Die Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraussetzungen von Schülerinnen und Schülern

Eine exemplarische Analyse am Beispiel der Qualität der Lernmotivation

Inauguraldissertation zur Erlangung des Grades eines Doktors der Philosophie im Fachbereich Erziehungswissenschaften der Johann-Wolfgang-Goethe-Universität zu Frankfurt am Main

vorgelegt von

Stephanie Leininger (geb. Musow)

2023

Erstgutachter: Prof. Dr. Eckhard Klieme

Zweitgutachter: Prof. Dr. Johannes Hartig

Tag der Disputation: 13.06.2023

Inhaltsverzeichnis

Zusammenfassung

Die vorliegende Dissertation hat die Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraussetzungen von Schülerinnen und Schülern zum Thema. Die Instruktionssensitivität von Items bzw. Testaufgaben erfasst, ob diese in der Lage sind, Effekte von Unterricht auf die Leistungen der Schülerinnen und Schüler abzubilden. Der Begriff der individuellen Lernvoraussetzungen wird im Rahmen dieser Arbeit sehr breit gefasst und subsumiert unter anderem kognitive, metakognitive, motivationale und volitionale Merkmale (Brühwiler, 2014; Brühwiler et al., 2017). Ausgehend von den Lernvoraussetzungen, welche im Zusammenhang von Angebots-Nutzungs-Modellen (u.a. Brühwiler, 2014; Fend, 1981) konzeptionell aufgegriffen werden, wird in den daran anschließenden quantitativen Analysen ein besonderer Fokus auf die Qualität der Lernmotivation (Ryan & Deci, 2000) gelegt. Das empirische Ziel der Arbeit besteht in der exemplarischen Überprüfung, inwiefern Indikatoren der Instruktionssensitivität von Testitems durch die Qualität der Lernmotivation der Schülerinnen und Schüler beeinflusst werden.

Vor dem Hintergrund dieses Erkenntnisinteresses werden Parameterschätzungen aus längsschnittlichen Mehrebenen-Item-Response-Modellen mit unterschiedlichen Modellspezifikationen gegenübergestellt. Analysiert werden Daten von 832 Fünftklässlerinnen und Fünftklässlern aus dem Schweizer Kanton St. Gallen, die wiederholt an Schulleistungstests im Fach Mathematik teilgenommen haben. Unterscheiden sich die Parameterschätzungen zwischen den Spezifikationen, spricht dies dafür, dass die Qualität der Lernmotivation einen Einfluss auf die Schätzung der Instruktionssensitivität der Mathematikitems nimmt. Als Indikatoren der Instruktionssensitivität werden die differenzielle (Naumann et al., 2016) und die spezifische Sensitivität herangezogen. Angenommen wird, dass die Einbeziehung der Qualität der Lernmotivation als Kovariate die Schätzung dieser beiden Indikatoren beeinflusst.

Das Ergebnis der Analysen ist eindeutig: Keine der aufgestellten Hypothesen kann angenommen werden. Die Resultate sprechen dafür, dass die Parameterschätzungen zur Evaluation der Instruktionssensitivität von der Qualität der Lernmotivation nicht wesentlich beeinflusst werden. Diese Befundlage überrascht, da zahlreiche Studien darauf hindeuten, dass motivationale Merkmale von Schülerinnen und Schülern einen Einfluss auf deren schulische Leistungen nehmen (u.a. Kriegbaum et al., 2015; Taylor et al., 2014) und für die Schätzung der Indikatoren der Instruktionssensitivität auf Daten von Schulleistungstests zurückgegriffen wurde. Die Ergebnisse werden aus inhaltlicher und methodischer Perspektive diskutiert.

Abstract

The evaluation of instructional sensitivity in consideration of individual learning prerequisites – An exemplary analysis focusing on students' quality of learning motivation

This dissertation focuses on the evaluation of instructional sensitivity of test items in consideration of students' individual learning prerequisites. The instructional sensitivity of items or test tasks reflects whether they are capable of capturing the effects of teaching. The term individual learning prerequisites is broadly defined and subsumes among others cognitive, metacognitive, motivational, and volitional characteristics (Brühwiler, 2014; Brühwiler et al., 2017). Considering individual learning prerequisites which are addressed in the context of utilization of learning opportunities models (e.g. Brühwiler, 2014; Fend, 1981), a particular focus of the subsequent quantitative analyses is on the quality of students' learning motivation (Ryan & Deci, 2000). The aim of these analyses is to examine, by way of example, the extent to which indicators of instructional sensitivity of test items are influenced by the quality of learning motivation.

To this end, parameter estimates from longitudinal multilevel item response theory models with different model specifications are compared. Data from 832 fifth graders from the Swiss canton of St. Gallen who repeatedly participated in school achievement tests in mathematics are analysed. If the estimated parameters differ, this suggests that the quality of learning motivation influences the estimation of the instructional sensitivity. The indicators of instructional sensitivity used are differential (Naumann et al., 2016) and specific sensitivity. It is assumed that including the quality of learning motivation as a covariate has an impact on the estimate of these two indicators.

The result of the analyses is clear: none of the hypotheses can be accepted. This finding indicates that the estimated parameters for evaluating instructional sensitivity are not significantly influenced by the quality of learning motivation. This is surprising given that numerous studies suggest that students' motivational characteristics have an impact on their academic performance (e.g. Kriegbaum et al., 2015; Taylor et al., 2014) and that data from school achievement tests were used to estimate the indicators of instructional sensitivity. The results are discussed from both a substantive and a methodological perspective.

1 Einleitung

Die Relevanz standardisierter Schulleistungstests ist im deutschsprachigen Raum insbesondere in den letzten zwei Jahrzehnten deutlich gestiegen. So werden Leistungstestdaten der Schülerinnen und Schüler beispielsweise herangezogen, um diese für eine testdatenbasierte Schul- und Unterrichtsentwicklung zu nutzen (u.a. Altrichter, Moosbrugger & Zuber, 2016; Ramsteck & Maier, 2015), eine testdatenbasierte Evaluation von Schule und Unterricht (Bildungsmonitoring) durchzuführen (u.a. Altrichter et al., 2016; Ramsteck & Maier, 2015), Schul- und Unterrichtseffektivitätsforschung zu betreiben (u.a. Hattie, 2009; Scheerens & Bosker, 1997) oder sie werden zur evidenzbasierten Steuerung des Bildungssystems verwendet (u.a. Altrichter & Maag Merki, 2016; für einen Überblick zu Funktionen standardisierter Leistungstest siehe auch Tresch, 2007). Schmette und Haase (2014) erachten standardisierte Schulleistungstests dabei als „Kernstück einer outputorientierten Steuerung von Bildungsqualität“, die – „vermittelt über die Rückmeldung von Testergebnissen – die Leistungsfähigkeit des Bildungssystems sichern bzw. steigern [sollen d. Verf.].“ (S. 48; siehe auch Fend, 2011, S. 8).

Um basierend auf solchen Leistungstestdaten der Schülerinnen und Schüler valide Rückschlüsse über Schule und Unterricht ziehen zu können, ist es wichtig, dass die Instruktionssensitivität von Tests bzw. Testitems sichergestellt ist (u.a. Cox & Vargas, 1966; Haladyna & Roid, 1981; Popham, 2007). Die Instruktionssensitivität eines Tests oder Testitems stellt ein wichtiges Validitätskriterium dar, das angibt, ob ein Test oder ein Testitem in der Lage ist, Effekte von Unterricht auf die Leistungen der Schülerinnen und Schüler abzubilden (u.a. Popham, 2007). Ist das Validitätskriterium der Instruktionssensitivität nicht hinreichend sichergestellt, können basierend auf den Testdaten der Schülerinnen und Schüler keine gültigen Rückschlüsse über Schule und Unterricht gezogen werden.

In Anbetracht der Erkenntnisse aus der Schul- und Unterrichtsforschung zu den Determinanten des schulischen Lernens (u.a. Brühwiler, Helmke & Schrader, 2017; Fend, 1981; Hattie, 2009; Helmke & Weinert, 1997) erscheint es jedoch „problematisch, dass in den Ansätzen zur Evaluation der Instruktionssensitivität häufig von einer unmittelbaren Wirkbeziehung zwischen Unterricht und Lernertrag ausgegangen wird“ (Musow, Naumann, Hochweber & Hartig, 2017, Abs. 1). Merkmale der Schülerinnen und Schüler, welche die Nutzung des Unterrichtsangebots (mit-)bedingen (Hintergrundmerkmale, individuelle Lernvoraussetzungen; siehe u.a. Brühwiler et al., 2017; Helmke & Weinert, 1997; Seidel, 2014), finden in den bisherigen Analysen zur

Instruktionssensitivität nur sehr eingeschränkt Berücksichtigung. So werden Hintergrundmerkmale, wie z.B. sozioökonomischer Status (SES) oder Migrationshintergrund, nur teilweise einbezogen. Den individuellen Lernvoraussetzungen wird kaum Beachtung geschenkt. Vor diesem Hintergrund geht es in der vorliegenden Arbeit um die Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraussetzungen der Schülerinnen und Schüler. Es soll überprüft werden, inwiefern die Instruktionssensitivität von Testitems durch die Qualität der Lernmotivation – als ein Beispiel einer individuellen Lernvoraussetzung – beeinflusst wird. Dahinter steht die Frage, ob Modelle zur Evaluation der Instruktionssensitivität von Testitems zu stark vereinfacht sind, wenn individuelle Lernvoraussetzungen in den Analysen unberücksichtigt bleiben.

Der Aufbau dieser Arbeit gestaltet sich wie folgt: In Kapitel 2 steht das Konzept der Instruktionssensitivität von Tests und Testitems im Vordergrund. Es geht um eine Klärung, was unter der Instruktionssensitivität verstanden werden kann, in welchen Kontexten das Konzept von Relevanz ist und welche Möglichkeiten zur Überprüfung der Instruktionssensitivität bestehen. Aufbauend auf den Ausführungen zur Instruktionssensitivität erfolgt in Kapitel 3 eine Erweiterung der Perspektive: die Evaluation der Instruktionssensitivität unter Berücksichtigung individueller Lernvoraussetzungen. Mithilfe des *Value-Addded-Ansatzes* und den Angebots-Nutzungs-Modellen wird der Versuch einer konzeptionellen Einbettung dieser Perspektivenerweiterung unternommen. Am Ende des Kapitels erfolgen die Ableitung der Fragestellung und die Darlegung der Hypothesen. Die Methode zur Beantwortung der Fragestellung wird in Kapitel 4 beschrieben. Zur Überprüfung der Hypothesen wird ein innovatives Verfahren angewendet, welches im bayesianischen Kontext zu verorten ist und auf *längsschnittlichen Mehrebenen-Item-Response-Theory-Modellen* basiert. Im Anschluss daran werden in Kapitel 5 die Ergebnisse präsentiert. Aus Gründen der Leserfreundlichkeit werden die Hypothesen im gleichen Kapitel überprüft, d.h. direkt im Anschluss an die jeweilige Ergebnisdarstellung. Die Ergebnisdiskussion sowie das Aufzeigen der Limitationen dieser Arbeit stehen im Mittelpunkt des Kapitels 6. Ansätze zur Erklärung der Ergebnisse dieser Dissertation sowie methodische Aspekte werden diskutiert. Ein wichtiges Anliegen im Rahmen dieser Arbeit ist zudem, über den Evaluationscharakter hinauszugehen und nach Begründungen für die Sensitivität bzw. Insensitivität von Testitems zu suchen. Ein Ausblick bezogen auf mögliche Implikationen für Forschung und Praxis wird in Kapitel 7 gegeben. Die Dissertation schließt mit einem Schlusswort in Kapitel 8, in dem u.a. für eine verstärkte interdisziplinäre

Zusammenarbeit und die Weitung des Blickwinkels auf die Instruktionssensitivität plädiert wird, um die Forschung in diesem Bereich voranzutreiben.

2 Informationen zum Hintergrund und zum Forschungsstand zur Instruktionssensitivität von Tests und Testitems

Das vorliegende Kapitel hat zum Ziel, einen allgemeinen Überblick über die Instruktionssensitivität von Tests bzw. Testitems zu vermitteln. Zunächst soll in Kapitel 2.1 der Begriff der Instruktionssensitivität definiert, anschließend in Kapitel 2.2 der Ursprung des Konzepts verdeutlicht und in Kapitel 2.3 die Relevanz des Konstrukts hierzulande dargelegt werden. In einem nächsten Schritt erfolgen in Kapitel 2.4 Ausführungen zur Operationalisierung der Instruktionssensitivität von Tests bzw. Testitems sowie die Beschreibung des aktuellen Forschungsstands. Daran anknüpfend wird in Kapitel 2.5 herausgearbeitet, was Hinweise auf eine unzureichende Instruktionssensitivität von Tests und Testitems sein könnten. Das Kapitel schließt mit einer Zusammenfassung und einem Fazit in Kapitel 2.6.

2.1 Der Begriff der Instruktionssensitivität von Tests bzw. Testitems

Bezogen auf die Instruktionssensitivität von Tests bzw. Testitems gibt es verschiedene Definitionen. Eine Definition, die hier aufgeführt werden soll, stammt von Haladyna und Roid (1981). Sie betrachten die Instruktionssensitivität als die Tendenz eines Testitems, infolge von Unterricht in der (Item-)Schwierigkeit zu variieren (S. 40). Dahinter steht die (zu überprüfende) Annahme, dass in einem *Pretest-Posttest-Design* ein instruktionssensitives Item für Schülerinnen und Schüler über die Zeit leichter werden sollte (siehe auch Cox & Vargas, 1966). An dieser Stelle sei allerdings darauf hingewiesen, dass Testitems natürlich kein Eigenleben führen. Mit Aussagen, dass ein Testitem über die Zeit leichter wird oder es sensitiv reagiert, ist im Grunde gemeint, dass sich die Antworten der Schülerinnen und Schüler infolge von Unterricht verändern. Letzteres kann mithilfe von längsschnittlich erhobenen Schulleistungstestdaten aufgezeigt werden, vorausgesetzt die Items sind instruktionssensitiv. Deutlich wird damit auch, dass Tests und Testitems sich hinsichtlich ihrer Sensitivität unterscheiden können. Mithilfe von Indikatoren der Instruktionssensitivität und unter Hinzuziehung von statistischen Tests ist eine Beurteilung möglich, ob diese als sensitiv oder insensitiv einzustufen sind.

Diese Ausführungen zur Definition der Instruktionssensitivität liefern insofern eine erste Idee darüber, was unter Instruktionssensitivität verstanden werden kann. Nachfolgend sollen

Definitionen aufgezeigt werden, die den Aspekt des Unterrichts stärker in den Blick nehmen. Eine erste Definition stammt von Popham (2007). Er versteht unter Instruktionssensitivität den Grad, in dem sich die Qualität des Unterrichts in den Leistungen der Schülerinnen und Schüler korrekt widerspiegelt (S. 1). Niemi, Wang, Steinberg, Baker und Wang (2007) gehen hinsichtlich der Definition in eine ähnliche Richtung. Sie fokussieren aber nicht auf die Unterrichtsqualität, sondern bleiben allgemeiner. Ihrer Ansicht nach zeichnet sich ein instruktionssensitiver Test dadurch aus, dass er in der Lage ist, Effekte aus dem vorangegangenen Unterricht zu messen (S. 216). Neben Effekten von (1) Unterrichtsqualität[1] kann es aber auch um Effekte bezogen auf (2) die Unterrichtsinhalte bzw. die im Unterricht verfolgten Lernziele, (3) die Gewichtung bzw. die Quantität von Unterrichtsmerkmalen und (4) didaktische Aspekte hinsichtlich der Gestaltung des Unterrichts gehen (siehe z.B. D'Agostino, Welsh & Corson, 2007; Wiley & Yoon, 1995).

Diese ausgewählten Definitionen sollen die verschiedenen Blickwinkel auf die Instruktionssensitivität von Tests und Testitems verdeutlichen. Zum einen werden einzelne Aspekte zur Messung der Instruktionssensitivität aufgegriffen. Zum anderen wird die Instruktionssensitivität mit bildungswissenschaftlichen Aspekten in Verbindung gebracht, indem Effekte von Unterricht mit einem Lernzuwachs bei Schülerinnen und Schülern in Beziehung gesetzt werden. Die verschiedenen Blickwinkel ergänzen sich gegenseitig und sind für die Entwicklung einer Vorstellung über das Konstrukt der Instruktionssensitivität hilfreich. Eine letzte Definition, die hier eine übergeordnete Rolle einnimmt und deshalb dargelegt werden soll, stammt von Polikoff (2010), der die Instruktionssensitivität als psychometrische Eigenschaft eines Tests oder eines Testitems beschreibt. Er verortete damit die Instruktionssensitivität erstmals explizit im Bereich der Psychometrie. Letzteres bedeutet aber nicht, dass das Thema ausschließlich im Bereich der Psychologie von Interesse ist (siehe hierzu Naumann, Musow, Aichele, Hochweber & Hartig, 2019). In Kapitel 2.3 wird auf die verschiedenen Kontexte eingegangen, in denen die Instruktionssensitivität von Tests und Testitems von Relevanz ist.

Insgesamt kann festgehalten werden, dass trotz der dargelegten unterschiedlichen Perspektiven auf die Instruktionssensitivität von Tests und Testitems von einem gemeinsamen Grundverständnis auszugehen ist. Dieses allgemeine Grundverständnis, das auch dieser Arbeit

[1] Der Begriff der Unterrichtsqualität kann sehr unterschiedlich definiert werden. In den Studien zur Instruktionssensitivität wird i.d.R. zwischen den vier genannten Aspekten differenziert.

zugrunde liegt, lautet: *Die Instruktionssensitivität eines Tests bzw. eines Testitems beschreibt das Ausmaß, in dem ein Test oder ein Testitem in der Lage ist, Effekte von Unterricht auf Leistungen der Schülerinnen und Schüler abzubilden.* Dahinter steht die Annahme, dass der Unterricht Lernprozesse bei Schülerinnen und Schülern anregt, die sich in den Tests oder Testitems niederschlagen sollten. Eine wichtige Annahme dieser Arbeit ist zudem, dass sich der Unterricht zwischen den Klassen unterscheidet und dadurch Schülerinnen und Schüler aus verschiedenen Klassen bei der Bearbeitung von Tests oder Testitems erfolgreicher oder weniger erfolgreicher abschneiden (siehe auch Naumann, Hartig & Hochweber, 2017).

2.2 Ursprung der Instruktionssensitivität von Tests und Testitems

Der Ursprung des Konzepts der Instruktionssensitivität ist in den 1960er/70er Jahren zu verorten und ging mit dem Aufkommen kriteriumsorientierter Tests einher (Cox & Vargas, 1966; Ebel, 1962; Glaser 1963; Millman, 1970; Polikoff, 2010; Popham & Husek, 1969). Unter einem kriteriumsorientierten Test versteht man einen Test, dem Standards zugrunde liegen (sachliche Bezugsnorm, z.B. Bildungsstandards, curriculare Ziele) und dessen Erreichen überprüft werden soll (Drechsel, Prenzel, & Seidel, 2015; Heller & Hany, 2014; Rheinberg, 2014). Normorientierte Test dienen hingegen dem Vergleich zwischen den Leistungen einer Person oder einer Gruppe mit sich selbst oder anderen Personen bzw. Gruppen (individuelle und soziale Bezugsnormorientierung; siehe Heller & Hany, 2014; Rheinberg, 2014). Während zur Konstruktion der bis dato dominierenden normorientierten Tests klassische Indizes wie die Trennschärfe oder auch die Itemschwierigkeit zur Itemselektion herangezogen wurden, erschienen diese traditionellen Maße zur Evaluation der Instruktionssensitivität weniger adäquat (Haladyna & Roid, 1981; Kosecoff & Klein, 1974; Polikoff, 2010). Es brauchte neue Methoden, mit denen überprüft werden konnte, inwiefern sich Tests oder auch Testitems eigneten (d.h. u.a. sich als sensitiv erwiesen), um Unterschiede hinsichtlich des Unterrichts (z.B. Inhalt, Qualität, Gewichtung von Unterricht) zu erfassen (Brennan, 1972; Cox & Vargas, 1966; Helmstadter, 1972; Kosecoff & Klein, 1974; Popham, 1971; Roudabush, 1974). Mit dem Aufkommen kriteriumsorientierter Tests wurde die Instruktionssensitivität von Tests bzw. Testitems so zu einem wichtigen Testgütekriterium.

In den 1980er Jahren rückte das Konzept der Instruktionssensitivität zunehmend in den Hintergrund zugunsten des Konzepts der Instruktionsvalidität (Polikoff, 2010). Auslöser stellte eine Klage gegen den US-Bundesstaat Florida dar, welche in den Vereinigten Staaten viel

Aufsehen erregte (D'Agostino et al., 2007). In Florida wurde 1981 ein Test an Highschools eingeführt, der für einen erfolgreichen Highschool-Abschluss bestanden werden musste (Debra P. v. R. D. Turlington, 1979/1984). Zahlreiche Schülerinnen und Schüler, insbesondere Personen mit afroamerikanischer Herkunft, verfehlten jedoch die Mindestanforderungen zum Bestehen des Tests und zogen deshalb vor Gericht (ebd.). Dem Bundesstaat wurde damals die unzureichende Vorbereitung auf den Test im Rahmen des Unterrichts vorgeworfen (ebd.). Zwar gewann der Staat den Prozess (ebd.), doch führte er zu einer verstärkten Aufmerksamkeit für Fragen der Validität in Verbindung mit dem Einsatz standardisierter Schulleistungstests (Airasian & Madaus, 1983; Linn, 1983a, 1983b). Neue Instrumente wurden entwickelt, die insbesondere der Erfassung von Lerngelegenheiten (*Opportunities to Learn*, OTL) dienten, um die Passung zwischen dem implementierten Curriculum und den Testinhalten (*content coverage*) zu prüfen und sicherzustellen. Darüber hinaus wurden Instrumente zur Erfassung der Gewichtung von Unterrichtsinhalten (*emphasis*) und zur Untersuchung der Beziehung zwischen Lerngelegenheiten und den Leistungen der Schülerinnen und Schüler konzipiert (u.a. Gamoran, Porter, Smithson & White, 1997; Wiley & Yoon, 1995; Yoon, Burstein & Gold, 1991; Yoon & Resnick, 1998). Fragen zur Testfairness (Linn & Harnish, 1981) kamen auf, welche später mithilfe von *Differential Item Functioning* (DIF) untersucht wurden (Holland & Wainer, 1993). Aufgrund des Fokus auf OTL und Testfairness gab es in den 1990er Jahren nur wenige Autoren, die sich explizit mit dem Thema der Instruktionssensitivität von Tests und Testitems befassten und die Verfahren weiterentwickelten (Baker, 1994; Muthén et al., 1995; Muthén et al., 1991).

Erst Anfang des 21. Jahrhunderts erfuhr die Debatte um die Instruktionssensitivität standardisierter Schulleistungstests neuen Aufschwung (Popham, 2007; Ruiz-Primo et al., 2012). Als Ursache hierfür kann der *No Child Left Behind Act* ([NCLB], 2002) betrachtet werden, der mit einem verstärken Ausbau des *high-stakes testing* und der Forcierung eines *Accountability*-Systems (Rechenschaftslegung) einherging. Mit dem *high-stakes testing* wird eine Art Wettbewerbslogik verfolgt. Schneiden Schülerinnen und Schüler und damit die Lehrpersonen und Schulen in den standardisierten Tests besonders gut ab, kann dies positive Aspekte mit sich bringen (z.B. Stipendien, positive Bilanz bei Schulrankings, Sonderzahlungen; Amrein & Berliner, 2002; siehe auch Kelly, 2022 sowie Stecher, 2002). Ein besonders schlechtes Abschneiden der Schülerinnen und Schüler kann hingegen negative Konsequenzen zur Folge haben (z.B. Verfehlen eines Abschlusses, schlechte Bilanz bei Schulrankings; Kürzung finanzieller Mittel, Schulreformen; Amrein & Berliner, 2002; Kelly,

2022; siehe auch Minarechová, 2012 sowie Stecher, 2002). Aufgrund möglicher negativer Konsequenzen infolge schlechter Testergebnisse ist es nachvollziehbar, dass das *high-stakes testing* mit einem erhöhten Druck bei den Beteiligten einhergeht (Stecher, 2002). Die Instruktionssensitivität von Tests und Testitems rückte daher mit dem *No Child Left Behind Act* (2002) und den damit verbundenen Entwicklungen im Bildungssystem wieder vermehrt in den Fokus von Wissenschaftlerinnen und Wissenschaftlern, da sie ein wichtiges Gütekriterium standardisierter Schulleistungstests darstellt (Chen, 2012; D'Agostino et al., 2007; Ing, 2008, 2016, 2017; Li, Qin & Lei, 2017; Niemi et al., 2007; Pham, 2009; Polikoff, 2010, 2016; Popham, 2014; Popham et al., 2014; Popham & Ryan, 2012; Rodriguez, 2017; Ruiz-Primo & Li, 2015; Ruiz-Primo et al., 2012; Ruiz-Primo, Shavelson, Hamilton & Klein, 2002). Denn es erschien aufgrund der weitreichenden Konsequenzen wichtig, dass die gezogenen Rückschlüsse über Schule und Unterricht basierend auf den Leistungstestdaten der Schülerinnen und Schüler auch valide (siehe auch Messick, 1989, 1995) waren.

Die Ausführungen machen ersichtlich, dass in den USA die Instruktionssensitivität von Tests bzw. Testitems seit längerem Aufmerksamkeit erfährt und diskutiert wird (u.a. Chen, 2012; D'Agostino et al., 2007; Haladyna & Roid, 1981; Hanson, McMorris & Bailey, 1986; Ing, 2008, 2016; Kosecoff & Klein, 1974; Muthén et al., 1995; Muthén et al., 1991; Polikoff, 2010, 2016; Popham & Ryan, 2012; Rodriguez, 2017; Ruiz-Primo et al., 2012). Im deutschsprachigen Raum findet das Gütekriterium hingegen wenig Beachtung (Deutscher & Winther, 2018; Naumann, 2013; Naumann et al., 2017; Naumann, Hochweber & Hartig, 2014; Naumann, Klieme & Hochweber, 2016; Naumann, Musow et al., 2019; Naumann, Musow & Katstaller, 2020; Naumann, Rieser, Musow, Hochweber & Hartig, 2019). Ein Grund hierfür könnte sein, dass hierzulande eher ein *low-stakes testing* praktiziert wird (Bach, Wurster, Thillmann, Pant & Thiel, 2014; Grünkorn, Klieme & Stanat, 2019). Ein *low-stakes testing* soll in erster Linie der internen Qualitätsentwicklung auf Schul- und Unterrichtsebene dienen (Bach et al., 2014; Grünkorn et al., 2019; Maier, 2008). Der Druck bezogen auf das Abschneiden im Test ist bei einem *low-stakes testing* für die Beteiligten geringer und dennoch gibt es Hinweise auf unerwünschte Wirkungen durch die Einführung derartiger standardisierter Tests, wenngleich diese nicht so gut belegt sind wie in den USA (Jäger, 2012; Lind, 2009; Maier & Kuper, 2012).

2.3 Relevanz der Instruktionssensitivität von Tests und Testitems

Zunächst soll vorweggenommen werden, dass das Gütekriterium der Instruktionssensitivität nicht für jeden Test oder jedes Testitem gleichermaßen von Relevanz ist. Die Sicherstellung der Instruktionssensitivität von Schulleistungstest bzw. von Testitems ist von zentraler Bedeutung, wenn basierend auf Testdaten der Schülerinnen und Schüler Rückschlüsse über Schule und Unterricht gezogen werden sollen. Es ist also zu überlegen, wofür der Test oder die jeweiligen Testitems genutzt bzw. in welcher Hinsicht die Ergebnisse interpretiert werden sollen (siehe auch Kane, 2013). Soll ein Test oder das jeweilige Testitem dazu dienen, Rückschlüsse über die Schule oder den Unterricht zu ziehen, was wäre dann die Konsequenz, wenn das Gütekriterium der Instruktionssensitivität nicht hinreichend sichergestellt ist? Im letztgenannten Fall würde dies bedeuten, dass die Lehrperson mit einem hohen Zeitaufwand und in hoher Qualität Unterricht gestalten könnte, ohne dass sich dies in den Schulleistungstests der Schülerinnen und Schüler widerspiegelt. Die Validität der Testwertnutzung bzw. der Testwertinterpretation wäre im letztgenannten Fall anzuzweifeln. Deutlich wird an dieser Stelle, dass die Relevanz des Gütekriteriums der Instruktionssensitivität vor dem Hintergrund der Testwertnutzung bzw. der Testwertinterpretation zu beurteilen ist. Ungeklärt ist noch die Frage, inwiefern diese Form der Testwertnutzung und Testwertinterpretation in Deutschland von Bedeutung ist. In welchen Kontexten werden standardisierte Schulleistungstests genutzt, um Rückschlüsse über Schule und Unterricht zu ziehen? Erste Überlegungen hierzu sind bei Naumann, Musow et al. (2019) bzw. Naumann et al. (2020) zu finden, die an dieser Stelle noch weiter systematisiert und ausgeführt werden sollen.

Insgesamt sollen vier Kontexte herausgestellt werden: (1) die testdatenbasierte Schul- und Unterrichtseffektivitätsforschung, (2) die testdatenbasierte Evaluation von Schule, Unterricht und Lehrpersonen, (3) die testdatenbasierte Sicherstellung und Weiterentwicklung der Qualität von Schule und Unterricht und (4) die evidenzbasierte Steuerung des Bildungssystems. Allen vier Kontexten ist gemeinsam, dass die Leistungstestdaten als eine Art Feedback genutzt werden (Altrichter et al., 2016). Die einzelnen Kategorien gehen teilweise ineinander über, da standardisierte Schulleistungstests mitunter nicht nur eine, sondern mehrere Funktionen haben. Ein Beispiel hierfür, auf das an späterer Stelle noch näher eingegangen wird, stellen die Vergleichsarbeiten in Schulen dar (VERA; siehe Helmke, Hosenfeld & Schrader, 2004).

Zu (1): Schul- und Unterrichtseffektivitätsforschung (im Englischen: *educational effectiveness research* bzw. *school or teaching effectiveness research*; Creemers & Kyriakides, 2008; Hattie,

2009; Scheerens & Bosker, 1997) ist auch in Deutschland ein zentrales Thema, was anhand der zahlreichen Forschungsbemühungen ersichtlich wird (siehe u.a. Klieme, 2016; Klieme & Rakoczy, 2008). Im Zentrum der Schul- und Unterrichtseffektivitätsforschung steht die Frage nach den Wirkfaktoren, welche im Kontext von Schule und Unterricht zum Tragen kommen und mit dem Lernen von Schülerinnen und Schülern in Zusammenhang gebracht werden. Als Beispiel können Interventionsstudien mit quasi-experimentellem Design zur Überprüfung der Effekte unterschiedlicher Unterrichtsmethoden auf die kognitiven und nicht kognitiven Leistungen der Schülerinnen und Schüler genannt werden (siehe z.B. die IGEL-Studie (Individuelle Förderung und adaptive Lern-Gelegenheiten in der Grundschule); Decristan, Hondrich et al., 2015). Um valide Rückschlüsse über die Wirksamkeit von Schule und Unterricht ziehen zu können, ist die Sicherstellung der Instruktionssensitivität von Tests bzw. Testitems in diesem Kontext von hoher Bedeutung.

Zu (2): Auch die Evaluation von Unterricht und Schule hat in Deutschland einen hohen Stellenwert (Altrichter et al., 2016; Bohl & Kiper, 2009; Isaac, Halt, Hosenfeld, Helmke & Ophoff, 2006; Lorenz, 2005; Ramsteck & Maier, 2015). Bei der Evaluation geht es zunächst um eine Bewertung von Schule und Unterricht basierend auf den Testergebnissen der Schülerinnen und Schüler, die intern oder auch extern durchgeführt werden (Kiper, 2009). Auf Grundlage der Evaluationsergebnisse können die Testdaten der Schülerinnen und Schüler dann für weitere Zwecke – wie beispielsweise einem Bildungsmonitoring zur Überprüfung des Erreichens von Bildungsstandards oder der Weiterentwicklung des Unterrichts, der Schule und des Bildungssystems – Verwendung finden (Maier & Kuper, 2012; Ramsteck & Maier, 2015). Sollten die Schülerinnen und Schüler einer Klasse bzw. einer Schule unterdurchschnittlich abschneiden, drohen aber keine vergleichenbaren Sanktionen wie in den USA zu Zeiten des NCLB (Maier & Kuper, 2012; Pinkerton, Scott, Buell & Kober, 2004; Ramsteck & Maier, 2015). Die Sicherstellung der Instruktionssensitivität von Tests bzw. Testitems ist aber auch in diesem Kontext von hoher Bedeutung.

Zu (3): Die Gewährleistung und Weiterentwicklung der Qualität von Schule und Unterricht (im Englischen: *school and teaching improvement*) basierend auf Testdaten der Schülerinnen und Schüler ist ein Thema, das auch in Deutschland von besonderer Relevanz ist (Altrichter et al., 2016; Maier & Kuper, 2012; Ramsteck & Maier, 2015). Dazu wurden Instrumente entwickelt, mit denen ein selbstgesteuerter Prozess zur Qualitätssicherung und Optimierung von Schule und Unterricht ermöglicht werden soll (Altrichter et al., 2016). Grünkorn et al. (2019) sprechen

in diesem Zusammenhang auch von einem Systemmonitoring im weiteren Sinne, einen Begriff, auf den am Ende dieses Unterkapitels noch eingegangen wird. Darüber hinaus ist an dieser Stelle anzumerken, dass die beiden Aspekte mit der Evaluation eng verbunden sind. Die Evaluationsergebnisse können eine entsprechende Grundlage bilden, um die Qualität von Schule und Unterricht zu sichern und/oder weiterzuentwickeln. Insofern ist das Gütekriterium der Instruktionssensitivität auch hier von besonderer Relevanz.

Zu (4): Einen letzten Aspekt stellt die evidenzbasierte Steuerung des Bildungssystems dar, welche auf dem Gedanken fußt, sich Erkenntnisse aus der empirischen Bildungsforschung für bildungspolitische Entscheidungen zunutze zu machen (Altrichter & Maag Merki, 2016). Standardisierte Schulleistungstests werden in diesem Zusammenhang als „Kernstück einer outputorientierten Steuerung der Bildungsqualität“ erachtet, die – „vermittelt über die Rückmeldung von Testergebnissen – die Leistungsfähigkeit des Bildungssystems sichern bzw. steigern [sollen d. Verf.]“ (Schmette & Hasse, 2014, S. 48; siehe auch Fend, 2011, S. 8). Grünkorn et al. (2019) sprechen in diesem Kontext von einem Systemmonitoring im engeren Sinne. Häufig wird aber auch von einem Bildungsmonitoring gesprochen (siehe auch Bromme, Prenzel & Jäger, 2016).

Da der Begriff des Monitorings bereits mehrfach gefallen ist, soll dieser noch kurz erläutert werden. Grünkorn et al. (2019) differenzieren zwischen einem Systemmonitoring im engeren und im weiteren Sinne. Das Systemmonitoring im weiteren Sinne zielt auf eine testdatenbasierte Rückmeldung zur Selbstevaluation von Schule und Unterricht ab. Es geht also um die Sicherstellung und Weiterentwicklung der Qualität der jeweiligen Schule und des Unterrichts in der jeweiligen Klasse (ebd.). Das Systemmonitoring im engeren Sinne kommt hingegen auf Ebene des Bildungssystems zum Tragen und ist mit dem Begriff des „Bildungsmonitorings“ – wie es beispielsweise im Zusammenhang mit PISA (Program for International Student Assessment), TIMSS (Trends in International Mathematics and Science Study) oder IGLU (Internationale Grundschul-Lese-Untersuchung) kommuniziert wird – gleichzusetzen. Altrichter et al. (2016) merken zudem an, dass sich internationale und nationale Schulleistungsstudien hinsichtlich der primären Ziele unterscheiden. Internationale Vergleichsstudien besitzen die primäre Funktion eines Bildungsmonitorings im engeren Sinne und werden für bildungspolitische Entscheidungsprozesse herangezogen (ebd.). Dagegen sollen nationale Vergleichsstudien insbesondere der internen Qualitätssicherung und Qualitäts-weiterentwicklung von Schule und Unterricht dienen (ebd.). Der Begriff des Monitorings wird

folglich im bildungswissenschaftlichen Kontext unterschiedlich verwendet. Insgesamt ist aber festzuhalten, dass auch hier die Sicherstellung der Instruktionssensitivität von Tests und Testitems eine wichtige Rolle spielen kann.

Die Ausführungen des Kapitels 2.3 machen damit deutlich, dass es zahlreiche Kontexte in Deutschland gibt, bei denen zu überlegen ist, inwiefern die Sicherstellung der Instruktionssensitivität von Tests oder Testitems von Relevanz ist. Gleichzeitig ist darauf hinzuweisen, dass nicht jeder Test oder jedes Testitem instruktionssensitiv sein muss. Es ist stets zu überlegen, in welcher Hinsicht die Testdaten später verwendet werden sollen.

2.3.1 Beispiel VERA – Schul- und Unterrichtsentwicklung, Bildungsmonitoring & Individualdiagnostik

Die Einführung nationaler Bildungsstandards resultierte aus der Erkenntnis internationaler Bildungsmonitorings (insbesondere PISA/PISA-E), dass die erfolgreichen Länder im Rahmen des Bildungsmonitorings im Vergleich zu den weniger erfolgreichen Ländern einer Outputorientierung folgen und Bildungsstandards vorweisen konnten (Klieme, 2004; Ramsteck & Maier, 2015). Mit der Einführung nationaler Bildungsstandards in Deutschland (Klieme et al., 2003) kamen erste Instrumente auf den Markt, mit denen die Implementierung von Bildungsstandards fortan überprüft und evaluiert werden sollte (Jäger, 2012; Klieme, 2004; Lorenz, 2005; Ramsteck & Maier, 2015). Ein prominentes Beispiel stellen hierfür die Vergleichsarbeiten (VERA) dar, welche gemäß den Ausführungen von Ramsteck und Maier (2015) auf den verschiedenen Ebenen der Akteure zum Tragen kommen.

Auf der Ebene der Unterrichtsentwicklung erhalten Lehrpersonen mittels VERA eine Rückmeldung zum Lernstand bzw. zum Lernprofil ihrer Schülerinnen und Schüler mit diversen Vergleichsmöglichkeiten. Das Instrument eröffnet die Möglichkeit einer Kriteriums- (durch den Abgleich mit dem Erwartungshorizont bezogen auf Bildungsstandards) und einer Normorientierung (durch den Vergleich mit Referenzgruppen), welche zur Reflexion der Unterrichtspraxis herangezogen werden können (Isaac et al., 2006; Lorenz, 2005; Maier 2008; Ramsteck & Maier, 2015). VERA ermöglicht damit den Lehrpersonen mittels Auswertung und im kollegialen Austausch wichtige Anhaltspunkte zur Optimierung des Unterrichts zu erhalten. Auf der Ebene der Schulentwicklung kann die jeweilige Schulleitung bzw. die jeweilige Fachkonferenz (ggf. auch in Kooperation mit der Schulaufsicht) die Testdaten der Schülerinnen und Schüler heranziehen, um eine schulinterne Evaluation durchzuführen (Isaac et al., 2006;

Kiper, 2009; Lorenz, 2005; Ramsteck & Maier, 2015). VERA soll entsprechend die Autonomie der jeweiligen Schulen unterstützen bzw. unterstreichen, die Gestaltungsfreiräume aufzeigen (Isaac et al., 2006; Ramsteck & Maier, 2015) und neue Impulse geben, ohne dabei den Anspruch zu erheben, pauschale Empfehlungen bereitzustellen (Isaac et al., 2006).

Neben der Funktion der Schul- bzw. Unterrichtsentwicklung wird VERA auch eine Funktion des Systemmonitorings im engeren Sinne zugeschrieben (Jäger, 2012; Lorenz, 2005; Ramsteck & Maier, 2015). Mit VERA wird das Ziel verfolgt, die Qualität des Bildungssystems zu prüfen, zu sichern und weiterzuentwickeln. Die Vergleichsarbeiten können daher zu Zwecken der Rechenschaftslegung und der Transparenz herangezogen werden (Ramsteck & Maier, 2015). Darüber hinaus wird VERA auch mit dem Aspekt der Individualdiagnostik in Verbindung gebracht. Dementsprechend können Ergebnisse aus VERA zur individuellen Förderung, zur Zertifizierung von Kompetenzen und zur Orientierung u.a. für Übertrittsentscheidungen genutzt werden (Lorenz, 2005; Maier, 2008). Das Bildungsmonitoring im engeren Sinne und die Individualdiagnostik werden bei VERA allerdings als Sekundärziele beschrieben.

VERA ist somit ein gutes Beispiel, um zu verdeutlichen, dass auch in Deutschland Leistungstestdaten der Schülerinnen und Schüler herangezogen werden, um Rückschlüsse über Schule und Unterricht zu ziehen. In den Bundesländern werden Vergleichsarbeiten in den Schulen durchgeführt. Die Sicherstellung der Gütekriterien der VERA-Tests ist daher von hoher Relevanz.

2.3.2 Beispiele – Die Instruktionssensitivität kann nicht per se angenommen werden

Dass die Instruktionssensitivität von Tests oder Testitems nicht per se angenommen werden kann, sondern variiert, soll anhand von zwei Beispielen verdeutlicht werden. Ein erstes Beispiel, das stellvertretend für zahlreiche Studien zur Instruktionssensitivität steht, stellt die Studie von Polikoff (2016) dar. In der Untersuchung wurde die Sensitivität eines Mathematik- und eines Englischtests unter Hinzuziehung von vier verschiedenen Instrumenten zur Erfassung der Unterrichtsqualität überprüft. Ziel war ein Erkenntnisgewinn, inwiefern die Tests bezogen auf die Daten zur Unterrichtsqualität sensitiv sind. Das Ergebnis zeigte, dass der Nachweis eines Zusammenhangs zwischen dem Abschneiden im Schulleistungstest und der gemessenen Unterrichtsqualität zum einen vom gewählten Instrument zur Erhebung der Unterrichtsqualität abhängt und zum anderen in Verbindung mit den Leistungstestdaten aus dem jeweiligen Bezirk

(engl. *district*) steht (ebd.). Darüber hinaus gibt es zahlreiche andere Studien, welche zeigen, dass die Instruktionssensitivität von Tests bzw. Testitems variiert und nicht per se gegeben ist (siehe auch Ing, 2017; Ruiz-Primo et al., 2012).

Ein diesbezügliches zweites Beispiel ist bei Naumann, Rieser et al. (2019) zu finden. In ihrer Studie wurden 286 Testversionen erstellt, indem aus einem Pool von 13 Items in 286 Durchgängen jeweils zehn Items in unterschiedlichen Kombinationen gezogen und zu einem Test zusammengestellt wurden. Die in der Studie verwendeten Items stammten aus der IGEL-Studie (Decristan, Hertel et al., 2014). Naumann, Rieser et al. (2019) zeigten mit ihren Analysen auf, inwiefern die Instruktionssensitivität der Tests aufgrund der unterschiedlichen Testzusammensetzung variiert. Für die Überprüfung wurde das Unterrichtsqualitätsmerkmal der kognitiven Aktivierung herangezogen. Ist der Test sensitiv, würde dies bedeuten, dass Schülerinnen und Schüler, die einen kognitiv aktivierenden Unterricht erlebt haben, im Schulleistungstest erfolgreicher abschneiden als Schülerinnen und Schüler, deren Unterricht weniger kognitiv aktivierend gestaltet war. Das Ergebnis dieser Studie ist eindrücklich. Von den 286 zusammengestellten Tests erwiesen sich insgesamt 38 Tests als sensitiv bezogen auf die kognitive Aktivierung. Im Umkehrschluss bedeutet es, dass die große Mehrheit der Tests auf das Qualitätsmerkmal der kognitiven Aktivierung nicht sensitiv reagierte (ebd.).

Die Ausführungen der beiden Beispiele machen deutlich, dass die Wahl des Erhebungsinstruments bzw. die Auswahl der Testitems für die Konzipierung eines Tests von erheblicher Relevanz ist, wenn die Instruktionssensitivität sichergestellt werden soll.

2.4 Methoden zur Überprüfung der Instruktionssensitivität von Tests und Testitems

Mit dem Wissen über die Definition zur Instruktionssensitivität von Tests und Testitems und deren Relevanz stellt sich die Frage, wie die Instruktionssensitivität empirisch überprüft werden kann. Dafür gibt es zahlreiche Verfahren, die jedoch an dieser Stelle nicht im Detail ausgeführt werden sollen. Stattdessen soll in einem ersten Schritt eine Systematisierung der verschiedenen Verfahren gemäß den Varianzquellen und den herangezogenen Informationsquellen zur Überprüfung der Instruktionssensitivität von Tests und Testitems erfolgen.

2.4.1 Systematisierung der Verfahrensweisen gemäß den Varianzquellen

Naumann et al. (2017) haben ein erweitertes *Framework* konzipiert (Abbildung 1), das Aufschluss über die Verortung der verschiedenen Verfahrensweisen zur Evaluation der Instruktionssensitivität unter Berücksichtigung der Varianzquellen und der zugrunde liegenden Hypothesen gibt. Hinsichtlich der Varianzquellen unterscheiden die Autoren drei Perspektiven, nämlich (1) die Zeitpunkte-Perspektive, (2) die Gruppen-Perspektive und (3) die Zeitpunkte-x-Gruppen-Perspektive (ebd.). Mit der Zeitpunkte-Perspektive ist gemeint, dass sich die Itemschwierigkeiten über die Zeit verändern können. Ein instruktionssensitives Item zeichnet sich demnach dadurch aus, dass dieses einen Lernzuwachs (über die Zeit) infolge von Unterricht auf die Leistungen der Schülerinnen und Schüler abbildet (ebd.). Als Beispiel für ein Verfahren, das der Zeitpunkte-Perspektive zugeordnet werden kann, ist der *pretest-posttest difference index* (PPDI, Cox & Vargas, 1966) zu nennen. Mit der Gruppen-Perspektive ist hingegen gemeint, dass sich die Itemschwierigkeiten aufgrund von Gruppenzugehörigkeiten unterscheiden können (Naumann et al., 2017). Beispielsweise können Schulklassen bei der Bearbeitung instruktionssensitiver Testitems je nach Unterricht erfolgreicher oder weniger erfolgreich abschneiden. Ein Beispiel für ein Verfahren, das der Gruppen-Perspektive zugeordnet werden kann, stellt der *Differential-Item-Functioning-Ansatz* (DIF, Linn & Harnisch, 1981) oder auch der *Multilevel-Differential-Item-Functioning-Ansatz* (Multilevel DIF, Robitzsch, 2009) dar. Die Interaktion der beiden bereits genannten Perspektiven kommt in der Zeitpunkte-x-Gruppen-Perspektive zum Ausdruck. Im Fokus stehen gruppenspezifische Veränderungen in den Itemschwierigkeiten (Naumann et al., 2017). Als Beispiel für ein Verfahren, das dieser Perspektive zugeordnet werden kann, ist das *längsschnittliche Mehrebenen-Item-Response-Theory-Modell* zu nennen (LMLIRT-Modell, ebd.). Das LMLIRT-Modell ist für die vorliegende Arbeit von hoher Bedeutung (Kapitel 4.5). Aus diesem Grund soll bereits an dieser Stelle kurz auf die Indikatoren der Instruktionssensitivität von Testitems eingegangen werden, die mithilfe des LMLIRT-Modells geschätzt werden können. Zwei Indikatoren sind im Zusammenhang mit dem LMLIRT-Modell zu nennen: (1) die globale und (2) die differenzielle Sensitivität (ebd.). Die globale Sensitivität entspricht der mittleren Veränderung der klassenspezifischen Itemschwierigkeit über die Zeit (Naumann et al., 2014). Die differenzielle Sensitivität entspricht hingegen der Varianz der Veränderung der klassenspezifischen Itemschwierigkeit (ebd.). Beide Indikatoren umfassen gruppenspezifische Veränderungen in den Itemschwierigkeiten, wobei sich der eine Indikator auf die Mittelwerte bezieht, während sich der andere Indikator auf die Varianz stützt (Naumann et al., 2017).

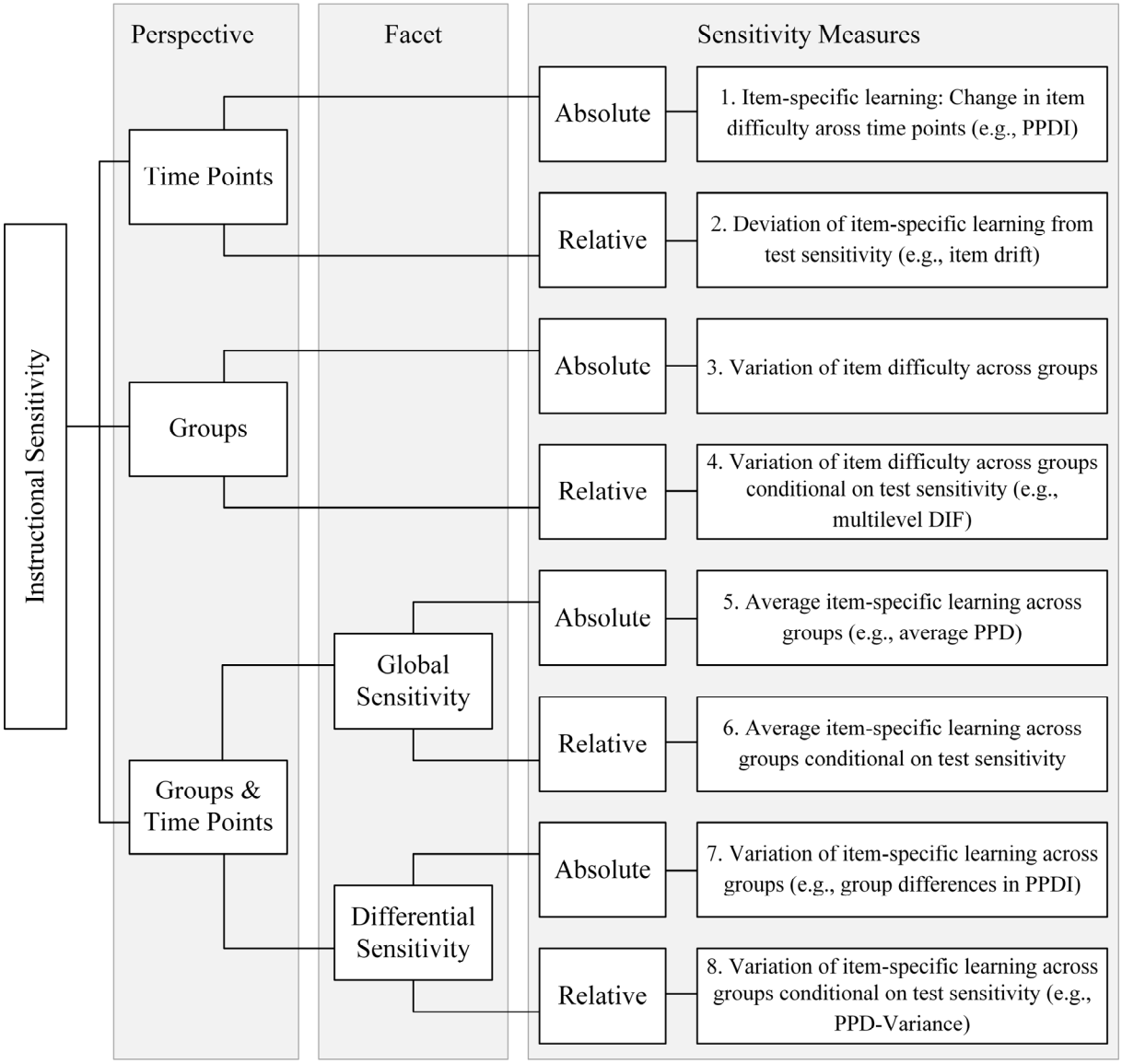


Abbildung 1. Systematik zur Verortung der Verfahrensweisen zur Messung von Instruktionssensitivität (Naumann et al., 2017, S. 681).

Ein weiterer Aspekt zur Einordnung der Verfahren zur Evaluation der Instruktionssensitivität liegt in der Unterscheidung zwischen relativer und absoluter Sensitivität. Die Autoren arbeiten heraus, dass mit der Gesamtheit an Verfahrensweisen zwei unterschiedliche Zielsetzungen verfolgt werden können (ebd). Während ein Teil der Verfahren der Überprüfung der absoluten Sensitivität dient, eignet sich der andere Teil zur Untersuchung der relativen Sensitivität. Mit der absoluten Sensitivität wird erfasst, inwieweit ein einzelnes Item unter Berücksichtigung der Zeitpunkte-, der Gruppen oder der Zeitpunkte-x-Gruppen-Perspektive absolut sensitiv ist (ebd.). D.h.: Je mehr der Wert von null abweicht, desto höher ist die angenommene Sensitivität eines Testitems. Mit der relativen Sensitivität wird hingegen erfasst, inwiefern ein Testitem

relativiert an der Sensitivität eines Tests sensitiv ist (ebd.). Letzteres bedeutet: Ist das Item ähnlich sensitiv wie der Test (durchschnittliche Sensitivität aus der Gesamtheit an Testitems), nimmt dieses einen Wert von null an (ebd.). Je stärker die Sensitivität eines Testitems von der Testsensitivität abweicht, desto stärker weicht der Wert auch von null ab. Anhand eines relativen Maßes kann folglich nicht auf die allgemeine Instruktionssensitivität geschlossen werden. Darüber hinaus ist mit Bezug zu Abbildung 1 anzumerken, dass nicht alle aufgeführten Verfahren bisher Anwendung erfahren haben. Teilweise handelt es sich auch um eine theoretische Erweiterung der Verfahrensweisen (ebd.).

2.4.2 Systematisierung der Verfahrensweisen gemäß den Informationsquellen

Folgt man den Gedanken von Polikoff (2010), können drei Ansätze zur Evaluation der Instruktionssensitivität differenziert werden: (1) ein Ansatz basierend auf reinen Itemstatistiken, (2) ein Ansatz basierend auf Itemstatistiken unter Hinzuziehung von Unterrichtsmerkmalen und (3) ein Ansatz basierend auf Expertinnen- und Expertenratings.

Reine Itemstatistiken

Für den Ansatz der reinen Itemstatistiken ist zunächst charakteristisch, dass für die Evaluation der Instruktionssensitivität ausschließlich Leistungstestdaten der Schülerinnen und Schüler herangezogen werden. Zahlreiche Studien zur Evaluation der Instruktionssensitivität unter Verwendung reiner Itemstatistiken liegen vor (z.B. Brennan, 1972; Brennan & Stolurow, 1971; Cox & Vargas, 1966; Haladyna, 1974; Helmstadter, 1972; Kosecoff & Klein, 1974). Basierend auf diesen Leistungstestdaten wird die statistische Variabilität in den Itemantworten überprüft. Hierfür werden Maße wie die Itemschwierigkeit herangezogen (Haladyna & Roid, 1981). Das prominenteste Verfahren, das dem Bereich der reinen Itemstatistiken zuzuordnen ist, stellt der PPDI dar (Cox & Vargas, 1966). Der PPDI entspricht der Differenz aus dem Anteil an Schülerinnen und Schülern, die das Testitem im Posttest korrekt gelöst haben, und dem Anteil an Schülerinnen und Schülern, die das Item im Pretest richtig beantwortet haben (ebd.). Darüber hinaus wurden weitere – dem PPDI sehr ähnliche – Verfahren entwickelt. Hierzu zählen beispielsweise der *normalized difference between item difficulty* (ZDIFF, Haldyna & Roid, 1981; siehe auch Polikoff, 2010), der *percentage of possible gain* (Brennan & Stolurow, 1971) und der *Brennan index* (Brennan, 1972). Eine weitere Gruppe, welche den Verfahren der reinen

Itemstatistiken zuzuordnen ist, sind die *Kontingenztabellen* (siehe z.B. Kosekoff & Klein, 1974). Mithilfe der Kontingenztabellen wird eine Verteilung korrekter und falscher Itemantworten, getrennt nach Pre- und Posttest, visualisiert, welche Hinweise bezogen auf die Sensitivität von Items liefern. Darüber hinaus stellen Verfahren basierend auf punktbiseralen Korrelationen eine weitere Gruppe dar (Haladyna, 1974). Als Beispiel können der *combined samples point-biseral index* (COMPBI; Haladyna, 1974) und der *internal sensitivity index* (ISI; Kosecoff & Klein, 1974) genannt werden, die sich aber kaum durchgesetzt haben. Zu erwähnen sind außerdem Verfahren, die dem bayesianischen Kontext (Helmstadter, 1974) oder auch dem Bereich des *differential item functioning* (Linn & Harnisch, 1981) zuzuordnen sind.

Itemstatistiken unter Berücksichtigung von Unterrichtsmerkmalen

Mithilfe reiner Itemstatistiken kann allerdings nicht sichergestellt werden, ob die statistische Variabilität in den Itemantworten wirklich auf den Unterricht zurückzuführen ist (van der Linden, 1981). Insofern wurden Verfahren entwickelt, welche auf Itemstatistiken basieren und zusätzlich Unterrichtmerkmale mit einbeziehen. Stehen spezifische Unterrichtsmerkmale mit der statistischen Variabilität in den Itemantworten in Beziehung, spricht dies für die Instruktionssensitivität eines Testitems. Erste Forschungsarbeiten fokussierten insbesondere auf die *Behandlung* (z.B. Haladyna & Roid, 1981; siehe auch Chen, 2012; Li et al., 2017; Muthén et al., 1991) bzw. die *Abdeckung von testrelevanten Unterrichtsinhalten* (*content coverage*) (Mehrens & Phillips, 1987; Yoon, Burstein, Chen & Kim, 1989; siehe auch Gamoran et al., 1997[2]). Aber nicht nur die Abdeckung von Unterrichtsinhalten, sondern auch die *Quantität hinsichtlich der Behandlung testrelevanter Inhalte bzw. von Klassenaktivitäten* wurden in Studien berücksichtigt, um diese Daten in die Analysen zur Evaluation der Instruktionssensitivität einzubeziehen (u.a. Hanson et al., 1986; Wiley & Yoon, 1995; Yoon & Resnick, 1998). Erfasst wurde beispielsweise, in wie vielen Lektionen Lehrpersonen spezifische Inhalte verfolgten (Hanson et al., 1986). Ein ähnliches Konzept stellt auch die *Gewichtung* (*emphasis*) *testrelevanter Inhalte* dar. Je nachdem, wie stark ein Inhalt gewichtet oder wie intensiv ein Lernziel verfolgt wurde, sollte sich dies im Antwortverhalten der Schülerinnen und Schüler auf spezifische Weise widerspiegeln, so die Annahme (D'Agostino et al., 2007; Muthén et al., 1995). Darüber hinaus erfolgte in den letzten zwei Dekaden

[2] In der Studie geht es nicht primär um die Überprüfung der Instruktionssensitivität, sondern um den Nachweis von Unterrichtseffekten auf die Leistungen der Schülerinnen und Schüler. Die Studie wird häufig in der Literatur zur Instruktionssensitivität herangezogen, um die Einbeziehung von Unterrichtsvariablen zu bekräftigen.

insbesondere die Berücksichtigung von Merkmalen zur *Unterrichtsqualität* (Ing, 2008, 2017; Naumann, Rieser et al., 2019; Polikoff, 2016; Polikoff & Porter, 2014; Popham, 2007). Hierbei wurden Unterrichtsmerkmale einbezogen, die beispielsweise auf die kognitive Aktivierung oder das Klassenmanagement abzielten (Ing, 2017; Naumann, Rieser et al., 2019). Des Weiteren wurden Merkmale des Unterrichts in die Analysen einbezogen, welche u.a. Aspekte der Didaktik (Yoon & Resnick, 1998) oder auch die Nähe der Testitems zum Unterricht fokussierten (Ruiz-Primo et al., 2012). Die beiden letztgenannten Gesichtspunkte kamen aber bisher nur sehr vereinzelnd bei der Evaluation der Instruktionssensitivität zum Tragen.

Expertinnen- und Expertenratings

Neben den zwei vorgestellten Ansätzen (reine Itemstatistiken, Itemstatistiken unter Hinzuziehung von Unterrichtsmerkmalen) zur Evaluation der Instruktionssensitivität von Schulleistungstests gibt es einige wenige Studien, welche mithilfe von Expertinnen- und Expertenratings das Ziel verfolgen, die Instruktionssensitivität von Testitems einzuschätzen. Popham (2007) war der erste, der Expertinnen- und Expertenratings als Ansatz zur Evaluation der Instruktionssensitivität in Betracht gezogen hat, gefolgt von Arbeiten von Chen (2012) und Musow et al. (2019). Aufgrund der wenigen Studien geht es bisher primär um die Frage, inwiefern Expertinnen und Experten überhaupt in der Lage sind, die Instruktionssensitivität von Testitems einzuschätzen. Verschiedene Rating-Instrumente wurden entwickelt und im Rahmen empirischer Studien eingesetzt. Das Rating von Chen (2012) basiert beispielsweise auf einer Globaleinschätzung, indem die Expertinnen und Experten gefragt wurden: *„Wie instruktionssensitiv ist die jeweilige Aufgabe?“* (S. 124, eigene Übersetzung). Bei Musow et al. (2019) wurde zum einen in Anlehnung an Chen (2012) ein globales Rating und zum anderen ein neu entwickeltes differenziertes Rating durchgeführt. Das differenzierte Rating zeichnet sich durch vier Aspekte aus, anhand derer ein instruktionssensitives von einem nicht instruktionssensitiven Testitem unterschieden werden soll. Ein Beispiel für einen solchen Aspekt stellen die Unterrichtsinhalte dar. Im Hinblick darauf nehmen die Expertinnen und Experten eine Einschätzung zu folgender Aussage vor: *„Die Schülerinnen und Schüler können die Aufgabe erst lösen, nachdem das Thema unterrichtet wurde.“* Erfasst wird, inwiefern der Aussage mit Blick auf das jeweilige Testitem zugestimmt wird. Um die Ergebnisse aus den Expertinnen- und Expertenratings zu validieren, wurden diese mit Ergebnissen basierend auf Itemstatistiken (Korrelationen, Regressionsanalysen, Cochran-Mantel-Haenszel-Test) abgeglichen. In den bisherigen Studien wird von moderaten Korrelationen zwischen den

Ergebnissen der Expertinnen- und Expertenratings und den Itemstatistiken berichtet (Chen, 2012; Musow et al., 2019).

Angesichts der vorgestellten Verfahrensweisen (reine Itemstatistiken, Itemstatistiken unter Hinzuziehung von Unterrichtsmerkmalen und Expertinnen- und Expertenratings; siehe Polikoff, 2010) stellt sich die Frage, welches für die Evaluation der Instruktionssensitivität von Tests und Testitems zu präferieren ist. Die Anwendung von Verfahren basierend auf reinen Itemstatistiken bringt den Nachteil mit sich, dass die statistische Variabilität in den Itemantworten auf unterrichtsbezogene Merkmale attribuiert, diese jedoch nicht empirisch geprüft wird. Diesem Kritikpunkt bezogen auf Verfahrensweisen reiner Itemstatistiken kann mittels Einbeziehung von Unterrichtsmerkmalen entgegnet werden. Eine Herausforderung, die sich jedoch dabei ergibt, ist der Nachweis einer Beziehung zwischen der statistischen Variabilität in den Itemantworten und den unterrichtsbezogenen Merkmalen. Letzteres gelang in einigen Studien (Chen, 2012; Gamoran et al., 1997; Yoon & Resnick, 1998), in anderen nur teilweise (Polikoff, 2016; Yoon et al., 1989) oder gar nicht, da die erwarteten Effekte ausblieben (Ing, 2008, 2017; Polikoff & Porter, 2014; Yoon et al., 1989). Darüber hinaus muss bei diesem Verfahren die Frage gestellt werden, welche Unterrichtsmerkmale vor dem Hintergrund des Erkenntnisinteresses als angemessen erachtet werden, um die Instruktionssensitivität von Tests und Testitems zu überprüfen. Die Expertinnen- und Expertenratings haben hingegen den Vorteil, dass sie sich als deutlich ökonomischer als die zwei anderen Ansätze erweisen könnten. Hintergrund ist, dass eine Erhebung von Leistungstestdaten der Schülerinnen und Schüler entweder deutlich reduziert oder auch ganz wegfallen könnte. Um Expertinnen- und Expertenratings zur Evaluation der Instruktionssensitivität gezielt einsetzen zu können, braucht es aber noch weiterer Forschung.

2.5 Studien zur Instruktionssensitivität und die (Nicht-)Berücksichtigung lernrelevanter Merkmale von Schülerinnen und Schülern

Im vorherigen Kapitel zur Systematisierung der Verfahrensweisen (Kapitel 2.4.2) wurde auf die Einbeziehung unterrichtsbezogener Merkmale hingewiesen, um sicherzustellen, dass die statistische Variabilität in den Itemantworten auch auf den Unterricht zurückzuführen ist. Allerdings ist davon auszugehen, dass die Testleistungen der Schülerinnen und Schüler nicht nur durch den Unterricht bedingt werden (siehe u.a. Carroll, 1963; Shulman, 1982). Lernrelevante Merkmale der Schülerinnen und Schüler (Brühwiler et al., 2017; Hosenfeld,

Helmke & Schrader, 2002) können die Ergebnisse von Schulleistungstests ebenfalls beeinflussen, was anhand zahlreicher Studien aus dem Bereich der Schul- und Unterrichtseffektivitätsforschung deutlich wird (Kapitel 3). Inwiefern es in Studien zur Instruktionssensitivität von Tests und Testitems Hinweise darauf gibt, dass die Einbeziehung von lernrelevanten Merkmalen der Schülerinnen und Schüler von Relevanz sein könnte, soll nachfolgend aufgezeigt werden. Darüber hinaus soll ein Überblick gegeben werden, welche lernrelevanten Merkmale der Schülerinnen und Schüler in welchem Ausmaß in Studien zur Evaluation der Instruktionssensitivität von Tests und Testitems Berücksichtigung gefunden haben.

2.5.1 Hinweise in der Literatur

Der Begriff der lernrelevanten Merkmale der Schülerinnen und Schüler umfasst in der vorliegenden Arbeit zwei Aspekte: Zum einen sind damit die *individuellen Lernvoraussetzungen* von Schülerinnen und Schülern gemeint, die das Lernen beeinflussen (Brühwiler et al., 2017; Hosenfeld et al., 2002; Wang, Haertel & Walberg, 1993). Hierzu zählen beispielsweise das Vorwissen, die Lernmotivation, die Lernstrategien sowie das Selbstkonzept der Schülerinnen und Schüler. Zum anderen werden damit *demografische Merkmale* von Schülerinnen und Schülern beschrieben, die ebenfalls mit einem Einfluss auf das Lernen in Verbindung gebracht werden (Brühwiler et al., 2017; Sutton & Soderstrom, 1999; Wang et al., 1993). Als Beispiel können das Alter, das Geschlecht, die Ethnie oder auch der SES genannt werden (Polikoff & Porter, 2014; Ruiz-Primo et al., 2012; Yoon et al., 1989).

In der Literatur zur Instruktionssensitivität von Tests und Testitems lassen sich vereinzelt Hinweise finden, die auf die Bedeutung der Einbeziehung von lernrelevanten Merkmalen der Schülerinnen und Schüler im Rahmen der Evaluation der Instruktionssensitivität hindeuten.

> The issues highlighted are that appropriate models for student performance and proper contextualization through the collection of auxiliary about students and their instructional programs are essential in the design, analysis and interpretation of multipurpose instructional assessments. Results from instructional assessments reflect both demographic and ability attributes of the students and instructional intents, opportunities, and execution of the setting in which students reside; moreover, these accrue and change over. Thus one needs to be cognizant of these

> circumstances and sensitive to their consequences for the validity of test interpretation. (Burstein, 1989, S. 7)

Die Aussage von Burstein (1989) scheint in Anbetracht empirischer Unterrichtsforschung und Modellen zur Erklärung von Schulleistungen wie den Angebots-Nutzungs-Modellen (Brühwiler, 2014; Brühwiler et al., 2017; Fend, 2002; Reusser & Pauli, 2010) immer noch aktuell zu sein. Auch Naumann et al. (2016) weisen beispielsweise in ihrem Artikel zur Instruktionssensitivität darauf hin, dass schulische Leistungen nicht ausschließlich auf die Instruktion zurückgeführt werden können: „Variation between groups on the item level or the test level may not be attributable solely to instruction, but may originate in other sources such as group composition (e.g. ability sorting)“ (ebd., S. 95).

Sowohl im Beitrag von Burstein (1989) als auch bei Naumann et al. (2016) wurde der Einfluss von lernrelevanten Merkmalen der Schülerinnen und Schüler auf Leistungstestdaten aufgegriffen. Eine empirische Überprüfung blieb in diesen beiden Fällen jedoch aus.[3] Die einzige Studie, in welcher der Einfluss lernrelevanter Merkmale der Schülerinnen und Schüler auf die Indikatoren der Instruktionssensitivität im Fokus stand und systematisch untersucht wurde, stammt von Muthén et al. (1991). Die Autoren bezogen Merkmale der Schülerinnen und Schüler wie das Geschlecht, die Ethnie, die familiären Hintergrundmerkmale, das fachspezifische Vorwissen und die Einstellung zum Fach Mathematik in die Analysen mit ein. Überprüft wurde, ob sich die Ergebnisse zur Instruktionssensitivität von Testitems mit und ohne Kontrolle für lernrelevante Merkmale der Schülerinnen und Schüler unterschieden. Muthén et al. (1991) kamen zu dem Ergebnis:

> When using the SIMS [Second International Mathematics Study d. Verf.] data to illustrate the approach to assessing instructional item sensitivity, Muthén (1989) included the OTL variables only and not the other background variables used here. If we assume that our present model including such background variables is true, omitting these other background variables would lead to biased estimates of the item parameters and their instructional sensitivity. (Muthén et al., 1991, S. 13)

In der Studie von Muthén et al. (1991) geht es also um die Frage von Verzerrungen in den Parameterschätzungen, welche aus den Analysen unter Berücksichtigung der getroffenen Modellannahmen und der Ergebnisinterpretation resultieren können. Mit den Verzerrungen

[3] Im nachfolgenden Kapitel 2.5.2 wird auf Studien zur Instruktionssensitivität, in denen lernrelevante Merkmale Berücksichtigung finden, noch näher eingegangen.

sind unerwünschte Einflüsse sogenannter Störvariablen auf eine oder mehrere Variablen gemeint (Eid, Gollwitzer & Schmitt, 2015; Elwert, 2013; Wu, Liu, Stone, Zou & Zumbo, 2017). Eid et al. (2013) unterscheiden dabei zwischen systematischen und unsystematischen Störvariablen. Von einer *unsystematischen Störvariable* kann gesprochen werden, wenn diese einen Einfluss nur auf die abhängige und nicht auf die unabhängige Variable nimmt. Hat die Störvariable sowohl auf die abhängige als auch auf die unabhängige Variable einen Einfluss, wird diese als eine *systematische Störvariable* bezeichnet (Eid et al., 2013; siehe auch Elwert, 2013). Wu et al. (2017) weisen zudem darauf hin, dass Drittvariablen nicht per se kontrolliert werden müssen, sondern nur, wenn diese einen unerwünschten Einfluss auf andere Variablen nehmen. Bezugnehmend auf die Studie von Muthén et al. (1991) wurden die lernrelevanten Merkmale der Schülerinnen und Schüler als Kovariaten ins Modell mit aufgenommen. Muthén et al. (1991) berichten dabei von kleinen Verzerrungseffekten in den Ergebnissen, die bei fehlender Berücksichtigung lernrelevanter Merkmale von Schülerinnen und Schülern zum Tragen kommen. In ihrer Studie konnten die Mathematikleistungen im Posttest zu 76% durch die einbezogenen lernrelevanten Merkmale der Schülerinnen und Schüler erklärt werden (ebd.). Die lernrelevanten Merkmale scheinen also einen nicht unerheblichen Einfluss auf die Mathematikleistungen zu nehmen, auch wenn sich die Autoren dafür aussprechen, die Bedeutsamkeit dieser Merkmale der Schülerinnen und Schüler im Hinblick auf die Evaluation der Instruktionssensitivität nicht zu überschätzen (ebd.).

2.5.2 Studien zur Instruktionssensitivität

Um der Frage nachzugehen, welche lernrelevanten Merkmale der Schülerinnen und Schüler in welchem Ausmaß in Studien zur Evaluation der Instruktionssensitivität von Tests und Testitems Berücksichtigung finden, wird nachfolgend der gegenwärtige Stand der einschlägigen Literatur aufgezeigt. Bei der Darstellung soll mit den am häufigsten genannten demografischen Merkmalen in Artikeln zur Instruktionssensitivität begonnen werden. Im Anschluss daran wird auf die individuellen Lernvoraussetzungen eingegangen.

Ein erstes demografisches Merkmal der Schülerinnen und Schüler, das bei der Evaluation der Instruktionssensitivität von Tests und Testitems mehrfach genannt wurde, betrifft den SES (D'Agostino et al., 2007; Mehrens & Phillips, 1986; Muthén et al., 1995; Muthén et al., 1991; Polikoff, 2016; Polikoff & Porter, 2014; Traynor, 2017; Yoon & Resnick, 1998). Der SES wird über die Bildung, den Beruf und/oder das Einkommen der Eltern operationalisiert, wobei die

Indikatoren zur Erfassung des SES erheblich variieren. Ein zweites Merkmal, welches in den Studien zur Evaluation der Instruktionssensitivität von Tests und Testitems Berücksichtigung findet, betrifft die Ethnie (D'Agostino et al., 2007; Muthén et al., 1995; Muthén et al., 1991; Polikoff, 2016; Polikoff & Porter, 2014; Traynor, 2017; Yoon & Resnick, 1998). Hintergrund ist, dass die Studien zur Instruktionssensitivität zumeist aus dem angloamerikanischen Raum stammen. Das schulische Lernen wird hier im Zusammenhang mit der Ethnie regelmäßig thematisiert und beispielsweise vor dem Hintergrund möglicher Benachteiligungen ethnischer Minderheiten diskutiert (Kapitel 2.2, Debra P. v. R. D. Turlington, 1979/1984). Ein drittes demografisches Merkmal der Schülerinnen und Schüler stellt den sprachlichen Hintergrund dar (D'Agostino et al., 2007; Li et al., 2017; Muthén et al., 1995; Polikoff, 2016; Polikoff & Porter, 2014; Traynor, 2017). Hierbei geht es um die Frage, ob die Landessprache für die jeweilige Schülerin bzw. den jeweiligen Schüler die Muttersprache oder eine Zweit- bzw. Fremdsprache darstellt. Ein weiteres und gleichzeitig letztes demografisches Merkmal der Schülerinnen und Schüler, das häufig in Studien zur Instruktionssensitivität Berücksichtigung findet, ist das Geschlecht (Li et al., 2017; Muthén et al., 1995; Polikoff, 2016; Polikoff & Porter, 2014). Letzteres ist darauf zurückzuführen, dass Leistungsunterschiede von Schülerinnen und Schülern auch im Zusammenhang mit dem Geschlecht diskutiert werden (siehe u.a. Hyde, Linderg, Linn, Ellis & Williams, 2008).

Bezugnehmend auf die Merkmale der individuellen Lernvoraussetzungen der Schülerinnen und Schüler ist zunächst die fachspezifische Vorleistung zu nennen. Dieses Merkmal nimmt im Kontext der Instruktionssensitivität eine Sonderstellung ein, da zahlreichen Studien die Zeitpunkte-Perspektive zugrunde liegt. Die fachspezifischen Vorleistungen werden dementsprechend in den Analysen der meisten Studien – zumindest indirekt – berücksichtigt (Brennan & Stolurow, 1971; Cox & Vargas, 1966; D'Agostino et al., 2007; Haladyna, 1974; Haladyna & Roid, 1981; Helmstadter, 1972, 1974; Hsu, 1971; Ing, 2008; Kosecoff & Klein, 1974; Mehrens & Phillips, 1986; Muthén et al., 1995; Muthén et al., 1991; Naumann, 2013; Naumann et al., 2020; Naumann et al., 2017; Naumann et al., 2014; Naumann, Rieser et al., 2019; Polikoff & Porter, 2014; Ruiz-Primo et al., 2012; Ruiz-Primo et al., 2002; Traynor, 2017; Yoon et al., 1989; Yoon et al., 1991). In den Studien von Naumann et al. (2020), Naumann, Rieser et al. (2019), Polikoff (2016) sowie Polikoff und Porter (2014) wird zudem für Merkmale der Intelligenz bzw. der Begabung kontrolliert. Muthén et al. (1991) beziehen die Einstellung zum Fach Mathematik in die Analysen mit ein, welche ebenfalls als individuelle Lernvoraussetzung von Schülerinnen und Schülern erachtet werden kann.

Darüber hinaus kann konstatiert werden, dass die in diesem Unterkapitel genannten Studien zahlreichen empirischen Studien zur Instruktionssensitivität gegenüberstehen, in denen auf eine Einbeziehung lernrelevanter Merkmale verzichtet wird (Brennan, 1972; Roudabush, 1974; Linn & Harnisch, 1981; Hanson et al., 1986; Mehrens & Phillips, 1987; Wiley & Yoon, 1995; Chen, 2012; Grossmann, Cohen, Ronfeld & Brown, 2014; Ing, 2016, 2017; Deutscher & Winther, 2018). Die Einbeziehung lernrelevanter Merkmale der Schülerinnen und Schüler bei der Evaluation der Instruktionssensitivität von Tests oder Testitems stellt folglich keine Selbstverständlichkeit dar. Hinsichtlich der Betrachtung der genannten Studien, in denen lernrelevante Merkmale eine Rolle spielen, ist festzuhalten, dass die von Muthén et al. (1991) einen besonderen Stellenwert einnimmt, da in ihr untersucht wurde, inwiefern sich die Sensitivitätsmaße mit und ohne Berücksichtigung lernrelevanter Merkmale der Schülerinnen und Schüler unterscheiden (Kapitel 2.5.1). Sie kann folglich als die zentrale Studie erachtet werden, wenn es um Hinweise zur Relevanz von Merkmalen der Schülerinnen und Schüler bei der Evaluation der Instruktionssensitivität geht.

2.6 Zusammenfassung und Fazit

In Kapitel 2 wurde in das Konzept der Instruktionssensitivität von Tests und Testitems eingeführt, mit dem Ziel, die für die vorliegende Arbeit zentralen Aspekte der Instruktionssensitivität zu erläutern. Als zentrale Aspekte wurden zunächst die Begriffsklärung zur Sicherstellung eines gemeinsamen Verständnisses und die Ausführungen zum Ursprung der Instruktionssensitivität erachtet. Letztere dienten insbesondere der Einbettung des Konzepts in einen Kontext, nämlich dem kriterienorientierten Testen im Rahmen eines *Accountability-Systems*. Mit diesem Kontext wird ein erstes Handlungsfeld ersichtlich, in dem die Instruktionssensitivität von Relevanz ist. Weitere Kontexte wurden herausgearbeitet. Dabei war es ein wichtiges Anliegen, die relevanten Kontexte bezogen auf den deutschsprachigen Raum zu identifizieren und zu beschreiben. Mit der Klärung der Relevanz der Instruktionssensitivität ging die Frage einher, wie die Instruktionssensitivität von Tests und Testitems überprüft werden kann. Dazu wurde ein Überblick über verschiedene Operationalisierungsansätze basierend auf Varianzquellen gegeben. Zu unterscheiden sind die Zeitpunkte-, die Gruppen- und die Zeitpunkte-x-Gruppen-Perspektive. Des Weiteren wurden Verfahrensweisen zur Überprüfung der Instruktionssensitivität gemäß den Informationsquellen kategorisiert. Die Verfahren können dabei in reine Itemstatistiken, Itemstatistiken unter Hinzuziehung von

Unterrichtsmerkmalen und Expertinnen- und Expertenratings eingeteilt werden. Mit den Ausführungen zu den Itemstatistiken unter Hinzuziehung der Unterrichtsmerkmale wurde aber auch deutlich, dass es nicht immer gelingt, die statistische Variabilität in den Itemantworten auf die einbezogenen Unterrichtsmerkmale zurückzuführen. Ein Erklärungsansatz besteht darin, dass Unterrichtsmerkmale nur bedingt die Itemantworten der Schülerinnen und Schüler erklären und weitere Aspekte, wie etwa lernrelevante Merkmale der Schülerinnen und Schüler bei der Evaluation der Instruktionssensitivität von Tests bzw. Testitems, eine Rolle spielen könnten. Bezugnehmend auf die lernrelevanten Merkmale kann hierbei zwischen demografischen Merkmalen und individuellen Lernvoraussetzungen unterschieden werden. Die demografischen Merkmale wurden (nur) teilweise in die Analysen zur Instruktionssensitivität einbezogen. Individuelle Lernvoraussetzungen blieben – mit Ausnahme der fachspezifischen Vorleistungen – häufig unberücksichtigt. Basierend auf dieser Erkenntnis des Literaturreviews soll der Versuch unternommen werden, die Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraussetzungen der Schülerinnen und Schüler konzeptionell einzubetten, da dies im Kontext der Evaluation der Instruktionssensitivität von Testitems noch aussteht. Es ist unklar, auf welche Ansätze, Modelle, Theorien und/oder empirische Erkenntnisse zurückgegriffen werden kann, um sich der Frage zu nähern, inwiefern individuelle Lernvoraussetzungen der Schülerinnen und Schüler die Indikatoren der Instruktionssensitivität beeinflussen.

3 Die Evaluation der Instruktionssensitivität unter Berücksichtigung individueller Lernvoraussetzungen der Schülerinnen und Schüler – Der Versuch einer konzeptionellen Einbettung

Um der Frage nachzugehen, inwiefern die Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraussetzungen der Schülerinnen und Schüler konzeptionell eingebettet werden kann, wird im Kapitel 3.1 zunächst auf die Beziehung zwischen der Evaluation der Instruktionssensitivität von Test und Testitems und der Schul- und Unterrichtsforschung eingegangen. Dieser Schritt ist notwendig, da für die konzeptionelle Einbettung der *Value-Added-Ansatz* und Angebots-Nutzungs-Modelle herangezogen werden. Der *Value-Added-Ansatz* wird in Kapitel 3.2 und die Angebots-Nutzungs-Modelle werden in Kapitel 3.3 beschrieben. Beide bilden zusammen einen Rahmen zur Konzeptualisierung der Evaluation der Instruktionssensitivität unter Berücksichtigung individueller Lernvoraussetzungen von Schülerinnen und Schülern. Anschließend werden in Kapitel 3.4 die individuellen Lernvoraussetzungen der Schülerinnen und Schüler fokussiert. Es geht um die Frage, was unter individuellen Lernvoraussetzungen der Schülerinnen und Schüler verstanden werden kann und inwiefern diese für das schulische Lernen von Relevanz sind. Auf zwei individuelle Lernvoraussetzungen wird dabei genauer eingegangen und es wird erläutert, wie diese mit der Instruktionssensitivität von Tests und Testitems in Beziehung stehen könnten. Neben den individuellen Lernvoraussetzungen werden zudem in Kapitel 3.5 die Unterrichtsmerkmale in den Blick genommen, die hinsichtlich der Evaluation der Instruktionssensitivität von Testitems Berücksichtigung finden. Es soll geklärt werden, welche spezifischen Unterrichtsmerkmale einbezogen werden und was unter den jeweiligen Unterrichtsmerkmalen verstanden werden kann. In Kapitel 3.6 werden dann die zentralen Aspekte aus diesem Kapitel nochmals zusammengefasst und ein Fazit gezogen. Anschließend erfolgen in Kapitel 3.7 die Ableitung der Fragestellung und die Erläuterung der Hypothesen, die dieser Arbeit zugrunde liegen.

3.1 Die Beziehung zwischen der Instruktionssensitivität von Testitems und der Schul- und Unterrichtseffektivitätsforschung

In dieser Arbeit wird auf den *Value-Added-Ansatz* und auf die Angebots-Nutzungs-Modelle zurückgegriffen, welche der Gestaltung eines konzeptionellen Rahmens zur Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraus-

setzungen von Schülerinnen und Schülern dienen. Der Rückgriff auf den *Value-Added-Ansatz* als einen Ansatz für Schul- und Unterrichtsevaluationen und die Angebots-Nutzungs-Modelle als ein Konzept zur Erklärung schulischer Leistungen liegt nahe, da diese der Evaluation der Instruktionssensitivität komplementär gegenüberstehen. In Abbildung 2 wird die Beziehung zwischen der Evaluation der Instruktionssensitivität von Tests und Testitems und der Schul- und Unterrichtseffektivitätsforschung sowie weiteren Handlungsfeldern, in denen basierend auf Testdaten Rückschlüsse über Schule und Unterricht gezogen werden, dargestellt. Die dort dargelegten Fragestellungen beziehen sich zum einen auf die Unterrichtseffektivitätsforschung (links in der Abbildung 2) und zum anderen auf die Evaluation der Instruktionssensitivität von Tests und Testitems (rechts in der Abbildung 2).

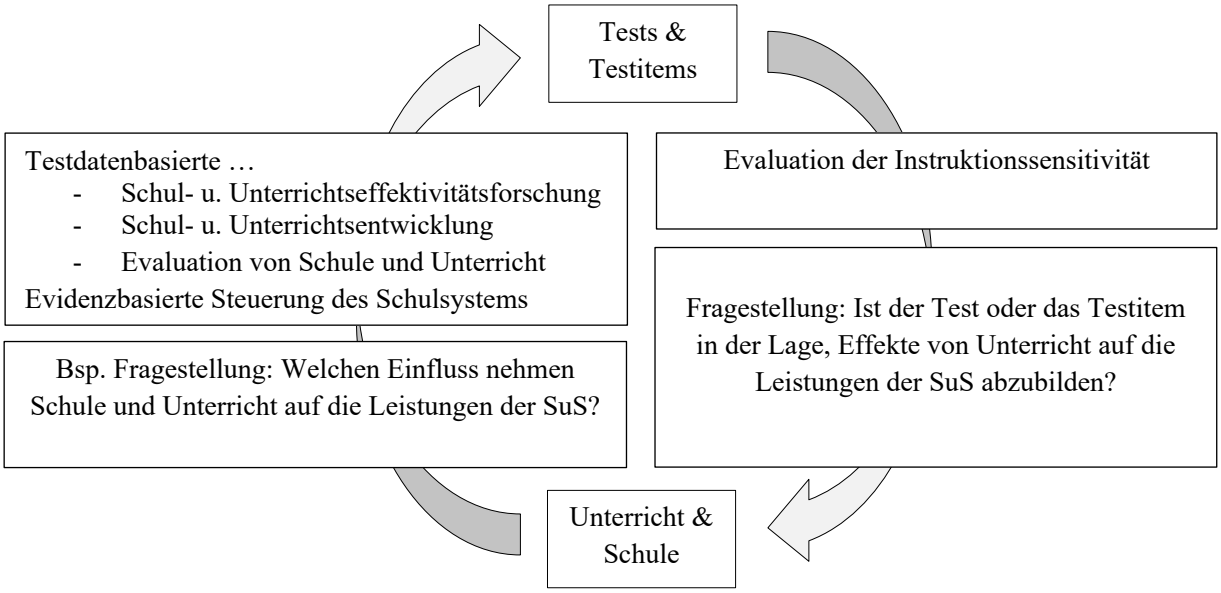


Abbildung 2. Beziehung zwischen der Evaluation der Instruktionssensitivität und der Schul- und Unterrichtseffektivitätsforschung sowie anderen Handlungsfeldern, in denen die Instruktionssensitivität von Relevanz ist (siehe auch Naumann et al., 2020).

Anmerkung: SuS: Schülerinnen und Schüler.

Während im Kontext der Evaluation der Instruktionssensitivität geprüft wird, ob Tests oder Testitems in der Lage sind, Effekte von Unterricht auf die Leistungen der Schülerinnen und Schüler abzubilden, werden beispielsweise im Kontext der Schul- und Unterrichtseffektivitätsforschung Leistungstestdaten der Schülerinnen und Schüler verwendet, um die Wirksamkeit von Schule und Unterricht zu überprüfen. Deutlich wird an dieser Stelle, dass sich die Ausgangspunkte der wissenschaftlichen Fragestellungen erheblich voneinander unterscheiden (Fokus auf Tests/Testitems vs. Schule/Unterricht). Zur Beantwortung der

jeweiligen Fragestellung werden aber in beiden Fällen häufig Daten aus standardisierten Schulleistungstests herangezogen. Insofern ist es eine Frage der Perspektive, die mit der Aussage, dass der *Value-Added-Ansatz* sowie die Angebots-Nutzungs-Modelle der Evaluation der Instruktionssensitivität komplementär gegenüberstehen, zum Ausdruck kommen soll.

3.2 Value-Added-Ansatz

Mit dem *Value-Added-Ansatz* wird im Bildungskontext das Ziel verfolgt, den durch die Lehrperson initiierten Leistungszuwachs von Schülerinnen und Schülern zu messen (Newton, Darling-Hammond, Haertel & Thomas, 2010). Die Idee dabei ist, Effekte des Unterrichts von Effekten, welche die Lehrperson nicht kontrollieren kann, zu isolieren (McPherson, 1993; Newton et al., 2010). Willms und Raudenbush (1989) unterscheiden in diesem Zusammenhang zwei Effekt-Typen: den Effekt-Typ-A und den Effekt-Typ-B (siehe auch Raudenbush & Willms, 1995). Der Effekt-Typ-A ist dienlich, um allgemeine Informationen zur Schuleffektivität zu erhalten, unabhängig davon, ob der Effekt beispielsweise auf die Unterrichtspraxis, auf die Kontextfaktoren der Schule oder auf Merkmale der Schülerinnen und Schüler zurückzuführen ist. So kann er beispielsweise für Eltern von Interesse sein, wenn diese wissen möchten, auf welcher Schule ihr Kind die besten Bedingungen hinsichtlich des Lernens hat (McPherson, 1993; Raudenbush & Willms, 1995; Willms & Raudenbush, 1989). Der Effekt-Typ-B ist hingegen hilfreich, um den ‚reinen', von der Lehrperson initiierten Unterrichtseffekt zu messen. Dieser kann von Interesse sein, wenn es um die Evaluation der Unterrichtspraxis bzw. der Lehrperson geht (McPherson, 1993; Raudenbush & Willms, 1995; Willms & Raudenbush, 1989). Mit dem Effekt-Typ-B soll beispielsweise ein möglichst fairer Vergleich bezogen auf Lehrpersonen bzw. deren Unterrichtspraxis ermöglicht werden, auch wenn die Voraussetzungen der Lehrpersonen, um Schülerinnen und Schüler zu möglichst guten Leistungen zu bringen, sehr unterschiedlich sind. Die methodische Umsetzung zur Analyse von Typ-B-Effekten erfolgt durch die Kontrolle möglicher Störeinflüsse. Die Auswahl der einzubeziehenden Kovariaten ist hierbei zentral. Ein diesbezüglich sehr eindrückliches Ergebnis, das die Relevanz einer gut durchdachten Einbeziehung von Kovariaten im Kontext des *Value-Added-Ansatzes* verdeutlicht, ist bei Paterson (1991) zu finden. Paterson (1991) hat mehrere Analysen bezogen auf das Bildungsniveau von Schülerinnen und Schülern am Ende der Sekundarstufe vorgenommen, bei denen die Einbeziehung von Kovariaten (Fähigkeiten und Fertigkeiten bei Eintritt in die Sekundarstufe, unterschiedliche Operationalisierungen des SES)

variiert wurde. Die Erkenntnis ist: Sowohl die Fähigkeiten und Fertigkeiten bei Eintritt in die Sekundarstufe als auch die verschiedenen Indikatoren des SES (z.B. Berufstätigkeit der Eltern, Bildungsniveau der Eltern) klären im unterschiedlichen Maße die Varianz im Bildungsniveau am Ende der Sekundarstufe auf (ebd.). Dementsprechend macht es beispielsweise einen Unterschied, ob und welche SES-Indikatoren in die Analysen einbezogen werden, wenn es im Rahmen des *Value-Added-Ansatzes* um Fragen des Typ-B-Effekts geht.

Mit der Intention, die Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraussetzungen zu evaluieren, stellt sich die Frage, wie individuelle Lernvoraussetzungen in den Modellen zur Evaluation der Instruktionssensitivität Berücksichtigung finden. Bezugnehmend auf den *Value-Added-Ansatz* könnten individuelle Lernvoraussetzungen als Kovariaten einbezogen werden, um diese im Sinne von Störeinflüssen zu kontrollieren. Gemäß Raudenbush und Willms (1995) würde diese Vorgehensweise dem Typ-B-Effekt entsprechen. Das Erkenntnisinteresse würde damit auf der Sensitivität von Testitems hinsichtlich eines ‚reinen', von der Lehrperson initiierten Unterrichtseffekt liegen. Nachfolgend wird dieser Effekt als *Instruktionssensitivität im engeren Sinne* bezeichnet. Dem gegenüber steht der Typ-A-Effekt, der ohne die Implementierung von Kovariaten auskommt. Bei ihm geht es um die Sensitivität der Items, unabhängig davon, ob diese auf das Handeln der Lehrperson, die Merkmale der Schülerinnen und Schüler und/oder auf die Interaktion beider Aspekte zurückzuführen ist. In der vorliegenden Arbeit wird dieser Effekt als *Instruktionssensitivität im weiteren Sinne* bezeichnet. Die Unterscheidung zwischen der Instruktionssensitivität im engeren oder weiteren Sinne[4] ist ein – im Rahmen dieser Arbeit – neu eingeführter Gedanke. Je nach intendierter Testwertnutzung und -interpretation wäre die Evaluation der Instruktionssensitivität dementsprechend unterschiedlich auszurichten.

3.3 Angebots-Nutzungs-Modelle

Das ursprüngliche Angebots-Nutzungs-Modell (in der Literatur auch *models of learning opportunities and uses of instruction*, Lipowsky et al., 2009; *multilevel supply-use model*, Brühwiler & Blatchford, 2011; *offer/take-up-model*, Göbel & Helmke, 2010; *utilization of learning opportunities model*, Seidel, 2014; Bieg et al., 2017 genannt) geht auf Fend (1981) zurück und wurde seither von verschiedenen Autoren adaptiert (u.a. Brühwiler, 2014;

[4] Alternativ könnte auch von einem bereinigten und einem unbereinigten (Unterrichts-)Effekt gesprochen werden.

Brühwiler & Blatchford, 2011; Fend, 2002; Helmke & Weinert, 1997; Lipowsky et al., 2009; Reusser & Pauli, 2010; Seidel 2014). In Abbildung 3 ist das mehrebenenanalytische Angebot-Nutzungs-Modell von Brühwiler (2014) als ein Beispiel für ein Angebots-Nutzungs-Modell aufgeführt. Der Grundgedanke dieses Ansatzes besteht darin, dass der Lernertrag von zwei Aspekten abhängt: zum einen davon, wie das Angebot gestaltet wird, und zum anderen davon, wie das Angebot genutzt wird (Fend, 2008). Bezogen auf den Schulkontext wird der Unterricht als das Lernangebot aufgefasst, welches durch die Lehrperson gestaltet wird. Die Nutzung von Lerngelegenheiten kann hingegen als individueller Verarbeitungsprozess der Schülerinnen und Schüler verstanden werden. Gemäß den Modellannahmen hängt der Lernertrag also nicht nur von der Qualität und/oder der Quantität des Lernangebots ab, sondern auch von den Schülerinnen und Schülern, die das Angebot auf unterschiedliche Weise nutzen. Den Schülerinnen und Schülern wird dabei im Sinne eines konstruktivistischen Lehr-Lern-Verständnisses eine aktive Rolle in der Aneignung der Welt zugeschrieben. Die Beziehung zwischen dem Lernangebot und dem Lernertrag wird folglich von den individuellen Lernvoraussetzungen und den individuellen Verarbeitungsprozessen bzw. den Lernaktivitäten direkt oder auch indirekt beeinflusst (Brühwiler, 2014; Brühwiler & Blatchford, 2011; Fend, 2002; Helmke, 2009; Helmke & Schrader, 2019; Seidel, 2014). Brühwiler (2014) führt bezogen auf die individuellen Lernvoraussetzungen die kognitiven, motivationalen, emotionalen und selbstbezogenen Kognitionen der Schülerinnen und Schüler an (siehe auch Brühwiler & Blatchford, 2011). Generell variieren jedoch die Ausführungen zu den individuellen Lernvoraussetzungen der Schülerinnen und Schüler innerhalb der Angebots-Nutzungs-Modelle (siehe auch Fend, 2002; Helmke, 2009; Seidel, 2014).

Neben dem Unterrichtsangebot, den individuellen Lernvoraussetzungen und den individuellen Verarbeitungsprozessen ist anhand der Abbildung 3 zu erkennen, dass weitere Faktoren direkt oder indirekt mit den Leistungen der Schülerinnen und Schüler in Beziehung stehen. Ein Aspekt, auf den an dieser Stelle noch eingegangen werden soll, sind die individuellen Lernumwelten. Dieser Aspekt wird relativ breit gefasst, indem die Familie und die Gleichaltrigen, aber auch die Medien hinzugezählt werden (Brühwiler, 2014; Brühwiler & Blatchford, 2011; Seidel, 2014). Zu den individuellen Lernumwelten gehören auch die demografischen Merkmale, die teilweise in Studien zur Evaluation der Instruktionssensitivität Berücksichtigung finden und in Kapitel 2.5.1 bereits erwähnt wurden. Mit Blick auf das Angebot-Nutzungs-Modell in Abbildung 3 wird davon ausgegangen, dass die demografischen Merkmale u.a. einen Einfluss auf die individuellen Lernvoraussetzungen der Schülerinnen

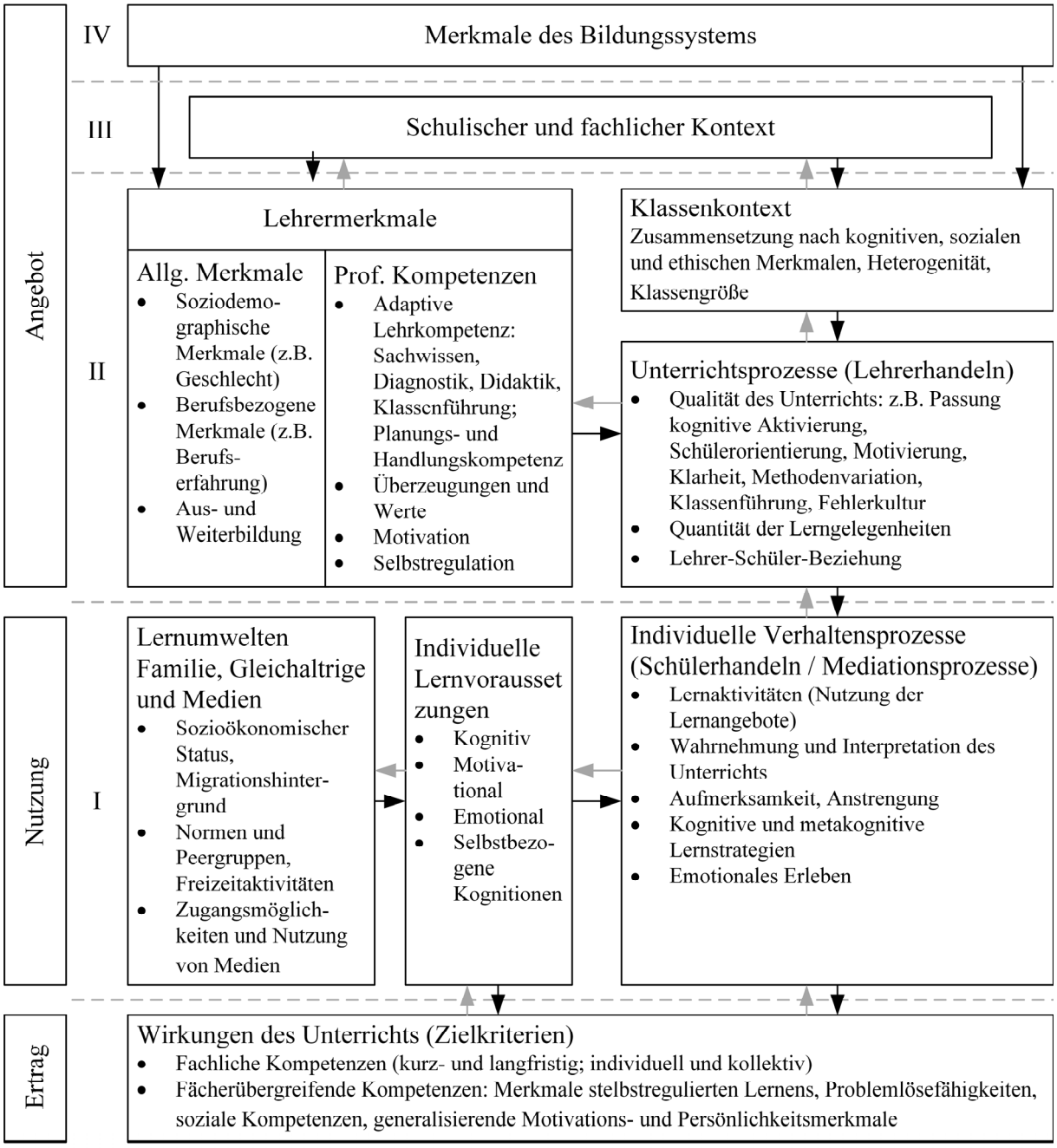


Abbildung 3. Mehrebenenanalytisches Angebots-Nutzungs-Modell des schulischen Lernens (Brühwiler, 2014, S. 23).

und Schüler nehmen (Brühwiler, 2014; siehe auch Brühwiler & Blatchford, 2011; Helmke & Schrader, 2019; Seidel, 2014). Insgesamt ist allerdings eine gewisse Vorsicht geboten, was die Darstellungen hinsichtlich der Wirkbeziehungen in Angebots-Nutzungs-Modellen anbelangt. Die Komplexität und der generische Charakter der Angebots-Nutzungs-Modelle bringen u.a. den Nachteil mit sich, dass diese im Sinne eines Modellcharakters stark vereinfacht sind und eine empirische Überprüfung des Gesamtmodells noch aussteht. Je nach Perspektive kann

konstatiert werden, dass Angebots-Nutzungs-Modelle jedoch bzw. lediglich durch theoretische Konstrukte fundiert werden, die im Kontext der Bildungsforschung als empirisch abgesichert gelten (Helmke, 2009). Spezifische Fragen zu den Wirkbeziehungen zwischen einzelnen Aspekten innerhalb der Angebots-Nutzungs-Modelle bleiben dabei aber sehr vage. Insofern werden in diesem Kapitel die Wirkbeziehungen innerhalb der Angebots-Nutzungs-Modelle nur auf einer sehr allgemeinen Ebene betrachtet. Erschwerend kommt hinzu, dass die genannten Lernvoraussetzungen in den Angebots-Nutzungs-Modellen variieren bzw. ausschließlich Überbegriffe genannt werden. Unklar bleibt, welche spezifischen Konstrukte hinsichtlich der individuellen Lernvoraussetzungen einzubeziehen sind. Die Ableitung konkreter Hypothesen ist dementsprechend anhand solcher Modelle kaum möglich (Helmke, 2009). Wird Letzteres intendiert, ist eine vertiefte Auseinandersetzung mit dem Modell, dessen Bestandteilen und deren Zusammenhängen unter Hinzuziehung von Theorien und empirischen Erkenntnissen unabdingbar. Nichtsdestotrotz liefern Angebots-Nutzungs-Modelle aber wertvolle Gedanken, die zur Evaluation der Instruktionssensitivität unter Berücksichtigung individueller Lernvoraussetzungen dienlich sind. Angebots-Nutzungs-Modelle unterstreichen den Gedanken, der bereits im Zusammenhang mit dem *Value-Added-Ansatz* zur Sprache kam: Leistungstestdaten der Schülerinnen und Schüler werden nicht ausschließlich von der Lehrperson bzw. dem Unterrichtsangebot bedingt, sondern auch von den Schülerinnen und Schülern, denen im Sinne eines konstruktivistischen Lehr-Lern-Verständnisses eine aktive Rolle zugeschrieben wird. Die Ausführungen liefern insofern erste Hinweise, wie individuelle Lernvoraussetzungen der Schülerinnen und Schüler einen Einfluss auf die Leistungstestdaten nehmen könnten. Inwieweit die individuellen Lernvoraussetzungen der Schülerinnen und Schüler die Indikatoren der Instruktionssensitivität beeinflussen, ist allerdings noch unklar.

3.4 Individuelle Lernvoraussetzungen der Schülerinnen und Schüler

Es gibt diverse Vorstellungen darüber, welche Faktoren auf Seiten der Schülerinnen und Schüler zur Erklärung von Schulleistungen heranzuziehen sind (siehe u.a. Brühwiler et al., 2017; Fend, 1981; Hattie, 2009; Helmke & Weinert, 1997; Wang et al., 1993). Um sich einen Überblick über die individuellen Lernvoraussetzungen der Schülerinnen und Schüler zu verschaffen, welche die schulischen Leistungen bedingen, kann u.a. auf die Determinanten von Schulleistungen zurückgegriffen werden (Brühwiler et al., 2017; siehe auch Fend, 1981; Hattie, 2009; Helmke & Weinert, 1997). Genannt werden kognitive, motivationale und volitionale

Merkmale (Brühwiler et al., 2017). Andere Autoren erachten die kognitiven, metakognitiven, motivationalen und affektiven Merkmale von Schülerinnen und Schülern als besonders relevant für das schulische Lernen (u.a. Wang et al., 1993). Da in der vorliegenden Arbeit nicht alle individuellen Lernvoraussetzungen aufgezeigt werden können, sollen zwei sehr zentrale herausgegriffen und näher ausgeführt werden. Hierbei handelt es sich zum einen um das fachspezifische Vorwissen als ein kognitives Merkmal und zum anderen um die Lernmotivation als ein motivationales Merkmal.

3.4.1 Fachspezifisches Vorwissen als ein kognitives Merkmal von Schülerinnen und Schülern

In zahlreichen Studien zur Evaluation der Instruktionssensitivität wird das fachspezifische Vorwissen indirekt kontrolliert (Kapitel 2.5). Was unter dem fachspezifischen Vorwissen verstanden werden kann und inwiefern dieses Merkmal die Leistungen der Schülerinnen und Schülern bedingt, soll nachfolgend geklärt werden.[5]

Unter dem Vorwissen soll im Rahmen dieser Arbeit die Wissensbasis von Schülerinnen und Schülern verstanden werden, welche zu einem bestimmten Zeitpunkt bei der Aneignung von neuem Wissen für die Schülerinnen und Schülern verfügbar ist (Dochy, 1992). Das Vorwissen stellt dabei zur Vorhersage schulischer Leistungen einen wichtigen Prädiktor dar (Ausubel, Novak & Hanesian, 1968; Dochy, 1992; Hattie, 2013; Kalyuga, 2007; Parkerson, Lornax, Schiller & Walberg, 1984[6]; Schraw, 2006; Schuler, Funke & Baron-Boldt, 1990[7]; Weinert, 1989; Weinert, Schrader & Helmke 1989). Dochy (1988a) nimmt an, dass das Vorwissen entscheidend dafür ist, inwiefern neue Informationen verstanden, gespeichert und genutzt werden können. Zur Verarbeitung neuer Informationen sind Assimilations- und Integrationsprozesse erforderlich, die dazu dienen, diese Informationen in bestehende Wissensstrukturen einzupflegen (Binder, Schmiemann, Theyßen, Sandmann & Sures, 2016; Hoz, Bowman & Kozminsky, 2001). Ziegler (2008) geht davon aus, dass das Vorwissen (operationalisiert über schulische Ausgangsleistungen) der wichtigste Prädiktor innerhalb der kognitiven Merkmale von Schülerinnen und Schülern zur Vorhersage von Schulleistungen ist und damit eine größere Vorhersagekraft aufweist als die Intelligenz. Die Annahme zur

[5] An dieser Stelle sei darauf hingewiesen, dass es neben dem fachspezifischen Vorwissen weitere kognitive Merkmale von Schülerinnen und Schülern gibt, die ebenfalls für die schulischen Leistungen prädiktiv sind. Als Beispiele können die Intelligenz oder auch die Lernstrategien genannt werden, auf deren Ausführung jedoch verzichtet wird.
[6] In der Studie wird das Vorwissen über die fachspezifischen Ausgangsleistungen operationalisiert.
[7] In der Studie wird das Vorwissen über die Schulnoten operationalisiert.

Bedeutung des Vorwissens kann u.a. mit dem Matthäus-Effekt (Merton, 1968) erklärt werden (Renkl, 1996). Demnach gehen im schulischen Kontext hohe Ausgangsleistungen mit einer schnelleren Leistungsentwicklung und geringe Ausgangsleistungen mit einer langsameren Leistungsentwicklung einher (Merton, 1986; Niklas, Segerer, Schmiedeler & Schneider, 2012). Darüber hinaus wird neben dem domänenübergreifenden auch die Bedeutung des domänenspezifischen Vorwissens betont (Dochy, 1988b). Binder et al. (2016) nehmen dabei an, dass die Erklärungskraft des domänenspezifischen Vorwissens in Abhängigkeit zum Fach und den verschiedenen Facetten des Vorwissens variiert. Wird das Vorwissen im Allgemeinen und nicht in Abhängigkeit zum Fach betrachtet, gibt die Metaanalyse von Hattie (2009) einen Anhaltspunkt, was den Einfluss dieses kognitiven Merkmals anbelangt. Er kommt in seinen Analysen zum Ergebnis, dass das Vorwissen mit einer Effektstäke von $d = .67$ einen nicht unerheblichen Einfluss auf die Schulleistungen nimmt.

Um das Vorwissen zu beschreiben, lassen sich in der Literatur verschiedene Kategorisierungen finden. Eine zentrale Unterscheidung ist die von Anderson (1980), der zwischen dem deklarativen und dem prozeduralen Wissen differenziert. Das deklarative Wissen („knowing what“) steht für das Wissen von Fakten, Strukturen, Prinzipien und Konzepten, während das prozedurale Wissen („knowing how“) dem Wissen von Handlungsweisen oder Methoden entspricht (Anderson, 1980, S. 222f.). Die Gedanken von Anderson (1980) wurden in zahlreichen Publikationen zum Vorwissen aufgegriffen und mit weiteren (Unter-)Kategorien in Verbindung gebracht. Auch Dochy (1988b) berücksichtigt in seinen Ausführungen zur Kategorisierung des domänenspezifischen Vorwissens das deklarative und das prozedurale Wissen. Insgesamt definiert er acht Kategorien, die sich wie folgt darstellen: (1) die Art und der Umfang des Vorwissens („nature and amount“), (2) die Abrufbarkeit des Vorwissens („availability“), (3) den Aufbau von kognitiven Strukturen („structuring“), (4) die Relevanz des Vorwissens bezogen auf den Lerngegenstand („relevance of information“), (5) das Lerntempo bzw. der zeitliche Umfang des Lernprozesses („pace or duration“), (6) die Beständigkeit bzw. Robustheit des Wissens bezogen auf Umstrukturierungsprozesse („durability of what is retained“), (7) die bisherige Bildung („previous education“) sowie (8) die Arbeitserfahrung und das Alter („work experience and age“) (Dochy, 1988b, S. 15). Mit dieser Einteilung ermöglichte er bereits Ende der 1980er Jahre eine sehr differenzierte Betrachtungsweise auf das Vorwissen, die später von ihm noch weiterentwickelt wurde (siehe u.a. Dochy, 1996). Eine weitere Kategorisierung stammt von Hailikari (2009), die auf den Gedanken von Anderson (1982, 1995) und Dochy (1992, 1996) aufbaute und zudem Gedanken von Bloom (1976) und Anderson

et al. (2001) zur Taxonomie kognitiver Lernzielen einbezog. Sie unterscheidet zwischen (1) „knowledge of facts“, (2) „knowledge of meaning“, (3) „integration of knowledge“ und (4) „application of knowledge“ (Hailikari, 2009, S. 25). Damit wird deutlich, dass das Vorwissen auf unterschiedlichen kognitiven Ebenen verortet werden kann. Die beiden erstgenannten Kategorien sind Hailikari (2009) zufolge vornehmlich dem deklarativen Wissen und die beiden letztgenannten Kategorien vornehmlich dem prozeduralen Wissen zuzuordnen.

Da im Rahmen dieser Arbeit das Vorwissen nur indirekt Berücksichtigung erfährt, sollen die Einblicke in Definition, Bedeutung und mögliche Kategorisierungen des Vorwissens an dieser Stelle genügen. Anders sieht es mit dem Merkmal der Lernmotivation aus, das im Rahmen dieser Arbeit im Fokus steht. Nachfolgend wird das motivationale Merkmal als ein weiterer Aspekt der individuellen Lernvoraussetzungen aufgegriffen und beschrieben.

3.4.2 Motivationale Merkmale der Schülerinnen und Schüler

Im Zusammenhang mit der Erklärung von Schulleistungen wird auch der *Motivation* von Schülerinnen und Schülern eine zentrale Rolle zugeschrieben. Zahlreiche Studien weisen darauf hin, dass sie in einer positiven Beziehung zu den schulischen Leistungen von Schülerinnen und Schülern steht (Boekaerts, 2007; Dresel, Backes & Lämmle, 2011; Kriegbaum, Jansen & Spinath, 2015; Schrader & Helmke, 2002). „Je höher die Lernmotivation, desto schneller wird eine Lernhandlung initiiert, gegen Widerstände oder bei Schwierigkeiten (langweilige Aufgabe, unlösbare Probleme) aufrechterhalten, und desto intensiver sind Lernengagement, Anstrengung und Persistenz“ (Schrader & Helmke, 2002, S. 266, siehe auch Dresel et al., 2011; Schunk, Meece & Pintrich, 2012). Der Lernmotivation wird infolgedessen eine hohe Relevanz zugeschrieben, wenn es um die Erklärung von Schulleistungen geht (siehe auch Krapp, 2003). Dabei ist darauf hinzuweisen, dass zahlreiche Theorien zur Motivation existieren und diverse Instrumente zur Erhebung von Motivationskonstrukten auf dem Markt sind (Boekaerts, 2007; Hoffmann, 2015; Krapp, 1999). Hoffman (2015) legt über 70 Motivationskonstrukte (z.B. Fähigkeitsselbstkonzept, Interesse, Zielorientierung) mit je einem exemplarischen Erhebungsinstrument dar. Boekaerts (2007) geht dabei davon aus, dass die jeweiligen Konstrukte miteinander in Beziehung stehen (siehe auch Dresel & Lämmle, 2011). Die Schaffung eines einheitlichen Rahmenmodells, welches die diversen Konstrukte miteinander in Beziehung setzt, wäre ein erstrebenswertes Ziel. Bislang existieren diesbezüglich nur erste Versuche der Systematisierung von Motivationskonstrukten, die

beispielsweise bei Krapp (1999) oder auch bei Heckhausen und Heckhausen (2010) aufgeführt werden.

Was die Vorhersagekraft verschiedenster Motivationskonstrukte für die mathematischen Testleistungen in PISA anbelangt, können die Analysen von Kriegbaum et al. (2015) Aufschluss geben. Einbezogen wurden Skalen zum Selbstkonzept, zur Selbstwirksamkeit, zum Interesse und zur Freude, zur instrumentellen Motivation und zur Zielorientierung. Basierend auf den querschnittlich und längsschnittlich angelegten Untersuchungen kamen die Autorinnen und Autoren zum Ergebnis, dass die meisten in dieser Studie eingesetzten Motivationsskalen PISA-Testergebnisse im Bereich Mathematik vorhersagen, selbst wenn die mathematischen Vorleistungen und die Intelligenz in den Analysen kontrolliert wurden. Kriegbaum et al. (2015) schlussfolgerten, dass die motivationalen Merkmale der Schülerinnen und Schülern prädiktiv für mathematische Leistungen sind. Die Bedeutsamkeit der Motivation zur Erklärung schulischer Leistungen wird zudem von den Ergebnissen der Meta-Analyse von Hattie und Zierer (2018) unterstrichen. Sie stellten fest, dass die Leistungsmotivation und die Leistungsorientierung mit einer Effektstäke von $d = .62$ einen zentralen Prädiktor für schulische Leistungen darstellen.

Mit der Darlegung erster empirischer Erkenntnisse zum Einfluss motivationaler Konstrukte auf die schulischen Leistungen soll nachfolgend der Blick auf die Motivationstheorien gerichtet werden. Zu ihren bekanntesten Vertretern dürften die Selbstbestimmungstheorie nach Ryan und Deci (2000), die Erwartungs-mal-Wert-Modelle nach Wigfield und Eccles (2000), das Erweiterte Kognitive Motivationsmodell nach Heckhausen und Reinberg (1980), das Risiko-Wahl-Modell nach Atkinson (1957), die Flow-Theorie nach Csikszentmihalyi (1990), die Theorien zur Zielorientierung, z.B. nach Dweck (1986), und die Interessenstheorien, wie z.B. die Person-Gegenstand-Konzeption nach Krapp (1992), zählen. In der vorliegenden Arbeit soll ein besonderes Augenmerk auf die Selbstbestimmungstheorie von Ryan und Deci (2000) gelegt werden, welche heute noch hohen Zuspruch erfährt, in international renommierten Journalen angeführt wird und u.a. im Zusammenhang mit den drei Basisdimensionen der Unterrichtsqualität (Klieme, Lipowsky, Rakoczy & Ratzka, 2006) und den Angebots-Nutzugs-Modellen (Seidel, 2014) aufgegriffen wird. In der Theorie von Ryan und Deci (2000) stehen die Gründe, aus denen eine Person heraus handelt, im Vordergrund (Motive einer Handlung bzw. einer Handlungssteuerung). Es kann sich leicht vorgestellt werden, dass Schülerinnen und Schüler im schulischen Kontext aus ganz unterschiedlichen Motiven heraus agieren (aus Angst

vor Bestrafung, aus Interesse etc.). Je nachdem, welches Motiv dem Handeln zugrunde liegt, ist die Qualität bezogen auf die Auseinandersetzung mit dem Lerngegenstand eine andere. Ryan und Deci (2000) argumentieren mit dem Grad an Selbst- bzw. Fremdbestimmung und arbeiten sechs Formen der Motivationsregulierung heraus: (1) die Amotivation, (2) die „externale Regulation“, (3) die „introjizierte Regulation“, (4) die „identifizierte Regulation“, (5) die „integrierte Regulation“ und (6) die „intrinsische Regulation“ (Dresel & Lämmle, 2011, S. 90; siehe auch Ryan & Deci, 2000). In Abbildung 4 wird die Selbstbestimmungstheorie von Ryan und Deci (2000) in leicht modifizierter Weise dargestellt (Dresel & Lämmle, 2017, S. 90), die an dieser Stelle näher ausgeführt werden soll.

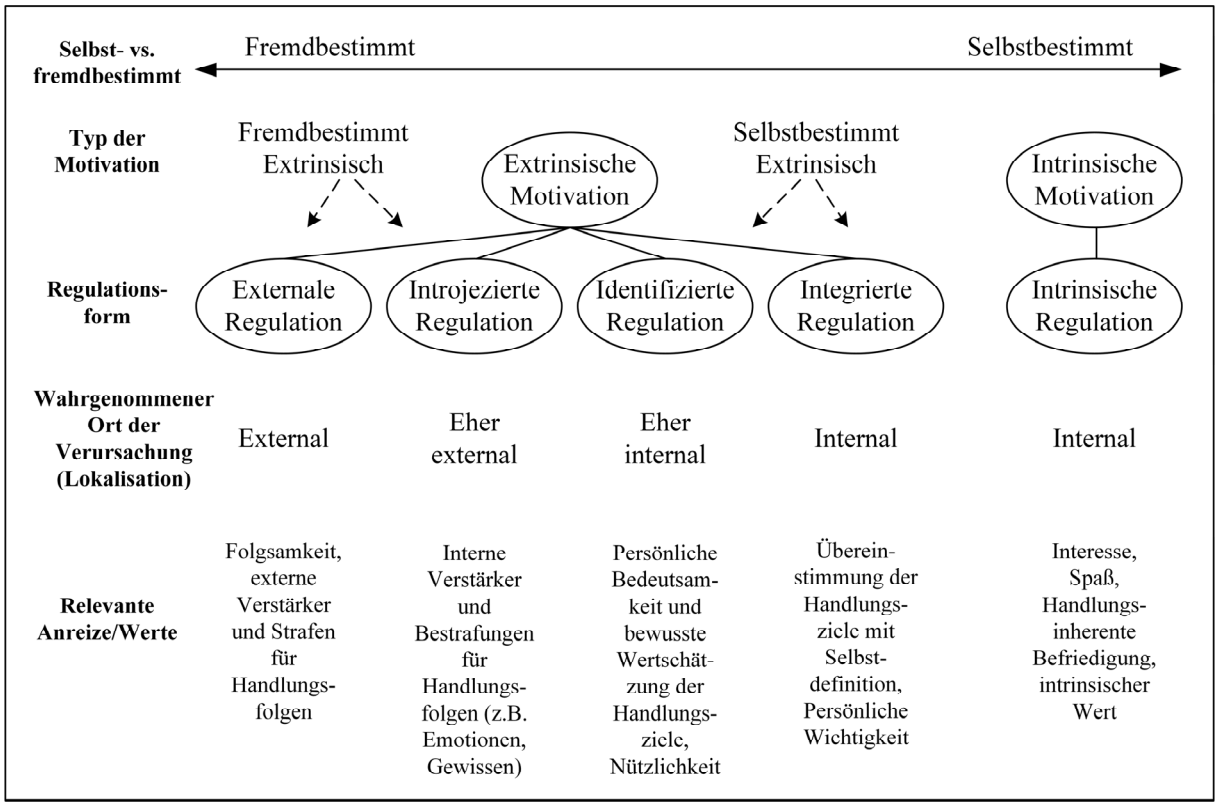


Abbildung 4. Selbstbestimmungstheorie gemäß Ryan und Deci (2000), modifiziert und ins Deutsche übersetzt von Dresel und Lämmle (2011, S. 90).

Anmerkung: Bei Dresel und Lämmle (2011) wird von introjezierter Motiviertheit gesprochen, während in anderen Publikationen es introjizierte Motiviertheit heißt (z.B. Frey et al., 2009); gemeint ist damit das gleiche Konstrukt.

Die Amotiviertheit (*amotivation*) ist zunächst von der extrinsischen und der intrinsischen Motivation abzugrenzen. Wie der Begriff bereits vermuten lässt, steht er für die fehlende Intention bezogen auf eine Handlung (Deci & Ryan, 1993; Ryan & Deci, 2000). Amotivation wird mit drei Ursachen in Verbindung gebracht: (1) Die Handlung ist der Person nicht wichtig (siehe auch persönlicher Wert nach Eccles, 1983), (2) die Person fühlt sich nicht ausreichend

kompetent, um die Handlung durchzuführen (siehe auch Seligman, 1975) und/oder (3) das Handeln wird als wenig lohnend eingeschätzt (siehe auch Wigfield & Eccles, 2000). Die externe Regulation (*external regulation*) zeichnet sich hingegen durch die Intention aus, negative Konsequenzen zu vermeiden und Regeln einzuhalten. Handlungen werden beispielweise aus Angst vor Bestrafung aufgenommen oder damit begründet, dass es die Regel ist (Deci & Ryan, 1993; Ryan & Cornell, 1989). Das Handeln einer Person kann dabei mit dem Gefühl der Kontrolle einhergehen (Ryan & Deci, 2000). Für die introjizierte Regulation (*introjected regulation*) ist kennzeichnend, dass die Person im Zuge des eigenen Gewissens, der Selbstdarstellung und Selbstbestätigung handelt (Deci & Ryan, 1993; Ryan & Cornell, 1989). Es geht um das eigene Ansehen, welches gestärkt oder aufrechterhalten werden soll (Ryan & Cornell, 1989). Bei der identifizierten Regulation (*identificated regulation*) steht dagegen die Bedeutsamkeit einer Handlung im Fokus (Deci & Ryan, 1993). Die Person identifiziert sich mit etwas und schreibt dem Handeln einen Wert zu (Ryan & Cornell, 1989). Dadurch wird ein erhöhtes Autonomie- und Selbstbestimmungserleben im Vergleich zu den vorherigen Motivationsformen möglich. Eine weitere und gleichzeitig letzte Form der extrinsischen Motiviertheit stellt die integrierte Regulation (*integrated regulation*) dar (Deci & Ryan, 1993; Ryan & Deci, 2000). Für diese ist charakteristisch, dass die Gründe des Handelns im Einklang mit der Person stehen (Ryan & Deci, 2000). Die Person kann ihr Handeln regulieren, weshalb diese Motivationsform auch den höchsten Grad an Autonomie- und Selbstbestimmungserleben innerhalb der extrinsischen Motiviertheit besitzt (ebd.). Da mit dem Handeln aber ein bestimmtes Ziel verfolgt wird, das nicht einzig im Erleben von Spaß oder Freude besteht, ist eine solche Motiviertheit als extrinsisch zu bewerten (ebd.). Von intrinsischer Regulation (*intrinsic regulation*) kann erst dann gesprochen werden, wenn eine Handlung ausgeübt wird, die dem reinen Selbstzweck dient. Das Erleben von Freude und Spaß ist dabei kennzeichnend für die intrinsische Motiviertheit (Ryan & Cornell, 1989; Ryan & Deci, 2000).

Soll in Studien die Motivation unter Berücksichtigung der Selbstbestimmungstheorie (Deci & Ryan, 1993; Ryan & Deci, 2000) erfasst werden, ist laut Taylor et al. (2014) die *Academic Motivation Scale* (u.a. Vallerand et al., 1992, 1993) die am häufigsten verwendete Skala. Mit ihrer Hilfe können Erkenntnisse gewonnen werden, die Aufschluss über die Qualität der Lernmotivation (Deci & Ryan; 1993; Ryan & Deci, 2000) als Prädiktor für schulische bzw. akademische Leistungen geben. Taylor et al. (2014) haben eine Meta-Analyse mit Studien unter Verwendung der *Academic Motivation Scale* durchgeführt. Die Probandinnen und Probanden

dieser einbezogenen Studien umfassten verschiedene Altersgruppen: Schülerinnen und Schüler der Grundschule, der Sekundarstufe (genauer: High School) und Studierende an Colleges und an Universitäten. Das zu Beginn vorgenommene Literaturreview der Autorinnen und Autoren hatte zum Ergebnis, dass es zahlreiche Studien gibt, in denen die Beziehung zwischen der Qualität der Lernmotivation und den schulischen bzw. akademischen Leistungen überprüft wird (z.B. Chatzisarantis, Hagger, Biddle, Smith & Wang, 2003; Deci, Vallerand, Pelletier & Ryan, 1991). Die Mehrheit der Untersuchungen wies dabei ein querschnittliches Design auf (Cokley, Bernard, Cunningham & Motoike, 2001; d'Ailly, 2003; Fortier, Vallerand & Guay, 1995; Grolnick, Ryan & Deci, 1991; Hardre & Reeve, 2003; Noels, Clement & Pelletier, 1999; Petersen, Louw & Dumont, 2009, Soenens & Vansteenkiste, 2005; Walls & Little, 2005). Nur wenige Studien hatten ein längsschnittliches Design und kontrollierten dabei die schulischen bzw. akademischen Leistungen im Ausgangswert (Baker, 2003; Black & Deci, 2000; Burton, Lydon, D'Alessandro & Koestner, 2006). Das Ergebnis der Meta-Analyse spricht dafür, dass die intrinsische Motiviertheit (z.T. auch die identifizierte Motiviertheit) einen positiven Effekt auf die schulischen bzw. akademischen Leistungen hat (auch bei Kontrolle der Vorleistungen), während die Amotiviertheit einen negativen Einfluss auf die schulischen bzw. akademischen Leistungen nimmt (Taylor et al., 2014). Bei der externalen und introjizierten Motiviertheit erwiesen sich die Ergebnisse als inkonsistent (ebd.).

Zusammenfassend kann konstatiert werden, dass es ein breites Spektrum an motivationalen Konstrukten gibt. Die verschiedenen Meta-Analysen liefern eine empirische Evidenz hinsichtlich der Vorhersagekraft motivationaler Merkmale von Schülerinnen und Schülern für schulische Leistungen (Hattie & Zierer, 2018; Kriegbaum et al., 2015; Taylor et al., 2014). Darüber hinaus hat die Meta-Analyse von Taylor et al. (2014) deutlich gemacht, dass die verschiedenen Formen der Qualität der Lernmotivationen in unterschiedlichem Maße Einfluss auf die schulischen Leistungen nehmen.

3.5 Unterrichtsmerkmale

Der Begriff der Unterrichtsmerkmale ist sehr breit gefasst und beinhaltet verschiedene Aspekte, von denen einige im Kapitel 2.4.2 bereits zur Sprache kamen (z.B. die Abdeckung von Unterrichtsinhalten, die Gewichtung von Unterrichtsinhalten und die Unterrichtsqualität). Während diese verschiedenen Aspekte in Kapitel 2.4.2 getrennt voneinander betrachtet wurden, kann gemäß Klieme (2019) die Unterrichtsqualität auch als Überbegriff fungieren, unter dem

„fachlich-inhaltliche Qualität[-saspekte]“ (S. 398), „spezifische Methoden und Lehr-Lern-Arrangements“ (S. 399) und „generische Grunddimensionen der Unterrichtsqualität“ (S. 400) subsumiert werden. Demzufolge kann „Unterrichtsqualität (…) als Gesamtheit der empirisch beobachtbaren Merkmale des Unterrichtsgeschehens [verstanden werden d. Verf.], die nachweislich mit einer Entwicklung der Lernenden im Sinne der Realisierung von Bildungs- und Erziehungszielen einhergehen“ (S. 396). In Anlehnung an Klieme (2019) und unter Berücksichtigung der Studien zur Evaluation der Instruktionssensitivität unter Einbeziehung von Unterrichtsmerkmalen sollen nachfolgend (1) das Konzept *Opportunity to Learn* (siehe u.a. Carroll, 1963) und (2) die „generische[n] Grunddimensionen der Unterrichtsqualität“ (Klieme, 2019, S. 400; siehe auch Klieme, Schümer & Knoll, 2001) beschrieben werden.

3.5.1 Opportunity to Learn

Der Begriff *Opportunity to Learn* (*Lerngelegenheiten*) ist bereits bei Carroll (1963) im Zusammenhang mit dem Modell zur Erklärung des schulischen Lernens zu finden. Gemäß diesem Modell ist schulisches Lernen von fünf Faktoren abhängig: (1) die Befähigung, also das Ausmaß an Zeit, die notwendig ist, um etwas zu lernen (unter der Voraussetzung optimaler Unterrichtsqualität), (2) die Fähigkeit von Schülerinnen und Schülern, dem Unterricht zu folgen, (3) die Ausdauer, also die Zeit, die Schülerinnen und Schüler bereit sind, für das Lernen zu investieren, (4) die Möglichkeit, also die Zeit, die Schülerinnen und Schülern zum Lernen zur Verfügung gestellt wird und (5) die Unterrichtsqualität[8] (Carroll 1963, S. 729). OTL definiert Carroll (1989) dabei als „the amount of time allowed for learning, for example by a school schedule or programm“ (S. 26).

Aber nicht nur der Zeitfaktor spielt bei den OLT-Konzepten eine entscheidende Rolle. Brewer und Stasz (1996) ordnen unterrichtsbezogene Variablen aus verschiedenen Studien drei Aspekten zu: (1) Einen ersten Aspekt stellen die curricularen Inhalte (*curriculum content*) dar. Gemeint sind damit die Abdeckung curricularer Inhalte im Unterricht, die Frage nach der aufgewendeten Zeit für die Vermittlung eines Inhalts und das Gewicht, das einem bestimmten Inhalt zugesprochen wird (ebd.). (2) Ein zweiter Aspekt liegt in den Strategien zur Vermittlung curricularer Ziele im Unterricht (*strategies of instruction*), d.h. wie Lehrpersonen die Inhalte

[8] Während bei Klieme (2019) das OTL-Konzept aus der Perspektive eines fachlich-inhaltlichen Qualitätsaspekts von Unterricht betrachtet wird, ist bei Carroll (1963) eine Trennung zwischen dem Inhaltlichen und dem Qualitativen zu erkennen, wie es auch bei zahlreichen späteren Studien zur Instruktionssensitivität unter Berücksichtigung von Unterrichtsmerkmalen vorzufinden ist.

des Curriculums vermitteln und wie Schülerinnen und Schüler im Unterricht einbezogen werden (ebd.). (3) Der dritte Aspekt umfasst schließlich die Ressourcen des Unterrichts (*instructional resources*), also die Lehrmittel, die PC-Ausstattung und z.T. auch das Humankapital bezogen auf die Lehrpersonen und deren Professionswissen (ebd.).

Ein etwas anderes OTL-Konzept, in das sich die genannten Aspekte von Brewer und Stasz (1996) durchaus einordnen lassen, ist bei Kurz und Elliot (2013) zu finden. Ihr Entwurf (Abbildung 5) erscheint auch in Anbetracht der genannten Unterrichtsmerkmale im Kontext der Instruktionssensitivität (Kapitel 2.4.2) sehr passend. Kurz und Elliott (2011, 2013) gehen von den folgenden drei Aspekten aus: (1) die Qualität des Unterrichts, (2) die Lerninhalte des Unterrichts und (3) der zeitliche Umfang, der für die Auseinandersetzung mit dem Lerngegenstand im Unterricht aufgewendet wird (siehe auch Elliott, 2015; Kurz, Elliott, Kettler & Yel, 2014). Die Autoren definieren entsprechend „OTL as the degree to which a teacher dedicates instructional time and content coverage to the intended curriculum objectives emphasizing high-order cognitive processes, evidence-based instructional practices, and alternative grouping formats“ (ebd., S. 1).

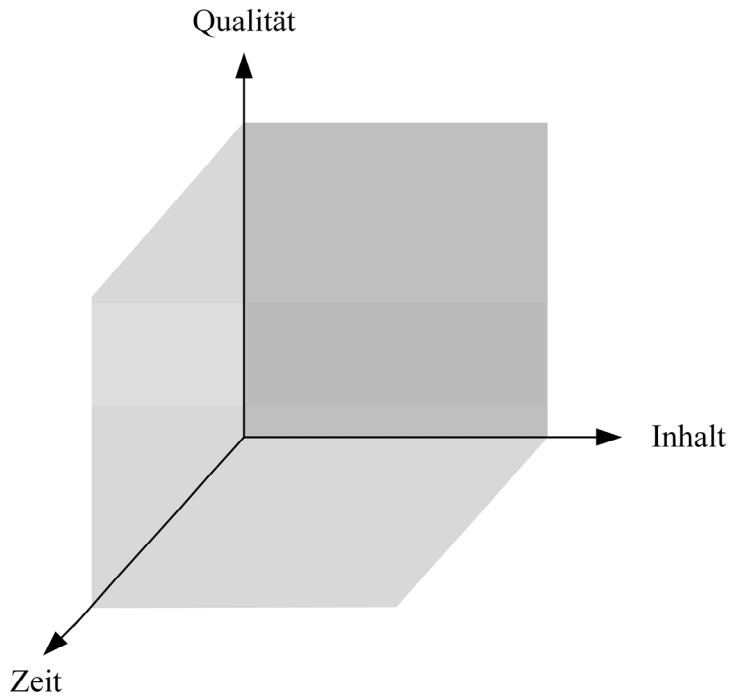


Abbildung 5. OTL-Konzept nach Kurz und Elliott (2013, S. 1), eigene Übersetzung und leicht modifizierte Darstellung.

Die Ausführungen lassen erkennen, wie vielfältig sich die OTL-Konzepte gestalten. Zur Beschreibung von Lerngelegenheiten wird von Aspekten oder auch von Dimensionen gesprochen, die z.T. klar voneinander abgrenzbar sind (ebd.). Andere Autoren betonen hingegen die Überscheidung der verschiedenen OTL-Aspekte bzw. OTL-Dimensionen (Brewer & Stasz, 1996). Bezugnehmend auf die Evaluation der Instruktionssensitivität unter Einbeziehung von Unterrichtsmerkmalen lassen sich viele Parallelen zum OTL-Konzept von Kurz und Elliott (2013) finden. In Anbetracht der Komplexität von Unterricht wird aber auch deutlich, dass im Kontext der Evaluation der Instruktionssensitivität immer nur Teilaspekte von Unterricht beleuchtet werden, was bei der späteren Nutzung und Interpretation solcher Daten zu berücksichtigen ist.

3.5.2 Generische Grunddimensionen der Unterrichtsqualität

In der empirischen Bildungsforschung wurde u.a. ein *Framework* zur Unterrichtsqualität entwickelt, welches auf drei generischen Grunddimensionen der Unterrichtsqualität basiert: (1) die Klassenführung, (2) die konstruktive Unterstützung bzw. das unterstützende Unterrichtsklima und (3) die kognitive Aktivierung (Klieme et al., 2001; Praetorius, Klieme, Herbert & Pinger, 2018). Die Herausarbeitung dieser drei Dimensionen geht zunächst auf deskriptive Arbeiten von Baumert et al. (1997) sowie von Stigler und Hiebert (1999) zurück. Klieme et al. (2001) gelang schließlich der Nachweis einer dreidimensionalen Struktur der Unterrichtsqualität mittels Faktoranalysen, aus dem das heute vorzufindende *Framework* zur Unterrichtsqualität resultiert. Klieme et al. (2006) merken jedoch an, dass bislang unklar ist, inwiefern die jeweiligen Qualitätsdimensionen auf die schulischen Leistungen der Schülerinnen und Schüler wirken. Die Autorinnen und Autoren konzipierten vor einiger Zeit ein Modell, dass die vermuteten Wirkbeziehungen aufzeigt (Abbildung 6, Klieme et al., 2006). Dabei wird darauf hingewiesen, dass die bisherige Vorgehensweise in der Unterrichtsforschung vornehmlich einer „induktiv[en]“ und nicht einer „theoriegeleitete[n]“ Vorgehensweise entspricht (ebd., S. 127). Die dreidimensionale Struktur konnte in zahlreichen Studien bestätigt werden (Fauth, Decristan, Rieser, Klieme & Büttner, 2014; Kunter & Voss, 2011; Tschannen-Moran & Hoy 2001). Zur Fundierung der Modellannahmen werden Theorien zur Motivation (Ryan & Deci, 2000) und zur kognitiven Aktivität (Mayer, 2004) herangezogen. Die Unterrichtsqualitätsdimensionen werden dabei als generisch bezeichnet, da diese fächerübergreifend von Relevanz sein sollen (Klieme, 2019). Bemerkenswert ist an dieser

Stelle, dass in den USA eine analoge Struktur herausgearbeitet wurde, die sich in „*classroom organization*“, „*emotional support*“ und „*instructional support*“ einteilen lässt (Pianta, La Paro & Hamre, 2008, S. 2; Pianta & Hamre, 2009, S. 111). Die Autorinnen und Autoren entwickelten in diesem Zusammenhang das Beobachtungsinstrument *Classroom Assessment Scoring System* (CLASS) zur Einschätzung der Unterrichtsqualität (Pianta et al., 2008). Auf inhaltlicher Ebene kommt ihre Arbeit dem *Framework* von Klieme et al. (2001) sehr nahe. Wie sich die generischen Basisdimensionen konkret definieren lassen, soll nachfolgend aufgezeigt werden.

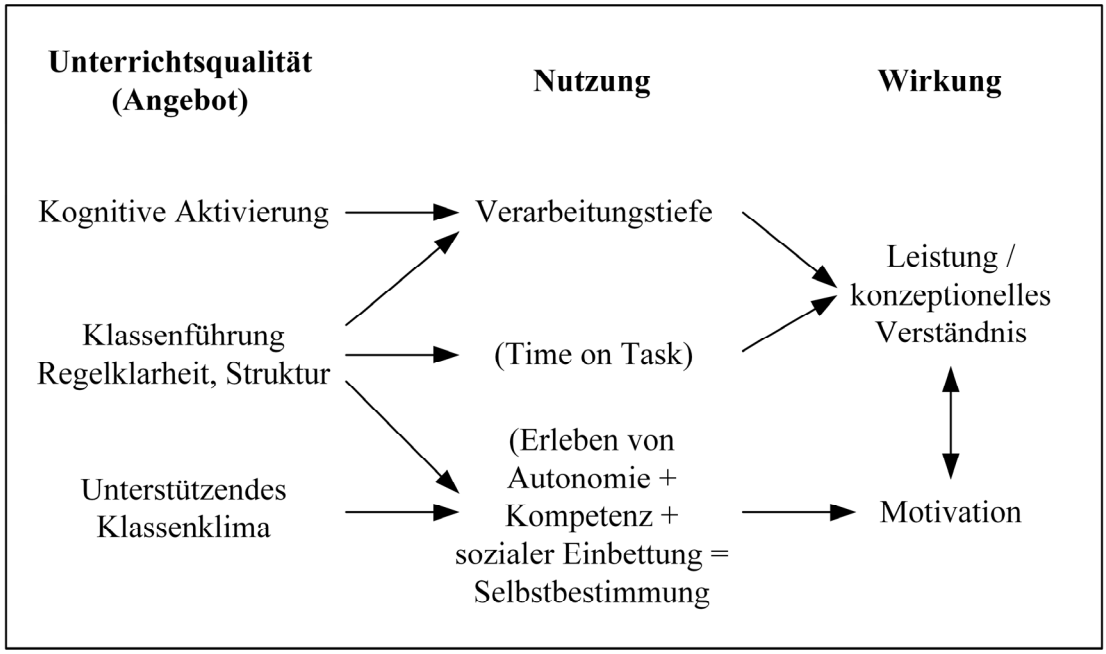


Abbildung 6. Basisdimensionen guten Unterrichts und deren vermutete Wirkung (Klieme et al., 2006, S. 131).

Beginnend mit der effizienten Klassenführung werden hierunter verschiedene Konstrukte subsumiert. Erste Arbeiten zur Klassenführung gehen auf Forschungsarbeiten von Kounin (1976, 2006) zurück, der die „Allgegenwärtigkeit“, die „Überlappung“, die „Reibungslosigkeit“ (bzw. „Sprunghaftigkeit“), den „Schwung“ (bzw. die „Verzögerung“), den „Gruppen-Fokus“ und die „programmierte Überdrussvermeidung“ als zentral Subdimensionen der Klassenführung herausgearbeitet hat (Kounin, 2006, S. 10). Die von Kounin (2006) formulierten Dimensionen zur Klassenführung sind heute noch aktuell, wenngleich diese z.T. etwas modifiziert wurden. So legten Praetorius et al. (2018) insgesamt vier Subdimensionen der effizienten Klassenführung dar. Zu nennen sind das Konstrukt der „Unterrichtsstörungen“ bzw. der „Disziplinprobleme“, der „effektiven Zeitnutzung“ bzw. „*time on task*“, das „Monitoring“ bzw. die „Allgegenwärtigkeit“ der Lehrperson sowie die „Regelklarheit“ bzw.

eingeübten „Routinen“ (Praetorius et al., 2018, S. 412, eigene Übersetzung). Ziel einer effizienten Klassenführung ist es, erwünschtes Verhalten zu fördern und unerwünschtes Verhalten zu vermeiden (Hochweber, Hosenfeld & Klieme, 2014). Für einen reibungslosen Ablauf ist es dienlich, klare Regeln und Erwartungen gegenüber den Schülerinnen und Schülern zu formulieren sowie Routinen einzuüben. Eine effiziente Klassenführung soll dabei Unterrichtsstörungen präventiv entgegenwirken, indem die Lehrperson die gesamte Klasse im Blick hat (Praetorius et al., 2018). Lehrpersonen haben dadurch die Möglichkeit, ins Geschehen einzugreifen, bevor größere Unruhe o.Ä. aufkommt. Darüber hinaus steht die effiziente Klassenführung auch für die optimale Nutzung der Zeit (*time on task*), um sich mit dem Lerngegenstand auseinanderzusetzen (ebd.). Fließende Übergänge verhindern Wartezeiten und minimieren das Ablenkungsrisiko. Ein gut geplanter und organisierter Unterricht ist dabei für eine effiziente Klassenführung unerlässlich.

Das unterstützende Klassenklima[9] ist als zweite Unterrichtsqualitätsdimension zu nennen, die insbesondere auf die Interaktion zwischen Schülerinnen und Schülern sowie die Interaktion zwischen Schülerinnen und Schülern und der Lehrperson abzielt (ebd.; siehe auch Fauth et al., 2014). Ein unterstützendes Klassenklima zeichnet sich durch ein Miteinander aus, das vom gegenseitigen Respekt, gegenseitiger Wertschätzung, gegenseitigem Interesse und gegenseitiger Unterstützung geprägt ist (Praetorius et al., 2018). Merkmale des unterstützenden Klassenklimas sind die Ermöglichung von Kompetenzerleben, die Bereitstellung von Differenzierungs- sowie adaptiven Unterstützungsangeboten, ein angemessenes Tempo der Instruktion, ein konstruktiver Umgang mit Fehlern und Misskonzepten, sachlich-konstruktives Feedback, Wertschätzung, Förderung des Gefühls von Autonomie, die Möglichkeit eines selbstbestimmten Handelns beispielsweise durch Wahlmöglichkeiten und die Unterstützung des Gefühls sozialer Eingebundenheit bzw. Verbundenheit (Praetorius et al., 2018, S. 412f.; siehe auch Assor, Kaplan & Roth, 2002; Fauth et al., 2014; Reyes, Brackett, Rivers, White & Salovey, 2012). Das unterstützende Klassenklima wird dabei insbesondere mit emotional-motivationalen Aspekten in Verbindungen gebracht. Argumentiert wird an dieser Stelle mit der von Ryan und Deci (2000) entwickelten Selbstbestimmungstheorie. Ein unterstützendes Klassenklima soll demnach das Gefühl von Kompetenz, Autonomie und einem sozialen Miteinander stärken und sich förderlich auf das Lernen auswirken.

[9] In neueren Beiträgen wird auch von der konstruktiven Unterstützung gesprochen, welche zum einen eine emotionale und zum anderen eine kognitive Komponente umfasst (Fauth, 2021).

Die dritte und letzte Dimension der Unterrichtsqualität umfasst die kognitive Aktivierung. Diese kommt darin zum Ausdruck, anspruchsvolle Aufgaben, Problemstellungen und Fragenstellungen zu formulieren, das Vorwissen zu erfassen und zu aktivieren, Schülervorstellungen zu eruieren und einzubeziehen sowie verschiedene Lösungswege zuzulassen (Praetorius et al., 2018). Schülerinnen und Schüler werden aufgefordert, sich aktiv mit dem Unterrichtsgegenstand auseinanderzusetzen, zu argumentieren und die eigenen Antworten zu begründen. Kognitive Aktivierung steht entsprechend für die Anregung von Denkprozessen sowie für die Förderung eines konzeptionellen Verständnisses, wird mit einem genetisch-sokratischen Unterricht in Verbindung gebracht und basiert auf einem konstruktivistischen Lehr-Lern-Verständnis (Klieme, 2019; Praetorius et al., 2018; siehe auch Heckmann, 1980; Klafki, 2019; Labudde, 1996). Die Förderung von Denkprozessen höherer Ordnung (siehe Bloomsche Taxonomie, Bloom, 1976; Taxonomie kognitiver Lernziele, Anderson et al., 2001) soll sich dabei positiv auf das Lernen im Sinne eines vertieften konzeptionellen Verständnisses auswirken (Hamre & Pianta, 2005; Klieme, Pauli & Reusser, 2009; Praetorius et al., 2018). Ferner sollen nicht nur die kognitiven, sondern auch die metakognitiven Prozesse angeregt werden (Praetorius et al., 2018).

Inwiefern die generischen Unterrichtsqualitätsdimensionen einen Einfluss auf die schulischen Leistungen der Schülerinnen und Schüler nehmen, wird im Artikel von Praetorius et al. (2018) ersichtlich. In zahlreichen Studien gelingt es, die angenommene Beziehung zwischen den spezifischen Dimensionen der Unterrichtsqualität und den Leistungen der Schülerinnen und Schüler nachzuweisen (siehe auch Decristan, Klieme et al., 2015). Eine einheitliche Befundlage gibt es jedoch nicht, da es auch Untersuchungen gibt, in denen der Nachweis einer solchen Beziehung nicht gelingt (Praetorius et al., 2018). Bezugnehmend auf die Evaluation der Instruktionssensitivität unter Hinzuziehung von Unterrichtsmerkmalen kann aber konstatiert werden, dass die generischen Grunddimensionen der Unterrichtsqualität Berücksichtigung finden (Kapitel 2.4.2).

3.6. Zusammenfassung und Fazit

Um die Instruktionssensitivität von Testitems unter Berücksichtigung von individuellen Lernvoraussetzungen der Schülerinnen und Schüler konzeptionell einzubetten, wurde zunächst in Kapitel 3.1 die Evaluation der Instruktionssensitivität mit der Schul- und Unterrichtseffektivitätsforschung in Beziehung gesetzt. Deutlich wurde an dieser Stelle, dass

sich das Konzept der Instruktionssensitivität und die Schul- und Unterrichtsforschung komplementär gegenüberstehen. In Anbetracht dieser Beziehung und mit der Intention, die Instruktionssensitivität unter Berücksichtigung individueller Merkmale der Schülerinnen und Schüler zu evaluieren, liegt der Rückgriff auf Ansätze, Modelle und empirische Erkenntnisse der Schul- und Unterrichtseffektivitätsforschung nahe.

Dementsprechend wurden in Kapitel 3.2 der *Value-Added-Ansatz* und in Kapitel 3.3 die Angebots-Nutzungs-Modelle vorgestellt. Mit dem *Value-Added-Ansatz* kann – übertragen auf den Kontext der Evaluation der Instruktionssensitivität – eine Unterscheidung zwischen der Instruktionssensitivität im engeren und im weiteren Sinne vorgenommen werden. Diese Unterscheidung stellt – in Anbetracht der Intention der Einbeziehung individueller Lernvoraussetzungen – einen wichtigen Gedanken dar, der in Kapitel 3.7 zur Ableitung der Fragestellung und den Hypothesen nochmals aufgegriffen und weiter ausgeführt wird. Anhand der Angebots-Nutzungs-Modelle in Kapitel 3.3 wurde ersichtlich, dass nicht nur der von der Lehrperson initiierte Unterricht, sondern auch die individuellen Lernvoraussetzungen der Schülerinnen und Schüler einen Einfluss auf die schulischen Leistungen nehmen. Es wird angenommen, dass sich Lernen in einer aktiven Auseinandersetzung mit der Umwelt durch die Schülerinnen und Schüler vollzieht. Zwei individuelle Lernvoraussetzungen, die das Lernen maßgeblich beeinflussen, wurden in Kapitel 3.4 exemplarisch vorgestellt: das fachspezifische Vorwissen und die Lernmotivation. Rückblickend auf die Ausführungen in Kapitel 2.5.2 ist ersichtlich, dass in den Analysen zur Instruktionssensitivität das Vorwissen häufig indirekt kontrolliert wird, die Lernmotivation hingegen aber weitgehend unberücksichtigt bleibt.

Fortfahrend mit den Unterrichtsmerkmalen wurden in Kapitel 3.5 das OTL-Konzept und die generischen Grunddimensionen der Unterrichtsqualität vorgestellt. Sowohl das OTL-Konzept als auch die Unterrichtsqualität finden in Studien zur Instruktionssensitivität häufig Berücksichtigung (Kapitel 2.4.2). Die Einbeziehung von Unterrichtsmerkmalen stellt dabei eine Validierungsstrategie dar (van der Linden, 1981). Die Ausführungen in Kapitel 2.4.2 haben aber auch deutlich gemacht, dass der Nachweis einer Beziehung zwischen den Unterrichtsmerkmalen und den Resultaten der Schülerinnen und Schüler bezogen auf Schulleistungstests nicht immer gelingt. Eine Ursache kann in der Komplexität von Unterricht erachtet werden, der man bei der Erfassung von Unterrichtsmerkmalen nur schwer gerecht wird. Eine andere Ursache liegt möglicherweise auch in den unterschiedlichen methodischen Herangehensweisen zur Messung von unterrichtsbezogenen Konstrukten begründet. Insgesamt

kann festgehalten werden, dass die Einbeziehung von Unterrichtsmerkmalen bei der Evaluation der Instruktionssensitivität sinnvoll erscheint, deren Umsetzung jedoch mit Herausforderungen verbunden ist.

3.7 Ableitung der Fragestellung und der Hypothesen

Die übergeordnete Fragestellung der vorliegenden Arbeit lautet: Inwieweit beeinflusst die Qualität der Lernmotivation[10] der Schülerinnen und Schüler die Schätzung der Indikatoren zur Instruktionssensitivität von Testitems? Es soll untersucht werden, inwiefern sich Verzerrungseffekte mit Anwendung verschiedener Varianten des von Naumann et al. (2017) vorgeschlagenen LMLIRT-Modells zur Evaluation der Instruktionssensitivität mit und ohne Einbeziehung der Qualität der Lernmotivation als Beispiel einer individuellen Lernvoraussetzung zeigen. Verzerrungseffekte können dabei sowohl in einer Unter- als auch in einer Überschätzung der Instruktionssensitivität von Testitems zum Ausdruck kommen. Günstige Ausprägungen der Qualität der Lernmotivation in der Gesamtstichprobe könnten mit einer Überschätzung, ungünstige Ausprägungen in der Gesamtstichprobe mit einer Unterschätzung der (Effekte auf die) Veränderung in den klassenspezifischen Itemschwierigkeiten einhergehen – vorausgesetzt es wird von einem Instruktionssensitivitäts-verständnisses im engeren Sinne ausgegangen und die Qualität der Lernmotivation wird nicht hinreichend kontrolliert. Darüber hinaus kann der Fall einer korrekten Schätzung eintreten. Die Möglichkeit einer korrekten Schätzung bzw. einer Über- oder Unterschätzung impliziert dabei die Existenz eines wahren Werts, vom dem abgewichen werden kann.

Eine empirische Überprüfung bezogen auf die übergeordnete Fragestellung, inwieweit die Qualität der Lernmotivation der Schülerinnen und Schüler die Schätzung der Indikatoren zur Instruktionssensitivität von Testitems beeinflusst, ist dabei unabdingbar. Zwar gibt es einige Studien zur Instruktionssensitivität, in denen lernrelevante Merkmale der Schülerinnen und Schüler einbezogen wurden (Kapitel 2.5.2). Der Frage nach möglichen Verzerrungseffekten, wenn lernrelevante Merkmale bei der Evaluation der Instruktionssensitivität unberücksichtigt bleiben, sind aber bisher nur Muthén et al. (1991) nachgegangen. Hinzu kommt, dass sich die Annahme möglicher Verzerrungseffekte primär auf der Übertragung von Erkenntnissen aus der Schul- und Unterrichtsforschung unter Berücksichtigung des *Value-Added-Ansatzes* und der

[10] Mit der *Qualität der Lernmotivation* wird an dieser Stelle auf die zugrunde liegende Theorie von Ryan und Deci (2000) verwiesen, die in Kapitel 3.4.2 vorgestellt wurde.

Angebots-Nutzungs-Modelle auf den Kontext der Instruktionssensitivität stützt. Ob sich ähnliche Ergebnisse wie bei Muthén et al. (1991) auch mit dem neueren LMLIRT- bzw. eLMLIRT-Modell *(erklärendes längsschnittliches Mehrebenen-Item-Response-Theory-Modell)* zur Evaluation der Instruktionssensitivität (Naumann et al., 2017; Naumann, Rieser et al., 2019) mit und ohne Hinzuziehung der Qualität der Lernmotivation als Kovariate zeigen, ist eine offene Frage. Die Studie von Muthén et al. (1991) war für die damalige Zeit inhaltlich und methodisch sehr fortschrittlich. Im Vergleich zu den heutigen Analyseverfahren weisen neuere Modelle wie das LMLIRT-Modell jedoch wertvolle Weiterentwicklungen auf. Mit dem LMLIRT bzw. dem eLMLIRT-Modell wird beispielsweise der geclusterten Datenstruktur Rechnung getragen, da dem Modell ein Mehrebenen-Ansatz zugrunde liegt. Darüber hinaus sind neue Indikatoren wie die *globale* und *differenzielle Sensitivität* zur Evaluation der Instruktionssensitivität verfügbar, welche um einen weiteren Indikator – den der *spezifischen Sensitivität* – im Rahmen dieser Arbeit ergänzt werden. Mithilfe der empirischen Befunde dieser Arbeit können Hinweise zur Validität der Testwertnutzung und -interpretation im Kontext der Evaluation der Instruktionssensitivität von Testitems gewonnen werden. Hinter diesem Erkenntnisinteresse steht auch die Frage, ob Modelle zur Evaluation der Instruktionssensitivität von Testitems zu stark vereinfacht sind, wenn von einem Instruktionssensitivitätsverständnis im engeren Sinne ausgegangen wird und lernrelevante Merkmale der Schülerinnen und Schüler unberücksichtigt bleiben. In Anbetracht dieser genannten Faktoren hat die vorliegende Arbeit einen entsprechenden Neuheitswert.

Bevor die einzelnen Hypothesen aufgeführt werden, sollen zwei Aspekte vorab erläutert werden, die für alle Hypothesen gleichermaßen relevant sind und die Argumentationslinie stützen. In einem ersten Schritt soll dargelegt werden, aus welcher Perspektive die Instruktionssensitivität von Testitems in der vorliegenden Arbeit betrachtet wird. Oder anders ausgedrückt: Welches Verständnis von Instruktionssensitivität liegt dieser Arbeit zugrunde? Hierfür wird auf den *Value-Added-Ansatz* und die Angebots-Nutzungs-Modelle Bezug genommen. In einem zweiten Schritt geht es um die Begründung, weshalb die Qualität der Lernmotivation als Beispiel einer individuellen Lernvoraussetzung im Rahmen dieser Arbeit fokussiert wird.

Perspektive auf die Instruktionssensitivität von Testitems

Je nach intendierter Testwertnutzung und -interpretation ist die Evaluation der Instruktionssensitivität unterschiedlich auszurichten, was sich in der Formulierung der Hypothesen widerspiegelt. In Kapitel 3.2 wurde die Unterscheidung zwischen der Instruktionssensitivität im *engeren* und im *weiteren Sinne* eingeführt. Sollen Testitems auf den Unterricht im Allgemeinen sensitiv reagieren, ist die Instruktionssensitivität im *weiteren Sinne* von Interesse. (Unterrichts-)Effekte resultieren in diesem Zusammenhang aus einem Zusammenspiel zwischen dem von der Lehrperson initiierten Unterricht und der Nutzung des Unterrichtsangebots durch die Schülerinnen und Schüler. Mediations- sowie Moderationsannahmen könnten in diesem Kontext aufgestellt werden. Zu überlegen wäre, ob das Zusammenspiel von Lehrperson, Schülerinnen- und Schülern in einen multiplikativen, additiven, kumulativen, kompensatorischen oder auch stabilen Effekt resultiert (Pham, 2018). Sollen Testitems auf den von der Lehrperson zumeist kontrollierbaren (Unterrichts-)Effekt sensitiv reagieren, ist hingegen die Instruktionssensitivität im *engeren Sinne* von Interesse. (Unterrichts-)Effekte basieren in diesem Zusammenhang in erster Linie auf dem von der Lehrperson initiierten Unterricht, also dem Unterrichtsangebot. Einflussfaktoren, wie z.B. lernrelevante Merkmale der Schülerinnen und Schüler, sollen demnach nicht ins Gewicht fallen. Die Einbeziehung von Kovariaten ist eine Möglichkeit, um Störeinflüsse herauszupartialisieren.

Argumente für das Interesse an einem Instruktionssensitivitätsverständnis im engeren und im weiteren Sinne gibt es gleichermaßen. Analysen zur Evaluation der Instruktionssensitivität, denen ein enges Verständnis der Instruktionssensitivität zugrunde liegt, können beispielsweise mit der Intention einer testdatenbasierten Überprüfung der Wirksamkeit spezifischer Unterrichtsmethoden oder einer testdatenbasierten Evaluation der Unterrichtsqualität einhergehen. Vor diesem Hintergrund wird der Arbeit ein enges Instruktionssensitivitäts-verständnis zugrunde gelegt.

Begründung für die Einbeziehung der Qualität der Lernmotivation als Beispiel einer individuellen Lernvoraussetzung bei der Evaluation der Instruktionssensitivität

Die Ausführungen zur Evaluation der Instruktionssensitivität unter Berücksichtigung individueller Lernvoraussetzungen der Schülerinnen und Schüler haben u.a. die Komplexität der Wirkbeziehungen zur Erklärung von Schulleistungen verdeutlicht. Vor diesem Hintergrund

ist davon auszugehen, dass nicht nur der durch die Lehrperson initiierte Unterricht, sondern auch andere Aspekte wie etwa individuelle Lernvoraussetzungen (z.B. Lernmotivation) der Schülerinnen und Schüler die schulischen Leistungen und damit auch die klassenspezifischen Veränderungen in den Itemschwierigkeiten bedingen.

In dieser Arbeit wird die Qualität der Lernmotivation als Beispiel für die Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraussetzungen herangezogen. Die Lernmotivation gilt neben dem Vorwissen als eine der bedeutsamsten individuellen Lernvoraussetzungen für das schulische Lernen und findet auch in Angebots-Nutzungs-Modellen Berücksichtigung (Brühwiler, 2014; Brühwiler et al., 2017; Seidel, 2014). Autoren wie Dresel et al. (2011), Dresel und Lämmle (2011) sowie Schrader und Helmke (2002) gehen davon aus, dass die Lernmotivation insbesondere für die Initiierung von Lernhandlungen, für die Ausdauer und Anstrengungsbereitschaft während einer Handlung, für die Persistenz bei auftretenden Schwierigkeiten und das Hervorrufen handlungsrelevanter Emotionen von hoher Bedeutung ist. Die Motivation gilt entsprechend als Regulator des eigenen Handelns (Ryan & Deci, 2000) und ist entscheidend für erfolgreiches Lernen (Klieme et al., 2006). Ryan and Deci (2000) haben bezogen auf die Motivation unterschiedliche Motive des Handelns herausgearbeitet, die von Seidel, Prenzel und Kobarg (2005) als Qualität der Lernmotivation beschrieben wurden. Inwiefern die Qualität der Lernmotivation als eine individuelle Lernvoraussetzung von Schülerinnen und Schülern auf die Schätzung zur Instruktionssensitivität von Testitems Einfluss nimmt, soll in der vorliegenden Arbeit überprüft werden.

Hypothesen

Nachfolgend werden die Hypothesen aufgezeigt und erläutert. Die Abfolge der Hypothesen erfolgt analog zum Aufbau des Kapitels zur Analysestrategie (Kapitel 4.5.2).

Hypothese 1: Die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der differenziellen Sensitivität im Kontext der Evaluation der Instruktionssensitivität von Testitems.

Die Überprüfung der Hypothese 1 erfolgt mithilfe der differenziellen Sensitivität als ein Indikator zur Evaluation der Instruktionssensitivität von Testitems, welcher die Zeitpunkte-x-Gruppen-Perspektive berücksichtigt. Aktuell ist noch unklar, inwiefern die Schätzung der

differenziellen Sensitivität durch die Qualität der Lernmotivation beeinflusst wird. Es kann aber davon ausgegangen werden, dass die Qualität der Lernmotivation mit dem schulischen Lernen in Beziehung steht und somit auch die Schätzung des Indikators beeinflusst. Inwiefern dies der Fall ist, soll mithilfe der Hypothese 1 überprüft werden. Angenommen wird, dass sich die Schätzung der differenziellen Sensitivität mit und ohne Kontrolle der Qualität der Lernmotivation voneinander unterscheidet.

Hypothese 2: Die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der spezifischen Sensitivität im Kontext der Evaluation der Instruktionssensitivität von Testitems.

Die Hypothese 2 stützt sich auf den Gedanken von van der Linden (1981), der sich bereits Anfang der 1980er Jahre für die Einbeziehung der Unterrichtsmerkmale bei der Evaluation der Instruktionssensitivität ausgesprochen hat (siehe auch Naumann et al., 2016). So kann sichergestellt werden, dass die statistische Variabilität in den Itemantworten auch auf den Unterricht zurückgeführt werden kann (Naumann et al., 2016; van der Linden, 1981). Für die Überprüfung der Hypothese 2 wird also nicht nur die Qualität der Lernmotivation der Schülerinnen und Schüler berücksichtigt, sondern es werden auch Unterrichtsmerkmale mit einbezogen. So kann überprüft werden, inwiefern sich die Schätzung der spezifischen Sensitivität mit und ohne Kontrolle der Qualität der Lernmotivation voneinander unterscheidet.

Die spezifische Sensitivität stellt dabei einen Indikator dar, der im Rahmen dieser Arbeit neu eingeführt wird. Er bezieht sich auf die Schätzung der Effekte von Unterrichtsmerkmalen (z.B. der kognitiven Aktivierung) auf die Veränderung der klassenspezifischen Itemschwierigkeiten. Die spezifische Sensitivität kommt also nur zum Tragen, wenn Unterrichtsmerkmale in die Evaluation der Instruktionssensitivität von Testitems einbezogen werden.

Welche Unterrichtsmerkmale zu inkludieren sind, ist vor dem Hintergrund des Erkenntnisinteresses zu entscheiden. In der vorliegenden Arbeit wird auf Unterrichtsmerkmale zurückgegriffen, die in der Literatur zur Evaluation der Instruktionssensitivität häufig Anwendung finden: (1) Maße zur Unterrichtsqualität, d.h. in diesem Fall einen Aspekt der Klassenführung und die kognitive Aktivierung[11] und (2) Maße zu den Unterrichtsinhalten, d.h.

[11] Zum unterstützenden Klassenklima wird keine Hypothese formuliert, da davon ausgegangen wird, dass dieses Merkmal, mediiert über die Motivation, einen Effekt auf die Indikatoren der Instruktionssensitivität hat (u.a. Klieme et al., 2006; Yıldırım, 2012). Basierend auf der Annahme der Mediatorfunktion erscheint es wenig

die im Unterricht verfolgen inhaltsbezogenen Lernziele und deren Gewichtung. Die beiden letztgenannten Aspekte (Inhalt & Gewichtung) werden als Interaktionsterm (Inhalt x Gewichtung) in die Analysen einbezogen. Aus Gründen der besseren Lesbarkeit wird folglich von den „im Unterricht verfolgten Lernzielen" gesprochen. Die Gewichtung wird hierbei stets mitgedacht. Nachfolgend wird die Hypothese 2 im Zuge der genannten Unterrichtsmerkmale in drei Unterhypothesen ausdifferenziert.

Hypothese 2.1: Die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der spezifischen Sensitivität bezogen auf Unterrichtsstörungen im Kontext der Evaluation der Instruktionssensitivität von Testitems.

Hypothese 2.2: Die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der spezifischen Sensitivität bezogen auf die kognitive Aktivierung im Kontext der Evaluation der Instruktionssensitivität von Testitems.

Hypothese 2.3: Die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der spezifischen Sensitivität bezogen auf die im Unterricht verfolgten Lernziele im Kontext der Evaluation der Instruktionssensitivität von Testitems.

Die Zusammenhänge zwischen den spezifischen Unterrichtsmerkmalen und den schulischen Leistungen wurden bereits dargelegt. Bezugnehmend auf die Hypothese 2.1 wird angenommen, dass sich ein möglichst störungsfreier Unterricht positiv auf die Leistungen der Schülerinnen und Schüler auswirkt (siehe Ausführungen in Kapitel 3.5.2). Daran anknüpfend wird davon ausgegangen, dass ein möglichst störungsfreier Unterricht einen Effekt auf die Veränderung der klassenspezifischen Itemschwierigkeiten hat, welcher den Indikator der spezifischen Sensitivität darstellt. Inwiefern die Qualität der Lernmotivation von Schülerinnen und Schülern einen Einfluss auf die Schätzung der spezifischen Sensitivität bezogen auf Unterrichtsstörungen nimmt, soll mit der Hypothese 2.1 überprüft werden. Angenommen wird, dass sich die Schätzung der spezifischen Sensitivität mit und ohne Kontrolle der Qualität der Lernmotivation voneinander unterscheidet.

Fortfahrend mit Hypothese 2.2 wird vermutet, dass ein kognitiv aktivierender Unterricht einen positiven Einfluss auf die schulischen Leistungen der Schülerinnen und Schüler nimmt (siehe Ausführungen in Kapitel 3.5.2). Insofern wird als wahrscheinlich angesehen, dass ein kognitiv

sinnvoll, die Qualität der Lernmotivation als Kovariate in den Analysen zum unterstützenden Klassenklima einzubeziehen.

aktivierender Unterricht einen Effekt auf die Schätzung der spezifischen Sensitivität hat. Inwiefern die Qualität der Lernmotivation von Schülerinnen und Schülern einen Einfluss auf die Schätzung der spezifischen Sensitivität bezogen auf die kognitive Aktivierung nimmt, soll mit der Hypothese 2.2 überprüft werden. Angenommen wird, dass sich die Schätzung der spezifischen Sensitivität mit und ohne Kontrolle der Qualität der Lernmotivation voneinander unterscheidet.

Eine letzte Hypothese 2.3 bezieht sich auf die Unterrichtsinhalte bzw. die im Unterricht verfolgten inhaltsbezogenen Lernziele und deren Gewichtung. Es wird die Auffassung vertreten, dass die im Unterricht verfolgten Lernziele in Kombination mit deren Gewichtung einen Einfluss auf die schulischen Leistungen der Schülerinnen und Schüler nehmen (D'Agostino et al., 2007; siehe auch Ausführungen in Kapitel 3.5.1). Insofern wird davon ausgegangen, dass die im Unterricht verfolgten Lernziele einen Effekt auf die Veränderung der klassenspezifischen Itemschwierigkeiten haben, welcher den Indikator der spezifischen Sensitivität darstellt. Inwiefern die Qualität der Lernmotivation von Schülerinnen und Schülern einen Einfluss auf die Schätzung der spezifischen Sensitivität bezogen auf die im Unterricht verfolgten Lernziele hat, soll mit der Hypothese 2.3 überprüft werden. Angenommen wird, dass sich die Schätzung der spezifischen Sensitivität mit und ohne Kontrolle der Qualität der Lernmotivation voneinander unterscheidet.

Sollten sich die Schätzungen der Sensitivität mit und ohne Einbeziehung der Qualität der Lernmotivation voneinander unterscheiden, würde dies dafür sprechen, dass dieses lernrelevante Merkmal die Indikatoren der Instruktionssensitivität beeinflusst. Die genauen Kriterien zur Hypothesenprüfung werden in Kapitel 4.5.4 dargelegt.

4 Methode

In diesem Kapitel steht das methodische Vorgehen zur Beantwortung der Fragestellung im Fokus. Hierfür werden in Kapitel 4.1 die INSE-Studie kurz vorgestellt und anschließend das Forschungsdesign in Kapitel 4.2 dargelegt. Die Stichprobenbeschreibung ist Inhalt des Kapitels 4.3, gefolgt von der Erläuterung der zum Einsatz gekommenen Instrumente in Kapitel 4.4. Zu den Instrumenten gehören der Klassencockpit-Test, zwei Schülerinnen- und Schüler-Fragebögen und der Lehrpersonen-Fragebogen. In Kapitel 4.5 stehen schließlich die Analysen im Vordergrund. In diesem Zusammenhang werden die statistischen Modelle aufgezeigt, die Analysestrategie beschrieben und Hinweise zu den Berechnungen bzw. Schätzverfahren gegeben. Darüber hinaus wird auf die Hypothesen und deren Überprüfung Bezug genommen.

4.1 Die INSE-Studie

Das Projekt „Instruktionssensitivität von Testitems in der Pädagogisch-Psychologischen Diagnostik“ (INSE) war ein binationales Forschungsprojekt zwischen der Pädagogischen Hochschule St. Gallen (PHSG) (Leitung: Prof. Dr. Jan Hochweber) und dem DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation (Leitung: Prof. Dr. Johannes Hartig). Gefördert durch den Schweizerischen Nationalfonds (SNF) und die Deutsche Forschungsgemeinschaft (DFG) wurden mit dem Projekt vier Hauptziele verfolgt: (1) „die Weiterentwicklung des [*längsschnittlichen mehrebenen differential item functioning models*] LML-DIF-Modells, (2) die Untersuchung der Qualität und Bedingungen der Parameterschätzungen, (3) die Validierung von statistischen Indikatoren der Instruktionssensitivität und (4) die Entwicklung von Kriterien zur Klassifikation von Instruktionssensitivität auf Test- und Itemebene“ (Hochweber & Leininger, o. D.). Die vorliegende Dissertation ist im Rahmen des INSE-Projekts entstanden.

4.2 Forschungsdesign

Zunächst wurden vom Bildungsdepartment des Kantons St. Gallen alle Schulleitungen der Mittel- und Unterstufe im Kanton per Mail angeschrieben. Die Schulleitungen wurden auf diesem Wege über das INSE-Projekt informiert. Die Rekrutierung der teilnehmenden Klassen erfolgte über die Schulleitungen, welche die Kontaktdaten von interessierten Lehrpersonen an

das St. Galler Forschungsteam weitergaben. Schulleitungen, die sich innerhalb eines vierwöchigen Zeitraums nicht gemeldet hatten, wurden vom St. Galler Forschungsteam telefonisch kontaktiert. Die Anzahl an Lehrpersonen, die sich zur Teilnahme am Projekt freiwillig bereit erklärten, konnte dadurch deutlich erhöht werden.

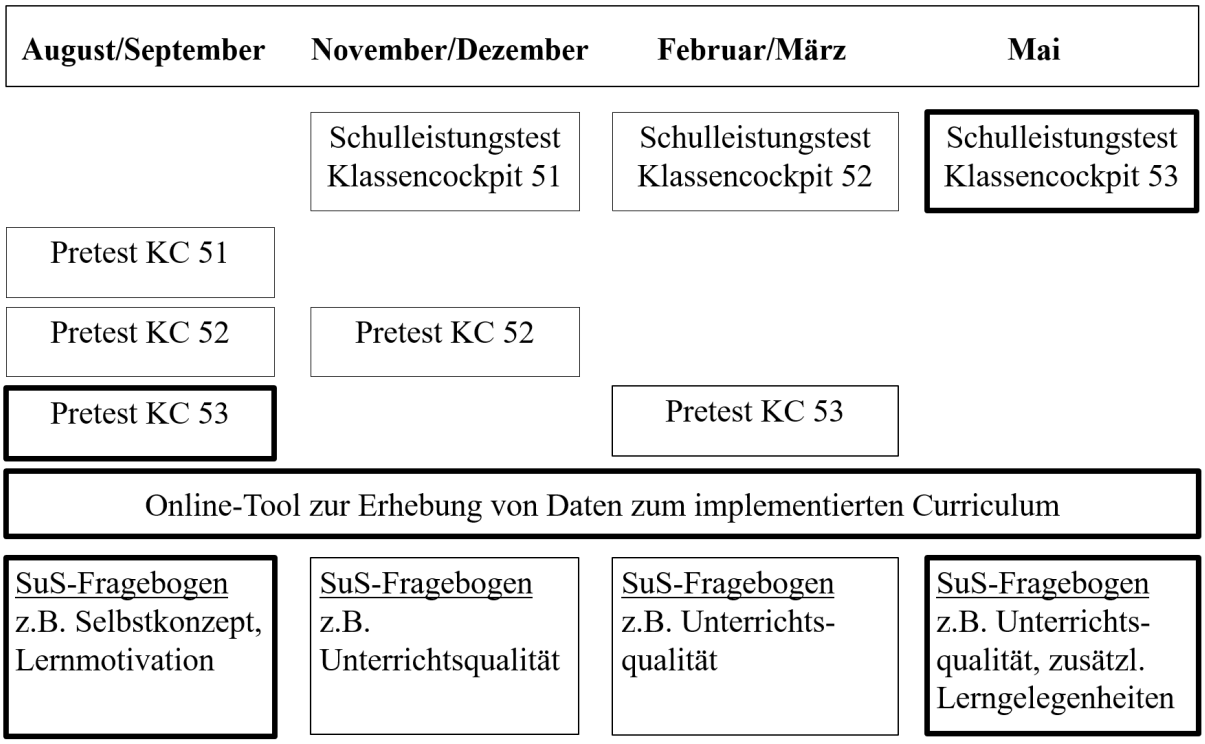


Abbildung 7. Forschungsdesign des Projekts INSE, Schuljahr 2016/17 (eigene Darstellung).

In der Studie wurde ein längsschnittliches Forschungsdesign fokussiert, das insbesondere auf Daten des Klassencockpit-Mathematiktests für die fünfte Klasse basiert (Abbildung 7). Klassencockpit ist ein seit ca. 20 Jahren implementierter Schulleistungstest, der die Fächer Mathematik und Deutsch abdeckt. Insofern wurde für die Studie auf langjährig erprobtes Testmaterial zurückgegriffen. Wie aus der Abbildung 7 ersichtlich wird, wurde jedem Klassencockpit-Test mindestens ein Pretest vorgeschaltet. Die Konstruktion der Pretests erfolgte durch das Forschungsteam. Das Vorgehen wird an späterer Stelle noch näher erläutert. Für die Dissertation wurden Daten des Klassencockpit 53-Pretests (*t1*, August/September) und Daten des Klassencockpit 53-Posttests (*t4*, Mai) verwendet. Die Zahl 53 steht dabei für das dritte Klassencockpitmodul in der fünften Klassenstufe.

Zusätzlich zu den Leistungstestdaten wurden Fragebögen eingesetzt, um von verschiedenen Zielgruppen Informationen zu Unterrichtsmerkmalen zu erhalten. Die Lehrpersonen nahmen ca. alle zwei Wochen an einer Online-Befragung zu den inhaltsbezogenen Lernzielen und zur

Gewichtung dieser Lernziele im Unterricht teil (insgesamt 17 Messzeitpunkte). Hierbei ging es insbesondere um die Frage, welche inhaltsbezogenen Lernziele zu welchem Zeitpunkt verfolgt wurden und ob diese Lernziele ein zentrales Unterrichtsziel darstellten. Von den Schülerinnen und Schülern wurden im gleichen Schuljahr Daten zur Unterrichtsqualität mittels Paper-Pencil-Methode zu drei Messzeitpunkten erhoben. Für die Analysen unter Einbeziehung der Unterrichtsqualität wurden nur die Daten zu Messzeitpunkt vier (*t4*) herangezogen. Diese können als retroperspektivische Einschätzung der Unterrichtsqualität erachtet werden. Darüber hinaus fand eine Befragung der Schülerinnen und Schüler ebenfalls mittels Paper-Pencil-Methode statt, um Daten zu den Ausprägungen lernrelevanter Merkmale der Schülerinnen und Schüler zu erhalten. Diese Daten wurden zum ersten Messzeitpunkt (*t1*) erhoben.

Alle für die Dissertation verwendeten Daten sind in Abbildung 7 mit einer fettgedruckten Umrahmung markiert. Die Datenerhebungen mittels Klassencockpit-Pretest, den regulären Klassencockpit-Tests und den Schülerinnen- und Schüler-Fragebögen wurden von geschulten Testadministratorinnen und -administratoren durchgeführt. Mit Ende des Schuljahres waren auch die Erhebungen in den Klassen abgeschlossen. Im Anschluss daran wurden noch sogenannte Expertinnen- und Experten-Ratings zur Einschätzung der Instruktionssensitivität von Testitems durchgeführt. An dieser Erhebung nahmen sowohl Lehrpersonen der beteiligten Klassen als auch Fachdidaktikerinnen und Fachdidaktiker aus der deutschsprachigen Schweiz teil. Die Expertinnen- und Expertenratings werden in der Dissertation jedoch nicht aufgegriffen (siehe hierfür Musow et al., 2019).

4.3 Stichprobenbeschreibung

In Tabelle 1 werden die Gesamtstichprobe sowie die Analysestichprobe beschrieben. Die Gesamtstichprobe konstituiert sich aus allen Schülerinnen und Schülern, zu denen anhand von Klassenlisten durch die Lehrperson Informationen ermittelt werden konnten. Die Sorgeberechtigten der Schülerinnen und Schüler wurden per Brief über das Projekt informiert und hatten die Möglichkeit, ihr Kind von der Studie abzumelden. Insgesamt neun Sorgeberechtigte machten von dieser Möglichkeit Gebrauch. Ferner wurden Schülerinnen und Schüler von den Analysen im Rahmen der Dissertation ausgeschlossen, wenn diese aufgrund von besonderen Lernbedürfnissen im Fach Mathematik lernzielbefreit waren. Auf eine Lernzielbefreiung im Fach Mathematik wurde geschlossen, wenn die Lehrperson ein besonderes Lernbedürfnis bei der Schülerin bzw. dem Schüler vermerkt und die Schülerin oder

der Schüler keine Zeugnisnote im Fach Mathematik hatte (Daten aus dem Vorjahr). Die Abfrage erfolgte ebenfalls über Klassenlisten. Die Anzahl an gültigen Fällen innerhalb der Stichprobe kann in Abhängigkeit zu den jeweiligen Messzeitpunkten variieren, da Schülerinnen und Schüler während des Schuljahres beispielsweise zu- oder auch weggezogen sind. Die Analysestichprobe der vorliegenden Arbeit basiert ausschließlich auf Fünftklässlerinnen und Fünftklässler, die im Fach Mathematik am regulären Unterricht teilnahmen und nicht lernzielbefreit waren. Insgesamt basieren die Analysen auf 832 Schülerinnen und Schüler.

Tabelle 1
Stichrobe der INSE-Studie

Schulen & Klassen		
Teilnehmende Schulen im Kanton St. Gallen	37	
Anzahl an Klassen	48	
Altersdurchmischte Klassen	12	
Durchschnittliche Klassengröße (Gesamtstichprobe/Anzahl an Klassen)	17.3	
Schülerinnen- und Schülerschaft		
Gesamtstichprobe	961	
Teilstichprobe Fünftklässler/-innen	850	
Ausschluss von Fünftklässler/-innen (Mehrfachnennung möglich)		
Von der Studie ausgeschlossen aufgrund besonderer Lernbedürfnisse	18	
Analysestichprobe		
Analysestichprobe Fünftklässler/-innen	832	
Geschlecht		
Analysestichprobe der Fünftklässler/-innen (w/m)	48.80%	51.20%

4.4 Instrumente

Im Rahmen des Dissertationsprojekts sind drei Instrumente zentral: (1) der Klassencockpit-Test, (2) die Schülerinnen- und Schüler-Fragebögen und (3) der Lehrpersonen-Fragebogen. Diese drei Instrumente sollen nachfolgend erläutert werden.

4.4.1 Klassencockpit-Test

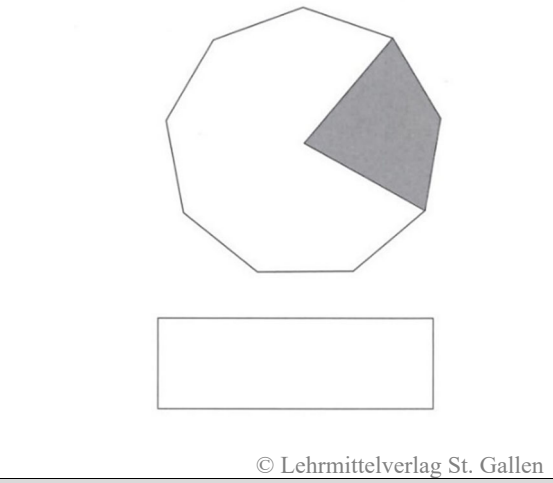
Klassencockpit ist ein Instrument zur Evaluation des Unterrichts, das auf eine 20-jährige Geschichte zurückblickt. In der deutschsprachigen Schweiz entwickelt, wurde dieses Instrument Mitte der 1990er Jahre für die Fächer Mathematik und Deutsch in mehreren Kantonen eingeführt und fortlaufend weiterentwickelt. Mittlerweile wurde das Klassencockpit durch das neu entwickelte Instrument Lernlupe abgelöst. Die Implementierung eines neuen Instruments ist mit Blick auf die Einführung des Lehrplans 21 in der Schweiz zu begrüßen. Inzwischen ist an allen Schulen die Umstellung auf den Lehrplan 21 erfolgt, die mit großen Herausforderungen verbunden war. Eine zentrale Neuerung bestand in der Homogenisierung der Lehrpläne unterschiedlicher Kantone und einer kompetenzorientierten Ausrichtung, die ein Umdenken in den Köpfen, neue didaktische Konzepte, passende Lehrmittel und adäquate Evaluationsinstrumente erforderte.

Zum Zeitpunkt des INSE-Projekts wurde noch nach dem St. Galler Bildungs- und Lehrplan unterrichtet und die Durchführung des Klassencockpits war bis zum Schuljahr 2018/19 im Kanton St. Gallen hochgradig empfohlen bzw. obligatorisch. Das vordergründige Ziel von Klassencockpit bestand für Lehrpersonen, Schülerinnen und Schüler in einer Standortbestimmung über den Lernerfolg der Schülerinnen und Schüler. Über ein Online-Portal konnten Lehrpersonen die Testscores der Schülerinnen und Schüler in ein System eingeben, das eine sofortige Auswertung für die Lehrperson bereitstellte. Lehrpersonen erhielten Informationen über das Abschneiden ihrer Klasse im Vergleich zur Normierungsstichprobe, konnten die Auswertung getrennt nach Geschlecht abrufen, bekamen Auskunft über Leistungsniveaus (drei Niveaustufen) und erhielten gemäß den Testleistungen der Schülerinnen und Schüler Notenvorschläge. Die Bereitstellung von Notenvorschlägen basierend auf den Testdaten der Schülerinnen und Schüler war mit der Entwicklung von Klassencockpit nicht intendiert. Vielmehr war es der ausdrückliche Wunsch zahlreicher Lehrpersonen, dem schließlich Rechnung getragen wurde. Die Durchführung des Klassencockpits im Fach Deutsch und Mathematik war jeweils 3x im Jahr à 2x ca. 45 Minuten vorgesehen und entsprechend mit einem hohen Zeitaufwand verbunden. Das Schreiben zusätzlicher Klassenarbeiten zu Zwecken der Notenvergabe hätte den zeitlichen Aufwand fürs Testen noch einmal deutlich erhöht. Für Schülerinnen und Schüler, die Noten auf Basis ihrer Klassencockpit-Ergebnisse erhalten haben, glich dies einem *high-stakes testing* (Baumert & Demmrich, 2001; Rutkowski & Wild, 2015).

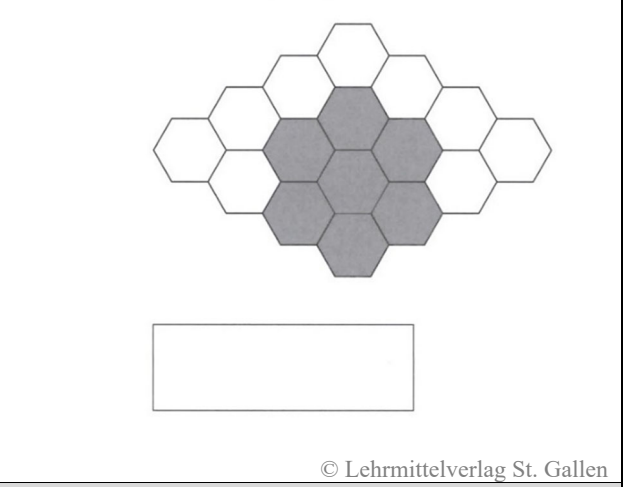
Die Entwicklung des Klassencockpits zur Erfassung mathematischer Kompetenzen wurde von Expertinnen und Experten gemäß wissenschaftlichen Kriterien vorgenommen. Die im Rahmen der Dissertation verwendeten Testitems stammen aus dem originalen Klassencockpit-Test und sind gut erprobt. Folgende wissenschaftliche Standards wurden bei der Entwicklung des Klassencockpits berücksichtigt: (1) die Analyse des Curriculums, (2) die Konstruktion der Testitems unter Berücksichtigung einer adäquaten Itemgestaltung (z.B. kontextuelle Einbettung), die Wahl eines adäquaten Antwortformats und die Repräsentativität der Items in Anbetracht des Curriculums, (3) die Überprüfung der Items auf psychometrische Eigenschaften und (4) die Normierung (Moser, 2003; Roick, 2008). Darüber hinaus wurde das Klassencockpit von der Universität Zürich wissenschaftlich evaluiert (Moser, 2003) und durch die Pädagogische Hochschule St. Gallen wissenschaftlich begleitet.

Für die Umsetzung des Dissertationsprojekts war es zum einen wichtig, eine hinreichend große und genestete Datenstruktur zu haben. Zum anderen brauchte es eine längsschnittliche Perspektive hinsichtlich der Leistungstestdaten. Beides sind Voraussetzungen, um die Testitems mit Blick auf die Gruppen-x-Zeitpunkte-Perspektive zu analysieren. Zur Umsetzung eines *Pretest-Posttest-Designs* wurden aus den regulären Klassencockpit-Tests für das Fach Mathematik nach inhaltlichen Kriterien und psychometrischen Eigenschaften der Testitems jeweils 15 bis 17 Testitems pro Klassencockpitmodul (51, 52, 53) ausgewählt und zu einem Pretest zusammengestellt.

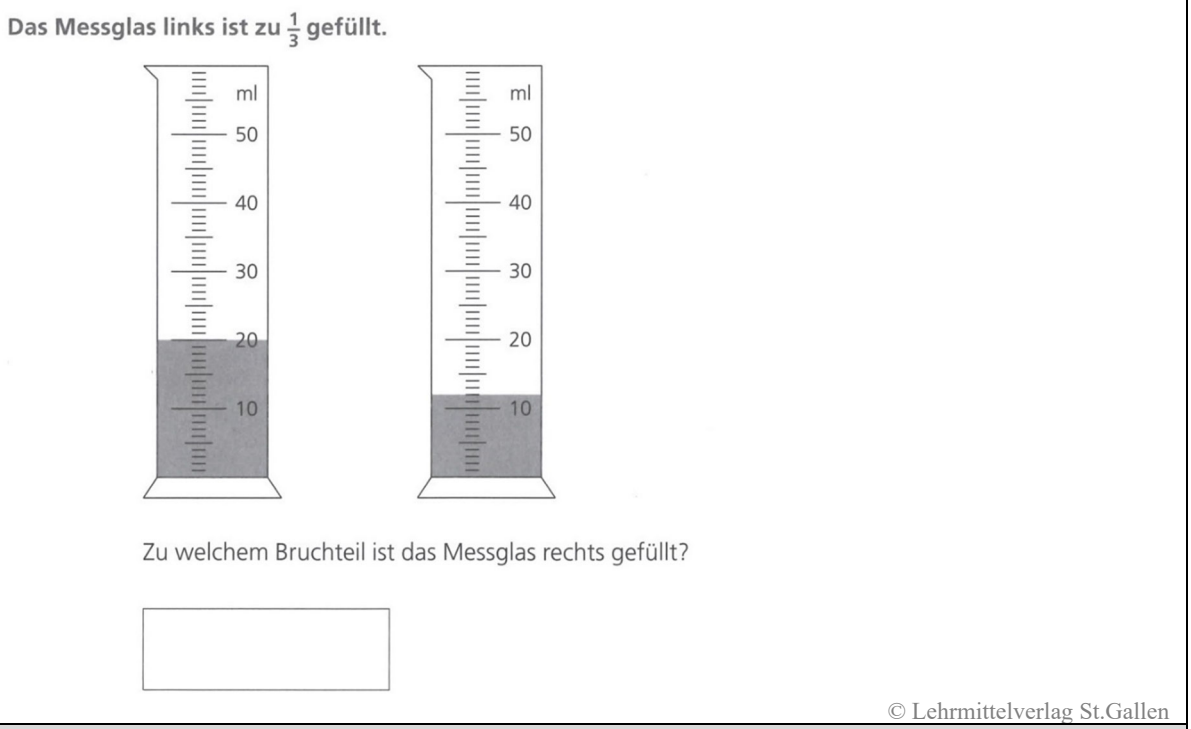
Auf inhaltlicher Ebene hat sich das Projektteam für die Abdeckung von Themen der Arithmetik für die fünfte Klassestufe und gegen die Bereiche Geometrie sowie Funktionen und Relationen entschieden. Der Bereich der Arithmetik hält insbesondere zwei Themen bereit, die in der fünften Klasse neu eingeführt werden: (1) dezimale Zahlen und (2) Brüche. In der vorliegenden Arbeit wird der Fokus auf die in Abbildung 8 dargelegten Testitems gelegt. Ein Argument für die Entscheidung liegt in der Alltagsrelevanz begründet. Dezimale Zahlen haben im Alltag eine deutlich höhere Präsenz als Brüche (Padberg & Wartha, 2017). Vor diesem Hintergrund ist anzunehmen, dass der Alltag weniger Lerngelegenheiten im Umgang mit Brüchen als mit dezimalen Zahlen bietet. Die Bedeutung des Unterrichts zum Erwerb von Kompetenzen im Umgang mit Brüchen sollte dementsprechend hoch sein (ebd.). Es wird daher davon ausgegangen, dass sich das Thema Brüche für exemplarische Analysen zur Evaluation der Instruktionssensitivität von Testitems besonders eignet. Mit Blick auf die fünfte Klassenstufe soll der Unterricht insbesondere darauf abzielen, mathematische Grundvorstellungen zu



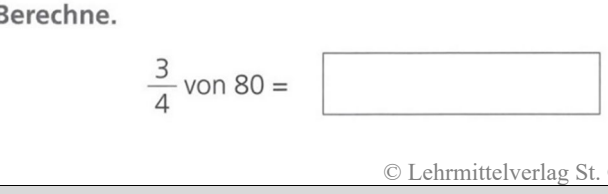
Item A01



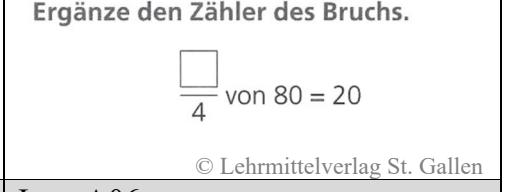
Item A02



Item A03



Item A05



Item A06

Abbildung 8. Klassencockpit-Testitems, welche in die Analysen zur Instruktionssensitivität eingegangen sind.

100 Personen wurden zu ihren Tätigkeiten in der Freizeit befragt.

Freizeitaktivitäten

Tätigkeit	täglich	wöchentlich	monatlich	seltener	nie
Freunde, Kollegen und Bekannte treffen	12	61	20	5	2
Lesen	70	19	5	3	3
Sport treiben	6	50	12	4	28
Kurse besuchen	0	11	6	9	74
Musizieren, Singen	8	15	5	3	69
Handarbeit, Werken, Gartenarbeit	12	38	16	8	26
Kino besuchen	0	3	27	36	34
Theater, Opern oder Ausstellungen besuchen	0	2	24	42	32
Sportanlässe besuchen	0	6	18	22	54
Fernsehen	83	8	4	3	2
Internet benützen	75	12	2	0	?

Welche Aussage ist richtig?

- ☐ Ein Drittel der befragten Personen besuchen monatlich ein Kino.
- ☐ Drei Fünfundzwanzigstel der Befragten benützen wöchentlich das Internet.
- ☐ Ein Fünftel der Befragten singt täglich.
- ☐ Ein Viertel der Befragten besucht nie einen Sportanlass.

Item A21

Abbildung 8. Klassencockpit-Testitems, welche in die Analysen zur Instruktionssensitivität eingegangen sind (Fortsetzung).

entwickeln, spezifische Schreibweisen und Symbole zu erlernen und prozedurales Wissen (Rechentechniken) aufzubauen. Der Schwerpunkt liegt dabei gemäß dem obligatorischen Lehrmittel auf der Entwicklung mathematischer Grundvorstellungen (Noser, Oester Schläppi, Scherer & Züger, 2005a, 2005c). Letzterer kommt ein besonderer Stellenwert zu, da dieser Bereich auch für spätere mathematische Themen (z.B. Wahrscheinlichkeitslehre, Gleichungslehre und Algebra) von großer Bedeutung ist (Padberg & Wartha, 2017).

Neben den inhaltlichen Überlegungen wurden auch psychometrische Eigenschaften der Items überprüft, um diese ebenfalls in die Itemauswahl einfließen zu lassen (siehe u.a. Wright & Masters, 1982). Die psychometrischen Eigenschaften wurden anhand der Normierungsstichprobe des Klassencockpits für die fünfte Klassenstufe im Fach Mathematik

analysiert. Die Parameterschätzungen erfolgten mit dem Paket TAM (Robitzsch, Kiefer & Wu, 2021)[12] in R (R Core Team, 2017) und wurden zusammen mit dem Team vom DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation durchgeführt. Die Itemselektion wurde insbesondere unter Berücksichtigung der Itemschwierigkeit und der Trennschärfe (Kalava & Moosbrugger, 2012) vorgenommen. Bezogen auf die Itemschwierigkeiten wurde darauf geachtet, dass sich diese in einem Bereich von ≥ .20 und ≤ .80 bewegten (siehe auch Roick, 2008). Darüber hinaus wurden nur Items ausgewählt, deren Trennschärfe bei ≥ .30 lag (siehe auch Roick, 2008). Ein weiteres Kriterium stellten die Infit- und Outfit-Werte dar, um die Passung zwischen den empirischen Werten und den Annahmen des Rasch-Modells zu überprüfen. Nur Items, deren Infit- und Outfit-Werte sich im Bereich zwischen 0.75 und 1.2 bewegten, wurden berücksichtigt (in Anlehnung an Bond & Fox, 2015). Darüber hinaus wurde zudem die Ökonomie bei der Bearbeitung der Testitems berücksichtigt. Da die Zeit zum Testen von Schülerinnen und Schülern begrenzt war, sollten die Items möglichst ökonomisch hinsichtlich der Bearbeitungsdauer sein.

Nach der Auswahl der Testitems (Abbildung 8), die im Längsschnitt erhoben werden sollten, stellte sich noch die Frage nach deren Anordnung innerhalb der Tests. Sowohl der Pretest als auch der Posttest wurden für die Erhebungen zusammengestellt bzw. Testitems wurden in einer neuen Reihenfolge angeordnet. Einen wichtigen Aspekt, den es zu berücksichtigen galt, stellten die Itempositionseffekte dar (Nagy, Frey, Nagengast, Becker & Rose, 2016). Um möglichst aussagekräftige Ergebnisse zu erlangen, wurden die Pretest-Items nicht gebündelt an den Anfang oder an das Ende des Posttests (Klassencockpit 51 bzw. 52) gestellt, sondern in ihn integriert. Damit sollte vermieden werden, dass sich mit fortschreitender Testbearbeitung ein Abfall in der Leistungsfähigkeit auf ein Itembündel (z.B. Pretest-Items, wenn diese ans Testende verortet werden) stark auswirkt und die Ergebnisse dadurch verzerrt würden. Bei der Anordnung der Items wurde außerdem darauf geachtet, dass verhältnismäßig leichte Items am Anfang des Tests stehen, um den Einstieg in die Testbearbeitung zu erleichtern (Pospeschill, 2010). Des Weiteren wurde bei der Anordnung der Testitems berücksichtigt, dass in den regulären Klassencockpit-Tests oftmals mehrere Items thematisch zusammenhängen. Die ursprüngliche Anordnung der Items wurde – trotz der Integration der Pretest-Items – dahingehend gewahrt, dass Items mit derselben thematischen Einbettung oder den gleichen mathematischen Inhalten in kleinere Einheiten gebündelt und im Test verortet wurden. Dieser

[12] TAM Paket, Version 3.1-45 von Robitzsch et al. (2019). Die Webseite mit der damaligen Version ist nicht mehr abrufbar, weshalb die neuere Quelle zitiert wurde.

Aspekt wirkt einer Fragmentierung der einzelnen Testitems entgegen und ist insbesondere dann sinnvoll, wenn sich mehrere Aufgaben auf dieselbe Abbildung beziehen. In Anbetracht dessen, dass es sich um eine längsschnittlich angelegte Studie handelt, wurden die Itempositionen in den jeweiligen Tests über die verschiedenen Messzeitpunkte möglichst konstant gehalten. Dies soll der Vergleichbarkeit von Ergebnissen auf Itemebene (insbesondere Itemschwierigkeiten) über die Zeit dienen.

4.4.2 Schülerinnen- und Schüler-Fragebogen zur Erfassung individueller Merkmale und der Unterrichtsqualität

Neben dem Klassencockpit kamen Schülerinnen- und Schüler-Fragebögen zu vier Messzeitpunkten zum Einsatz. Die Fragebogenerhebungen bei den Schülerinnen und Schülern wurden jeweils im Anschluss an die Klassencockpit-Erhebung von geschulten Testadministratorinnen und Testadministratoren durchgeführt. Analog zur Klassencockpit-Erhebung erfolgte die Bearbeitung der Schülerinnen- und Schüler-Fragebögen via Paper-Pencil-Methode.

Zum ersten Messzeitpunkt im August/September wurden demografische Daten sowie lernrelevante Merkmale der Schülerinnen und Schüler abgefragt. Als demografische Merkmale wurden das Alter, das Geschlecht, die Herkunft, die Erstsprache, der Sprachgebrauch, die Bildungsressourcen sowie die Wohlstandsgüter erhoben. Hinsichtlich der lernrelevanten Merkmale der Schülerinnen und Schüler wurden Daten zur Qualität der Lernmotivation, zur Handlungskontrolle, zum akademischen Selbstkonzept, zur Zielorientierung, zu kognitiven und metakognitiven Lernstrategien und zur Mitarbeit im Unterricht erhoben. Bei den nachfolgenden Messzeitpunkten zwei bis vier lag der Fokus auf einer Datenerhebung zu den Basisdimensionen der Unterrichtsqualität.

Mit Blick auf die Dissertation sind die Skalen zur Qualität der Lernmotivation und zur Unterrichtsqualität von zentraler Bedeutung. Beginnend mit der Qualität der Lernmotivation beziehen sich die verschiedenen Subskalen auf die Selbstbestimmungstheorie (Deci & Ryan, 1993, 2008, Ryan & Deci, 2000), die bereits im Kapitel 3.4.2 ausgeführt wurde. Die im Rahmen dieser Arbeit verwendeten Skalen stammen von Frey et al. (2009), die laut der Autorinnen und Autoren auf die Arbeit von Seidel et al. (2005) zurückgehen. Der eigentliche Ursprung der Skalen liegt jedoch bei Ryan und Cornell (1989). Gemäß der Theorie von Ryan und Deci (2000) weist die Qualität der Lernmotivation verschiedene Subdimensionen auf: (1) die

Amotiviertheit, (2) die externale Motiviertheit, (3) die introjizierte Motiviertheit, (4) die identifizierte Motiviertheit, (5) die interessierte Motiviertheit und (6) die intrinsische Motiviertheit (Kapitel 3.4.2). Damit die Subskalen von Frey et al. (2009) im Rahmen der Dissertation angewendet werden konnten, wurden die Items entsprechend der Zielgruppe und dem Kontext modifiziert. Um eine Vorstellung von den jeweiligen Subdimensionen zu erhalten, wurde jeweils ein Itembeispiel für jede Subdimension angeführt (siehe Tabelle 2). Die Subdimensionen unterscheiden sich – wie im Kapitel 3.4.2 dargelegt – hinsichtlich des Grads an Selbstbestimmung.

Tabelle 2
Itembeispiele zu den Subskalen der Qualität der Lernmotivation

Subdimension	Itembeispiel
Amotiviertheit	Im Mathematikunterricht bin ich mit meinen Gedanken woanders.
Externale Motiviertheit	Im Mathematikunterricht arbeite ich nur mit, wenn ich dazu aufgefordert werde.
Introjizierte Motiviertheit	Im Mathematikunterricht strenge ich mich an, weil ich alles richtig machen möchte.
Identifizierte Motiviertheit	Im Mathematikunterricht arbeite ich mit, weil ich Mathematik später im Leben bestimmt brauche.
Interessierte Motiviertheit	Im Mathematikunterricht möchte ich gern mehr über einzelne Themen erfahren.
Intrinsische Motiviertheit	Im Mathematikunterricht macht mir der Unterricht Spaß.[13]

Anmerkung zum Antwortformat: 1=nie, 2=manchmal, 3=fast immer, 4=immer.

Neben den Skalen zur Qualität der Lernmotivation wurden Daten zu den Basisdimensionen der Unterrichtsqualität erhoben. Verwendet wurden die Daten des vierten Messzeitpunkts, sodass es sich um eine retroperspektivische Einschätzung der Unterrichtsqualität handelt. Auch die Basisdimensionen zur Unterrichtsqualität wurden bereits im Kapitel 3.5.2 erläutert, sodass hier lediglich Beispielitems zur Veranschaulichung der Skalen aufgeführt werden. Die verwendeten Skalen gehen auf Fauth et al. (2014) zurück. Eine deutsche Übersetzung der Skalen ist im unveröffentlichten Skalenbuch der IGEL-Studie (siehe Decristan, Hertel et al., 2014) zu finden. Da sich die Skalen von Fauth et al. (2014) im Original auf den Sachunterricht beziehen, wurden die Itemformulierungen auf den Kontext des Mathematikunterrichts angepasst. In Tabelle 3

[13] In der vorliegenden Arbeit wird die deutsche und nicht die schweizerische Rechtschreibung verwendet. Aus diesem Grund wurde die Rechtschreibung insbesondere bei der ausschnitthaften Darstellung von eingesetzten Instrumenten angepasst. Eine Ausnahme bilden Screenshots, welche der Originalfassung (schweizerische Rechtschreibung) entsprechen. Im Anhang A bis C sind die Instrumente ebenfalls im Original aufgeführt.

sind die Skalen zu den verschiedenen Basisdimensionen der Unterrichtsqualität mit je einem Itembeispiel aufgeführt. Das unterstützende Klassenklima wird hier ebenfalls aufgeführt, auch wenn diese Subdimension in den Hypothesen keine Berücksichtigung erfährt. Hintergrund ist, dass bei der späteren konfirmatorischen Mehrebenen-Faktorenanalyse das Gesamtmodell der Unterrichtsqualität hinsichtlich der Passung überprüft wird.

Tabelle 3
Itembeispiele zu den Subskalen der drei Basisdimensionen der Unterrichtsqualität

Subdimension	**Itembeispiel**
Unterrichtsstörung	„Im Mathematikunterricht stört keiner den Unterricht.“
Unterstützendes Klassenklima	„Unsere Lehrperson im Mathematikunterricht macht mir Mut, wenn ich eine Aufgabe schwierig finde.“
Kognitive Aktivierung	„Unsere Lehrperson in Mathematik stellt Fragen, über die ich ganz genau nachdenken muss.“

Anmerkung zum Antwortformat: 1=stimme gar nicht zu, 2=stimme eher nicht zu, 3=stimme eher zu, 4=stimme voll zu.

4.4.3 Lehrpersonen-Fragebogen zur Erhebung inhaltsbezogener Lernziele im Mathematikunterricht und Bildung des Prädiktors

Die Erfassung von Unterrichtsmerkmalen erfolgte nicht nur über die Schülerinnen und Schüler, sondern auch über die Lehrpersonen. In einem i.d.R. zweiwöchigen Rhythmus (Abweichungen gab es lediglich aufgrund von Schulferien) wurden die an der Studie beteiligten Lehrpersonen mithilfe eines Online-Fragebogens befragt, welche inhaltsbezogenen Lernziele sie im Unterricht verfolgt haben. Der in Zusammenarbeit mit Mathematik-Fachdidaktikern der Pädagogischen Hochschule St. Gallen entwickelte Online-Fragebogen orientierte sich dabei eng am St. Galler Bildungs- und Lehrplan.

Im Instrument spiegelt sich u.a. die Struktur des Bildungs- und Lehrplans wider, welcher sich in drei Teilbereiche aufgliedern lässt: (1) der Bereich Arithmetik und Algebra, (2) der Bereich Funktionen und Relationen und (3) der Bereich Geometrie. Jeder dieser Teilbereiche weist formulierte Richtziele auf wie beispielsweise „Zahlvorstellung entwickeln“ oder „Die Umwelt mit Größen erfassen“ (Abbildung 9). Ausgehend von den Richtzielen werden im Lehrplan wiederum Grobziele aufgeführt, auf deren Ebene die Erhebung der Lernziele stattfand. Um eine Vorstellung von den Grobzielen zu erhalten, soll auch hier ein Beispiel genannt werden. Das Beispielitem lautet in diesem Fall: „Dezimale Zahlen und Brüche als Erweiterung des Bereichs der natürlichen Zahlen erfahren“. Wie bereits dargelegt, wird in der vorliegenden Arbeit der

1 Arithmetik & Algebra			
	1.1 Zahlvorstellung entwickeln		
		1.1.1 Orientierung im Zahlenraum	
			L1Aza02: Dezimale Zahlen und Brüche als Erweiterung des Bereichs der natürlichen Zahlen erfahren
			L1Aza03: Einfache Brüche der Größe nach ordnen
			L1Aza04: Einfluss von Zähler und Nenner auf den Wert des Bruches erkennen
		1.1.2 Darstellung von Zahlen	
		1.1.3 Eigenschaften von Zahlen	
	1.2 Mit Größen die Umwelt erfassen		
		1.2.1 Kalender und Zeiten	
		1.2.2 Dezimale Größen	
		1.2.3 Zusammengesetzte Größen	
	1.3 Operationen verstehen und ausführen		
		1.3.1 Addition & Subtraktion	
			L1Aop06: Brüche kürzen und erweitern
			L1Aop07: Einfluss von Veränderungen im Zähler und Nenner auf den Wert von Brüchen erkennen
		1.3.2 Multiplikation & Division	
		1.3.3 Terme	
		1.3.4 Gleichungen	

Abbildung 9. Darlegung der Struktur im Fragebogen und den für die Dissertation relevanten Items (L1Aza02 bis L1Aop07) (eigene Darstellung).

Fokus auf den Lernzielen und Items zu Brüchen liegen. Insofern wird an dieser Stelle nur ein entsprechender Ausschnitt des Instruments dargestellt.[14] Mit Blick auf Abbildung 9 gilt es zu bedenken, dass sich die Lernziele sowohl auf die fünfte als auch auf die sechste Klassenstufe beziehen, da im Bildungs- und Lehrplan jeweils zwei Klassenstufen zusammengefasst werden.

Die Erhebung der Daten erfolgte mithilfe zweier Fragestellungen (Abbildung 10). (1) „Welche Lernziele haben Sie in Ihrem Unterricht in den letzten beiden Wochen verfolgt? Bitte tragen Sie je Kalenderwoche die Anzahl an Lektionen (1-5) in die jeweiligen Kästchen ein.“ (2)

[14] Für einen Überblick über das gesamte Instrument siehe Anhang C.

„Welche Lernziele waren in Ihrem Unterricht in den letzten beiden Wochen zentral? Bitte kreuzen Sie die zentralen Lernziele an.“ Die Lernziele wurden im Fragebogen aufgeführt, sodass das Instrument eine hohe Standardisierung aufwies. Darüber hinaus wurden zu jedem Lernziel ein Aufgabenbeispiel sowie ein Verweis auf die entsprechende Seite im Lehrmittel „Logisch 5“ gegeben. Für die Bearbeitung des Fragebogens hatten die Lehrpersonen jeweils eine Woche Zeit. Den Zeitpunkt zum Ausfüllen des Fragebogens konnten sie frei wählen. Wurde der Fragebogen innerhalb des genannten Zeitfensters nicht ausgefüllt, bekamen die entsprechenden Lehrpersonen eine automatisierte Erinnerungsmail mit der Bitte, den Fragebogen innerhalb der nächsten sechs Tage zu bearbeiten. Ein besonderer Dank gilt den Lehrpersonen, die insgesamt 17 Mal den umfangreichen Fragebogen ausgefüllt haben.

Beispiel zum Ausfüllen des Fragebogens
1.1.1 Orientierung im Zahlenraum

	Welche Lernziele haben Sie in Ihrem Unterricht in den letzten beiden Wochen verfolgt? Bitte tragen Sie je Kalenderwoche die Anzahl an Lektionen (1-5) in die jeweiligen Kästchen ein.		**Welche Lernziele waren in Ihrem Unterricht in den letzten beiden Wochen zentral?** Bitte kreuzen Sie die zentralen Lernziele an.
	KW 35	KW 36	KW 35/36
Dezimale Zahlen und Brüche als Erweiterung des Bereichs der natürlichen Zahlen erkennen	1	3	✓
Einfache Brüche der Größe nach ordnen		1	

Abbildung 10. Ausschnitt aus dem Fragebogen zur Erhebung der im Unterricht verfolgten Lernziele (eigene Darstellung).

Im Unterricht verfolgte Lernziele – Bildung des Prädiktors

Im obligatorischen Lehrmittel „Logisch 5“ der fünften Klassenstufe im Kanton St. Gallen (Stand Schuljahr: 2016/17) sind insgesamt zwei Kapitel zur Bruchrechnung enthalten. Der Fokus in der vorliegenden Arbeit wird auf dem ersten Buchkapitel „Brüche“ liegen, in dem es um den Aufbau von mathematischen Grundvorstellungen zu diesem Thema geht (Noser et al., 2005a, 2005c). Ein wesentliches Ziel, das mit dem Unterricht zu diesem Buchkapitel verfolgt wird, besteht in der Entwicklung einer Vorstellung, dass der Zähler und der Nenner der Beschreibung eines Bruchs dienen (Noser et al., 2005a, 2005c).

Fünf Lernziele des St. Galler Bildungs- und Lehrplans werden mit dem Kapitel „Brüche“ in Verbindung gebracht (Noser et al., 2005c): (1) „Orientierung im Zahlenraum: Dezimalzahlen und Brüche als Erweiterung des Bereichs der natürlichen Zahlen erfahren“ (Lernziel: Aza02), (2) „Einfache Brüche der Größe nach ordnen“ (Lernziel: Aza03), (3) „Einfluss von Zähler und Nenner auf den Wert des Bruchs erkennen“ (Lernziel: Aza04), (4) „Brüche kürzen und erweitern“ (Lernziel: Aop06) und (5) „Einfluss von Veränderungen im Zähler und Nenner auf den Wert von Brüchen erkennen“ (Lernziel: Aop07). Die Lernziele wurden in Anlehnung an das Buchkapitel „Brüche“ gebündelt und stellen einen Prädiktor dar. Jedes Lernziel separat ins Modell aufzunehmen, wäre nicht zielführend gewesen, da sich die Zuordnung zwischen dem jeweiligen Lernziel des St. Galler Bildungs- und Lehrplans und den jeweiligen Items des Klassencockpit-Tests als schwierig und oftmals nicht eindeutig erwies. Das Problem liegt vor allem darin, dass mit den jeweiligen Aufgaben mehrere Lernziele verfolgt werden können, aber nicht müssen. Darüber hinaus besteht ein gewisser Interpretationsspielraum: Je nach Ansicht können den jeweiligen Testitems ein oder mehrere Lernziel/e zugeordnet werden. Der Lehrmittelverlag, der sowohl für das Lehrmittel als auch für die Klassencockpit-Tests verantwortlich war, stellte eine solche Zuordnung zwischen den Lernzielen des St. Galler Bildungs- und Lehrplans und den Klassencockpit-Items ebenfalls nicht bereit.

Welche Daten werden zur Bildung des Prädiktors herangezogen?

Es wurden zwei Arten von Daten zur Bildung des Prädiktors herangezogen: zum einen Daten, die angaben, wie häufig ein Lernziel in der jeweiligen Woche verfolgt wurde, und zum anderen Daten, die angaben, welche der im Unterricht verfolgten Lernziele für den Unterricht zentral waren (Abbildung 10). Bezugnehmend auf die Literatur zur Instruktionssensitivität sind beide Maße (Quantität: z.B. Hanson et al., 1986; Wiley & Resnick, 1998; Gewichtung: Muthén et al., 1995; D'Agostino et al., 2007) stark vertreten. Darüber hinaus könnten mit den vorliegenden Daten auch Angaben im Sinne von behandelt/nicht behandelt generiert werden, was ebenfalls in der Literatur vorzufinden ist (Haladyna & Roid, 1981). Keines der drei genannten Unterrichtsmerkmale (Quantität, Gewichtung, Abdeckung von inhaltsbezogenen Lernzielen) kann dabei als überlegen erachtet werden. Die Entscheidung in der vorliegenden Arbeit fällt auf die Verwendung von Daten zur Quantität der im Unterricht verfolgten Lernziele und ob die jeweils im Unterricht verfolgten Lernziele als zentral für den Unterricht erachtet wurden (siehe auch D'Agostino et al., 2007). Der Prädiktor wird infolgedessen als Interaktionsterm gestaltet basierend auf Angaben zur Quantität und zur Gewichtung der fünf genannten Lernziele.

Erhebungszeitraum

In Abbildung 11 sind die Zeiträume zur Erhebung der im Unterricht verfolgten Lernziele, der Klassencockpit-Erhebungen und der Schülerinnen- und Schüler-Fragebogen-Erhebungen abgebildet. Bezugnehmend auf das obligatorische Lehrmittel „Logisch 5“ und den Klassencockpit-Test ist das Kapitel „Brüche“ im Zeitraum zwischen der Klassencockpit-Erhebung von *t3* zu *t4* zu verorten. Abweichungen hinsichtlich der Reihenfolge zu den Unterrichtsthemen sind aber durchaus vorstellbar. Insofern werden Daten zu den im Unterricht verfolgten Lernzielen vom ersten (*t1*) bis zum 17. Messzeitpunkt (*t17*) einbezogen.

Mo	August/September					Oktober							November/Dezember			
KW	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
KC SuS	*t1*	*t1*	*t1*			FER	FER	FER						*t2*	*t2*	*t2*
LZ		*t1*		*t2*	*t3*	FER	FER	FER		*t4*		*t5*		*t6*		*t7*

Mo	Dezember/Januar					Februar										März/April
KW	51	52	1	2	3	4	5	6	7	8	9	10	11	12	13	14
KC SuS		FER	FER			z.T. FER	z.T. FER	z.T. FER		*t3*	*t3*	*t3*				
LZ	*t8*	FER	FER		*t9*	z.T. FER	z.T. FER	*t10*		*t11*		*t12*		*t13*		*t14*

Mo	April/Mai						Juni	
KW	15	16	17	18	19	20	21	22
KC SuS	FER	FER				*t4*	*t4*	*t4*
LZ	FER	FER		*t15*		*t16*		*t17*

Abbildung 11. Erhebungszeitpunkte gemäß Monat (Mo) und Kalenderwochen (KW) im Schuljahr 2016/17 (eigene Darstellung).

Anmerkung: Mo: Monat; KW: Kalenderwoche; KC SuS: Klassencockpit-Testung & Schülerinnen- und Schüler-Fragebogen-Erhebung; LZ: Erhebung der im Unterricht verfolgten Lernziele; FER: Ferien; z.T. FER: zum Teil Ferien. Der Zeitpunkt der Sportferien variiert zwischen den Schulen im Kanton St. Gallen, weshalb die jeweiligen Wochen mit z.T. FER gekennzeichnet wurden.

Rücklauf

Der Rücklauf der jeweiligen Umfrage ist der Tabelle 4 zu entnehmen. In den Zeilen sind die Identifikationsnummern der Lehrpersonen vermerkt, in den Spalten die Messzeitpunkte. Bei der Darstellung der deskriptiven Ergebnisse wurden alle 48 Klassen berücksichtigt. Wie anhand der Tabelle 4 zu erkennen ist, schwankte die Rücklaufquote zwischen 75% (36 Lehrpersonen) und 92% (44 Lehrpersonen). Die Teilnahme ist vor dem Hintergrund, dass die Lehrpersonen über ein gesamtes Schuljahr i.d.R. alle zwei Wochen um Angaben gebeten wurden, als sehr

zufriedenstellend zu bewerten. Im Mittel lag die Rücklaufquote bei 80.76%. Von den insgesamt 48 Lehrpersonen haben 20 an allen 17 Erhebungen teilgenommen.

Tabelle 4

Fragebogenrücklauf pro Messzeitpunkt und pro Lehrperson

ID	*t1*	*t2*	*t3*	*t4*	*t5*	*t6*	*t7*	*t8*	*t9*	*t10*	*t11*	*t12*	*t13*	*t14*	*t15*	*t16*	*t17*	Summe
1011	X	1	1	1	1	1	X	1	1	X	X	X	1	X	X	1	X	**9**
1012	1	1	1	1	1	1	1	X	1	1	1	1	1	1	1	1	1	**16**
1021	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
1022	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
2031	X	X	1	1	X	1	1	1	1	X	1	X	1	X	1	X	1	**10**
2041	1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	**1**
2042	1	1	1	1	1	X	1	1	1	1	1	1	1	1	1	1	1	**16**
2051	1	1	1	X	1	X	1	X	X	X	X	X	X	X	X	X	X	**5**
2052	1	1	1	1	1	1	1	1	1	X	1	1	1	1	1	1	1	**16**
4061	1	1	1	X	1	1	1	1	1	1	1	1	1	1	1	1	1	**16**
6071	1	1	1	1	1	1	1	1	X	1	1	1	X	X	1	1	1	**14**
7081	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
7091	1	1	1	1	X	X	X	X	X	X	X	X	X	X	X	X	X	**4**
7092	1	1	X	1	X	X	X	X	X	X	X	X	X	X	X	X	X	**3**
7111	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
7112	1	1	X	X	1	X	1	1	1	X	X	1	1	1	1	1	X	**11**
8121	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
8131	1	X	X	1	1	1	1	1	1	1	1	X	X	1	1	1	1	**13**
10141	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
10152	1	1	1	X	1	1	X	1	1	X	X	X	X	1	1	1	X	**10**
10153	1	1	1	1	1	X	1	1	X	1	1	1	1	1	1	1	X	**14**
11161	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
12171	1	X	X	1	1	X	X	X	X	1	1	1	X	1	1	1	X	**9**
14191	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
15201	1	1	1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	**3**
15212	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	**0**
18221	1	1	1	1	1	1	1	1	1	1	1	1	1	X	1	1	1	**16**
19222	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
19231	1	1	1	1	1	1	1	1	X	1	1	1	X	1	1	1	1	**15**
19241	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
19261	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
19262	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
20271	1	1	1	1	1	1	X	1	1	1	1	1	1	1	1	1	1	**16**
20281	1	X	1	1	1	1	1	1	1	X	1	1	1	1	1	1	1	**15**
21291	1	1	1	1	1	1	X	X	X	1	X	1	1	1	1	1	1	**13**
22301	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
23311	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
23312	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
23321	X	X	X	1	1	1	1	1	1	1	1	X	1	X	1	1	1	**12**
23322	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
23331	1	X	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**16**
23332	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
23341	1	1	1	1	X	1	1	1	1	1	1	1	X	1	1	1	1	**15**
23351	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
23352	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
23353	1	1	1	1	1	1	1	X	1	1	X	1	1	1	1	1	1	**15**
23354	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**17**
23361	1	1	1	1	1	1	X	1	1	1	1	1	1	1	1	1	1	**16**
Summe	**44**	**40**	**41**	**41**	**41**	**38**	**37**	**38**	**37**	**36**	**37**	**37**	**36**	**37**	**41**	**41**	**37**	

Anmerkung: Eine „1“ bedeutet, dass der Fragebogen zum entsprechenden Messzeitpunkt ausgefüllt wurde, während ein „X“ darauf hinweist, dass von einem Ausfüllen des Fragebogens abgesehen wurde.

Umgang mit fehlenden Werten

Insgesamt gibt es 17 Erhebungen zu den im Unterricht verfolgten Lernzielen (*t1* bis *t17*), die in die Analysen eingegangen sind. Wie zu erwarten, sind zum Teil fehlende Werte in diesem Datensatz vorhanden. Zwei Aspekte kamen im Umgang mit fehlenden Werten zum Tragen: (1) ein Fallausschluss und (2) ein Imputationsverfahren.

Um die Stichprobe nicht zu stark zu reduzieren, fand zunächst ein sogenanntes 52.94-%-Kriterium Anwendung. Das 52.94-%-Kriterium bedeutet, dass von Lehrpersonen mindestens neun von 17 Fragebögen ausgefüllt vorliegen mussten, damit diese nicht ausgeschlossen wurden. Das Kriterium erfüllten 42 der insgesamt 48 Lehrpersonen dieser Stichprobe. Wäre beispielsweise das 100-%-Kriterium zur Anwendung gekommen, hätte sich die Stichprobe von 48 auf insgesamt 20 Lehrpersonen erheblich reduziert, sodass von einem derart strengen Kriterium abgesehen wurde.

Bei den Fällen, die das 52.94-%-Kriterium erfüllten, aber fehlende Werte aufwies, kam ein Imputationsverfahren zur Anwendung. Dazu wurden die fehlenden Werte an den entsprechenden Stellen ersetzt, indem auf Basis der gültigen Fälle eine Berechnung erfolgte, wie häufig die Lehrpersonen das jeweilige Lernziel zum jeweiligen Messzeitunkt im Mittel fokussierten. Analog dazu wurden Mittelwerte zur Gewichtung des jeweiligen Lernziels berechnet. Die Imputation von Mittelwerten erschien generell adäquater zu sein, als eine Null (niedrigster Wert → Unterschätzung) oder eine Eins (höchster Wert → Überschätzung) zu vergeben.

Vorgehensweise bei der Prädiktorbildung

Zunächst fand eine Berechnung statt, wie oft das jeweilige Lernziel im Unterricht im Zeitraum von *t1* zu *t17* verfolgt wurde (Unterricht nicht verfolgt: Codierung = 0, im Unterricht verfolgt: Codierung = 1 bis 5 Lektionen). Anschließend wurde der Wert durch die Anzahl der Lektionen innerhalb des genannten Zeitraums geteilt (32 Schulwochen x 5 Lektionen pro Woche = 160 Lektionen innerhalb des Zeitraumes). Analog erfolgte die Vorgehensweise zur Gewichtung der Lernziele. Hierfür wurden die Daten zur Gewichtung der jeweiligen Lernziele addiert (kein zentrales Lernziel in den letzten zwei Wochen: Codierung = 0, ein zentrales Lernziel in den letzten zwei Wochen: Codierung = 1) und anschließend durch die Anzahl an Erhebungen dividiert. In einem letzten Schritt wurde der Wert zur Quantität mit dem Wert zur Gewichtung des jeweiligen Lernziels multipliziert (siehe auch D'Agostino et al., 2007). Dadurch ergab sich

für jede Klasse ein Wert pro Lernziel. Die Lernziele zum Buchkapitel „Brüche“ wurden anschließend gebündelt, indem die Summe der Interaktionsterme gebildet wurde. Letzteres stellt den Prädiktor dar, der nachfolgend als die im Unterricht verfolgten Lernziele bezeichnet wird.

4.5 Analysen

Nachfolgend werden die statistischen Modelle zur Bearbeitung der empirischen Fragestellung beschrieben. In diesem Zusammenhang werden auch die Gedanken zur Konzeptualisierung aus Kapitel 3 nochmals aufgegriffen. Die Erläuterungen zu den Modellen und zur Konzeptualisierung dienen der Fundierung der Modellspezifikationen, sodass die in dieser Hinsicht getroffenen Entscheidungen nachvollziehbar werden. In Anbetracht der späteren Hypothesenprüfung wird zudem die Analysestrategie vorgestellt und es werden Informationen zum Schätzverfahren und den relevanten Kennwerten dargelegt. Darüber hinaus werden weitere relevante Analyseverfahren aufgeführt, die später der Beschreibung der Daten und der Prüfung spezifischer Voraussetzungen dienen.

4.5.1 Statistische Modelle

Zur Evaluation der Instruktionssensitivität von Testitems sollen zunächst zwei Modelle vorgestellt werden: (1) das *längsschnittliche Mehrebenen-Item-Response-Theory-Modell* (LMLIRT-Modell, Naumann et al., 2017) und (2) die Erweiterung des Modells als *erklärendes längsschnittliches Mehrebenen-Item-Response-Theory-Modell* (eLMLIRT-Modell, Naumann, Rieser et al., 2019). Diese beiden Modelle sind für die Analysen und zur Beantwortung der Fragestellung zentral.

4.5.1.1 LMLIRT-Modell

Das LMLIRT-Modell (Naumann et al., 2017) stellt eine Erweiterung des Rasch-Modells (Rasch, 1961) innerhalb des *generalized linear mixed-model frameworks* dar (Rijmen, Tuerlinckx, De Boeck & Kuppens, 2003; van den Noortgate, De Boeck & Meulders, 2003). Wie im Rasch-Modell hängt auch im LMLIRT-Modell die Lösungswahrscheinlichkeit von der Differenz von Fähigkeits- und Schwierigkeitsparametern ab. Die Erweiterung des Rasch-

Modells kommt in der Mehrebenenstruktur und in der längsschnittlichen Perspektive zum Ausdruck.

Die Mehrebenenstruktur ermöglicht die Berücksichtigung einer geschachtelten Datenstruktur mit Schülerinnen und Schülern innerhalb von Klassen, was für die Schätzung der Varianz zwischen den Lerngruppen zentral ist (Gruppen-Perspektive; Naumann et al., 2016). Die Annahme dahinter lautet, dass sich Schülerinnen und Schüler aufgrund ihrer Klassenzugehörigkeit und dem erlebten Unterricht in ihrem Antwortverhalten bei der Testbearbeitung unterscheiden (Robitzsch, 2009; Kapitel 2.4.1). Die Erweiterung des Modells in Form einer längsschnittlichen Perspektive erfolgt vor dem Hintergrund, dass Unterricht mit einem Lernzuwachs in Verbindung gebracht wird und sich dieser auf die Leistungen abbilden lässt (Zeitpunkte-Perspektive; Naumann et al., 2016). Die längsschnittliche Perspektive ermöglicht folglich die Schätzung der Veränderung in den Itemschwierigkeiten über die Zeit (Cox & Vagas, 1966; Kapitel 2.4.1). Mithilfe des LMLIRT-Modells können Schätzungen aus der Zeitpunkte- sowie der Gruppen-Perspektive bereitgestellt werden (Naumann et al., 2016). Eine besondere Stärke des Modells liegt aber in der Möglichkeit, beide Varianzquellen (Gruppe und Zeit) im Sinne einer Gruppen-x-Zeitpunkte-Perspektive (Interaktionsterm) zu berücksichtigen. Dieser Interaktionsterm ermöglicht Aussagen zur globalen und differenziellen Sensitivität von Testitems (ebd.).

Den Gedanken von Naumann et al. (2017) folgend kann mit dem LMLIRT-Modell die Wahrscheinlichkeit *P* einer korrekten Antwort X einer Person *i* in Klasse *c* auf ein Item *k* zum Zeitpunkt *t* anhand der nachfolgenden Gleichung geschätzt werden:

$$logit(P[X_{tcik} = 1]) = \sum_{u=1}^{T} q_{tu}(\theta_{tc} + \theta_{tci} - \beta_{tck}) \quad (1)$$

Der Wert q_{tu} stammt aus einer vordefinierten Matrix **Q** und dient dazu, den Beitrag der Parameter in den Klammern ein- und auszuschalten (ebd.). In der **Q**-Matrix steht der Wert eins dafür, dass der Parameter am gegebenen Zeitpunkt zur Schätzung der Antwortwahrscheinlichkeit beitragen soll. Der Wert null steht für den umgekehrten Fall. Um ein Veränderungsmodell zu schätzen, werden die Werte von **Q** so festgesetzt, dass die Parameter von früheren Messzeitpunkten zur Modellierung späterer Antworten beitragen.

Im Rahmen der vorliegenden Arbeit gehen Daten zu drei Zeitpunkten in die Analysen ein, sodass **Q** einer 3x3-Matrix mit den Dimensionen *t* und *u* entspricht. Die Zeilen *t* von **Q** stehen

für die Gewichtung der zeitpunktspezifischen Parameter, die Spalten *u* von **Q** stehen für die Zeitpunkte. Die erste Spalte steht folglich für den Messzeitpunkt *t1*, die zweite Spalte für den Messzeitpunkt *t3* und die dritte Spalte für den Messzeitpunkt *t4*. In der ersten Zeile werden die Werte eins, null und null aufgeführt (Gleichung 2). Damit wird deutlich, dass nur der Parameter bezogen auf den ersten Messzeitpunkt für die Schätzung der Antwortwahrscheinlichkeiten herangezogen wird (→ Schätzung von Werten zu *t1*). In der zweiten Zeile werden die Werte eins, eins und null angegeben. Entsprechend werden zusätzlich zu den Parametern des ersten Messzeitpunkts noch Parameter des dritten Messzeitpunkts für die Schätzung der Antwortmöglichkeiten herangezogen (→ Schätzung von Differenzwerten $t3-t1$). In der dritten Zeile werden die Werte eins, null und eins vergeben. Dementsprechend werden zusätzlich zu den Parametern des ersten Messzeitpunkts noch Parameter des vierten Messzeitpunkts für die Schätzung der Antwortmöglichkeiten einbezogen (→ Schätzung von Differenzwerten $t4-t1$).

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{2}$$

Deutlich wird an dieser Stelle, dass die Veränderungswerte immer mit Bezug zum Ausgangswert (*t1*, *t3 – t1*, *t4 – t1*) gebildet werden (siehe auch Steyer, Eid & Schwenkmezger, 1997). [15] Im Erkenntnisinteresse steht folglich, inwiefern sich die klassenspezifischen Itemschwierigkeiten von Testitems innerhalb eines Schuljahres verändern. Der Grund für die Entscheidung einer Modellierung mit Bezug zum Ausgangswert *t1* liegt zum einen in der unterrichtlichen Praxis, da Lehrpersonen eine gewisse Flexibilität in der zeitlichen Verortung spezifischer Unterrichtsthemen obliegt. Es ist dementsprechend möglich, dass Lehrpersonen das Thema „Brüche“ und „Bruchteile von Größen“ vorziehen, sodass unter Umständen eine Datenerhebung zu *t3* nicht zwingend einem Pretest entsprechen muss. Mit der Modellierung der Differenzwerte (*t4 – t1*) haben die im Unterricht verfolgten Lernziele (Kapitel 4.4.3) einen umfassenden Informationsgehalt, da Daten aus 17 Messzeitpunkten einfließen. Ein weiteres Argument für die Modellierung von Veränderungswerten mit Bezug auf den Ausgangswert stellt die Qualität der Lernmotivation dar, welche als Kovariate ins Modell mit aufgenommen wird. Bei den Analysen zur Evaluation der Instruktionssensitivität von Testitems soll die Qualität der Lernmotivation (Ausgangswert) kontrolliert werden. Die präferierte Modellierung

[15] Alternativ könnten die Differenzwerte auch mit dem jeweils benachbarten Messzeitpunkt (*t2–t1*, *t3–t2*, *t4–t3*) gebildet werden, was von einer inhaltlichen Perspektive aus betrachtet aber als weniger sinnvoll anzusehen ist.

hat den Vorteil, dass die Prädiktoren und Leistungsdaten einen gemeinsamen Bezugsrahmen aufweisen.[16]

Je nach Verortung in der **Q**-Matrix haben die Fähigkeits- und Schwierigkeitsparameter (siehe Gleichung 1: θ_{tc}, θ_{tci}, β_{tck}) folglich unterschiedliche Bedeutungen. Werden beispielsweise die Paramater gemäß den Angaben in der ersten Zeile geschätzt, entspricht θ_{tc} dem Ausgangswert der mittleren Fähigkeit einer Klasse *c* zum Zeitpunkt *t*=1, θ_{tci} dem Ausgangswert der individuellen Fähigkeit einer Schülerin bzw. eines Schülers *i* in Klasse *c* zum Zeitpunkt *t*=1 und β_{tck} dem Ausgangswert der Schwierigkeit eines Items *k* in einer Klasse *c* zum Ausgangswert *t*=1. Bezugnehmend auf die Parameter, die gemäß den Angaben in der zweiten Zeile geschätzt werden, entspricht θ_{tc} der Veränderung der Fähigkeit einer Klasse *c* vom Zeitpunkt *t*3 – *t*1, θ_{tci} entspricht der Abweichung der Fähigkeit einer Schülerin bzw. eines Schülers *i* von der mittleren Klassenfähigkeit *c* vom Zeitpunkt *t3 – t1* und β_{tck} der Veränderung der Schwierigkeit eines Items *k* in einer Klasse *c* vom Zeitpunkt *t3 – t1*. Analog verhält es sich mit der dritten Zeile, nur mit dem Unterschied, dass *t4* anstelle von *t3* steht.

In Abhängigkeit davon, wie das Modell restringiert wird, können mit dem LMLIRT-Modell sowohl absolute als auch relative Sensitivitätsmaße geschätzt werden (Kapitel 2.4.1). Die absolute Sensitivität gibt an, inwiefern das Item – unabhängig von der Sensitivität des gesamten Tests – in der Lage ist, sensitiv auf Unterrichtsmerkmale zu reagieren (Naumann et al., 2017). Die relative Sensitivität gibt an, inwiefern die Sensitivität eines Items von der Sensitivität des gesamten Tests abweicht (ebd.). In der vorliegenden Arbeit wird – wie auch in anderen Studien (Naumann, Rieser et al., 2019) – ausschließlich auf die absolute Sensitivität fokussiert. Dabei wird davon ausgegangen, dass die Parameter θ_{tc} und θ_{tci} multivariat normalverteilt sind (Gleichung 3 und 4):

$$\theta_{tc} \sim MN(\theta_t, \mathbf{\Lambda}) \quad (3)$$

$$\theta_{tci} \sim MN(0, \sum) \quad (4)$$

Die Mittelwerte θ_t der Verteilungen von θ_{tc} sind zeitpunktspezifisch, während die Mittelwerte von θ_{tci} auf den Wert null fixiert sind. Die Kovarianzmatrix von θ_{tc} wird als $\mathbf{\Lambda}$ und die Kovarianzmatrix von θ_{tci} wird als $\sum$ bezeichnet. Bezugnehmend auf β_{tck} wird ebenfalls

[16] Bei einer Modellierung von Veränderungswerten mit Bezug auf den benachbarten Messzeitpunkt wäre ein gemeinsamer Bezugsrahmen nicht gegeben.

angenommen, dass die klassen- und zeitpunktspezifischen Itemschwierigkeiten multivariat normalverteilt sind (Gleichung 5):

$$\beta_{tck} \sim MN(\beta_{tk}, \mathbf{\Phi}) \quad (5)$$

Dabei wird von zeitpunkt- und itemspezifischen Mittelwerten β_{tk} und einer Kovarianzmatrix $\mathbf{\Phi}$ ausgegangen. Die Ausgangs- und Veränderungswerte aller Items dürfen kovariieren. Das β_{tk} steht für die mittlere Veränderung der klassenspezifischen Schwierigkeit eines Items *k* über die Zeit (d.h. $t > 1$) und entspricht der globalen Sensitivität des Items. Die Diagonalelemente der Kovarianzmatrix $\mathbf{\Phi}$ stehen für die Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten und entsprechen der differenziellen Sensitivität ϕ_k^2 eines Items *k*.

Zur zusätzlichen Illustration ist in Abbildung 12 ein exemplarisches Modell zur Schätzung der absoluten Sensitivität abgebildet. Zu erkennen ist zunächst eine Mehrebenenstruktur, welche die Individual- und die Klassenebene umfasst. Auf der Individualebene sind die manifesten Itemantworten X_{tk} getrennt nach Messzeitpunkt *t1* und *t4* aufgeführt. Die Benennung der Messzeitpunkte gestaltet sich analog zum Studiendesign.

Mit Blick auf die Individualebene in Abbildung 12 wird ersichtlich, dass die Itemantworten von der Fähigkeit einer Person in einer Klasse θ_{1ci} bzw. von der Veränderung der Fähigkeit einer Person in einer Klasse $\Delta\theta_{4ci}$ beeinflusst werden. Auf der Klassenebene sind die latenten Antworten X_{tk} aufgeführt, welche die Indikatoren für die klassenspezifischen Itemschwierigkeiten β_{1c} bzw. für die Veränderung der klassenspezifischen Itemschwierigkeiten $\Delta\beta_{4c}$ darstellen. Ausgehend von den Itemantworten auf der Individualebene ermöglicht das Modell die Schätzung der Parameter auf Klassenebene. Hierzu zählen insbesondere die Schätzungen der $\Delta\beta_{4c}$, auf denen die Indikatoren der Instruktionssensitivität basieren (siehe fettgedruckter Kasten in Abbildung 12), also die mittlere Veränderung der klassenspezifischen Itemschwierigkeiten $M(\Delta\beta_{4c})$ über die Zeit (globale Sensitivität) sowie die Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten $Var(\Delta\beta_{4c})$ über die Zeit (differenzielle Sensitivität). Die klassenspezifischen Fähigkeiten θ_{1c}, die Veränderung der klassenspezifischen Fähigkeiten $\Delta\theta_{4c}$ und die dazugehörige Kovarianzmatrix wurden in der Abbildung 12 nicht dargestellt. Hintergrund ist, dass zur Schätzung der absoluten Sensitivität diese Parameter sowie die Kovarianzmatrix auf null restringiert werden und aus Gründen der besseren Übersichtlichkeit ausgespart wurden.

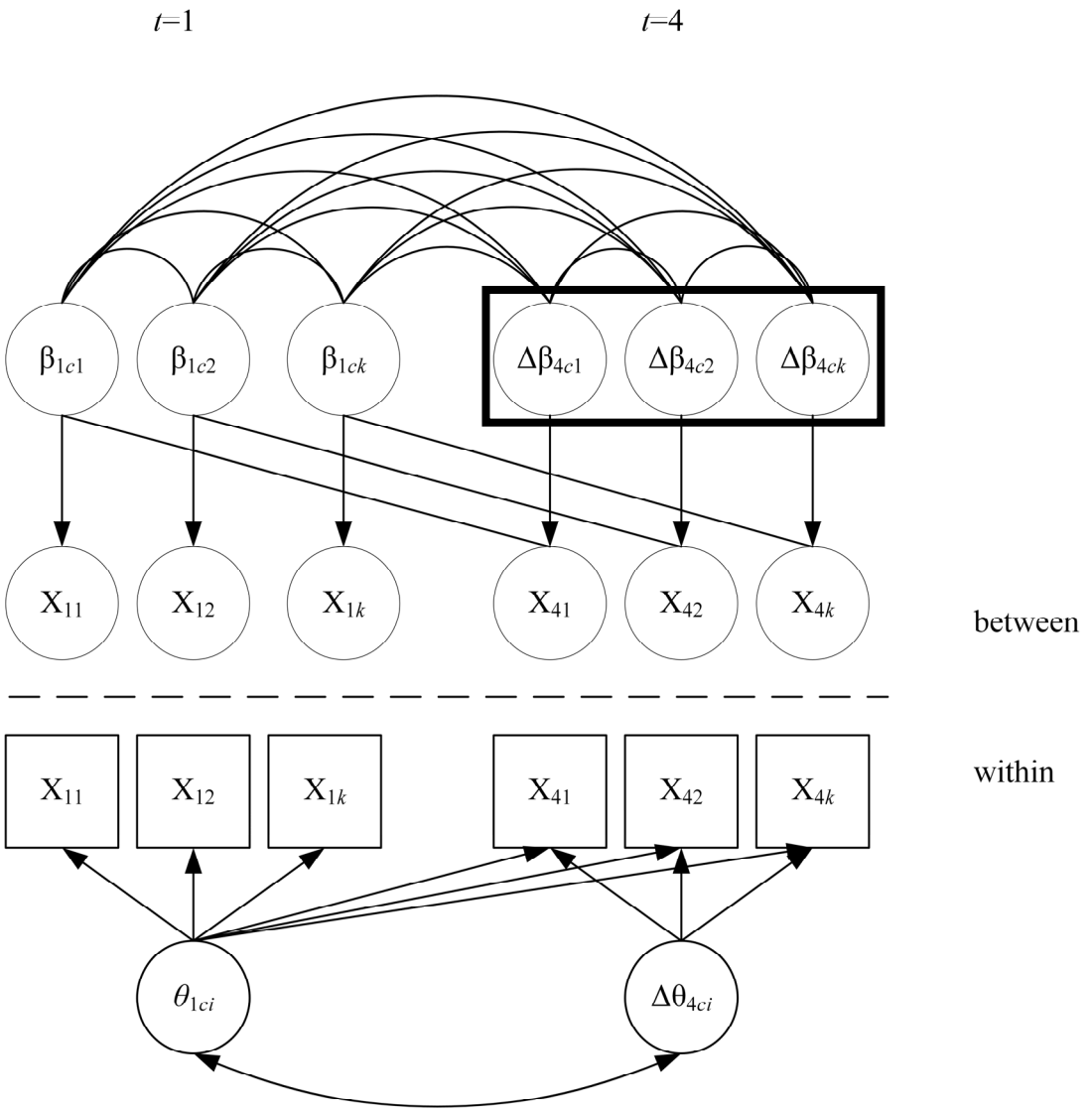


Abbildung 12. LMLIRT-Modell zur Schätzung der absoluten Sensitivität (in Anlehnung an Naumann, Rieser et al., 2019, S. 46).

4.5.1.2 eLMLIRT-Modell

Das eLMLIRT-Modell steht für ein *erklärendes längsschnittliches Mehrebenen-Item-Response-Theory-Modell*. Diese Bezeichnung verweist auf die Einbeziehung von einem oder mehreren Prädiktor/en. Im Kontext der Instruktionssensitivität kann damit geprüft werden, ob das jeweilige Item auf ein bestimmtes Unterrichtsmerkmal (z.B. auf einen kognitiv aktivierenden Unterricht) sensitiv reagiert. Darüber hinaus kann das eLMLIRT-Modell auch verwendet werden, um im Rahmen der Evaluation der Instruktionssensitivität von Testitems

bestimmte Einflussfaktoren zu kontrollieren. Prädiktoren und/oder Kovariaten können in diesem Zusammenhang leicht ins eLMLIRT-Modell integriert werden.

Bezugnehmend auf die empirische Fragestellung und den Gedanken zu den exemplarischen Analysen (Kapitel 3.7) sollen mithilfe des eLMLIRT-Modells spezifische Unterrichtsmerkmale und/oder die Qualität der Lernmotivation in die Analysen einbezogen werden. Bei welchen Variablen dies wie geschieht, geht aus Tabelle 5 hervor. Das jeweilige Unterrichtsmerkmal wird als erklärende Variable ins Modell aufgenommen, während die Qualität der Lernmotivation im Modell als Kovariate fungiert (Kapitel 3.7)

Tabelle 5

Leitfragen zur Konzeptualisierung der Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung lernrelevanter Merkmale der Schülerinnen und Schüler

Konzeptionelle Leitfragen zur Modellierung statistischer Modelle	**Entscheidungen hinsichtlich der exemplarischen Analysen**
Welche Unterrichtsmerkmale sind in die Analysen zur Evaluation der Instruktionssensitivität einzubeziehen?	o Unterrichtsqualität o Inhalt und Gewichtung der im Unterricht verfolgten Lernziele
Welche lernrelevanten Merkmale der Schülerinnen und Schüler sind bei der Evaluation der Instruktionssensitivität von Testitems zu berücksichtigen?	o Qualität der Lernmotivation
Auf welche Art von Effekt sollen die Testitems sensitiv reagieren?	o Effekt-Typ B (Kapitel 3.2) o Die Unterrichtsmerkmale werden als erklärende Variable, die Qualität der Lernmotivation als Kovariate einbezogen (Kapitel 3.7)

Einbeziehung einer kontrollierenden oder erklärenden Variable

Um eine kontrollierende oder erklärende Variable mit ins Modell aufzunehmen, kann die Gleichung 6 zur Verteilung der klassen- und zeitpunktspezifischen Itemschwierigkeiten des LMLIRT-Modells um einen Regressionsterm ergänzt werden. In diesem Fall wird von einem eLMLIRT-Modell gesprochen, dessen allgemeine Gleichung unter Berücksichtigung der Qualität der Lernmotivation auf Klassenebene Z_c bzw. eines Unterrichtsmerkmals Z_c wie folgt lautet:

$$\beta_{tck} \sim MN(\gamma_{0tk} + \gamma_{1tk} * Z_c, \mathbf{\Phi}) \quad (6)$$

Das β_{tck} steht dabei für die Schwierigkeit eines Items *k* zum Messzeitpunkt *t* in Klasse *c*. Das *Intercept* γ_{0tk} entspricht der Schwierigkeit eines Items *k* zu einem Messzeitpunkt *t*, wenn die

Variable Z_c den Wert null annimmt. Das Regressionsgewicht γ_{1tk} steht für den Einfluss einer Variable Z_c auf die klassen- und zeitpunktspezifische Itemschwierigkeit. Das **Φ** liefert im eLMLIRT-Modell Informationen zur Residualvarianz und unterscheidet sich damit vom **Φ** im LMLIRT-Modell, wo es der Gesamtvarianz entspricht. Wie beim LMLIRT-Modell besteht auch beim eLMLIRT-Modell die Annahme, dass β_{tck} multivariat normalverteilt ist. Die Modellrestriktionen erfolgen analog zum LMLIRT-Modell.

In Abbildung 13 ist ein eLMLIRT-Modell unter Berücksichtigung der Qualität der Lernmotivation und in Abbildung 14 ein eLMLIRT-Modell unter Berücksichtigung eines Unterrichtsmerkmals dargestellt. Soll bei der Schätzung der differenziellen Sensitivität die Qualität der Lernmotivation kontrolliert werden (Qualität der Lernmotivation als Kovariate), interessieren die ϕ^2-Werte (siehe Analysestrategie zum LMLIRT-Modell 1.1 und zum eLMLIRT-Modell 1.2 im nachfolgenden Kapitel 4.5.2). Die Darstellung der Modelle orientiert sich an Naumann, Rieser et al. (2019), sodass nur die Regressionsgewichte und nicht die Konstante (*Intercepts*) im Modell abgebildet werden.[17]

Soll die Veränderung der klassenspezifischen Itemschwierigkeiten durch eine Unterrichtsvariable erklärt werden (Unterrichtsmerkmal als Prädiktor), ist der Indikator der spezifischen Sensitivität von Interesse. In diesem Fall geht es um die Frage von Effekten, die in den Werten von γ_{1tk} zum Ausdruck kommen (siehe Analysestrategie zum eLMLIRT-Modell 2.1 und zum eLMLIRT-Modell 2.2 im nachfolgenden Kapitel 4.5.2). Im Modell der Abbildung 14 stehen die Pfeile ausgehend von der Qualität der Lernmotivation zu $\Delta\beta_{4ck}$, für γ_{1tk}.

Anhand der Abbildungen 13 und 14 wird zudem deutlich, dass sowohl die Qualität der Lernmotivation als auch die Unterrichtsmerkmale auf Klassenebene verortet sind. Was ist der Grund für diese Entscheidung? Aus einer inhaltlichen Perspektive kann zunächst festgehalten werden, dass es für die Verortung der Qualität der Lernmotivation sowohl für die Individual- als auch für die Klassenebene Argumente hinsichtlich der Modellspezifikation gibt (Individualebene: z.B. *Aptitude-Treatment-Interaktion*, Cronbach & Snow, 1977; auf der Klassenebene: z.B. spezifische Mediationsmodelle im Mehrebenenkontext, Preacher, Zyphur & Zhang, 2010; auf der Individual- und Klassenebene insbesondere Kompositionseffekte, Coleman et al., 1966). Bezogen auf die Qualität der Lernmotivation kann davon ausgegangen

[17] Um das Regressionsgewicht und die Konstante im Modell darzustellen, könnte sich die Darstellung an latenten Wachstumsmodellen orientieren, was aber in bisherigen Darstellungen aus der einschlägigen Literatur unüblich ist.

werden, dass diese nicht nur auf Individualebene, sondern auch auf Klassenebene unterschiedlich ausgeprägt ist. Wird die Qualität der Lernmotivation auf Klassenebene kontrolliert, können Aussagen getroffen werden, inwiefern die Items sensitiv reagieren, wenn die Schülerinnen und Schüler der jeweiligen Klassen gleichermaßen motiviert sind. Letzteres ist ein Aspekt, der im Rahmen dieser Arbeit besonders interessiert.

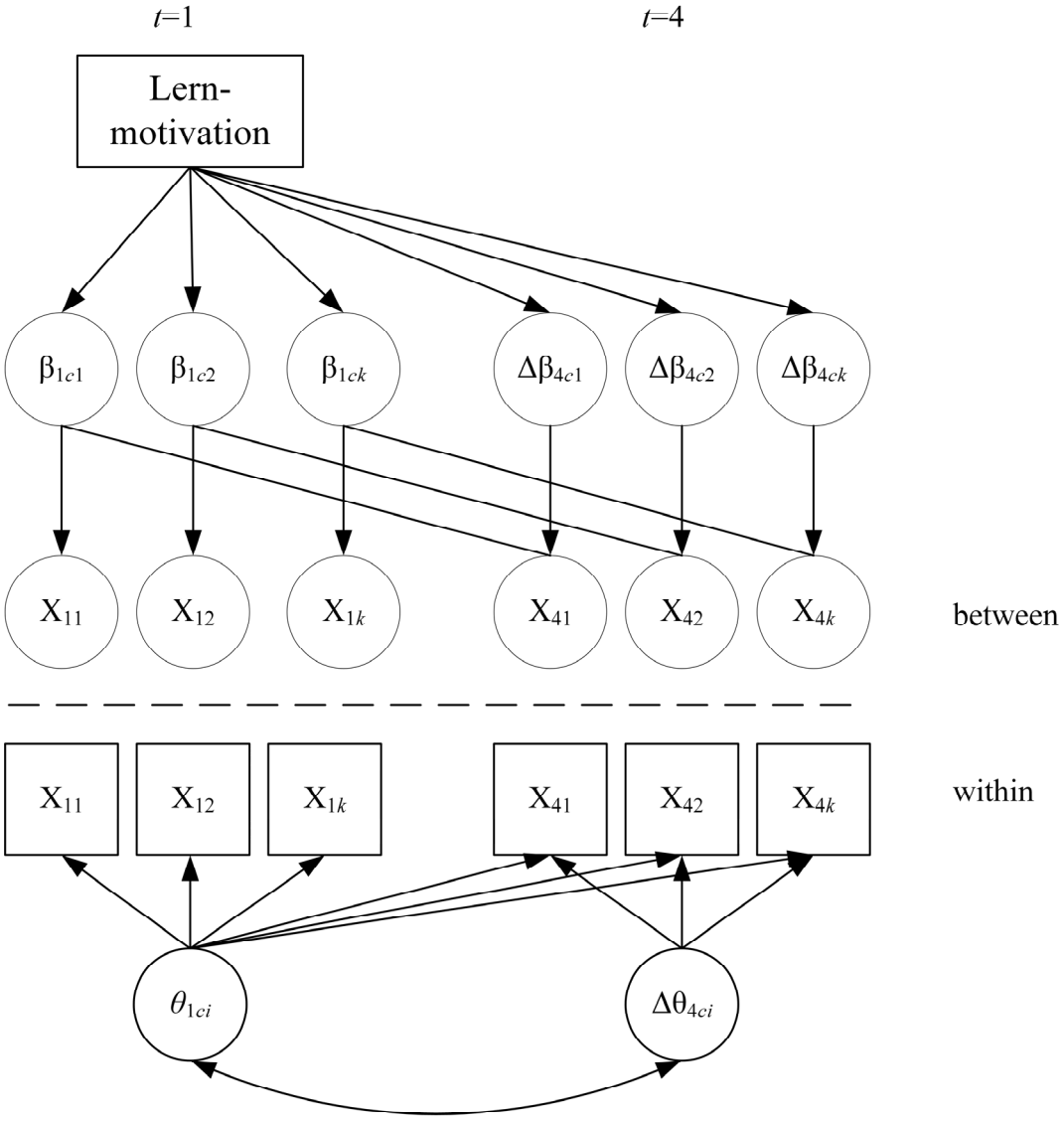


Abbildung 13. eLMLIRT-Modell unter Berücksichtigung der Qualität der Lernmotivation (in Anlehnung an Naumann, Rieser et al., 2019, S. 47).

Anmerkung: Die Ausgangs- und Veränderungswerte der klassenspezifischen Itemschwierigkeiten aller Items dürfen kovariieren. Aus Gründen der besseren Übersichtlichkeit werden diese nicht dargestellt.

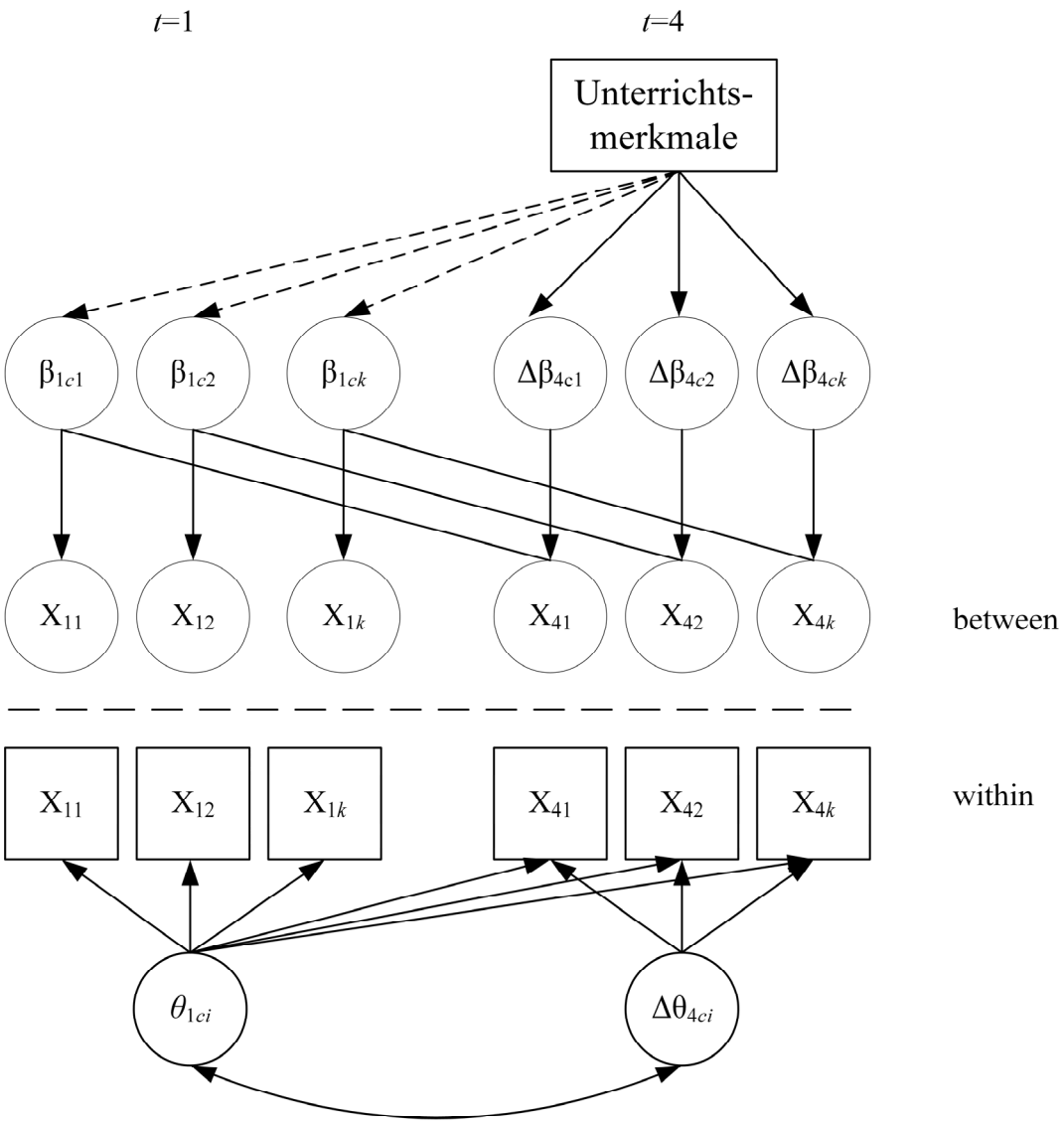


Abbildung 14. eLMLIRT-Modell unter Berücksichtigung eines Unterrichtsmerkmals (in Anlehnung an Naumann, Rieser et al., 2019, S. 47).

Anmerkung: Die Ausgangs- und Veränderungswerte der klassenspezifischen Itemschwierigkeiten aller Items dürfen kovariieren. Aus Gründen der besseren Übersichtlichkeit werden diese nicht dargestellt.

Einbeziehung einer kontrollierenden und einer erklärenden Variable

Soll die Veränderung der klassenspezifischen Itemschwierigkeiten durch ein Unterrichtsmerkmal und unter Kontrolle der Qualität der Lernmotivation erklärt werden, kann die Gleichung wie folgt erweitert werden:

$$\beta_{tck} \sim MN(\gamma_{0tk} + \gamma_{1tk} * Z_{1c} + \gamma_{2tk} * Z_{2c,} \mathbf{\Phi}) \quad (7)$$

Die Gleichung (7) ist weitgehend identisch mit der Gleichung (6). Lediglich die Anzahl der Variablen hat sich erhöht, was mit Z_{1c} und Z_{2c} zum Ausdruck kommt. Angenommen Z_{1c} steht für die zu kontrollierende Variable der Qualität der Lernmotivation und Z_{2c} für die erklärende Variable eines Unterrichtsmerkmals, so wären in diesem Fall insbesondere die Ergebnisse des Regressionsgewichts γ_{2tk} von Interesse, die für den Einfluss des Unterrichtsmerkmals Z_{2c} (erklärende Variable) auf die Veränderung der klassenspezifischen Itemschwierigkeiten unter Kontrolle der Qualität der Lernmotivation Z_{1c} (kontrollierende Variable) stehen. Das Modell ist in Abbildung 15 dargestellt.

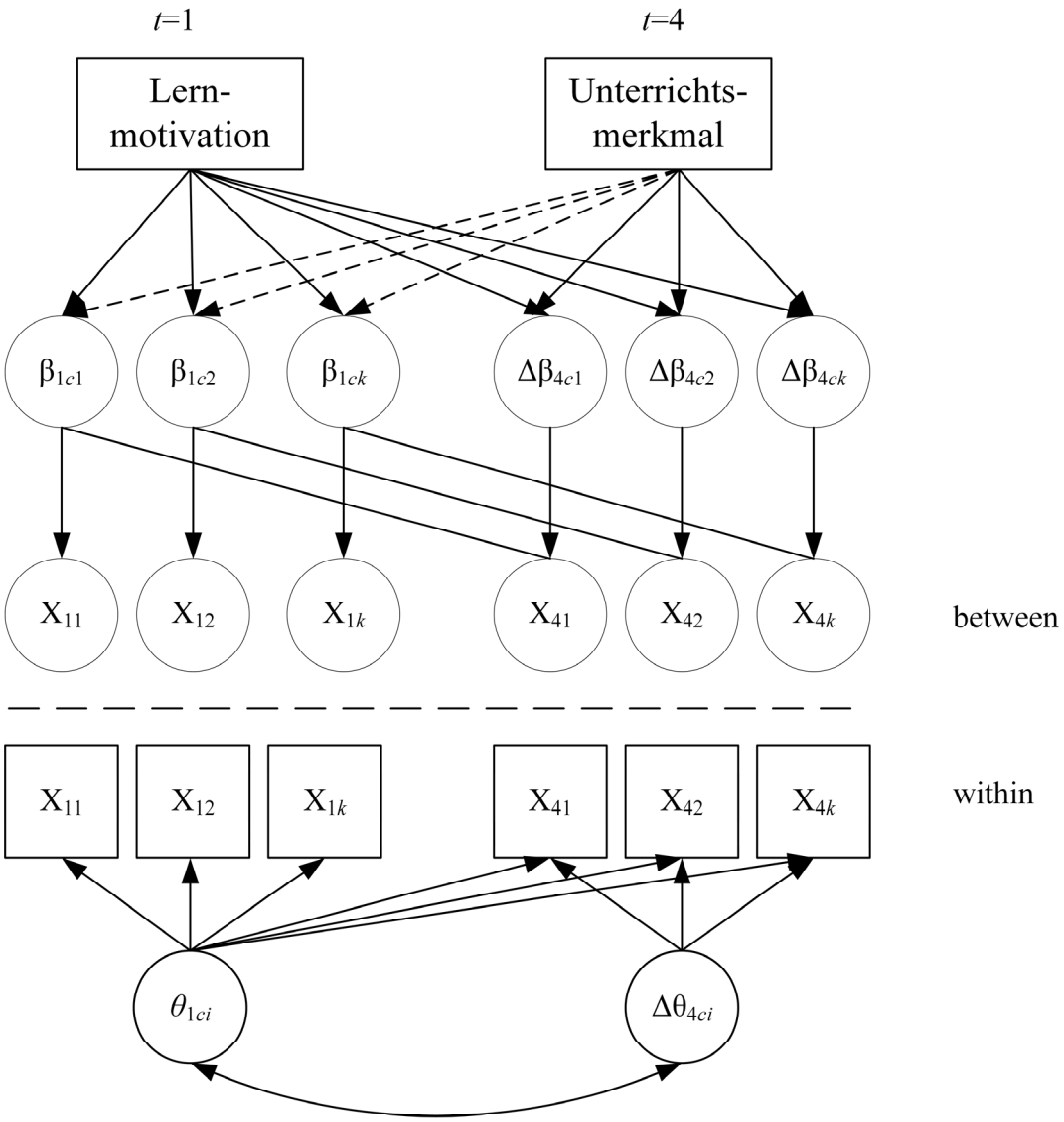


Abbildung 15. eLMLIRT-Modell unter Berücksichtigung der Qualität der Lernmotivation und eines Unterrichtsmerkmals (in Anlehnung an Naumann, Rieser et al., 2019, S. 47).

Anmerkung: Die Ausgangs- und Veränderungswerte der klassenspezifischen Itemschwierigkeiten aller Items dürfen kovariieren. Aus Gründen der besseren Übersichtlichkeit werden diese aber nicht dargestellt.

4.5.2 Analysestrategie

Basierend auf den Beschreibungen zum LMLIRT-Modell und zum eLMLIRT-Modell wurden vier Modellvarianten vorgestellt: das LMLIRT-Modell 1.1 ohne Einbeziehung jeglicher erklärender oder kontrollierender Variablen (Abbildung 12), das eLMLIRT-Modell 1.2 unter Einbeziehung der kontrollierenden Variable der Qualität der Lernmotivation (Abbildung 13), das eLMLIRT-Modell 2.1 unter Einbeziehung eines erklärenden Unterrichtsmerkmals (Abbildung 14) und das eLMLIRT-Modell 2.2 unter Hinzuziehung eines erklärenden Unterrichtsmerkmals und unter Kontrolle der Qualität der Lernmotivation (Abbildung 15). Die allgemeine Strategie besteht darin, die Schätzung der Indikatoren der Instruktionssensitivität von Testitems resultierend aus zwei unterschiedlichen Modellspezifikationen gegenüberzustellen (Abbildung 16). In einem ersten Schritt werden die Schätzungen der Indikatoren der Instruktionssensitivität resultierend aus dem LMLIRT-Modell 1.1 und dem eLMLIRT-Modell 1.2 verglichen. In einem zweiten Schritt geht es um einen Vergleich der Schätzungen resultierend aus dem eLMLIRT-Modell 2.1 und dem eLMLIRT-Modell 2.2. Der Vergleich basiert also immer auf einem statistischen Modell mit und ohne Kontrolle der Qualität der Lernmotivation der Schülerinnen und Schüler. Dabei soll überprüft werden, inwiefern sich die Schätzungen zur Instruktionssensitivität von Testitems zwischen den beiden Modellen jeweils unterscheiden.

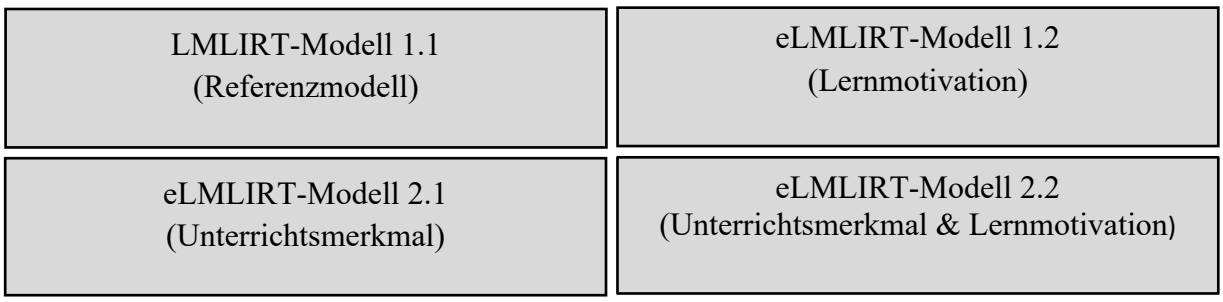


Abbildung 16. Spezifizierung der Modellvarianten (eigene Darstellung)**.**

Welche Indikatoren werden für welche Vergleiche hinsichtlich der Parameterschätzungen herangezogen?

Bezogen auf den Vergleich zwischen der Schätzung resultierend aus dem LMLIRT-Modell 1.1 und dem eLMLIRT-Modell 1.2 unter Kontrolle der Qualität der Lernmotivation wird die *differenzielle Sensitivität* in den Blick genommen. Analysen unter Berücksichtigung der *globalen Sensitivität* wären ebenso spannend gewesen; dies hätte aber eine andere Fragestellung

und ein anderes Forschungsdesign verlangt. Für den Vergleich zwischen der Schätzung resultierend aus dem eLMLIRT-Modell 2.1 zum Effekt des jeweiligen Unterrichtsmerkmals auf die Indikatoren der Instruktionssensitivität und der Schätzung resultierend aus dem eLMLIRT-Modell 2.2 zum Effekt des jeweiligen Unterrichtsmerkmals auf die Indikatoren der Instruktionssensitivität unter Kontrolle der Qualität der Lernmotivation wird die *spezifische Sensitivität* herangezogen. Darüber hinaus werden zu Beginn des Ergebnisteils Angaben zur Beschreibung der Daten gemacht. In diesem Zusammenhang werden zunächst Ergebnisse dargelegt, inwiefern die Items als *global* und/oder *differenziell sensitiv* eingestuft werden können. Die Ergebnisse zur globalen und differenziellen Sensitivität dienen dabei aber nicht der Beantwortung der Fragestellung, sondern um einen ersten allgemeinen Eindruck von den Daten zu erhalten (Tabelle 6).

Tabelle 6
Zusammenfassung der Analysestrategie

Analyse-strategie	**Indikator**	**Beschreibung des Indikators**	**Einbeziehung von Variablen**	**Kapitel**
Beschreibung der Daten	Globale Sensitivität	mittlere Veränderung der klassenspezifischen Itemschwierigkeiten	ohne Einbeziehung von erklärenden bzw. kontrollierenden Variablen	Kapitel 5.5.1
Beschreibung der Daten	Differenzielle Sensitivität	Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten	ohne Einbeziehung von erklärenden bzw. kontrollierenden Variablen	Kapitel 5.5.2
LMLIRT-Modell 1.1& eLMLIRT-Modell 1.2	Differenzielle Sensitivität	Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten	mit und ohne Einbeziehung einer kontrollierenden Variable	Kapitel 5.6
eLMLIRT-Modell 2.1 & eLMLIRT-Modell 2.2	Spezifische Sensitivität	Effekt einer Variable auf die Veränderung der klassenspezifischen Itemschwierigkeiten	Einbeziehung einer erklärenden Variable mit und ohne Einbeziehung einer kontrollierenden Variable	Kapitel 5.7

4.5.3 Berechnungen

Die Analysen unter Hinzuziehung des LMLIRT- bzw. des eLMLIRT-Modells sind in einem bayesianischen *Framework* zu verorten (Fox, 2010). Durchgeführt werden diese in R 3.5.1 (R

Core Team, 2017) unter Anwendung von JAGS 4.3.0 (Plummer, 2003, 2017a, 2017b) und dem Paket runjags (Denwood, 2016). Als Schätzmethode wird *Markov-Chain Monte Carlo* (MCMC) verwendet. Für jedes Modell werden vier *Markov-Chains* mit jeweils 67.500 Iterationen geschätzt. Die ersten 5.000 Iterationen werden als *Burn-in* verworfen. Um die Autokorrelation zu reduzieren, wird nur jede 25. Iteration in die Analysen einbezogen. Die Stichprobe umfasst damit (zunächst) 2.500 Fälle pro *Markov-Chain* ((67.500−5.000) / 25 = 2.500). Damit die Werte der Gelman-Rubin-Statistik über alle Analysen hinweg zufriedenstellend ausfallen, wird bei Bedarf auf weitere 1.000 Fälle pro *Markov-Chain* ausgeweitet. Sollten auch damit die Werte der Gelman-Rubin-Statistik immer noch nicht zufriedenstellend sein, wird nochmals um 500 Fälle pro *Markov-Chain* erhöht.

Für die Analysen werden nicht informative Prior-Verteilungen verwendet. Die Mittelwerte der Parameter zur Fähigkeit einer Person (hier die Schülerinnen und Schüler) werden auf null festgelegt. Die Varianz wird frei geschätzt unter Verwendung eines inversen *Wishart-Priors* mit einem Freiheitsgrad, der um eins höher als die Zahl der Messzeitpunkte ist (t+1). Zur Überprüfung, ob das Modell konvergiert ist und das Mixing zufriedenstellend ausfällt, empfiehlt sich die visuelle Betrachtung der *Traceplots* (Lynch, 2007). Darüber hinaus werden zur Überprüfung der Modellgüte die psrf-Werte (*potential scale reduction factor*) betrachtet, welche ≤ 1.05 sein sollten.

4.5.3.1 Überprüfung der globalen Sensitivität

Um die globale Sensitivität zu prüfen, werden die statistischen Maße zur mittleren Veränderung der klassenspezifischen Itemschwierigkeiten (β) herangezogen. Berichtet werden Modalwerte[18], die im bayesianischen Kontext dem wahrscheinlichsten Wert einer Schätzung entsprechen. Der Modalwert für die globale Sensitivität kann Werte ≥ und ≤ null annehmen. Anhand des 95-%-bayesianischen-Kredibilitätsintervalls (*BCI*) kann geprüft werden, ob sich die mittleren Itemschwierigkeiten über die Zeit (Indikator für globale Sensitivität) verändern. Umschließt das *BCI* nicht die Null, kann davon ausgegangen werden, dass der Wert mit einer Wahrscheinlichkeit von 95% verschieden von null ist. Letzteres spricht für die globale Sensitivität des Items. Im umgekehrten Fall (wenn die Null im *BCI* eingeschlossen ist) kann

[18] Die Begriffe Modalwert und Modus werden synonym verwendet.

nicht davon ausgegangen werden, dass mit einer Wahrscheinlichkeit von 95% der Wert von null verschieden ist. Das jeweilige Item würde demnach als global nicht sensitiv gelten.

4.5.3.2 Überprüfung der differenziellen Sensitivität

Die Überprüfung der differenziellen Sensitivität erfolgt in Anlehnung an Verhagen und Fox (2013) bzw. Verhagen, Levy, Millsap und Fox (2016). Verhagen und Fox (2013) schlagen einen Bayes-Faktor-Test basierend auf der Varianz von Itemparametern zwischen Gruppen vor, mit dem Hypothesen zur Messinvarianz beziehungsweise zur Nicht-Messinvarianz überprüft werden können (Verhagen et al., 2016). Die Berechnung des Bayes-Faktors erfolgt mittels *Savage-Dickey density ratio* (Dickey, 1971; Wagenmakers, Lodewyckx, Kuriyal & Grasman, 2010), mit der zwei Bayes-Faktoren (BF_{01}, BF_{10}) berechnet werden können. Bayes-Faktoren ermöglichen eine evidenzbasierte Überprüfung sowohl der Null- als auch der Alternativhypothese. Der Bayes-Faktor BF_{01} dient dabei der Überprüfung der Nullhypothese, der Bayes-Faktor BF_{10} hingegen der Überprüfung der Alternativhypothese. Ist der Bayes-Faktors $BF_{01} > 3$, spricht dies dafür, dass das Item differenziell nicht sensitiv ist (→ Annahme Nullhypothese, siehe Jeffreys, 1961; Verhagen et al., 2016). Ist der Bayes-Faktor $BF_{01} < .33$, spricht dies gegen die Annahme der Nullhypothese. Bezogen auf den zweiten Bayes-Faktor spricht ein Wert von $BF_{10} > 3$ dafür, dass das Item differenziell sensitiv ist (→ Annahme der Alternativhypothese; Jeffreys, 1961; Verhagen et al., 2016). Ist der Bayes-Faktor $BF_{10} < .33$ spricht dies gegen die Annahme der Alternativhypothese. Die Beziehung zwischen dem BF_{01} und dem BF_{10} wird anhand der Gleichung 9 aufgezeigt.

$$BF_{10} = \frac{1}{BF_{01}} \tag{9}$$

BF_{01} und BF_{10} sind demnach unmittelbar gekoppelt. Je kleiner BF_{01}, desto größer ist BF_{10}. Nachfolgend soll aufgezeigt werden, wie der Bayes-Faktor BF_{01} berechnet werden kann, um diesen zur Hypothesentestung zu nutzen. Basierend auf dem Satz von Bayes kann der Bayes-Faktor anhand der folgenden Gleichung kalkuliert werden:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)}{p(D|M_2)} * \frac{p(M_1)}{p(M_2)} \tag{10}$$

$p(M_1|D)/p(M_2|D)$ entspricht dabei der Posterior-Chance, die sich aus dem Produkt des Quotienten der (marginalen) *Likelihood* $p(D|M_1)/p(D|M_2)$ und dem Quotienten der Priors

beider Modelle $p(M_1)/p(M_2)$ ergibt. Sind im Modell keine „unbekannten Parameter" enthalten, entspricht der Bayes-Faktor dem „*Likelihood-Quotienten*" (Held, 2008). Gibt es unbekannte Parameter im Modell, so ist die marginale *Likelihood* vorab zu bestimmen (ebd.). $p(D|M_1)/p(D|M_2)$ entspricht folglich dem Bayes-Faktor und liefert Evidenz, ob ein Item als differenziell sensitiv einzuschätzen ist ($BF_{10} > 3$, $BF_{01} <. 33$) oder nicht ($BF_{10} < .33$, $BF_{01} > 3$).

4.5.3.3 Überprüfung der spezifischen Sensitivität

Die spezifische Sensitivität stellt einen weiteren Indikator dar, bei dem es um den Effekt (im Sinne eines Regressionskoeffizienten) einer erklärenden Variable (hier das jeweilige Unterrichtsmerkmal) auf die Veränderung der klassenspezifischen Itemschwierigkeiten geht. Für die Angabe der Regressionskoeffizienten γ_1 werden auch hier Modalwerte verwendet. Es gilt zu überprüfen, ob spezifische Unterrichtsmerkmale einen Effekt auf die Veränderungen in den klassenspezifischen Itemschwierigkeiten haben. Die Überprüfung erfolgt anhand des 95-%-*BCI*. Umschließt das *BCI* nicht die Null, ist davon auszugehen, dass der Wert verschieden von null ist. Letzteres spricht für die Annahme der spezifischen Sensitivität des Testitems.

4.5.4 Bezug zu den Hypothesen

Hypothese 1: Die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der differenziellen Sensitivität im Kontext der Evaluation der Instruktionssensitivität von Testitems.

Die Hypothese 1 wird bekräftigt, wenn sich die Schätzung der differenziellen Sensitivität mittels LMLIRT-Modell 1.1 im Vergleich zum eLMLIRT-Modell 1.2 unter Berücksichtigung der Qualität der Lernmotivation (Kapitel 4.5.2) voneinander unterscheidet.

Um die Hypothese zu überprüfen, erfolgt in einem ersten Schritt ein Vergleich zwischen den $\boldsymbol{\phi^2}$ -Werten desselben Items, das mittels LMLIRT- und eLMLIRT-Modell geschätzt wird. Differenzwerte werden basierend auf den $\boldsymbol{\phi^2}$ -Werten des jeweiligen Items gebildet ($\Delta\phi^2$-Werte). Je grösser die $\Delta\phi^2$-Werte des jeweiligen Items, desto stärker unterscheiden sich die geschätzten $\boldsymbol{\phi^2}$ -Werte beider Modelle. Ein Unterschied in der Schätzung der differenziellen Sensitivität gilt als wahrscheinlich, wenn die $\boldsymbol{\phi^2}$ -Werte des jeweiligen Items, resultierend aus der Schätzung mittels eLMLIRT-Modell 1.2, außerhalb des *BCI* desselben Testitems,

resultierend aus der Schätzung mittels LMLIRT-Modell 1.1, liegen. Ist Letzteres der Fall, kann davon ausgegangen werden, dass sich die ϕ^2 -Werte unterscheiden.

Um die Nullhypothese zu falsifizieren und die Alternativhypothese zu stützen, sollten sich bei mindestens einem von sechs Items die Schätzungen aus dem Modellvergleich (d.h. die ϕ^2 -Werte unter Berücksichtigung der *BCI*) unterscheiden. Anderenfalls ist die Nullhypothese beizubehalten. Darüber hinaus soll als Ergänzung – aber nicht zur Hypothesenprüfung – der Bayes-Faktor herangezogen werden. Er ermöglicht eine Einschätzung, ob das jeweilige Item – unter Berücksichtigung der unterschiedlichen Modellspezifikationen – als differenziell sensitiv betrachtet werden kann. Es ist vorstellbar, dass basierend auf den Parameterschätzungen je nach Modell unterschiedlich viele Testitems sich als differenziell sensitiv erweisen. Das Verfahren soll mögliche Konsequenzen einer anderen, aber durchaus üblichen Herangehensweise zur Evaluation der Instruktionssensitivität dienen.

Hypothese 2 bzw. Hypothese 2.1 bis 2.3: Die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der spezifischen Sensitivität im Kontext der Evaluation der Instruktionssensitivität von Testitems.

Die Hypothese 2.1 bis 2.3 wird bekräftigt, wenn die Schätzung der spezifischen Sensitivität mittels eLMLIRT-Modell 2.1 unter Einbeziehung des jeweiligen Unterrichtsmerkmals im Vergleich zum eLMLIRT-Modell 2.2 unter Einbeziehung des jeweiligen Unterrichtsmerkmals und der Qualität der Lernmotivation sich voneinander unterscheidet (Kapitel 4.5.2).

Um die Hypothesen zu überprüfen, erfolgt in einem ersten Schritt ein Vergleich zwischen den γ-Werten desselben Items, geschätzt mittels LMLIRT- und eLMLIRT-Modell. Differenzwerte werden basierend auf den γ-Werten des jeweiligen Items gebildet (Δγ-Werte). Je grösser die Δγ-Werte des jeweiligen Items, desto stärker unterscheiden sich die geschätzten γ-Werten beider Modelle. Ein Unterschied hinsichtlich der Schätzung der spezifischen Sensitivität gilt als wahrscheinlich, wenn die γ-Werte des jeweiligen Items, resultierend aus der Schätzung mittels eLMLIRT-Modell 2.2, außerhalb des *BCI* desselben Testitems, resultierend aus der Schätzung des eLMLIRT-Modells 2.1, liegen. Ist Letzteres der Fall, kann davon ausgegangen werden, dass sich die γ-Werte unterscheiden.

Um die Nullhypothese zu falsifizieren und die Alternativhypothese zu stützen, sollten sich bei mindestens einem von sechs Items die Schätzungen aus dem Modellvergleich (γ-Werte unter

Berücksichtigung der *BCI*) entsprechend unterscheiden. Anderenfalls ist die Nullhypothese beizubehalten.

4.5.5 Weitere Analyseverfahren zur Beschreibung der Daten und zur Prüfung von Voraussetzungen

Neben den bereits dargelegten bayesianischen Verfahren, die in dieser Arbeit zur Hypothesenprüfung herangezogen werden, finden zudem Verfahren des Frequentismus Anwendung. Die ausgewählten frequentistischen Verfahren werden sowohl zur Beschreibung der Daten als auch zur Überprüfung spezifischer Voraussetzung für das methodische Vorgehen herangezogen.

4.5.5.1 Konfirmatorische Faktorenanalyse

Die konfirmatorische Faktorenanalysen (KFA, im Englischen CFA) ist im Kontext von Strukturgleichungsmodellen zu verorten und zählt zu den strukturüberprüfenden Verfahren. Konkret dient die KFA der Überprüfung der Beziehung latenter Konstrukte und deren Indikatoren (*confirm* → bestätigen). Für die Anwendung einer KFA ist es wichtig, dass bereits Wissen über das latente Konstrukt (Faktor) und dessen Indikatoren (Testitems) vorhanden ist. Sollte dies nicht der Fall sein, wäre i.d.R. eine explorative Faktorenanalyse (EFA) vorzuziehen. Im Rahmen dieser Arbeit soll überprüft werden, ob sich die angenommene Struktur latenter Konstrukte und deren Indikatoren in den Daten widerspiegelt.

Neben der klassischen konfirmatorischen Faktorenanalyse kommt zudem ein Verfahren zur Anwendung, das ebenfalls der Überprüfung der Beziehung eines latenten Konstrukts und den Indikatoren dient und zudem der geclusterten Datenstruktur Rechnung trägt. Gemeint ist die konfirmatorische Mehrebenen-Faktorenanalyse (KMFA, Muthén, 1991, 1994). Diese ermöglicht eine Überprüfung der Passung theoretischer Modellannahmen und empirischer Daten unter Berücksichtigung der Individual- (Level 1) und der Gruppenebene (Level 2). Die KMFA ist insbesondere dann von Relevanz, wenn eine geclusterte bzw. hierarchische Datenstruktur zugrunde liegt. Ein Beispiel wäre das Konstrukt der Unterrichtsqualität im Sinne der drei Basisdimensionen, welches mithilfe von Einschätzungen der Schülerinnen und Schüler erhoben und dessen angenommene Struktur überprüft werden soll.

Die Berücksichtigung der geclusterten Datenstruktur begründet sich daraus, dass Einschätzungen von Schülerinnen und Schülern einer Klasse je nach interessierendem Merkmal ggf. ähnlicher ausfallen als Einschätzungen von Schülerinnen und Schülern unterschiedlicher Klassen. Die Daten sind folglich nicht voneinander unabhängig (Lüdtke, Trautwein, Schnyder & Niggli, 2007). Diesem Aspekt wird bei Anwendung der KMFA Rechnung getragen. Des Weiteren ist bei einer geclusterten Datenstruktur zu beachten, dass Unterschiede in den jeweiligen Einschätzungen der Schülerinnen und Schüler auf die Individual- und/oder auf die Klassenebene zurückzuführen sind. Diese Unterscheidung ermöglicht eine differenzierte Betrachtungsweise der Varianzquellen (Unterschiede zwischen den Individuen innerhalb der jeweiligen Klasse vs. Unterschiede zwischen den Klassen). Die Grundidee der KMFA soll anhand der Gleichung 11 dargelegt werden:

$$Y_T = Y_W + Y_B \tag{11}$$

Y_T steht dabei für die jeweils individuelle Einschätzung einer Schülerin bzw. eines Schülers, welche sich aus der Summe des Klassenmittelwerts Y_B und der individuellen Abweichung vom Klassenmittelwert Y_W ergibt (ebd.). Insofern wird bei der KMFA zwischen zwei Varianzquellen differenziert. Basierend auf den Kovarianzmatrizen $\sum$ werden Zusammenhänge innerhalb der Klassen $\sum_W$ und Zusammenhänge zwischen den Klassen $\sum_B$ untersucht (ebd.):

$$S_B = \sum\nolimits_W + c * \sum\nolimits_B \tag{12}$$

S_B entspricht dabei der Stichprobenkovarianzmatrix der Klassenebene, $\sum_W$ der Kovarianzmatrix auf Individualebene, $\sum_B$ der Kovarianzmatrix auf Klassenebene und c der Gruppengröße (Lüdtke et al., 2007; Muthén, 1991, 1994). Die Einbeziehung von c ist notwendig, um einer Verzerrung bei der Schätzung von S_B entgegenzuwirken (Lüdtke et al., 2007; Muthén, 1994). Sowohl die KFA als auch die KMFA werden mithilfe des R-Pakets „lavaan“ (Version 0.6–3; Rosseel et al., 2018; siehe auch Rosseel, 2012) angewendet.

4.5.5.2 Cronbachs Alpha

Des Weiteren wird die Reliabilität der eingesetzten Skalen mithilfe von Cronbachs Alpha überprüft. Der Bildung von Skalenmittelwerten pro Fall liegt die Bedingung zugrunde, dass für

mindestens zwei von drei Items ein gültiger Wert pro Fall existieren muss. Die Gleichung zur Berechnung von Cronbachs Alpha (Cronbach, 1951) lautet folgendermaßen:

$$\alpha = \frac{p}{p-1} \cdot \left[1 - \frac{\sum_{i=1}^{p} \sigma_{Y_i}^2}{\sigma_X^2}\right] \text{ mit } X = \sum_{i=1}^{p} Y_i \tag{13}$$

p steht dabei für die Anzahl an Items (bzw. Teilskalen) einer Skala, $\sigma_{Y_i}^2$ für die Varianz des Items i und σ_X^2 für die Varianz des beobachteten Testwerts. Cronbachs Alpha beschreibt also, inwiefern die einzelnen Items mit der Gesamtskala in Beziehung stehen, und wird daher auch als ein Maß für die interne Konsistenz bezeichnet.

Die Berechnungen von Cronbachs Alpha basieren auf den lavaan-Objekten (R-Paket „lavaan“, Version 0.6–3; Rosseel et al., 2018; siehe auch Rosseel, 2012), die im Rahmen der konfirmatorischen Faktorenanalyse gebildet werden. Für die Analysen wird das R-Paket „semTool“ (Version 0.5-1; Jorgensen, Pornprasertmanit, Schoemann & Rosseel, 2018) herangezogen.

4.5.5.3 Intraklassenkorrelation

Darüber hinaus werden mit den Daten der Schülerinnen- und Schüler-Fragebögen und den Leistungstestdaten Analysen zu drei Formen der Intraklassenkorrelation vorgenommen: die $ICC1_{FB}$, die $ICC2_{FB}$ und die ICC_{KC}. Das tiefgestellte $_{FB}$ steht dabei für Daten aus den Fragebögen, während das tiefgestellte $_{KC}$ auf Daten aus dem Leistungstest Klassencockpit verweist. Beginnend mit der $ICC1_{FB}$ kann hiermit eine Aussage getroffen werden, wie hoch der Anteil an Varianz in den Daten ist, der auf die Gruppen- bzw. in diesem Fall auf die Klassenzugehörigkeit zurückgeführt werden kann. Es ist davon auszugehen, dass sich der Unterricht zwischen den Klassen unterscheidet. Dies sollte sich u.a. in den Schülerinnen- und Schüler-Fragebogendaten zur Unterrichtsqualität widerspiegeln. Die Varianz zwischen den Klassen ist dabei eine wichtige Voraussetzung zur Evaluation der Instruktionssensitivität. Zur Berechnung der $ICC1_{FB}$ werden die Varianz zwischen den Klassen τ^2 und die Varianz innerhalb der Klassen σ^2 herangezogen. Die Gleichung zur Kalkulation der $ICC1_{FB}$ lautet wie folgt:

$$ICC1_{FB} = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{14}$$

τ^2 steht dabei für die Varianz zwischen den Klassen und σ^2 für die Varianz innerhalb der jeweiligen Klassen (Lüdtke et al., 2007; Muthén 1991; Snijders & Bosker, 2012). Die Summe aus τ^2 und σ^2 entspricht folglich der Gesamtvarianz in den Daten. Nimmt die $ICC1_{FB}$ einen Wert von null an, bedeutet dies, dass sich die Gruppenmittelwerte nicht voneinander unterscheiden. Beträgt die $ICC1_{FB}$ eins, so ist die Varianz in den Daten ausschließlich auf Unterschiede zwischen den Klassen zurückzuführen. Die Analysen zur $ICC1_{FB}$ erfolgten mithilfe des R-Pakets «lme4» (Version 1.1–27.1; Bates, Mächler, Bolker & Walker, 2015).

Basierend auf dem $ICC1_{FB}$-Wert und der Anzahl an Schülerinnen und Schüler (n) kann die $ICC2_{FB}$ berechnet werden (Raykov & Marcoulides, 2006; Snijders & Bosker, 2012). Die $ICC2_{FB}$ gibt Auskunft über die Varianz in den Einschätzungen in Anhängigkeit zur Stichprobengröße und wird zur Überprüfung der Reliabilität eingesetzt. Raykov & Marcoulides (2006) beschreiben die $ICC2_{FB}$ als Indikator, mit dem gemessen wird, inwiefern die aggregierten Daten als repräsentativ gelten können. Die Gleichung lautet wie folgt:

$$ICC2_{FB} = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} = \frac{n \cdot ICC1}{1 + (n-1) \cdot ICC1} \tag{15}$$

Ein niedriger $ICC1_{FB}$-Wert und eine kleine Stichprobengröße führen zu einer niedrigen $ICC2_{FB}$, während ein hoher $ICC1_{FB}$-Wert und eine große Stichprobe in eine hohe $ICC2_{FB}$ resultieren. Bei einem moderaten $ICC1_{FB}$-Wert kann theoretisch eine zufriedenstellende $ICC2_{FB}$ erreicht werden, wenn die Stichprobe entsprechend groß ist (Marsh et al., 2012).

Eine weitere Form der Intraklassenkorrelation findet bei der Berechnung der $ICC1_{KC}$ für die Klassencockpit-Testitems Anwendung. Den Klassencockpitdaten liegt – im Unterschied zu den Fragebogendaten – eine binäre Datenstruktur zugrunde (Item korrekt gelöst = 1, Item nicht korrekt gelöst = 0). Die Berechnungen der $ICC1$ für Daten mit binärer Datenstruktur kann gemäß folgender Gleichung vorgenommen werden:

$$ICC1_{KC} = \frac{\tau^2}{\tau^2 + \frac{\pi^2}{3}} \tag{16}$$

Ein Vergleich zwischen den Gleichungen 15 und 16 macht deutlich, dass diese sehr ähnlich sind. In Gleichung 16 stehen τ^2 für die Varianz zwischen den Klassen und $\frac{\pi^2}{3}$ für die Varianz innerhalb der jeweiligen Klassen (Rodriguez & Elo, 2003; Wu, Crespi & Wong, 2012). Die

Analysen zur $ICC1_{KC}$ wurden ebenfalls mit dem R-Paket lme4 (Version 1.1–27.1; Bates et al., 2015) durchgeführt.

5 Ergebnisdarstellung

Im vorliegenden Kapitel werden die Ergebnisse der statistischen Analysen zur Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung der Qualität der Lernmotivation von Schülerinnen und Schülern dargestellt. Hierfür werden Ergebnisse zur Qualität der Lernmotivation in Kapitel 5.1, zur Unterrichtsqualität in Kapitel 5.2, zu den im Unterricht verfolgten Lernzielen in Kapitel 5.3 und zu den Klassencockpitaufgaben in Kapitel 5.4 berichtet. Die Ausführungen in Kapitel 5.1 bis einschließlich 5.4 dienen dazu, sich ein Bild von den Daten zu verschaffen und zentrale Voraussetzungen für das spätere methodische Vorgehen zu überprüfen. Im Anschluss daran werden die Resultate basierend auf bayesianischen Schätzungen zur Instruktionssensitivität von Testitems dargelegt. In Kapitel 5.5. werden zunächst Ergebnisse zur globalen und differenziellen Sensitivität berichtet, welche auf Parameterschätzungen aus dem Nullmodell resultieren. Anschließend folgen Auswertungen zu den Modellvergleichen, die der Hypothesenprüfung dienen. In Kapitel 5.6 wird ein Vergleich zwischen den Parameterschätzungen resultierend aus dem LMLIRT-Modell 1.1 (Nullmodell) und dem eLMLIRT-Modell 1.2 unter Hinzuziehung der Qualität der Lernmotivation vorgenommen, während in Kapitel 5.7 ein Vergleich zwischen den Parameterschätzungen resultierend aus dem eLMLIRT-Modell 2.1 unter Hinzuziehung von Unterrichtsmerkmalen und dem eLMLIRT-Modell 2.2 unter Hinzuziehung von Unterrichtsmerkmalen und der Qualität der Lernmotivation fokussiert wird.

5.1 Qualität der Lernmotivation als ein lernrelevantes Merkmal der Schülerinnen und Schüler

Die Qualität der Lernmotivation wird als exemplarisches Beispiel für ein lernrelevantes Merkmal von Schülerinnen und Schülern in die Analysen einbezogen. Das Konstrukt basiert gemäß den Annahmen in der Literatur auf mehreren Subskalen: Zu nennen sind die Amotivierheit (S1ma), die externale Motiviertheit (S1mx), die introjizierte Motiviertheit (S1mt), die identifizierte Motiviertheit (S1md), die interessierte Motiviertheit (S1mi) und die intrinsische Motiviertheit (S1mr) (Kapitel 3.4.2). Die Überprüfung der Dimensionalität des Konstrukts ist Ziel des Kapitels. Darüber hinaus sollen deskriptive Statistiken und Reliabilitätsmaße dargelegt werden. Eine Übersicht über die Items kann dem Anhang A entnommen werden.

Ergebnisse zur Lernmotivation

Inwiefern die genannten Subskalen das Konstrukt der Qualität der Lernmotivation abbilden, soll analysiert werden. Zur Überprüfung des Messmodells des Konstrukts zur Qualität der Lernmotivation werden konfirmatorische Faktorenanalysen mit einer ein-, drei- und sechsfaktoriellen Lösung durchgeführt. Inhaltlich lassen sich insbesondere die drei- und die sechsfaktorielle Lösung gut begründen und finden in unterschiedlicher Zusammenstellung der (Sub-)Skalen Anwendung (Bartholomew et al., 2018; Frey et al., 2009; Prenzel, Seidel & Drechsel, 2004; Ryan & Connell, 1989; Ryan & Deci, 2000; Seidel et al., 2005). Für die Beurteilung des Modellfits werden folgende Indikatoren herangezogen: χ^2, χ^2/df, p (χ^2), der *Comparative Fit Index* (*CFI*), der *Trucker Lewis Index* (*TLI*), *root mean square error of approximation* (*RMSEA*), deren 90-%-Konfidenzintervall ($CI_{90\%}$) und der *standardized root mean square residual* (*SRMR*). Gemäß Bollen (1989), Bollen und Long (1993) sowie Hoelter (1983) gibt es unterschiedliche Empfehlungen zum Quotienten aus χ^2/df, die für einen guten Modellfit sprechen. Sie reichen von $\chi^2/df < 2$ bis hin zu $\chi^2/df < 5$. Da das Kriterium < 5 (Wheaton, Muthén, Alvin und Summers, 1977) als zu wenig streng erachtet wird, wird der Empfehlung von Carmines und McIver (1981) gefolgt. Demnach sollte der Quotient aus χ^2/df zwischen 1.0 und 3.0 liegen. Darüber hinaus sprechen die Ergebnisse für einen adäquaten Modellfit, wenn der $RMSEA \leq .06$ und der $SRMR \leq .08$ ist und der *CFI* sowie der *TLI* nahe .95 sind (Hu & Bentler, 1999). Eine Gegenüberstellung der Modelle mit den jeweiligen Kennwerten soll zeigen, welches Modell basierend auf den empirischen Daten bekräftigt wird. Des Weiteren werden für das favorisierte Modell standardisierte Faktorladungen λ der einzelnen Items und die Korrelation zwischen den Faktoren berichtet.

Zunächst wird das Konstrukt der Qualität der Lernmotivation im Sinne einer einfaktoriellen Lösung spezifiziert und empirisch überprüft.[19] Das Modell mit der einfaktoriellen Lösung zeigt einen nicht akzeptablen Modellfit (Tabelle 7). Dort fallen der Quotient aus χ^2/df sowie der Wert des *RMSEA* und des *SRMR* zu hoch und die *CFI*- und *TLI*-Werte zu niedrig aus (Carmines & McIver, 1981; Hu & Bentler, 1999). Die standardisierten Faktorladungen der 17 Items erstrecken sich von $\lambda_{Range} = .19$ bis .77 und sind signifikant von null verschieden (Tabelle 8).

Eine dreifaktorielle Lösung wird von Bartholomew et al. (2018) vorgeschlagen. Für die Modellspezifikation werden die Items zur Amotiviertheit (erster Faktor), zur kontrollierten

[19] Items zur Amotiviertheit und zur externalen Motiviertheit wurden für die Analysen zur konfirmatorischen Faktorenanalyse recodiert, da die originalen Itemformulierungen negativ sind.

Motiviertheit (externale & introjizierte Motiviertheit, zweiter Faktor) und zur autonomen Motiviertheit (identifizierte und intrinsische Motiviertheit, dritter Faktor) zusammengefasst (siehe Bartholomew et al., 2018). Der Modellfit spricht für eine nicht zufriedenstellende Modellpassung (Tabelle 7). Auch mit der dreifaktoriellen Lösung fallen der Quotient aus χ^2/df sowie der Wert des *RMSEA* und des *SRMR* zu hoch und die *CFI*- und *TLI*-Werte zu niedrig aus (Carmines & McIver, 1981; Hu & Bentler, 1999). Die standardisierten Faktorladungen der 17 Items erstrecken sich von $\lambda_{Range} = .23$ bis .74 und sind signifikant von null verschieden (Tabelle 8).

Ein drittes Modell wird in Anlehnung an Ryan und Deci (2000) mit einer sechsfaktoriellen Lösung spezifiziert (Tabelle 7). Da der Quotient aus $\chi^2/df < 3.0$, der *RMSEA*-Wert < .06 und der *SRMR*-Wert < .08 sind und der *CFI* sowie der *TLI* ungefähr bei .95 liegen, kann der Modellfit als zufriedenstellend erachtet werden (Carmines & McIver, 1981; Hu & Bentler, 1999). Die standardisierten Faktorladungen der 17 Items erstrecken sich von $\lambda_{Range} = .34$ bis .81 und sind signifikant von null verschieden (Tabelle 8). Die jeweiligen latenten Faktoren korrelieren miteinander (Tabelle 9).

Tabelle 7
Modellvergleich mittels konfirmatorischer Faktorenanalyse bzgl. der Qualität der Lernmotivation

	$\chi 2$	*df*	χ^2/df	*p*	*CFI*	*TLI*	*RMSEA*	$CI_{90\%}$	*SRMR*
1-Faktor-M.	1186.32	119	9.97	<.001	.72	.68	.11	[.10, .11]	.09
3-Faktor-M.	767.58	87	8.82	<.001	.76	.71	.10	[.09, .11]	.08
6-Faktor-M.	290.41	104	2.79	<.001	.95	.94	.05	[.04, .05]	.04

Anmerkung: Die Analysen basieren auf *N* = 832, die verwendete Stichprobe fällt dabei geringer aus (siehe Angaben im Text).
1-Faktor-M.=einfaktorielles Modell, 3-Faktor-M.=dreifaktorielles Modell, 6-Faktor-M.=sechsfaktorielles Modell.

Tabelle 8
Standard. Faktorladungen der 17 Items auf das Konstrukt der Qualität der Lernmotivation

	1			2			3			4			5		6		
Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Faktor-ladungen	.70	.70	.39	.34	.66	.69	.63	.65	.64	.55	.64	.76	.65	.81	.80	.70	.72

Anmerkung: Aus Platzgründen wurden die Items fortlaufend durchnummeriert. Im Anhang A sind die Items in der Gesamtheit mit Variablennamen und Itemformulierungen aufgelistet.

Tabelle 9
Korrelationen der latenten Faktoren

	F1	*F2*	*F3*	*F4*	*F5*	*F6*
F1	1.0					
F2	.65***	1.0				
F3	.35***	.35***	1.0			
F4	.19***	.11*	.71***	1.0		
F5	.47***	.25***	.64***	.64***	1.0	
F6	.53***	.38***	.66***	.52***	.96***	1.0

Anmerkung: Die Analysen basieren auf N = 832, die verwendete Stichprobe fällt dabei geringer aus (siehe Angaben im Text); F1: Amotiviertheit, F2: externale Motiviertheit, F3: introjizierte Motiviertheit, F4: identifizierte Motiviertheit, F5: interessierte Motiviertheit, F6: intrinsische Motiviertheit.
$p < .05$ *, $p < .01$ **, $p < .001$ ***

Der Vergleich der Modellpassung bezogen auf eine ein-, drei- und sechsfaktoriellen Lösung spricht für das sechsfaktorielle Modell und wird gegenüber den anderen zwei Modellspezifikationen – wenn es um die Abbildung des Konstrukts zur Qualität der Lernmotivation geht – präferiert.

Subskalen

In einem nächsten Schritt sollen die Subskalen zur Qualität der Lernmotivation näher betrachtet werden. In Tabelle 10 werden die Mittelwerte, die Standardabweichungen, Cronbachs α, die $ICC1_{FB}$ und die $ICC2_{FB}$ zu den jeweiligen Subskalen aufgeführt (Kapitel 4.5.5). Bezugnehmend auf die Werte zu Cronbachs α fällt der Wert zur Skala der externalen Motiviertheit (S1mx) auf, der mit .55 gemäß Murphy & Davidshofer (1988, 2013) als nicht akzeptabel zu bewerten ist. Die Reliabilität der Skalen zur Amotiviertheit (S1ma), zur introjizierten Motiviertheit (S1mt), zur identifizierten Motiviertheit (S1md) und zur interessierten Motiviertheit (S1mi) können als niedrig bis moderat bewertet werden (für einen Überblick über die verschiedenen Empfehlungen zur Interpretation von Cronbachs α siehe Peterson, 1994). Anhand der $ICC1_{FB}$ wird deutlich, dass der Wert zur Skala der externalen Motiviertheit mit einem Wert von $ICC1_{FB}$ = .01 sehr gering ausfällt. Ein Wert von $ICC1_{FB}$ = .01 bedeutet, dass nur 1% der Varianz in den Daten auf die Klassenzugehörigkeit zurückzuführen ist. Werte von ≥ .05 gelten gemäß LeBreton und Senter (2008) als Anscheinsbeweis („*prima facie evidence*“) eines Gruppeneffekts und werden im vorliegenden Kontext als nennenswert betrachtet (S. 838).

Da die Subskala zur externalen Motiviertheit (S1mx) keine zufriedenstellenden Werte bezüglich des Cronbachs α, der $ICC1_{FB}$ und der $ICC2_{FB}$ aufweist, wird diese von den Analysen der vorliegenden Arbeit zur Instruktionssensitivität von Testitems unter Berücksichtigung der Qualität der Lernmotivation ausgeschlossen.

Tabelle 10
Interne Konsistenz und Intraklassenkorrelation zur Qualität der Lernmotivation

Item	*N*	*M*	*SD*	*Cronbachs α*	$ICC1_{FB}$	$ICC2_{FB}$
S1ma	787/832	1.76	0.71	.60	.05	.48
S1mx	785/832	1.66	0.85	.55	.01	.14
S1mt	778/832	3.06	0.91	.67	.08	.62
S1md	783/832	2.99	1.01	.68	.05	.45
S1mi	775/832	2.62	0.96	.69	.05	.48
S1mr	784/832	2.82	0.91	.79	.07	.58

Anmerkung zum Antwortformat: 1=nie, 2=manchmal, 3=fast immer, 4=immer.

5.2 Unterrichtsqualität

Die Ergebnisdarstellung zur Unterrichtsqualität erfolgt analog zur Ergebnisdarstellung der Qualität der Lernmotivation. Im Unterschied zu den Berechnungen zur Qualität der Lernmotivation werden hier konfirmatorische Mehrebenen-Faktorenanalysen angewendet. Dies hängt damit zusammen, dass die Daten zu den Unterrichtsmerkmalen klassenspezifisch sind, während die Qualität der Lernmotivation einem Individualmerkmal entspricht. Im Kapitel 5.2.1 werden die einzelnen Subskalen zur Unterrichtsqualität aufgeführt, verschiedene Messmodelle zum Konstrukt der Unterrichtsqualität überprüft und die deskriptiven Ergebnisse sowie die Reliabilitätsmaße der jeweiligen Subskalen berichtet. Eine Übersicht über die Items ist dem Anhang B zu entnehmen.

5.2.1 Ergebnisse zur Unterrichtsqualität

Konzipiert wird zunächst ein Messmodell, das die drei Basisdimensionen der Unterrichtsqualität umfasst: Unterrichtsstörungen als Subdimension der Klassenführung, das unterstützende Klassenklima und die kognitive Aktivierung [20] (Kapitel 3.5.2). Die drei

[20] Wohlwissend, dass diese Konstrukte nur eine Auswahl der Unterrichtsqualitätsskalen darstellen (Praetorius et al., 2018).

genannten Basisdimensionen werden in einem Messmodell zur Unterrichtsqualität integriert. Ausgangspunkt bildet die Annahme, dass sowohl der Individual- als auch der Klassenebene eine dreifaktorielle Lösung zugrunde liegt.

Das 3/3-Messmodell konvergiert zunächst nicht. Dies gelingt erst mit der schrittweisen Einführung von Restriktionen, indem negative Residualvarianzen auf null gesetzt werden. Die Kennwerte des geschätzten Modellfits weichen dabei deutlich von den Empfehlungen von Hu und Bentler (1999) ab (Tabelle 11; siehe auch Tabelle 12 und 13). Ersichtlich wird dies insbesondere am $SRMR_{between}$. Darüber hinaus gehen aus den Parameterschätzungen weitere Hinweise hervor, die auf einen Missfit hindeuten. Das Modell wird daraufhin auf verschiedene Art und Weise modifiziert (Ausschluss von Items mit Querladungen, Ausschluss von Items mit

Tabelle 11
Konfirmatorische Faktorenanalyse bzgl. der drei Unterrichtsqualitätsskalen

	$\chi 2$	*df*	$\chi 2/df$	*p*	*CFI*	*TLI*	*RMSEA*	$SRMR_{within}$	$SRMR_{between}$
Unterrichts-qualität	386.21	237	1.63	<.001	.95	.94	.03	.05	.21

Anmerkung: Die Analysen basieren auf N = 832, die verwendete Stichprobe fällt dabei geringer aus (siehe Angaben im Text).

Tabelle 12
Stand. Faktorladungen der Items zu den drei Unterrichtsqualitätsskalen

·	1					2						3					
Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Unterrichts-qualität-w	.58	.76	.79	.66	.68	.65	.76	.76	.61	.63	.48	.59	.38	.20	.47	.55	.37
Unterrichts-qualität-b	.93	.97	1.0	.84	.97	.75	1.0	1.0	.81	.80	1.00	1.0	.98	.54	.44	.25	.44

Anmerkung: Aus Platzgründen wurden die Items fortlaufend durchnummeriert. Im Anhang B sind die Items in der Gesamtheit mit Variablennamen und den Itemformulierungen aufgelistet.
Unterrichtsqualität-w=$\text{Unterrichtsqualität}_{within}$, Unterrichtsqualität-b= $\text{Unterrichtsqualität}_{between}$.

Tabelle 13
Korrelationen der latenten Faktoren

	$F1_{within}$	$F2_{within}$	$F3_{within}$		$F1_{between}$	$F2_{between}$	$F3_{between}$
$F1_{within}$	1.0			$F1_{between}$	1.0		
$F2_{within}$	.38***	1.0		$F2_{between}$	.15	1.0	
$F3_{within}$	.32***	.82***	1.0	$F3_{between}$	−.02	.57**	1.0

Anmerkung: Die Analysen basieren auf N = 832, die verwendete Stichprobe fällt dabei geringer aus (siehe Angaben im Text); F1: Unterrichtsstörungen, F2: unterstützendes Klassenklima, F3: kognitive Aktivierung.
$p < .05$ *, $p < .01$ **, $p < .001$ ***

Faktorladungen < .30, Zulassung von Kovarianzen). Die Parameterschätzungen sprechen jedoch auch bei diesen Varianten für keine gute Passung zwischen den empirischen Daten und dem spezifizierten Messmodell. Aus diesem Grund werden in einem nächsten Schritt zwei Messmodelle konzipiert, in denen die Unterrichtsstörungen und die kognitive Aktivierung getrennt voneinander aufgegriffen werden (Kapitel 5.2.2 und 5.2.3). Von der Darlegung der Ergebnisse zum unterstützenden Klassenklimas wird abgesehen, da diese nur für die Abbildung des Gesamtkonstrukts der drei Basisdimensionen zur Unterrichtsqualität von Interesse gewesen wären.

5.2.2 Ergebnisse zu Unterrichtsstörungen

Zunächst soll das Konstrukt der Unterrichtsstörungen hinsichtlich des Messmodells geprüft werden. Die Anwendung der konfirmatorischen Mehrebenen-Faktorenanalyse unter Annahme einer einfaktoriellen Lösung auf Individual- und Gruppenebene zeigt einen zufriedenstellenden Modellfit (Tabelle 14). Das Item S4qus3r wird in diesem Zusammenhang auf null restringiert, da die Residualvarianz des Items negativ und nahe null ist. Insgesamt können die standardisierten Faktorladungen der fünf Items als akzeptabel erachtet werden. Die frei geschätzten standardisierten Faktorladungen erstrecken sich auf der Individualebene von λ_{Range} = .57 bis .78 und auf der Klassenebene von λ_{Range} = .85 bis 1.00[21] (Tabelle 15) und sind signifikant von null verschieden.

In Tabelle 16 werden zudem der Mittelwert, die Standardabweichung, das Cronbachs α, die $ICC1_{FB}$ und die $ICC2_{FB}$ für die Skala zu Unterrichtsstörungen (S4qus) dargelegt. Anhand der Werte wird deutlich, dass die Schülerinnen und Schüler hinsichtlich eines störungsfreien Unterrichts im Mittel die dritte Kategorie „stimme eher zu“ wählen. Die interne Konsistenz fällt mit einem Wert von Cronbachs α = .85 zufriedenstellend aus. Anhand des $ICC1_{FB}$-Werts wird deutlich, dass dieser mit einer $ICC1_{FB}$ = .21 für das Vorhandensein von Varianz zwischen den Klassen spricht. Der $ICC2_{FB}$-Wert zur Reliabilität des aggregierten Urteils fällt mit $ICC2_{FB}$ = .82 ebenfalls befriedigend aus.

[21] Der Wert ist auf die Restriktion der Residualvarianz zurückzuführen.

Tabelle 14
Konfirmatorische Faktorenanalysen bzgl. der Unterrichtsqualitätsskalen

	χ2	*df*	*χ2/ df*	*p*	*CFI*	*TLI*	*RMSEA*	*CI90%*	*SRMR within*	*SRMR between*
Unterrichts-störungen	26.88	11	2.44	<.01	.99	.98	.04	[.02, .07]	.02	.05
Kognitve Aktivierung	22.16	18	1.23	.23	.99	.98	.02	[.00, .04]	.02	.20

Anmerkung: Die Analysen basieren auf *N* = 832, die verwendete Stichprobe fällt dabei geringer aus (siehe Angaben im Text).

Tabelle 15
Stand. Faktorladungen der Items der jeweiligen Unterrichtsqualitätsskalen

	1	2	3	4	5	6
Unterrichtsstörungen$_{within}$	.57	.76	.78	.66	.67	-
Unterrichtsstörungen$_{between}$	.95	.96	1.00	.85	.99	-
Kognitive Aktivierung$_{within}$	.61	.41	.19	.49	.50	.41
Kognitive Aktivierung$_{between}$	1.00	.96	.53	.29	.02	.41

Anmerkung: Aus Platzgründen wurden die Items in fortlaufend durchnummeriert. Im Anhang B sind die Items in der Gesamtheit mit Variablennamen und den Itemformulierungen aufgelistet.

Tabelle 16
Deskriptive Ergebnisse und Reliabilitätsmaße zu den Unterrichtsqualitätsskalen

	N	*M*	*SD*	*Cronbachs α*	*ICC1*	*ICC2*
Unterrichtsstörungen	761/832	2.63	.78	.85	.21	.82
Kognitive Aktivierung	763/832	2.92	.81	.61	.07	.57

Anmerkung zum Antwortformat: 1=stimme gar nicht zu, 2=stimme eher nicht zu, 3=stimme eher zu, 4=stimme zu.

5.2.3 Ergebnisse zur kognitiven Aktivierung

In einem weiteren Schritt wird das Messmodell zur Überprüfung des Konstrukts der kognitiven Aktivierung herangezogen. Angenommen wird eine einfaktorielle Lösung auf der Individual- und der Klassenebene. Die Anwendung der konfirmatorischen Mehrebenen-Faktorenanalyse hat zum Ergebnis, dass das Modell zunächst nicht konvergiert. Erst mit der Restriktion einer negativen Residualvarianz eines Items auf Klassenebene, welches nahe null ist (S4quk1r), wird das Modell vollständig geschätzt. Darüber hinaus wird als weitere Modifikation die Kovarianz zwischen den Items S4quk2r und S4quk3r auf Individualebene zugelassen. Trotz Modifikationsversuchen (Ausschluss von Items, Zulassung von Kovarianzen) weicht der Modellfit von den Empfehlungen von Hu und Bentler (1999) ab (Tabelle 14). Ersichtlich wird dies am $SRMR_{between}$-Wert, der deutlich zu hoch ist. Die frei geschätzten standardisierten

Faktorladungen der sechs Items erstrecken sich auf der Individualebene von λ_{Range} = .19 bis .61 und auf der Klassenebene von λ_{Range} = .02 bis 1.00[22] (Tabelle 15). Die Faktorladungen sind damit z.T. als nicht akzeptabel zu bewerten (Backhaus, Erichson, Plinke & Weiber, 2016) und auch nur z.T. signifikant von null verschieden. Insgesamt bleibt festzuhalten, dass die Parameterschätzungen für eine unzureichende Passung zwischen den empirischen Daten und den Modellannahmen sprechen. Die Skala zur kognitiven Aktivierung wird daher von den Analysen zur Instruktionssensitivität ausgeschlossen.

5.3 Unterrichtsinhalte

Im nachfolgenden Kapitel werden zunächst die Ergebnisse zu den einzelnen Items der im Unterricht verfolgten Lernziele beschrieben. Anschließend wird auf den gebildeten Prädiktor eingegangen (Kapitel 4.4.3), der in den Analysen zur Instruktionssensitivität von Testitems Berücksichtigung findet. Die Ausführungen zum Prädiktor beschränken sich dabei auf deskriptive Ergebnisse. Von einer Überprüfung des Messmodells und dem Berichten von Reliabilitätsmaßen wird abgesehen, da mit den Testitems nicht der Anspruch erhoben wird, dass diese eine Skala zur Abbildung eines Konstrukts darstellen. Eine Übersicht über die Items ist dem Anhang C zu entnehmen.

Ergebnisse zu den im Unterricht verfolgten inhaltsbezogenen Lernzielen

Die Darstellung der Ergebnisse zu den im Unterricht verfolgten Lernzielen erfolgt mit Bezug zum gebildeten Prädiktor. Die Prädiktorbildung wurde in Kapitel 4.4.3 ausführlich beschrieben. Hierfür werden Daten von 42 der 48 Lehrpersonen verwendet, welche in die nachfolgenden Analysen eingehen. In Tabelle 17 sind die Lernziele mit den dazugehörigen Variablenbezeichnungen aufgeführt, welche zur Bildung des Prädiktors herangezogen werden. Bezugnehmend auf die im Unterricht verfolgten Lernziele sollen zunächst Mittelwerte, Standardabweichungen, Minimum und Maximum dargelegt werden. Dabei ist zu beachten, dass vor der Datenanalyse ein Imputationsverfahren angewendet wurde (Kapitel 4.4.3). In der linken Tabellenhälfte werden die Ergebnisse zur Häufigkeit, in der rechten Tabellenhälfte die Ergebnisse zur Gewichtung der im Unterricht verfolgten Lernziele aufgezeigt (Tabelle 18). Anhand der Minimum-Werte zur Häufigkeit wird ersichtlich, dass die Angaben einzelner

[22] Der Wert ist auf die Restriktion der Residualvarianz zurückzuführen.

Lehrpersonen darauf hindeuten, dass spezifische Lernziele im Unterricht nicht verfolgt wurden. Gleichzeitig gibt es Lehrpersonen, die gemäß den Angaben das jeweilige Lernziel in bis zu 14, 18, 19 bzw. 23.64 [23] von insgesamt 160 Lektionen fokussieren. Im Mittel haben die Lehrpersonen die jeweiligen Lernziele in ein bis sieben Lektionen verfolgt.

Tabelle 17
Items zur Erhebung der im Unterricht verfolgten Lernziele

Lernzielformulierung aus dem St. Galler Bildungs- und Lehrplan	**Kürzel**
Dezimalzahlen und Brüche als Erweiterung des Bereichs der natürlichen Zahlen erfahren	Aza02
Einfache Brüche der Größe nach ordnen	Aza03
Einfluss von Zähler und Nenner auf den Wert des Bruchs erkennen	Aza04
Brüche kürzen und erweitern	Aop06
Einfluss von Veränderungen im Zähler und Nenner auf den Wert von Brüchen erkennen	Aop07

Tabelle 18
Deskriptive Ergebnisse zu den im Unterricht verfolgten Lernzielen

	Wie häufig wurde das jeweilige Lernziel im Unterricht verfolgt?				Wie häufig wurde das jeweilige Lernziel für den Unterricht als zentral erachtet?			
Item	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*
Aza02	0	23.64	6.93	6.65	0	6.03	1.41	1.38
Aza03	0	18	4.53	4.01	0	3	0.97	0.75
Aza04	0	18	4.31	3.69	0	3.03	1.08	0.95
Aop06	0	19	1.42	3.66	0	3	0.22	0.55
Aop07	0	14	1.64	3.36	0	2	0.24	0.46

Anmerkung: Die Dezimalzahlen bei den Maximum-Werten sind auf die Anwendung des Imputationsverfahrens zurückzuführen.

Der rechten Tabellenhälfte zur Gewichtung der Lernziele kann entnommen werden, dass es Lehrpersonen gibt, die von einer Angabe, ob das jeweilige Lernziel im Unterricht zentral war, abgesehen haben. Gleichzeitig gibt es Lehrpersonen, die gemäß den Angaben das jeweilige Lernziel in bis zu zwei, drei bzw. sechs von insgesamt 17 Beobachtungszeiträumen als zentral erachtet haben. Die Angaben zur Gewichtung beziehen sich – im Unterschied zu den Häufigkeiten der im Unterricht verfolgten Lernziele – auf den Beobachtungszeitraum der

[23] Die Dezimalzahlen (Max: 23.64, 6.03, 3.03) sind auf das Imputationsverfahren zurückzuführen.

Lehrpersonen und nicht auf die einzelnen Lektionen. In der Regel entspricht ein Beobachtungszeitraum zwei Wochen, in Ausnahmefällen – aufgrund von Ferienzeiten – auch einer Woche. Im Mittel haben die Lehrpersonen das jeweilige Lernziel in keinem bis einem Beobachtungszeitraum als zentral erachtet (von insgesamt 17 Beobachtungszeiträumen).

In Abbildung 17 werden zudem die kumulierten Häufigkeiten zu den Lernzielen berichtet. Zum einen soll aufgezeigt werden, wann und in wie vielen Klassen angegeben wurde, dass das jeweilige Lernziel zum Brüche-Kapitel erstmals verfolgt wurde (Abbildung 17, hellgraue Säulendiagramme). Zum anderen soll veranschaulicht werden, wann und in wie vielen Klassen dieses Lernziel erstmals als zentral galt (Abbildung 17, dunkelgraue Säulendiagramme).

Ein Vergleich zwischen den Abbildungen zur Häufigkeit (hellgraue Säulendiagramm) und zur Gewichtung (dunkelgraue Säulendiagramme) lässt Ähnlichkeiten erkennen. Beim Lernziel Aza03 und Aza04 wird deutlich, dass diese bereits zu Beginn des Schuljahres von einzelnen Lehrpersonen verfolgt, sie aber am Anfang des Schuljahres noch nicht als zentral erachtet wurden. Die kumulierten Häufigkeiten zu den im Unterricht verfolgten Lernzielen fallen etwas höher aus als die zu den zentralen Lernzielen. Im Mittel verfolgten $M = 26.4$ ($SD = 12.26$) Lehrpersonen die jeweiligen Lernziele, $M = 19.4$ ($SD = 11.37$) Lehrpersonen gaben zudem an, dass die jeweiligen Lernziele zentral waren.

Ein Vergleich zwischen den kumulierten Häufigkeiten ($n_{verfolgt}$, hellgraue Säulendiagramme) bzw. der kumulierten Gewichtungen ($n_{zentral}$, dunkelgraue Säulendiagramme) der im Unterricht verfolgten Lernziele zeigt außerdem, dass die Lernziele Aza02, Aza03 und Aza04 am Ende des Schuljahres von deutlich mehr Lehrpersonen verfolgt bzw. als zentral erachtet wurden als die Lernziele Aop06 und Aop07. So gaben beim Lernziel Aza02 insgesamt $n_{verfolgt} = 35$ bzw. $n_{zentral} = 27$ Lehrpersonen, beim Lernziel Aza03 $n_{verfolgt} = 35$ bzw. $n_{zentral} = 29$ Lehrpersonen und beim Lernziel Aza04 $n_{verfolgt} = 26$ bzw. $n_{zentral} = 27$ Lehrpersonen an, dass sie dieses Lernziel im Unterricht verfolgt haben bzw. es im Unterricht zentral war. Hinsichtlich der Lernziele Aop06 und Aop07 fallen die kumulierten Häufigkeiten deutlich geringer aus. Bezogen auf das Lernziel Aop06 sagten $n_{verfolgt} = 12$ bzw. $n_{zentral} = 6$ Lehrpersonen, dass sie dieses Lernziel im Unterricht verfolgt bzw. es als zentral erachtet haben. Beim Lernziel Aop07 waren es $n_{verfolgt} = 14$ bzw. $n_{zentral} = 8$ Lehrpersonen.

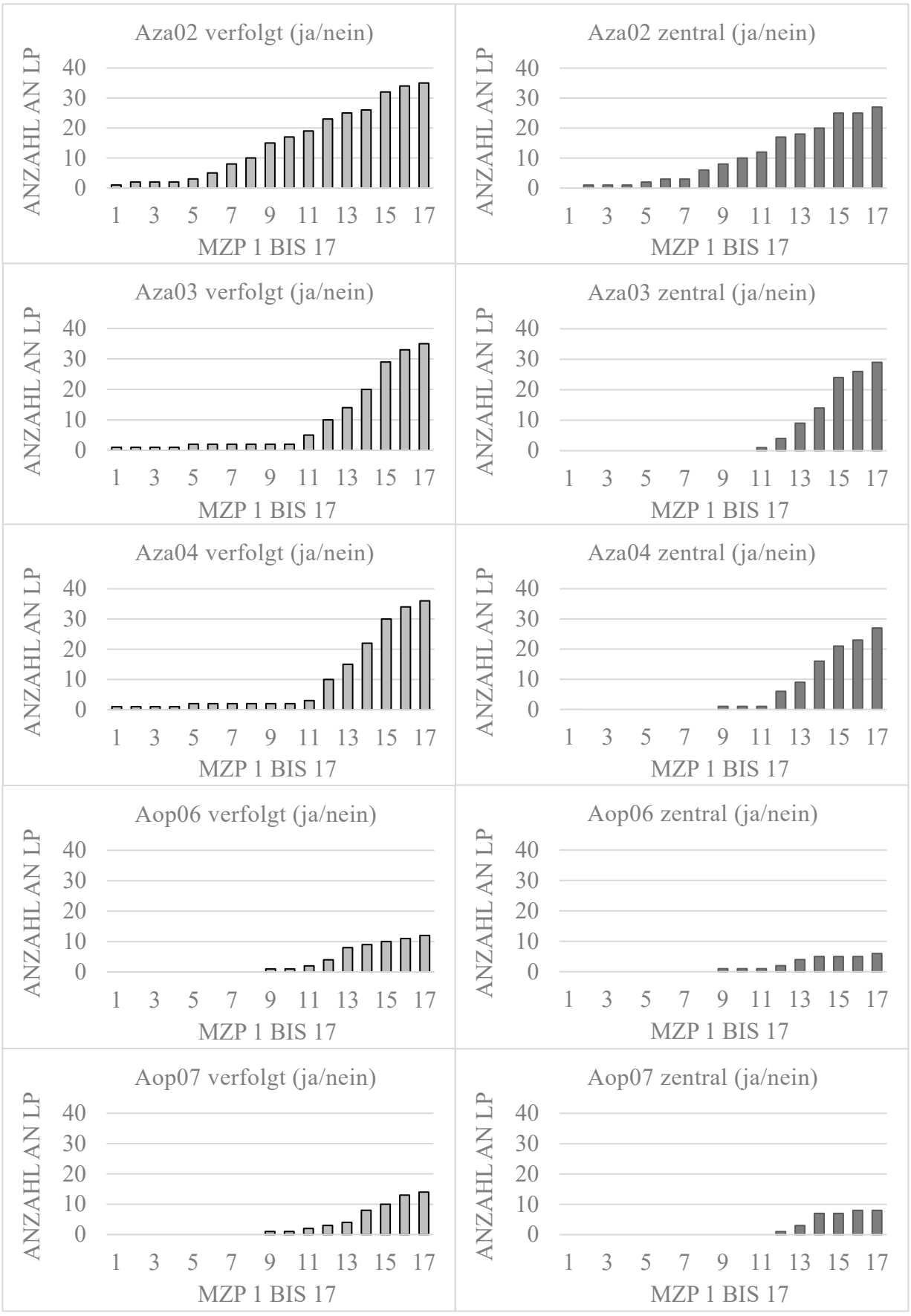


Abbildung 17. Kumulierte Häufigkeiten (hellgraue Säulendiagramme) und kumulierte Gewichtung (dunkelgraue Säulendiagramme) der im Unterricht verfolgten Lernziele.
Anmerkung: MZP: Messzeitpunkt; Anzahl an LP: Anzahl an Lehrpersonen.

Mit dem Wissen über die Ergebnisse zu den Häufigkeiten und zur Gewichtung der im Unterricht verfolgten Lernziele soll nachfolgend auf die deskriptiven Ergebnisse des Prädiktors eingegangen werden (Kapitel 4.4.3). Der Prädiktor wurde im Sinne eines Interaktionsterms zwischen der Häufigkeit und der Gewichtung der im Unterricht verfolgten Lernziele gebildet. Dafür wurden nicht nur die absoluten Werte, sondern auch die an der Gesamtzahl von 160 Lektionen bzw. an der Gesamtzahl von 17 Beobachtungszeiträumen relativierten Werte verwendet. In Tabelle 19 sind die relativen Häufigkeiten der im Unterricht verfolgten Lernziele, die relativen Häufigkeiten zur Gewichtung und der daraus resultierende Interaktionsterm aus Häufigkeit und Gewichtung abgebildet. Die Mittelwerte können als prozentualer Anteil zur Gesamtheit an Lektionen bzw. der Beobachtungszeiträume über alle Lehrpersonen hinweg (Mittelwert) gedeutet werden. Das Lernziel Aza02 wurde beispielsweise im Mittel in 4% der Lektionen (≈ 6 von insgesamt 160 Lektionen) innerhalb des Schuljahres verfolgt. In 8% der Beobachtungszeiträume (≈ 1 von insgesamt 17 Beobachtungszeiträumen) wurde das Lernziel Aza02 als zentral erachtet. Darüber hinaus wurden die Werte zur mittleren Häufigkeit mit den Werten zur mittleren Gewichtung multipliziert, welche in der Spalte zum Interaktionsterm vermerkt sind.

Tabelle 19
Relative Häufigkeiten zur Bildung des Interaktionsterms

	Häufigkeit (LZ verfolgt)		Gewichtung (LZ zentral)		Interaktion (LZ verfolgt * LZ zentral)	
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*
Aza02	.04	.04	.08	.08	.006	.009
Aza03	.03	.03	.06	.04	.002	.003
Aza04	.03	.02	.06	.06	.002	.003
Aop06	.01	.02	.01	.03	.001	.003
Aop07	.01	.02	.01	.03	.001	.002

In Abbildung 18 sind außerdem die individuellen Werte des Prädiktors der einzelnen Lehrpersonen aufgeführt. Die Werte ergeben sich aus der Summe der individuellen Interaktionsterme zu den fünf Lernzielen der jeweiligen Lehrperson. Diese lehrpersonenspezifischen Werte werden in die Analysen zur Instruktionssensitivität unter Berücksichtigung der im Unterricht verfolgten Lernziele einbezogen.

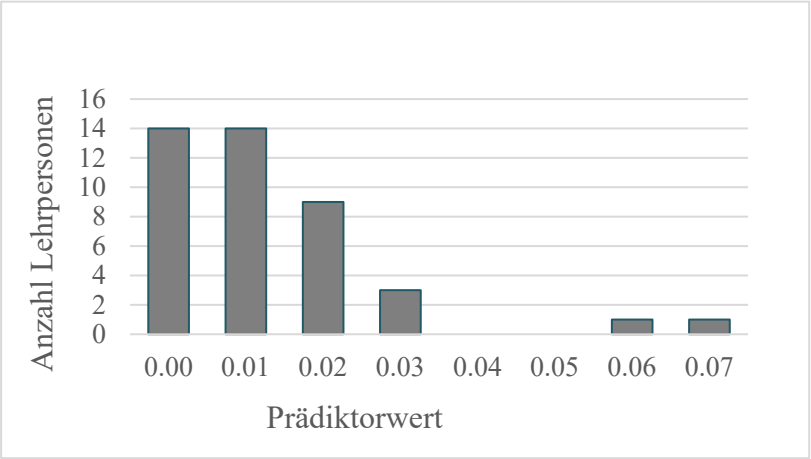


Abbildung 18. Prädiktorwert der jeweiligen Lehrpersonen.

5.4 Klassencockpit

Nachfolgend werden deskriptive Ergebnisse zu den sechs Testitems dargelegt (Itemanalyse, siehe Kelava & Moosbrugger, 2012). Berichtet werden zunächst die Itemschwierigkeiten zu den verschiedenen Messzeitpunkten. Diese geben insbesondere darüber Auskunft, inwiefern sich die Schwierigkeit der jeweiligen Items über die Zeit verändert. Des Weiteren werden Ergebnisse zur Intraklassenkorrelation präsentiert. Letzteres zeigt auf, inwiefern Varianz zwischen den Klassen vorhanden ist. Die Veränderung der Itemschwierigkeit und die Varianz zwischen den Klassen geben erste Anhaltspunkte bezogen auf die Instruktionssensitivität und sind daher besonders interessant. An dieser Stelle sei noch angemerkt, dass im Rahmen dieser Arbeit nur Items einbezogen wurden, die sich als Rasch-konform erwiesen haben.

Klassencockpit-Items

Die Analysen zu den Itemschwierigkeiten basieren auf Daten der Gesamtstichprobe zu drei Messzeitpunkten. Die Ergebnisse sind der Tabelle 20 zu entnehmen und werden entlang der Messzeitpunkte referiert. Bezugnehmend auf den Schwierigkeitsindex (*P,* siehe u.a. Kelava & Moosbrugger, 2012) zu den jeweiligen Messzeitpunkten (P_{t1}, P_{t2}, P_{t3}) wird deutlich, dass die Items über die Zeit leichter wurden. Um ein Beispiel zu nennen: Das Item A02 wurde zu *t1* (Pretest) von 19.3% der Schülerinnen und Schüler korrekt gelöst. Im Posttest (*t4*) waren es 71.5%. Die Itemschwierigkeit nahm also im Verlauf des Schuljahres im Mittel ab. Die Differenzwerte dienen in diesem Zusammenhang der Verdeutlichung, wie stark sich die Itemschwierigkeiten über die Zeit verändern. Insgesamt ist zu konstatieren, dass die

Veränderungswerte je nach Item sehr unterschiedlich ausfallen. Die geringste Veränderung bezogen auf die Itemschwierigkeit weist das Item A21, die größte Veränderung wird bei Item A02 ersichtlich (Tabelle 20).

Tabelle 20
Itemschwierigkeiten zu den jeweiligen Messzeitpunkten der sechs Testitems

Item	P_{t1}	P_{t3}	P_{t4}	Differenz *t3–t1*	Differenz *t4–t1*
A01	12.3	21.1	45.6	8.8	33.3
A02	19.2	32.6	71.5	13.4	52.3
A03	12.9	18.9	44.7	6.0	31.8
A05	22.8	32.6	51.8	9.8	29.0
A06	23.0	31.8	52.7	8.9	29.7
A21	21.9	28.7	28.9	6.8	7.0
P_{Mti} *(*P_{SDti}*)*	18.7 (4.9)	27.6 (6.11)	49.2 (13.9)		

Darüber hinaus werden in Tabelle 21 die Intraklassenkorrelationen ($ICC1_{KC}$) zu den sechs Testitems der jeweiligen Messzeitpunkte berichtet. Die Analyse zur $ICC1_{KC}$ ist von Bedeutung, da mit der Einbeziehung von Prädiktoren die Varianz auf der Klassenebene aufgeklärt werden soll. Gäbe es kaum Varianz auf der Klassenebene, würden die geplanten Analysen wenig sinnvoll erscheinen. Anhand der Werte in Tabelle 21 wird deutlich, dass Varianz in den Daten vorhanden ist ($ICC1_{KC} \geq .05$). Vereinzelt gibt es Items, die zu *t1* und/oder *t3* eine Varianz $ICC1_{KC} < .05$ aufweisen (insbesondere bei den Items A05, A06, A21 zu *t1* und *t2*). Wird die mittlere $ICC1_{KC}$ über alle sechs Items berechnet, zeigen sich zufriedenstellende $ICC1_{KC}$*-Werte* (Tabelle 21). Dementsprechend können zu *t1* im Mittel $ICC1_{KC\,Mt1} = 8\%$, zu *t3* im Mittel $ICC1_{KC\,Mt2} = 9\%$ und zu *t4* im Mittel $ICC1_{KC\,Mt4} = 13\%$ der Unterschiede auf die Klassenzugehörigkeit zurückgeführt werden.

Tabelle 21
Varianz zwischen den Klassen bezogen auf die Itemschwierigkeit der sechs Testitems

Item	$ICC_{KC\,t1}$	$ICC_{KC\,t3}$	$ICC_{KC\,t4}$
A01	.23	.15	.13
A02	.14	.31	.26
A03	.07	.04	.09
A05	< .01	.01	.13
A06	< .01	< .01	.09
A21	.02	.03	.07
$ICC_{KC\,Mti}$ *(*$ICC_{KC\,SDti}$*)*	.08 (.09)	.09 (.12)	.13 (.07)

5.5 LMLIRT-Modell

Bevor sich der Frage genähert wird, inwiefern die Qualität der Lernmotivation der Schülerinnen und Schüler die Schätzung der Instruktionssensitivität beeinflusst, soll vorab geklärt werden, ob sich die Items als global und differenziell sensitiv erweisen. Anschließend werden die Ergebnisse aus den LMLIRT- bzw. eLMLIRT-Modellen den Ergebnissen aus den eLMLIRT-Modellen unter Berücksichtigung der Qualität der Lernmotivation gegenübergestellt. Der Modellvergleich dient letztlich dem Erkenntnisgewinn, inwiefern die Qualität der Lernmotivation einen Einfluss auf die Indikatoren der Instruktionssensitivität von Testitems nimmt.

5.5.1 Globale Sensitivität

In den linken Spalten in Tabelle 22 sind die mittleren Ausgangswerte (*t1*) der klassenspezifischen Itemschwierigkeiten der Testitems angeführt. Was bedeuten nun die Werte zu den Itemschwierigkeiten (β), die in der Spalte zur Maximum-a-posteriori-Schätzung (MAP) angegeben sind? Allgemein lässt sich sagen: Je höher die Itemschwierigkeit zum ersten Messzeitpunkt (*t1*) ausfällt, desto schwieriger ist das Item im Mittel über alle Klassen (zu diesem Zeitpunkt). Umschließt das dazugehörige 95-%-Kredibilitätsintervall (95-%-*BCI*) nicht die Null, ist davon auszugehen, dass das Item mit einer Wahrscheinlichkeit von 95% schwerer ist als die mittlere Fähigkeit der Stichprobe. Letzteres würde sich im Antwortverhalten der Schülerinnen und Schüler widerspiegeln, indem die Mehrheit von ihnen nicht in der Lage wäre, das jeweilige Item zu *t1* korrekt zu lösen. Deutlich wird an dieser Stelle, dass die modellbasierten Itemschwierigkeiten β in enger Beziehung zu den in Kapitel 5.4 berichteten Itemschwierigkeiten *P* stehen. Anhand der Werte in Tabelle 22 ist ersichtlich, dass alle sechs Items für die Schülerinnen und Schüler zu *t1* verhältnismäßig schwer ausfallen (Range der Itemschwierigkeiten: β_{Range_t1} = 1.51 bis 2.50, mittlere Ausgangsschwierigkeit: β_{M_t1} = 1.89, Standardabweichung der mittleren Ausgangsschwierigkeiten: β_{SD_t1} = .44).

In den Spalten der rechten Tabellenhälfte werden die mittleren Veränderungswerte der klassenspezifischen Itemschwierigkeiten (*t4–t1*) aufgezeigt. Anhand der negativen Werte wird deutlich, dass die Items über die Zeit leichter werden. Das *BCI* gibt in diesem Zusammenhang an, ob die Items als *global sensitiv* betrachtet werden können. Das jeweilige Item gilt als global sensitiv, wenn das *BCI* die Null nicht umschließt. Die Ergebnisse sprechen dafür, dass die Items

in der Lage sind, einen Lernzuwachs innerhalb des Zeitraums von *t1* zu *t4* auf die Leistungen der Schülerinnen und Schüler abzubilden (Range der Veränderungswerte: $\beta_{Range\_t4\text{-}t1} = -0.47$ bis -3.09, mittlere Veränderung der klassenspezifischen Itemschwierigkeiten: $\beta_{M\_t4\text{-}t1} = -1.86$, Standardabweichung der Veränderung der klassenspezifischen Itemschwierigkeiten: $\beta_{SD\_t4\text{-}t1} = .86$). Die Ergebnisse zur Überprüfung der globalen Sensitivität sind in Tabelle 22 zusammengefasst und dienen einem späteren Vergleich mit den Ergebnissen zur differenziellen Sensitivität. Von einer Überprüfung des Einflusses der Qualität der Lernmotivation auf die Schätzung der globalen Sensitivität als Indikator der Instruktionssensitivität wird im Rahmen dieser Arbeit abgesehen (Kapitel 4.5.2).

Tabelle 22
Ausgangs- und Veränderungswerte der klassenspezifischen Itemschwierigkeiten resultierend aus dem LMLIRT-Modell

	β Ausgangswerte				β Veränderungswerte			
Item	*MAP* (*SD*) *t1*		95-%-*BCI* *t1*		*MAP* (*SD*) *t4–t1*		95-%-*BCI* *t4–t1*	
A01	2.50	(.17)	[2.19,	2.85]	−2.24	(.18)	[−2.59,	−1.89]
A02	1.85	(.15)	[1.55,	2.14]	−3.09	(.18)	[−3.44,	−2.73]
A03	2.35	(.15)	[2.06,	2.64]	−2.07	(.17)	[−2.41,	−1.75]
A05	1.52	(.12)	[1.29,	1.75]	−1.63	(.17)	[−1.95,	−1.30]
A06	1.51	(.12)	[1.29,	1.75]	−1.67	(.16)	[−1.99,	−1.35]
A21	1.60	(.12)	[1.37,	1.85]	−0.47	(.15)	[−0.77,	−0.18]

Anmerkung: β: Itemschwierigkeiten; *MAP*: Maximum-a-posteriori-Schätzung; *SD*: Standardabweichung des Mittelwerts der Posterior-Verteilung; 95-%-*BCI*: 95-%-Kredibilitätsintervall.

5.5.2 Differenzielle Sensitivität

Die Ergebnisse zur differenziellen Sensitivität sind in Tabelle 23 abgebildet. Dort werden die Schätzungen aus dem LMLIRT-Modell zur Varianz der Ausgangs- und Veränderungswerte der klassenspezifischen Itemschwierigkeiten (ϕ^2) aufgezeigt. Das Ausmaß der ϕ^2-Werte erlaubt eine Einschätzung, ob die Items in den verschiedenen Klassen zu *t1* unterschiedlich schwer sind. Eine hohe Varianz würde dafür sprechen, dass das Antwortverhalten die Schülerinnen und Schüler der jeweiligen Klasse variiert. Demnach müssten einige Klassen beim Lösen der jeweiligen Items erfolgreicher sein als andere. Die ϕ^2-Werte stehen somit in enger Beziehung zu den in Kapitel 5.4.1 berichteten $ICC1_{KC}$-Werten. Gemäß den Daten in Tabelle 23 sprechen die ϕ^2-Werte zunächst dafür, dass sich die Itemschwierigkeiten bezogen auf die Ausgangswerte

insbesondere bei den Items A01 ($\phi^2_{A01} = .44$) und A02 ($\phi^2_{A02} = .40$) zwischen den Klassen unterscheiden.[24] Bei den anderen vier Items fällt die Varianz zwischen den Klassen bezogen auf die Ausgangswerte geringer aus ($\phi^2_{A03} = .21$, $\phi^2_{A05} = .13$, $\phi^2_{A06} = .13$, $\phi^2_{A21} = .14$). An dieser Stelle ist jedoch anzumerken, dass es derzeit keine Konvention gibt, ab welchem ϕ^2-Wert von einer hohen oder niedrigen Varianz in den Daten gesprochen werden kann.

Tabelle 23

Varianz in den Ausgangs- und Veränderungswerten der klassenspezifischen Itemschwierigkeiten resultierend aus dem LMLIRT-Modell

	ϕ^2 *MAP (SD)*		ϕ^2 95-%-*BCI*		ϕ^2 *MAP (SD)*		ϕ^2 95-%-*BCI*			
Item	*t1*		*t1*		*t4–t1*		*t4–t1*		ϕ^2 BF_{01}	ϕ^2 BF_{10}
A01	.44	(.18)	[.20,	.86]	.28	(.16)	[.10,	.68]	0.25	4.02
A02	.40	(.16)	[.19,	.76]	.41	(.20)	[.16,	.91]	0.47	2.12
A03	.21	(.10)	[.09,	.46]	.18	(.10)	[.07,	.44]	0.52	1.94
A05	.13	(.06)	[.06,	.26]	.40	(.19)	[.17,	.86]	0.26	3.88
A06	.13	(.06)	[.06,	.28]	.37	(.18)	[.16,	.81]	0.12	8.34
A21	.14	(.06)	[.06,	.30]	.17	(.09)	[.07,	.41]	1.84	0.54

Anmerkung: ϕ^2: Varianz der klassenspezifischen Itemschwierigkeiten; *MAP*: Maximum-a-posteriori-Schätzung; *SD*: Standardabweichung des Mittelwerts der Posterior-Verteilung; 95-%-*BCI*: 95-%-Kredibilitätsintervall; *BF:* Bayes-Faktoren.

In der rechten Tabellenhälfte sind die Werte zur differenziellen Sensitivität dargelegt. Hierbei handelt es sich um die Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten (*t4−t1*). Vereinfacht gesagt, geht es also um das Maß, inwiefern sich die Klassen hinsichtlich des Leistungszuwachses bei einer Aufgabe unterscheiden. Je höher die jeweiligen ϕ^2-Werte ausfallen, desto höher ist die differenzielle Sensitivität. Zur Beurteilung, ob die Items differenziell sensitiv sind, werden die Bayes-Faktoren BF_{01} und BF_{10} herangezogen. Mit dem BF_{01} erfolgt die Überprüfung der Nullhypothese, also der Frage, ob das Item insensitiv ist. Der BF_{10} dient der Überprüfung der Alternativhypothese, also ob das jeweilige Item sensitiv ist. Anhand der Werte in Tabelle 23 wird ersichtlich, dass keines der Items einen $BF_{01} > 3.0$ hat. Dementsprechend liefern die Daten keine empirische Evidenz hinsichtlich der Annahme, dass die Items differenziell insensitiv sind. Bezogen auf den BF_{10}-Wert haben drei von sechs Items einen BF_{10}-Wert > 3.0. Die Daten liefern folglich empirische Evidenz, dass die drei Items A01, A05 und A06 differenziell sensitiv sind. Bei den Items A02, A03 und A21 bleibt die

[24] Hier geht es noch nicht um differentielle Sensitivität, da es sich um Ausgangswerte und nicht um Veränderungswerte handelt.

differenzielle Sensitivität unklar, da die Werte des BF_{01} und die Werte des BF_{10} weder für die Annahme der Nullhypothese noch für die Annahme der Alternativhypothese sprechen. Es gibt also basierend auf dem Konzept der Bayes-Faktoren bei den drei letztgenannten Items keine hinreichende Evidenz, um diese Items als sensitiv oder als insensitiv zu klassifizieren.

Abschließend sei noch angemerkt, dass die hier dargelegten Ergebnisse zur differenziellen Sensitivität resultierend aus den Analysen mittels LMLIRT-Modell später als Referenz für den Vergleich mit den Ergebnissen resultierend aus dem eLMLIRT-Modell unter Einbeziehung der Qualität der Lernmotivation dienen (Kapitel 5.6).

5.6 Vergleich der Parameterschätzungen resultierend aus dem LMLIRT- und dem eLMLIRT-Modell unter Einbeziehung der Qualität der Lernmotivation

Im vorherigen Kapitel wurden Informationen zur globalen und differenziellen Sensitivität basierend auf den Resultaten des LMLIRT-Modells (Nullmodell) dargelegt. Die Ergebnisse haben erste Hinweise geliefert, inwiefern sich die Items als instruktionssensitiv erweisen. Eine Antwort auf die Forschungsfrage liefern diese Ergebnisse jedoch noch nicht. Um zu überprüfen, inwiefern die Qualität der Lernmotivation von Schülerinnen und Schülern im Mathematikunterricht die Indikatoren der Instruktionssensitivität beeinflusst, werden nachfolgend verschiedene Modelle zur Evaluation der Instruktionssensitivität herangezogen. Begonnen wird mit einem Vergleich zwischen der Schätzung resultierend aus dem LMLIRT-Modell (Nullmodell) und der Schätzung resultierend aus dem eLMLIRT-Modell (erklärendes Modell) unter Einbeziehung der Qualität der Lernmotivation. Die Gegenüberstellung soll zeigen, ob sich die Schätzung der Instruktionssensitivität mit und ohne Hinzuziehung der Qualität der Lernmotivation voneinander unterscheidet (siehe Analysestrategie, Kapitel 4.5.2). Die Qualität der Lernmotivation wird hierbei als Kovariate ins Modell einbezogen. Sollten die Schätzungen der Instruktionssensitivität mit und ohne Hinzunahme der Qualität der Lernmotivation voneinander abweichen, kann dies als ein Hinweis für den Einfluss der Qualität der Lernmotivation auf die Indikatoren der Instruktionssensitivität gedeutet werden (Kapitel 3.7). Als Indikator zur Evaluation der Instruktionssensitivität von Testitems wird die differenzielle Sensitivität herangezogen (Kapitel 3.7, 4.5.2 und 5.5.2).

In Tabelle 24 ist die Schätzung resultierend aus dem LMLIRT-Modell zur Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten (*t4–t1*) und die Schätzung resultierend aus dem eLMLIRT-Modell zur Varianz der Veränderung der klassenspezifischen

Itemschwierigkeiten unter Kontrolle der Qualität der Lernmotivation (*t4–t1*) dargelegt. Während im LMLIRT-Modell die ϕ^2-Werte für die Gesamtvarianz stehen, beschreiben die ϕ^2-Werte aus dem eLMLIRT-Modell die Residualvarianz. Eine erste Annahme lautet, dass mit Hinzunahme der Qualität der Lernmotivation in das eLMLIRT-Modell Varianz aufgeklärt wird und sich im Zuge dessen die Residualvarianz (ϕ^2) im eLMLIRT-Modell reduziert. Ist Letzteres der Fall, sollten die ϕ^2-Werte im eLMLIRT-Modell im Vergleich zu den ϕ^2-Werten im LMLIRT-Modell geringer ausfallen. Darüber hinaus kann nicht ausgeschlossen werden, dass sich mit der Hinzunahme der Qualität der Lernmotivation die Residualvarianz auch vergrößert (Gelmann & Hill, 2007).

Tabelle 24

Vergleich der Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten resultierend aus dem LMLIRT-Modell und dem eLMLIRT-Modell unter Kontrolle der Qualität der Lernmotivation

	LMLIRT-Modell 1.1		eLMLIRT-Modell mit Lernmotivation 1.2		Δ LMLIRT-Modell 1.1 – eLMLIRT-Modell 1.2
	ϕ^2	ϕ^2	ϕ^2	ϕ^2	
	MAP (*SD*)	95-%-*BCI*	*MAP (SD)*	95-%-*BCI*	$\Delta\phi^2$
Item	*t4–t1*	*t4–t1*	*t4–t1*	*t4–t1*	*t4–t1*
A01	.28 (.16)	[.10, .68]	.24 (.16)	[.08, .63]	.04
A02	.41 (.20)	[.16, .91]	.32 (.19)	[.12, .78]	.09
A03	.18 (.10)	[.07, .44]	.18 (.10)	[.06, .42]	.00
A05	.40 (.19)	[.17, .86]	.41 (.21)	[.15, .90]	−.01
A06	.37 (.18)	[.16, .81]	.39 (.20)	[.15, .86]	−.02
A21	.17 (.09)	[.07, .41]	.16 (.09)	[.07, .40]	.01

Anmerkung: ϕ^2: Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten; *MAP*: Maximum-a-posteriori-Schätzung; *SD*: Standardabweichung des Mittelwerts der Posterior-Verteilung; 95-%-*BCI*: 95-%-Kredibilitätsintervall, $\Delta\phi^2$: Differenz der ϕ^2-Werte beider Modelle.

Um die Hypothese 1 zu überprüfen, werden zunächst die $\Delta\phi^2$-Werte, die ϕ^2-Werte und das *BCI* genauer betrachtet. Die Differenzwerte (Δ) basieren auf den ϕ^2-Werten der einzelnen Items, welche mithilfe des LMLIRT- und des eLMLIRT-Modells geschätzt werden ($\Delta\phi^2$). Sie sollen einen ersten Eindruck bezogen auf mögliche Unterschiede zwischen den Parameterschätzungen bei unterschiedlicher Modellspezifikation geben. Anhand der Werte in Tabelle 24 ist erkennbar, dass sich die ϕ^2-Werte resultierend aus den Parameterschätzungen beider Modelle kaum unterscheiden (*Range*$\Delta\phi^2$ = –.02 bis .09, *M*$\Delta\phi^2$ = .02). Dieser Eindruck, dass die ϕ^2-Werte

beider Modelle nahezu identisch sind, bestätigt sich auch bei der näheren Betrachtung der geschätzten ϕ^2-Werte der einzelnen Items mittels eLMLIRT-Modell und des geschätzten *BCI* des gleichen Items mittels LMLIRT-Modell. Würde der geschätzte Modalwert eines Items des eLMLIRT-Modells außerhalb des geschätzten *BCI* desselben Items des LMLIRT-Modells liegen, könnte mit ausreichender Sicherheit davon ausgegangen werden, dass sich die geschätzten Modalwerte beider Modelle voneinander unterscheiden. Letzteres ist jedoch nicht der Fall.

Die Hypothese 1, *die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der differenziellen Sensitivität im Kontext der Evaluation der Instruktionssensitivität von Testitems*, **ist daher abzulehnen**.

Das Kriterium, dass die ϕ^2-Werte resultierend aus dem eLMLIRT-Modell außerhalb des *BCI* des LMLIRT-Modells liegen (müssen), um die Hypothese zu bestätigen, ist nicht leicht zu erfüllen. Dies wird insbesondere bei der Betrachtung des Wertebereichs der ϕ^2-Werte und den *BCI* der hier vorliegenden Ergebnisse deutlich. Anhand der gebildeten $\Delta\phi^2$-Werte wird allerdings auch ersichtlich, dass die ϕ^2-Werte resultierend aus dem LMLIRT- und dem eLMLIRT-Modell sehr ähnlich ausfallen. Es ist also keinesfalls so, dass das Kriterium zur Annahme der Hypothese 1 nur knapp verfehlt wurde.

Die Überprüfung der Hypothese 1 mithilfe der ϕ^2-Werte und den *BCI* wurde in dieser Arbeit als die zu präferierende Strategie erachtet. Darüber hinaus wäre es auch möglich gewesen, die Hypothese 1 vor dem Hintergrund der Sensitivität der Testitems zu überprüfen. Im Unterschied zu den Analysen der ϕ^2-Werte unter Berücksichtigung des *BCI* geht es in diesem Fall um die Frage, ob die Ergebnisse aus den Analysen resultierend aus unterschiedlichen Modellspezifikationen zu voneinander abweichenden Schlussfolgerungen hinsichtlich der Sensitivität der Items führen. Vor diesem Hintergrund sollen ergänzend (nicht zur Überprüfung) zur Hypothese 1 Bayes-Faktoren herangezogen werden.

Bezugnehmend auf die Werte in Tabelle 25 wird ersichtlich, dass unabhängig davon, welche Modellspezifizierung den Analysen zugrunde liegt, keines der Items einen $BF_{01} > 3.0$ aufweist. Dementsprechend liefern die Daten keine hinreichende empirische Evidenz für die Annahme, dass die Items differenziell insensitiv sind. Fortfahrend mit dem BF_{10} haben im LMLIRT-Modell drei von sechs Items und im eLMLIRT-Modell keines der sechs Items einen BF_{10}-Wert > 3.0. Im LMLIRT-Modell können folglich drei Items (A01, A05 und A06) und im eLMLIRT-

Modell kann kein Item als differenziell sensitiv klassifiziert werden. Deutlich wird an dieser Stelle, dass je nach Modellspezifikation das Ergebnis zur Überprüfung der differenziellen Sensitivität mithilfe der Bayes-Faktoren bei drei von sechs Items unterschiedlich ausfällt.

Tabelle 25
Prüfung der Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten von t4 zu t1 mithilfe von Bayes-Faktoren

	LMLIRT		eLMLIRT	
Item	ϕ^2 BF_{01}	ϕ^2 BF_{10}	ϕ^2 BF_{01}	ϕ^2 BF_{10}
A01	0.25	4.02	1.66	0.60
A02	0.47	2.12	0.96	1.05
A03	0.52	1.94	0.82	2.21
A05	0.26	3.88	1.06	0.95
A06	0.12	8.34	0.35	2.86
A21	1.84	0.54	1.15	0.87

Anmerkung: ϕ^2: Varianz der klassenspezifischen Itemschwierigkeiten; *BF:* Bayes-Faktoren.

Abschließend soll noch angemerkt werden, dass das Kriterium, um ein Item als sensitiv bzw. insensitiv einzustufen, nur knapp erfüllt oder nur knapp verfehlt sein kann. Minimale Unterschiede in den Parameterschätzungen können folglich zu sehr unterschiedlichen Schlussfolgerungen führen. Dieser Aspekt stellte auch den Grund dar, weshalb die Hypothesenprüfung anhand der ϕ^2-Werte und der *BCI* und nicht anhand der Bayes-Faktoren erfolgte. Im Diskussionsteil in Kapitel 6.4.2 wird die Strategie der Hypothesenprüfung nochmals aufgegriffen.

5.7 Vergleich der Parameterschätzungen resultierend aus dem eLMLIRT-Modell unter Einbeziehung der Unterrichtsmerkmale und dem eLMLIRT-Modell unter Einbeziehung von Unterrichtsmerkmalen und der Qualität der Lernmotivation

Im Zusammenhang mit der Evaluation der Instruktionssensitivität von Testitems weist van der Linden (1981) auf die Bedeutung der Einbeziehung von Unterrichtsmerkmalen hin. So kann sichergestellt werden, dass die Varianz der Veränderung der klassenspezifischen Itemschwierigkeiten auch auf den Unterricht zurückzuführen ist (siehe auch Naumann, Rieser et al., 2019). In diesem Unterkapitel werden Ergebnisse zur Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung von Daten zu

Unterrichtsmerkmalen dargestellt. Hierfür wird die bereits aufgezeigte Strategie verfolgt, die Ergebnisse aus den Analysen mittels unterschiedlicher Modellspezifikationen miteinander zu vergleichen: Parameterschätzungen resultierend aus dem eLMLIRT-Modell unter Einbeziehung von Unterrichtsmerkmalen und den Parameterschätzungen resultierend aus dem eLMLIRT-Modell unter Einbeziehung von Unterrichtsmerkmalen und der Qualität der Lernmotivation der Schülerinnen und Schüler.

5.7.1 Unterrichtsstörung & Qualität der Lernmotivation

Eine erste Unterrichtsqualitätsdimension stellen Unterrichtsstörungen [25] dar, welche als Subdimension der Klassenführung gelten (Klieme et al., 2001; Praetorius et al., 2018). Die Ergebnisdarstellung (Tabelle 26) erfolgt in Anlehnung an das vorherige Unterkapitel. Verglichen werden Schätzungen resultierend aus dem eLMLIRT-Modell unter Einbeziehung der Unterrichtsstörungen und Schätzungen resultierend aus dem eLMLIRT-Modell unter Einbeziehung der Unterrichtsstörungen und der Qualität der Lernmotivation. Die γ-Werte

Tabelle 26
Vergleich des Effekts eines störungsfreien Unterrichts auf die Veränderung der klassenspezifischen Itemschwierigkeiten mit und ohne Kontrolle der Qualität der Lernmotivation

	eLMLIRT-Modell mit Unterrichtsstörung 2.1			eLMLIRT-Modell mit Unterrichtsstörung & Lernmotivation 2.2			Δ eLMLIRT-Modell 2.1 – eLMLIRT-Modell 2.2
	γ			γ			
	MAP (SD)		95-%-*BCI*	*MAP (SD)*		95-%-*BCI*	Δγ
Item	*t4 – t1*		*t4 – t1*	*t4 – t1*		*t4 – t1*	*t4 – t1*
A01	−.35	(.19)	[−.72, .00]	−.37	(.20)	[−.77, .00]	.02
A02	−.19	(.18)	[−.53, .18]	−.12	(.19)	[−.49, .24]	−.07
A03	−.03	(.17)	[−.34, .31]	−.02	(.18)	[−.38, .31]	−.01
A05	−.19	(.16)	[−.50, .14]	−.23	(.18)	[−.56, .13]	.04
A06	−.14	(.16)	[−.45, .19]	−.12	(.17)	[−.47, .21]	−.02
A21	−.03	(.16)	[−.33, .28]	−.01	(.16)	[−.34, .30]	−.02

Anmerkung: γ: Effekt des Unterrichtsmerkmals auf die klassenspezifischen Itemschwierigkeiten; *MAP*: Maximum-a-posteriori-Schätzung; *SD*: Standardabweichung des Mittelwerts der Posterior-Verteilung; 95-%-*BCI*: 95-%-Kredibilitätsintervall, Δγ : Differenz der γ-Werte beider Modelle.

[25] Siehe Skala zur Unterrichtsstörung im Anhang B.

stehen in beiden Modellen für die Regressionskoeffizienten und geben Auskunft über den Einfluss von Unterrichtsstörungen auf die Veränderung der klassenspezifischen Itemschwierigkeiten mit und ohne Kontrolle der Qualität der Lernmotivation. Negative Regressionskoeffizienten bedeuten, dass ein störungsfreier Mathematikunterricht dazu führt, dass die Items über die Zeit leichter werden.

Um die Hypothese 2.1 zu überprüfen, werden die $\Delta\gamma$-Werte, die γ-Werte und das *BCI* genauer betrachtet. Die Differenzwerte (Δ) basieren auf den γ-Werten der einzelnen Items, welche mithilfe der beiden eLMLIRT-Modelle geschätzt wurden ($\Delta\gamma$). Sie sollen einen ersten Eindruck über mögliche Unterschiede in den Parameterschätzungen aufgrund verschiedener Modellspezifikationen geben. Anhand der Werte in Tabelle 26 ist erkennbar, dass sich die geschätzten γ-Werte beider Modelle kaum voneinander unterscheiden (*Range*$\Delta\gamma = -.07$ bis .04, $M\Delta\gamma = -.02$). Dieser Eindruck, dass die geschätzten γ-Werte nahezu identisch sind, bestätigt sich auch bei der näheren Betrachtung der geschätzten γ-Werte der einzelnen Items mittels eLMLIRT-Modell 2.2 und des geschätzten *BCI* desselben Items mittels eLMLIRT-Modell 2.1. Würde der geschätzte Modalwert eines Items des eLMLIRT-Modell 2.2 außerhalb des geschätzten *BCI* desselben Items des eLMLIRT-Modells 2.1 liegen, könnte mit ausreichender Sicherheit davon ausgegangen werden, dass sich die geschätzten Modalwerte beiden Modelle voneinander unterscheiden. Letzteres ist jedoch nicht der Fall.

Hypothese 2.1, *die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der spezifischen Sensitivität bezogen auf Unterrichtsstörungen im Kontext der Evaluation der Instruktionssensitivität von Testitems*, **ist daher abzulehnen**.

Wie bereits bei der Überprüfung der Hypothese 1 angemerkt wurde, ist auch hier das Kriterium zur Überprüfung der Hypothese 2.1 nicht leicht zu erfüllen. Dies wird bei der Betrachtung der γ-Werte und der *BCI* der hier vorliegenden Ergebnisse deutlich. Die gebildeten $\Delta\gamma$-Werte sprechen aber dafür, dass das Kriterium zur Annahme der Hypothese 2.1 nicht nur knapp, sondern deutlich verfehlt wurde. Die Ablehnung der Hypothese 2.1 wird dadurch nochmals bekräftigt.

Als Ergänzung (nicht zur Überprüfung) zur Hypothese 2.1 soll auch hier zusätzlich kontrolliert werden, ob die Ergebnisse aus den Analysen resultierend aus den zwei unterschiedlichen Modellspezifikationen zu unterschiedlichen Schlussfolgerungen hinsichtlich der spezifischen Sensitivität der Items führen. Hierfür wird das *BCI* des jeweiligen Items resultierend aus dem

eLMLIRT-Modell 2.1 und dem eLMLIRT-Modell 2.2 herangezogen. Anhand der *BCI* in Tabelle 26 wird ersichtlich, dass unabhängig von der Modellspezifikation das Ausmaß von Unterrichtsstörungen keinen Effekt auf die Veränderung der klassenspezifischen Itemschwierigkeiten hat. Deutlich wird zudem, dass unabhängig von der Modellspezifikation das Ergebnis zur Überprüfung der spezifischen Sensitivität mithilfe des *BCI* in diesem Fall identisch ausfällt.

5.7.2 Unterrichtsinhalte & Qualität der Lernmotivation

Fortfahrend mit den Ergebnissen zu den im Unterricht verfolgten Lernzielen werden auch hier die Parameterschätzungen resultierend aus unterschiedlichen Modellspezifikationen miteinander verglichen: das eLMLIRT-Modell unter Einbeziehung der im Unterricht verfolgten Lernziele und das eLMLIRT-Modell unter Einbeziehung der im Unterricht verfolgten Lernziele sowie der Qualität der Lernmotivation (Tabelle 27). Die γ-Werte geben dabei Auskunft über den Einfluss der im Unterricht verfolgten Lernziele auf die Veränderung der klassenspezifischen Itemschwierigkeiten mit und ohne Kontrolle der Qualität der Lernmotivation. Der Prädiktor umfasst ausschließlich inhaltsbezogene Lernziele zur

Tabelle 27
Vergleich des Effekts der im Unterricht verfolgten inhaltsbezogenen Lernziele auf die Veränderung der klassenspezifischen Itemschwierigkeiten mit und ohne Kontrolle der Qualität der Lernmotivation

	eLMLIRT-Modell mit Unterrichtsinhalten 2.1			eLMLIRT-Modell mit Unterrichtsinhalten & Lernmotivation 2.2			Δ eLMLIRT-Modell 2.1 – eLMLIRT-Modell 2.2
	γ			γ			
	MAP (SD)		95-%-*BCI*	*MAP (SD)*		95-%-*BCI*	Δγ
Item	*t4–t1*		*t4–t1*	*t4–t1*		*t4–t1*	*t4–t1*
A01	.00	(.18)	[–.35, .34]	.00	(.19)	[–.36, .38]	.00
A02	–.04	(.19)	[–.41, .33]	.09	(.20)	[–.31, .46]	–.13
A03	–.14	(.17)	[–.47, .17]	–.16	(.18)	[–.51, .18]	.02
A05	–.32	(.16)	[–.65, –.01]	–.38	(.18)	[–.73, –.03]	.06
A06	–.32	(.17)	[–.66, –.02]	–.33	(.18)	[–.69, .03]	.01
A21	.09	(.15)	[–.22, .37]	.12	(.17)	[–.20, .47]	–.03

Anmerkung: γ: Effekt des Unterrichtsmerkmals auf die klassenspezifischen Itemschwierigkeiten; *MAP*: Maximum-a-posteriori-Schätzung; *SD*: Standardabweichung des Mittelwerts der Posterior-Verteilung; 95-%-*BCI*: 95-%-Kredibilitätsintervall, Δγ: Differenz der γ-Werte beider Modelle.

Bruchrechnung (Kapitel 4.4.3). Bei den Testitems handelt es sich ebenfalls ausschließlich um Aufgaben zur Bruchrechnung (Kapitel 5.4.1). Es wird erwartet, dass der Prädiktor einen Effekt auf die Veränderung der klassenspezifischen Itemschwierigkeiten hat. Die Darstellung und Interpretation der Ergebnisse erfolgt analog zu den Analysen der Unterrichtsstörungen in Kombination mit der Qualität der Lernmotivation (Kapitel 5.7.1).

Um die Hypothese 2.3 zu überprüfen, werden auch hier die Δγ-Werte, die γ-Werte und das *BCI* genauer betrachtet. Anhand der Werte in Tabelle 27 ist erkennbar, dass die γ-Werte resulierend aus den unterschiedlichen Modellspezifikationen kaum voneinander abweichen (*Range*Δγ = –.13 bis .06, *M*Δγ = –.01). Dieser Eindruck, dass die geschätzten γ-Werte nahezu identisch sind, bestätigt sich auch bei der näheren Betrachtung der geschätzten γ-Werte der einzelnen Items mittels eLMLIRT-Modell 2.2 und der geschätzten *BCI* derselben Items mittels eLMLIRT-Modell 2.1. Würde der geschätzte Modalwert eines Items des eLMLIRT-Modells 2.2 außerhalb des geschätzten *BCI* desselben Items des eLMLIRT-Modells 2.1 liegen, könnte auch hier mit hinreichender Sicherheit davon ausgegangen werden, dass sich die Parameterschätzungen voneinander unterscheiden. Letzteres ist jedoch nicht der Fall.

Hypothese 2.3, *die Einbeziehung der Qualität der Lernmotivation als Kovariate beeinflusst die Schätzung der spezifischen Sensitivität bezogen auf die im Unterricht verfolgten Lernziele im Kontext der Evaluation der Instruktionssensitivität von Testitems*, **ist daher abzulehnen**.

Analog zu den vorherigen Ausführungen wird auch hier anhand der gebildeten Δγ-Werte ersichtlich, dass die γ-Werte resultierend aus den Analysen mithilfe der eLMLIRT-Modelle mit und ohne Kontrolle der Qualität der Lernmotivation sich kaum voneinander unterscheiden. Es ist also nicht davon auszugehen, dass das Kriterium zur Annahme der Hypothese 2.3 nur knapp verfehlt wurde.

Als Ergänzung zur Hypothese 2.3 soll auch hier überprüft werden, ob die Ergebnisse aus den Analysen resultierend aus unterschiedlichen Modellspezifikationen zu voneinander abweichenden Schlussfolgerungen hinsichtlich der spezifischen Sensitivität der Items führen. Hierfür werden erneut die *BCI* der jeweiligen Items resultierend aus dem eLMLIRT-Modell 2.1 und dem eLMLIRT-Modell 2.2 herangezogen. Die Ergebnisse in Tabelle 27 lassen erkennen, dass im eLMLIRT-Modell 2.1 das Ausmaß der im Unterricht verfolgten Lernziele einen Effekt auf die Veränderung der klassenspezifischen Itemschwierigkeiten bei zwei der sechs Items hat. Es handelt sich hierbei um das Item A05 und das Item A06. Im eLMLIRT-Modell 2.2 erweist

sich bezogen auf die im Unterricht verfolgten Lernziele eines der sechs Items als sensitiv. Es handelt sich hierbei um das Items A05. Das Item A06 verfehlt nur knapp das Kriterium, dass das *BCI* die Null nicht umschließt. Deutlich wird an dieser Stelle, dass je nach Modellspezifikation das Ergebnis zur Überprüfung der spezifischen Sensitivität unterschiedlich ausfallen kann. Anhand dieses Beispiels wird erneut ersichtlich, dass kleine Unterschiede in den Parameterschätzungen zu sehr verschiedenen Schlussfolgerungen führen können.

6 Diskussion

In Kapitel 6.1 wird zunächst eine kurze Zusammenfassung gegeben, was in der vorliegenden Arbeit untersucht wurde und welche Ergebnisse aus den Analysen hervorgegangen sind. Anschließend erfolgt die Diskussion der Resultate in den Kapiteln 6.2 bis 6.6. In Kapitel 6.2 soll zunächst erörtert werden, welche Erklärungsansätze es für das insgesamt unerwartete Ergebnis der vorliegenden Arbeit gibt. In diesem Zusammenhang wird nochmals auf die Beziehung zwischen motivationalen Merkmalen von Schülerinnen und Schülern und den schulischen Leistungen sowie den Indikatoren der Instruktionssensitivität eingegangen. Anschließend erfolgt in Kapitel 6.3 eine Auseinandersetzung mit der Instruktionssensitivität aus der Perspektive von Sensitivität vs. Insensitivität. Wird diese Perspektive zugrunde gelegt, wäre das Ergebnis, dass die Qualität der Lernmotivation einen Einfluss auf den Indikator der Instruktionssensitivität nimmt. Anhand zweier Beispiele soll dabei die potenzielle Wirkweise der Qualität der Lernmotivation auf die Instruktionssensitivität verdeutlicht werden. Fortfahrend mit Kapitel 6.4 wird die Ergebnisdiskussion aus einer methodischen Sichtweise betrachtet. Nach der inhaltlichen und methodischen Diskussion der Ergebnisse zur Qualität der Lernmotivation im Zusammenhang mit den Indikatoren der Instruktionssensitivität soll in Kapitel 6.5 noch ein spezifischer Blick auf die herangezogenen Unterrichtsmerkmale gerichtet werden. Ziel ist eine Auseinandersetzung mit den Unterrichtsmerkmalen, welche zur Überprüfung der spezifischen Sensitivität Berücksichtigung fanden und sich als wenig prädiktiv erwiesen. Der Diskussionsteil schließt in Kapitel 6.6 mit einer kritischen Reflexion der eingesetzten Erhebungsinstrumente.

6.1 Zusammenfassung[26]

Die hier vorliegende Arbeit hat die Evaluation der Instruktionssensitivität von Testitems unter Berücksichtigung individueller Lernvoraussetzungen von Schülerinnen und Schülern zum Thema. Der Begriff der individuellen Lernvoraussetzungen wurde dabei sehr breit gefasst und subsumiert kognitive, metakognitive, motivationale und volitionale Merkmale (Brühwiler, 2014; Brühwiler et al., 2017). Ausgehend von diesen verschiedenen Aspekten individueller Lernvoraussetzungen wurde ein besonderer Fokus auf die Qualität der Lernmotivation (Ryan & Deci, 2000) als ein motivationales Merkmal gelegt, welches in den Analysen zur

[26] Die Zusammenfassung enthält Textpassagen, die aus der Zusammenfassung (Seite 6) stammen.

Instruktionssensitivität von Testitems Berücksichtigung fand. Das Ziel dieser Arbeit bestand in der Überprüfung, inwiefern die Indikatoren der Instruktionssensitivität von Testitems durch die Qualität der Lernmotivation der Schülerinnen und Schüler beeinflusst werden. Die Arbeit hat folglich einen exemplarischen Charakter. Eine allumfassende Antwort auf die Frage, inwiefern lernrelevante Merkmale einen Einfluss auf die Indikatoren der Instruktionssensitivität nehmen, kann im Rahmen dieser Arbeit nicht gegeben werden. Mit der dargelegten Konzeptualisierung zur Einbeziehung von Merkmalen der Schülerinnen und Schüler bei der Evaluation der Instruktionssensitivität ist es jedoch möglich, hieran anzuknüpfen und weitere Analysen unter Hinzuziehung anderer lernrelevanter Merkmale durchzuführen. Letzteres kann als eine besondere Stärke dieser Arbeit erachtet werden.

Inwieweit beeinflusst also die Qualität der Lernmotivation der Schülerinnen und Schüler die Schätzung der Indikatoren der Instruktionssensitivität von Testitems? Kann davon ausgegangen werden, dass Modelle zur Evaluation der Instruktionssensitivität zu stark vereinfacht sind? Um dies zu überprüfen, wurden Parameterschätzungen resultierend aus Analysen mit unterschiedlicher Modellspezifikation gegenübergestellt. Unterscheiden sich die Parameter, spricht dies dafür, dass die Qualität der Lernmotivation einen Einfluss auf die Schätzung der Instruktionssensitivität von Testitems nimmt. Drei Hypothesen wurden folglich im Rahmen dieser Arbeit überprüft. Das Ergebnis ist eindeutig: Keine der aufgestellten Hypothesen konnte angenommen werden. Die Ergebnisse sprechen dafür, dass die Parameterschätzungen zur Evaluation der Instruktionssensitivität von der Qualität der Lernmotivation nicht wesentlich beeinflusst werden. Diese Befundlage überrascht, da es in zahlreichen Studien Hinweise darauf gibt, dass motivationale Merkmale von Schülerinnen und Schülern einen Einfluss auf deren schulische Leistungen nehmen (u.a. Boekaerts, 2007; Kriegbaum et al., 2015).

6.2 Die Qualität der Lernmotivation und das Ausbleiben eines Einflusses auf die Indikatoren der Instruktionssensitivität – zwei Erklärungsansätze

Erklärungsansatz: konstanter Leistungsvorsprung

Ein erster Erklärungsansatz ist möglicherweise darin zu sehen, dass die Analysen mithilfe des LMLIRT- bzw. des eLMLIRT-Modells auf Schätzungen von Veränderungswerten basieren. In diesem Zusammenhang ist denkbar, dass die Qualität der Lernmotivation keinen Einfluss auf die Veränderungswerte, jedoch auf die Querschnittsdaten zum jeweiligen Zeitpunkt (z.B. Posttest-Daten) hat. Ein vorstellbares Szenario wäre, dass Klassen mit Schülerinnen und

Schülern, die mehrheitlich eine hohe Qualität der Lernmotivation zu *t1* aufweisen, bereits höhere Leistungen (bzw. geringere Itemschwierigkeiten) zum ersten Messzeitpunkt haben (im Vergleich zu Klassen mit Schülerinnen und Schülern, die mehrheitlich eine geringere Qualität der Lernmotivation aufweisen). Bleibt dieser Leistungsvorsprung über die Zeit bestehen, spricht dies für einen Haupteffekt der Qualität der Lernmotivation auf die schulische Leistung (Abbildung 19). Der Interaktionseffekt zwischen der Qualität der Lernmotivation und der schulischen Leistung (Abbildung 20) steht in diesem Fall für die Annahme, dass Schülerinnen und Schüler mit einer hohen Qualität der Lernmotivation einen höheren Leistungszuwachs aufweisen als Schülerinnen und Schüler mit einer niedrigen Qualität der Lernmotivation. Die Ausgangsleistungen können dabei entweder vergleichbar (auf eine Darstellung wurde verzichtet) oder auch unterschiedlich (Abbildung 20) ausfallen. Um zu klären, ob der in Abbildung 19 dargestellte Erklärungsansatz tragfähig ist, soll nochmals auf die Beziehung zwischen motivationalen Merkmalen der Schülerinnen und Schüler und den schulischen Leistungen bzw. den Indikatoren der Instruktionssensitivität eingegangen werden.

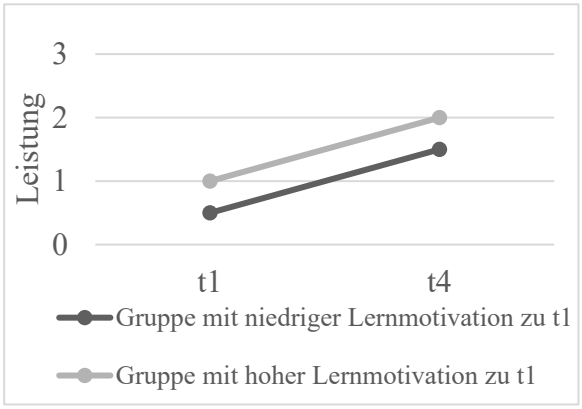


Abbildung 20. Schülerinnen und Schüler mit einer hohen Qualität der Lernmotivation haben höhere Ausgangsleistungen, der Leistungszuwachs ist aber bei Schülerinnen und Schüler mit hoher und niedriger Lernmotivation vergleichbar (Haupteffekt) (eigene Darstellung).

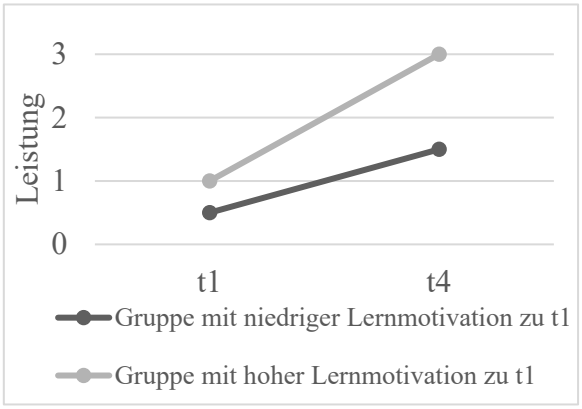


Abbildung 19. Schülerinnen und Schüler mit einer hohen Qualität der Lernmotivation haben einen höheren Leistungszuwachs als Schülerinnen und Schüler mit einer niedrigen Qualität der Lernmotivation (Interaktionseffekt) (eigene Darstellung).

Bezugnehmend auf das Literaturreview von Taylor et al. (2014), das bereits in Kapitel 3.4.2 zur Sprache kam, wurde deutlich, dass die Mehrheit an Studien, in denen die *Academic Motivation Scale* eingesetzt wurde, ein querschnittliches Design aufweist (siehe auch Chatzisarantis et al., 2003). Darüber hinaus gibt es eine Reihe längsschnittlicher Studien, jedoch lag lediglich dreien, die in die Meta-Analyse eingingen, ein längsschnittliches Design zugrunde, in denen die Ausgangsleistung kontrolliert wurde (Taylor et al., 2014; siehe auch Baker, 2003; Black &

Deci, 2000; Burton et al., 2006). Taylor et al. (2014) erachten daher die Erkenntnisse aus den längsschnittlich angelegten Studien unter Verwendung der *Academic Motivation Scale* noch als zu wenig belastbar, um allgemeine Aussagen über den Forschungsgegenstand treffen zu können und plädieren daher für weitere Untersuchungen. Die bisherigen Erkenntnisse der längsschnittlich angelegten Studien deuten aber darauf hin, dass die intrinsische Motiviertheit einen positiven Effekt und die Amotiviertheit einen negativen Effekt auf die Entwicklung der akademischen Leistungen aufweisen, was für die Annahme von Interaktionseffekten spricht. Einschränkend muss jedoch angemerkt werden, dass die Meta-Analyse von Taylor et al. (2014) schulische bzw. akademische Leistungen im Allgemeinen und nicht das Fach Mathematik im Speziellen fokussiert.

Beschränkt man sich nicht nur auf Studien unter Verwendung der *Academic Motivation Scale*, sondern bezieht motivationale Konstrukte im Allgemeinen ein, zeigen die längsschnittlichen Analysen von Kriegbaum et al. (2015), dass motivationale Konstrukte – unter Kontrolle der Vorleistung und der Intelligenz – eine Vorhersagekraft für die Entwicklung von mathematischen Kompetenzen haben. Eine Besonderheit an der Studie von Kriegbaum et al. (2015) ist, dass Ergebnisse aus querschnittlich und längsschnittlich angelegten Analysen gegenübergestellt werden. Den Ergebnissen zufolge ist der Einfluss motivationaler Merkmale auf die Entwicklung mathematischer Kompetenzen geringer als in querschnittlich angelegten Studien ohne Kontrolle der Vorleistung. Auch dieser Aspekt könnte in der vorliegenden Arbeit eine Rolle gespielt haben. Einfache lineare Regressionsanalysen wurden durchgeführt, um den Einfluss der Qualität der Lernmotivation auf die klassenspezifischen Itemschwierigkeiten zu *t1, t4* und auf die Veränderung der klassenspezifischen Itemschwierigkeiten (*t4–t1*) zu prüfen. Der Erklärungsansatz kann jedoch basierend auf den vorliegenden empirischen Daten nicht hinreichend gestützt werden.

Erklärungsansatz: ICC1

Es wurde angenommen, dass die Qualität der Lernmotivation einen Einfluss auf die differenzielle Sensitivität nimmt. Dieser Einfluss wurde mit einem Vergleich der Parameterschätzungen mittels LMLIRT- und eLMLIRT-Modell überprüft. In Kapitel 5.6 wurde dargelegt, dass sich mit der Hinzunahme der Qualität der Lernmotivation im eLMLIRT-Modell die ϕ^2-Werte verändern könnten. Hintergrund ist, dass mit der Einbeziehung der Qualität der Lernmotivation möglicherweise Varianz aufgeklärt wird, sodass sich die

Residualvarianz reduziert. Darüber hinaus ist auch der umgekehrte Fall – also die Zunahme der Residualvarianz – nicht auszuschließen. Die Ergebnisse zur differenziellen Sensitivität sprechen allerdings dafür, dass unabhängig davon, ob die Qualität der Lernmotivation in den Analysen kontrolliert wird oder nicht, die Varianz der Veränderung in den klassenspezifischen Itemschwierigkeiten ähnlich ausfällt. Eine mögliche Erklärung ist in den Ergebnissen zur Intraklassenkorrelation (*ICC1*) zu sehen. Die *ICC1*-Werte zu den Subskalen der Qualität der Lernmotivation (Kapitel 5.1) weisen darauf hin, dass nur ein kleiner prozentualer Anteil der Varianz auf Unterschiede zwischen den Klassen zurückzuführen ist ($ICC1_{Range}$: .05 bis .08). Der deutlich größere Anteil an Varianz kann mit Unterschieden auf der Individualebene in Zusammenhang gebracht werden. Es ist also davon auszugehen, dass wenig Varianz zwischen den Klassen vorhanden ist, wenn es um die Qualität der Lernmotivation geht.

Analog verhält es sich mit den unerwarteten Ergebnissen zur spezifischen Sensitivität (Kapitel 5.7). Angenommen wurde, dass der Effekt der jeweiligen Unterrichtsmerkmale auf die Veränderung der klassenspezifischen Itemschwierigkeiten durch die Qualität der Lernmotivation beeinflusst wird. Zur Überprüfung wurde ein Vergleich von Parameterschätzungen mittels eLMLIRT-Modellen mit und ohne Kontrolle der Qualität der Lernmotivation vorgenommen. Würden sich die Parameterschätzungen unterscheiden, wäre dies ein Indiz dafür, dass die Qualität der Lernmotivation einen Einfluss auf die spezifische Sensitivität nimmt. Die Ergebnisse zur spezifischen Sensitivität sprechen allerdings dafür, dass unabhängig davon, ob die Qualität der Lernmotivation der Schülerinnen und Schülern kontrolliert wird oder nicht, die Effekte der jeweiligen Unterrichtsmerkmale auf die Veränderung der klassenspezifischen Itemschwierigkeiten ähnlich ausfallen. Ein möglicher Erklärungsansatz kann auch hier in den geringen *ICC1*-Werten der Qualität der Lernmotivation gesehen werden. Mit einer *ICC1* von .05 bis .08 liegt es allerdings auch nicht ganz außerhalb des Möglichen, einen Effekt nachzuweisen (LeBreton & Senter, 2008). Die homogene Stichprobe wird aber als zentraler Faktor erachtet, wenn es um die Erklärung der unerwarteten Ergebnisse geht. Um mehr Varianz in den Daten zu haben, wäre eine Ausweitung der Datenerhebungen über den Kanton St. Gallen hinaus erstrebenswert gewesen. Die Umsetzung war im Rahmen des INSE-Projekts aber leider nicht möglich.

6.3 Ergebnisse zur spezifischen Sensitivität von Testitems aus der Perspektive *Sensitivität vs. Insensitivität* – Wie könnte die Qualität der Lernmotivation auf den Indikator wirken?

In den Kapiteln 5.6 und 5.7 wurde die Hypothesenprüfung der vorliegenden Arbeit dargelegt. Liegt der geschätzte Modalwert eines Items resultierend aus dem LMLIRT- bzw. dem eLMLIRT-Modell außerhalb des geschätzten *BCI* desselben Items resultierend aus dem eLMLIRT-Modell, kann davon ausgegangen werden, dass sich die geschätzten Modalwerte mittels beider Modelle voneinander unterscheiden. Eine andere Herangehensweise, um zu analysieren, inwiefern die Qualität der Lernmotivation einen Einfluss auf die Indikatoren der Instruktionssensitivität nimmt, geht mit der Perspektive auf die *Sensitivität vs. Insensitivität* der Items einher. Umschließt das *BCI* nicht die Null, gilt das Item als sensitiv. Im nachfolgenden Abschnitt sollen Ergebnisse zur spezifischen Sensitivität bezogen auf die im Unterricht verfolgten Lernziele bei den Items A05 und A06 aus der Perspektive *Sensitivität vs. Insensitivität* diskutiert werden (Tabelle 28; Kapitel 5.7.2). Beide Items erwiesen sich – im Unterscheid zu den anderen Items – bei den Analysen ohne Kontrolle der Qualität der Lernmotivation als sensitiv. Wurde die Qualität der Lernmotivation jedoch kontrolliert, reagierte ausschließlich das Item A05 sensitiv.[27] Wie kann dieses Ergebnis erklärt werden? Welche Mechanismen könnten in Bezug auf die Qualität der Lernmotivation zum Tragen kommen? Dieser Frage soll nachfolgend nachgegangen werden.

Tabelle 28
Ausschnitt aus Tabelle 27 in Kapitel 5.7.2

	eLMLIRT-Modell mit Unterrichtsinhalten 2.1				eLMLIRT-Modell mit Unterrichtsinhalten & Lernmotivation 2.2			
	γ				γ			
	MAP (SD)		95-%-*BCI*		*MAP (SD)*		95-%-*BCI*	
Item	*t4–t1*		*t4–t1*		*t4–t1*		*t4–t1*	
A05	–.32	(.16)	[–.65,	–.01]	–.38	(.18)	[–.73,	–.03]
A06	–.32	(.17)	[–.66,	–.02]	–.33	(.18)	[–.69,	.03]

Anmerkung: γ: Effekt des Unterrichtsmerkmals auf die klassenspezifischen Itemschwierigkeiten; *MAP*: Maximum-a-posteriori-Schätzung; *SD*: Standardabweichung des Mittelwerts der Posterior-Verteilung; 95-%-*BCI*: 95-%-Kredibilitätsintervall.

[27] Bei Item A05 wurde die Null vom *BCI* nicht umschlossen, was jedoch bei Item A06 der Fall war.

Bei Schrader und Helmke (2002) heißt es: „Je höher die Lernmotivation, desto schneller wird eine Lernhandlung initiiert, gegen Widerstände oder bei Schwierigkeiten [langweilige Aufgabe, unlösbare Probleme d. Verf.] aufrechterhalten und desto intensiver sind Lernengagement, Anstrengung und Persistenz“ (S. 266). Bei einer genauen Betrachtung der Items A05 und A06 ist zunächst ersichtlich, dass diese sehr ähnlich sind (Abbildung 21). Allerdings: Während bei Item A05 *vom Teil und dem Ganzen auf einen Anteil* geschlossen werden muss, geht es bei Item A06 darum *vom Ganzen und dem Anteil auf einen Teil* zu schließen. Solche Umkehraufgaben, also das Vertauschen von *gegeben* und *gesuch*t, sind in der Mathematikdidaktik ein gängiges Mittel zur Aufgabenvariation. Zum Lösen des Items A05 ist i.d.R. ein prozedurales oder ein prozedurales und konzeptionelles Wissen erforderlich. Item A06 ist vom Aufgabentyp her anspruchsvoller, da es i.d.R. nicht ausreichend sein wird, das prozedurale Wissen abzurufen und anzuwenden. Es ist vielmehr auch konzeptionelles Wissen erforderlich. Insofern könnte eine hohe Qualität der Lernmotivation zum Lösen des Items A06 von höherer Bedeutung sein als bei Item A05, da bei A06 mitunter erst ein Lösungsweg gesucht werden muss. Die Anstrengungsbereitschaft, die Ausdauer und die Persistenz, bei auftretenden Schwierigkeiten dranzubleiben, wären in einem solchen Fall von besonderem Wert. So ließe sich das Ergebnis erklären, dass bei Item A06 die Qualität der Lernmotivation einen Einfluss auf den Indikator der Instruktionssensitivität nimmt und es infolgedessen einen Unterschied macht, ob die Qualität der Lernmotivation kontrolliert wird oder nicht.

Berechne. $\frac{3}{4}$ von 80 = ☐ © Lehrmittelverlag St. Gallen	**Ergänze den Zähler des Bruchs.** $\frac{\square}{4}$ von 80 = 20 © Lehrmittelverlag St. Gallen
A05	A06

Abbildung 22. Item A05 (links) und A06 (rechts) der Klassencockpit-Testitems (siehe auch Abbildung 8 in Kapitel 4.4.1).

Das Beispiel soll an dieser Stelle zeigen, in welcher Hinsicht die Qualität der Lernmotivation im Kontext der Instruktionssensitivität zum Tragen kommen könnte. Die Befundlage ist allerdings zu wenig belastbar, um den Erklärungsansatz für die aufgeführten Ergebnisse heranzuziehen. Darüber hinaus soll mit dem Erklärungsversuch bezogen auf die Items A05 und A06 dargelegt werden, dass die Wirkweisen der Qualität der Lernmotivation itemspezifisch – und nicht nur testspezifisch (alle Items eines Tests) – betrachtet werden können. Den Fokus auf

das jeweilige Testitem zu richten, kann eine gewinnbringende Perspektive bei der Interpretation der Ergebnisse zur Instruktionssensitivität darstellen (Kapitel 6.5). In den bisherigen Studien finden spezifische Merkmale der Items zur Erklärung von *Sensitivität vs. Insensitivität* jedoch kaum Berücksichtigung.

6.4 Methodische Ergebnisdiskussion

Nachdem die Diskussion bis hierhin vornehmlich aus einer inhaltlichen Perspektive geführt wurde, soll nun eine methodische Auseinandersetzung mittels zweier Aspekte erfolgen. Es handelt sich hierbei zum einen um die Unsicherheit in den Parameterschätzungen und zum anderen um methodische Ansätze zur Hypothesentestung.

6.4.1 Unsicherheit in den Parameterschätzungen

Wie in Kapitel 6.3 beschrieben, wurden zur Überprüfung der Hypothesen die geschätzten Modalwerte und das jeweilige *BCI* herangezogen. Im Ergebnisteil (Kapitel 5.6) wurde bereits angedeutet, dass dieses Kriterium der Hypothesenprüfung nicht leicht zu erreichen war. Was damit gemeint ist, wird bei der Betrachtung der *BCI* zur differenziellen Sensitivität deutlich. Die ϕ^2-Werte können generell Werte zwischen null und eins annehmen. Mit Blick auf die *BCI* zur differenziellen Sensitivität wird ersichtlich, dass diese häufig einen Großteil des Wertebereichs abdecken. Die Werte sprechen dafür, dass viel Unsicherheit in den Parameterschätzungen vorhanden ist. Um die jeweilige Hypothese anzunehmen, müssten die jeweiligen ϕ^2-Werte sehr niedrig oder sehr hoch ausfallen.

Mit dem Ziel weniger Unsicherheit in den Parameterschätzungen zu haben, könnten verschiedene Ansätze verfolgt werden. Eine Möglichkeit besteht in der Wahl eines informativen Priors (Held, 2008; Kaplan 2014). Die Einbeziehung von Vorinformationen in die Analysen ist eine besondere Stärke der bayesianischen Statistik (Held, 2008; Kaplan 2014). Wurden bereits ähnliche Analysen vorgenommen oder ist entsprechendes Expertinnen- und Expertenwissen vorhanden, auf das zurückgegriffen werden kann, sollte dieses Wissen nicht vernachlässigt werden (Held, 2008; Kaplan 2014). In den Analysen der vorliegenden Arbeit fiel stattdessen die Wahl auf einen nicht informativen Prior (Kapitel 4.5.3). Der nicht informative Prior hat zur Folge, dass (1) alle möglichen Werte gleichermaßen wahrscheinlich sind und (2) bei den Parameterschätzungen den empirischen Daten ein hohes Gewicht

zugesprochen wird (Held, 2014; Kaplan, 2014).[28] Nun kann die Frage gestellt werden, ob es sinnvoll gewesen wäre, die Wahl des Priors anders zu gestalten. Aufgrund des Neuheitswerts der vorliegenden Arbeit mit innovativer Methode gab es jedoch ein starkes Argument für die Wahl eines nicht informativen Priors. Sofern aber in einem ähnlichen Kontext weitere Analysen hinsichtlich des Einflusses motivationaler Merkmale auf die Indikatoren der Instruktionssensitivität von Testitems angestrebt werden sollten, ist die Verwendung eines informativen Priors zukünftig vorstellbar. Die Spezifizierung eines informativen Priors muss aber unter Berücksichtigung von bereits gewonnenen/vorhandenen Informationen gut durchdacht werden.

6.4.2 Methodische Ansätze zur Hypothesenprüfung

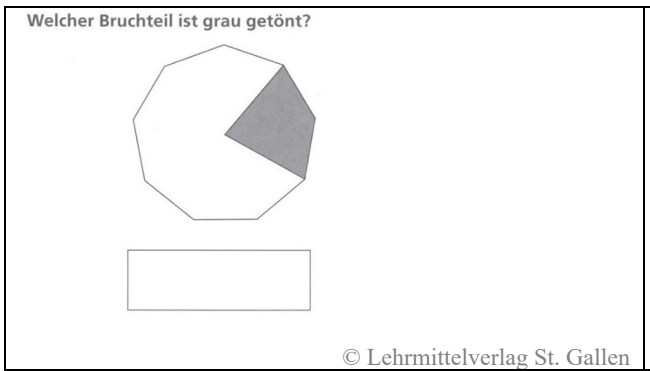
In den Kapiteln 5.6 und 5.7 wurde ergänzend zur Hypothesenprüfung die *Sensitivität vs. Insensitivität* der einzelnen Items resultierend aus den Analysen mit unterschiedlicher Modellspezifikation verglichen. Die Erkenntnis: Je nach Modellspezifikation (mit und ohne Einbeziehung der Qualität der Lernmotivation als Kovariate) unterscheiden sich die Parameterschätzungen geringfügig, was bei einzelnen Items zu verschiedenen Schlussfolgerungen hinsichtlich der *Sensitivität vs. Insensitivität* führt. Deutlich wurde dies anhand der Ergebnisse zu den im Unterricht verfolgten Lernzielen, auf die bereits in Kapitel 6.3 eingegangen wurde. Basierend auf den Analysen mittels eLMLIRT-Modell 2.1 hat das Ausmaß der im Unterricht verfolgten Lernziele bei den Items A05 und A06 einen Effekt auf die Veränderung der klassenspezifischen Itemschwierigkeiten. Bei den Analysen mithilfe des eLMLIRT-Modells 2.2 erweist sich bezogen auf die im Unterricht verfolgten Lernziele hingegen nur das Item A05 als sensitiv. Das Item A06 verfehlt das Kriterium, dass das *BCI* die Null nicht umschließt, nur knapp. Insofern wird deutlich, dass bereits geringe Unterschiede in den Parameterschätzungen zu ganz unterschiedlichen Schlussfolgerungen führen und eine Betrachtung der Ergebnisse aus der Perspektive *Sensitivität vs. Insensitivität* – wie es in vielen Studien zur Instruktionssensitivität der Fall ist – dazu verleiten kann, die Berücksichtigung der Qualität der Lernmotivation sehr wohl als relevant zu erachten. In Anbetracht der geringen Differenzwerte Δ und den Erkenntnissen bezogen auf die Unsicherheit in den Parameterschätzungen erscheint das Verfahren *Sensitivität vs. Insensitivit*ät zur Überprüfung

[28] Eine sehr eindrückliche Darstellung zum Einfluss eines informativen bzw. eines nicht informativen Priors im Rahmen baysianischer Schätzungen ist bei Kruschke (2015) oder auch bei MacKenzie et al. (2018) zu finden.

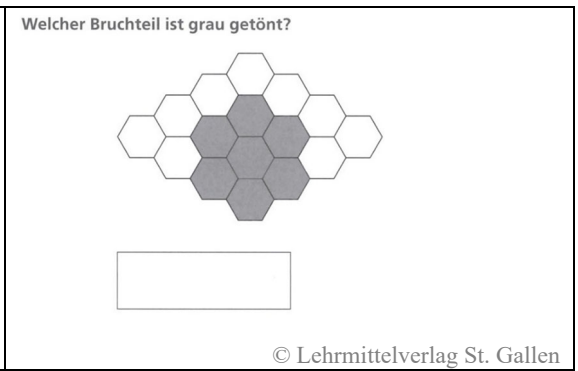
der Hypothesen als nicht angemessen. Letzteres bekräftigt das gewählte Verfahren zur Hypothesenprüfung, wenngleich dieses deutlich konservativer ist.

6.5 Warum sind die Unterrichtsmerkmale nicht prädiktiv?

Eine weitere Erkenntnis der vorliegenden Arbeit, die indirekt mit den Hypothesen in Verbindung steht, betrifft die Beziehung zwischen den Unterrichtsmerkmalen und den Indikatoren der Instruktionssensitivität. Der Nachweis eines Zusammenhangs zwischen spezifischen Unterrichtsmerkmalen und den Indikatoren der Instruktionssensitivität stellt häufig, wie auch diese Arbeit zeigt, eine Herausforderung dar. Zwar lassen die deskriptiven Ergebnisse erkennen, dass die Testitems (Abbildung 22) über die Zeit leichter werden (Kapitel 5.4.1) und sich im Zuge dessen als global sensitiv erweisen (Kapitel 5.5.1). Bei der Überprüfung der spezifischen Sensitivität wurde jedoch deutlich, dass nur wenige Items auf die spezifischen Unterrichtsmerkmale sensitiv reagieren. Das Item A05 bzw. die Items A05 und A06 zeigten sich bei den im Unterricht verfolgten Lernzielen sensitiv, während sich die Items A02, A03 und A21 in diesem Kontext als nicht sensitiv erwiesen. Hinzu kommt, dass keines der Items auf das Unterrichtsqualitätsmerkmal der Unterrichtsstörungen sensitiv reagierte. Welche Erklärungen können für die *Sensitivität* bzw. *Insensitivität* der Items herangezogen werden? Die im Zuge der inhaltlichen Auseinandersetzung gewonnenen Erkenntnisse bezogen auf die Klassencockpit-Testitems sowie die Unterrichtsmerkmale sollen zur Beantwortung dieser Frage aufgezeigt werden (Kapitel 6.5.1 und 6.5.2) und als Anregung zukünftiger Forschungsarbeiten dienen.



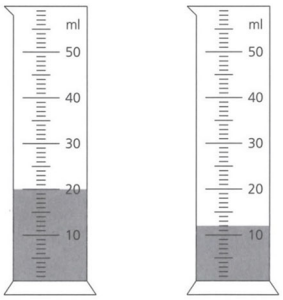
Item A01



Item A02

Abbildung 22. Klassencockpit-Testitems, welche in die Analysen zur Instruktionssensitivität einbezogen wurden (Kapitel 4.4).

Das Messglas links ist zu $\frac{1}{3}$ gefüllt.



Zu welchem Bruchteil ist das Messglas rechts gefüllt?

Item A03

Berechne.

$\frac{3}{4}$ von 80 = ☐

Item A05

Ergänze den Zähler des Bruchs.

$\frac{\square}{4}$ von 80 = 20

Item A06

100 Personen wurden zu ihren Tätigkeiten in der Freizeit befragt.

Freizeitaktivitäten

Tätigkeit	täglich	wöchentlich	monatlich	seltener	nie
Freunde, Kollegen und Bekannte treffen	12	61	20	5	2
Lesen	70	19	5	3	3
Sport treiben	6	50	12	4	28
Kurse besuchen	0	11	6	9	74
Musizieren, Singen	8	15	5	3	69
Handarbeit, Werken, Gartenarbeit	12	38	16	8	26
Kino besuchen	0	3	27	36	34
Theater, Opern oder Ausstellungen besuchen	0	2	24	42	32
Sportanlässe besuchen	0	6	18	22	54
Fernsehen	83	8	4	3	2
Internet benützen	75	12	2	0	?

Welche Aussage ist richtig?

- ☐ Ein Drittel der befragten Personen besuchen monatlich ein Kino.
- ☐ Drei Fünfundzwanzigstel der Befragten benützen wöchentlich das Internet.
- ☐ Ein Fünftel der Befragten singt täglich.
- ☐ Ein Viertel der Befragten besucht nie einen Sportanlass.

Item A21

Abbildung 23. Klassencockpit-Testitems, welche in die Analysen zur Instruktionssensitivität einbezogen wurden (Kapitel 4.4).

6.5.1 Inhaltliche Auseinandersetzung zur spezifischen Sensitivität bezogen auf Unterrichtsstörungen

Beginnend mit dem Unterrichtsmerkmal der Unterrichtsstörungen ist anhand der Ausführungen in Kapitel 3.5.2 deutlich geworden, dass das Ausmaß an Unterrichtsstörungen mit der aktiven Lernzeit in Beziehung gebracht wird, die sich wiederum – so die Annahme – auf die Leistungen der Schülerinnen und Schüler auswirkt. Bezugnehmend auf den Kontext der Instruktionssensitivität wurde infolgedessen angenommen, dass ein Unterricht frei von Störungen einen positiven Einfluss auf die Veränderung der klassenspezifischen Itemschwierigkeiten hat. Das Item A01 verfehlte – im Gegensatz zu den anderen fünf Testitems – nur äußerst knapp das Kriterium, um als spezifisch sensitiv zu gelten. Bei der genauen Betrachtung der γ-Werte und den *BCI* wurde dies deutlich (Tabelle 29). Vor diesem Hintergrund soll eine Auseinandersetzung erfolgen, was dieses Item von den anderen möglicherweise unterscheidet.

Tabelle 29
Ausschnitt aus Tabelle 26 in Kapitel 5.7.1

	eLMLIRT-Modell mit Unterrichtsstörung 2.1				eLMLIRT-Modell mit Unterrichtsstörung & Lernmotivation 2.2			
	γ				γ			
	MAP (SD)		95-%-*BCI*		*MAP (SD)*		95-%-*BCI*	
Item	*t4 – t1*		*t4 – t1*		*t4 – t1*		*t4 – t1*	
A01	−.35	(.19)	[−.72,	.00]	−.37	(.20)	[−.77,	.00]
A02	−.19	(.18)	[−.53,	.18]	−.12	(.19)	[−.49,	.24]
A03	−.03	(.17)	[−.34,	.31]	−.02	(.18)	[−.38,	.31]
A05	−.19	(.16)	[−.50,	.14]	−.23	(.18)	[−.56,	.13]
A06	−.14	(.16)	[−.45,	.19]	−.12	(.17)	[−.47,	.21]
A21	−.03	(.16)	[−.33,	.28]	−.01	(.16)	[−.34,	.30]

Anmerkung: γ: Effekt des Unterrichtsmerkmals auf die klassenspezifischen Itemschwierigkeiten; *MAP*: Maximum-a-posteriori-Schätzung; *SD*: Standardabweichung des Mittelwerts der Posterior-Verteilung; 95-%-*BCI*: 95-%-Kredibilitätsintervall.

Das Item A01 weist im Vergleich zu den anderen Items ein hohes Potenzial für einen Fehler auf, was möglicherweise aus einer unzureichenden Analyse der Aufgabendarstellung resultiert (Radatz, 1980). Lassen sich Schülerinnen und Schüler vom ersten Eindruck leiten oder schauen nicht genau hin bzw. wenden spezifische Techniken nicht an, die gegebenenfalls im Unterricht gelernt wurden (z.B. Geobrett, Falten, Linien ziehen), könnte bei Item A01 schnell als Ergebnis

$\frac{1}{4}$ statt $\frac{2}{9}$ angeben werden (Abbildung 22). Hintergrund ist, dass die Strukturierung fehlt (Teile eines Ganzen) und die Figur eventuell mit einem Kreis assoziiert bzw. verwechselt wird.[29] Ein Blick in die Rohdaten verdeutlicht, dass die Antwort $\frac{1}{4}$ zu allen Messzeitpunkten die mit Abstand am häufigsten falsch genannte Antwort darstellt. Zum ersten Messzeitpunkt haben über 100 Schülerinnen und Schüler diese Antwort gegeben, im Verlauf des Schuljahres stieg dieser Anteil sogar weiter an.

Radatz (1980) zufolge können mehrere Ursachen zur Erklärung eines solchen Fehlers bei der Aufgabe A01 herangezogen werden. Eine Möglichkeit liegt beispielsweise im „Gebundensein […] spezifische[r] Repräsentationsformen" (S. 47). Das heißt: Wenn sehr vertraute Darstellungen Anwendung finden, diese aber minimal verändert werden, kann dies zu einem vermehrten Auftreten von Fehlern führen. Da davon auszugehen ist, dass Schülerinnen und Schüler bereits früh eine Vorstellung von Brüchen als *Teile eines Ganzen* haben (siehe Lehrmittel logisch 4; Noser, Oester Schläppi, Scherer & Züger, 2004) und die Angabe $\frac{1}{4}$ im Alltag geläufig ist, ist es nicht verwunderlich, dass diese Antwort zu den jeweiligen Messzeitpunkten genannt wird. Ein zweiter spannender Aspekt, wenn es um die Ursache von Fehlern beim Lösen dieses Klassencockpit-Testitems geht, betrifft die „kognitive Stildimension Impulsivität vs. Reflektiertheit" (Radatz, 1980, S. 50; siehe auch Kagan & Kogan, 1970; Radatz, 1976). Gemeint sind damit …

> individuelle Unterschiede in der Art und in der Geschwindigkeit, mit der Informationen verarbeitet sowie Lösungshypothesen evaluiert und angeboten werden. (…) Impulsive zeigen [im Vergleich zu stärker Reflektierten d. Verf.] kurze Antwortzeiten und bieten sehr oft die erstbeste Hypothese an, ohne diese zuvor ausreichend auf ihre Angemessenheit und Gültigkeit geprüft zu haben. Das ist verbunden mit einer höheren Fehlerwahrscheinlichkeit in komplexeren Situationen. (Radatz, 1980, S. 50)

Die zwei genannten Ursachen (Gebundensein, kognitive Stildimension) können eine Erklärung für den am häufigsten vorkommenden Fehler bei Item A01 darstellen, wohlwissend, dass die Argumentationslinie nur eine Auswahl an möglichen Fehlerquellen[30] berücksichtigt. Wie kann aber der Erklärungsansatz zur Ursache des spezifischen Fehlers bei Item A01 mit den

[29] Im Lehrmittel weisen ähnliche Aufgaben eine Strukturierung von (gleichen) Teilen eines Ganzen auf. Lediglich beim Uhren-Beispiel wird eine Strukturierung von (gleichen) Teilen eines Ganzen nur angedeutet.
[30] Eine weitere interessante fachdidaktische Perspektive: Beide Items fokussieren auf Anteile eines Ganzen. A01 aber räumlich-kontinuierlich und A02 numerisch-diskret.

Erkenntnissen zur spezifischen Sensitivität bezogen auf Unterrichtsstörungen – Item A01 verfehlt äußerst knapp das Kriterium zur spezifischen Sensitivität – in Beziehung gesetzt werden? Kann es sein, dass insbesondere in Klassen mit einem vermehrten Aufkommen an Unterrichtsstörungen der hier beschriebene Fehler (Antwort: $\frac{1}{4}$) bezogen auf das Item A01 häufiger vorzufinden ist?

Aktuelle Publikationen deuten darauf hin, dass Unterrichtsstörungen auch eine Frage der Klassenkomposition sind. Gemäß Göllner, Fauth, Lenske, Praetorius und Wagner (2020) ist das Ausmaß an Unterrichtsstörungen nicht allein vom Handeln der Lehrperson, sondern auch von den Schülerinnen und Schülern bzw. der Klassenkomposition anhängig (u.a. Fauth, Göllner, Lenske, Praetorius & Wagner, 2020). Eine Beziehung zwischen der Impulsivität von Schülerinnen und Schülern und dem Aufkommen an Unterrichtsstörungen ist denkbar. Höflich (2019) vertritt die Ansicht, dass die Impulskontrolle nicht nur für das Lern- und Arbeitsverhalten (siehe auch Lintorf, 2012; Roth, 2011), sondern auch für das Sozialverhalten von hoher Bedeutung ist (siehe auch Roth, 2011). Während bei Radatz (1980) die Impulsivität mit dem Lern- bzw. Arbeitsverhalten und den daraus resultierenden Fehlerursachen in Zusammenhang gebracht wird, kann mit den Ausführungen von Höflich (2019) und Roth (2011) auch eine Beziehung zwischen der Impulsivität und dem Sozialverhalten hergestellt werden. Insofern liefern diese Aussagen einen möglichen Erklärungsansatz, welcher ein spezifisches Merkmal der Schülerinnen und Schüler, die Fehlerpotenziale von Testitems unter Berücksichtigung von Informationsaufnahme und -verarbeitungsprozessen sowie ein Unterrichtsqualitätsmerkmal gleichzeitig in Betracht zieht. Eine gewisse Komplexität ist dementsprechend erkennbar. Die Heranziehung einer sehr allgemeinen Erklärung zum Effekt von Unterrichtsstörungen auf die Leistungen der Schülerinnen und Schüler erscheint in diesem spezifischen Kontext zur Evaluation der Instruktionssensitivität von Testitems nicht hinreichend. Hintergrund ist, dass der Aufgabentyp A01 (bzw. A02) im obligatorischen Lehrmittel sehr zentral positioniert ist und dem Aufbau von mathematischen Grundvorstellungen dient (vom Hofe, 1995; Noser et al., 2005a, 2005c). Folglich sollte diesem Thema ein hoher Stellenwert beigemessen werden und es ist davon auszugehen, dass die große Mehrheit der Schülerinnen und Schüler mit derartigen Items vertraut ist, selbst wenn es sich um einen von Störungen geprägten Unterricht handelt. Darüber hinaus ist anzumerken, dass die anderen fünf Testitems weit davon entfernt sind, als spezifisch sensitiv bezogen auf Unterrichtsstörungen zu gelten. Diese Erkenntnis kann in die insgesamt inkonsistente

Befundlage bezogen auf Daten zu den drei generischen Grunddimensionen der Unterrichtsqualität eingeordnet werden (Kapitel 6.6).

6.5.2 Inhaltliche Auseinandersetzung zur spezifischen Sensitivität bezogen auf die im Unterricht verfolgten Lernziele

Bezugnehmend auf die im Unterricht verfolgten Lernziele erweisen sich das Item A05 und z.T. das Item A06 als sensitiv. Welche Erklärungsansätze es dafür gibt, dass diese sensitiv auf die im Unterricht verfolgten Lernziele reagierten, während sich die anderen Items diesbezüglich als insensitiv herausstellten, soll vor dem Hintergrund einer inhaltlichen Auseinandersetzung mit dem einbezogenen Prädiktor aufgezeigt werden.

Beginnend mit einem detaillierten Blick auf die im Unterricht verfolgten Lernziele (gebündelt) als Prädiktor zur Vorhersage der Veränderung der klassenspezifischen Itemschwierigkeiten (Abbildung 23) ist zunächst zu berücksichtigen, dass sich dieser auf das Buchkapitel „Brüche“ im obligatorischen Lehrmittel bezieht (Noser et al., 2005a, S. 46-49, siehe auch Noser et al., 2005c). Dies bedeutet allerdings nicht, dass sich alle Klassencockpit-Testitems, die in ähnlicher Form im Buchkapitel „Brüche“ vorkommen, auf die im Unterricht verfolgten Lernziele als sensitiv und alle Klassencockpit-Testitems, die in ähnlicher Form im Buchkapitel „Bruchteile von Größen“ abgebildet sind (Noser et al., 2005a, S. 62-64, siehe auch Noser et al., 2005c), sich als insensitiv erweisen sollten. Die Gründe hierfür werden kurz dargelegt.

Lernziele zur Bildung des Prädiktors
Dezimale Zahlen und Brüche als Erweiterung des Bereichs der natürlichen Zahlen erfahren
Einfache Brüche der Größe nach ordnen
Einfluss von Zähler und Nenner auf den Wert des Bruches erkennen
Brüche kürzen und erweitern
Einfluss von Veränderungen im Zähler und Nenner auf den Wert von Brüchen erkennen

Abbildung 24. Lernziele bezogen auf das Kapitel „Brüche“ im Lehrmittel „Logisch 5“, die zur Bildung des Prädiktors herangezogen wurden (eigene Darstellung).

Mit dem Buchkapitel „Brüche“ wird primär der Aufbau von mathematischen Grundvorstellungen intendiert (Noser et al., 2005a, 2005c). Wird dieses Ziel im Unterricht intensiv verfolgt, kann dies für nachfolgende, darauf aufbauende Themen vorteilhaft sein. Aus diesem Grund ist es vorstellbar, dass nicht nur die Klassencockpit-Testitems bezogen auf das

Buchkapitel „Brüche“, sondern auch bezogen auf das Buchkapitel „Bruchteile von Größen“ auf die im Unterricht verfolgten Lernziele sensitiv reagieren.[31]

Warum allerdings die Items A01 und A02 auf die im Unterricht verfolgten Lernziele nicht sensitiv reagieren, ist schwer zu erklären, da beide als zentral für das Buchkapitel „Brüche“ erachtet werden. Die Items verlangen ein prozedurales Wissen, das mithilfe der Inhalte aus dem Kapitel „Brüche“ aufgebaut werden soll. Während Item A01 mit der Aufgabe „Bruchteile auf einer Uhr“ assoziiert werden kann, ist das Item A02 mit der Aufgabe „Bruchteile mit Geomat-Plättchen“ im Kapitel „Brüche“ des Lehrmittels „Logisch 5“ in Verbindung zu bringen (Noser et al., 2005a, S. 46 & S. 48, 2005c). Anhand der vorherigen Ausführungen wurde jedoch deutlich, dass das Item A01 aufgrund des spezifischen Fehlerpotenzials (der erste Eindruck täuscht) nicht ganz unproblematisch ist. Hierin könnte ein limitierender Faktor liegen. Letzteres würde aber nicht erklären, warum das Item A02 nicht sensitiv auf die im Unterricht verfolgten Lernziele reagiert. Ein Erklärungsansatz kann darin gesehen werden, dass der Prädiktor zu wenig auf diesen spezifischen Aufgabentyp (Item A01 und A02) ausgerichtet ist und dies möglicherweise den Grund für die Insensitivität dieser Items darstellt.

Die Items A03, A05 und A06 können als zentral für das Buchkapitel „Bruchteile von Größen“ erachtet werden (Noser et al., 2005a, S. 64, 2005c, S. 129). Hierbei geht es insbesondere um die Bestimmung der Bruchteile von Größen, also z.B. Bruchteile von 1 Liter. Im Unterschied zu den Items A05 und A06 ist Item A03 kontextualisiert und verlangt einen Darstellungswechsel (Brüche, Dezimalzahlen). Warum Item A05 und z.T. Item A06 sensitiv reagieren, das Item A03 jedoch nicht, kann an dieser Stelle nicht beantwortet werden. Ein Erklärungsansatz ist aber auch hier im Prädiktor zu erachten, der sich möglicherweise in diesem Kontext nur bei einem bestimmten Aufgabentyp (Aufgabenstellung, Aufgabenbearbeitung) als passend erweist.

Ein letztes Item, das in die Analysen eingegangen ist, stellt das Item A21 dar. Dieses Item kommt weder im Kapitel „Brüche“ noch im Kapitel „Bruchteile von Größen“ in ähnlicher Form vor (Noser et al., 2005a, 2005c). Anhand der Itemschwierigkeiten wird ersichtlich, dass sich diese im eher geringen Maße über die Zeit verändern (Kapitel 5.4). Dementsprechend ist das Item A21 auch zum Messzeitpunkt von *t4* für viele Schülerinnen und Schüler sehr

[31] Die Datenerhebung zu *t4* fand zu einem Zeitpunkt statt, zu dem sowohl das Buchkapitel „Brüche“ als auch das Buchkapitel „Bruchteile von Größen“ gemäß intendiertem Curriculum bereits im Unterricht behandelt worden waren.

anspruchsvoll. Letzteres stellt aber noch keine hinreichende Begründung für die Insensitivität des Items dar. Ein möglicher Erklärungsansatz kann auch hier in den bereits dargelegten Ausführungen zur Passung erachtet werden. Hinzu kommt, dass die Kontextualisierung des Items nicht ideal ist. „Drei Fünfundzwanzigstel“ ist eine Angabe, die im Alltag kaum vorkommen wird und daher sehr konstruiert wirkt.

Somit bleibt abschließend festzuhalten, dass die Bildung des Prädiktors eine mögliche Erklärung für die erwartungswidrigen Ergebnisse darstellt.[32] Auch wenn dem Prädiktor differenzierte Daten zu den Lernzielen zugrunde liegen, ist die Herstellung einer Beziehung zu den Items schwierig. Eventuell wäre ein feineres Maß zur Bildung des Prädiktors zielführend gewesen. Die Ausführungen in diesem und dem vorherigen Unterkapitel 6.5.1 geben darüber hinaus eine Idee von der Notwendigkeit einer vertieften inhaltlichen Auseinandersetzung, wenn es um die Erklärung von Sensitivität bzw. Insensitivität geht. Diese umfasst die Auseinandersetzung mit den Items, den Anforderungen an die Schülerinnen und Schüler, die sich aus der Aufgabenstellung heraus ergeben, Vorstellungen über die Herausforderungen und Hindernisse bei der Aufgabenbearbeitung sowie die Beziehung, in welcher die Items mit Unterrichtsmerkmalen stehen (könnten). Mit dem Zusammenführen dieser Informationen könnten mitunter interessante Erklärungsansätze herausgearbeitet werden.

6.6 Kritische Reflexion eingesetzter Fragebogenskalen und Erhebungsinstrumente

Nachfolgend werden ausgewählte Aspekte der eingesetzten Fragebogenskalen bzw. der Instrumente zur Erhebung der Daten reflektiert.[33] Im Fokus des Abschnitts stehen (1) die Ausführungen möglicher Ursachen der inkonsistenten Befundlage zu den Basisdimensionen der Unterrichtsqualität, insbesondere mit Blick auf die Klassenführung als übergeordnetem Konstrukt der Unterrichtsstörungen, (2) die kritische Auseinandersetzung mit dem Instrument zur Erhebung der im Unterricht verfolgten Lernziele und (3) den Items des Klassencockpit-Tests vor dem Hintergrund möglicher Testungseffekte. Auf eine kritische Reflexion der Fragebogenskala zur Erhebung der Qualität der Lernmotivation wird verzichtet, da sich diese

[32] Siehe zudem die kritische Reflexion des Erhebungsinstruments in Kapitel 6.6.
[33] Die Auswahl ist damit zu begründen, dass bereits einzelne Punkte zur Sprache kamen, die an dieser Stelle nicht wiederholt werden sollen (Kapitel 6.5).

auf sehr allgemeine Aspekte von Fragebogenerhebungen (u.a. Raab-Steiner & Benesch, 2015) beziehen würde.

Fragebogenskala zur Erfassung von Unterrichtsstörungen

Studien, die das im deutschsprachigen Raum am meisten genutzte *Framework* zur Erfassung der Unterrichtsqualität (drei Basis- bzw. Grunddimensionen der Unterrichtsqualität) zur Grundlage haben, weisen gemäß den Erkenntnissen aus einem Literaturreview eine inkonsistente Befundlage auf (Praetorius et al., 2018). So konnte herausgestellt werden, dass nur in etwas mehr als der Hälfte der Studien ein Effekt der Klassenführung (als übergeordnetem Konstrukt der Unterrichtsstörungen) auf die Leistungen der Schülerinnen und Schüler nachweisbar ist (ebd.; siehe auch Praetorius, Klieme et al., 2020). Als Ursache für diese inkonsistente Befundlage werden mehrere Faktoren in Betracht gezogen (Fauth et al., 2020; Göllner et al., 2020; Praetorius et al., 2018; Praetorius, Klieme et al., 2020). Ersichtlich wird zunächst anhand der Arbeit von Praetorius et al. (2018), dass die jeweiligen Unterrichtsqualitätsdimensionen unterschiedlich operationalisiert werden. Die Skalen zur Klassenführung, zu denen auch die Unterrichtsstörungen zählen, beziehen sich auf unterschiedliche Akteurinnen und Akteure (Fauth et al., 2020). Folglich kann sich die vorzunehmende Einschätzung auf das Verhalten von Schülerinnen und Schülern, auf das Verhalten von Lehrpersonen oder auch auf die Interaktion beider Personengruppen beziehen (ebd.). Ein weiterer Faktor ist, dass die Einschätzungen von Schülerinnen und Schülern, von Lehrpersonen oder auch von externen Beobachterinnen und Beobachtern vorgenommen werden (ebd.). Unterschiedliche Zugänge zu spezifischem Wissen und zum eigenen Erleben können hierbei eine Rolle spielen (Clausen, 2002). Zu bedenken ist auch, dass die Einschätzungen zum Teil auf einen sehr begrenzten Ausschnitt des regulären Unterrichts basieren (Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014). Variierende Stichprobengrößen, verschiedene Schulfächer und Schulformen, vielfältige Analysestrategien (z.B. Einbeziehung von Kovariaten) und Unterschiede in den abhängigen Variablen werden als weitere potenzielle Ursachen für die inkonsistente Befundlage betrachtet (Praetorius et al., 2018).

Kritisch zu hinterfragen ist auch, ob die Erfassung von generischen Merkmalen der Unterrichtsqualität das Mittel der Wahl ist oder ob die Fachspezifität mithilfe der Indikatoren stärker berücksichtigt werden müsste (u.a. Dreher & Leuders, 2021). Praetorius, Herrmann et

al. (2020) arbeiten in ihrem Artikel heraus, dass auch bei der Unterrichtsqualitätsdimension der Klassenführung die Fachspezifität eine Rolle zu spielen scheint. Die aktuellen Indikatoren der Klassenführung sind vornehmlich auf den Frontalunterricht ausgerichtet, sodass kooperative Lernformen weniger Berücksichtigung finden (Praetorius, Herrmann et al., 2020). Auch dieser Aspekt kann eine Ursache für das unerwartete Ergebnis dieser Arbeit darstellen.

Abschließend bleibt festzuhalten, dass in der Unterrichtsforschung ein verstärktes Interesse darin besteht, die Ursachen der inkonsistenten Ergebnisse zu erklären. Es geht um das Erlangen eines tieferen Verständnisses der zugrunde liegenden Mechanismen unter Berücksichtigung der verschiedenen Bedingungen bei der Erhebung von Daten zur Unterrichtsqualität. Bezugnehmend auf die in der vorliegenden Arbeit eingesetzten Items zur Erfassung von Unterrichtsstörungen ging es um eine Einschätzung des Verhaltens von Schülerinnen und Schülern innerhalb einer Klasse (Beispielitem: „Im Mathematikunterricht stört keiner den Unterricht“, siehe Fauth et al., 2014). Es ist denkbar, dass mit Verwendung einer anderen Skala oder gar eines anderen Instruments die Ergebnisse erwartungskonformer ausgefallen wären. Bislang gibt es aber noch keine richtungsweisenden Erkenntnisse, welche Skala bzw. welches Verfahren zur Erfassung von Unterrichtsstörungen (in welchem Kontext) zu präferieren ist (siehe auch Praetorius et al., 2018).

Fragebogen zur Erfassung der im Unterricht verfolgten Lernziele

Neben dem generischen Unterrichtsqualitätsmerkmal der Unterrichtsstörungen wurden auch die im Unterricht verfolgten Lernziele der Schülerinnen und Schüler mithilfe eines Lehrpersonen-Fragebogens erhoben. Der Fragebogen ist eng am St. Galler Bildungs- und Lehrplan orientiert und wurde zusammen mit Mathematik-Fachdidaktikern der PHSG entwickelt. Da die Datenerhebungen ca. alle zwei Wochen über einen Zeitraum eines gesamten Schuljahres stattfanden, ist der Datensatz sehr umfangreich und detailliert. Für die Analysen in der vorliegenden Arbeit wurden ausschließlich Daten zu den im Unterricht verfolgten Lernzielen einbezogen, die sich inhaltlich auf die Bruchrechnung beziehen und mit den Items (des Klassencockpit-Tests) zur Bruchrechnung in Beziehung gesetzt werden können. Trotz der sehr sorgfältigen Entwicklung des Instruments gingen mit der Verwendung des Fragebogens im Rahmen des INSE-Projekts einige Herausforderungen einher, die sich auf die Datenqualität ausgewirkt haben könnten. So war der Fragebogen sehr umfangreich und Lehrpersonen waren aufgefordert, mit einer Fülle an Informationen umzugehen. Um die Lehrpersonen hierbei zu

unterstützen, wurde zu Beginn des Fragebogens die Gliederung der thematischen Frageblöcke aufgezeigt. Die Lehrpersonen konnten sich dann durch den Fragebogen navigieren. Dabei mussten nur Angaben zu Lernzielen gemacht werden, die im Unterricht verfolgt wurden. Anderenfalls wäre das Ausfüllen des Fragebogens deutlich zeitaufwendiger geworden. Das Verfahren, Kästchen nur anzukreuzen, wenn das jeweilige Lernziel in den letzten ca. zwei Wochen relevant war, birgt die Gefahr, dass Eintragungen vergessen werden. Um Lehrpersonen bei der Zuordnung von Unterrichtsaktivitäten und den Lernzielen im Fragebogen zu unterstützen, konnten im Fragebogen zu jedem Lernziel Beispielaufgaben und Angaben zur Verortung im obligatorischen Lehrmittel abgerufen werden. Auch umgekehrt war diese Zuordnung möglich, also vom didaktischen Kommentar des Lehrmittels hin zum Lehrmittel und den Lernzielen im St. Galler Bildungs- und Lehrplan. Insofern wurde versucht, den mit diesem Instrument einhergehenden Herausforderungen bestmöglich entgegenzuwirken. Eine gewisse Fehleranfälligkeit blieb aber dennoch mit diesem Instrument bestehen. Andere Datenquellen wie Unterrichtsartefakte oder Videodaten weisen diesbezüglich Vorteile auf, bringen aber andere Herausforderungen mit sich (z.B. Rekrutierung der Stichprobe wird schwieriger) und konnten im Rahmen des INSE-Projekts nicht realisiert werden.

Klassencockpit-Testitems zur Erfassung der Entwicklung mathematischer Kompetenzen

Auch wenn die Klassencockpit-Testitems nach wissenschaftlichen Standards entwickelt wurden, sind diese nicht ohne Schwächen. Dies liegt auch daran, dass die Weiterentwicklung der Klassencockpit-Tests zugunsten der Entwicklung neuer, für den zukunftsweisenden Lehrplan 21 kompatibler Schulleistungstests eingestellt wurde. In Kapitel 6.5.1 wurde auf einzelne Mängel bei spezifischen Testitems hingewiesen. Aber auch in der Gesamtheit weist der Klassencockpit-Test für die fünfte Jahrgangsstufe Schwachstellen auf. So kann im Rahmen der kritischen Reflexion der Klassencockpit-Testitems die zentrale Frage gestellt werden, ob die Items prinzipiell in der Lage waren, Effekte von Unterricht zu erfassen. Die Ergebnisse zur globalen Sensitivität sprechen dafür, dass dies der Fall ist. Die Testitems differenzieren aber wenig zwischen Lerngruppen. Eine zentrale Ursache ist darin zu sehen, dass nach Anderson et al. (2001) die Klassencockpit-Testitems vermehrt auf den unteren drei Stufen (Erinnern, Verstehen und Anwenden) der Taxonomie kognitiver Lernziele und seltener auf den oberen drei Stufen (Analysieren, Evaluieren und Kreieren) zu verorten sind. Im Nachhinein ist dieser Aspekt unter Berücksichtigung des methodischen Ansatzes als nicht ideal zu bewerten.

Bezüglich der Bearbeitung der Klassencockpit-Testitems ist zusätzlich noch das Auftreten möglicher Testungseffekte anzumerken (Roediger & Karpicke, 2006a, 2006b; Roediger & Marsh, 2005). Roediger und Karpicke (2006a, 2006b) zeigen auf, dass in verschiedenen Settings (z.B. Erinnern von Bildern oder Wörtern) ein wiederholtes Testen dazu führt, dass der Lerngegenstand auch zu einem späteren Zeitpunkt besser erinnert werden kann. Im Rahmen des INSE-Projekts haben Schülerinnen und Schüler zu mehreren Messzeitpunkten die gleichen Items bearbeitet, wobei i.d.R. zwei Testungen vor und eine nach der Vermittlung der Testinhalte stattfand. Roediger und Marsh (2005) legen anhand eines Multiple-Choice-Tests zum Leseverstehen im Fach Englisch dar, dass Schülerinnen und Schüler von wiederholten Testungen lernen können, auch wenn die Inhalte noch nicht vermittelt wurden. Erhalten Schülerinnen und Schüler keine detaillierte Rückmeldung zu den Testergebnissen, besteht die Gefahr, dass Misskonzepte unerkannt bleiben und infolgedessen nicht korrigiert werden (Roediger & Marsh, 2005). Im INSE-Projekt wurde eine detaillierte Rückmeldung nur zum Zeitpunkt des (intendierten) Posttests gegeben, also dem letzten Messzeitpunkt (*t4*). Dementsprechend bleibt festzuhalten, dass positive und/oder negative Testungseffekte im Rahmen des INSE-Projekts möglich sind und Veränderungen in den Itemenschwierigkeiten nicht nur mit dem Unterricht, sondern auch mit einer wiederholten Testung im Zusammenhang stehen könnten. Hierin ist eine Limitation der vorliegenden Arbeit zu sehen, welche sich in Anbetracht der Rahmenbedingungen (Stichprobengröße, begrenzter Itempool) des INSE-Projekts nicht vermeiden ließ. Für zukünftige Forschungsprojekte wäre es vorteilhaft, ein Matrixdesign zu verwenden, sodass Schülerinnen und Schüler zu den unterschiedlichen Messzeitpunkten verschiedene Items bearbeiten. Mittels längsschnittlicher Skalierung könnten dann die geschätzten Werte auf einer gemeinsamen Skala verortet werden, sodass Aussagen zum Kompetenzniveau der Schülerinnen und Schüler möglich werden würden (Hartig, 2007; Hartig & Kühnbach, 2006).

7 Ausblick

Nachfolgend wird ein Ausblick über Implikationen gegeben, die aus dieser Arbeit resultieren. In Kapitel 7.1 geht es zunächst um die Implikationen für die Praxis. Im Fokus steht die Verantwortung von Beteiligten, um eine valide Testwertnutzung und -interpretation zu gewährleisten. Anschließend werden in Kapitel 7.2 die Implikationen für die weitere Forschung thematisiert. Zwei Aspekte werden in diesem Zusammenhang aufgegriffen: (1) Die Frage, ob im Kontext der Evaluation der Instruktionssensitivität auch Moderations- und Mediationsannahmen eine Rolle spielen könnten und (2) die Frage nach der Modellierung von Veränderungswerten. Diese zwei genannten Aspekte könnten dazu beitragen, die Evaluation der Instruktionssensitivität maßgeblich weiterzuentwickeln.

7.1 Implikationen für die Praxis

Leistungstestdaten von Schülerinnen und Schülern bieten ein großes Potenzial zur Gewinnung von Informationen über Schule und Unterricht. Der in der vorliegenden Arbeit herangezogene Klassencockpit-Test wird gemäß einer Umfrage (1) zur Standortbestimmung der Klasse, (2) zur Objektivierung der Beurteilung, (3) zur Information über Leistungen der Schülerinnen und Schüler, (4) zur Überprüfung der Zielerreichung des Lehrplans, (5) zur Optimierung des Unterrichts und (6) zur Anpassung der Beurteilung verwendet (Moser, 2003). Deutlich wird anhand dieser von Moser (2003) publizierten Umfrageergebnisse, dass die Testdaten der Schülerinnen und Schüler aus dem Klassencockpit-Test u.a. dazu verwendet werden, um Rückschlüsse über den Unterricht zu ziehen (z.B. zur Optimierung des Unterrichts). Die Sicherstellung der Instruktionssensitivität von Tests bzw. Testitems durch die Entwicklerinnen und Entwickler ist dabei von hoher Bedeutung, da sich Personen aus der Forschung, der Bildungspolitik und der pädagogischen Praxis z.T. darauf verlassen, dass die jeweiligen Tests bzw. die jeweiligen Testitems in spezifischen Settings funktionieren. Um eine valide Testwertnutzung und -interpretation sicherzustellen (Kane, 2013), stehen aber nicht nur die Entwicklerinnen und Entwickler, sondern auch die Anwenderinnen und Anwender (AERA, APA & NCME, 2014; Hartig, Frey & Jude, 2020) sowie die Nutzerinnen und Nutzer der gewonnenen bzw. bereitgestellten Testdaten in der Verantwortung. Sind sie hinsichtlich einer validen Testwertnutzung und -interpretation nicht ausreichend geschult, kann es zu Fehlinterpretationen kommen. Ein Beispiel hierfür wäre, wenn Nutzerinnen und Nutzer die

Testdaten standardisierter Schulleistungstests als etwas betrachten, das einzig auf den Unterricht zurückgeführt werden kann und weitere potenzielle Einflussfaktoren unberücksichtigt bleiben. Vor diesem Hintergrund ist ein umsichtiger Umgang mit Tests, Testresultaten und deren Interpretation von zentraler Bedeutung. Es wäre wertvoll, wenn Personen, die Testdaten standardisierter Schulleistungstests nutzen, um Rückschlüsse über Schule und Unterricht zu ziehen, ein Verständnis hinsichtlich des Konzepts der Instruktionssensitivität haben. Eine Ausweitung des Bekanntheitsgrads des Konzepts der Instruktionssensitivität wäre in diesem Zusammenhang ein wichtiges Ziel. Dies schließt auch den Aufbau eines Bewusstseins ein, dass zum einen nicht jeder Test bzw. jedes Testitems per se instruktionssensitiv ist, und zum anderen, dass nicht jeder Test bzw. jedes Testitem instruktionssensitiv sein muss. Wünschenswert wären zudem Kompetenzen im Hinblick auf einen adäquaten Umgang mit Daten aus standardisierten Schulleistungstests. In Anbetracht der unterschiedlichen Berufsgruppen (Bildungsforschung, Bildungspolitik, pädagogische Praxis), die Testdaten nutzen, um Rückschlüsse über Schule und Unterricht zu ziehen, kann Letzteres nicht als eine Selbstverständlichkeit betrachtet werden.

Selbstkritisch ist an dieser Stelle anzumerken, dass die Thematik der Instruktionssensitivität von Tests und Testitems für Personen außerhalb der wissenschaftlichen Community relativ schwer zugänglich ist. Hintergrund ist, dass bisherige Publikationen zumeist sehr methodisch ausgerichtet und abstrakt sind. Hinzu kommt, dass die Beiträge vornehmlich in internationalen Fachzeitschriften publiziert werden. Um das Konzept der Instruktionssensitivität zugänglicher zu machen, müssten andere Formate mit dem Ziel eines Transfers zwischen Wissenschaft und Praxis (z.B. Weiterbildungen, praxisorientierte Tagungen, Literatur) bewusst angesteuert und Beiträge in einer auch für Personen anderer Berufsgruppen verständlichen Sprache bereitgestellt werden.

7.2 Implikationen für die Forschung

Bezugnehmend auf die Implikationen für die Forschung bietet die Evaluation der Instruktionssensitivität von Tests und Testitems ein großes Potenzial zur weiteren konzeptionellen und methodischen Ausdifferenzierung. Vor diesem Hintergrund wird auf die Einbeziehung lernrelevanter Merkmale als Mediator oder Moderator eingegangen. Anschließend geht es um die Modellspezifikation im Zusammenhang mit der Bildung von Veränderungswerten. Letzteres ist der Tatsache geschuldet, dass Studien nicht nur ein *Pretest-*

Posttest-Design, sondern zudem einen weiteren Messzeitpunkt (z.B. *Follow-up-Test*) vorsehen können. Dadurch eröffnen sich neue Fragen im Zusammenhang mit der Evaluation der Instruktionssensitivität.

7.2.1 Instruktionssensitivität: Kovariate, Moderation & Mediation

Aus Kapitel 2.5 geht hervor, dass bei der Evaluation der Instruktionssensitivität zum Teil lernrelevante Merkmale von Schülerinnen und Schülern als Kovariaten einbezogen werden. Die Kovariate steht für eine Variable, die konstant gehalten wird, um unter gleichbleibenden Bedingungen den Einfluss von einer unabhängigen auf eine abhängige Variable zu messen (Bortz & Schuster, 2010). Bezugnehmend auf die Ausführungen in Kapitel 3 zu den Annahmen zur Beziehung zwischen Unterrichtsmerkmalen, lernrelevanten Merkmalen von Schülerinnen und Schülern und deren Leistungen (Brühwiler & Blatchford, 2014; Seidel, 2014) kann die Frage gestellt werden, ob Moderations- und/oder Mediationseffekte (Baron & Kenny, 1986; Hayes, 2018) im Kontext der Instruktionssensitivität ebenfalls eine Rolle spielen könnten.

Mediationsannahme

Um eine Idee von Mediationsannahmen im Kontext der Evaluation der Instruktionssensitivität zu erhalten, soll hier ein Beispiel zum unterstützenden Klassenklima herangezogen werden. Angenommen wird, dass ein unterstützendes Klassenklima – vermittelt über die Qualität der Lernmotivation – sich positiv auf die schulischen Leistungen der Schülerinnen und Schüler auswirkt (Klieme et al., 2006). Um diesen Effekt zu überprüfen, wird – rein hypothetisch gedacht – ein neues Instrument entwickelt. Bevor dieses im Rahmen einer Unterrichtseffektivitätsstudie eingesetzt wird, soll geprüft werden, ob das Instrument auch in der Lage ist, Effekte des unterstützenden Klassenklimas – vermittelt über die Qualität der Lernmotivation – auf die Leistungen von Schülerinnen und Schülern abzubilden. Im dargelegten Beispiel wäre anzunehmen, dass das unterstützende Klassenklima keinen direkten Effekt auf die Leistungen der Schülerinnen und Schüler bzw. die Indikatoren der Instruktionssensitivität nimmt. Es erscheint daher sinnvoll, die Qualität der Lernmotivation in die Analysen zur Evaluation der Instruktionssensitivität mit einzubeziehen.[34]

[34] Eine Analyse mit der Qualität der Lernmotivation als Mediator konnte im Rahmen dieser Arbeit nicht umgesetzt werden, da die Daten zur Qualität der Lernmotivation auf die Ausgangsvoraussetzungen (zu *t1* erhoben)

In den Untersuchungen zur Evaluation der Instruktionssensitivität könnten folglich das unterstützende Klassenklima als Prädiktor, die Qualität der Lernmotivation als Mediator und die klassenspezifischen Itemschwierigkeiten als Kriterium einbezogen werden (siehe Klieme et al., 2006). Ein solches Vorgehen hätte den Vorteil, dass genaue Informationen gewonnen werden könnten, inwiefern die Indikatoren der Instruktionssensitivität auf das unterstützende Klassenklima sowie die Qualität der Lernmotivation sensitiv reagieren. Zum einen könnte dadurch – im Vergleich zu *reinen Itemstatistiken* – der Nachteil umgangen werden, dass eine gewisse Unsicherheit bestehen bleibt, ob sich die Veränderung in den Itemschwierigkeiten auch tatsächlich auf den Unterricht zurückführen lässt (Kapitel 2.4.2). Zum anderen kann bei der Vorgehensweise mittels *Itemstatistiken unter Berücksichtigung von Unterrichtsmerkmalen* die Problematik vermieden werden, dass ein Unterrichtsmerkmal als Prädiktor einbezogen wird, wenngleich kein direkter Effekt auf den Indikator der Instruktionssensitivität zu erwarten ist (Kapitel 2.4.2). Die Einbeziehung von Mediationsannahmen im Kontext der Evaluation der Instruktionssensitivität könnte daher in spezifischen Fällen gewinnbringend sein. Im hier genannten Beispiel müsste sich die Evaluation der Instruktionssensitivität nicht nur auf leistungsbezogene Testitems beziehen (Klassencockpit-Testitems), sondern würde um nicht kognitive Testitems (z.B. Qualität der Lernmotivation) erweitert werden.

Moderationsannahme

Um eine Idee von Moderationsannahmen (Baron & Kenny, 1986; Sharma, Durand & Gur-Arie, 1981) im Kontext der Evaluation der Instruktionssensitivität zu erhalten, soll in diesem Zusammenhang das *Framework* zur *Aptitude-Treatment-Interaction* (ATI) aufgegriffen werden. Das *ATI-Framework* wird Studien zugrunde gelegt, in denen es um die Überprüfung der Wirksamkeit von *Treatments* (z.B. Unterricht) unter Berücksichtigung von *Aptitudes* (individuellen Lernvoraussetzungen) geht. Es ist ein relationales Konstrukt, mit dem die Passung zwischen Situations- und Personenmerkmal fokussiert und basierend darauf individuelles Lernen erklärt wird (Snow & Cronbach, 1977; Snow, 1991). Mit dem *ATI-Framework* wird die Annahme vertreten, dass Personen mit unterschiedlichen Merkmalsausprägungen von unterschiedlichen *Treatments* (z.B. Unterricht) profitieren (Snow, 1991). Letzteres kann im Rahmen einer Moderationsannahme überprüft werden.

fokussieren, während sich die Daten zur Unterrichtsqualität auf einen späteren Zeitpunkt (zu *t4* erhoben) beziehen. Die Leistungstestdaten liegen sowohl zu *t1* als auch zu *t4* vor.

Als Beispiels kann auf Fuchs et al. (2014) verwiesen werden, denen der Nachweis der ATI im Zusammenhang mit der Vermittlung von Bruchrechnung gelang. Anhand ihrer Studie konnte gezeigt werden, dass für Schülerinnen und Schüler mit ungünstiger Ausprägung bezogen auf das Arbeitsgedächtnis sich eine ergänzende Unterrichtsintervention als gewinnbringend erweist, welche auf ein konzeptionelles Verstehen (*conceptual practise*) abzielt. Schülerinnen und Schüler mit einer günstigen Ausprägung bezogen auf das Arbeitsgedächtnis profitieren von einer ergänzenden Unterrichtsintervention, welche das prozedurale Wissen stärkt und zur Automatisierung mathematischer Verfahren dient (ebd.).

Übertragen auf den Kontext der Evaluation der Instruktionssensitivität kann das *Framework* wertvolle Ideen für mögliche Interaktionen zwischen Merkmalen der Schülerinnen und Schüler und Unterrichtsmerkmalen sowie deren Beziehung zu Leistungstestdaten der Schülerinnen und Schüler hervorbringen. So wäre unter Berücksichtigung der ATI davon auszugehen, dass Unterricht nicht für alle Schülerinnen und Schüler innerhalb einer Klasse gleichermaßen wirksam ist, da deren Merkmalsausprägungen untereinander variieren. In den bisherigen Studien zur Evaluation der Instruktionssensitivität wird diesem Gedanken (noch) keine Berücksichtigung geschenkt. Insofern liefert das *ATI-Framework* eine interessante Perspektive, welche der konzeptionellen und methodischen Erweiterung der Evaluation der Instruktionssensitivität dienen könnte. Die Berücksichtigung komplexer Wirkbeziehungen wird dadurch machbar.

Deutlich wird damit: Die hier aufgezeigten Beispiele bieten die Möglichkeit, das Konzept der Instruktionssensitivität zukünftig – unter Berücksichtigung aktueller Erkenntnisse aus der Schul- und Unterrichtsforschung – differenzierter zu betrachten. Es ist zu überlegen, auf was der Test oder das Testitem sensitiv reagieren soll. Die Einbeziehung von Prädiktoren, ggf. auch von Moderatoren oder Mediatoren, ist in Betracht zu ziehen. Darüber hinaus ist zu prüfen, ob Kovariaten in das Analysemodell mit aufzunehmen sind, um möglichen Störeinflüssen entgegenzuwirken.

7.2.2 Instruktionssensitivität: Modellspezifizierung bezogen auf Veränderungswerte

In Kapitel 4.5.1.1 wurde auf die Modellspezifizierung hinsichtlich der Veränderungswerte zu den klassenspezifischen Itemschwierigkeiten eingegangen. Hierbei wurde auf Steyer et al. (1997) Bezug genommen, die zwischen zwei Typen unterscheiden. Während beim Typ I

Veränderungswerte immer mit Bezug auf den Ausgangswert berechnet werden (z.B. *t2–t1*, *t3–t1*, *t4–t1*), ist für den Typ II charakteristisch, Differenzwerte mit Bezug auf den jeweils nächstgelegenen Messzeitpunkt zu analysieren (*t2–t1*, *t3–t2*, *t4–t3*). Steyer et al. (1997) schlagen diese zwei Typen im Zusammenhang mit Untersuchungen interindividueller Unterschiede durch intraindividuelle Veränderungen vor. Auf Modelle zur Schätzung von gruppenspezifischen Veränderungswerten wird nicht eingegangen. Die Gedanken von Steyer et al. (1997) sind aber dennoch bezogen auf das INSE-Projekt interessant, da mit dem Projekt – im Vergleich zu anderen Instruktionssensitivitätsprojekten – die Besonderheit einherging, dass längsschnittliche Daten von mehr als zwei Messzeitpunkten vorlagen (Kapitel 4.1). Entsprechend stellte sich die Frage, wie die Veränderungswerte sinnvollerweise zu bilden sind (Kapitel 4.5.1). Es kann sich leicht vorgestellt werden, dass die Ergebnisse je nach Modellspezifizierung (Typ I oder Typ II) sehr unterschiedlich ausfallen (könnten).

Im INSE-Projekt wurden die Erhebungszeitpunkte so gewählt, dass es sich zweimal um einen Pretest und einmal um einem Posttest handeln sollte. Wie bereits in Kapitel 4.5.1.1 dargelegt, ist jedoch damit zu rechnen, dass Lehrpersonen vom Stoffverteilungsplan abweichen, da dieser lediglich einen Empfehlungscharakter hat. Letzteres ist bei der Modellierung von Veränderungswerten mitzudenken. Gleiches gilt für die Frage nach dem Zeitraum, in dem ein Lernzuwachs zu erwarten ist, und inwiefern ein Interesse an der Sensitivität von Tests bzw. Testitems für Kurzzeit- und/oder Langzeiteffekte besteht. In Anbetracht dieser genannten Gesichtspunkte wurde sich im INSE-Projekt für die Typ-I-Modellierung entschieden.

Ein anderes Beispiel für die Modellierung von Veränderungswerten mit mehr als zwei Messzeitpunkte stellen Forschungsvorhaben mit einem *Pretest-*, *Posttest-* und *Follow-up-Test-Design* dar. Die Gedanken hierzu wurden bereits im Beitrag von Naumann et al. (2020) dargelegt, sodass auf eine erneute Ausführung verzichtet wird. Generell bleibt aber festzuhalten, dass ein Bewusstsein für die Thematik hinsichtlich der Bildung von Veränderungswerten im Kontext der Instruktionssensitivität insbesondere in Anbetracht möglicher Konsequenzen wünschenswert ist. Aus diesem Grund wurden die Überlegungen zur Modellierung hier aufgezeigt.

8 Schlusswort

Im Rahmen dieser Arbeit wurde überprüft, inwiefern die Qualität der Lernmotivation einen Einfluss auf die Indikatoren der Instruktionssensitivität nimmt. Das Ergebnis fiel entgegen den Erwartungen aus, was jedoch nicht bedeutet, dass die Qualität der Lernmotivation fortan und generell ohne Bedenken bei der Evaluation der Instruktionssensitivität von Tests oder Testitems unberücksichtigt bleiben kann. Weitere Forschung ist diesbezüglich notwendig, um die Erkenntnisse dieser Arbeit auch in Anbetracht der genannten Limitationen zu überprüfen.

Darüber hinaus bleibt festzuhalten, dass auch wenn das Ergebnis – trotz einer vielversprechend angelegten Studie mit innovativer Methode – eher ernüchternd ausfiel, im Rahmen dieser Dissertation wichtige Erkenntnisse gewonnen wurden. Hierzu zählen insbesondere die konzeptionellen Überlegungen, um die Instruktionssensitivität von Tests und Testitems unter Berücksichtigung individueller Lernvoraussetzungen von Schülerinnen und Schülern zu evaluieren. Es erscheint dabei spannend und zielführend, bei zukünftigen Forschungsarbeiten über die motivationalen Merkmale hinaus den Einfluss weiterer lernrelevanter Merkmale von Schülerinnen und Schülern (z.B. Vorwissen, Intelligenz, SES) auf die Indikatoren der Instruktionssensitivität hin zu überprüfen. Dies würde ein umfassendes Bild ermöglichen, um evidenzbasiert entscheiden zu können, inwiefern lernrelevante Merkmale unter Berücksichtigung spezifischer Fragestellungen im Kontext der Instruktionssensitivität von Tests und Testitems einzubeziehen sind. Des Weiteren sei auch auf die Erkenntnisse hinsichtlich methodischer Erweiterungsmöglichkeiten verwiesen, die im Rahmen dieser Arbeit herausgearbeitet wurden. Das LMLIRT bzw. eLMLIRT-Modell hat den Vorteil, dass es mit wenig Aufwand modifiziert werden kann und dadurch vielfältige Möglichkeiten zur Einbeziehung von lernrelevanten Merkmalen und/oder den Unterrichtsmerkmalen im Kontext der Instruktionssensitivität zur Verfügung stehen. Daher können beispielsweise Moderations- oder auch Mediationsannahmen ohne Weiteres umgesetzt werden. Insofern liefern die methodischen zusammen mit den konzeptionellen Überlegungen wertvolle Anregungen, um zukünftig und in Anbetracht komplexer Wirkbeziehungen neue Wege bei der Evaluation der Instruktionssensitivität von Tests und Testitems zu gehen.

Rückblickend ist bezüglich der Thematik der vorliegenden Arbeit noch anzumerken, dass diese auf die eingebrachte erziehungswissenschaftliche Perspektive zurückzuführen ist (u.a. geprägt von komplexen Modellen zur Erklärung schulischer Leistungen). Die bisherige Ausrichtung unseres, aber auch vieler anderer Forscherteams orientiert sich (häufig) stark an Gedanken, die

aus der Psychometrie stammen. Die interdisziplinäre Zusammenarbeit hat in diesem Fall eine neue Fragestellung hervorgebracht, der im Rahmen dieser Arbeit nachgegangen wurde. Das Zusammenwirken von Personen aus den Erziehungswissenschaften, der Psychologie sowie den Fachdidaktiken gestaltete sich aber auch vor dem Hintergrund der Erklärung der Instruktionssensitivität von Tests oder Testitems als äußerst wertvoll. Bisher steht bei einem Großteil der Studien der Evaluationsgedanke stark im Vordergrund. Mit dem Ziel, zukünftig instruktionssensitive Tests oder Testitems zu konstruieren, braucht es jedoch mehr als eine Evaluation. Es geht um das Verstehen, warum Tests oder Testitems sensitiv bzw. insensitiv sind. Eine verstärkte interdisziplinäre Zusammenarbeit von Personen verschiedener Fachdisziplinen ist anzustreben, um die Forschung zur Instruktionssensitivität von Tests und Testitems diesbezüglich voranzutreiben.

9 Literaturverzeichnis

Airasian, P. W. & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*(2), 103–118.

Altrichter, H. & Maag Merki, K. (Hrsg.). (2016). *Handbuch Neue Steuerung im Schulsystem* (2. Aufl.). Wiesbaden: Springer. doi: 10.1007/978-3-531-18942-0_10

Altrichter, H., Moosbrugger, R. & Zuber, J. (2016). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (2. Aufl., S. 235–277). Wiesbaden: Springer.

American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME] (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Amrein, A. L. & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1–74. doi: https://doi.org/10.14507/epaa .v10n18.2002

Anderson, J. R. (1980). *Cognitive psychology and its implications*. San Francisco: W. H. Freeman & Co Ltd.

Anderson, J.R. (1982). Acquisition of cognitive skill. *Psychological review*, *89*(4), 369–406. doi: https://doi.org/10.1037/0033-295X.89.4.369

Anderson, J.R. (1995). *Cognitive psychology and its implications* (5. Aufl.). New York: Wroth.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K., A. , Mayer, R. E., Pintrich, P. R., . . . Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Assor, A., Kaplan, H. & Roth, G. (2002). Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork. *British Journal of Educational Psychology, 72*(2), 261–278. doi: https://doi.org/10.1348/000709902158883

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological review, 64*(6 Pt. 1), 359–372. doi: https://doi.org/10.1037/h0043445

Ausubel, D. P., Novak, J. D. & Hanesian, H. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart and Winston

Bach, A., Wurster, S., Thillmann, K., Pant, H. A. & Thiel, F. (2014). Vergleichsarbeiten und schulische Personalentwicklung – Ausmaß und Voraussetzungen der Datennutzung. *Zeitschrift für Erziehungswissenschaft, 17*(1), 61–84. doi: https://doi.org/10.1007/ s11618-014-0486-5

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (14. Aufl.). Berlin & Heidelberg: Springer Gabler. doi: 10.1007/978-3-662-46076-4

Baker, E. L. (1994). Making performance assessment work: The road ahead. *Educational Leadership, 51*(6), 58–62.

Baker, S. R. (2003). A prospective longitudinal investigation of social problem-solving appraisals on adjustment to university, stress, health, and academic motivation and performance. *Personality and Individual Differences, 35*(3), 569–591. doi: https://doi.org/10.1016/S0191-8869(02)00220-9

Baron, R. M. & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. doi: 10.1037//0022-3514.51.6.1173

Bartholomew, K. J., Ntoumanis, N., Mouratidis, A., Katartzi, E., Thøgersen-Ntoumani, C. & Vlachopoulos, S. (2018). Beware of your teaching style: A school-year long investigation of controlling teaching and student motivational experiences. *Learning and Instruction, 53*, 50–63. doi: https://doi.org/10.1016/j.learninstruc.2017.07.006

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi: 10.18637/jss.v067.i01

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462. doi: https://doi.org/10.1007/BF03173192

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., . . . Neubrand, J. (1997). *TIMSS. Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde*. Opladen: Leske+ Budrich. doi: 10.1007/978-3-322-95096-3

Bieg, M., Goetz, T., Sticca, F., Brunner, E., Becker, E., Morger, V. & Hubbard, K. (2017). Teaching methods and their impact on students' emotions in mathematics: an experience-sampling approach. *ZDM - Mathematics Education, 49*, 411–422. doi: https://doi.org/10.1007/s11858-017-0840-1

Binder, T., Schmiemann, P., Theyßen, H., Sandmann, A. & Sures, B. (2016). Fachspezifisches Vorwissen und Studienerfolg in Biologie und Physik. In C. Maurer (Hrsg.), *Authentizität und Lernen - das Fach in der Fachdidaktik. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Berlin 2015* (S. 395–397). Regensburg: Universität Regensburg.

Black, A. E. & Deci, E. L. (2000). The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective. *Science Education, 84*(6), 740–756. doi: https://doi.org/10.1002/1098-237X(200011)84:6<740::AID-SCE4>3.0.CO;2-3

Bloom, B. S. (1976). Human characteristics and school learning. In R. E. Slavin (Hrsg.), *Educational psychology: Theory and practice* (S. 285–287). Boston: Allyn & Bacon.

Boekaerts, M. (2007). What have we learned about the link between motivation and learning/performance? *Zeitschrift für Pädagogische Psychologie, 21*(3/4), 263–269. doi: https://doi.org/10.1024/1010-0652.21.3.263

Bohl, T. & Kiper, H. (Hrsg.) (2009). *Lernen aus Evaluationsergebnissen – Verbesserungen planen und implementieren*. Bad Heilbrunn: Klinkhardt.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons. doi: 10.1002/9781118619179

Bollen, K. A. & Long, J. S. (Hrsg.) (1993). *Testing structural equation models*. Newbury Park u.a.: Sage.

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch model. Fundamental measurement in the human sciences* (3. Aufl.). New York: Routledge.

Bortz, J. & Schuster, C. (2010). *Statistik für Human-und Sozialwissenschaftler* (7. Aufl.). Berlin & Heidelberg: Springer.

Brennan, R. L. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement, 32*(2), 289–303. doi: https://doi.org/10.1177/001316447203200206

Brennan, R. L. & Stolurow, L. M. (1971, February). *An elementary decision process for formative evaluation of an instructional system*. Referat anlässlich der Jahrestagung der American Educational Research Association, New York, NY. Verfügbar unter: https://eric.ed.gov/?id=ED048343

Brewer, D. J. & Stasz, C. (1996). Enhancing opportunity to learn measures in NCES data. In E. G. Hochlander, J. E. Griffith & J. H. Ralph (Hrsg.), *From data to information: New*

directions for the National Center for Education Statistics (S. 1–28). Washington, DC: US Department of Education.

Bromme, R., Prenzel, M. & Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Zeitschrift für Erziehungswissenschaft, 17*(4), 3–54. doi: https://doi.org/10.1007/s11618-014-0514-5

Brühwiler, C. (2014). *Adaptive Lehrkompetenz und schulisches Lernen. Effekte handlungssteuernder Kognitionen von Lehrpersonen auf Unterrichtsprozesse und Lernergebnisse der Schülerinnen und Schüler*. Münster: Waxmann.

Brühwiler, C. & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21*(1), 95–108. doi:10.1016/j.learninstruc.2009.11.004

Brühwiler, C., Helmke, A. & Schrader, F.-W. (2017). Determinanten der Schulleistung. In M. H. W. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion* (3. Aufl., S. 291–314). Wiesbaden: Springer.

Burstein, L. (1989). *Conceptual Considerations in Instructionally Sensitive Assessment*. Referat anlässlich der Jahrestagung der American Education Research Association, San Francisco, CA. Verfügbar unter: https://files.eric.ed.gov/fulltext/ED341737.pdf

Burton, K. D., Lydon, J. E., D'Alessandro, D. U. & Koestner, R. (2006). The differential effects of intrinsic and identified motivation on well-being and performance: prospective, experimental, and implicit approaches to self-determination theory. *Journal of Personality and Social Psychology, 91*(4), 750–762. doi: https://doi.org/10.1037/0022-3514.91.4.750

Carmines, E. G. & McIver, J. P. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. W. Bohrnstedt & E. F. Borgatta (Hrsg.), *Social measurement: Current issues* (S. 65–115). Beverly Hills & London: Sage.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723–733.

Carroll, J. B. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher*, *18*(1), 26–31. doi: https://doi.org/10.3102/0013189X018001026

Chatzisarantis, N. L., Hagger, M. S., Biddle, S. J., Smith, B. & Wang, J. C. (2003). A meta-analysis of perceived locus of causality in exercise, sport, and physical education contexts. *Journal of Sport and Exercise Psychology, 25*(3), 284–306. doi: https://doi.org/10.1123/jsep.25.3.284

Chen, J. (2012). *Impact of instructional sensitivity on high-stakes achievement test items: A comparison of methods* (Doktorarbeit), University of Kansas, Lawrence, USA. Verfügbar unter: http://hdl.handle.net/1808/10132

Clausen, M. (2002). Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität. Münster u.a.: Waxmann.

Cokley, K. O., Bernard, N., Cunningham, D. & Motoike, J. (2001). A psychometric investigation of the academic motivation scale using a United States sample. *Measurement and Evaluation in Counseling and Development, 34*(2), 109–119. doi: https://doi.org/10.1080/07481756.2001.12069027

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D. & York, R. L. (1966). *Equality of educational opportunity (No. FS 5.238:38001)*. Washington, DC, USA: U.S. Government Printing Office. Verfügbar unter: https://files.eric.ed.gov/fulltext/ED012275.pdf

Cox, R. C. & Vargas, J. S. (1966, February). *A comparison of item selection techniques for norm-referenced and criterion-referenced tests*. Referat anlässlich der Jahrestagung des National Council on Measurement in Education, Chicago, IL. Verfügbar unter: https://files.eric.ed.gov/fulltext/ED010517.pdf

Creemers, B. M. & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London & New York: Routledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. doi: https://doi.org/10.1007/BF02310555

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

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row.

D'Agostino, J. V., Welsh, M. E. & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*(1), 1–22. doi: https://doi.org/10.1080/10627190709336945

d'Ailly, H. (2003). Children's autonomy and perceived control in learning: A model of motivation and achievement in Taiwan. *Journal of Educational Psychology, 95*(1), 84–96. doi: https://doi.org/10.1037/0022-0663.95.1.84

Debra P v. R. D. Turlington, 474 F. Supp. 244 (Civ. T. C. 1979), aff'd, 730 F.2d 1405 (U.S. Ct. App. 11th Cir. 1984).

Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik, 39*(2), 223–238.

Deci, E. L. & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology/Psychologie canadienne, 49*(3), 182–185. doi: https://doi.org/10.1037/a0012801

Deci, E. L., Vallerand, R. J., Pelletier, L. G. & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*(3-4), 325–346. doi: https://doi.org/10.1080/00461520.1991.9653137

Decristan, J., Hertel, S., Klieme, E., (Projektleitung DIPF), Büttner, G., Hardy, I., . . . (Projektleitung Goethe-Universität). (2014). *Projekt IGEL: Individuelle Förderung und adaptive Lern-Gelegenheiten in der Grundschule. Unveröffentlichtes Skalenbuch der IGEL-Studie*. IDeA-Forschungszentrum. Frankfurt a. M.

Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M., . . . Hardy, I. (2015). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research, 108*(5), 358–370. doi: https://doi.org/10.1080/00220671.2014.899957

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., . . . Hardy, I. (2015). Embedded formative assessment and classroom process quality. How do they interact in promoting science understanding? *American Educational Research Journal, 52*(6), 1133–1159. doi:10.3102/0002831215596412

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software, 71*(9), 1–25. doi: 10.18637/jss.v071.i09

Deutscher, V. & Winther, E. (2018). Instructional sensitivity in vocational education. *Learning and Instruction, 53*, 21–33. doi: https://doi.org/10.1016/j.learninstruc.2017.07.004

Deutsches Zentrum für Lehrkräftebildung Mathematik [DZLM] (o. D.). Bauernhofaufgaben. Verfügbar unter: https://kira.dzlm.de/probleml%C3%B6sen-co/prozessbezogene-kompetenzen-f%C3%B6rdern/gr%C3%B6%C3%9Fen-und-messen/bauernhofaufgaben, Abruf: 22.12.2022

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics, 42*(1), 204–223. doi:10.1214/aoms/1177693507

Dochy, F. J. R. C. (1988a). *The" prior knowledge state" of students and its facilitating effect on learning: Theories and research (OTIC-RR-1.2)*. Heerlen, NL: Open University, Centre for Educational Technological Innovation. Verfügbar unter: https://files.eric.ed.gov /fulltext/ED387486.pdf

Dochy, F. J. R. C. (1988b). *Variables influencing the indexation of the" prior knowledge state" concept and a conceptual model for research (OTIC-RR-2.2)*. Heerlen, NL: Open University, Centre for Educational Technological Innovation. Verfügbar unter: https://eric.ed.gov/?id=ED387487

Dochy, F. J. R. C. (1992). *Assessment of prior knowledge as a determinant for future learning: The use of prior knowledge state tests and knowledge profiles*. Utrecht: Lemma B. V.

Dochy, F. J. R. C. (1996). Assessment of domain-specific and domain-transcending prior knowledge: Entry assessment and the use of profile analysis. In M. Birenbaum & F. J. R. C. Dochy (Hrsg.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (S. 227–264). Boston: Kluwer

Drechsel, B., Prenzel, M. & Seidel, T. (2015). Nationale und internationale Schulleistungsstudien. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (2. Aufl., S. 343–368). Berlin & Heidelberg: Springer. doi: https://doi.org/10.1007/978-3-642-41291-2_15

Dreher, A. & Leuders, T. (2021). Fachspezifität von Unterrichtsqualität – aus der Perspektive der Mathematikdidaktik. *Unterrichtswissenschaft 49*, 285–292. doi: https://doi.org/10 .1007/ s42010-021-00116-9

Dresel, M., Backes, C. & Lämmle, L. (2011). Zur Bedeutung von Motivation und Selbstregulation für Leistungen im durchschnittlichen und im exzellenten Bereich: Eine Einführung. In M. Dresel & L. Lämmle (Hrsg.), *Motivation, Selbstregulation und Leistungsexzellenz* (S. 1–10). Münster: LIT.

Dresel, M. & Lämmle, L. (2011). Motivation. In T. Götz (Hrsg.), *Emotion, Motivation und selbstreguliertes Lernen* (Bd. 3481, S. 79–142). Paderborn: Ferdinand Schöningh.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040–1048. doi: https://doi.org/10.1037/0003-066X.41.10.1040

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement, 22*(1), 15–25. doi: https://doi.org/10.1177/001316446202200103

Eccles, J. S. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Hrsg.), *Achievement and Achievement Motives: Psychological and Sociological Approaches* (S. 75–146). San Francisco: W. H. Freeman.

Eid, M., Gollwitzer, M. & Schmitt, M. (2015). *Statistik und Forschungsmethoden* (4. Aufl). Weinheim & Basel: Beltz.

Elliott, S. N. (2015). Measuring opportunity to learn and achievement growth: Key research issues with implications for the effective education of all students. *Remedial and Special Education, 36*(1), 58–64. doi: https://doi.org/10.1177/0741932514551282

Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Hrsg.), *Handbook of causal analysis for social research* (S. 245–273). Dodrecht: Springer. doi: 10.1007/978-94-007-6094-3 13

Fauth, B. (2021, Juni). *Merkmale und Erfassung von Unterrichtsqualität*. Referat anlässlich der internen Weiterbildung an der Pädagogischen Hochschule St. Gallen, St. Gallen.

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9. doi: http://dx.doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K. & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different

perspectives. *Zeitschrift für Pädagogik. Beiheft, 66*(1), 138–155. doi: https://doi.org/10.3262/ZPB2001138

Fend, H. (1981). *Theorie der Schule* (2. Aufl.). München u.a.: Urban & Schwarzenberg.

Fend, H. (2002). Mikro- und Makrofaktoren eines Angebot-Nutzungsmodells von Schulleistungen. Gasteditorial. *Zeitschrift für Pädagogische Psychologie, 16*(3/4), 141–149. doi: https://doi.org/10.1024//1010-0652.16.34.141

Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Fend, H. (2011). Die Wirksamkeit der neuen Steuerung – theoretische und methodische Probleme ihrer Evaluation. *Zeitschrift für Bildungsforschung, 1*(1), 5–24. doi: 10.1007/s35834-011-0003-3

Fortier, M. S., Vallerand, R. J. & Guay, F. (1995). Academic motivation and school performance: Toward a structural model. *Contemporary Educational Psychology, 20*(3), 257–274. doi: https://doi.org/10.1006/ceps.1995.1017

Fox, J.-P. (2010). *Bayesian item response modeling. Theory and applications*. New York: Springer Science + Business Media. doi: 10.1007/978-1-4419-0742-4

Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E. & Pekrun, R. (Hrsg.) (2009). *PISA 2006. Skalenhandbuch. Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.

Fuchs, L. S., Schumacher, R. F., Sterba, S. K., Long, J., Namkung, J., Malone, A., . . . Changas, P. (2014). Does working memory moderate the effects of fraction intervention? An aptitude–treatment interaction. *Journal of Educational Psychology, 106*(2), 499–514. doi: https://doi.org/10.1037/a0034341

Gamoran, A., Porter, A. C., Smithson, J. & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis, 19*(4), 325–338. doi: https://doi.org/10.3102/01623737019004325

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press. doi: https://doi.org/10.1017/CBO9780511790942

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*(8), 519–521. doi: https://doi.org/10.1037/h0049294

Göbel, K. & Helmke, A. (2010). Intercultural learning in English as foreign language instruction: The importance of teachers' intercultural experience and the usefulness of precise instructional directives. *Teaching and Teacher Education, 26*(8), 1571–1582. doi: https://doi.org/10.1016/j.tate.2010.05.008

Göllner, R., Fauth, B., Lenske, G., Praetorius, A.-K. & Wagner, W. (2020). Do student ratings of classroom management tell us more about teachers or about classroom composition? *Zeitschrift für Pädagogik. Beiheft, 66*(1), 156–172. doi: https://doi.org/10.3262/ZPB2001156

Grolnick, W. S., Ryan, R. M. & Deci, E. L. (1991). Inner resources for school achievement: Motivational mediators of children's perceptions of their parents. *Journal of Educational Psychology, 83*(4), 508–517. doi: https://doi.org/10.1037/0022-0663.83.4.508

Grossman, P., Cohen, J., Ronfeldt, M. & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293–303. doi: https://doi.org/10.3102/0013189X14544542

Grünkorn, J., Klieme, E. & Stanat, P. (2019). Bildungsmonitoring und Qualitätssicherung. In O. Köller, M. Hasselhorn, F. W. Hesse, K. Maaz, J. Schrader, H. Solga, C. K. Spieß & K. Zimmer (Hrsg.), *Das Bildungswesen in Deutschland: Bestand und Potenziale* (S. 245–280). Bad Heilbrunn: Klinkhardt

Hailikari, T. (2009). *Assessing university students' prior knowledge: Implications for theory and practice.* (Doktorarbeit), University of Helsinki, Helsinki, Finnland. Verfügbar unter: http://urn.fi/URN:ISBN:978-952-10-5946-9

Haladyna, T. M. (1974). Effects of different samples on item and test characteristics of criterion-referenced tests. *Journal of Educational Measurement, 11*(2), 93–99. doi: https://doi.org/10.1111/j.1745-3984.1974.tb00977.x

Haladyna, T. & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement, 18*(1), 39–53. doi: 10.1111/j.1745-3984.1981.tb00841.x

Hamre, B. K. & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967. doi: https://doi.org/10.1111/j.1467-8624.2005.00889.x

Hanson, R. A., McMorris, R. F. & Bailey, J. D. (1986). Differences in instructional sensitivity between item formats and between achievement test items. *Journal of Educational Measurement, 23*(1), 1–12. doi: https://doi.org/10.1111/j.1745-3984.1986.tb00230.x

Hardre, P. L. & Reeve, J. (2003). A motivational model of rural students' intentions to persist in, versus drop out of, high school. *Journal of Educational Psychology, 95*(2), 347–356. doi: https://doi.org/10.1037/0022-0663.95.2.347

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

Hartig, J., Frey, A. & Jude, N. (2020). Validität von Testwertinterpretationen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 529–545). Berlin & Heidelberg: Springer. doi: https://doi.org/10.1007/978-3-662-61532-4_21

Hartig, J. & Kühnbach, O. (2006). Schätzung von Veränderung mit „plausible values “in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merkens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft* (S. 27–44). Wiesbaden: VS Verlag für Sozialwissenschaften. doi: https://doi.org/10.1007/978-3-531-90502-0_3

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hattie, J. (2013). *Lernen sichtbar machen. Überarbeitete deutschsprachige Ausgabe "Visible Learning" besorgt von Wolfgang Beywl und Klaus Zierer* (3. Aufl.). Baltmannsweiler: Schneider Verlag Hohengehren.

Hattie, J. & Zierer, K. (2018). *Visible Learning: Auf den Punkt gebracht*. Baltmannsweiler: Schneider Verlag Hohengehren.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis. A regression-based approach* (2. Aufl.). New York: Guilford Press.

Heckhausen, H. & Rheinberg, F. (1980). Lernmotivation im Unterricht, erneut betrachtet. *Unterrichtswissenschaft, 8*(1), 7–47.

Heckhausen, J., Heckhausen, H. & (Hrsg.) (2010). *Motivation und Handeln* (4. Aufl.). Berlin & Heidelberg: Springer. doi:10.1007/978-3-642-12693-2

Heckmann, G. (1980). *Das Sokratische Gespräch. Erfahrungen in philosophischen Hochschulseminaren*. Hannover: Schroedel.

Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes*. Heidelberg: Spektrum Akademischer Verlag.

Heller, K. A. & Hany, E. A. (2014). Standardisierte Schulleistungsmessungen. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schule* (3 Aufl., S. 87–102). Weinheim & Basel: Beltz.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett Kallmeyer.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), *Schulleitung und Schulentwicklung: Voraussetzungen, Bedingungen, Erfahrungen* (S. 119–144). Hohengehren: Schneider-Verlag.

Helmke, A. & Schrader, F. (2019). Angebot-Nutzungs-Modell der Wirkfaktoren akademischer Leistungen. In M. A. Wirtz (Hrsg.), Dorsch - Lexikon der Psychologie. Verfügbar unter: https://portal.hogrefe.com/dorsch/angebots-nutzungs-modell-der-wirkfaktoren-akademischer-leistungen/.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), *Enzyklopädie der Psychologie: Psychologie der Schule und des Unterrichts* (Bd. 3, S. 71–176). Göttingen: Hogrefe.

Helmstadter, G. C. (1972, September). *A comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance oriented instruction*. Referat anlässlich der American Psychological Association Convention, Honolulu, HI. Verfügbar unter: https://files.eric.ed.gov/fulltext/ED069728.pdf

Helmstadter, G. C. (1974, ohne Monatsangabe). *A Comparison of Bayesian and Traditional Indexes of Test Item Effectiveness*. Referat anlässlich der Jahrestagung des National Council on Measurement in Education, Chicago, IL. Verfügbar unter: https://eric.ed.gov/?id=ED087821

Hochweber, J., Hosenfeld, I. & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology, 106*(1), 289–300. doi: 10.1037/a0033829

Hochweber, J. & Leininger, S. (o. D.). Instruktionssensitivität von Testitems in der Pädagogisch-Psychologischen Diagnostik (INSE). Verfügbar unter: https://www.phsg.ch/de/forschung/projekte/instruktionssensitivitaet-von-testitems-der-paedagogisch-psychologischen, Abruf: 18.12.2022

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods & Research, 11*(3), 325–344. doi: https://doi.org/10.1177/0049124183011003003

Hoffman, B. (2015). *Motivation for learning and performance*. San Diego & Oxford: Academic Press.

Höflich, S. (2019). Unter Kontrolle: Die Bedeutung der Impulskontrolle in der Schule. *R&E-SOURCE*(11), 1–9.

Holland, P. W. & Wainer, H. (Hrsg.) (1993). *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates.

Hosenfeld, I., Helmke, A. & Schrader, F.-W. (2002). Diagnostische Kompetenz: Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrkräfte in der Unterrichtsstudie SALVE. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen. (Zeitschrift für Pädagogik. Beiheft 45)* (S. 65–82). Weinheim & Basel: Beltz.

Hoz, R., Bowman, D. & Kozminsky, E. (2001). The differential effects of prior knowledge on learning: A study of two consecutive courses in earth sciences. *Instructional Science, 29*(3), 187–211. doi: https://doi.org/10.1023/A:1017528513130

Hsu, T.-C. (1971, February). *Empirical data on criterion-referenced tests*. Referat anlässlich der Jahrestagung der American Educational Research Association, New York, NY. Verfügbar unter: https://eric.ed.gov/?id=ED050139

Hu, L-t. & 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. doi: https://doi.org/10.1080/10705519909540118

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B. & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*(5888), 494–495. doi: 10.1126/science.1160364

Ing, M. (2008). Using instructional sensitivity and instructional opportunities to interpret students' mathematics performance. *Journal of Educational Research & Policy Studies, 8*(1), 23–43.

Ing, M. (2016). Initial consideration when applying an instruction sensitivity framework: Partitioning the variation between and within classrooms for two mathematics assessments. *Applied Measurement in Education, 29*(2), 122–131. doi: https://doi.org/10.1080/08957347.2016.1138957

Ing, M. (2017). What about the "instruction" in instructional sensitivity? Raising a validity issue in research on instructional sensitivity. *Educational and Psychological Measurement, 78*(4), 635–652. doi: 10.1177/0013164417714846

Isaac, K., Halt, A. C., Hosenfeld, I., Helmke, A. & Groß Ophoff, J. (2006). VERA: Qualitätsentwicklung und Lehrerprofessionalisierung durch Vergleichsarbeiten. *Die Deutsche Schule, 98*(1), 107–111.

Jäger, S. (2012). *Rezeption und Nutzung von Diagnose- und Vergleichsarbeiten an Schulen. Eine Interviewstudie mit baden-württembergischen Lehrkräften an Haupt-, Realschulen und Gymnasien.* (Doktorarbeit), Pädagogische Hochschule Schwäbisch Gmünd, Schwäbisch Gmünd, Deutschland. Verfügbar unter: https://phsg.bsz-bw.de/frontdoor/deliver/index/docId/8/file/Dissertation_Sibylle_Jaeger_Opus.pdf

Jeffreys, H. (1961). *Theory of probability* (3. Aufl.). Oxford: Oxford University Press.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M. & Rosseel, Y. (2018). semTools: Useful tools for structural equation modeling. R package version 0.5–1. Verfügbar unter: https://CRAN.R-project.org/package=semTools

Kagan, J. & Kogan, N. (1970). Individuality and cognitive performance. In P. H. Mussen (Hrsg.), *Carmichael's manual of child psychology* (3. Aufl., Bd. 1, S. 1273–1365). New York u.a.: John Wiley & Sons, Inc.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*(4), 509–539. doi: https://doi.org/10.1007/s10648-007-9054-3

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi: https://doi.org/10.1111/jedm.12000

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York: Guilford Press.

Kelava, A. & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 75–102). Berlin & Heidelberg: Springer. doi: https://doi.org/10.1007/978-3-642-20072-4_4

Kelly, A. L. (2022). A guide to high-stakes standardized testing in the United States. Leiden: Brill. doi: 10.1163/9789004511736_001

Kiper, H. (2009). Schulentwicklung im Rahmen von Kontextsteuerung: Welche Hinweise geben (durch Evaluation und Vergleichsarbeiten gewonnene) Daten für ihre Ausrichtung? In T. Bohl, H. K. (Hrsg.), *Lernen aus Evaluationsergebnissen: Verbesserungen planen und implementieren* (S. 13–28). Bad Heilbrunn: Klinkhardt.

Klafki, W. (2019). Zur Frage nach der Pädagogischen Bedeutung des Sokratischen Gesprächs und neuerer Diskurstheorien. In K.-H. Braun, F. Stübig & H. Stübig (Hrsg.), *Allgemeine Erziehungswissenschaft. Systematische und historische Abhandlungen* (S. 61–70). Wiesbaden: VS Verlag für Sozialwissenschaften. doi: https://doi.org/10.1007/978-3-658-23165-1_3

Klieme, E. (2004). Begründung, Implementation und Wirkung von Bildungsstandards: Aktuelle Diskussionslinien und empirische Befunde. Einführung in den Thementeil. *Zeitschrift für Pädagogik, 50*(5), 625–634.

Klieme, E. (2016). Schulqualität, Schuleffektivität und Schulentwicklung: Welche Erkenntnis eröffnet empirische Forschung? In U. Steffens & T. Bargel (Hrsg.), *Schulqualität – Bilanz und Perspektiven.* (S. 45–64). Münster & New York: Waxmann.

Klieme, E. (2019). Unterrichtsqualität. In M. Gläser-Zikuda, M. Harring & C. Rohlfs (Hrsg.), *Handbuch Schulpädagogik* (S. 393–408). Münster & New York: Waxmann.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., . . . Vollmer, H., J. . (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise* (Bd. 1). Bonn & Berlin BMBF.

Klieme, E., Lipowsky, F., Rakoczy, K. & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht: Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule* (S. 127–146). Münster: Waxmann.

Klieme, E., Pauli, C. & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In Janik, T. & Seidel, T. (Hrsg.), *The power of video studies in investigating teaching and learning in the classroom* (S. 137–160). Münster u.a.: Waxmann.

Klieme, E. & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik, 54*(2), 222–237.

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

Kosecoff, J. B. & Klein, S. P. (1974, April). *Instructional sensitivity statistics appropriate for objectives-based test items: CSE Report No. 91*. Referat anlässlich der Jahrestagung des National Council on Measurement in Education, Chicago, IL. Verfügbar unter: https://files.eric.ed.gov/fulltext/ED103458.pdf

Kounin, J. S. (2006). *Techniken der Klassenführung. (Original der deutschen Ausgabe, 1976)* (Bd. 3). Münster u.a.: Waxmann.

Krapp, A. (1992). Das Interessenkonstrukt. Bestimmungsmerkmale der Interessenhandlung und des individuellen Interesses aus der Sicht einer Person-Gegenstands-Konzeption. In A. Krapp & M. Prenzel (Hrsg.), *Interesse, Lernen, Leistung: Neuere Ansätze der pädagogisch-psychologischen Interessenforschung* (S. 297–329). Münster: Aschendorff.

Krapp, A. (1999). Intrinsische Lernmotivation und Interesse: Forschungsansätze und konzeptionelle Überlegungen. *Zeitschrift für Pädagogik, 45*(3), 387–406.

Krapp, A. (2003). Die Bedeutung der Lernmotivation für die Optimierung des schulischen Bildungssystems. *Politische Studien, 54*(3), 91–105.

Kriegbaum, K., Jansen, M. & Spinath, B. (2015). Motivation: A predictor of PISA's mathematical competence beyond intelligence and prior test achievement. *Learning and Individual Differences, 43*, 140–148. doi: https://doi.org/10.1016/j.lindif.2015.08.026

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2. Aufl.). London u.a.: Elsevier.

Kunter, M. & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 85–113). Münster u.a.: Waxmann.

Kurz, A. & Elliott, S. (2011). Overcoming barriers to access for students with disabilities. Testing accommodations and beyond. In M. Russell & M. Kavanaugh (Hrsg.), *Assessing students in the margins: Challenges, strategies, and techniques*, 31–58. Verfügbar unter: https://ebookcentral.proquest.com/lib/phsg/reader.action?docID=3315589&ppg=48

Kurz, A. & Elliott, S. (2013). *MyiLOGS Guidebook: Advancing the measurement of instruction to help teachers grow and increase students' opportunities to learn (Version 3.0)*. Tempe: Arizona State University.

Kurz, A., Elliott, S. N., Kettler, R. J. & Yel, N. (2014). Assessing students' opportunity to learn the intended curriculum using an online teacher log: Initial validity evidence. *Educational Assessment, 19*(3), 159–184. doi: https://doi.org/10.1080/10627197.2014.934606

Labudde, P. (1996). Genetisch-sokratisch-exemplarisches Lernen im Lichte der neueren Wissenschaftstheorie. *Beiträge zur Lehrerinnen- und Lehrerbildung, 14*(2), 170–174.

LeBreton, J. M. & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. doi: https://doi.org/10.1177/1094428106296642

Li, H., Qin, Q. & Lei, P.-W. (2017). An examination of the instructional sensitivity of the TIMSS math items: A hierarchical differential item functioning approach. *Educational Assessment, 22*(1), 1–17. doi: http://dx.doi.org/10.1080/10627197.2016.1271702

Lind, G. (2009). Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen: Verbesserungen planen und implementieren* (S. 61–79). Bad Heilbrunn: Klinkhardt.

Linn, R. L. (1983a, April). *Measuring school effectiveness: How achievement data can and cannot be used*. Referat anlässlich der Jahrestagung der American Educational Research Association, Montreal, QC. Verfügbar unter: https://eric.ed.gov/?q=Measuring+School+Effectiveness%3a+How+Achievement+Data&id=ED231852

Linn, R. L. (1983b). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*(2), 179–189. doi: https://doi.org/10.1111/j.1745-3984.1983.tb00198.x

Linn, R. L. & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*(2), 109–118. doi: https://doi.org/10.1111/j.1745-3984.1981.tb00846.x

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten? Prädiktionskraft individueller und kontextspezifischer Merkmale*. Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/978-3-531-94339-8

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E. & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of

the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527–537. doi: https://doi.org/10.1016/j.learninstruc.2008.11.001

Lorenz, J. H. (2005). Zentrale Lernstandsmessung in der Primarstufe — Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern. *ZDM – Mathematics Education, 37*(4), 317–323. doi: 10.1007/bf02655818

Lüdtke, O., Trautwein, U., Schnyder, I. & Niggli, A. (2007). Simultane Analysen auf Schüler- und Klassenebene. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 39*(1), 1–11. doi: https://doi-org.ezproxy.phsg.ch/10.1026/0049-8637.39.1.1

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer. doi:10.1007/978-0-387-71265-9

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. & Hines, J. E. (2017). *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence* (2. Aufl.). London u.a.: Elsevier.

Maier, U. (2008). Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. *Zeitschrift für Pädagogik, 54*(1), 95–117.

Maier, U. & Kuper, H. (2012). Vergleichsarbeiten als Instrumente der Qualitätsentwicklung an Schulen. Überblick zum Forschungsstand. *Die Deutsche Schule, 104*(1), 88–99. doi:10.31244/dds.2012.01.07

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S. & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124. doi: https://doi.org/10.1080/00461520.2012.670488

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*(1), 14–19. doi: 10.1037/0003-066X.59.1.14

McPherson, A. (1993). Measuring added value in schools. *Education Economics, 1*(1), 43–51. doi: https://doi.org/10.1080/09645299300000006

Mehrens, W. A. & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357–370. doi: https://doi.org/10.1111/j.1745-3984.1987.tb00286.x

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63. doi:10.1126/science.159.3810.56

Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (3. Aufl., S. 13–104). New York: National Council on Education and Macmillan

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. doi: https://doi.org/10.1037/0003-066X.50.9.741

Millman, J. (1970). Reporting student progress: A case for a criterion-referenced marking system. *The Phi Delta Kappan, 52*(4), 226–230.

Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy, 3*(1), 82–100. doi: 10.2478/v10159-012-0004-x

Moser, U. (2003). *Klassencockpit im Kanton Zürich. Ergebnisse einer Befragung von Lehrerinnen und Lehrern der 6. Klasse über ihre Erfahrungen im Rahmen der Erprobung von Klassencockpit im Schuljahr 2002/03*. Zürich, CH: Universität Zürich, Kompetenzzentrum für Bildungsevaluation und Leistungsmessung. Verfügbar unter: http://www.ibe.uzh.ch/dam/jcr:ffffffff-b12d-25dc-0000-0000177f55e5/Evaluation_Klassencockpit.pdf

Murphy, K. R. & Davidshofer, C. O. (1988). *Psychological testing: Principles and applications*. Englewood Cliffs: Prentice-Hall.

Murphy, K. R. & Davidshofer, C. O. (2013). *Psychological testing. Principles and applications* (6. Aufl.). Harlow: Pearson Education

Musow, S., Naumann, A., Hochweber, J. & Hartig, J. (2017, September). *Evaluation der Instruktionssensitivität als Gütekriterium von Schulleistungstests*. Referat anlässlich der Jahrestagung der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF), Tübingen (D).

Musow, S., Naumann, A., Hochweber, J. & Hartig, J. (2019, April). *Multilevel IRT as a validation strategy for expert judgements on instructional sensitivity*. Referat anlässlich der Jahrestagung des National Council on Measurement in Education, New York, NY.

Muthén, B. O. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika, 54*(3), 385–396. doi: https://doi.org/10.1007/BF02294624

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338–354. doi: https://doi.org/10.1111/j .1745-3984.1991.tb00363.x

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376–398. doi: https://doi.org/10.1177/0049124194022003006

Muthén, B. O., Huang, L.-C., Jo, B., Khoo, S.-T., Goff, G. N., Novak, J. R. & Shih, J. C. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis, 17*(3), 371–403. doi: https://doi.org/10.3102/0162 3737017003371

Muthén, B. O., Kao, C. F. & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*(1), 1–22. doi: https://doi.org/10.1111 /j.1745-3984.1991.tb00340.x

Nagy, G., Nagengast, B., Frey, A., Becker, M. & Rose, N. (2016). Itempositionseffekte in Large-Scale-Assessments. In Bundesministerium für Bildung und Forschung (Hrsg.), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessment* (S. 121–139). Berlin: Bundesministerium für Bildung und Forschung.

Naumann, A. (2013). *Instruktionssensitivität von Leistungstestaufgaben.* (Doktorarbeit), Goethe-Universität Frankfurt a. M., Deutschland. Verfügbar unter: http://d-nb.info/ 1059759225/04

Naumann, A., Hartig, J. & Hochweber, J. (2017). Absolute and relative measures of instructional sensitivity. *Journal of Educational and Behavioral Statistics, 42*(6), 678–705. doi: https://doi.org/10.3102/1076998617703649

Naumann, A., Hochweber, J. & Hartig, J. (2014). Modeling instructional sensitivity using a longitudinal multilevel differential item functioning approach. *Journal of Educational Measurement, 51*(4), 381–399. doi: 10.1111/jedm.12051

Naumann, A., Hochweber, J. & Klieme, E. (2016). A psychometric framework for the evaluation of instructional sensitivity. *Educational Assessment, 21*(2), 89–101. doi: https://doi.org/10.1080/10627197.2016.1167591

Naumann, A., Musow, S., Aichele, C., Hochweber, J. & Hartig, J. (2019). Instruktionssensitivität von Tests und Items. *Zeitschrift für Erziehungswissenschaft, 22*(1), 181–202. doi: https://doi.org/10.1007/s11618-018-0832-0

Naumann, A., Musow, S. & Katstaller, M. (2020). Instructional sensitivity as a prerequisite for determining the effectiveness of interventions in educational research. In H. Astleitner (Hrsg.), *Intervention Research in Educational Practice: Alternative Theoretical Frameworks and Application Problems* (S. 147–170). Münster & New York: Waxmann.

Naumann, A., Rieser, S., Musow, S., Hochweber, J. & Hartig, J. (2019). Sensitivity of test items to teaching quality. *Learning and Instruction, 60*, 41–53. doi: https://doi.org/10.1016/j.learninstruc.2018.11.002

Newton, X. A., Darling-Hammond, L., Haertel, E. & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23), 1–27. doi: https://doi.org/10.14507/epaa.v18n23.2010

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L. & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment, 12*(3 & 4), 215–237. doi: 10.1080/10627190701578271

Niklas, F., Segerer, R., Schmiedeler, S. & Schneider, W. (2012). Findet sich ein „Matthäus-Effekt "in der Kompetenzentwicklung von jungen Kindern mit oder ohne Migrationshintergrund? *Frühe Bildung, 1*(1), 26–33. doi:10.1026/2191-9186/a000022

No Child Left Behind Act of 2001, Pub. L. 107–110, 20 U.S.C. § 6319 (2002).

Noels, K. A., Clément, R. & Pelletier, L. G. (1999). Perceptions of teachers' communicative style and students' intrinsic and extrinsic motivation. *The Modern Language Journal, 83*(1), 23–34. doi: https://doi.org/10.1111/0026-7902.00003

Noser, F., Oester Schläppi, M., Scherer, T. & Züger, R. (2004). *Logisch 4. Das Buch. Mathematik-Lehrmittel für die 4. Klasse*. Rorschach: Kantonaler Lehrmittelverlag St. Gallen.

Noser, F., Oester Schläppi, M., Scherer, T. & Züger, R. (2005a). *Logisch 5. Das Buch. Mathematik-Lehrmittel für die 5. Klasse*. Rorschach: Kantonaler Lehrmittelverlag St. Gallen.

Noser, F., Oester Schläppi, M., Scherer, T. & Züger, R. (2005b). *Logisch 5. Das Heft. Mathematik-Lehrmittel für die 5. Klasse*. Rorschach: Kantonaler Lehrmittelverlag St. Gallen.

Noser, F., Oester Schläppi, M., Scherer, T. & Züger, R. (2005c). *Logisch 5. Der Kommentar. Mathematik-Lehrmittel für die 5. Klasse*. Rorschach: Kantonaler Lehrmittelverlag St. Gallen.

Padberg, F. & Wartha, S. (2017). *Didaktik der Bruchrechnung. Mathematik Primarstufe und Sekundarstufe I + II* (5. Aufl.). Berlin: Springer Spektrum. doi: 10.1007/978-3-662-52969-0

Parkerson, J. A., Lomax, R. G., Schiller, D. P. & Walberg, H. J. (1984). Exploring causal models of educational achievement. *Journal of Educational Psychology, 76*(4), 638–646. doi: https://doi.org/10.1037/0022-0663.76.4.638

Paterson, L. (1991). Socio-economic status and educational attainment: a multi-dimensional and multi-level study. *Evaluation & Research in Education, 5*(3), 97–121. doi: https://doi.org/10.1080/09500799109533303

Petersen, I. h., Louw, J. & Dumont, K. (2009). Adjustment to university and academic performance among disadvantaged students in South Africa. *Educational Psychology, 29*(1), 99–115. doi: https://doi.org/10.1080/01443410802521066

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(2), 381–391. doi: https://doi.org/10.1086/209405

Pham, G. (2018). *Complex interplay effects of classroom instructional and context factors on student growth in English as a foreign language in Vietnam: The case of nonlinear relationship, regularized regression and the relevance of the scaling model.* Universität Koblenz-Landau, Landau, Deutschland. Verfügbar unter: https://kola.opus.hbz-nrw.de/opus45-kola/frontdoor/deliver/index/docId/1617/file/Pham__Dissertation_Publication.pdf

Pham, V. H. (2009). *Computer modeling of the instructionally insensitive nature of the Texas assessment of knowledge and skills (TAKS) exam.* (Doktorarbeit), University of Texas Austin, USA. Verfügbar unter: https://repositories.lib.utexas.edu/bitstream/ handle/ 2152/10602/phamv53572.pdf?sequence=2

Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119. doi: https://doi.org/10.3102/0013189X09332374

Pianta, R. C., La Paro, K. M. & Hamre, B. K. (2008). *Classroom Assessment Scoring System. Manual K-3*. Baltimore: Paul H. Brookes Publishing Co.

Pinkerton, E., Scott, C., Buell, B. & Kober, N. (2004). *From the Capital to the Classroom. Year 2 of the No Child Left Behind Act*. Washington, DC, USA: Center on Education Policy. Verfügbar unter: https://nepc.colorado.edu/sites/default/files/EPRU-0504-122-OWI.pdf

Plummer, M. (2003, March). *JAGS. A Program for analysis of Bayesian graphical models using Gibbs sampling.* Paper presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria.

Plummer, M. (2017a). JAGS version 4.3.0 user manual. Verfügbar unter: http://www.stat.yale.edu/~jtc5/238/materials/jags_4.3.0_manual_with_distributions.pdf

Plummer, M. (2017b). JAGS. (Version 4.3. 0) [Computer Software]. Verfügbar unter: http://mcmc-jags.sourceforge.net/

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*(4), 3–14. doi:10.1111/j.1745-3992.2010.00189.x

Polikoff, M. S. (2016). Evaluating the instructional sensitivity of four states' student achievement tests. *Educational Assessment, 21*(2), 102–119. doi:10.1080/10627197.2016.1166342

Polikoff, M. S. & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*(4), 399–416. doi: https://doi.org/10.3102/0162373714531851

Popham, W. J. (1971). Indices of adequacy for criterion-referenced test items. In W. J. Popham (Hrsg.), *Criterion-referenced measurement: An introduction* (S. 79–98). Englewood Cliffs, NJ: Educational Technology.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89*(2), 146–155. doi: https://doi.org/10.1177/003172170708900211

Popham, W. J. (2014). The right test for the wrong reason. *Phi Delta Kappan, 96*(1), 46–52. doi: https://doi.org/10.1177/0031721714547862

Popham, W. J., Berliner, D. C., Kingston, N. M., Fuhrman, S. H., Ladd, S. M., Charbonneau, J. & Chatterji, M. (2014). Can today's standardized achievement tests yield instructionally useful data? Challenges, promises and the state of the art. *Quality Assurance in Education, 22*(4), 303–316. doi: https://doi.org/10.1108/QAE-07-2014-0033

Popham, W. J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1–9. doi: https://doi.org/10.1111/j.1745-3984.1969.tb00654.x

Popham, W. J. & Ryan, J. M. (2012, April). *Determining a high-stakes test's instructional sensitivity*. Referat anlässlich der Jahrestagung des National Council on Educational Measurement, Vancouver, BC. Verfügbar unter: https://www.minnetonka .k12.mn.us/administration/teachingandlearning/staffdev/2012%20Handouts/June%202012/Popham_article_DETECTING%20THE%20INSTRUCTIONAL%20SENSITIVITY.pdf

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation*. München: Ernst Reinhardt.

Praetorius, A.-K., Klieme, E., Herbert, B. & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM Mathematics Education, 50*(3), 407–426. doi: https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A. K., Herrmann, C., Gerlach, E., Zülsdorf-Kersting, M., Heinitz, B., & Nehring, A. (2020). Unterrichtsqualität in den Fachdidaktiken im deutschsprachigen Raum – zwischen Generik und Fachspezifik. *Unterrichtswissenschaft*, *48*(3), 409–446. doi: https://doi.org/10.1007/s42010-020-00082-8

Praetorius, A.-K., Klieme, E., Kleickmann, T., Brunner, E., Lindmeier, A., Traut, S. & Charalambous, C. (2020). Towards developing a theory of generic teaching quality: origin, current status, and necessary next steps regarding the Three Basic Dimensions model. *Zeitschrift für Pädagogik. Beiheft, 66*, 15–36. doi: https://doi.org/10.3262/ZPB2001015

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12. doi: https://doi.org/10.1016/j.learninstruc.2013.12.002

Preacher, K. J., Zyphur, M. J. & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209–233. doi: https://doi.org/10.1037/a0020141

Prenzel, M., Seidel, T. & Drechsel, B. (2004). Autonomie in Wissensprozessen. In G. Reinmann-Rothmeier & H. Mandl (Hrsg.), *Psychologie des Wissensmanagements: Perspektiven, Theorien und Methoden* (S. 102–113). Göttingen: Hogrefe

R Core Team. (2017). R: A language and environment for statistical computing [Computer Software]. Verfügbar unter: https://www.R-project.org/

Raab-Steiner, E. & Benesch, M. (2015). *Der Fragebogen. Von der Forschungsidee zur SPSS-Auswertung* (4. Aufl.). Wien: Facultas-Verlag.

Radatz, H. (1976). *Individuum und Mathematikunterricht: eine Untersuchung zum konzeptuellen Tempo seiner Bedeutung für den Bereich der Erziehung*. Hannover Schroedel.

Radatz, H. (1980). *Fehleranalysen im Mathematikunterricht*. Braunschweig: Friedr. Vieweg & Sohn Verlagsgesellschaft.

Ramsteck, C. & Maier, U. (2015). Testdatenbasierte Schul- und Unterrichtsentwicklung. Analyse von Handlungsmustern bei der Rezeption und Nutzung von Vergleichsarbeitsdaten. In J. Schrader, J. Schmid, K. Amos & A. Thiel (Hrsg.), *Governance von Bildung im Wandel: Interdisziplinäre Zugänge* (S. 119–144). Wiesbaden: VS Verlag für Sozialwissenschaften.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Hrsg.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Bd. 4, S. 321–333). Berkeley & Los Angeles: University of California Press. doi: https://doi.org/10.1002/bimj.19640060422

Raudenbush, S. W. & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307–335. doi: https://doi.org/10.3102/10769986020004307

Raykov, T., & Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling*, *13*(1), 130–141. doi: https://doi.org/10.1207/s15328007sem1301_7

Renkl, A. (1996). Vorwissen und Schulleistung. In J. Möller & O. Köller (Hrsg.), *Emotionen, Kognitionen und Schulleistung* (S. 175–190). Weinheim: Psychologie Verlags Union.

Reusser, K. & Pauli, C. (2010). Unterrichtsgestaltung und Unterrichtsqualität – Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht: Einleitung und Überblick. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 9–32). Münster: Waxmann

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M. & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*(3), 700–712. doi: https://doi.org/10.1037/a0027268

Rheinberg, F. (2014). Bezugsnormen und schulische Leistungsbeurteilung In Weinert, F. E. (Hrsg.), *Leistungsmessungen in Schulen* (3 Aufl., S. 59–71). Weinheim & Basel: Beltz.

Rijmen, F., Tuerlinckx, F., De Boeck, P. & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185–205. doi: https://doi.org/10.1037/1082-989X.8.2.185

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In A. Bremerich-Vos, D. Ganzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 42–106). Weinheim: Beltz Pädagogik.

Robitzsch, A., Kiefer, T. & Wu, M. (2021). TAM: Test analysis modules. R package version 3.7–16. Verfügbar unter: https://CRAN.R-project.org/package=TAM

Rodriguez, G. & Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal, 3*(1), 32–46. doi: https://doi.org/10.1177/1536867X0300300102

Rodriguez, M. (2017, April). *Item and test design considering instructional sensitivity*. Referat anlässlich der Jahrestagung des National Council on Measurement in Education, San Antonio, TX. Verfügbar unter: http://www.edmeasurement.net/presentations/ Rodriguez-2017-instructional-sensitivity.pdf

Roediger III, H. L. & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. doi: https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger III, H. L. & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. doi: https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger III, H. L. & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155–1159. doi: https://doi.org/10.1037/0278-7393.31.5.1155

Roick, T. (2008). Standardisierte Schulleistungstests. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 271–281). Göttingen: Hogrefe.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software, 48*(2), 1–36. doi:10.18637/jss.v048.i02

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., . . . Barendse, M. (2018). lavaan: Latent variable analysis. R package version 0.6-3. Verfügbar unter: https://cran.r-project.org/package=lavaan

Roth, G. (2011). *Bildung braucht Persönlichkeit. Wie Lernen gelingt*. Stuttgart: Klett-Cotta.

Roudabush, G. E. (1974, April). *Models for a Beginning Theory of Criterion-Referenced Tests*. Referat anlässlich der Jahrestagung des National Council on Measurement in Education, Chicago, IL. Verfügbar unter: https://files.eric.ed.gov/fulltext/ED107702.pdf

Ruiz-Primo, M. A. & Li, M. (2015). The relationship between item context characteristics and student performance: The case of the 2006 and 2009 PISA science items. *Teachers College Record, 117*(1), 1–36.

Ruiz-Primo, M. A., Min, L., Wills, K., Giamellaro, M., Lan, M. C., Mason, H. & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science.

Journal of Research in Science Teaching, 49(6), 691–712. doi: https://doi.org/10.1002/tea.21030

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369–393. doi: https://doi.org/10.1002/tea.10027

Rutkowski, D. & Wild, J. (2015). Stakes matter: Student motivation and the validity of student assessments for teacher evaluation. *Educational Assessment, 20*(3), 165–179. doi: https://doi.org/10.1080/10627197.2015.1059273

Ryan, R. M. & Connell, J. P. (1989). Perceived locus of causality and internalization: examining reasons for acting in two domains. *Journal of Personality and Social Psychology, 57*(5), 749–761. doi: https://doi.org/10.1037/0022-3514.57.5.749

Ryan, R. M. & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54–67. doi: https://doi.org/10.1006/ceps.1999.1020

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

Schmette, M. & Haase, H.-M. (2014). *Forschungsbericht 2010 - 2012.* Schwäbisch Gmünd: Pädagogische Hochschule Schwäbisch Gmünd. Verfügbar unter: http://www.ph-gmuend.de/fileadmin/redakteure/ph-hauptseite/redakteure/daten/download/forschung/Forschungsbericht_2012-2014_Gesamtdokument_final_Maerz_15_Internet.pdf

Schrader, F.-W. & Helmke, A. (2002). Motivation, Lernen und Leistung. In A. Helmke & R. S. Jäger (Hrsg.), *Das Projekt MARKUS. Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext* (S. 257–324). Landau: Empirische Pädagogik.

Schraw, G. (2006). Knowledge: Structures and processes. In P. A. Alexander & P. H. Winne (Hrsg.), *Handbook of educational psychology* (2. Aufl., S. 245–264). New York: Routledge. doi: https://doi.org/10.4324/9780203874790

Schuler, H., Funke, U. & Baron-Boldt, J. (1990). Predictive validity of school grades - A meta-analysis. *Applied Psychology, 39*(1), 89–103. doi: https://doi.org/10.1111/j.1464-0597.1990.tb01039.x|

Schunk, D. H., Meece, J. R. & Pintrich, P. R. (Hrsg.). (2014). *Motivation in education: Theory, research, and applications* (4. Aufl.). London: Pearson.

Seidel, T. (2014). Angebots-Nutzungs-Modelle in der Unterrichtspsychologie: Integration von Struktur- und Prozessparadigma. *Zeitschrift für Pädagogik, 60*(6), 850–866.

Seidel, T., Prenzel, M. & Kobarg, M. (2005). *How to run a video study. Technical report of the IPN Video Study*. Münster: Waxmann.

Seligman, M. E. P. (1975). *Helplessness. On depression, development, and death*. San Francisco: W. H. Freeman

Sharma, S., Durand, R. M. & Gur-Arie, O. (1981). Identification and analysis of moderator variables. *Journal of Marketing Research, 18*(3), 291–300. doi: https://doi.org/10.1177/002224378101800303

Shulman, L. S. (1982). Educational psychology returns to school. In A. G. Kraut (Hrsg.), *The G. Standley Hall lecture series* (Bd. 2, S. 77–117). Washington, DC: American Psychological Association. doi: https://doi.org/10.1037/10087-002

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling* (2. Aufl.). Los Angeles u.a.: Sage.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59*(2), 205–216. doi: https://doi.org/10.1037/0022-006X.59.2.205

Soenens, B. & Vansteenkiste, M. (2005). Antecedents and outcomes of self-determination in 3 life domains: The role of parents' and teachers' autonomy support. *Journal of Youth and Adolescence, 34*(6), 589–604. doi: 10.1007/s10964-005-8948-y

Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher & S. P. Klein (Hrsg.), *Making sense of test-based accountability in education* (S. 79–100). Santa Monica u.a.: RAND.

Steyer, R., Eid, M. & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online, 2*(1), 21–33.

Stigler, J. W. & Hiebert, J. (1999). *The teaching gap. Best ideas from the world's teacher for improving education in the classroom*. New York: The Free Press.

Sutton, A. & Soderstrom, I. (1999). Predicting elementary and secondary school achievement with school-related and demographic factors. *The Journal of Educational Research, 92*(6), 330–338. doi: https://doi.org/10.1080/00220679909597616

Taylor, G., Jungert, T., Mageau, G. A., Schattke, K., Dedic, H., Rosenfield, S. & Koestner, R. (2014). A self-determination theory approach to predicting school achievement over time: The unique role of intrinsic motivation. *Contemporary Educational Psychology, 39*(4), 342–358. doi: https://doi.org/10.1016/j.cedpsych.2014.08.002

Traynor, A. (2017). Does test item performance increase with test-to-standards alignment? *Educational Assessment, 22*(3), 171–188. doi: https://doi.org/10.1080/10627197.2017.1344092

Tresch, S. (2007). *Potenzial Leistungstest. Wie Lehrerinnen und Lehrer Ergebnisrückmeldungen zur Sicherung und Steigerung ihrer Unterrichtsqualität nutzen*. Bern: h.e.p.

Tschannen-Moran, M. & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783–805. doi: https://doi.org/10.1016/S0742-051X(01)00036-1

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Brière, N. M., Senécal, C. & Vallières, E. F. (1992). The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement, 52*(4), 1003–1017. doi: https://doi.org/10.1177/0013164492052004025

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Brière, N. M., Senécal, C. & Vallières, E. F. (1993). On the assessment of intrinsic, extrinsic, and amotivation in education: Evidence on the concurrent and construct validity of the Academic Motivation Scale. *Educational and Psychological Measurement, 53*(1), 159–172. doi: https://doi.org/10.1177/ 0013164493053001018

van den Noortgate, W., De Boeck, P. & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369–386. doi: https://doi.org/10.3102/10769986028004369

van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research, 51*(3), 379–402. doi: https://doi.org/10.3102/00346543051003379

van der Linden, W. J. (2016). *Handbook of item response theory* (Bd. 1 Models). New York: Capman and Hall CRC Press. doi: https://doi.org/10.1201/9781315374512

van der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 1–28). New York: Springer. doi: https://doi.org/10.1007/978-1-4757-2691-6_1

Verhagen, A. J. & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology, 66*(3), 383–401. doi: https://doi.org/10.1111/j.2044-8317.2012.02059.x

Verhagen, J., Levy, R., Millsap, R. E. & Fox, J.-P. (2016). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology, 72*, 171–182. doi: https://doi.org/10.1016/j.jmp.2015.06.005

vom Hofe, R. (1995). *Grundvorstellungen mathematischer Inhalte*. Heidelberg u.a.: Spektrum Akademischer Verlag.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H. & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology, 60*(3), 158–189. doi: https://doi.org/10.1016/j.jmp.2015.06.005

Walls, T. A. & Little, T. D. (2005). Relations among personal agency, motivation, and school adjustment in early adolescence. *Journal of Educational Psychology, 97*(1), 23–31. doi: https://doi.org/10.1037/0022-0663.97.1.23

Wang, M. C., Haertel, G. D. & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*(3), 249–294. doi: https://doi.org/10.3102/00346543063003249

Weinert, F. E. (1989, März). *Is the past the best predictor of the future. Short-and long-term predictability of individual differences in childrens cognitive achievement.* Referat anlässlich der Jahrestagung der American Educational Research Association, San Francisco, CA.

Weinert, F. E., Schrader, F.-W. & Helmke, A. (1989). Quality of instruction and achievement outcomes. *International Journal of Educational Research, 13*(8), 895–914. doi: https://doi.org/10.1016/0883-0355(89)90072-4

Wheaton, B., Muthen, B., Alwin, D. F. & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology, 8*, 84–136. doi: http://dx.doi.org/10.2307/270754

Wigfield, A. & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. doi: https://doi.org/10.1006/ceps.1999.1015

Wiley, D. E. & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis, 17*(3), 355–370. doi: https://doi.org/10.3102/01623737017003355

Willms, J. D. & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*(3), 209–232. doi: https://doi.org/10.1111/j.1745-3984.1989.tb00329.x

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis. Rasch measurement*. Chicago: Mesa Press.

Wu, A. D., Liu, Y., Stone, J. E., Zou, D. & Zumbo, B. D. (2017). Is difference in measurement outcome between groups differential responding, bias or disparity? A methodology for detecting bias and impact from an attributional stance. *Frontiers in Education, 2*, 1–12. doi: https://doi.org/10.3389/feduc.2017.00039

Wu, S., Crespi, C. M. & Wong, W. K. (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials, 33*(5), 869–880. doi: 10.1016/j.cct.2012.05.004

Yıldırım, S. (2012). Teacher support, motivation, learning strategy use, and achievement: A multilevel mediation model. *The Journal of Experimental Education, 80*(2), 150–172. doi: https://doi.org/10.1080/00220973.2011.596855

Yoon, B., Burstein, L., Chen, Z. & Kim, K.-S. (1989). *Patterns in teacher reports of topic coverage and their effects on math achievement: Comparisons across years*.

Washington, DC, USA: Graduate School of Education, University of California. Verfügbar unter: https://files.eric.ed.gov/fulltext/ED316418.pdf

Yoon, B., Burstein, L. & Gold, K. (1991). *Assessing the content validity of teachers' reports of content coverage and its relationship to student achievement (CSE Report No. 328)*. Los Angeles: University of Los Angeles & Center for Research on Evaluation, Standards, and Student Testing.

Yoon, B. & Resnick, L. B. (1998). *Instructional validity, opportunity to learn and equity: New standards examinations for the California mathematics renaissance (CSE Technical Report 484)*. Los Angeles, CA, USA: California University, Center for the Study of Evaluation & Center for Research on Evaluation, Standards, and Student Testing. Verfügbar unter: https://files.eric.ed.gov/fulltext/ED426071.pdf.

Ziegler, A. (2008). *Hochbegabung*. München: Ernst Reinhardt.

10 Abbildungsverzeichnis

11 Tabellenverzeichnis

Anhang

Anhang A: Skalen zur Qualität von Lernmotivation

Amotiviertheit

Die verwendete Skala stammt von Frey et al. (2009). Der Ursprung der Skala ist allerdings auf die Arbeiten von Ryan und Cornell (1989) sowie Seidel et al. (2005) zurückzuführen. Um die Skalen von Frey et al. (2009) verwenden zu können, wurden die Items umformuliert, um diese auf die Zielgruppe und den Kontext des Mathematikunterrichts abzustimmen. Die Items S1ma1 und S1ma2 wurden in der Formulierung verändert. Auch die Antwortkategorien von Frey et al. (2009) wurden modifiziert (Original: 1=fast immer, 2=oft, 3=manchmal, 4=nicht bearbeitbar, 5=nicht gültig, 6=nicht bearbeitet).

Tabelle 30

Skala: Qualität der Lernmotivation – Amotiviertheit

Instruktion	**12) Im Mathematikunterricht ...** *(Gib bitte an, wie sehr die folgenden Aussagen auf dich zutreffen.)*
Anzahl Items	3
Codierung	1 nie
	2 manchmal
	3 fast immer
	4 immer
Recodierte Items	-
Variablenname	**Text**
S1ma1	... habe ich keine Lust, mitzuarbeiten
S1ma2	... habe ich keine Lust, mich mit den Themen zu beschäftigen
S1ma3	... bin ich mit meinen Gedanken woanders

Tabelle 31

Deskriptive Ergebnisse zur Amotiviertheit

Item	***Itemnr. (Tabelle 8)***	***N***	***M***	***SD***	***Min***	***Max***
S1ma1	1	785/832	1.74	.72	1	4
S1ma2	2	768/832	1.68	.70	1	4
S1ma3	3	780/832	1.84	.70	1	4

Anmerkung zum Antwortformat: 1=nie, 2=manchmal, 3=fast immer, 4=immer.

Externale Motiviertheit

Die verwendete Skala stammt von Frey et al. (2009). Der Ursprung der Skala ist allerdings auf die Arbeiten von Ryan und Cornell (1989) sowie Seidel et al. (2005) zurückzuführen. Um die Skalen von Frey et al. (2009) verwenden zu können, wurden die Items umformuliert, um diese auf die Zielgruppe und den Kontext des Mathematikunterrichts abzustimmen. Die Items S1mx04 und S1mx06 wurden aufgrund der Erkenntnisse aus der Pilotierung angepasst. Auch die Antwortkategorien von Frey et al. (2009) wurden modifiziert (Original: 1=fast immer, 2=oft, 3=manchmal, 4=nicht bearbeitbar, 5=nicht gültig, 6=nicht bearbeitet).

Tabelle 32

Skala: Qualität der Lernmotivation – Externale Motiviertheit

Instruktion	**12) Im Mathematikunterricht ...** *(Gib bitte an, wie sehr die folgenden Aussagen auf dich zutreffen.)*
Anzahl Items	3
Codierung	1 nie
	2 manchmal
	3 fast immer
	4 immer
Recodierte Items	-
Variablenname	**Text**
S1mx4	… mache ich nicht mehr, als die Lehrperson verlangt
S1mx5	… arbeite ich nur mit, wenn ich dazu aufgefordert werde
S1mx6	… mache ich nur mit, wenn es nicht anders geht

Tabelle 33

Deskriptive Ergebnisse zur externalen Motiviertheit

Item	***Itemnr. (Tabelle 8)***	***N***	***M***	***SD***	***Min***	***Max***
S1mx4	4	777/832	1.99	.89	1	4
S1mx5	5	781/832	1.58	.82	1	4
S1mx6	6	780/832	1.40	.73	1	4

Anmerkung zum Antwortformat: 1=nie, 2=manchmal, 3=fast immer, 4=immer.

Introjizierte Motiviertheit

Die verwendete Skala stammt von Frey et al. (2009). Der Ursprung der Skala ist allerdings auf die Arbeiten von Ryan und Cornell (1989) sowie Seidel et al. (2005) zurückzuführen. Um die Skalen von Frey et al. (2009) verwenden zu können, wurden die Items umformuliert, um diese auf die Zielgruppe und den Kontext des Mathematikunterrichts abzustimmen. Ein Item wurde aufgrund der Erkenntnisse aus der Pilotierung angepasst. Auch die Antwortkategorien von Frey et al. (2009) wurden modifiziert (Original: 1=fast immer, 2=oft, 3=manchmal, 4=nicht bearbeitbar, 5=nicht gültig, 6=nicht bearbeitet).

Tabelle 34

Skala: Qualität der Lernmotivation – Introjizierte Motiviertheit

Instruktion	**12) Im Mathematikunterricht ...** *(Gib bitte an, wie sehr die folgenden Aussagen auf dich zutreffen.)*
Anzahl Items	3
Codierung	1 nie
	2 manchmal
	3 fast immer
	4 immer
Recodierte Items	-
Variablenname	**Text**
S1mt15	… arbeite ich mit, weil ich es immer so mache
S1mt16	… strenge ich mich an, weil ich alles richtig machen möchte
S1mt17	… beteilige ich mich, weil man das als Schülerin oder Schüler so macht

Tabelle 35

Deskriptive Ergebnisse zur introjizierten Motiviertheit

Item	***Itemnr. (Tabelle 8)***	***N***	***M***	***SD***	***Min***	***Max***
S1mt15	7	776/832	2.99	.90	1	4
S1mt16	8	777/832	3.32	.82	1	4
S1mt17	9	775/832	2.88	.97	1	4

Anmerkung zum Antwortformat: 1=nie, 2=manchmal, 3=fast immer, 4=immer.

Identifizierte Motiviertheit

Die verwendete Skala stammt von Frey et al. (2009). Der Ursprung der Skala ist allerdings auf die Arbeiten von Ryan und Cornell (1989) sowie Seidel et al. (2005) zurückzuführen. Um die Skalen von Frey et al. (2009) verwenden zu können, wurden die Items umformuliert, um diese auf die Zielgruppe und den Kontext des Mathematikunterrichts abzustimmen. Die Items S1md7, S1md8 und S1md9 wurden in der Formulierung verändert. Auch die Antwortkategorien von Frey et al. (2009) wurden modifiziert (Original: 1=fast immer, 2=oft, 3=manchmal, 4=nicht bearbeitbar, 5=nicht gültig, 6=nicht bearbeitet).

Tabelle 36

Skala: Qualität der Lernmotivation – Identifizierte Motiviertheit

Instruktion	**12) Im Mathematikunterricht ...** *(Gib bitte an, wie sehr die folgenden Aussagen auf dich zutreffen.)*
Anzahl Items	3
Codierung	1 nie
	2 manchmal
	3 fast immer
	4 immer
Recodierte Items	-
Variablenname	**Text**
S1md7	... arbeite ich mit, weil ich Mathematik später im Leben bestimmt brauche
S1md8	... lerne ich wichtige Dinge, die im Alltag hilfreich sind
S1md9	... mache ich mit, damit ich mich später in Mathematik auskenne

Tabelle 37

Deskriptive Ergebnisse zur identifizierten Motiviertheit

Item	***Itemnr. (Tabelle 8)***	***N***	***M***	***SD***	***Min***	***Max***
S1md7	10	776/832	2.82	1.12	1	4
S1md8	11	780/832	3.12	0.90	1	4
S1md9	12	779/832	3.03	0.99	1	4

Anmerkung zum Antwortformat: 1=nie, 2=manchmal, 3=fast immer, 4=immer.

Interessierte Motiviertheit

Die verwendete Skala stammt von Frey et al. (2009). Der Ursprung der Skala ist allerdings auf die Arbeiten von Ryan und Cornell (1989) sowie Seidel et al. (2005) zurückzuführen. Um die Skalen von Frey et al. (2009) verwenden zu können, wurden die Items umformuliert, um diese auf die Zielgruppe und den Kontext des Mathematikunterrichts abzustimmen. Ein Item wurde in der Formulierung verändert. Auch die Antwortkategorien von Frey et al. (2009) wurden modifiziert (Original: 1=fast immer, 2=oft, 3=manchmal, 4=nicht bearbeitbar, 5=nicht gültig, 6=nicht bearbeitet).

Tabelle 38

Skala: Qualität der Lernmotivation – Interessierte Motiviertheit

Instruktion	**12) Im Mathematikunterricht ...** *(Gib bitte an, wie sehr die folgenden Aussagen auf dich zutreffen.)*
Anzahl Items	2
Codierung	1 nie
	2 manchmal
	3 fast immer
	4 immer
Recodierte Items	-
Variablenname	**Text**
S1mi10	… möchte ich gern mehr über einzelne Themen erfahren
S1mi11	… bekomme ich Lust, mich weiter damit zu beschäftigen

Tabelle 39

Deskriptive Ergebnisse zur interessierten Motiviertheit

Item	***Itemnr. (Tabelle 8)***	***N***	***M***	***SD***	***Min***	***Max***
S1mi10	13	781/832	2.60	.95	1	4
S1mi11	14	779/832	2.64	.97	1	4

Anmerkung zum Antwortformat: 1=nie, 2=manchmal, 3=fast immer, 4=immer.

Intrinsische Motiviertheit

Die verwendete Skala stammt von Frey et al. (2009). Der Ursprung der Skala ist allerdings auf die Arbeiten von Ryan und Cornell (1989) sowie Seidel et al. (2005) zurückzuführen. Um die Skalen von Frey et al. (2009) verwenden zu können, wurden die Items umformuliert, um diese auf die Zielgruppe und den Kontext des Mathematikunterrichts abzustimmen. Ein Item wurde in der Formulierung verändert. Auch die Antwortkategorien von Frey et al. (2009) wurden modifiziert (Original: 1=fast immer, 2=oft, 3=manchmal, 4=nicht bearbeitbar, 5=nicht gültig, 6=nicht bearbeitet).

Tabelle 40

Skala: Qualität der Lernmotivation – Intrinsische Motiviertheit

Instruktion	**12) Im Mathematikunterricht ...** *(Gib bitte an, wie sehr die folgenden Aussagen auf dich zutreffen.)*
Anzahl Items	3
Codierung	1 nie
	2 manchmal
	3 fast immer
	4 immer
Recodierte Items	-
Variablenname	**Text**
S1mr12	… finde ich die Themen richtig spannend
S1mr13	… bin ich mit Freude dabei
S1mr14	… macht mir der Unterricht Spass

Tabelle 41

Deskriptive Ergebnisse zur intrinsischen Motiviertheit

Item	**Itemnr. (Tabelle 8)**	***N***	***M***	***SD***	***Min***	***Max***
S1mr12	15	782/832	2.63	.92	1	4
S1mr13	16	780/832	2.88	.91	1	4
S1mr14	17	779/832	2.95	.87	1	4

Anmerkung zum Antwortformat: 1=nie, 2=manchmal, 3=fast immer, 4=immer.

Anhang B: Unterrichtsqualitätsdimensionen

Unterrichtsstörungen

Die verwendete Skala stammt von Fauth et al. (2014). Bei der Übersetzung vom Englischen ins Deutsche wurde sich am unveröffentlichten Skalenbuch der IGEL-Studie von 2014 unter der Leitung von Decristan, Hertel, Klieme (damals am DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation) sowie Büttner, Hardy, Kunter und Lühken (damals an der Goethe-Universität) orientiert. Während dort die Skala zu Unterrichtsstörungen bzw. zum *Classroom Management* aus sechs Items besteht, wurden bei Fauth et al. (2014) fünf Items zur Abbildung des Konstrukts verwendet. Im INSE-Projekt wurde sich primär an Fauth et al. (2014) orientiert. Im Original wurde auf den Sachunterricht Bezug genommen (ebd.), sodass die Itemformulierungen im Rahmen des INSE-Projekts auf den Kontext des Mathematikunterrichts abgestimmt wurden.

Tabelle 42
Skala Unterrichtsqualität – Unterrichtsstörungen

Instruktion	**18) Im Mathematikunterricht ...** *(Gib bitte an, wie sehr du den Aussagen zustimmst.)*
Anzahl Items	5
Kategorien	1 stimme gar nicht zu
	2 stimme eher nicht zu
	3 stimme eher zu
	4 stimme zu
Recodierte Items	-
Variablenname	**Text**
S4qus1	… stört keiner den Unterricht
S4qus2	… sind die Schülerinnen und Schüler still, wenn die Lehrperson spricht
S4qus3	… hören alle zu und sind leise
S4qus4	… quatscht niemand dazwischen
S4qus5	… hören alle auf die Lehrperson

Tabelle 43

Deskriptive Ergebnisse zu Unterrichtsstörungen

Item	*Itemnr. (Tabelle 12)*	*Itemnr. (Tabelle 15)*	*N*	*M*	*SD*	*Min*	*Max*
S4qus1	1	1	759/832	2.42	.76	1	4
S4qus2	2	2	759/832	2.75	.75	1	4
S4qus3	3	3	754/832	2.64	.74	1	4
S4qus4	4	4	757/832	2.48	.80	1	4
S4qus5	5	5	756/832	2.88	.77	1	4

Anmerkung zum Antwortformat: 1=stimme gar nicht zu, 2=stimme eher nicht zu, 3=stimme eher zu, 4=stimme zu.

Unterstützendes Klassenklima

Die verwendete Skala stammt von Fauth et al. (2014). Bei der Übersetzung vom Englischen ins Deutsche wurde sich am unveröffentlichten Skalenbuch der IGEL-Studie von 2014 unter der Leitung von Decristan, Hertel, Klieme (damals am DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation) sowie Büttner, Hardy, Kunter und Lühken (damals an der Goethe-Universität) orientiert. Im Original wurde auf den Sachunterricht Bezug genommen (Fauth et. al., 2014), sodass die Itemformulierungen im Rahmen des INSE-Projekts auf den Kontext des Mathematikunterrichts abgestimmt wurden.

Tabelle 44

Skala: Unterrichtsqualität – Unterstützendes Klassenklima

Instruktion	**20) Unsere Lehrperson in Mathematik ...** *(Gib bitte an, wie sehr du den Aussagen zustimmst.)*
Anzahl Items	9
Kategorien	1 stimme gar nicht zu
	2 stimme eher nicht zu
	3 stimme eher zu
	4 stimme voll zu
Recodierte Items	-
Variablenname	**Text**
S4quu1	... ist freundlich zu mir, auch wenn ich einen Fehler mache
S4quu2	... kümmert sich um mich
S4quu3	... macht mir Mut, wenn ich eine Aufgabe schwierig finde
S4quu4	... sagt mir, wie ich es besser machen kann, wenn ich einen Fehler gemacht habe
S4quu5	... mag mich
S4quu6	... sagt mir, was ich bereits gut kann und was ich noch üben muss
S4quu7	... ist freundlich zu mir
S4quu8	... lobt mich, wenn ich etwas gut gemacht habe
S4quu9	... denkt, dass ich auch schwierige Aufgaben lösen kann

Tabelle 45

Deskriptive Ergebnisse zum unterstützenden Klassenklima

Item	*Itemnr. (Tabelle 12)*	*N*	*M*	*SD*	*Min*	*Max*
S1quu1	6	767/832	3.55	.71	1	4
S1quu3	7	763/832	3.25	.81	1	4
S1quu4	8	764/832	3.38	.72	1	4
S1quu6	9	764/832	3.09	.84	1	4
S1quu8	10	760/832	3.16	.85	1	4
S1quu9	11	760/832	3.41	.69	1	4

Anmerkung zum Antwortformat: 1=stimme gar nicht zu, 2=stimme eher nicht zu, 3=stimme eher zu, 4=stimme zu.

Kognitive Aktivierung

Die verwendete Skala stammt von Fauth et al. (2014). Bei der Übersetzung vom Englischen ins Deutsche wurde sich am unveröffentlichten Skalenbuch der IGEL-Studie von 2014 unter der Leitung von Decristan, Hertel, Klieme (damals am DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation) sowie Büttner, Hardy, Kunter und Lühken (damals an der Goethe-Universität) orientiert. Im Original wurde auf den Sachunterricht Bezug genommen (Fauth et. al., 2014), sodass die Itemformulierungen im Rahmen des INSE-Projekts auf den Kontext des Mathematikunterrichts abgestimmt wurden.

Tabelle 46

Skala: Unterrichtsqualität – Kognitive Aktivierung

Instruktion	**21) Unsere Lehrperson in Mathematik ...** *(Gib bitte an, wie sehr du den Aussagen zustimmst.)*
Anzahl Items	6
Kategorien	1 stimme gar nicht zu
	2 stimme eher nicht zu
	3 stimme eher zu
	4 stimme voll zu
Recodierte Items	-
Variablenname	**Text**
S4quk1	... fragt genau nach, was ich verstanden habe und was nicht
S4quk2	... stellt Fragen, über die ich ganz genau nachdenken muss
S4quk3	... stellt uns Aufgaben, die mir am Anfang schwierig vorkommen
S4quk4	... fragt uns, was wir zu einem neuen Thema schon wissen
S4quk5	... stellt uns Aufgaben, über die ich gerne nachdenke
S4quk6	... möchte, dass ich meine Antworten auch erklären kann

Tabelle 47

Deskriptive Ergebnisse zur kognitiven Aktivierung

Item	***Itemnr. (Tabelle 12)***	***Itemnr. (Tabelle 15)***	***N***	***M***	***SD***	***Min***	***Max***
S4quk1	12	1	765/832	3.03	.72	1	4
S4quk2	13	2	763/832	2.94	.71	1	4
S4quk3	14	3	760/832	2.93	.80	1	4
S4quk4	15	4	762/832	3.14	.89	1	4
S4quk5	16	5	759/832	2.74	.84	1	4
S4quk6	17	6	764/832	3.28	.74	1	4

Anmerkung zum Antwortformat: 1=stimme gar nicht zu, 2=stimme eher nicht zu, 3=stimme eher zu, 4=stimme zu.

Anhang C: Fragebogen zur Erhebung der im Unterricht verfolgten Lernziele und deren Gewichtung

Arithmetik & Algebra		
Zahlvorstellung entwickeln		**Beispielaufgaben** im „Logisch 4, 5, 6“
Orientierung im Zahlenraum		
	Zahlen situationsgerecht runden	Runde Zahlen des Alltags (z.B. Kosten für den Einkauf, Einwohnerzahlen) (SB Logisch 5: S. 59).
	Dezimale Zahlen und Brüche als Erweiterung des Bereichs der natürlichen Zahlen erfahren	Beispiele für dezimale Grössen im Alltag finden und erklären (SB Logisch 5: S. 33); Beispiele für Brüche im Alltag finden und erklären (SB Logisch 5: S. 46).
	Einfache Brüche der Grösse nach ordnen	Entscheide zwischen zwei Brüchen, welcher grösser oder kleiner ist (Heft Logisch 5: S. 98); vergleiche die Brüche mit einem Ganzen (1). Setze >, < oder = ein (z.B. $\frac{3}{4}$ □ $\frac{1}{2}$) (SB Logisch 6: S. 27).
	Einfluss von Zähler und Nenner auf den Wert des Bruches erkennen	Die Bestandteile eines Bruchs erläutern und die Bedeutung von Zähler und Nenner anhand eines Beispiels aus dem Alltag erklären (SB Logisch 5: S. 46).
Darstellung von Zahlen		
	Einfache Brüche in Dezimalbrüche umwandeln und umgekehrt	Überführe die als Bruch angegebene Flüssigkeitsmenge (im Messbecher) in die Dezimalschreibweise (SB Logisch 5: S. 64); bestimme den Bruchteil von einem Franken (SB Logisch 5: S. 64). Anmerkung: Die Umwandlung kann sowohl diskrepante als auch kontinuierliche Grössen umfassen, z.B. 2,065 m = 2 m $\frac{65}{1000}$ m (SB Logisch 6: S. 41).
	Den Zusammenhang zwischen Brüchen und Dezimalbrüchen erkennen	Erkläre den Zusammenhang von Brüchen und Dezimalzahlen anhand eines Beispiels (z.B. Bruchteile von 1 m; Bruchteile von 1 kg). (SB Logisch 5: S. 63; SB Logisch 6: S. 40).
	Den Aufbau des Zehnersystems verstehen	Den Aufbau der Stellenwerttafel erklären (SB Logisch 5: S. 5); erkläre die Funktion eines Kilometerzählers unter Berücksichtigung der Stellenwerttafel; erkläre das Bündeln von Mengen im Zusammenhang mit dem Zehnersystem → 1+1+1+1+1+1+1+1+1+1=10).

	Prozentangaben in Dezimalbrüche umwandeln	Wandle die folgenden Prozentangaben in Dezimalbrüche um: (z.B. 17% = 0,17; 56% = 0,56) (SB Logisch 6: S. 68).
	Prozentangaben als Hundertstel auffassen	Erkläre anhand von Schildern des Sonderverkaufs (z.B. 50% auf T-Shirts), was diese bedeuten (SB Logisch 6: S. 67).
Eigenschaften von Zahlen		
	Teilbarkeitsregeln anwenden	Teilbarkeitsregeln anwenden: Teilbarkeit durch 3: Quersumme bilden und prüfen, ob diese durch 3 teilbar ist (z.B. 6.021 : 3 → durch 3 teilbar oder nicht teilbar) (SB Logisch 6: S. 48).
	Teilbarkeitsregeln für 2, 3, 4, 5, und 10 erkennen	Teilbarkeitsregeln erklären und Argumente für die Gültigkeit dieser Regeln finden (z.B.: Warum kann man anhand der letzten Ziffer erkennen, ob eine Zahl durch 2 teilbar ist oder nicht?) (SB Logisch 6: S. 48).
	Verschiedene Aspekte der Brüche erfahren (Masszahlaspekt, Relationsaspekt, Operatoraspekt, Quotientenaspekt)	Die verschiedenen Aspekte der Brüche erklären, z.B. den Masszahlaspekt an einem Alltagsbeispiel erläutern: $\frac{1}{2}$ Apfel (SB Logisch 5: S. 46); den Buchteile von einer Grösse bestimmen (SB Logisch 5: S. 62), (Heft Logisch 5: S. 131). Erläuterungen zu den verschiedenen Aspekten der Brüche: Masszahlaspekt: $\frac{1}{2}$ von 1 Operatoraspekt: $\frac{1}{6}$ x 6 = Quotientenaspekt: $\frac{1}{6}$: 6 = Relationsaspekt: $\frac{1}{6}$: — = 1
Mit Grössen die Umwelt erfassen		
Kalender und Zeiten		
	Situationsgerechte Zeitmasse verwenden	Berechne die Zeitdauer und überführe das Ergebnis in ein angemessenes Zeitmass (in Anlehnung an SB Logisch 4: S. 52).
	Zeitmasse kennen: s, min, Monat, Jahr und Bruchteile davon	Die dargestellten Bruchteile der jeweiligen Zeitmasse auf einer Uhr erklären (SB Logisch 5: S. 48), (in Anlehnung an das Heft Logisch 5: S. 95).
	Zeitmasse in Nachbareinheiten umrechnen	(8 min 25 s = s) (in Anlehnung an SB Logisch 6: S. 24); zähle die Sekunden einer Minute (SB Logisch 4: S. 52).

	Mit Zeitangaben mündlich Grundoperationen ausführen	Anhand von Abfahrts- und Ankunftszeiten die Fahrzeit mit der Bahn berechnen (SB Logisch 5: S. 40); Berechnung der benötigen Fahrzeit mit dem Fahrrad (SB Logisch 5: S. 21); Joel steigt um 12:28 in Bruggen in den Zug und fährt nach Wil. Wann kommt er an und wie lange dauert die Fahrt? (SB Logisch 6: S. 24).
	Zeitpunkt und Zeitdauer unterscheiden	Erkläre den Unterschied zwischen Fahrtdauer und Abfahrts- bzw. Ankunftszeit anhand eines SBB-Fahrplans (in Anlehnung an SB Logisch 5: S. 40), (SB Logisch 4: S. 52 → Unterscheidung zwischen Zeitpunkt und Zeitdauer).
	Zeitplane lesen und anfertigen	Den Fahrplan der SBB lesen (SB Logisch 5: S. 40).
Dezimale Grössen		
	Längen, Flächen, Rauminhalte, Massen und Temperauren schätzen und messen.	Schätze die Grösse von Personen (SB Logisch 5: S. 8); vergleiche die Grösse und das Gewicht von Hühnerei und Straussenei (SB Logisch 5: S. 9); bestimme die Fläche von Figuren und vergleiche diese miteinander (Heft Logisch 5: S. 61); berechne den Flächeninhalt eines Rechtecks (SB Logisch 6: S. 31); bestimme das Volumen von Würfeln (SB Logisch 6: S. 57).
	Einheiten für Länge, Fläche, Volumen, Massen und Temperaturen kennen	Erkläre, was unter einer Längeneinheit bzw. unter einer Flächeneinheit zu verstehen ist (SB Logisch 5: S. 32; SB Logisch 6: S. 30); erkläre den Unterschied zwischen cm, dm, m und km (SB Logisch 5: S. 33).
	Grundoperationen mit dezimalen Grössen ausführen	Berechne die Gesamtlänge (z.B. 19,125 m + 2,754 m + 0,459 m + 7 m = m) (SB Logisch 5: S. 35); 8 x 375 ml = ml (SB Logisch 5: S. 36).
	Bedeutung der Vorsilben Milli-, Zenti-, Kilo- kennen	Erkläre, wofür die Vorsilben Milli-, Zenti- und Kilo- stehen? Julia meint, 75 kg sind 0,75 t. Hat Julia Recht? Erkläre deine Antwort.
	Grössenangaben situationsgerecht runden	Runde die Einwohnerzahlen (SB Logisch 5: S. 59); runde die Geldbeträge (Heft Logisch 5: S. 123); runde die Zahlen auf die nachfolgend angegebene Stelle (Zehner, Hunderter etc.) (Heft Logisch 5: S. 123).
	Zweifach benannte Grössen als Dezimalzahlen darstellen	1450 kg = 1 t 450 g = 1,450 t (SB Logisch 5: S. 33); 25 kg – 6 kg 70 g = kg (SB Logisch 5: S. 36).
	Bruchteile von Grössen erkennen	Erkläre, wie Brüche im Alltag vorkommen (SB Logisch 5: S. 46).
Zusammengesetzte Grössen		
	Zusammengesetzte Grössen vergleichen.	Vergleiche den Benzinverbrauch (l/km) von unterschiedlichen Autos (Heft Logisch 5: S. 104); stelle die unterschiedlichen km/h-Angaben gegenüber und vergleiche (Heft Logisch 5: S. 58). Anmerkung: Das genannte Beispiel passt auch in die Kategorie „Funktionen/Relationen“.

	Zusammengesetzte Grössen kennen.	Erkläre die Zusammensetzung von Grössen anhand eines Beispiels (z.B. km/h) und erläutere ihre Bedeutung (in Anlehnung an das Heft Logisch 5: S. 54).
Operationen verstehen und ausführen		
Addition und Subtraktion		
	Summen und Differenzen von Dezimalzahlen schätzen und berechnen	Schätze die Kosten für einen Einkauf und berechne anschliessend die genaue Summe des Einkaufs (in Anlehnung an SB Logisch 5: S. 8; SB Logisch 5: S. 35).
	Einsicht in den Zusammenhang zwischen dem Stellenwertsystem und den schriftlichen Rechenverfahren gewinnen	Stelle die Additionsaufgabe 138 + 145 mithilfe von Zehnerblöcken (Hunderterplatte, Zehnerstange und Würfel) dar und vergleiche die Darstellung mit dem schriftlichen Additionsverfahren; erkläre anhand der Gegenüberstellung den Zusammenhang zwischen dem Stellenwertsystem und dem schriftlichen Rechenverfahren (in Anlehnung an SB Logisch 5: S. 11), (Padberg & Benz, 2011, S. 226). Findest du Regelmässigkeiten in den Zahlen? Kannst du diese erklären? a) 1221 + 2112 b) 5445 + 4554 c) 7117 + 1771 Quelle: http://pikas.dzlm.de/upload/Material/Haus_7_-_Gute_-_Aufgaben/UM/Umkehrzahlen/Schueler-Material/Variationen/UM_UZ_Schueler_ANNA.pdf
	Rechengesetze zur Vereinfachung von Termen und zur Gewinnung von Rechenvorteilen anwenden	Anwendung von Rechengesetzen zur Vereinfachung von Termen, z.B. die Subtraktion ist die Umkehroperation der Addition (Heft Logisch 5: S. 24); die Anwendung von Rechengesetzen zur Vereinfachung von Termen kann sich z.B. auf das Kommutativgesetz beziehen.
	Rechengesetze als Grundlage der schriftlichen Addition und Subtraktion erkennen	Vergleiche die beiden Rechenverfahren und erkläre, inwiefern das Distributivgesetz bei der schriftlichen Addition eine Rolle spielt. 57 + 84 = 141 57 + 84 = (5 x 10 + 7 x 1) + (8 x 10 + 4 x 1) ➔ Mit Kommutativ- und Assoziativgesetz umordnen = 5 x 10 + 8 x 10 + 7 x 1 + 4 x 1 ➔ Anwendung des Distributivgesetzes = (5 + 8) x 10 + (7 + 4) x 1 = 13 x 10 + 11 x 1 Anmerkung: Aufgaben, die sich auf das Kommutativ-, Assoziativ- oder Distributivgesetz beziehen und im Zusammenhang mit der schriftlichen Addition stehen.

	Eigene schriftliche Rechenverfahren entwickeln und mit den Normalverfahren vergleichen	Eigene Lösungswege für eine Aufgabe zum schriftlichen Subtrahieren suchen, ausprobieren und vergleichen (SB Logisch 5: S. 11).
	Brüche kürzen und erweitern	Mit welchem Faktor wurde erweitert? (z.B. $\frac{3}{8} = \frac{6}{16}$) Mit welchem Divisor wurde gekürzt? (z.B. $\frac{20}{50} = \frac{2}{5}$) (SB Logisch 6: S. 26).
	Einfluss von Veränderungen im Zähler und Nenner auf den Wert von Brüchen erkennen	Was passiert mit einem Bruch, wenn der Zähler (Nenner) grösser (kleiner) wird (z.B. Kuchenstücke und Anzahl an Personen).
	Einfache Brüche addieren und subtrahieren	Addiere die Brüche, indem du diese in Dezimalschreibweise umwandelst (Heft Logisch 5: S. 132); addiere gleichnamige Brüche (SB Logisch 6: S. 33).
	Addition und Subtraktion von gleichnamigen Brüchen mit derjenigen von Grössen gleicher Einheit verbinden	(z. B. $\frac{3}{4}l + \frac{1}{4}l = \quad l$)
	Gleichnamigkeit als Voraussetzung für die Addition und Subtraktion erkennen	Erkläre, warum ungleichnamige Brüche nicht einfach addiert werden können, sondern zunächst gleichnamige Nenner sicherzustellen sind (in Anlehnung an SB Logisch 6: S. 33).
Multiplikation & Division		
	Produkte und Quotienten von Dezimalzahlen schätzen, mit und ohne Taschenrechner berechnen	Berechne die Kosten für drei Helme (3 x 43.50 Fr.) (SB Logisch 5: S. 36).
	In halbschriftlichen Verfahren die Zerlegbarkeit der Operationen im Stellenwertsystem erkennen	Erkläre anhand eines Beispiels die Zerlegbarkeit der Operatoren im Stellenwertsystem (z.B. 23 x 36) (SB Logisch 5: S. 28).
	Produkte mit einem zweistelligen Faktor, Quotienten mit zweistelligem Divisor schriftlich berechnen	Löse die Multiplikationsaufgabe mit zweistelligem Faktor schriftlich (z.B. 32 x 145) (SB Logisch 5: S. 29).
	Das Distributivgesetz als Grundlage der schriftlichen Multiplikation und Division erkennen.	Erkläre die Anwendung des Distributivgesetzes am Beispiel der folgenden Aufgabe: 23 x 15 = (20 + 3) x (10 + 5) (SB Logisch 5: S. 28).
	Eigene schriftliche Rechenverfahren entwickeln und mit den Normalverfahren vergleichen	Entwickle verschiedene Verfahren der Multiplikation und probiere diese aus. Was stellst du fest, wenn du deine Verfahren mit dem Normalverfahren vergleichst? (SB Logisch 5: S. 28), (Heft Logisch 5: S. 86).

	Grössenordnungen beim Operieren mit Zehnerpotenzen erfahren	Löse die folgende Multiplikationsaufgabe/Divisionsaufgabe (z.B. 64‘000 : 8‘000 =). Was stellst du fest? (SB Logisch 5: S. 7). Wie viele Kühe bräuchte man, um die ganze Schule eine Woche lang mit Milch zu versorgen (Fermi-Aufgabe) Quelle: DZLM (o. D.)
	Stufenfolge der Operationen kennen	Erkläre und begründe, in welcher Reihenfolge du bei einer Rechenaufgabe vorgehen musst, die sowohl Addition/Subtraktion als auch Multiplikation/Division beinhaltet (Rechenregel: Punkt-vor-Strich) (SB Logisch 5: S. 58); (z.B. 30 – 5 x 6) (Heft Logisch 5: S. 118).
Terme		
	Terme in die Umgangssprache übersetzen und umgekehrt	Terme als Zahlenrätsel, z.B.: Addiere zur Differenz von 83 und 17 die Zahl 9 (Heft Logisch 5: S. 121); denk dir eine Textaufgabe aus, aus der der folgende Term resultiert: 46 : 2 + 18 (in Anlehnung an SB Logisch 6: S. 50).
	Aufbau von Termen mit einer Variablen durchschauen	Erkläre, wofür die Variable x in einem Term steht? (in Anlehnung an SB Logisch 6: S. 49).
	Terme mit einer Variablen bilden und vereinfachen	27 + 45 = x (SB Logisch 6: S. 50).
	Rechengesetze als Regeln zur Umformung und Vereinfachung von Termen anwenden	Rechnen mit Klammern (SB Logisch 5: S. 58), (Heft Logisch 5: S. 122).
Gleichungen		
	Einfache Gleichungen aus Sachzusammenhängen gewinnen	Auf einer Waage, die sich im Gleichgewicht befindet, sind mehrere Pakete abgelegt. Auf einem Paket ist das Gewicht nicht mehr lesbar. Wie schwer ist das Paket? Anmerkung: Die Gewichte der anderen Pakete sind einer Grafik zu entnehmen (SB Logisch 6: S. 49).
	Zusammenhang zwischen Sachverhalt und Zahlverhalt erkennen	Warum ist die Waage ein gutes Beispiel zur Erklärung des Prinzips einer Gleichung? Suche nach einer Erklärung und tausche dich mit deinem Sitznachbarn aus.
	Mit einem Umformungsschritt auflösbare Gleichungen lösen	528 g + x =728 g + 126 g (SB Logisch 6: S. 49).