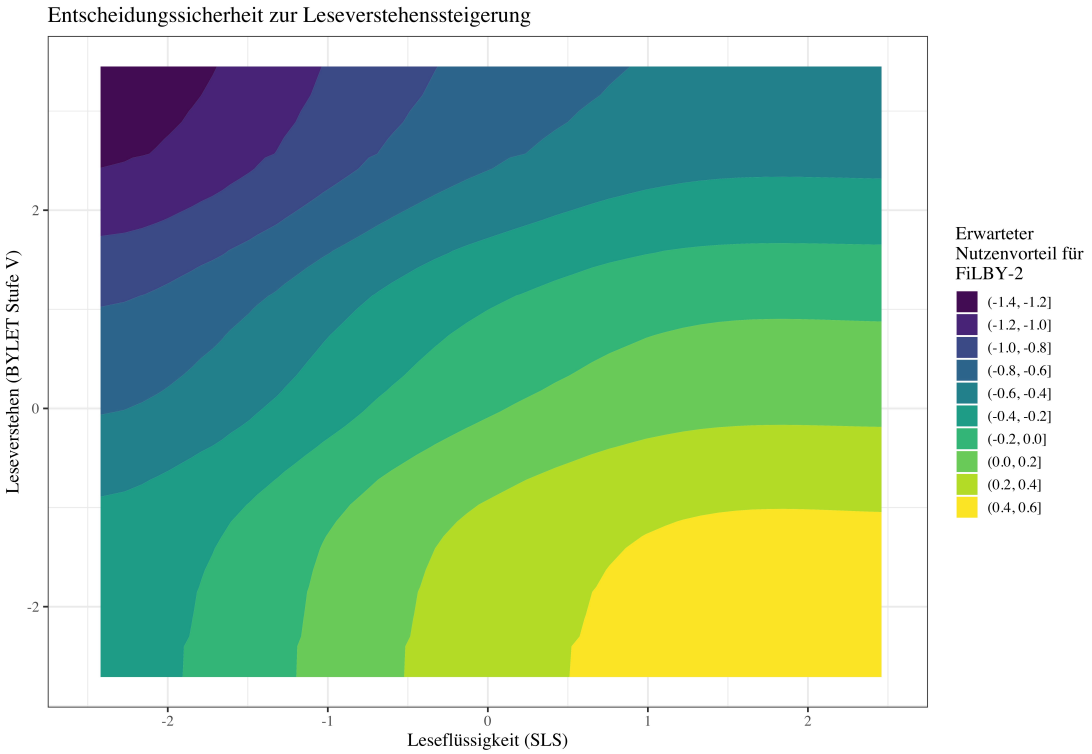


Abbildung 34: Trainingsempfehlungssicherheit zur Leseverstehenssteigerung

Ein ähnliches Bild zeigte sich auch bei der Auswertung der Entscheidungssicherheit zur Steigerung des Leseverstehens (vgl. Abbildung 34). Hier wird Kindern, die bereits sehr flüssig lesen mit hoher Sicherheit ein weiteres Training mit dem FiLBY-2-Training empfohlen, während Kindern mit hohem Leseverständnis das FiLBY-3-Training empfohlen wird. Dies deutet darauf hin, dass beide Trainings tendenziell die starken Kinder weiter fördern, aber zu wenig auf die Bedürfnisse der schwachen Schülerinnen und Schüler zugeschnitten sind.

5.6.3 Entscheidungen für den Leseflüssigkeitszuwachs unter Maximin

Grundsätzlich zeigt sich unter der Maximin-Entscheidung ein ähnliches Bild wie bei der Erwartungsnutzen-Entscheidung. Die etwas gröbere Auflösung und weniger glatte Natur der Funktion ergibt sich dadurch, dass in der Maximin-Entscheidung Extremwerte betrachtet wurden. Wie Abbildung 35 darstellt, zeigte sich wieder die FiLBY-3-Trainingsempfehlung für die schnell Lesenden, sowie für diejenigen Kinder, deren Leseflüssigkeit und Leseverstehen schwach ausgeprägt waren. Für langsame Leser:innen, die über ein gutes Leseverstehen verfügten, wurde nach der Maximin-Regel jedoch durchgängig das FiLBY-2-Training empfohlen.

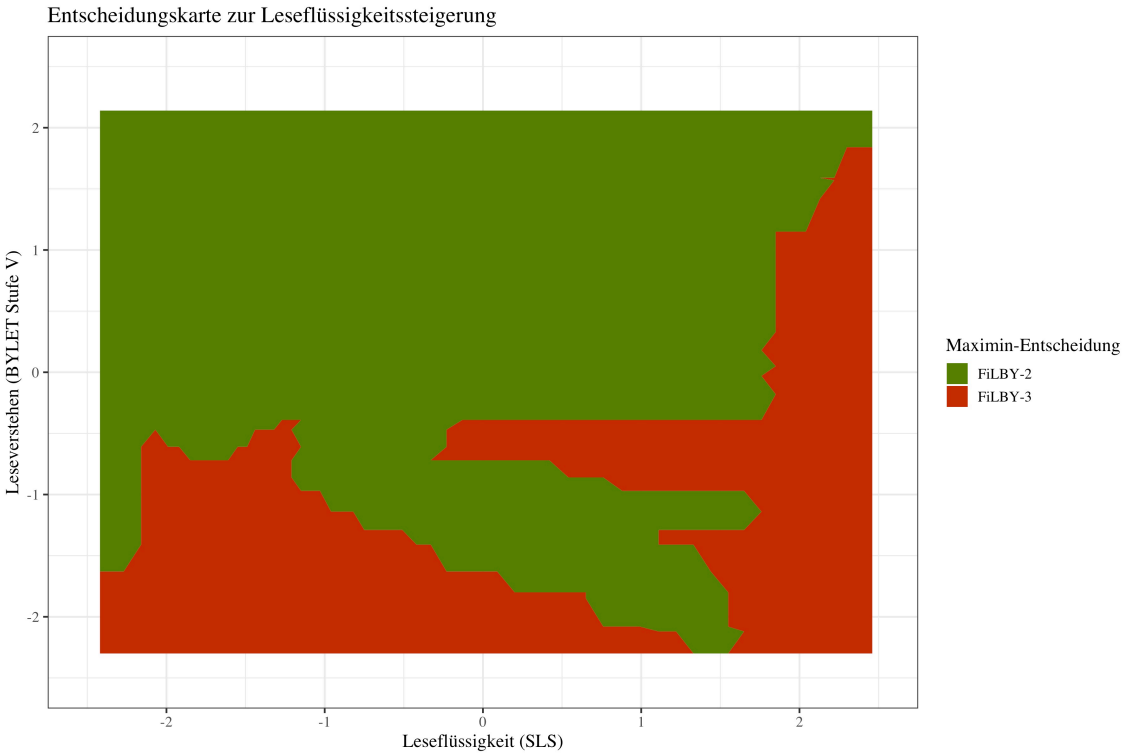


Abbildung 35: Trainingsempfehlung zur Leseflüssigkeitssteigerung

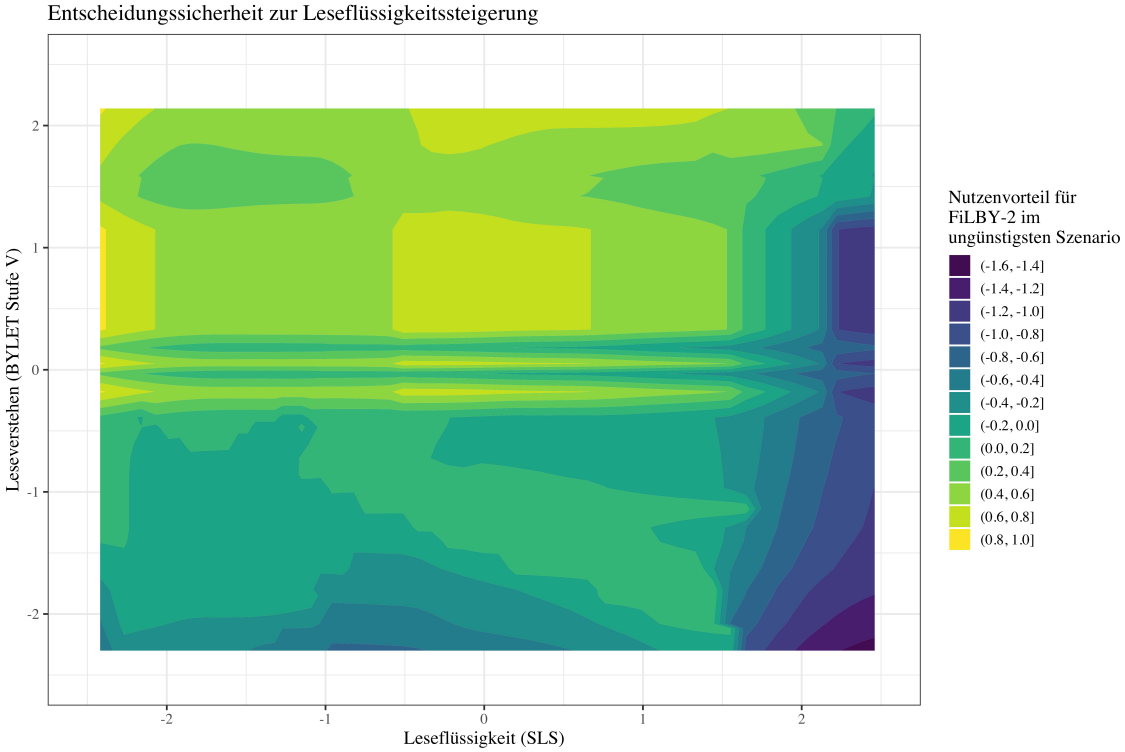


Abbildung 36: Trainingsempfehlungssicherheit zur Leseflüssigkeitssteigerung

Dass für Kinder mit mittlerer Leseflüssigkeit und eher schwach ausgeprägtem Leseverstehen an manchen Stellen, das FiLBY-2-, an anderen Stellen jedoch das FiLBY-3-Training empfohlen wurde, erscheint nicht besonders plausibel und ist mit Sicherheit auf die Betrachtung des ungünstigsten Falls der Maximin-Regel zurückzuführen. Hier erhalten einzelne Extremstellen der Nutzenfunktion eine maximale Gewichtung in der Entscheidung, wodurch Entscheidungsregeln eine weniger glatte Gestalt annehmen. Bei Betrachtung der Entscheidungssicherheit (vgl. Abbildung 36), zeigte sich jedoch wieder, dass gerade für die inhaltlich nicht gut erklärbaren Entscheidungsbereiche, die Entscheidungssicherheit sehr gering war. Darüber hinaus lag unter der Maximin-Regel überwiegend in den Randbereichen eine hohe Entscheidungssicherheit vor, während sich in den Leistungsbereichen mit der höchsten Beobachtungsdichte, beide Trainings als fast gleichwertig zeigten.

5.6.4 Entscheidungen für den Leseverstehenszuwachs unter Maximin

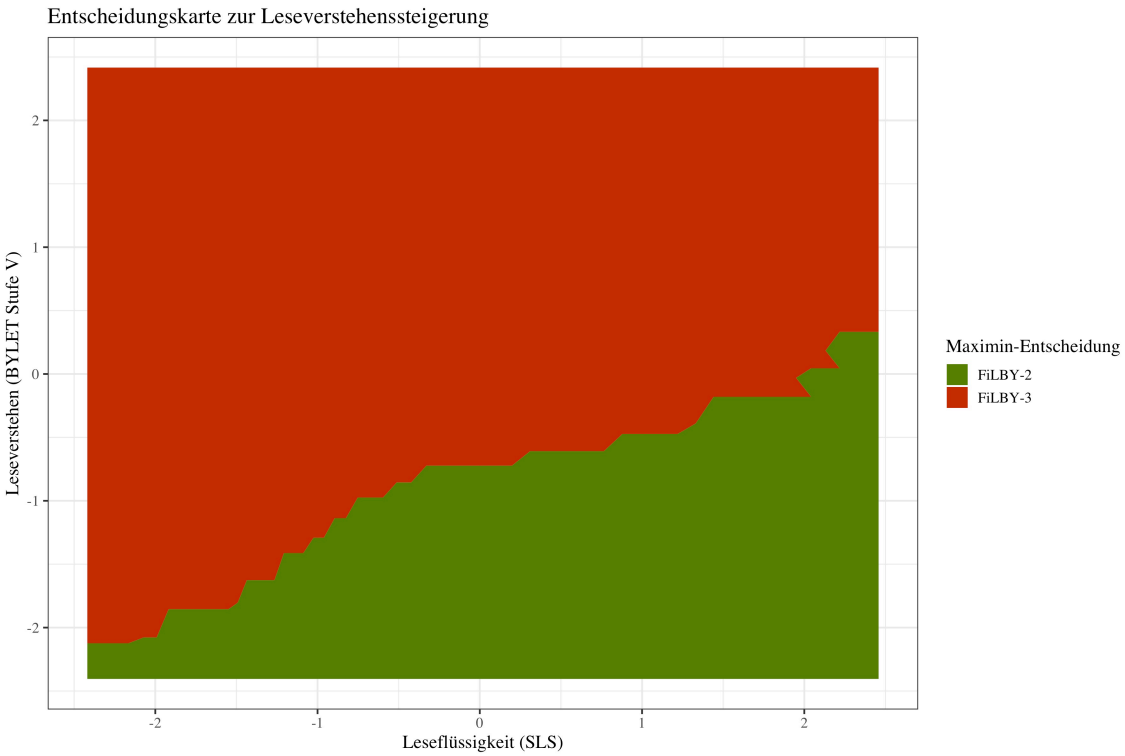


Abbildung 37: Trainingsempfehlung zur Leseverstehenssteigerung

Auch unter Anwendung der Maximin-Regel zeigte sich für die Entscheidung zum Leseverstehenszuwachs eine Schwelle des Ausgangsleseverstehens, welches eine höhere Effizienz des Lesestrategietrainings bedingte (s. Abbildung 37). Interessant ist jedoch, dass unter der Maximin-Entscheidung diese niedriger zu liegen scheint, als bei der Erwartungsnutzen-Entscheidung. Nimmt der:die Entscheidungsträger:in also eine konservative Haltung ein, so sollte schon bei geringer ausgeprägtem

Leseverstehen mit dem FiLBY-3-Training trainiert werden. Zusätzlich fällt auf, dass für Kinder mit extrem schwach ausgeprägtem Leseverstehen und extrem schwach ausgeprägter Leseflüssigkeit, erwartungskonform das FiLBY-2-Training empfohlen wird.

Die Entscheidungssicherheit im ungünstigsten Fall zeichnet nur an den Rändern ein eindeutiges Bild (vgl. Abbildung 38). In weiten Teilen des Parameterraums schneiden im ungünstigsten Fall beide Trainings gleich gut ab. Es bot sich hier also im Wesentlichen keine sehr eindeutige Entscheidungslage.

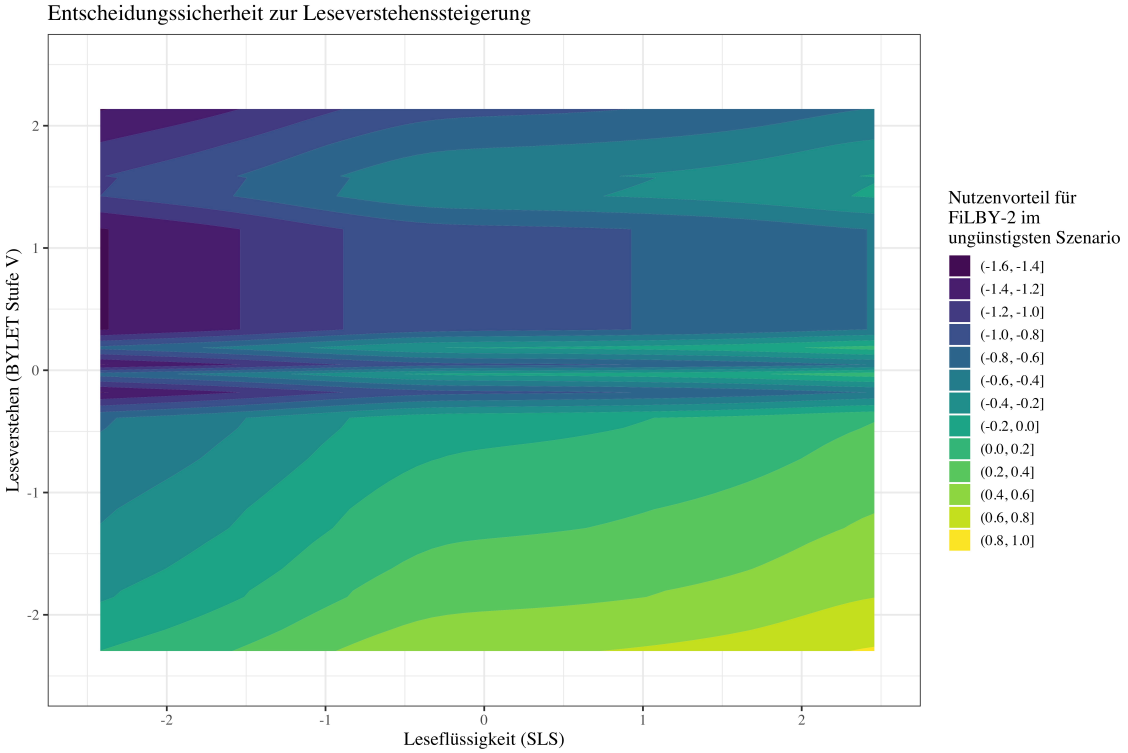


Abbildung 38: Trainingsempfehlungssicherheit zur Leseverstehenssteigerung

5.6.5 Unterschiede in den Entscheidungsfunktionen

Legt man die Entscheidungskarten der Maximin- und der Erwartungswertentscheidung übereinander, so erkennt man, dass beide Entscheidungsregeln für beide Zielkriterien in weiten Teilen des betrachteten Parameterraums zu denselben Empfehlungen gelangen. Man erkennt zudem in Abbildung 39, dass die Erwartungsnutzen-Entscheidung das FiLBY-3-Training für einen kleineren Kompetenzbereich empfiehlt, als die Maximin-Entscheidung.

Beim Leseverstehen zeigt Abbildung 40 ein ähnliches Bild. Allerdings empfiehlt hier die Maximin-Entscheidung das FiLBY-2-Training in einem weiteren Bereich. Es scheint sich also als generelles Muster Fogendes abzuzeichnen. Die konservativere Entscheidungshaltung präferiert das aus theo-

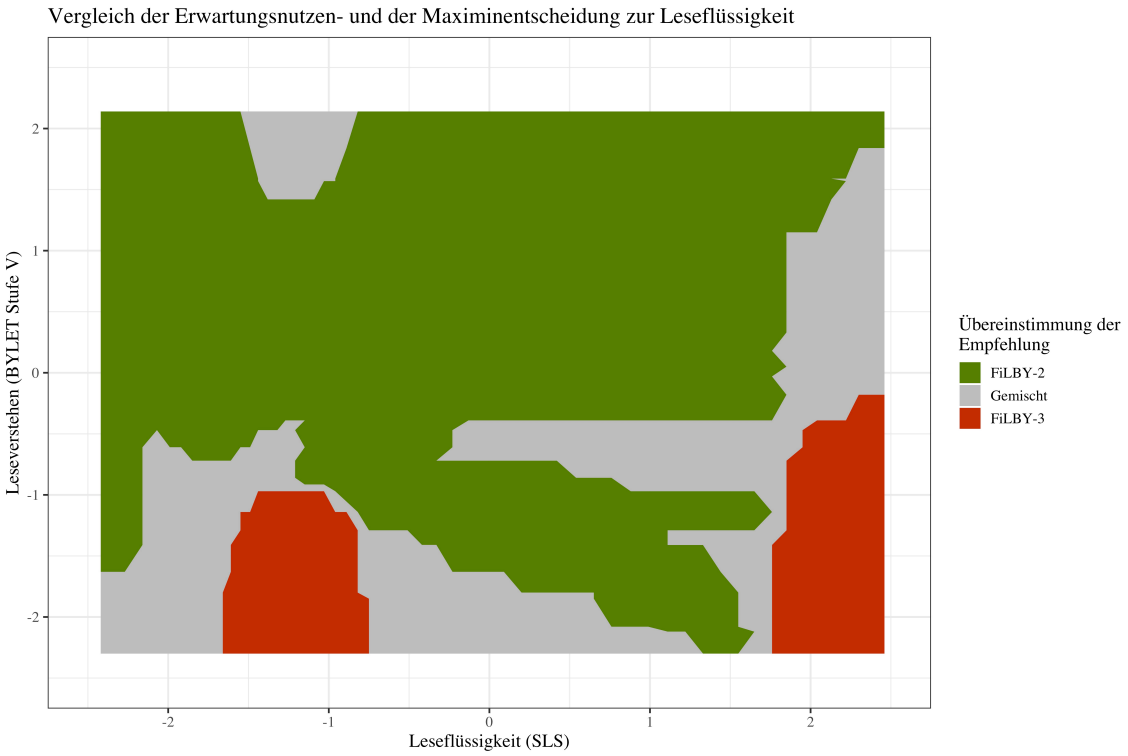


Abbildung 39: Vergleich der Entscheidungsfunktionen für die Leseflüssigkeit

retischer Sicht eher ungeeignetere Training, während sich aus der Erwartungsnutzen-Entscheidung eher theoriekonforme Entscheidungsgrenzen ableiten.

Als Fazit bleibt festzuhalten, dass unter zusätzlicher Berücksichtigung der Entscheidungssicherheiten, sich für pessimistisch eingestellte Entscheidungsträger:innen beide Trainings gleichermaßen als empfehlenswert erwiesen. Für optimistische Entscheidungsträger:innen, die eine Erwartungsnutzen-Entscheidung bevorzugen, ergab sich hingegen durchaus eine zuverlässigere Trainingsempfehlung. Darüber hinaus zeigte sich eine Robustheit in weiten Bereichen der Lesekompetenz, in denen beide Entscheidungsfunktionen zur selben Empfehlung gelangen.

5.7 Diskussion

In der folgenden Diskussion werden zunächst die Ergebnisse zusammengefasst und anschließend auf inhaltlicher Ebene und mit Bezug zur Entscheidungstheorie diskutiert.

5.7.1 Zusammenfassung der Ergebnisse

Gerade im Vergleich der Erwartungsnutzenmaximierung mit der Maximin-Entscheidung zeigte sich für beide Entscheidungsfunktionen ein ähnliches Bild. Ist die Leseflüssigkeitssteigerung das Trai-

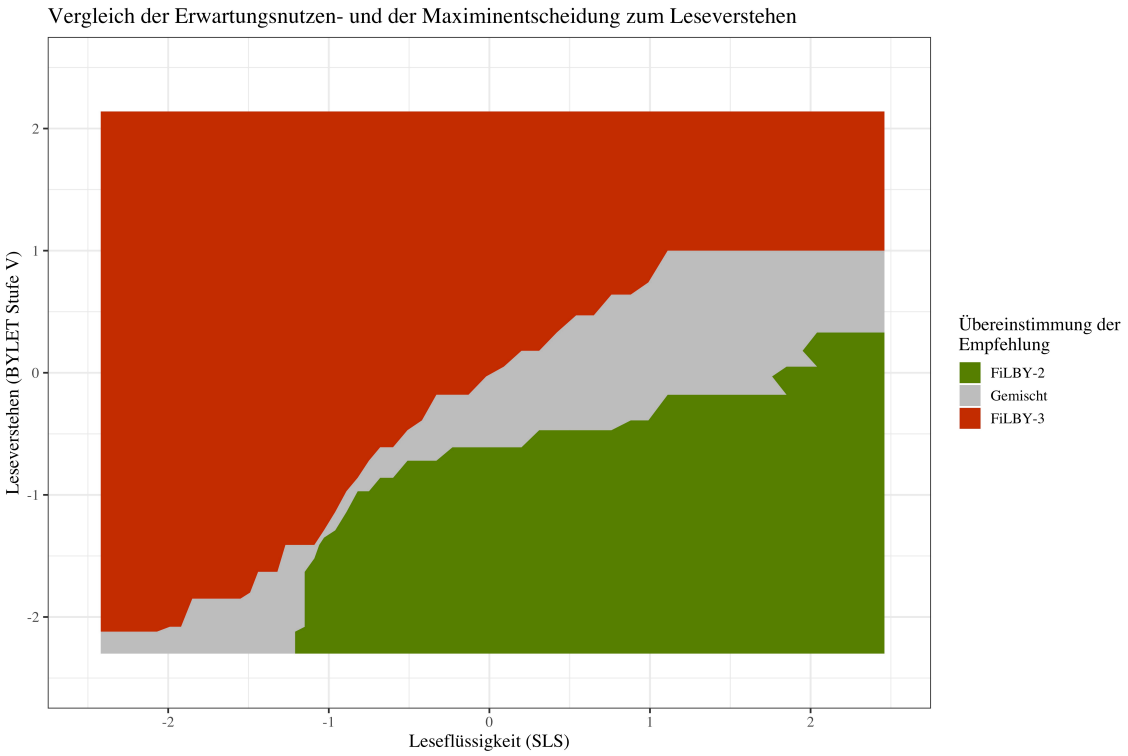


Abbildung 40: Vergleich der Entscheidungsfunktionen für das Leseverstehen

ningsziel, so bietet sich für fast alle Lesenden das FiLBY-2-Training an. Eine Ausnahme bilden hier sehr flüssig Lesende, für die das FiLBY-3-Training wirksamer ist und sehr schlecht Lesende, für die beide Trainings wenig wirksam waren. Ist jedoch das Ziel, das Leseverstehen zu steigern, so muss ein gewisses Grundlevel an Leseverstehen bereits gegeben sein, damit das FiLBY-3-Training wirksamer ist, als das FiLBY-2-Training. Dies stellt ein inhaltlich sehr plausibles Ergebnis dar. Ist nämlich das Leseverstehen noch sehr schwach ausgeprägt, so sollte nach gängigen Empfehlungen (Chall, 1983; Pikulski & Chard, 2005), erst die Leseflüssigkeit trainiert werden, da diese oft eine Flaschenhals-Funktion im Leselernprozess einnimmt. Lesestrategietrainings wie das FiLBY-3-Training setzen eine Orientierung im Text voraus. Es müssen etwa Schlüsselwörter identifiziert werden und Textabschnitte den verschiedenen erzählerischen Elementen wie Figur, Situation und Ereignis zugeordnet werden. Ist das Leseverstehen dabei noch schwach ausgeprägt und muss viel kognitive Aktivität auf das Entschlüsseln der einzelnen Wörter verwendet werden, so bleibt zu wenig kognitive Kapazität für die die anspruchsvolle Inferenzleistung des Leseverstehens übrig. Das Leseverstehen misslingt (Sweller, 2011).

5.7.2 Diskussion aus entscheidungstheoretischer Perspektive

Aus entscheidungstheoretischer Perspektive spricht die geringe Unterschiedlichkeit der beiden Entscheidungsfunktionen für eine Robustheit der Ergebnisse. Zwar treffen die Erwartungsnutzenmaximierung und die Maximin-Entscheidung sehr unterschiedliche Vorannahmen über die Beschaffenheit der Umweltzustände, dennoch kommen die Analysen zu ähnlichen Entscheidungen. Dabei zeigte sich jedoch, dass die Entscheidungssicherheit für die Maximin-Entscheidung nicht so hoch war, wie für die Erwartungsnutzen-Entscheidung. Wird also wenig Vertrauen in die Messung der Ausgangsleistungen gelegt und die Maximin-Regel gewählt, so sinkt die Entscheidungssicherheit bezüglich der gewählten Trainings. Auch dies ist ein plausibles Ergebnis, wenn man bedenkt, dass die Nutzenfunktion ja in Abhängigkeit der Ausgangsleistungen geschätzt wurde. So setzt sich die Unsicherheit über die Prämissen im Entscheidungsproblem direkt in der Unsicherheit in der Treatmententscheidung fort.

Der Einbezug von Unsicherheit auf Ebene der Messung der Kovariablen zeigt, was Cronbach & Gleser (1965) bereits in den 60er Jahren erkannten. In der Treatmententscheidung kommt es nicht so stark auf die Reliabilität der Messung an, wenn die Nutzenfunktionen der Treatmentoptionen nur unterschiedlich genug sind. Auch in der Betrachtung der Entscheidungskarten in dieser Studie sieht man, dass in der Anwendung für spezifische Entscheidungen nur an den Schnittstellen der Nutzenfunktionen eine ausreichende Reliabilität benötigt wird. Wie groß dieser Bereich ist, bestimmt die Validität, beziehungsweise der differential Payoff der Nutzenfunktionen. Sind die Winkel an den Schnittstellen der Nutzenfunktionen klein, und die funktionale Form überwiegend linear so benötigt man in einem größeren Kovariablenbereich eine hohe Reliabilität. Denn dann liegen die beiden Nutzenfunktionen nah beieinander und die an der Messungenauigkeit gewichteten prädiktiven Verteilungen verfügen über sehr ähnliche Erwartungswerte. Ist jedoch der differential Payoff groß, so sind auch die Erwartungswerte der gewichteten prädiktiven Verteilungen unterschiedlich und die Entscheidungssicherheit kann auch bei geringerer Reliabilität, also größerer Varianz der prädiktiven Verteilungen gewährleistet werden.

Damit ergeben sich für die Konstruktion neuer für die Treatmententscheidung konstruierte Messinstrumente neue Anforderungen. Es sollten Messinstrumente konstruiert werden, welche in ganz spezifischen Leistungsbereichen über eine sehr hohe Reliabilität verfügen, idealerweise also adaptive Messinstrumente.

Auch in der konkreten Anwendung des TreaDeMs liefern die Kennwerte der Entscheidungssicherheit wichtige Informationen. Zum einen kann eine hohe Entscheidungsunsicherheit als Hinweis auf den Bedarf für weitere diagnostische Untersuchungen verstanden werden. Dann sollte im Sinne des

diagnostischen Prozesses, weitere Diagnostik zur Entscheidungsfindung eingesetzt werden. Möglicherweise ist eine weitere Diagnostik jedoch nicht möglich, oder nicht erwünscht. Dann könnte eine weitere Implikation einer hohen Entscheidungsunsicherheit sein, dass man das ausgewählte Treatment früher als üblich evaluiert, um so etwaige Fehlentscheidungen schneller korrigieren zu können. Es kann also nicht nur die Treatmententscheidung selbst, sondern auch die aus dem Modell gewonnene Unsicherheit der Treatmententscheidung praktisch genutzt werden.

5.7.3 Ausblick

Neben dem Vergleich verschiedener Entscheidungsfunktionen, gibt es weitere Vorannahmen, die Einfluss auf die finale Entscheidungsfunktion nehmen. Verfolgt man das Ziel, ein robustes und transparentes Treatmententscheidungsmodell zu formulieren, so sollten alle Vorannahmen, die an irgendeiner Stelle der Modellierung getroffen werden expliziert werden. So wird die Robustheit überprüfbar. Denn wenn mit (fast) beliebiger Vorannahme, beliebiger Nutzenfunktion und beliebiger Entscheidungsregel immer dieselbe oder eine sehr ähnliche Entscheidungsfunktion ermittelt wird, dann ist dies ein starker Indikator für die externe Validität der Entscheidungsfunktion. Variiert die Entscheidung jedoch stark in Abhängigkeit der Vorannahmen, so sollte das Modell weiter verbessert werden. Dabei sollten das Finden geeigneter Kovariablen in der Nutzenfunktion und eine spezifisch reliable Messung dieser im Fokus stehen.

In diesem Zusammenhang wäre es interessant in zukünftigen Studien zu untersuchen, welche Reliabilität man mindestens braucht, um noch vernünftige Entscheidungsfunktionen zu erhalten. Dafür könnte man aus den gewichteten prädiktiven Vorhersageverteilungen die Fehler erster und zweiter Art bestimmen, für diese ein Maximalniveau festlegen und über dieses auf eine Mindestreliabilität rückschließen. Gerade im Kontext von Kurzskalen und computergesteuerten Lernprogrammen ist dieser Ansatz interessant, da hier mit nur wenigen Items versucht wird, eine:n Lernende:n effizient durch ein Programm zu führen, welches im Grunde als eine Folge von kleinen Treatments betrachtet werden kann. Deren Auswahl könnte so evidenzbasiert erfolgen. In Bezug auf viele aktuelle Lernprogramme, wäre das eine Neuerung, denn hier ist die Treatmentauswahl in der Regel hardgecoded und auf Experteneinschätzung beruhend implementiert (vgl. Tzouveli, Mylonas, & Kollias, 2008; Vandewaetere & Clarebout, 2014).

Zudem sollte man in der Zukunft ein Treatmententscheidungsmodell entwickeln, welches alle Unsicherheiten - auch die der Nutzenfunktion - berücksichtigt. So könnte man Zielvorgaben erarbeiten, wie viel Varianz in der Nutzenfunktion aufgeklärt werden sollte, ob man bei Beobachtungsstudien überhaupt eine Aussage machen kann und mit welcher Reliabilität eine Kovariable gemessen wer-

den muss (Cronbach & Gleser, 1965), damit man im Anschluss für möglichst viele Proband:innen Trainingsempfehlungen aussprechen kann, die über Raten hinausgehen.

6 Generaldiskussion

Abschließend soll nun die Arbeit als Ganzes diskutiert werden. Dabei soll der Blick geweitet werden, das TreaDeM als evidenzbasierte Anwendung in der Praxis in den Blick genommen werden und auch besprochen werden, was die entscheidungstheoretische Denk- und Herangehensweise für das Feld der Diagnostik und der Bildungswissenschaften leisten und nicht leisten kann.

6.1 Beitrag (falscher) Modelle zu Entscheidungen

"Alle Modelle sind falsch, aber manche Modelle sind hilfreich" lautet ein viel zitierter Ausspruch von George Box, einem renommierten britischen Statistikprofessor. Dies gilt auch für Entscheidungsmodelle und auch für das in dieser Arbeit entwickelte Treatmententscheidungsmodell. Doch was bedeutet "falsch" sein? Es bedeutet in der Regel, dass Modelle nicht perfekt sind. Die Realität ist multikausal, multidimensional und komplex. Ein Modell berücksichtigt immer nur eine endliche Anzahl an Variablen und wird auf einer endlichen Stichprobe mit endlicher Rechenleistung optimiert. Dennoch kann man nicht sagen, dass Modelle "falsch" in einem tieferen philosophischen Verständnis sind. Gewiss kann auch ein nicht optimal passendes Treatmententscheidungsmodell zumindest nach der Datenlage offensichtlich schlechte Entscheidungen abwenden. Ein zweiter Ansatz, der zum Beispiel von Gilboa und Schmeidler (1995) verfolgt wird, ist es, Modelle als eine Art Fallstudie zu verstehen, also als eine kompakte Beschreibung von spezifischer Vergangenheit. Ob ein Modell daher für zukünftige Entscheidungen hilfreich ist, hängt auch maßgeblich davon ab, wie vergleichbar, wie ähnlich die vergangene einer zukünftigen Situation ist.

Ein schönes Beispiel ist ein Münzwurf mit einer gezinkten Münze. Kennt man die letzten 100 Würfe und hat darunter neunzigmal Kopf beobachtet, so erscheint die Wahrscheinlichkeit für einen Kopfwurf als hoch. Demgegenüber ist etwa die Vorhersage, ob eine Person eine rote Ampel übertreten wird nur auf Basis der vergangenen Übertritte bei roter und grüner Ampel zu kurz gegriffen. Intuitiv ist klar, dass die Entscheidung etwa von der Anwesenheit von Polizist:innen oder von der Verkehrsdichte abhängig sein wird. Während also zukünftige Münzwürfe vergangenen Münzwürfen sehr ähnlich sind, ist das erneute Überqueren einer Ampel möglicherweise sehr unähnlich, wenn etwa die Gegend, die Anwesenheit anderer Personen etc. variiert werden. Modelle können also, wenn alle ihre Vorannahmen und Entstehungsbedingungen klar kommuniziert wurden als kollektiver Erfahrungswert verstanden werden. Gleichzeitig gilt, dass Modelle auch Robustheit in den aus ihnen gezogenen Schlüssen zeigen können. Wenn Modelle aus unterschiedlichen Stichproben, Zeiten und mit unterschiedlichen Messinstrumenten zu der selben finalen Entscheidung gelangen, dann gibt

es wohl eine eindeutige Befundlage. In der eben beschriebenen Sichtweise der Fallstudien gibt es dann für fast jede erdenkliche zukünftige Situation eine kompakte Beschreibung von spezifischer Vergangenheit, die zu übereinstimmenden Schlussfolgerungen kommt und dennoch nie exakt der aktuellen oder zukünftigen (Entscheidungs-)Situation entspricht.

6.2 Beitrag statistischer Verfahren der Entscheidungstheorie in der Praxis

In der diagnostischen Praxis gibt es bisher kaum formalisierte Entscheidungsprozesse. Gleichzeitig finden Studien fast übereinstimmend, dass Entscheidungen, die auf statistischen Modellen beruhen, im Schnitt besser sind als "klinische Urteile" (z. B. Grove et al., 2000; Ægisdóttir et al., 2006) und diagnostische Einschätzungen von Lehrkräften stark von standardisierten diagnostischen Ergebnissen abweichen (Anmarkrud & Bråten, 2012; Bremerich-Vos et al., 2017; Österbauer et al., 2020). Dabei kommt es natürlich auch darauf an, wie die Entscheidungsgüte definiert und gemessen wird. Grundsätzlich gilt aber, dass Entscheidungsmodelle die menschlichen Entscheidungen zumindest unterstützen und kontrastieren können und so auch Reflexion über das eigene Entscheiden anregen (z. B. Collins, Murphy, & Bierman, 2004). In einer eigenen Untersuchung zur Auswahl von Lesefördermaßnahmen bei Grundschullehrkräften etwa zeigte sich, dass sich keine Zusammenhänge zwischen der Wahl der Lesefördermaßnahmen und den Leistungs- oder demografischen Merkmalen der unterrichteten Kinder feststellen ließen. Dafür wählten Lehrkräfte fast immer dieselben oder sehr ähnliche Maßnahmen, wie ihre Koleg:innen (Kraus, Wild, Schilcher, & Hilbert, 2020). Eine Explizierung dieses Entscheidungsmodells könnte Lehrkräfte dazu anregen, auch auf subjektiv wahrgenommene Charakteristika ihrer Kinder einzugehen oder Misskonzepte über die Passung von Lesefördermaßnahmen und Leistungsmerkmalen sichtbar zu machen (vgl. auch Cronbach & Gleser, 1965). Ein weiterer Vorteil von statistischen Entscheidungsmodellen ist, dass sie eine Information über die Klarheit oder die Ambivalenz in einer Entscheidung liefern können. Durch die Bestimmung der Entscheidungssicherheit im Modell können Entscheidungen in klare und ambigue Entscheidungen unterteilt werden und so im diagnostischen Prozess erkannt werden, für welche Einzelfälle weitere Diagnostik erfolgen sollte. Darüber hinaus kann die Entscheidungssicherheit auch zur Steuerung des Evaluationsverhaltens verwendet werden. Werden etwa Treatmententscheidungen mit geringer Sicherheit getroffen, so sollten diese nach kürzerer Zeit evaluiert werden, als wenn die Entscheidungen mit einer hohen Sicherheit getroffen wurden.

6.3 Beitrag der Entscheidungstheorie für das Feld der Diagnostik

Ein entscheidungstheoretischer Zugang zur diagnostischen Entscheidung ermöglicht eine neue Dimension der Testbeurteilung, denn der (praktische) Nutzen eines Tests wird bisher von den Hauptgütekriterien nicht berücksichtigt (Cronbach & Gleser, 1965). Tatsächlich kommen die Hauptgütekriterien aus entscheidungstheoretischer Perspektive an anderen Stellen zum Tragen. Reliabilität zum Beispiel schlägt in der Einzelfallentscheidung nur an den Stellen zu Buche, wo sich Nutzenfunktionen kreuzen. Die Validität eines Tests wirkt sich auf die saubere Trennung der einzelnen Kurven aus, ist jedoch genauso stark mit dem persönlichen Wertesystem des:r Entscheidungsträger:in verknüpft. Angenommen, eine Trainingseinheit erhöht valide einen bestimmten, aber sehr begrenzten Wissensbereich. Nun nehme man an, dass dieser durch einen Test exakt und valide gemessen werden kann, so bleibt trotzdem in der diagnostischen Entscheidung, ob das Training in Zukunft eingesetzt wird, eine Bewertung über den persönlichen Nutzen dieses sehr begrenzten Wissensbestands beim:bei der Entscheidungsträger:in. Schlussendlich läuft alles auf die Entscheidung heraus, ob der:die Entscheidungsträger:in das Zielkriterium als solches anerkennt und wie er:sie es gegenüber anderen Kriterien gewichtet. Validität ist also zwar ein notwendiges Kriterium für den Einsatz zur diagnostischen Entscheidungsfindung, aber kein hinreichendes Kriterium.

Grundsätzlich gibt die Formalisierung eines in der realen Welt existierenden Problems immer einen Rahmen, um verschiedene Ansätze (z. B. Empfehlungen) objektiv miteinander zu vergleichen. So treten an die Stelle der rhetorischen Überzeugungskraft objektive Kriterien und numerische Kennwerte. Sicherlich birgt der Prozess der Formalisierung und der Parametrisierung ebenfalls viele Streitpunkte. Hat jedoch diesbezüglich erst einmal eine Einigung stattgefunden, so weicht eine rhetorische Argumentation einer mathematischen - einer Argumentation, die durch zwingende Logik und einen faktischen Charakter besticht (Gilboa, 2009). Bei gleichzeitiger Formalisierung aller Unsicherheit ermutigen die statistischen Modelle zu einer sehr zurückhaltenden Position. Unter Berücksichtigung von Schätzunsicherheit, Messunsicherheit und designbedingter Unsicherheit zeigt sich nur zu oft, dass statistische Modelle gerade in den Sozialwissenschaften kaum starke Aussagen und Empfehlungen für Individuen rechtfertigen (Fisher, Medaglia, & Jeronimus, 2018). Die Messung ist maßgeblich vom Messfehler bestimmt, Modelle klären selten mehr als 40% der Varianz auf - selbst bei flexibler Modellierung - und nur partiell identifizierte Verteilungen, wie sie durch nicht zufällige fehlende Werte oder nicht-experimentelle Designs entstehen, werden bisher in den Modellierungen der angewandten Forschung gar nicht berücksichtigt.

6.4 Grenzen statistischer Verfahren der Entscheidungstheorie

Statistische Verfahren können Zusammenhänge zwischen Variablen absichern. Sie können Kosten, Voraussetzungen und erwartete Outcomes in Verbindung setzen und die Zusammenhangssicherheit quantifizieren. Ein Wertesystem, das definiert, wie wichtig welche Outcomes sind und wie viel Risiko man eingehen will, muss der:die Entscheidende selbst festlegen. Nur dann kann eine formal gesehen optimale Entscheidung abgeleitet werden. (Cronbach & Gleser, 1965)

Statistische Verfahren können zudem keine "once and for all" Lösung bieten. Tatsächlich findet sich hier bei Cronbach & Gleser (1965) schon der Verweis darauf, dass Modelle aktuell gehalten werden müssen und sich den aktuellen Entwicklungen anpassen müssen. Zudem gilt, dass ein Entscheidungsmodell, sobald es einmal in Gebrauch ist, nur noch bedingt auf Basis der dann gewonnenen Daten verbessert oder evaluiert werden kann (Stoye, 2009). Da alle im Einsatz des Modells gewonnenen Daten auf einer nicht-zufälligen Zuordnung der Personen zu den Treatmentbedingungen basieren, sind alle Verteilungen der Zielvariablen partiell identifiziert. Anders ausgedrückt: die gemessenen Effekte basieren maßgeblich auf dem durch das Entscheidungsmodell ausgelösten Selektionseffekt. Es muss also zur Qualitätssicherung der Entscheidungsmodelle immer wieder rekursiv auf Basis von experimentellen Daten die Optimalität der Modelle getestet werden.

6.5 Gründe für den zurückhaltenden Einsatz entscheidungstheoretischer Modelle in der Diagnostik

Tatsächlich erscheint es nach einer fast dreijährigen Auseinandersetzung mit entscheidungstheoretischen Themen erstaunlich, dass diese elegante und fundierte Denkweise bisher kaum Einzug in Forschung und Anwendung im Bereich der Diagnostik gefunden hat, obwohl dies immer wieder als wünschenswert propagiert wurde (vgl. Cronbach & Gleser, 1965; Irtel, 1996; Schmidt-Atzert & Amelang, 2012). Möglicherweise könnte ein Mangel an Interdisziplinarität ursächlich sein. Es zeigt sich, dass entscheidungstheoretische Überlegungen überwiegend in einer stark formalisierten und mathematischen Darstellung präsentiert und diskutiert werden (vgl. Manski, Rudner, Stoye, Cronbach, etc.). Es scheint bisher keine Darstellung und keine Implementation zu geben, welche die Konzepte für die Praxisforschung anwendbar aufbereitet. Diese Arbeit hofft, dahingehend einen ersten Schritt zu machen.

7 Literatur

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Amendum, S. J., Conradi, K., & Hiebert, E. (2018). Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension. *Educational Psychology Review*, *30*(1), 121–151.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140.

Anmarkrud, Ø., & Bråten, I. (2012). Naturally-occurring comprehension strategies instruction in 9th-grade language arts classrooms. *Scandinavian Journal of Educational Research*, *56*(6), 591–623.

Archer, M. (2013). *Social origins of educational systems*. London: Routledge.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme,..., & M. Weiß (Eds.), *PISA 2000* (pp. 69–137). Opladen: VS Verlag für Sozialwissenschaften.

Artinger, F. M., Artinger, S., & Gigerenzer, G. (2019). CYA: Frequency and causes of defensive decisions in public administration. *Business Research*, *12*(1), 9–25.

Asparouhov, T., Muthén, B., & Muthén, B. O. (2006). Robust chi square difference testing with mean and variance adjusted test statistics. *Matrix*, *1*(5), 1–6.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. Boca Raton: CRC Press.

Baldwin, R. S., Peleg-Bruckner, Z., & McClintock, A. H. (1985). Effects of topic interest and prior knowledge on reading comprehension. *Reading Research Quarterly*, *20*(4), 497–504.

Baumert, J., Stanat, P., & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In J. Baumert, E. Klieme,..., & M. Weiß (Eds.), *PISA 2000* (pp. 15–68). Opladen: VS Verlag für Sozialwissenschaften.

Beckmann, J. S., & Lew, D. (2016). Reconciling evidence-based medicine and precision medicine in the era of big data: Challenges and opportunities. *Genome Medicine*, *8*(1), 134.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63–88.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis*. New York: Springer Science & Business Media.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. Commentarii academiae scientiarum imperialis. *Petropolitanae*, *5*, 175–192.

Bernoulli, D. (1738/1954). Exposition of a new theory of risk evaluation. *Econometrica*, *22*(1), 23–36.

Birnbaum, Z. W. (1968). *On the importance of different components in a multicomponent system*. Washington University Seattle Lab of Statistical Research.

Blavatskyy, P. R. (2005). Back to the St. Petersburg paradox? *Management Science*, *51*(4), 677–678.

Bliese, P. (2016). *Multilevel: Multilevel functions*. Retrieved from https://CRAN.R-project.org/package=multilevel

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah: Psychology Press.

Bos, W., Strietholt, R., Goy, M., Stubbe, T. C., Tarelli, I., & Hornberg, S. (2010). *IGLU 2006. Dokumentation der Erhebungsinstrumente*. München: Waxmann.

Boudreau, J. W. (1991). Utility analysis in human resource management decisions. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology (2.ed)* (Vol. 2, pp. 621–745). Palo Alto, CA: Consulting Psychologists Press.

Bremerich-Vos, A., Wendt, H., & Bos, W. (2017). Lesekompetenzen im internationalen Vergleich: Testkonzeption und Ergebnisse. In A. Hußmann, H. Wendt, W. Bos, Bremerich-Vos A., & R. Valtin (Eds.), *IGLU 2016* (pp. 79–142). Münster: Waxmann.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*, *2*(2), 171–183.

Brown, L. (1986). *Fundamentals of statistical exponential families: With applications in statistical decision theory*. Hayward, CA: Institute of Mathematical Statistics.

Brown, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Testing Structural Equation Models*, *154*(1), 136–162.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA Journal of Applied Mathematics*, *6*(1), 76–90.

Bühner, M. (2011). *Einführung in die Test-und Fragebogenkonstruktion (3rd ed.)*. München: Pearson Studium.

Bühner, M. (2021). *Einführung in die Test-und Fragebogenkonstruktion (4th ed.)*. München: Pearson Studium.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition*, *29*(6), 850–859.

Cattaneo, M. E. (2013). Likelihood decision functions. *Electronic Journal of Statistics*, *7*, 2924–2946.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research*, *49*(2), 278–293.

Chall, J. S. (1983). *Stages of reading development*. New York: McGraw-Hill.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chamberlain, G. (2000). Econometrics and decision theory. *Journal of Econometrics*, *95*(2), 255–283.

Clemen, R. T. (1996). *Making hard decisions: An introduction to decision analysis*. Boston: Cole Publishing Company.

Coles, M., & Hall, C. (2002). Gendered readings: Learning from children's reading choices. *Journal of Research in Reading*, *25*(1), 96–108.

Collins, L. M., Murphy, S. A., & Bierman, K. L. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science*, *5*(3), 185–196.

Connell, M. W., Sheridan, K., & Gardner, H. (2003). On abilities and domains. In R. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 126–155). Cambridge: Cambridge University Press.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Chicago: University of Illinois Press.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Cummins, J. (2012). The intersection of cognitive and sociocultural factors in the development of reading comprehension among immigrant students. *Reading and Writing*, *25*(8), 1973–1990.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*(3), 277–299.

Dehaene, S. (2014). *Lesen: Die größte Erfindung der Menschheit und was dabei in unseren Köpfen passiert*. München: Albrecht Knaus Verlag.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, *125*(1-2), 141–173.

Dossey, J., Hartig, J., Klieme, E., & Wu, M. (2004). *Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003*. Paris: OECD Publications.

Duffy, G. G. (1993). Rethinking strategy instruction: Four teachers' development and their low achievers' understandings. *The Elementary School Journal*, *93*(3), 231–247.

Eberwein, M., Schui, G., & Krampen, G. (2006). Zur Entwicklung deutschsprachiger Testverfahren in der 2. Hälfte des 20. Jahrhunderts. *Diagnostica*, *52*(4), 199–207.

Edwards, W. (1977). How to use multiattribute utility measurement for social decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, *7*(5), 326–340.

Engelmayer, O. (1968). *Das Soziogramm in der modernen Schule*. München: Ehrenwirth.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2007). *Regression*. Berlin: Springer.

Feller, W. (2008). *An introduction to probability theory and its applications*. New York: John Wiley & Sons.

Ferguson, T. S. (1967). *Mathematical statistics—a decision theoretic approach*. New York: Academic Press New York.

Finney, D. (1962). The statistical evaluation of educational allocation and selection. *Journal of the Royal Statistical Society: Series A (General)*, *125*(4), 525–549.

Fishburn, P. C., & Keeney, R. L. (1974). Seven independence concepts and continuous multiattribute utility functions. *Journal of Mathematical Psychology*, *11*(3), 294–327.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), 6106–6115.

Fisseni, H. J. (2004). *Lehrbuch der psychologischen Diagnostik: Mit Hinweisen zur Intervention*. Göttingen: Hogrefe.

Freebody, P., & Tirre, W. C. (1985). Achievement outcomes of two reading programmes: An instance of aptitude-treatment interaction. *British Journal of Educational Psychology*, *55*(1), 53–60.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*(1), 55–77.

Fuchs, D., Kearns, D. M., Fuchs, L. S., Elleman, A. M., Gilbert, J. K., Patton, S., … Compton, D. L. (2019). Using moderator analysis to identify the first-grade children who benefit more and less from a reading comprehension program: A step toward aptitude-by-treatment interaction. *Exceptional Children*, *85*(2), 229–247.

Galuschka, K., Rothe, J., & Schulte-Körne, G. (2015). Die methodische Beurteilung und qualitative Bewertung psychometrischer Tests am Beispiel aktueller Verfahren zur Erfassung der Lese-und/oder Rechtschreibleistung. *Zeitschrift für Kinder-und Jugendpsychiatrie und Psychotherapie*, *43*(5), 317–334.

García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*(1), 74–111.

Ghelani, K., Sidhu, R., Jain, U., & Tannock, R. (2004). Reading comprehension and reading related abilities in adolescents with reading disabilities and attention-deficit/hyperactivity disorder. *Dyslexia*, *10*(4), 364–384.

Gilboa, I. (2009). *Theory of decision under uncertainty* (Vol. 45). Cambridge: Cambridge University Press.

Gilboa, I., & Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics*, *110*(3), 605–639.

Gold, A., Trenk-Hinterberger, I., & Souvignier, E. (2009). Kapitel 11 „Die Textdetektive "–Ein strategieorientiertes Programm zur Förderung des Leseverständnisses. In W. Lenhard (Ed.), *Diagnostik und Förderung des Leseverständnisses* (p. 207). Göttingen: Hogrefe.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574.

Gomez-Uribe, C. A., & Hunt, N. (2015). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, *6*(4), 1–19.

Good, J., I. (1952). Rational decisions. *Journal of the Royal Statistical Society Series B*, *14*, 107–114.

Groeben, N. (1982). *Leserpsychologie*. Münster: Aschendorff.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19.

Gulliksen, H. (2009). *Theory of mental tests*. New York: Routledge.

Harris, T. L., & Hodges, R. E. (1995). *The literacy dictionary: The vocabulary of reading and writing*. Newark: International Reading Association.

Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Ed.), *Leistung und Leistungsdiagnostik* (pp. 127–143). Berlin: Springer.

Hasselhorn, M., Schneider, W., & Marx, H. (2000). *Diagnostik von Lese-Rechtschreibschwierigkeiten. Tests und Trends*. Göttingen: Hogrefe.

Hastie, T. J., & Tibshirani, R. J. (2017). *Generalized Additive Models*. London: Routledge.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hayden, B. Y., & Platt, M. L. (2009). The mean, the median, and the St. Petersburg paradox. *Judgment and Decision Making*, *4*(4), 256.

Hilbert, S., Pargent, F., Kraus, E., Naumann, F., Eichhorn, K., Ungar, P., & Bühner, M. (2020). What's the measure? An empirical investigation of self-ratings on response scales. *International Journal of Social Research Methodology*, 1–20.

Hirano, K., & Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, *77*(5), 1683–1701.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher*, *58*(8), 702–714.

Hudson, R. F., Pullen, P. C., Lane, H. B., & Torgesen, J. K. (2008). The complex nature of reading fluency: A multidimensional view. *Reading & Writing Quarterly*, *25*(1), 4–32.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In M. J. Gierl & J. P. Leighton (Eds.), *Cognitive diagnostic assessment for education* (pp. 19–61). Cambridge: Cambridge University Press.

Indrisano, R., & Chall, J. S. (1995). Literacy development. *Journal of Education*, *177*(1), 63–83.

Ingenkamp, K.-H. (2007). *Lehrbuch der pädagogischen Diagnostik*. Basel: Beltz.

Irtel, H. (1996). *Entscheidungs-und testtheoretische Grundlagen der psychologischen Diagnostik*. Frankfurt a.M.: P. Lang.

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

Jank, W., & Meyer, H. (2002). *Didaktische Modelle*. Berlin: Cornelsen-Scriptor.

Jia, J., & Dyer, J. S. (1996). A standard measure of risk and risk-value models. *Management Science*, *42*(12), 1691–1705.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350.

Keefer, D. L. (1991). Resource allocation models with risk aversion and probabilistic dependence: Offshore oil and gas bidding. *Management Science*, *37*(4), 377–395.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163.

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

Kirkwood, C. W. (1997). *Strategic Decision Making*. Belmont: Duxbury Press.

Kirkwood, C. W. (2004). Approximating risk aversion in decision analysis applications. *Decision Analysis*, *1*(1), 51–67.

Klauer, K. J. (1982). *Handbuch der pädagogischen Diagnostik (2.ed.)*. Düsseldorf: Pädagogischer Verlag Schwann.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52*(6), 876–903.

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme,..., & M. Weiß (Eds.), *PISA 2000* (pp. 139–190). Opladen: VS Verlag für Sozialwissenschaften.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie*, *216*(2), 61–73.

Kraus, E., Wild, J., Schilcher, A., & Hilbert, S. (2020). *Reading related activities in second grade and their effect on fluency- an evaluation of teachers' classroom practices*. Retrieved from psyarxiv.com/g2f5x

Kraus, E., Wild, J., Schilcher, A., & Hilbert, S. (2021). *Research Report zur Pilotierung des Bayerischen Lesetests (BYLET)*. Retrieved from https://osf.io/bmp54/?view_only=3db22c0b3cec4383820514243%0Ae431b1b

Kroll, Y., Levy, H., & Markowitz, H. M. (1984). Mean-variance versus direct utility maximization. *The Journal of Finance*, *39*(1), 47–61.

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, *45*(2), 230–251.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, *95*(1), 3.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323.

Lee, J., & Park, O. (2008). Adaptive instructional systems. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 469–484). Mahwah: Lawrence Erlbaum Associates Publishers.

Lei, P.-W., & Wu, Q. (2007). Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice*, *26*(3), 33–43.

Lenhard, W., & Lenhard, A. (2014-2017). *Berechnung des Lesbarkeitsindex LIX nach Björnson*. Verfügbar unter: http://www.psychometrica.de/lix.html. Bibergau: Psychometrica.

Lenhard, W., & Schneider, W. (2009). *Diagnostik und Förderung des Leseverständnisses* (Vol. 7). Göttingen: Hogrefe.

Lenhard, W., Schneider, W., Lenhard, A., & Schneider, W. (2018). *ELFE II: Ein Leseverständnistest für Erst-bis Siebtklässler-Version II*. Göttingen: Hogrefe.

Lervåg, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry*, *51*(5), 612–620.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949.

Linacre, J. (1996). The Rasch model cannot be "disproved." *Rasch Measurement Transactions*, *10*(3), 512–514.

Logan, S., & Johnston, R. (2009). Gender differences in reading ability and attitudes: Examining where these differences lie. *Journal of Research in Reading*, *32*(2), 199–214.

Logan, S., & Johnston, R. (2010). Investigating gender differences in reading. *Educational Review*, *62*(2), 175–187.

Lukesch, N. (1998). *Einführung in die pädagogisch-psychologische Diagnostik (2.ed.)*. Regensburg: Roderer.

Mair, P., Hatzinger, R., Maier, M. J., Rusch, T., & Mair, M. P. (2020). *eRm: Extended Rasch Modeling. 1.0-2*. Retrieved from https://cran.r-project.org/package=eRm

Manski, C. F. (2003). *Partial identification of probability distributions*. New York: Springer.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, *72*(4), 1221–1246.

Manski, C. F., & Tetenov, A. (2007). Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *Journal of Statistical Planning and Inference*, *137*(6), 1998–2010.

Maranell, G. (2017). *Scaling : A sourcebook for behavioral scientists*. Abingdon: Routledge.

Markowitz, H. (2014). Mean–variance approximations to expected utility. *European Journal of Operational Research*, *234*(2), 346–355.

Marzano, R. J., & Kendall, J. S. (2006). *The new taxonomy of educational objectives*. Thousand Oaks: Corwin Press.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127.

McDonald Connor, C., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., … Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child× instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100.

McNamara, D. S., Graesser, A. C., Cai, Z., & Kulikowich, J. M. (2011). Coh-metrix easability components: Aligning text difficulty with theories of text comprehension. *Annual meeting of the american educational research association*, *10*. New Orleans.

McNamara, D. S., Graesser, A., & Louwerse, M. (2012). Sources of text difficulty: Across genres and grades. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Lanham: R&L Education.

Merisuo-Storm, T. (2006). Girls and boys like to read and write different texts. *Scandinavian Journal of Educational Research*, *50*(2), 111–125.

Mesmer, H. A. E. (2008). *Tools for matching readers to texts: Research-based practices*. New York: Guilford Press.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383.

Millard, E. (2002). *Differently literate: Boys, girls and the schooling of literacy*. London Philadelphia: Routledge.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195.

Morrow, L. (2019). *Best practices in literacy instruction*. New York: The Guilford Press.

Morton, A., & Fasolo, B. (2009). Behavioural decision theory for multi-criteria decision analysis: A guided tour. *Journal of the Operational Research Society*, *60*(2), 268–275.

Mulaik, S. (1972). A mathematical investigation of some multidimensional rasch models for psychological tests. *Annual meeting of the psychometric society*. Princeton, NJ.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, B. O. (1998). *Mplus technical appendices*. Los Angeles: Muthén & Muthén.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, 286–295.

Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (2015). *Handbook of Practical Program Evaluation*. San Francisco: John Wiley & Sons.

Nitko, A. J., & Hsu, T.-C. (1974). Using domain-referenced tests for student placement, diagnosis and attainment in a system of adaptive, individualized instruction. *Educational Technology*, *14*(6), 48–54.

Novick, M. R., & Lindley, D. V. (1978). The use of more realistic utility functions in educational applications. *Journal of Educational Measurement*, *15*(1), 181–191.

O'Connor, R. E., Bell, K. M., Harty, K. R., Larkin, L. K., Sackor, S. M., & Zigmond, N. (2002). Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. *Journal of Educational Psychology*, *94*(3), 474–485.

Oakhill, J. V., & Petrides, A. (2007). Sex differences in the effects of interest on boys' and girls' reading comprehension. *British Journal of Psychology*, *98*(2), 223–235.

Österbauer, V., Bachinger, A., Winter, B. O., Paasch, D., & Illetschko, M. (2020). Leseförderung revisited–Sind die verschiedenen Verfahren zur Leseförderung im österreichischen Deutschunterricht der 4. Schulstufe angekommen? *Leseforum. Ch*, *1*, 1–26.

Panayides, P., Robinson, C., & Tymms, P. (2010). The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal*, *36*(4), 611–626.

Passow, F. (1852/2008). Handwörterbuch der Griechischen Sprache, (2.ed). *Leipzig: Academic in Wissenschaftliche Buchgesellschaft*.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Perfetti, C. A., Marron, M. A., & Foltz, P. W. (1996). *Sources of comprehension failure: Theoretical perspectives and case studies.* Mahwah: Lawrence Erlbaum Associates Publishers.

Petermann, F., & Eid, M. (2006). *Handbuch der psychologischen Diagnostik*. Göttingen: Hogrefe.

Peterson, M. (2017). *An introduction to decision theory*. Cambridge: Cambridge University Press.

Pikulski, J. J., & Chard, D. J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, *58*(6), 510–519.

Prenzel, M., Rost, J., Senkbeil, M., Häußler, P., & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme,..., & M. Weiß (Eds.), *PISA 2000* (pp. 191–248). Opladen: VS Verlag für Sozialwissenschaften.

Preston, R. (2006). Review of UNESCO, education for all global monitoring report. *International Journal of Educational Development*, *26*(1), 668–669.

Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, *39*(2), 1180–1210.

R Core Team. (2020). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Danmarks Paedagogiske Institut.

Rasinski, T. V., Reutzel, D. R., Chard, D., & Linan-Thompson, S. (2011). Reading fluency. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research, volume IV* (pp. 312–345). Mahwah: Routledge.

Reckase, M. (1972). *Development and application of a multivariate logistic latent trait model.* (Doctoral dissertation). Syracuse University.

Reckase, M. (2009). *Multidimensional Item Response Theory.* New York London: Springer.

Rettig, D. (2012). *Im Zwiespalt – Warum fallen leichte Entscheidungen schwer?* Retrieved from %7Bhttps://www.alltagsforschung.de/im-zwiespalt-warum-fallen-leichte-entscheidungen-schwer/%7D

Reulecke, W., & Rollett, B. (1976). Pädagogische Diagnostik und lernzielorientierte Tests. In K. Pawlik (Ed.), *Diagnose der Diagnostik* (pp. 177–202). Stuttgart: Klett.

Richter, T., & Christmann, U. (2002). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. *Lesekompetenz: Bedingungen, Dimensionen, Funktionen*, *1*(1), 25–49.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Rosebrock, C., & Nix, D. (2017). *Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung.* Hohengehren: Schneider.

Rosenhead, J., Elton, M., & Gupta, S. K. (1972). Robustness and optimality as criteria for strategic decisions. *Journal of the Operational Research Society*, *23*(4), 413–431.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

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

Rost, J., & Spada, H. (1978). Learning tests: Psychometric and psychological considerations on a method to get diagnostic information from process data. *XIXth international congress of applied psychology.* München.

Roth, M., & Herzberg, P. Y. (2008). Psychodiagnostik in der Praxis: State of the art? *Klinische Diagnostik und Evaluation*, *1*(1), 5–18.

Rudner, L. M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research, and Evaluation*, *14*(1), 8.

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, *46*(253), 55–67.

Savage, L. J. (1972). *The foundations of statistics*. New York: Dover Publications.

Schilcher, A., Wild, J., & Steinert, M. (2019). FiLBY-2 Lehrerhandreichung. Fachintegrierte Leseförderung Bayern. *Regensburg. O.V.*

Schmidt-Atzert, L., & Amelang, M. (2012). *Psychologische Diagnostik*. Berlin: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, *24*(111), 647–656.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, *28*(1), 1–39.

Stelter, A., & Miethe, I. (2019). Forschungsmethoden im Lehramtsstudium–aktueller Stand und Konsequenzen. *Erziehungswissenschaft*, *30*(1), 7–8.

Stenner, A. J. (1996). Measuring reading comprehension with the lexile framework. *California comparability symposium*. Burlingame.

Stevens, S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stoye, J. (2009). Partial identification and robust treatment choice: An application to young offenders. *Journal of Statistical Theory and Practice*, *3*(1), 239–254.

Stoye, J. (2012). Minimax regret treatment choice with limited validity of experiments or with covariates. *Journal of Econometrics*, *166*(1), 138–156.

Swanson, H. L., Cochran, K. F., & Ewers, C. A. (1989). Working memory in skilled and less skilled readers. *Journal of Abnormal Child Psychology*, *17*(2), 145–156.

Sweller, J. (2011). *Cognitive Load Theory*. New York: Springer.

Sympson, J. B. (1978). A model for testing with multidimensional items. *Proceedings of the 1977 computerized adaptive testing conference*. Minnesota.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, *2*(1), 167–195.

Thissen, D. E., & Wainer, H. E. (2001). *Test scoring.* Mahwah: Lawrence Erlbaum Associates Publishers.

Tzouveli, P., Mylonas, P., & Kollias, S. (2008). An intelligent e-learning system based on learner profiling and learning resources adaptation. *Computers & Education*, *51*(1), 224–238.

Van der Linden, W. J. (2016). *Handbook of Item Response Theory: Volume 1: Models.* Monterey: CRC Press.

Van Der Linden, W. J. (1987). *Applications of decision theory to test-based decision making. Project psychometric aspects of item banking no. 23. Research report 87-9.* University of Twente.

Van der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement*, *4*(4), 469–492.

Van der Linden, W. J. (1981). Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery scores. *Psychometrika*, *46*(3), 257–274.

Van der Linden, W. J. (1998). A decision theory model for course placement. *Journal of Educational and Behavioral Statistics*, *23*(1), 18–34.

Van der Linden, W. J., & Mellenbergh, G. J. (1978). Coefficients for tests from a decision theoretic point of view. *Applied Psychological Measurement*, *2*(1), 119–134.

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

Van Keer, H., & Verhaeghe, J. P. (2005). Comparing two teacher development programs for innovating reading comprehension instruction with regard to teachers' experiences and student outcomes. *Teaching and Teacher Education*, *21*(5), 543–562.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology* (pp. 300–319). Oxford: Oxford University Press.

Vandewaetere, M., & Clarebout, G. (2014). Advanced technologies for personalized learning, instruction, and performance. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 425–437). Heidel-

berg: Springer.

Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, *47*(2), 247–275.

Vrijhof, B. J., Mollenbergh, G. J., & Van den Brink, W. P. (1983). Assessing and studying utility functions in psychometric decision theory. *Applied Psychological Measurement*, *7*(3), 341–357.

Wald, A. (1950). *Statistical decision functions.* London: Wiley Publications.

Walter, J. (2013). *VSL: Verlaufsdiagnostik sinnerfassenden Lesens*. Göttingen: Hogrefe.

Wei, G. C., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704.

Weinert, F. E. (2001). *Leistungsmessungen in Schulen*. Basel: Beltz.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*(4), 479–494.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wild, J., & Pissarek, M. (2018). *Ratte. Regensburger Analysetool für Texte*. Retrieved from http://bit.%20ly/2UpRUBb

Wild, J., Schilcher, A., & Steinert, M. (2019). FiLBY-3 Lehrerhandreichung. Fachintegrierte Leseförderung Bayern. *Regensburg. O.V.*

Wimmer, H., & Mayringer, H. (2016). *SLS 2-9: Salzburger Lese-Screening für die Schulstufen 2-9; Manual*. Göttingen: Hogrefe.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Boca Raton: CRC press.

Wood, S., & Wood, M. S. (2015). Package 'mgcv.' *R Package Version*, 1–7.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428.

Ziegler, M., & Bühner, M. (2012). *Grundlagen der psychologischen Diagnostik*. Wiesbaden: Springer.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21–29.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., … Cohen, G. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382.

Anhang A

Anhang B

Anhang A

Inhaltsverzeichnis

Deskriptive Ergebnisse

Lösungshäufigkeiten in Abhängigkeit des Alters BYLET-A

Tabelle 1: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
Mitte Klasse 2	0.63	0.59	0.49	0.06	0.64
Ende Klasse 2	0.72	0.69	0.60	0.09	0.75
Mitte Klasse 3	0.74	0.75	0.61	0.20	0.79
Ende Klasse 3	0.76	0.76	0.67	0.19	0.79
Anfang Klasse 4	0.83	0.81	0.70	0.20	0.86
Weihnachten Klasse 4	0.78	0.86	0.78	0.36	0.86
Anfang Klasse 6	0.93	0.93	0.94	0.58	0.97
Anfang Klasse 7	0.98	0.98	0.91	0.75	0.99

Tabelle 2: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 6-10

	Item 6	Item 7	Item 8	Item 9	Item 10
Mitte Klasse 2	0.38	0.23	0.06	0.38	0.23
Ende Klasse 2	0.48	0.30	0.10	0.44	0.33
Mitte Klasse 3	0.53	0.27	0.19	0.49	0.38
Ende Klasse 3	0.52	0.30	0.16	0.54	0.40
Anfang Klasse 4	0.57	0.34	0.21	0.62	0.44
Weihnachten Klasse 4	0.53	0.28	0.33	0.58	0.39
Anfang Klasse 6	0.75	0.68	0.53	0.86	0.76
Anfang Klasse 7	0.89	0.67	0.70	0.88	0.78

Tabelle 3: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
Mitte Klasse 2	0.10	0.24	0.32	0.23	0.06
Ende Klasse 2	0.13	0.31	0.42	0.31	0.08
Mitte Klasse 3	0.20	0.39	0.55	0.41	0.12
Ende Klasse 3	0.23	0.42	0.54	0.41	0.12
Anfang Klasse 4	0.29	0.46	0.64	0.49	0.16
Weihnachten Klasse 4	0.17	0.56	0.72	0.36	0.31
Anfang Klasse 6	0.72	0.90	0.94	0.78	0.36
Anfang Klasse 7	0.74	0.87	0.99	0.89	0.49

Tabelle 4: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
Mitte Klasse 2	0.05	0.24	0.03	0.15	0.10
Ende Klasse 2	0.09	0.31	0.03	0.19	0.16
Mitte Klasse 3	0.13	0.46	0.10	0.26	0.23
Ende Klasse 3	0.18	0.46	0.10	0.26	0.22
Anfang Klasse 4	0.19	0.55	0.12	0.33	0.26
Weihnachten Klasse 4	0.44	0.50	0.19	0.36	0.42
Anfang Klasse 6	0.49	0.79	0.40	0.53	0.76
Anfang Klasse 7	0.55	0.84	0.38	0.63	0.82

Lösungshäufigkeiten in Abhängigkeit des Alters BYLET-B

Tabelle 5: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
Mitte Klasse 2	0.43	0.57	0.51	0.30	0.50
Ende Klasse 2	0.41	0.62	0.55	0.30	0.54
Mitte Klasse 3	0.49	0.70	0.59	0.33	0.66
Ende Klasse 3	0.48	0.74	0.66	0.38	0.70

Anfang Klasse 4	0.58	0.78	0.68	0.41	0.74
Weihnachten Klasse 4	0.42	0.75	0.65	0.32	0.68
Anfang Klasse 6	0.86	0.96	0.94	0.70	0.97
Anfang Klasse 7	0.88	0.96	0.98	0.78	0.96

Tabelle 6: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 6-10

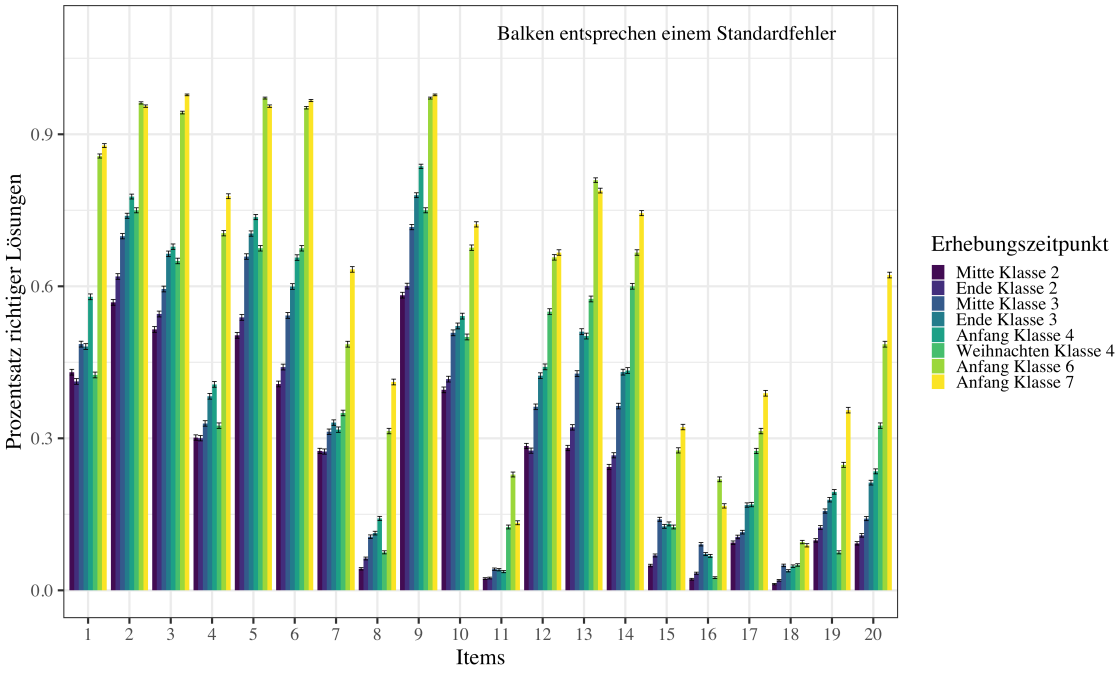
	Item 6	Item 7	Item 8	Item 9	Item 10
Mitte Klasse 2	0.41	0.28	0.04	0.58	0.40
Ende Klasse 2	0.44	0.27	0.06	0.60	0.42
Mitte Klasse 3	0.54	0.31	0.11	0.72	0.51
Ende Klasse 3	0.60	0.33	0.11	0.78	0.52
Anfang Klasse 4	0.66	0.32	0.14	0.84	0.54
Weihnachten Klasse 4	0.68	0.35	0.07	0.75	0.50
Anfang Klasse 6	0.95	0.49	0.31	0.97	0.68
Anfang Klasse 7	0.97	0.63	0.41	0.98	0.72

Tabelle 7: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
Mitte Klasse 2	0.02	0.28	0.28	0.24	0.05
Ende Klasse 2	0.02	0.28	0.32	0.27	0.07
Mitte Klasse 3	0.04	0.36	0.43	0.36	0.14
Ende Klasse 3	0.04	0.42	0.51	0.43	0.13
Anfang Klasse 4	0.04	0.44	0.50	0.43	0.13
Weihnachten Klasse 4	0.12	0.55	0.58	0.60	0.12
Anfang Klasse 6	0.23	0.66	0.81	0.67	0.28
Anfang Klasse 7	0.13	0.67	0.79	0.74	0.32

Tabelle 8: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
Mitte Klasse 2	0.02	0.09	0.01	0.10	0.09
Ende Klasse 2	0.03	0.11	0.02	0.12	0.11
Mitte Klasse 3	0.09	0.11	0.05	0.16	0.14
Ende Klasse 3	0.07	0.17	0.04	0.18	0.21
Anfang Klasse 4	0.07	0.17	0.05	0.19	0.23
Weihnachten Klasse 4	0.03	0.28	0.05	0.07	0.32
Anfang Klasse 6	0.22	0.31	0.10	0.25	0.49
Anfang Klasse 7	0.17	0.39	0.09	0.36	0.62



Lösungshäufigkeiten in Abhängigkeit des Alters BYLET-C

Tabelle 9: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
Mitte Klasse 2	0.80	0.53	0.39	0.14	0.56
Ende Klasse 2	0.82	0.54	0.39	0.15	0.63
Mitte Klasse 3	0.90	0.61	0.40	0.21	0.73
Ende Klasse 3	0.90	0.66	0.39	0.22	0.78
Anfang Klasse 4	0.95	0.73	0.43	0.20	0.84
Weihnachten Klasse 4	0.91	0.70	0.42	0.20	0.79
Anfang Klasse 6	1.00	0.92	0.32	0.17	0.98
Anfang Klasse 7	1.00	0.98	0.51	0.15	0.98

Tabelle 10: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 6-10

	Item 6	Item 7	Item 8	Item 9	Item 10
Mitte Klasse 2	0.47	0.37	0.04	0.40	0.23
Ende Klasse 2	0.55	0.43	0.06	0.42	0.29
Mitte Klasse 3	0.62	0.42	0.10	0.54	0.34
Ende Klasse 3	0.68	0.44	0.11	0.57	0.39
Anfang Klasse 4	0.74	0.48	0.11	0.67	0.40
Weihnachten Klasse 4	0.73	0.39	0.11	0.64	0.42
Anfang Klasse 6	0.97	0.45	0.25	0.89	0.70
Anfang Klasse 7	0.99	0.45	0.27	0.92	0.78

Tabelle 11: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
Mitte Klasse 2	0.01	0.17	0.42	0.36	0.03

Ende Klasse 2	0.01	0.18	0.51	0.43	0.04
Mitte Klasse 3	0.02	0.19	0.71	0.56	0.06
Ende Klasse 3	0.02	0.20	0.73	0.61	0.07
Anfang Klasse 4	0.01	0.17	0.82	0.63	0.07
Weihnachten Klasse 4	0.01	0.17	0.83	0.64	0.08
Anfang Klasse 6	0.02	0.10	0.98	0.67	0.08
Anfang Klasse 7	0.03	0.12	0.98	0.75	0.10

Tabelle 12: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Alters Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
Mitte Klasse 2	0.02	0.17	0.01	0.18	0.18
Ende Klasse 2	0.02	0.19	0.01	0.24	0.23
Mitte Klasse 3	0.02	0.24	0.02	0.32	0.33
Ende Klasse 3	0.03	0.26	0.01	0.35	0.38
Anfang Klasse 4	0.03	0.29	0.02	0.37	0.41
Weihnachten Klasse 4	0.02	0.34	0.01	0.40	0.47
Anfang Klasse 6	0.05	0.46	0.01	0.50	0.76
Anfang Klasse 7	0.00	0.56	0.02	0.57	0.86

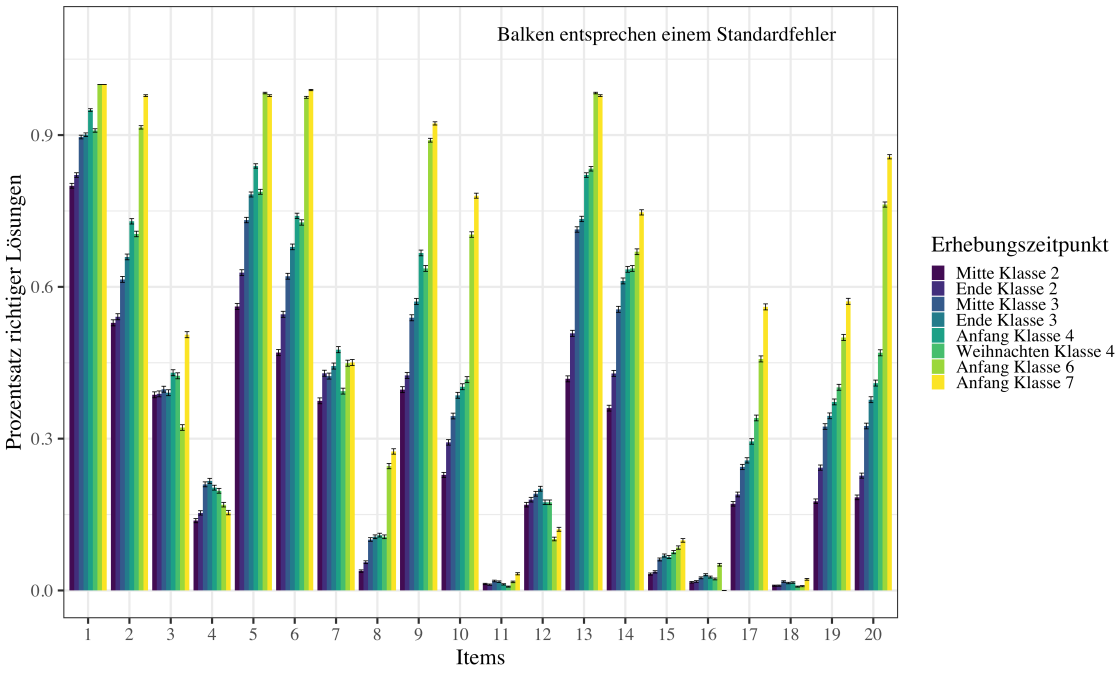


Abbildung 1: BYLET-C: Lösungshäufigkeit in Abhängigkeit des Alters

Lösungshäufigkeit in Abhängigkeit des Geschlechts BYLET-A

Tabelle 13: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
männlich	0.72	0.69	0.59	0.14	0.74
weiblich	0.71	0.71	0.6	0.15	0.76

Tabelle 14: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 6-10

	Item 6	Item 7	Item 8	Item 9	Item 10
männlich	0.48	0.28	0.13	0.48	0.35
weiblich	0.48	0.3	0.14	0.49	0.34

Tabelle 15: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
männlich	0.17	0.36	0.47	0.37	0.11
weiblich	0.19	0.34	0.47	0.34	0.11

Tabelle 16: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
männlich	0.13	0.39	0.08	0.25	0.2
weiblich	0.12	0.36	0.07	0.2	0.18

Lösungshäufigkeit in Abhängigkeit des Geschlechts BYLET-B

Tabelle 17: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
männlich	0.46	0.67	0.56	0.37	0.59
weiblich	0.48	0.66	0.63	0.32	0.61

Tabelle 18: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 6-10

	Item 6	Item 7	Item 8	Item 9	Item 10
männlich	0.5	0.3	0.08	0.66	0.46
weiblich	0.52	0.31	0.1	0.69	0.46

Tabelle 19: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
männlich	0.03	0.33	0.4	0.34	0.1
weiblich	0.04	0.35	0.37	0.32	0.09

Tabelle 20: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
männlich	0.05	0.14	0.03	0.15	0.16
weiblich	0.05	0.11	0.03	0.13	0.15

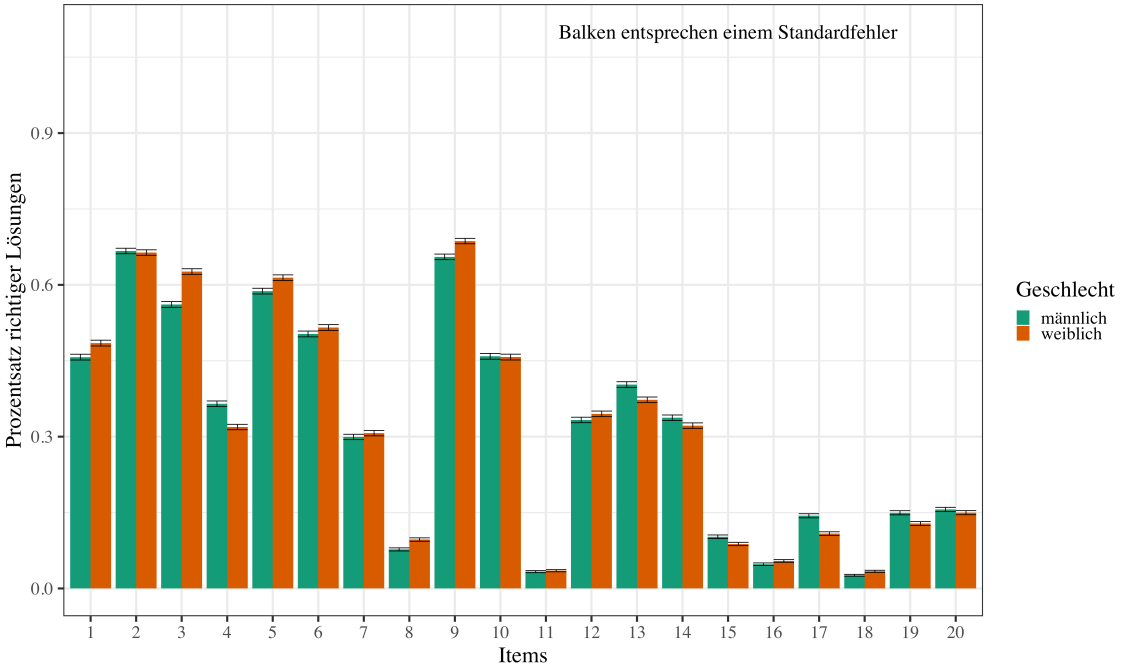


Abbildung 2: BYLET-B: Lösungshäufigkeit in Abhängigkeit des Geschlechts

Lösungshäufigkeit in Abhängigkeit des Geschlechts BYLET-C

Tabelle 21: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
männlich	0.84	0.58	0.4	0.16	0.68
weiblich	0.89	0.64	0.4	0.2	0.72

Tabelle 22: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 6-10

	Item 6	Item 7	Item 8	Item 9	Item 10
männlich	0.6	0.42	0.08	0.49	0.37
weiblich	0.62	0.43	0.09	0.54	0.31

Tabelle 23: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
männlich	0.01	0.18	0.61	0.5	0.05
weiblich	0.01	0.18	0.63	0.5	0.06

Tabelle 24: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Geschlechts Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
männlich	0.03	0.23	0.01	0.3	0.3
weiblich	0.02	0.24	0.01	0.28	0.3

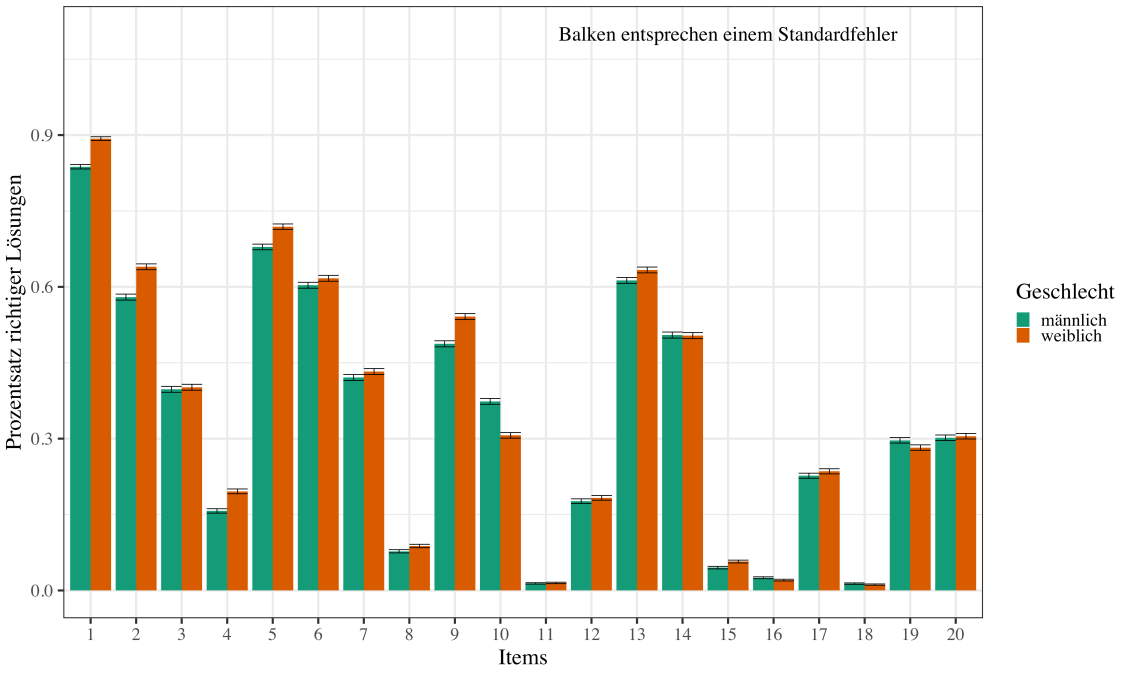


Abbildung 3: BYLET-C: Lösungshäufigkeit in Abhängigkeit des Geschlechts

Lösungshäufigkeit in Abhängigkeit des Migrationshintergrunds BYLET-A

Tabelle 25: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
kein Migrationshintergrund	0.72	0.72	0.62	0.15	0.76
Migrationshintergrund	0.68	0.59	0.48	0.12	0.71

Tabelle 26: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 6-10

	Item 6	Item 7	Item 8	Item 9	Item 10
kein Migrationshintergrund	0.49	0.3	0.14	0.51	0.35
Migrationshintergrund	0.42	0.24	0.13	0.38	0.27

Tabelle 27: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
kein Migrationshintergrund	0.19	0.37	0.49	0.36	0.11
Migrationshintergrund	0.15	0.27	0.34	0.28	0.08

Tabelle 28: BYLET-A: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
kein Migrationshintergrund	0.13	0.38	0.07	0.23	0.19
Migrationshintergrund	0.08	0.35	0.07	0.23	0.18

Lösungshäufigkeit in Abhängigkeit des Migrationshintergrunds BYLET-B

Tabelle 29: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
kein Migrationshintergrund	0.48	0.68	0.61	0.35	0.62
Migrationshintergrund	0.41	0.59	0.5	0.26	0.51

Tabelle 30: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 6-10

	Item 6	Item 7	Item 8	Item 9	Item 10
kein Migrationshintergrund	0.53	0.32	0.09	0.69	0.47
Migrationshintergrund	0.34	0.19	0.08	0.59	0.36

Tabelle 31: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
kein Migrationshintergrund	0.03	0.36	0.4	0.34	0.1
Migrationshintergrund	0.04	0.22	0.3	0.26	0.07

Tabelle 32: BYLET-B: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
kein Migrationshintergrund	0.05	0.13	0.03	0.14	0.16
Migrationshintergrund	0.04	0.12	0.02	0.15	0.12

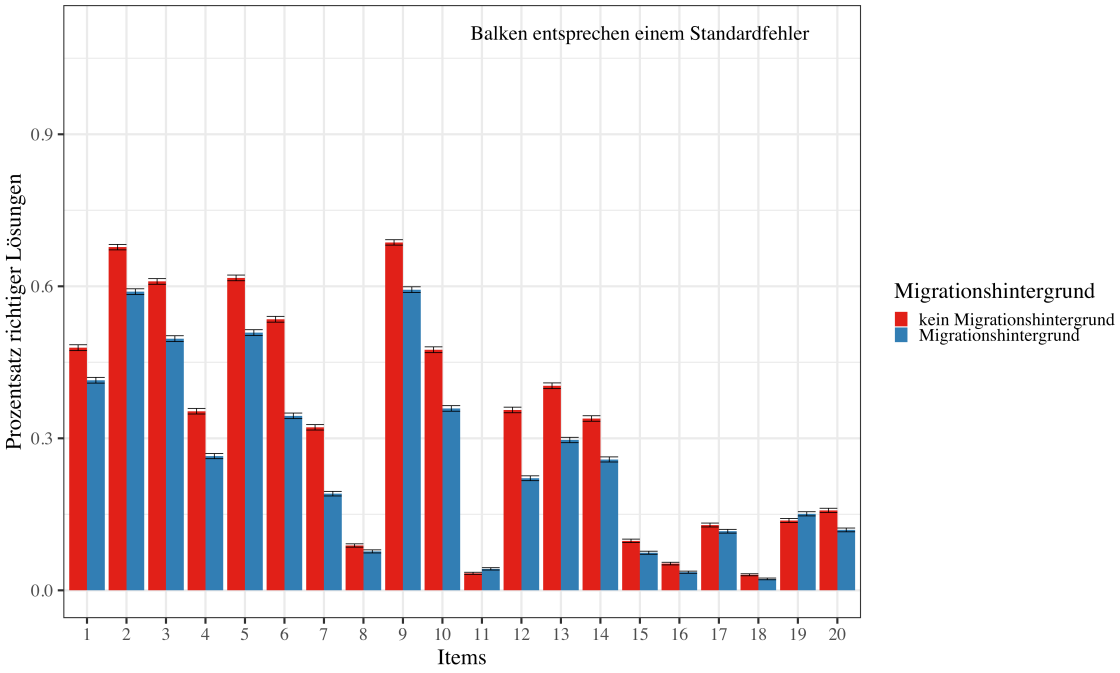


Abbildung 4: BYLET-B: Lösungshäufigkeit in Abhängigkeit des Migrationshintergrunds

Lösungshäufigkeit in Abhängigkeit des Migrationshintergrunds BYLET-C

Tabelle 33: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 1-5

	Item 1	Item 2	Item 3	Item 4	Item 5
kein Migrationshintergrund	0.88	0.63	0.4	0.18	0.71
Migrationshintergrund	0.75	0.47	0.38	0.13	0.61

Tabelle 34: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 6-10

	Item 6	Item 7	Item 8	Item 9	Item 10
kein Migrationshintergrund	0.63	0.44	0.08	0.53	0.35
Migrationshintergrund	0.48	0.34	0.06	0.39	0.27

Tabelle 35: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 11-15

	Item 11	Item 12	Item 13	Item 14	Item 15
kein Migrationshintergrund	0.01	0.18	0.65	0.52	0.05
Migrationshintergrund	0.01	0.21	0.49	0.43	0.04

Tabelle 36: BYLET-C: Itemlösungshäufigkeiten in Abhängigkeit des Migrationshintergrunds Items 16-20

	Item 16	Item 17	Item 18	Item 19	Item 20
kein Migrationshintergrund	0.02	0.24	0.01	0.3	0.31
Migrationshintergrund	0.01	0.18	0.02	0.23	0.24

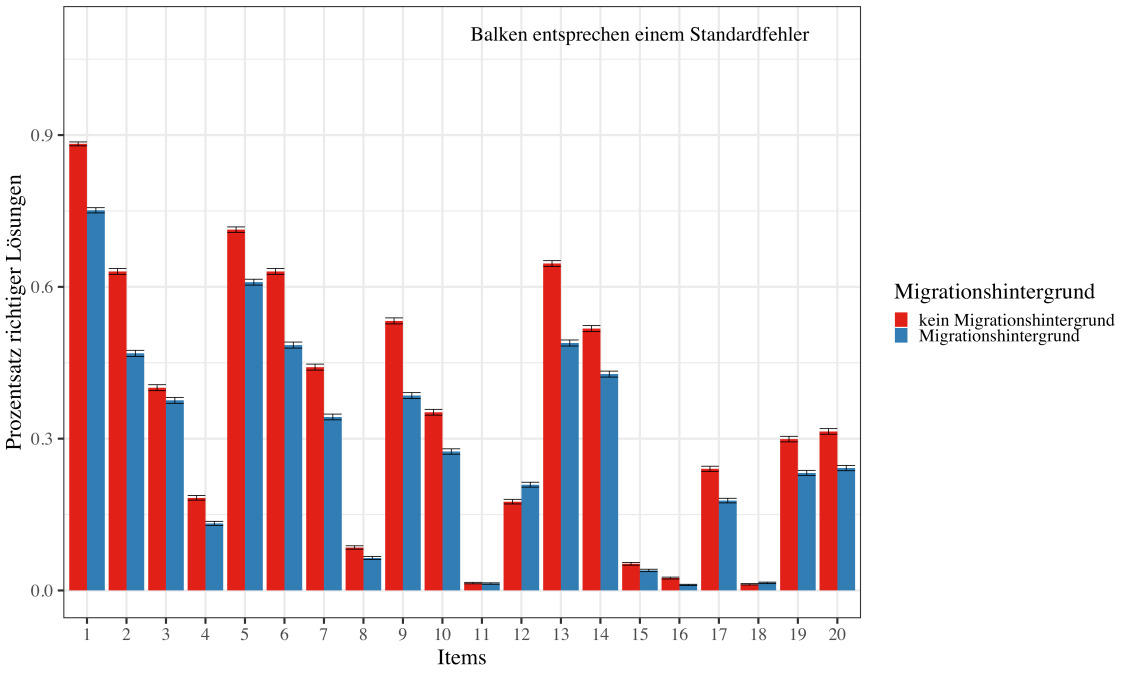


Abbildung 5: BYLET-C: Lösungshäufigkeit in Abhängigkeit des Migrationshintergrunds

Itemkorrelationen mit Hinblick auf die angenommene Faktorstruktur BYLET-B

Tabelle 37: Vergleich der Mediane der Itemkorrelationen der Textschwierigkeitsfaktoren des BYLET-B

Itemgruppen	Abschnitt A	Abschnitt B	Abschnitt C	Abschnitt D
$MD_{innerhalb}$	0.27	0.26	0.27	0.33
$MD_{außerhalb}$	0.23	0.24	0.23	0.24

Anmerkung. Vergleich der Mediane (MD) der Itemkorrelationen, innerhalb und außerhalb, der durch die Faktoren bestimmten Itemgruppen des BYLET-B.

Tabelle 38: Vergleich der Mediane der Itemkorrelationen der Kompetenzfaktoren des BYLET-B

Itemgruppen	Stufe II	Stufe III	Stufe IV	Dekodieren	Sprachverständnis
$MD_{innerhalb}$	0.34	0.26	0.18	0.34	0.22
$MD_{außerhalb}$	0.25	0.23	0.21	0.25	0.25

Anmerkung. Vergleich der Mediane (MD) der Itemkorrelationen, innerhalb und außerhalb, der durch die Faktoren bestimmten Itemgruppen des BYLET-B.

Itemkorrelationen mit Hinblick auf die angenommene Faktorstruktur BYLET-C

Tabelle 39: Vergleich der Mediane der Itemkorrelationen der Textschwierigkeitsfaktoren des BYLET-C

Itemgruppen	Abschnitt A	Abschnitt B	Abschnitt C	Abschnitt D
$MD_{innerhalb}$	0.19	0.15	-0.01	0.21
$MD_{außerhalb}$	0.13	0.13	0.11	0.13

Anmerkung. Vergleich der Mediane (MD) der Itemkorrelationen, innerhalb und außerhalb, der durch die Faktoren bestimmten Itemgruppen des BYLET-C.

Tabelle 40: Vergleich der Mediane der Itemkorrelationen der Kompetenzfaktoren des BYLET-C

Itemgruppen	Stufe II	Stufe III	Stufe IV	Dekodieren	Sprachverständnis
$MD_{innerhalb}$	0.47	0.12	0.01	0.47	0.09
$MD_{außerhalb}$	0.20	0.13	0.09	0.20	0.20

Anmerkung. Vergleich der Mediane (MD) der Itemkorrelationen, innerhalb und außerhalb, der durch die Faktoren bestimmten Itemgruppen des BYLET-C.

Modellwahl per Kreuzvalidierung der MIRT-Modellierung

BYLET-B

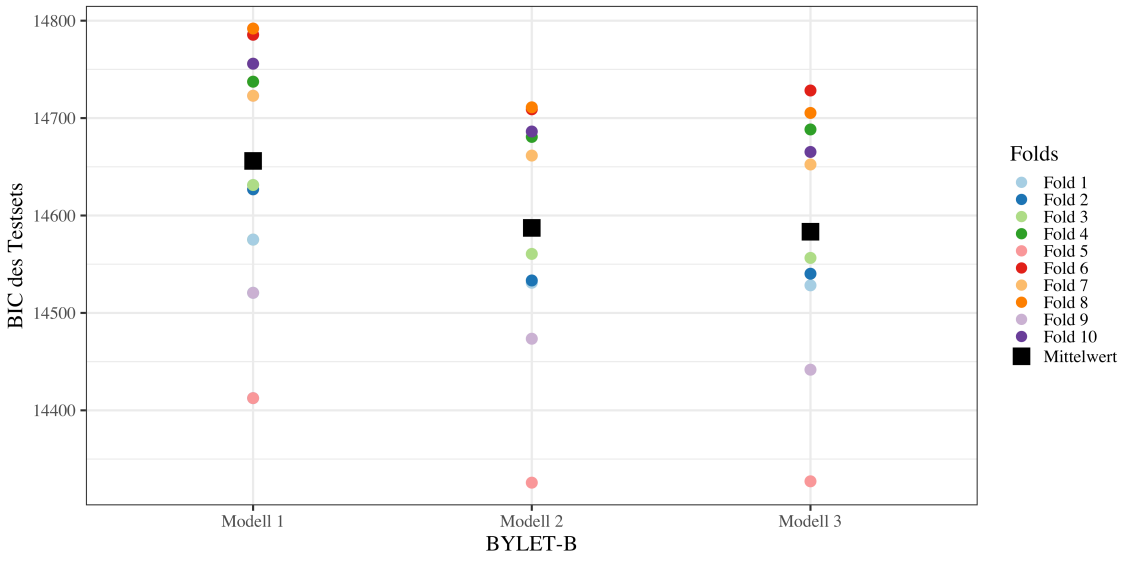


Abbildung 6: Modellwahl MIRT

BYLET-C

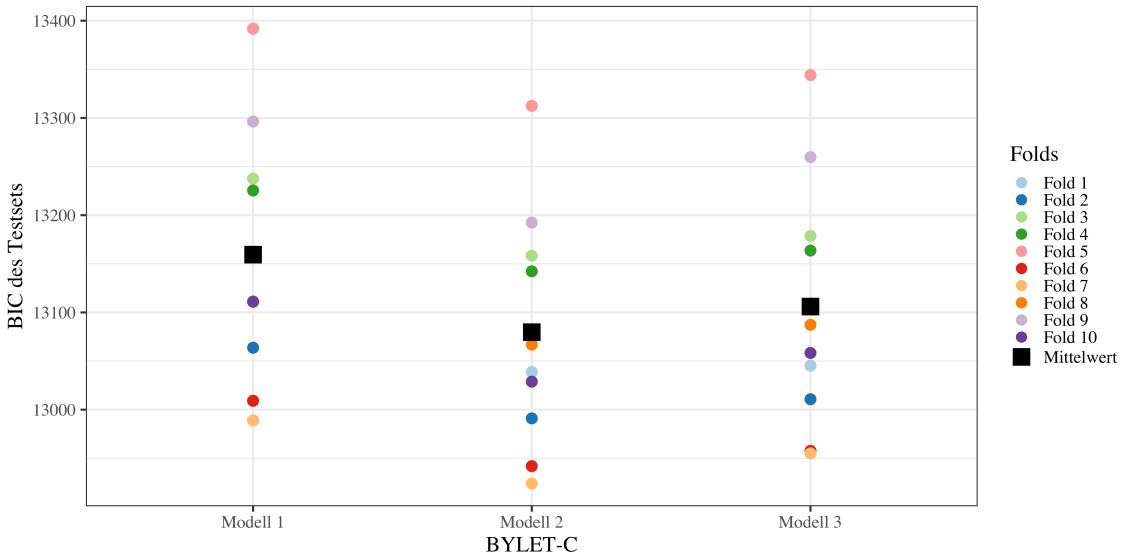


Abbildung 7: Modellwahl MIRT

Kennwerte der finalen MIRT-Modelle

BYLET-B

Tabelle 41: Fitstatistiken BYLET-B

BIC	Loglikelihood	df	p	RMSEA	CFI
143004	-71238	151	< 0.001	0.03	0.95

Anmerkung. Kennwerte des finalen MIRT-Modells für den BYLET-B; df = Freiheitsgrade; $p = p$-Wert.

Tabelle 42: Personenparameterkorrelationen BYLET-B

	II	III	IV	V
II	1	0.67	0.34	-
III	0.67	1	0.32	-
IV	0.34	0.32	1	-
V	-	-	-	1

Anmerkung. Römische Ziffern beziehen sich auf die Personenparameter der jeweiligen Kompetenzstufe.

Tabelle 43: Ladungsmatrix BYLET-B

	II	III	IV	V	Kommunalität
Item 1	0.4	-	-	0.39	0.31
Item 2	-	0.32	-	0.46	0.31
Item 3	-	0.24	-	0.49	0.29
Item 4	-	-	0.14	0.41	0.19
Item 5	0.38	-	-	0.58	0.47
Item 6	-	0.38	-	0.68	0.61
Item 7	-	0.15	-	0.42	0.20
Item 8	-	-	0.24	0.34	0.18
Item 9	0.26	-	-	0.62	0.45

Tabelle 43: Ladungsmatrix BYLET-B

	II	III	IV	V	Kommunalität
Item 10	-	0.07	-	0.51	0.26
Item 11	-	0.25	-	0.34	0.18
Item 12	-	-	0.11	0.53	0.30
Item 13	-0.25	-	-	0.72	0.59
Item 14	-	-0.23	-	0.68	0.52
Item 15	-	0.13	-	0.53	0.30
Item 16	-	-	0.21	0.41	0.21
Item 17	-	-	-	0.44	0.19
Item 18	-	-	-	0.34	0.12
Item 19	-	-	-	0.29	0.08
Item 20	-	-	-	0.53	0.28

Anmerkung. Römische Ziffern bezeichnen die Kompetenzstufen.

BYLET-C

Tabelle 44: Fitstatistiken BYLET-C

BIC	Loglikelihood	df	p	RMSEA	CFI
128057	-63767	151	< 0.001	0.03	0.94

Anmerkung. Kennwerte des finalen MIRT-Modells für den BYLET-B; df = Freiheitsgrade; $p = p$-Wert.

Tabelle 45: Personenparameterkorrelationen BYLET-C

	II	III	IV	V
II	1	0.64	0.31	-
III	0.64	1	0.35	-
IV	0.31	0.35	1	-
V	-	-	-	1

Anmerkung. Römische Ziffern beziehen sich auf die Personenparameter der jeweiligen Kompetenzstufe.

Tabelle 46: Ladungsmatrix BYLET-C

	II	III	IV	V	Kommunalität
Item 1	0.51	-	-	0.50	0.51
Item 2	-	0.48	-	0.45	0.43
Item 3	-	0.1	-	0.18	0.04
Item 4	-	-	0.35	0.16	0.15
Item 5	0.51	-	-	0.48	0.5
Item 6	-	0.4	-	0.54	0.45
Item 7	-	0.02	-	0.24	0.06
Item 8	-	-	0.5	0.20	0.29
Item 9	0.31	-	-	0.58	0.43
Item 10	-	0.1	-	0.42	0.19
Item 11	-	-0.05	-	-0.13	0.02
Item 12	-	-	-0.2	0.21	0.08
Item 13	0.14	-	-	0.83	0.7
Item 14	-	-0.14	-	0.73	0.56
Item 15	-	0.32	-	0.23	0.16
Item 16	-	-	0.29	0.06	0.09
Item 17	-	-	-	0.50	0.25
Item 18	-	-	-	-0.03	-
Item 19	-	-	-	0.49	0.24
Item 20	-	-	-	0.61	0.37

Anmerkung. Römische Ziffern bezeichnen die Kompetenzstufen.

Modellwahl per Kreuzvalidierungder SEM Modellierung

BYLET-B

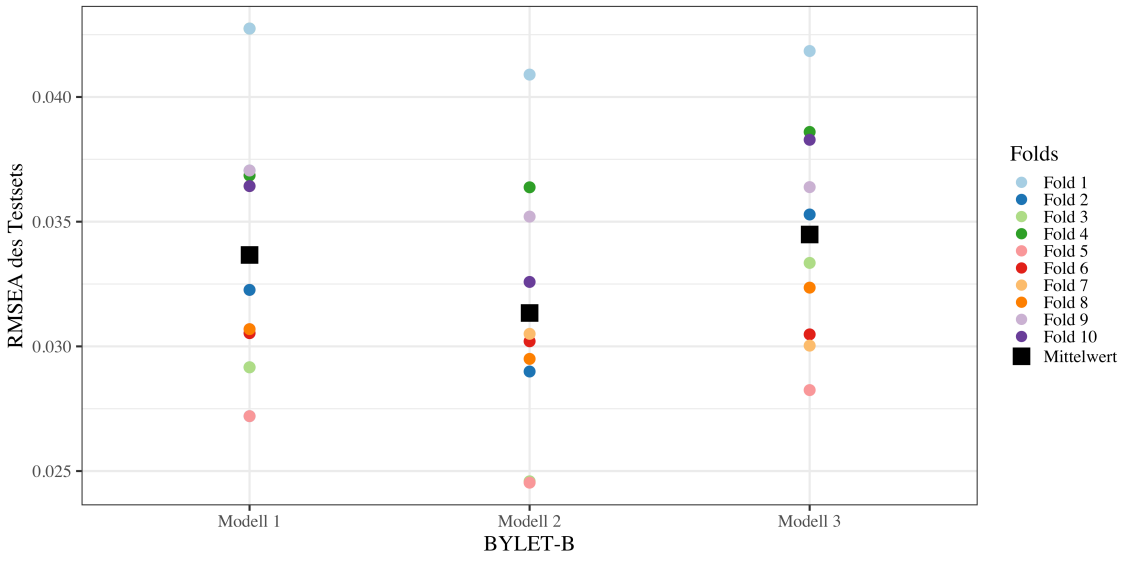


Abbildung 8: Modellwahl SEM RMSEA

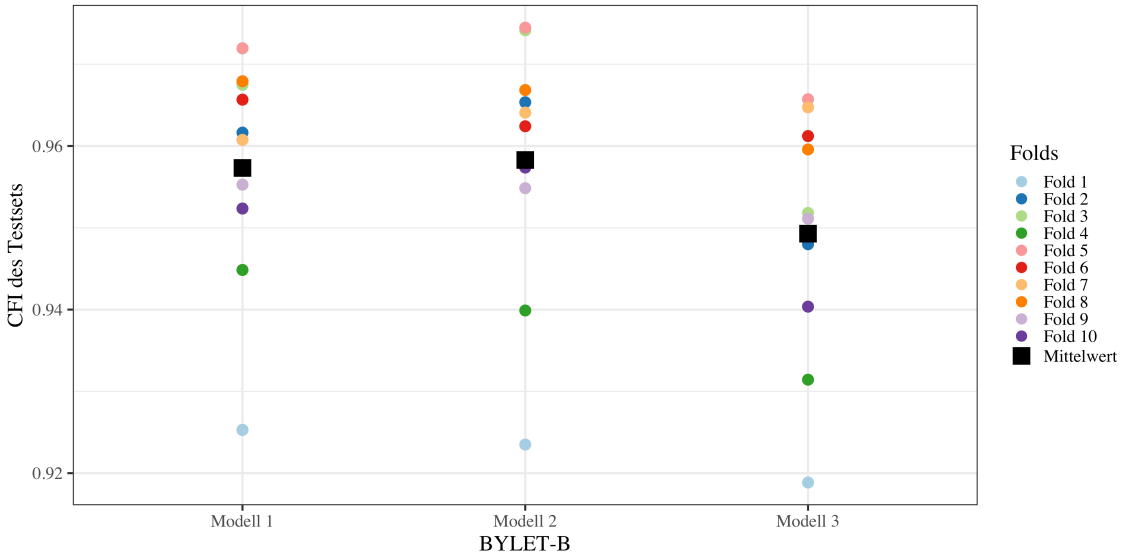


Abbildung 9: Modellwahl SEM CFI

BYLET-C

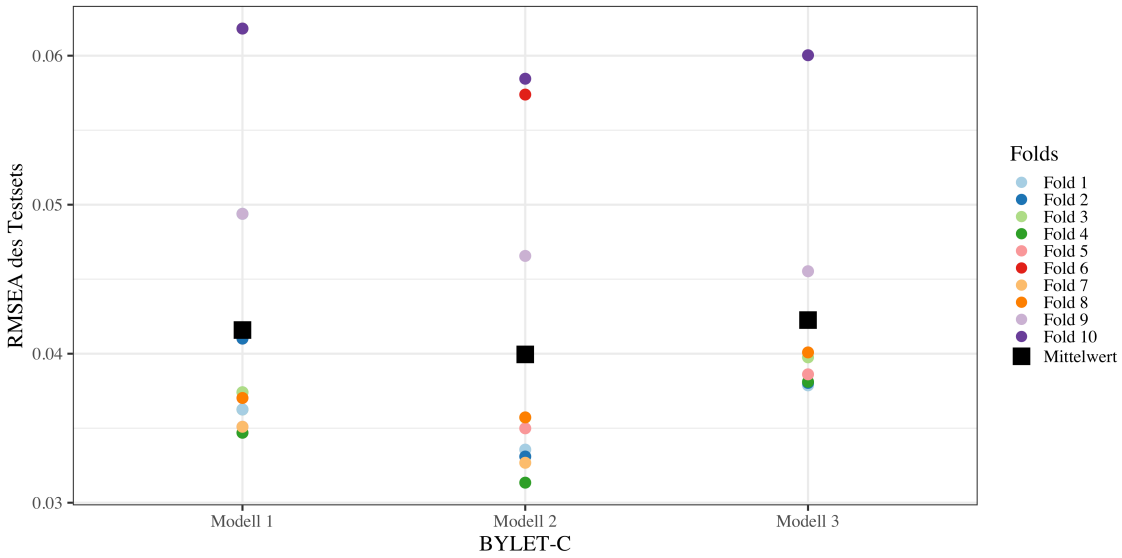


Abbildung 10: Modellwahl SEM RMSEA

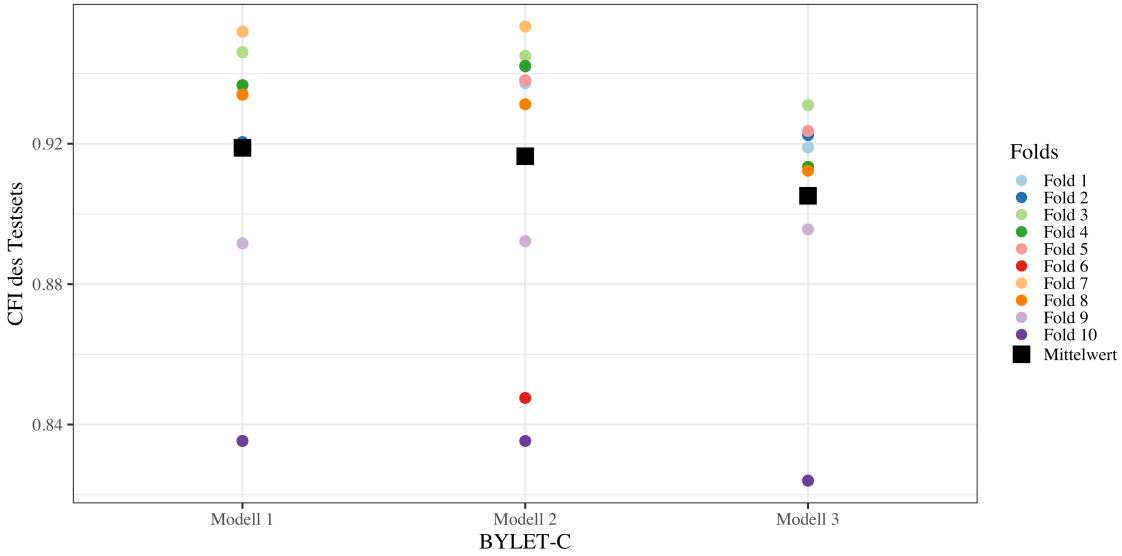


Abbildung 11: Modellwahl SEM CFI

Kennwerte der finalen SEM-Modelle

BYLET-B

Tabelle 47: Modellkennwerte des finalen Modells mit WLSMV-Schätzung

χ^2	df	p-Wert
609.2	151	< 0.001

Anmerkung.

df = Freiheitsgrade.

Tabelle 48: Ladungsmatrix BYLET-B

	II	III	IV	V
Item 1	0.49	-	-	0.23
Item 2	-	0.45	-	0.31
Item 3	-	0.37	-	0.39
Item 4	-	-	0.18	0.40
Item 5	0.52	-	-	0.40
Item 6	-	0.57	-	0.50
Item 7	-	0.28	-	0.34
Item 8	-	-	0.18	0.28
Item 9	0.46	-	-	0.46
Item 10	-	0.29	-	0.43
Item 11	-	0.23	-	0.21
Item 12	-	-	0.27	0.47
Item 13	0.02	-	-	0.77
Item 14	-	0.06	-	0.68
Item 15	-	0.22	-	0.46
Item 16	-	-	0.11	0.36
Item 17	-	-	-	0.46
Item 18	-	-	-	0.31
Item 19	-	-	-	0.30

Tabelle 48: Ladungsmatrix BYLET-B

	II	III	IV	V
Item 20	-	-	-	0.57

Anmerkung. Römische Ziffern bezeichnen die Kompetenzstufen.

Tabelle 49: Faktorkorrelationen BYLET-B

	II	III	IV	V
II	1.00	1.07	0.95	0
III	1.07	1.00	1.04	0
IV	0.95	1.04	1.00	0
V	0.00	0.00	0.00	1

Anmerkung. Römische Ziffern beziehen sich auf die Faktorwerte der jeweiligen Kompetenzstufe. Korrelationen > 1 entstehen durch Schätzfehler.

BYLET-C

Tabelle 50: Modellkennwerte des finalen Modells mit WLSMV-Schätzung

χ^2	df	p-Wert
639.09	151	< 0.001

Anmerkung.

df = Freiheitsgrade.

Tabelle 51: Ladungsmatrix BYLET-C

	II	III	IV	V
Item 1	0.51	-	-	0.43
Item 2	-	0.49	-	0.37
Item 3	-	0.17	-	0.15
Item 4	-	-	0.43	0.14

Tabelle 51: Ladungsmatrix BYLET-C

	II	III	IV	V
Item 5	0.52	-	-	0.40
Item 6	-	0.49	-	0.44
Item 7	-	0.12	-	0.22
Item 8	-	-	0.51	0.16
Item 9	0.44	-	-	0.50
Item 10	-	0.16	-	0.40
Item 11	-	-0.07	-	-0.07
Item 12	-	-	-0.16	0.22
Item 13	0.25	-	-	0.81
Item 14	-	0.06	-	0.70
Item 15	-	0.18	-	0.17
Item 16	-	-	0.19	0.04
Item 17	-	-	-	0.51
Item 18	-	-	-	-0.03
Item 19	-	-	-	0.50
Item 20	-	-	-	0.64

Anmerkung. Römische Ziffern bezeichnen die Kompetenzstufen.

Tabelle 52: Faktorkorrelationen BYLET-C

	II	III	IV	V
II	1.00	1.05	0.35	0
III	1.05	1.00	0.24	0
IV	0.35	0.24	1.00	0
V	0.00	0.00	0.00	1

Anmerkung. Römische Ziffern beziehen sich auf die Faktorwerte der jeweiligen Kompetenzstufe.Korrelationen > 1 entstehen durch Schätzfehler.

Modifikationsindizes der finalen SEM-Modelle

BYLET-B

Tabelle 53: Modifikationsindizes des finalen Modells des BYLET-B

Finales Modell					Modifiziertes Modell (II=~ i6)				
Parameterset	mi	epc	power	dec	Parameterset	mi	epc	power	dec
II=~i6	184.73	0.54	1	r.	i9~~i10	75.23	0.16	1.00	n.r.
IV=~i6	180.29	0.55	1	r.	III=~i5	58.69	-2.21	0.18	r.
i5~~i6	140.43	0.23	1	n.r.	i5~~i6	58.17	0.18	1.00	n.r.
III=~i5	137.96	0.45	1	r.	i17~~i19	34.24	0.15	1.00	n.r.
IV=~i5	119.07	0.43	1	r.	i15~~i16	31.76	0.19	1.00	n.r.
i9~~i10	94.24	0.18	1	n.r.	IV=~i5	30.79	-5.76	0.06	r.
III=~i9	89.21	0.37	1	r.	IV=~i18	28.21	0.21	1.00	n.r.
III=~i1	88.97	0.36	1	r.	III=~i18	28.02	0.21	1.00	n.r.
IV=~i1	81.42	0.35	1	r.	II=~i18	27.86	0.21	1.00	n.r.
IV=~i9	80.15	0.36	1	r.	i1~~i2	25.52	0.11	1.00	n.r.
II=~i2	65.93	0.32	1	r.	II=~i19	25.20	-0.13	1.00	n.r.
IV=~i2	61.67	0.32	1	r.	i15~~i8	25.10	0.15	1.00	n.r.

Anmerkung. mi = Modifikationsindex; epc = erwarteter Parameterwert; delta (Grenzwert der Parameterrelevanz) = 0.3; dec = Entscheidung; n.r. = nicht relevant; r. = relevant; i. = indifferent.

Tabelle 54: Fit-Indizes nach Modifikation des finalen Modells des BYLET-B

Modell	RMSEA	CFI
Finales Modell	0.0200	0.9854
Modifiziertes Modell	0.0201	0.9854

Anmerkung. RMSEA = Root mean squared error of approximation; CFI = Comparative Fit Index; Angabe auf vier Nachkommastellen genau, um geringe Unterschiedlichkeit zu verdeutlichen.

BYLET-C

Tabelle 55: Modifikationsindizes des finalen Modells des BYLET-C

Finales Modell					Modifiziertes Modell 1 (IV=~i15)				
Parameterset	mi	epc	power	dec	Parameterset	mi	epc	power	dec
II=~i2	58.71	0.34	1.00	r.	IV=~i2	70.39	0.34	1.00	r.
II=~i3	0.64	0.04	1.00	n.r.	II=~i2	67.32	0.36	1.00	r.
II=~i6	52.13	0.32	1.00	r.	IV=~i13	57.63	0.28	1.00	n.r.
II=~i7	1.30	0.05	1.00	n.r.	III=~i5	56.13	0.38	1.00	r.
II=~i10	4.45	0.10	1.00	n.r.	II=~i6	54.69	0.32	1.00	r.
II=~i11	0.10	0.04	0.70	i.	IV=~i3	47.12	-0.22	1.00	n.r.
II=~i14	0.36	0.03	1.00	n.r.	i3~~i4	45.57	-0.15	1.00	n.r.
II=~i15	9.50	0.24	0.97	n.r.	i13~~i14	43.42	0.17	1.00	n.r.
II=~i4	1.75	0.10	0.98	n.r.	III=~i9	35.34	0.28	1.00	n.r.
II=~i8	1.08	0.09	0.92	n.r.	IV=~i11	31.21	0.40	0.99	r.
II=~i12	11.02	0.12	1.00	n.r.	i19~~i20	31.19	0.13	1.00	n.r.
II=~i16	0.09	0.02	1.00	n.r.	i9~~i6	26.71	0.11	1.00	n.r.
II=~i17	0.27	0.02	1.00	n.r.	III=~i1	23.97	0.29	1.00	n.r.

Anmerkung. mi = Modifikationsindex; epc = erwarteter Parameterwert; delta (Grenzwert der Parameterrelevanz) = 0.3; dec = Entscheidung; n.r. = nicht relevant; r. = relevant; i. = indifferent.

Tabelle 56: Modifikationsindizes des finalen Modells des BYLET-C

Modifiziertes Modell 2 (IV=~i2)					Modifiziertes Modell 3 (II=~i6)				
Parameterset	mi	epc	power	dec	Parameterset	mi	epc	power	dec
II=~i2	0.00	0.00	0.06	i.	II=~i14	8.51	-0.11	1.00	n.r.
II=~i3	32.88	-0.43	0.98	n.r.	II=~i15	1.72	0.38	0.18	i.
II=~i6	52.13	0.92	0.65	r.	II=~i8	2.33	0.06	1.00	n.r.
II=~i7	9.06	-0.21	0.99	n.r.	II=~i12	8.97	0.09	1.00	n.r.
II=~i10	14.77	0.28	0.98	n.r.	II=~i16	0.78	-0.05	1.00	n.r.
II=~i11	25.31	0.82	0.45	r.	II=~i17	0.08	0.01	1.00	n.r.

II=~i14	17.73	-0.33	0.97	n.r.	II=~i18	7.40	-0.17	1.00	n.r.
II=~i15	3.12	1.56	0.06	i.	II=~i19	17.15	-0.12	1.00	n.r.
II=~i4	0.16	0.02	1.00	n.r.	II=~i20	0.19	0.01	1.00	n.r.
II=~i8	3.39	0.08	1.00	n.r.	III=~i1	7.38	-0.13	1.00	n.r.
II=~i12	9.67	0.09	1.00	n.r.	III=~i5	26.23	0.24	1.00	n.r.
II=~i16	0.62	-0.04	1.00	n.r.	III=~i9	0.67	-0.03	1.00	n.r.
II=~i17	0.90	0.03	1.00	n.r.	III=~i13	19.41	-0.18	1.00	n.r.

Anmerkung. mi = Modifikationsindex; epc = erwarteter Parameterwert; delta (Grenzwert der Parameterrelevanz) = 0.3; dec = Entscheidung; n.r. = nicht relevant; r. = relevant; i. = indifferent.

Tabelle 57: Fit-Indizes nach Modifikation des finalen Modells des BYLET-C

Modell	RMSEA	CFI
Finales Modell	0.0215	0.9796
Modifiziertes Modell 1	0.0192	0.9838
Modifiziertes Modell 2	0.0179	0.9861
Modifiziertes Modell 3	0.0215	0.9886

Anmerkung. RMSEA = Root mean squared error of approximation; CFI = Comparative Fit Index; Angabe auf vier Nachkommastellen genau, um geringe Unterschiedlichkeit zu verdeutlichen.

Bestimmung der Hauptgütekriterien

Reliabilität BYLET-B

Tabelle 58: Reliabilitäten

	II	III	IV	V	Gesamtreliabilität
Ordinales Alpha	0.66	0.74	0.5	0.86	0.86

Anmerkung. Ordinales Alpha = Interne Konsistenz nach Zumbo und Kollegen (2007).

Reliabilität BYLET-C

Tabelle 59: Reliabilitäten

	II	III	IV	V	Gesamtreliabilität
Ordinales Alpha	0.77	0.51	0.08	0.77	0.77

Anmerkung. Ordinales Alpha = Interne Konsistenz nach Zumbo und Kollegen (2007).

Validitäten BYLET-B

Tabelle 60: Vergleich der Ladungen Stufe II und III

Items	II mirt	II sem	II sem p	III mirt	III sem	III sem p
1	0.4	0.49	< 0.001	-	-	-
2	-	-	-	0.32	0.45	< 0.001
3	-	-	-	0.24	0.37	< 0.001
4	-	-	-	-	-	-
5	0.38	0.52	< 0.001	-	-	-
6	-	-	-	0.38	0.57	< 0.001
7	-	-	-	0.15	0.28	< 0.001
8	-	-	-	-	-	-
9	0.26	0.46	< 0.001	-	-	-
10	-	-	-	0.07	0.29	< 0.001
11	-	-	-	0.25	0.23	< 0.001
12	-	-	-	-	-	-
13	-0.25	0.02	0.543	-	-	-
14	-	-	-	-0.23	0.06	0.035
15	-	-	-	0.13	0.22	< 0.001
16	-	-	-	-	-	-
17	-	-	-	-	-	-
18	-	-	-	-	-	-
19	-	-	-	-	-	-
20	-	-	-	-	-	-

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; p = p-Wert.

Tabelle 61: Vergleich der Ladungen Stufe IV und V

Items	IV mirt	IV sem	IV sem p	V mirt	V sem	V sem p	Kommunalität
1	-	-	-	0.39	0.23	< 0.001	0.31
2	-	-	-	0.46	0.31	< 0.001	0.31

3	-	-	-	0.49	0.39	< 0.001	0.29
4	0.14	0.18	< 0.001	0.41	0.40	< 0.001	0.19
5	-	-	-	0.58	0.40	< 0.001	0.47
6	-	-	-	0.68	0.50	< 0.001	0.61
7	-	-	-	0.42	0.34	< 0.001	0.20
8	0.24	0.18	< 0.001	0.34	0.28	< 0.001	0.18
9	-	-	-	0.62	0.46	< 0.001	0.45
10	-	-	-	0.51	0.43	< 0.001	0.26
11	-	-	-	0.34	0.21	< 0.001	0.18
12	0.11	0.27	< 0.001	0.53	0.47	< 0.001	0.30
13	-	-	-	0.72	0.77	< 0.001	0.59
14	-	-	-	0.68	0.68	< 0.001	0.52
15	-	-	-	0.53	0.46	< 0.001	0.30
16	0.21	0.11	0.005	0.41	0.36	< 0.001	0.21
17	-	-	-	0.44	0.46	< 0.001	0.19
18	-	-	-	0.34	0.31	< 0.001	0.12
19	-	-	-	0.29	0.30	< 0.001	0.08
20	-	-	-	0.53	0.57	< 0.001	0.28

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; p = p-Wert.

Tabelle 62: Korrelation der BYLET-B-Faktorwerte mit dem Salzburger Lesescreening

BYLET	SLS					
Stufe	Mitte Kl.2	Ende Kl.2	Mitte Kl.3	Ende Kl.3	Anfang Kl.4	Weihnachten Kl.4
IIsem	0.00	0.00	0.00	0.01	0.00	-0.15
IIIsem	0.00	0.00	0.00	-0.01	0.00	0.15
IVsem	0.00	0.01	0.00	0.01	0.00	-0.12
Vsem	0.00	0.01	0.04	-0.01	-0.04	0.05
IImirt	0.13	0.12	0.12	0.12	0.10	0.04
IIImirt	0.14	0.13	0.14	0.14	0.14	0.09

IVmirt	0.16	0.16	0.17	0.16	0.15	0.16
Vmirt	0.44	0.43	0.42	0.42	0.40	0.27

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; SLS = Salzburger Lesescreening; Kl. = Klasse., Stufen sind Kompetenzstufen des Bayerischen Lesetests

Tabelle 63: Korrelation der BYLET-B-Faktorwerte mit den Schulnoten

	Deutschnote	Mathenote
IIsem	-0.03	-0.04
IIIsem	0.03	0.04
IVsem	-0.03	-0.04
Vsem	0	0.01
IImirt	-0.21	-0.2
IIImirt	-0.23	-0.22
IVmirt	-0.21	-0.21
Vmirt	-0.41	-0.34
Deutschnote	1	
Mathenote	0.69	1

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; Schulnoten reichen von 1 bis 6. 1 ist die beste Note.

Validitäten BYLET-C

Tabelle 64: Vergleich der Ladungen Stufe II und III

Items	II mirt	II sem	II sem p	III mirt	III sem	III sem p
1	0.51	0.51	< 0.001	-	-	-
2	-	-	-	0.48	0.49	< 0.001
3	-	-	-	0.1	0.17	< 0.001
4	-	-	-	-	-	-

5	0.51	0.52	< 0.001	-	-	-
6	-	-	-	0.4	0.49	< 0.001
7	-	-	-	0.02	0.12	< 0.001
8	-	-	-	-	-	-
9	0.31	0.44	< 0.001	-	-	-
10	-	-	-	0.1	0.16	< 0.001
11	-	-	-	-0.05	-0.07	0.302
12	-	-	-	-	-	-
13	0.14	0.25	< 0.001	-	-	-
14	-	-	-	-0.14	0.06	0.029
15	-	-	-	0.32	0.18	< 0.001
16	-	-	-	-	-	-
17	-	-	-	-	-	-
18	-	-	-	-	-	-
19	-	-	-	-	-	-
20	-	-	-	-	-	-

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; p = p-Wert.

Tabelle 65: Vergleich der Ladungen Stufe IV und V

Items	IV mirt	IV sem	IV sem p	V mirt	V sem	V sem p	Kommunalität
1	-	-	-	0.50	0.43	< 0.001	0.51
2	-	-	-	0.45	0.37	< 0.001	0.43
3	-	-	-	0.18	0.15	< 0.001	0.04
4	0.35	0.43	< 0.001	0.16	0.14	< 0.001	0.15
5	-	-	-	0.48	0.40	< 0.001	0.5
6	-	-	-	0.54	0.44	< 0.001	0.45
7	-	-	-	0.24	0.22	< 0.001	0.06
8	0.5	0.51	< 0.001	0.20	0.16	< 0.001	0.29
9	-	-	-	0.58	0.50	< 0.001	0.43
10	-	-	-	0.42	0.40	< 0.001	0.19

11	-	-	-	-0.13	-0.07	0.155	0.02
12	-0.2	-0.16	< 0.001	0.21	0.22	< 0.001	0.08
13	-	-	-	0.83	0.81	< 0.001	0.7
14	-	-	-	0.73	0.70	< 0.001	0.56
15	-	-	-	0.23	0.17	< 0.001	0.16
16	0.29	0.19	0.018	0.06	0.04	0.328	0.09
17	-	-	-	0.50	0.51	< 0.001	0.25
18	-	-	-	-0.03	-0.03	0.544	-
19	-	-	-	0.49	0.50	< 0.001	0.24
20	-	-	-	0.61	0.64	< 0.001	0.37

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; p = p-Wert.

Tabelle 66: Korrelation der BYLET-C-Faktorwerte mit dem Salzburger Lesescreening

BYLET	SLS					
Stufe	Mitte Kl.2	Ende Kl.2	Mitte Kl.3	Ende Kl.3	Anfang Kl.4	Weihnachten Kl.4
IIsem	0.02	0.01	0.00	0.03	0.01	0.07
IIIsem	-0.02	-0.01	0.00	-0.03	-0.01	-0.04
IVsem	0.02	-0.01	0.00	-0.03	-0.01	0.07
Vsem	0.00	0.00	0.01	-0.01	0.00	-0.08
IImirt	0.21	0.21	0.22	0.25	0.21	0.26
IIImirt	0.19	0.18	0.20	0.22	0.19	0.18
IVmirt	0.17	0.17	0.18	0.18	0.19	0.06
Vmirt	0.41	0.41	0.40	0.40	0.34	0.28

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; SLS = Salzburger Lesescreening; Kl. = Klasse., Stufen sind Kompetenzstufen des Bayerischen Lesetests

Tabelle 67: Korrelation der BYLET-C-Faktorwerte mit den Schulnoten

	Deutschnote	Mathenote
IIsem	0	-0.02
IIIsem	0	0.02
IVsem	0.01	0.02
Vsem	0.01	-0.01
IImirt	-0.29	-0.26
IIImirt	-0.27	-0.25
IVmirt	-0.18	-0.17
Vmirt	-0.36	-0.32
Deutschnote	1	
Mathenote	0.7	1

Anmerkung. mirt = Schätzung als MIRT-Modell; sem = Schätzung als Strukturgleichungsmodell; römische Ziffern bezeichnen die Kompetenzstufen; Schulnoten reichen von 1 bis 6. 1 ist die beste Note.

Anhang B

Itemschwierigkeiten des Salzburger Lese-Screenings

Tabelle 1: Itemschwierigkeiten nach SLS-Versionen

Items	A	B1	B2	C
Item 01	0.98	0.99	0.99	0.97
Item 02	0.99	0.96	0.99	0.97
Item 03	0.98	0.99	0.98	0.87
Item 04	0.96	0.98	0.97	0.83
Item 05	0.97	0.94	0.93	0.84
Item 06	0.97	0.98	0.99	0.95
Item 07	0.98	0.97	0.99	0.95
Item 08	0.97	0.98	0.97	0.92
Item 09	0.97	0.98	0.97	0.95
Item 10	0.96	0.85	0.98	0.94
Item 11	0.96	0.96	0.85	0.92
Item 12	0.95	0.97	0.96	0.90
Item 13	0.95	0.95	0.97	0.93
Item 14	0.96	0.93	0.95	0.84
Item 15	0.96	0.96	0.92	0.90
Item 16	0.95	0.92	0.94	0.88
Item 17	0.92	0.91	0.93	0.84
Item 18	0.89	0.91	0.92	0.53
Item 19	0.93	0.93	0.94	0.71
Item 20	0.91	0.89	0.90	0.74
Item 21	0.93	0.87	0.89	0.70
Item 22	0.84	0.78	0.87	0.65
Item 23	0.87	0.85	0.86	0.55
Item 24	0.83	0.80	0.76	0.57
Item 25	0.83	0.81	0.84	0.49
Item 26	0.81	0.79	0.81	0.40

Tabelle 1: Itemschwierigkeiten nach SLS-Versionen

Items	A	B1	B2	C
Item 27	0.78	0.77	0.76	0.43
Item 28	0.76	0.72	0.77	0.37
Item 29	0.72	0.70	0.74	0.35
Item 30	0.72	0.65	0.70	0.30
Item 31	0.64	0.59	0.64	0.29
Item 32	0.61	0.56	0.60	0.24
Item 33	0.58	0.52	0.57	0.24
Item 34	0.53	0.48	0.53	0.17
Item 35	0.43	0.44	0.41	0.15
Item 36	0.46	0.36	0.42	0.14
Item 37	0.42	0.37	0.39	0.13
Item 38	0.38	0.33	0.35	0.12
Item 39	0.36	0.29	0.30	0.11
Item 40	0.32	0.27	0.28	0.10
Item 41	0.29	0.22	0.24	0.08
Item 42	0.24	0.20	0.21	0.08
Item 43	0.23	0.18	0.20	0.06
Item 44	0.21	0.16	0.17	0.05
Item 45	0.18	0.14	0.15	0.05
Item 46	0.16	0.12	0.13	0.04
Item 47	0.14	0.10	0.11	0.04
Item 48	0.13	0.09	0.10	0.02
Item 49	0.11	0.08	0.08	0.02
Item 50	0.09	0.07	0.07	0.02
Item 51	0.08	0.06	0.06	0.02
Item 52	0.07	0.05	0.06	0.02
Item 53	0.06	0.05	0.05	0.02
Item 54	0.05	0.04	0.04	0.01
Item 55	0.04	0.04	0.04	0.01

Tabelle 1: Itemschwierigkeiten nach SLS-Versionen

Items	A	B1	B2	C
Item 56	0.04	0.03	0.03	0.01
Item 57	0.03	0.03	0.03	0.01
Item 58	0.03	0.03	0.03	0.01
Item 59	0.02	0.02	0.02	0.01
Item 60	0.02	0.02	0.02	0.01
Item 61	0.02	0.02	0.01	0.01
Item 62	0.02	0.01	0.01	0.01
Item 63	0.01	0.01	0.01	0.01
Item 64	0.01	0.01	0.01	0.01
Item 65	0.01	0.01	0.01	0.01
Item 66	0.01	0.01	0.01	0.01
Item 67	0.01	0.01	0.01	0.01
Item 68	0.01	0.01	0.01	0.01
Item 69	0.01	0.01	0.01	0.00
Item 70	0.01	0.01	0.01	0.00
Item 71	0.01	0.01	0.01	0.00
Item 72	0.00	0.01	0.01	0.00
Item 73	0.00	0.01	0.01	0.00
Item 74	0.00	0.01	0.01	0.00
Item 75	0.00	0.01	0.00	0.00
Item 76	0.00	0.01	0.00	0.00
Item 77	0.00	0.01	0.00	0.00
Item 78	0.00	0.01	0.00	0.00
Item 79	0.00	0.01	0.00	0.00
Item 80	0.00	0.01	0.00	0.00
Item 81	0.00	0.01	0.00	0.00
Item 82	0.00	0.01	0.00	0.00
Item 83	0.00	0.01	0.00	0.00
Item 84	0.00	0.00	0.00	0.00

Tabelle 1: Itemschwierigkeiten nach SLS-Versionen

Items	A	B1	B2	C
Item 85	0.00	0.01	0.00	0.00
Item 86	0.00	0.01	0.00	0.00
Item 87	0.00	0.00	0.00	0.00
Item 88	0.00	0.00	0.00	0.00
Item 89	0.00	0.01	0.00	0.00
Item 90	0.00	0.00	0.00	0.00
Item 91	0.00	0.00	0.00	0.00
Item 92	0.00	0.00	0.00	0.00
Item 93	0.00	0.00	0.00	0.00
Item 94	0.00	0.00	0.00	0.00
Item 95	0.00	0.00	0.00	0.00
Item 96	0.00	0.00	0.00	0.00
Item 97	0.00	0.00	0.00	0.00
Item 98	0.00	0.00	0.00	0.00
Item 99	0.00	0.00	0.00	0.00
Item 100	0.00	0.00	0.00	0.00

Anmerkung. SLS = Salzburger Lese-Screening.