Instruktionssensitivität von Tests und Items [Instructional Sensitivity of Tests and Items]

Alexander Naumann[1,3], Stephanie Musow[2,3], Christine Aichele[1], Jan Hochweber[2,3] & Johannes Hartig[1,3]

This is a preprint, the definitive version is available at https://link.springer.com/article/10.1007/s11618-018-0832-0)

1) Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
2) Pädagogische Hochschule St. Gallen (PHSG)
3) IDeA Forschungszentrum Frankfurt

Please cite as:

Naumann, A., Musow, S., Aichele, C., Hochweber, J., & Hartig, J. (2019). Instruktionssensitivität von Tests und Items [Instructional Sensitivity of Tests and Items]. *Zeitschrift für Erziehungswissenschaft*. *22*(1), 181–202.

Alexander Naumann ist wissenschaftlicher Mitarbeiter im Arbeitsbereich Educational Measurement der Abteilung Bildungsqualität und Evaluation am Deutschen Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, 60486 Frankfurt am Main, Deutschland, Naumanna@dipf.de

Stephanie Musow ist wissenschaftliche Mitarbeiterin am Institut für Kompetenzdiagnostik an der Pädagogischen Hochschule St. Gallen (PHSG), Notkerstraße 27, 9000 St. Gallen, Schweiz, Stephanie.Musow@phsg.ch.

Christine Aichele ist wissenschaftliche Mitarbeiterin im Arbeitsbereich Educational Measurement der Abteilung Bildungsqualität und Evaluation am Deutschen Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, 60486 Frankfurt am Main, Deutschland, Aichele@dipf.de

Jan Hochweber ist Professor und Leiter des Instituts für Kompetenzdiagnostik an der Pädagogischen Hochschule St. Gallen (PHSG), Notkerstraße 27, 9000 St. Gallen, Schweiz, Jan.Hochweber@phsg.ch

Johannes Hartig ist Professor für Educational Measurement in der Abteilung Bildungsqualität und Evaluation am Deutschen Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, 60486 Frankfurt am Main, Deutschland, Hartig@dipf.de

Abstract

Testergebnisse von Schülerinnen und Schülern dienen regelmäßig als ein zentrales Kriterium für die Beurteilung der Effektivität von Schule und Unterricht. Gültige Rückschlüsse über Schule und Unterricht setzen voraus, dass die eingesetzten Testinstrumente mögliche Effekte des Unterrichts auffangen können, also instruktionssensitiv sind. Jedoch wird diese Voraussetzung nur selten empirisch überprüft. Somit bleibt mitunter unklar, ob ein Test nicht instruktionssensitiv oder ein Unterricht nicht effektiv war. Die Klärung dieser Frage erfordert die empirische Untersuchung der Instruktionssensitivität der eingesetzten Tests und Items.

Während die Instruktionssensitivität in den USA bereits seit Langem diskutiert wird, findet das Konzept im deutschsprachigen Diskurs bislang nur wenig Beachtung. Unsere Arbeit zielt daher darauf ab, das Konzept Instruktionssensitivität in den deutschsprachigen Diskurs über schulische Leistungsmessung einzubetten. Dazu werden drei Themenfelder behandelt, (a) der theoretische Hintergrund des Konzepts Instruktionssensitivität, (b) die Messung von Instruktionssensitivität sowie (c) die Identifikation von weiteren Forschungsbedarfen.

Keywords: Instruktionssensitivität, Validität, Unterrichtseffektivität

Abstract

Students' performance in assessments is regularly attributed to more or less effective teaching. Valid interpretation requires that outcomes are affected by instruction to a significant degree. Hence, instruments need to be capable of detecting effects of instruction, that is, instruments need to be instructionally sensitive. However, empirical investigation of the instructional sensitivity of tests and items is seldom in practice. In consequence, in many cases, it remains unclear whether teaching was ineffective or the instrument was insensitive.

While there is a living discussion on the instructional sensitivity of tests and items in the USA, the concept of instructional sensitivity is rather unknown in German-speaking countries. Thus, the present study aims at (a) introducing the concept of instructional sensitivity, (b) providing an overview on current approaches of measuring instructional sensitivity, and (c) identifying further research directions.

Keywords: Instructional sensitivity, validity, educational effectiveness

Instruktionssensitivität von Tests und Items [Instructional Sensitivity of Tests and Items]

Leistungsmessungen bei Schülerinnen und Schülern erfolgen üblicherweise in der Erwartung, aus den empirischen Daten nützliche Informationen für pädagogische und politische Entscheidungen ableiten zu können (Weinert, 2001; Hartig et al., 2008). Sie bilden das Kernstück einer Output-orientierten und evidenzbasierten Steuerung im Bildungswesen (Fend, 2011). Evidenzbasierte Praxis und Politik bedürfen allerdings der Entwicklung adäquater Messinstrumente als Grundlage für gültige Rückschlüsse über die Fähigkeiten von Schülerinnen und Schülern sowie die Qualität von Bildungseinrichtungen und –maßnahmen (Klieme und Leutner, 2006). Gültige Rückschlüsse über individuelle Fähigkeiten von Schülerinnen und Schülern sowie über die Qualität von Bildungseinrichtungen und –maßnahmen erfordern jeweils spezifische empirische Evidenz (AERA et al., 2014).

Im schulischen Kontext spielt je nach Zweck der Leistungsmessung die Betrachtung des Zusammenhangs der Testinstrumente mit dem, was die Schülerinnen und Schüler im Unterricht lernen sollen (intendiertes Curriculum), und/oder dem, was tatsächlich im Unterricht vermittelt wird (implementiertes Curriculum), eine besondere Rolle (Pellegrino, 2002). Testwerte können zum Beispiel hinsichtlich des Lernstands innerhalb eines Schulfaches oder bezüglich des Lernfortschritts aufgrund des Unterrichts interpretiert werden. Für solche Testwertinterpretationen stellt sich die Frage nach der Passung von (1) Test, (2) intendiertem sowie (3) implementiertem Curriculum (Porter, 2002). Informationen zur Passung dieser drei Elemente dienen als empirische Evidenz zur Unterstützung der jeweiligen Testwertnutzung und -interpretation (Kane, 2013).

Vor allem die Passung von Test und implementiertem Curriculum wird regelmäßig kontrovers diskutiert (Popham, 2007). Insbesondere im US-amerikanischen Raum stellte sich mit der flächendeckenden Einführung von high-stakes Accountability-Systemen im Zuge des *No Child Left Behind Acts* und der damit einhergehenden Beurteilung des Erfolgs von Schule

und Unterricht auf Basis der Schülerergebnisse in standardisierten Leistungstests die Frage, ob und in welchem Ausmaß Tests überhaupt dazu in der Lage sind, Effekte von Schule und Unterricht zu erfassen, also instruktionssensitiv sind (Polikoff, 2010). Zwar wird die Instruktionssensitivität der Tests häufig implizit angenommen, jedoch nur selten empirisch überprüft (D'Agostino et al., 2007).

Im deutschsprachigen Raum ist die Bedeutung schulischen Wissens für den Erfolg in Leistungstests ebenfalls umstritten (z.B. Arnold, 2005; Baumert et al., 2007; Rindermann, 2006). Die Passung von Test und implementiertem Curriculum – und damit Instruktionssensitivität und deren empirischer Überprüfung – findet bislang allerdings kaum Beachtung. Ein möglicher Grund liegt in der vergleichsweise kurzen Tradition flächendeckender standardisierter Leistungsmessungen in Schulen, welche erst im Zuge des nur mittelmäßigen Abschneidens der deutschen Schülerinnen und Schüler bei TIMSS 1995 und PISA 2000 an Bedeutung gewann (Weinert, 2001). Deutschland hatte sich bis zu den 90er Jahren weitgehend bei internationalen Vergleichsstudien enthalten (Drechsel et al., 2015). Die testdatenbasierte Schul- und Unterrichtsentwicklung ist daher noch vergleichsweise jung (Ramsteck und Maier, 2015; Altrichter et al., 2017). Ein zweiter Grund ist in der unterschiedlichen Verwendung der Testwerte und den damit verbundenen Konsequenzen zu vermuten. Während in den USA oftmals *high-stakes*-Testing eingesetzt wird, sind Tests im deutschen Bildungssystem in der Regel nicht mit vergleichbaren Konsequenzen für Schulen und Lehrpersonen verbunden (*Low-stakes*-Testing; z.B. Grünkorn et al., im Druck; Maag Merki, 2017). Altrichter, Moosbrugger und Zuber (2017) sprechen im Zusammenhang von *low-stakes*-Testing von einem evidenzbasierten Steuerungssystem, das der Überwachung und Weiterentwicklung von Bildungssystem, Schule und Unterricht dient.

Vor diesem Hintergrund bilden die Testdaten der Schülerinnen und Schüler in Deutschland in erster Linie die Grundlage für eine empiriegestützte Qualitätsentwicklung im

Sinne eines Systemmonitorings beziehungsweise einer Schulevaluation (Grünkorn et al., im Druck; Kultusministerkonferenz, 2006) oder sie dienen als abhängige Variablen innerhalb wissenschaftlicher Studien (z.B. Interventionen; Hascher und Schmitz, 2010). Beispielsweise sollen die in Deutschland flächendeckend eingeführten Vergleichsarbeiten (VERA; Kultusministerkonferenz, 2006) als ein Instrument des Bildungsmonitorings auf der Ebene der Schule und des Unterrichts ansetzen und sowohl der Bestandsaufnahme als auch der Vorbereitung pädagogischer und didaktischer Entscheidungen dienen, indem (a) Lehrpersonen Hinweise zur Unterrichtsreflexion und zu Handlungsbedarfen und (b) Schulen eine empirische Datenbasis für die Selbstevaluation erhalten (Spoden und Leutner, 2011; Ramsteck und Maier, 2015). In wissenschaftlichen Studien liegt der Fokus dagegen auf der Identifikation und Untersuchung von Merkmalen von Schule und Unterricht, die das Lernen der Kinder ermöglichen, fördern oder erschweren (Klieme, 2008). Dementsprechend nehmen Studien der empirischen Schul- und Unterrichtsforschung regelmäßig die Rolle von Kontext-, Bedingungs- und Prozessmerkmalen von Unterricht und ihre Wirkung auf den Lernertrag von Schülerinnen und Schülern in den Blick (z.B. Klieme, Pauli und Reusser, 2009). Auf ähnliche Weise beurteilen (quasi-)experimentelle Interventionsstudien im Unterricht für gewöhnlich die Wirkung ihrer Interventionsmaßnahmen auf das Lernen der Schülerinnen und Schüler anhand der erzielten Testwerte (z.B. Decristan et al., 2015). Das heißt sowohl Bildungspolitik als auch -wissenschaft erwarten, auf Basis der Testwerte empirisch fundierte Rückschlüsse über die von Schülerinnen und Schülern im Unterricht erworbenen Fähigkeiten und Kompetenzen ziehen zu können (vgl. Stanat und Pant, 2016). Testwerte bilden damit einerseits die Grundlage einer Output-orientierten Steuerung des Bildungswesens in Deutschland, andererseits stellt die wissenschaftliche Erklärung von Leistungszuwächsen einen zentralen Baustein zur Effizienzsteigerung schulischer Maßnahmen dar (Klieme, 2008).

Auch diese Testwertinterpretationen setzen voraus, dass die eingesetzten Testinstrumente in der Lage sind, Unterrichtseffekte zu erfassen.

Dass Rückschlüsse über die Effektivität von Unterricht aufgrund von Testeigenschaften variieren können, zeigten Grossman und Kollegen (2014) für den empirischen Zusammenhang zwischen Unterrichtsqualität und Schülerleistung, wenn Testwerte verschiedener Tests herangezogen werden. So hingen Unterrichtsqualität und Testwerte nur in bestimmten Tests positiv zusammen, während für andere Tests keine statistisch bedeutsamen Zusammenhänge erkennbar waren. Tests, die dasselbe Konstrukt abbilden sollen, können also unterschiedlich instruktionssensitiv sein – und damit zu unterschiedlichen Rückschlüssen über Unterricht führen.

Oft bleibt jedoch unklar, ob ein Test nicht instruktionssensitiv oder der Unterricht nicht effektiv war. In Deutschland gibt es beispielsweise eine Reihe von Leistungstests, mit denen entgegen der Erwartungen keine oder nur geringe Kompetenzzuwächse nachgewiesen werden konnten (z.B. Fischer et al., 2016; Lossen et al., 2016; Nagy et al., 2017). In diesen Fällen ist offen, ob die Schülerinnen und Schüler tatsächlich keinen Leistungszuwachs zu verzeichnen haben oder ob die eingesetzten Instrumente nicht in der Lage sind, Effekte von Unterricht auf die Schülerleistungen zu erfassen. Die Klärung dieser Frage erfordert die empirische Untersuchung der Instruktionssensitivität der eingesetzten Tests und Items, die hierzulande häufig ausbleibt.

Unsere Arbeit zielt daher darauf ab, das Konzept Instruktionssensitivität in den deutschsprachigen Diskurs über schulische Leistungsmessung einzubetten. Dazu werden im Folgenden drei Themenfelder behandelt:

a) der theoretische Hintergrund des Konzepts Instruktionssensitivität
b) die Messung von Instruktionssensitivität
c) die Identifikation von weiteren Forschungsbedarfen

Theoretischer Hintergrund und Herkunft des Konzepts Instruktionssensitivität

Polikoff (2010) definiert Instruktionssensitivität als die psychometrische Eigenschaft eines Tests oder einzelnen Items, Effekte von Unterricht zu erfassen. Spezifisch geht es um die Sensitivität hinsichtlich der Unterrichtsqualität und der vermittelten Lerninhalte. Das Verständnis von Instruktionssensitivität war jedoch nicht immer einheitlich, sondern unterlag seit den 1960er Jahren maßgeblichen Veränderungen.

Herkunft des Konzepts Instruktionssensitivität

Erste Überlegungen zur Instruktionssensitivität kamen mit der wachsenden Bedeutung kriterienorientierter Tests ab der Mitte der 1960er Jahre auf (D'Agostino et al., 2007). Kriterienorientierte Tests zielen im Gegensatz zu normorientierten Tests auf den individuellen Stand einer Person hinsichtlich eines Lernziels ab (Millman, 1970). Damit rückten auch die Frage nach der Wirkung des Unterrichts und die Messung seiner Effekte auf die kriterienbezogenen Lernfortschritte der Schülerinnen und Schüler stärker in den Vordergrund. Klassische Indizes zur Itemselektion wie Schwierigkeit und Trennschärfe sollten daher durch neue Sensitivitätsindizes ergänzt werden, die erfassen, inwiefern ein Item zwischen unterrichteten und nicht-unterrichteten Schülerinnen und Schülern differenzieren kann (Cox und Vargas, 1966). Das Ausmaß dieser Itemeigenschaft definierte man als Instruktionssensitivität oder Itemsensitivität (Kosecoff und Klein, 1974). Haladyna und Roid (1981) präzisierten diese Beschreibung und definierten Instruktionssensitivität als die Tendenz eines Items, in Abhängigkeit vom Unterricht in der Schwierigkeit zu variieren. Nach diesem Verständnis ist Instruktionssensitivität ein Konzept, das die Bedeutung von Unterricht für die Lösungswahrscheinlichkeit eines einzelnen Items hervorhebt. Allerdings berücksichtigt diese Definition nur einzelne Items, nicht gesamte Tests, und bezieht darüber hinaus keine Merkmale mit ein, welche die Quantität oder Qualität des Unterrichts abbilden.

Tests rückten erst ab den 1980er Jahren in den Fokus. Ausgangspunkt war eine gerichtliche Auseinandersetzung darüber, ob Schülerinnen und Schüler in Florida Unterricht in den für das Bestehen notwendigen Inhalten des bundesstaatlichen High-School-Abschlusstests erhielten (United States Court of Appeals, 1981). Charakteristisch war die Frage nach Chancengleichheit und der Angemessenheit des implementierten Curriculums für das erfolgreiche Abschneiden in standardisierten Abschlussprüfungen (McClung, 1979; Yoon und Resnick, 1998). Anstelle der Itemselektion stand die Verknüpfung von Test und Unterricht stärker im Mittelpunkt, also die *Instruktionsvalidität* (Airasian und Madaus, 1983). Instruktionsvalidität bezieht sich auf die Frage, ob und inwiefern Unterricht zur Testleistungen in standardisierten Tests beiträgt (z.B. Mehrens und Philips, 1987). Gleichzeitig stellte sich die Frage nach einem fairen Vergleich der Leistungen von Schülerinnen und Schülern, die einen unterschiedlichen Unterricht erfahren hatten (Muthén, 1989). Zwar war die Instruktionssensitivität als eine psychometrische Eigenschaft der Testinstrumente in der Diskussion eher nachrangig (Polikoff, 2010), doch wurde *Instruktionsbias* (Linn und Harnisch, 1981) in Testaufgaben als eher vorteilhaft angesehen, wenn es um Rückschlüsse über Unterricht ging.

Spätere Arbeiten griffen Gedanken zur Itemsensitivität und zur Instruktionsvalidität auf und bildeten die Grundlage für das heutige Verständnis von Instruktionssensitivität. Besonders Burstein (1989) sowie Muthén und Kollegen (1991) führten systematisch beide Denkrichtungen zusammen. Sie stellten einerseits den ursprünglichen Gedanken von Instruktionssensitivität als messbare Eigenschaft eines Tests beziehungsweise eines einzelnen Items wieder in den Mittelpunkt und verwiesen andererseits auf die Bedeutung von Instruktionssensitivität für die Interpretation des gemessenen Konstrukts vor dem Hintergrund schulischen Unterrichts. Instruktionssensitivität wurde in der Folge einerseits als Validitätshinweis für Rückschlüsse über Unterricht angesehen (z.B. Popham, 2007),

andererseits jedoch ebenso als Beeinträchtigung der Testfairness bei Rückschlüssen über individuelle Schülerleistungen (z.B. Geisinger und McCormick, 2010).

Aktuelle Arbeiten stellen dagegen heraus, dass Instruktionssensitivität nicht notwendig die Verletzung von Testfairness oder anderer Annahmen von Messinvarianz erfordert (Naumann et al., 2017). Die Rolle von Instruktionssensitivität als ein essentielles Validitätsargument für gültige Rückschlüsse über Schule und Unterricht bleibt davon unberührt. Auf die Rolle von Instruktionssensitivität für die Testwertinterpretation wird im Folgenden näher eingegangen.

Bedeutung von Instruktionssensitivität für die Testwertinterpretation

Schulische Leistungsmessungen stehen in der Erwartung, gültige Rückschlüsse über individuelle Fähigkeiten von Schülerinnen und Schüler als auch über die Qualität von Schule und Unterricht zu erlauben. Beispielsweise kann eine Interpretation der Testwerte hinsichtlich (a) der Leistungsfähigkeit einer Schülerin oder eines Schülers, (b) des Lernstands innerhalb einer Domäne oder eines Schulfaches oder aber (c) des Lernfortschritts aufgrund der Qualität von Schule und Unterricht gewünscht werden. Die valide Nutzung und Interpretation von Testwerten aus schulischen Leistungsmessungen hinsichtlich der individuellen Fähigkeiten von Schülerinnen und Schülern einerseits oder der Qualität von Schule und Unterricht andererseits erfordern jedoch jeweils spezifische empirische Evidenz (AERA et al., 2014). Grundlage für eine gültige Testwertinterpretation ist demnach je nach Fragestellung der Grad der Passung von (1) Test, (2) intendiertem sowie (3) implementiertem Curriculum. Abbildung 1 stellt das Verhältnis dieser drei Elemente graphisch in Form eines Dreiecks dar, wobei die Seiten des Dreiecks die Passung zwischen jeweils zwei Elementen beschreiben (adaptiert nach Anderson, 2002; Pellegrino, 2002).

[Abbildung 1 hier]

Sollen Testwerte den Grad widerspiegeln, in dem Schülerinnen und Schüler ein definiertes Lernziel erreicht haben, liefert die Passung von Test und intendiertem Curriculum Argumente für diese Interpretation (curriculare Validität; Hartig et al., 2012). Empirische Hinweise für curriculare Validität lassen sich beispielsweise durch den Abgleich von Testmaterial und formalen Dokumenten wie Lehrplänen ermitteln (AERA et al., 2014). Ein solcher Abgleich kann im Prinzip vor, während oder nach der eigentlichen Testung erfolgen.

Werden dagegen Interpretationen angestrebt, welche die Testwerte auf den von Schülerinnen und Schülern erhaltenen Unterricht zurückführen, zum Beispiel im Rahmen von Bildungsmonitoring oder Unterrichtsinterventionen, geht es um die Passung von Test und implementiertem Curriculum. Für solche Testwertinterpretationen sind empirische Maße zur Instruktionssensitivität als Validitätsevidenz besonders relevant (Popham, 2007; Yoon und Resnick, 1998). Nur wenn Tests instruktionssensitiv sind, ist sichergestellt, dass Unterrichtseffekte gültig interpretierbar sind. Bleibt beispielsweise ein erwarteter Unterrichtseffekt aus, ist andernfalls unklar, ob ein Unterricht ineffektiv oder der eingesetzte Test nicht sensitiv war (Naumann et al., 2016). Oftmals wird Instruktionssensitivität durch das Vorhandensein empirischer Belege für curriculare Validität implizit angenommen. Curriculare Validität bezieht sich allerdings auf das intendierte und damit nicht zwangsläufig implementierte Curriculum. Die curriculare Validität ist somit zwar eine notwendige, aber keine hinreichende Bedingung für Instruktionssensitivität. Instruktionssensitivität bezieht sich auf die Passung von Test und dem tatsächlich im Unterricht implementierten Curriculum. Das tatsächlich implementierte Curriculum und die Qualität dieser Implementation ist letztlich ausschlaggebend für den Beitrag des Unterrichts am Zustandekommen der Testwerte. Eine gültige Interpretation der Testwerte bezüglich der Effektivität von Schule und Unterricht wie im Bildungsmonitoring oder in empirischen Studien erfordert also die Messung von Instruktionssensitivität.

Die Messung von Instruktionssensitivität

Zentral für die Messung von Instruktionssensitivität ist die Beobachtung von Veränderung im Antwortverhalten von Schülerinnen und Schülern in Abhängigkeit vom erhaltenen Unterricht (Burstein, 1989). Ansätze zur Messung von Instruktionssensitivität greifen auf drei Datenquellen zurück (Polikoff, 2010): a) empirische Testdaten (Testwerte oder Itemantworten), b) empirische Maße über den Inhalt und die Qualität eines Unterrichts, sowie c) Expertenurteile. Auf dieser Basis findet entweder eine Beurteilung der Instruktionssensitivität eines gesamten Tests oder eines einzelnen Items statt.

Messung der Instruktionssensitivität von Tests und Items

Analysen der Instruktionssensitivität von Tests beziehen üblicherweise Testwerte und empirische Unterrichtsmaße als Datenquellen ein. Als Unterrichtsmaße dienen zum Beispiel die Abdeckung oder die Gewichtung von Lerninhalten (D'Agostino et al., 2007; Greer, 1995) oder die Unterrichtsqualität (Grossman et al., 2014). Die Erwartung ist, dass die Testwerte mit mehr oder höherwertigem Unterricht ansteigen (Baker, 1994). Die Prüfung dieser Annahme erfolgt in der Regel mittels hierarchisch linearer Modelle (Raudenbush und Bryk, 2002) mit den Testwerten als abhängiger Variable und den Unterrichtsmerkmalen als Prädiktoren. Zeigen die Unterrichtsmerkmale einen statistisch bedeutsamen positiven Zusammenhang mit den Testwerten, so gilt der Test als instruktionssensitiv (z.B. Ing, 2008).

Analysen der Instruktionssensitivität einzelner Items nutzen entweder die Itemantworten oder die Urteile geschulter Expertinnen und Experten. Expertenurteile zur Instruktionssensitivität eines Items können auf drei Arten erfolgen: a) global, das heißt anhand eines einzelnen Indikators (z.B. Chen, 2012), b) anhand mehrerer Indikatoren, die zu einem Gesamturteil führen (z.B. Popham, 2007; Popham und Ryan, 2012), oder c) anhand mehrerer Indikatoren, die ein differenziertes Urteil erlauben (Musow et al., 2018). Beispielsweise schlägt Popham (2007) ein Gesamturteil über die Instruktionssensitivität eines

Items mittels drei dichotomer Indikatoren vor: (1) der Einfluss des sozioökonomischen Status auf die Lösungswahrscheinlichkeit einer Aufgabe, (2) die Rolle der individuellen Begabung des Kindes, sowie (3) der Einfluss des Unterrichts auf die Lösungswahrscheinlichkeit. Ein Item wird dann als instruktionssensitiv angesehen, wenn die ersten beiden Indikatoren negativ beurteilt werden und dem dritten Indikator zugestimmt wird. Zwar betont Polikoff (2010) das Potential von Expertenurteilen, jedoch sind sie bisher kaum validiert und nur selten praktisch angewendet.

Auf Itemantworten basierende Ansätze bestimmen Instruktionssensitivität mittels Itemstatistiken. Seit Mitte der 1960er Jahre wurde eine Vielzahl von Maßen vorgeschlagen (für eine ausführliche Übersicht siehe Haladyna und Roid, 1981; Polikoff, 2010), die in der Regel auf der Trennschärfe oder der Itemschwierigkeit fußen (Haladyna, 2004). Im Wesentlichen leiten sich diese itemstatistischen Ansätze aus einer von zwei Traditionen ab: a) der Entwicklung von Itemstatistiken, die ursprünglich im Kontext des kriterienorientierten Testens traditionelle Itemanalyseverfahren ersetzen oder ergänzen sollten (z.B. Kosecoff und Klein, 1974), oder b) aus der Untersuchung von Differential Item Functioning (DIF; Holland und Wainer, 1993) aufgrund des Lernkontextes der Schülerinnen und Schüler (z.B. Clauser et al., 1996; Linn und Harnisch, 1981). DIF-Untersuchungen zur Instruktionssensitivität trugen maßgeblich dazu bei, die Verletzung von Messinvarianzannahmen als eine notwendige Voraussetzung für Instruktionssensitivität anzusehen (Naumann et al., 2014). Aus psychometrischer Sicht ist DIF jedoch keine Voraussetzung für Instruktionssensitivität (Naumann et al., 2017).

Allgemein werden Items dann als besonders instruktionssensitiv angenommen, wenn sich die Itemparameter stark über Messzeitpunkte verändern (z.B. Cox und Vargas, 1966) oder über Lerngruppen hinweg variieren (z.B. Robitzsch, 2009). Allerdings führt die gleichzeitige Anwendung dieser beiden Herangehensweisen zur Beurteilung der

Instruktionssensitivität eines Items nicht zu konsistenten Ergebnissen, da sie sich, wie nachfolgend dargestellt, auf unterschiedliche Varianzquellen (Naumann et al., 2016) und Hypothesen bezüglich der Itemsensitivität (Naumann et al., 2017) beziehen.

Ein psychometrischer Rahmen zur Messung von Instruktionssensitivität

Die Vielzahl an Verfahren zur Operationalisierung von Instruktionssensitivität und deren inkonsistenten Ergebnisse lassen leicht den Eindruck einer Fragmentierung entstehen (vgl. Polikoff, 2010). Zur Systematisierung der bisherigen Ansätze zur Messung von Instruktionssensitivität entwickelten Naumann, Hartig und Hochweber (2017) daher einen psychometrischen Rahmen (Abbildung 2). Dieser Rahmen erlaubt die Kategorisierung bestehender Ansätze und stellt deren Gemeinsamkeiten und Unterschiede heraus.

[Abbildung 2 hier]

Der Rahmen unterscheidet drei Perspektiven auf Instruktionssensitivität, die sich auf die unterschiedlichen Varianzquellen beziehen, welche üblicherweise als zentral zur Messung von Instruktionssensitivität angesehen werden (Naumann et al., 2016). Diese Varianzquellen sind a) die Zeitpunkte der Messung (vor oder nach einer Lerngelegenheit / Unterricht), b) die Zugehörigkeit zu Lerngruppen innerhalb einer Stichprobe (z.B. Klassen), sowie c) die Kombination aus den beiden vorher genannten Varianzquellen (Zeitpunkte × Gruppen). Die den Perspektiven zugehörigen Varianzquellen sind also ein erstes Unterscheidungsmerkmal der verschiedenen Ansätze zur Messung der Instruktionssensitivität von Tests und Items.

Ansätze aus der Zeitpunkte-Perspektive betrachten Veränderung oder Variation von Itemparametern beziehungsweise Testwerten über die Zeit. Meist werden Testdaten zu zwei Messzeitpunkten herangezogen. Beispielsweise misst der Pretest-Posttest-Difference Index (PPDI; Cox und Vargas, 1966) die Instruktionssensitivität eines Items als die Differenz der Itemschwierigkeiten vor und nach dem Unterricht. Je größer diese Differenz ist, umso höher die Instruktionssensitivität des Items. Tabelle 1A zeigt dies anhand eines Beispielitems mit

drei Stufen (Naumann et al., 2016). Während das Item zum ersten Messzeitpunkt nur von vergleichsweise wenigen Kindern gelöst wird, erreichen zum zweiten Zeitpunkt mehr Kinder die mittlere beziehungsweise obere Antwortkategorie. Die Schwierigkeit der Aufgabe verringerte sich also über die Zeit, so dass das Item als sensitiv aus der Zeitpunkte-Perspektive angesehen werden kann.

[Tabelle 1 hier]

Ansätze aus der Gruppen-Perspektive betrachten Unterschiede oder Variation von Itemparametern beziehungsweise Testwerten über Lerngruppen innerhalb einer Stichprobe. Grundlegend ist die Annahme, dass maßgebliche Unterschiede im erlebten Unterricht auf die Zugehörigkeit der Schülerinnen und Schüler zu einer bestimmten Lerngruppe wie beispielsweise einer Klasse oder Schule zurückgehen (z.B. Robitzsch, 2009). Im Gegensatz zur Zeitpunkte-Perspektive handelt es sich bei Ansätzen aus der Gruppen-Perspektive um rein querschnittliche Ansätze, die auf Testdaten nach dem Unterricht basieren. Unterscheidet sich der Unterricht in den Gruppen, sollten einerseits instruktionssensitive Items für die Gruppen unterschiedlich schwierig sein und andererseits Lerngruppen mit höherwertigerem Unterricht höhere Testwerte in instruktionssensitiven Tests aufweisen. Tabelle 1B zeigt Instruktionssensitivität aus der Gruppen-Perspektive anhand der unterschiedlichen Antworthäufigkeiten eines Items in zwei verschiedenen Klassen zum selben Zeitpunkt. Je stärker also Itemparameter oder Testwerte über Lerngruppen hinweg variieren, umso höher wird die Instruktionssensitivität angenommen. Auf Testebene wird üblicherweise die durch Unterrichtsmaße erklärte Varianz der Testwerte als Evidenz für Instruktionssensitivität angesehen (z.B. Grossman et al., 2014).

Ansätze aus der Gruppen × Zeitpunkte-Perspektive berücksichtigen sowohl die Lerngruppen innerhalb einer Stichprobe als auch die Messzeitpunkte als Varianzquellen in der Analyse der Instruktionssensitivität. Diese Verknüpfung ermöglicht die Ableitung von

zwei Facetten der Sensitivität – globale und differentielle Sensitivität –, welche sowohl die zeitpunkt- als auch die gruppenbezogenen Sensitivitätsaspekte abbilden (Naumann et al., 2014). Die globale Sensitivität gibt an, inwiefern sich die Itemschwierigkeit oder die Testwerte über Lerngruppen innerhalb einer Stichprobe hinweg im Mittel über die Zeit verändern. Die differentielle Sensitivität gibt an, inwiefern diese Veränderung über Lerngruppen hinweg variiert. Mit Bezug zu Items empfehlen Naumann und Kollegen (2014) beide Facetten zur Analyse der Instruktionssensitivität heranzuziehen, da die Beurteilung anderenfalls möglicherweise unvollständig oder irreführend ist. Tabelle 1C zeigt die Antworthäufigkeiten eines global und differentiell sensitiven Items. Im Mittel über beide Klassen wird die Aufgabe über die Zeit leichter (globale Sensitivität), während die Schwierigkeitsveränderung über die Klassen variiert (differentielle Sensitivität). In der Messung der Instruktionssensitivität von Tests wird oft nur die differentielle Sensitivität im Sinne der durch Unterrichtsmerkmale erklärbaren Zwischen-Gruppen-Varianz betrachtet (z.B. Ing, 2008), während die globale Sensitivität üblicherweise wenig Beachtung findet.

Innerhalb des psychometrischen Rahmens mit den drei genannten Perspektiven (Zeitpunkte-, Gruppen- und Gruppen × Zeitpunkte-Perspektive) wird jeweils zwischen absoluten und relativen Maßen von Instruktionssensitivität unterschieden (Naumann et al., 2017). Absolute und relative Maße erlauben das Testen zweier unterschiedlicher Hypothesen bezüglich der Instruktionssensitivität eines Items, nämlich A) ob ein Item für sich betrachtet sensitiv ist oder B) ob die Sensitivität eines einzelnen Items von der Sensitivität des Gesamttests abweicht. Naumann und Kollegen (2017) definieren dabei Testsensitivität als die Varianz der Testwerte, je nach Perspektive über die Zeit, über die Lerngruppen oder über beides. Während also die Prüfung von Hypothese A unabhängig von der Testzusammensetzung für jedes einzelne Item möglich ist, ändern sich die Resultate der

Prüfung von Hypothese B für jedes Item in Abhängigkeit von der Zusammensetzung des Tests.

Absolute Sensitivitätsmaße beschreiben demnach die Gesamtsensitivität eines einzelnen Items unabhängig von der Sensitivität der übrigen Testitems. Das heißt, absolute Maße beschreiben das gesamte Ausmaß, in dem die Itemparameter über Zeitpunkte, über Lerngruppen oder beides variieren. Absolute Sensitivitätsmaße nehmen entsprechend den Wert Null an, wenn die Itemparameter konstant sind, sich also beispielsweise zwischen Zeitpunkten nicht verändern oder über Lerngruppen hinweg nicht unterscheiden. Sie nehmen Werte ungleich Null an, wenn die Itemparameter über die Zeit, über die Lerngruppen oder beides hinweg variieren. Dementsprechend zeigt Tabelle 1 Beispiele für absolute Sensitivität aus den drei Perspektiven. Das bekannteste absolute Sensitivitätsmaß ist der PPDI.

DIF-Methoden sind dagegen ein prominentes Beispiel für relative Sensitivitätsmaße. Relative Sensitivitätsmaße beschreiben die Abweichung der Sensitivität eines einzelnen Items von der Testsensitivität. Das Ausmaß der relativen Sensitivität eines Items hängt damit von der absoluten Sensitivität der anderen Testitems ab. Relative Sensitivitätsmaße nehmen entsprechend den Wert Null an, wenn die absolute Sensitivität eines Items der Testsensitivität entspricht. Beträgt beispielsweise die absolute Sensitivität eines Items 0.5 und ist identisch mit der Testsensitivität, dann ist die relative Sensitivität dieses Items $0.5 - 0.5 = 0$. Das heißt, das Item ist nicht relativ sensitiv, obwohl es absolut sensitiv ist. Die Itemsensitivität unterscheidet sich also nicht von der Testsensitivität. Weicht dagegen die absolute Sensitivität eines Items von der Testsensitivität ab, nehmen die relativen Sensitivitätsmaße für dieses Item einen Wert ungleich Null an. Dies ist beispielsweise dann der Fall, wenn sich ein Item in seiner Schwierigkeit über die Zeit weniger stark verändert als der Test (Zeitpunkte-Perspektive). Wie Naumann und Kollegen (2017) herausstellen, kann ein Item jedoch auch dann relativ sensitiv sein, wenn es absolut insensitiv ist. Für sich genommen geben relative

Maße also keine Auskunft darüber, ob ein Item dazu in der Lage ist, Effekte von Schule und Unterricht aufzufangen.

Zusammenfassend sind auf der Itemebene also insgesamt acht verschiedene Sensitivitätsmaße konzipierbar (letzte Spalte Abbildung 2). Allerdings fanden bisher nicht alle dieser Varianten eine praktische Anwendung in empirischen Studien. Aus der Zeitpunkte-Perspektive werden regelmäßig sowohl absolute Maße wie zum Beispiel der PPDI als auch relative Maße wie das Ausmaß des Item Parameter Drifts (z.B. DeMars et al., 2004; French et al., 2016) zur Messung der Instruktionssensitivität von Items herangezogen. Aus der Gruppen-Perspektive dominieren dagegen die relativen Sensitivitätsmaße auf Basis von DIF-Methoden (z.B. Deutscher und Winther, 2017; Li et al., 2016). Da absolute und relative Maße unterschiedliche Hypothesen prüfen, schlagen Naumann und Kollegen (2017) vor, in einem ersten Schritt der Itemselektion die absolute Sensitivität zu prüfen und darauf aufbauend in einem zweiten Schritt der relativen Sensitivität der ausgewählten Items nachzugehen. Soll ein Testinstrument Rückschlüsse über Unterricht erlauben, sollte die absolute Sensitivität idealerweise möglichst hoch und die relative Sensitivität möglichst niedrig sein. Der Vorteil dieser Vorgehensweise besteht darin, zunächst Informationen über das Ausmaß zu erhalten, in dem jedes für einen Test infrage kommende Item Effekte des Unterrichts erfassen kann (absolute Sensitivität), und daran anknüpfend Erkenntnisse über die Konsequenzen der Testzusammenstellung zu erlangen. Konsequenzen der Testzusammenstellung können beispielsweise Verletzungen von Messinvarianzannahmen oder der Eindimensionalität sein, wenn einzelne oder mehrere Items von der Testsensitivität abweichen (relative Sensitivität).

Der Rahmen als Item-Response-Modell

Das psychometrische Framework lässt sich unmittelbar in ein längsschnittliches Mehrebenen-IRT-Modell übersetzen (LMLIRT Modell; Naumann et al., 2017). Im LMLIRT Modell ergibt sich die Wahrscheinlichkeit einer korrekten Antwort aus der mittleren Fähigkeit

von Gruppe c, θ_{tc}, der individuellen Fähigkeit von Person i, θ_{tci}, sowie der gruppen- und zeitpunktspezifischen Itemschwierigkeit β_{tck}. Für jeden Zeitpunkt $t > 1$ sind θ_{tc} und θ_{tci} die Veränderung der gruppenspezifischen und individuellen Fähigkeiten vom vorhergehenden Zeitpunkt $t-1$. Analog beschreibt β_{tck} den Ausgangswert für die gruppenspezifische Schwierigkeit des Items k zu Zeitpunkt $t = 1$ und die gruppenspezifische Veränderung der Schwierigkeit zu jedem Zeitpunkt $t > 1$. Abbildung 3 zeigt das LMLIRT Modell beispielhaft für fünf Items und zwei Messzeitpunkte.

[Abbildung 3 hier]

Die Itemparameter β_{tck} werden als multivariat normalverteilt angenommen mit Mittelwerten β_{tk} und itemspezifischen Kovarianzmatrizen $\mathbf{\Phi}_k$. Die Verteilung der gruppenspezifischen Itemparameter β_{tck} beinhaltet die Information zur globalen und differentiellen Sensitivität der Items. Während die in β_{tk} enthaltenen Mittelwerte die globale Sensitivität anzeigen, beschreiben die Diagonalelemente von $\mathbf{\Phi_k}$, also die Varianzparameter ϕ^2_{tk}, die differentielle Sensitivität eines Items. Ist β_{tk} für ein Item ungleich null für einen Zeitpunkt $t > 1$, also die mittlere Veränderung der Itemschwierigkeit über Gruppen entweder positiv oder negativ, dann wird das Item als global sensitiv für diese Zeitspanne angesehen. Das heißt, anhand dieses Items wird ein Lerneffekt in der Stichprobe zwischen den Zeitpunkten sichtbar. In gleicher Weise wird ein Item als differentiell sensitiv angesehen, umso höher ϕ^2_{tk} für $t > 1$ ist, also je stärker die Schwierigkeitsveränderung eines Items über Lerngruppen hinweg variiert. Das heißt, anhand dieses Items werden Unterschiede im Lernen zwischen Gruppen sichtbar. Je nach Vorgehen bei der Modellidentifikation erhält man absolute und relative Maße als statistische Indikatoren für die globale und differentielle Sensitivität (siehe Naumann et al., 2017).

Zusätzlich erlaubt das LMLIRT Modell, diese auf Itemantworten der Schülerinnen und Schüler basierenden statistischen Indikatoren für Instruktionssensitivität mit

Unterrichtsmerkmalen und/oder mit Itemmerkmalen in Beziehung zu setzen. So kann die Verteilung der klassen- und zeitpunktspezifischen Itemschwierigkeitsparameter in Gleichung 4 um Unterrichtsmerkmale als Prädiktoren ergänzt werden, um die differentielle Sensitivität zu erklären (z.B. Naumann et al., 2015). Ebenso lassen sich anstelle der itemspezifischen Parameter β_{tck} aufgabenmerkmalspezifische Koeffizienten im Sinne eines Linear Logistischen Testmodells (LLTM; Fischer, 1972) verwenden, um die globale Sensitivität von Aufgaben zu erklären (z.B. Hochweber et al., 2017).

Weiterer Forschungsbedarf

Mit dem vorgestellten psychometrischen Rahmen ist eine systematische Betrachtung der Instruktionssensitivität von Tests und Items möglich. Auf dem Weg zu Testinstrumenten, die zuverlässig Unterrichtseffekte erfassen können, lässt sich weiterer Forschungsbedarf in drei Bereichen identifizieren, nämlich bei (1) der Validierung von Aussagen über Instruktionssensitivität, (2) der Konstruktion instruktionssensitiver Items, sowie (3) der Bedeutung von Instruktionssensitivität für die Nutzung und Interpretation der Testwerte.

Validierung von Aussagen über Instruktionssensitivität

Die Gültigkeit von Expertenaussagen über Instruktionssensitivität ist bisher wenig untersucht. Expertenaussagen erscheinen in der Untersuchung von Instruktionssensitivität vorteilhaft, da sie im Vergleich zu auf Itemantworten basierenden Verfahren einen wesentlich geringeren Aufwand erfordern. Sie können ähnlich wie Untersuchungen zur curricularen Validität zu jedem beliebigen Zeitpunkt durchgeführt werden. Dagegen muss die Messung der Instruktionssensitivität eines Tests oder einzelner Items mittels Itemantworten von der Wirksamkeitsprüfung des Unterrichts getrennt erfolgen. Die Wirksamkeitsprüfung des Unterrichts und die Messung von Instruktionssensitivität verhalten sich in diesem Falle komplementär zueinander. Die Messung von Instruktionssensitivität erfolgt unter der Annahme, dass Unterricht effektiv ist, um die Sensitivität eines Tests oder Items zu

bestimmen, während die Wirksamkeitsprüfung des Unterrichts die Annahme erfordert, dass ein Instrument sensitiv ist, um den Effekt des Unterrichts zu bestimmen (Naumann et al., 2016). Anhand derselben Testdaten sind Effektivität und Sensitivität also nicht trennbar. Bei Expertenaussagen gibt es diese Konfundierung nicht. Jedoch stellt sich die Frage, inwiefern Expertinnen und Experten dazu in der Lage sind, valide interpretierbare Aussagen zur Instruktionssensitivität von Tests und Items zu treffen (Polikoff, 2010). Aktuell ist die Befundlage uneindeutig. So fand Chen (2012) zwar moderate Korrelationen ($r > .30$) zwischen einem globalen Expertenurteil zur Instruktionssensitivität von Items und deren gemessener relativen Sensitivität aus der Gruppen-Perspektive, sie folgerte aus diesem Befund jedoch, dass Expertenurteile nicht mit Itemstatistiken übereinstimmen. Für das Verhältnis zu anderen empirischen Sensitivitätsmaßen oder zu differenzierteren Expertenurteilen liegen zum jetzigen Zeitpunkt keine systematischen Untersuchungen vor.

Allerdings ist auch der Zusammenhang zwischen den meisten Itemstatistiken und den implementierten Unterrichtsinhalten sowie der Qualität der Implementation nach wie vor kaum empirisch untersucht. Bereits 1981 kritisierte van der Linden, dass Itemstatistiken nicht per se für Effekte des Unterrichts stehen. Einerseits ist die Variabilität in den Itemparametern eine notwendige Voraussetzung, andererseits ist sie kein hinreichender Beleg für die Instruktionssensitivität eines Items (Naumann et al., 2016). Die Variabilität kann ebenso von Merkmalen beeinflusst sein, die nicht unmittelbar mit dem Inhalt und der Qualität des Unterrichts im Zusammenhang stehen wie beispielsweise dem sozioökonomischen Status der Schülerinnen und Schüler. Tatsächlich berücksichtigen nur wenige Studien Merkmale des Unterrichts in ihren Itemanalysen (z.B. Muthén et al., 1991). Für Itemstatistiken bleibt entsprechend bislang häufig offen, welchen Anteil der Unterricht tatsächlich an der Variabilität in den Itemparametern hat. Der Einbezug von Unterrichtsmaßen und

Kontrollvariablen in die Analyse der Itemsensitivität könnte daher eine validere Nutzung und Interpretation der Itemstatistiken erlauben.

Selbst wenn Unterrichtsmaße einbezogen werden, könnte die Validität von Aussagen zur Instruktionssensitivität weiter gestärkt werden. Studien zur Instruktionssensitivität von Tests berücksichtigen beispielsweise regelmäßig empirische Maße zu behandelten Unterrichtsinhalten oder zur Qualität der Implementation (Polikoff und Porter, 2014; Ruiz-Primo et al., 2012). Die Auswahl der Unterrichtsmaße orientiert sich jedoch stark am Unterrichtsangebot. Bereits in der Literatur zur Instruktionssensitivität von Testitems lassen sich vereinzelt Hinweise finden, die auf eine Bedeutsamkeit des Einbezugs von Merkmalen der Schülerinnen und Schüler und der Klassenkomposition hindeuten (Muthén et al., 1991; Naumann et al., 2016). In Anbetracht theoretischer Modelle zur Erklärung des Zustandekommens von Schulleistungen wie den Angebots-Nutzungs-Modellen (Brühwiler, 2014; Fend, 2002; Helmke, 2012) erscheint der Nicht-Einbezug von Schülermerkmalen und Klassenkomposition zu stark vereinfachend. Entsprechend könnte der systematische Einbezug von Merkmalen der Schülerinnen und Schüler sowie der Klassenzusammensetzung eine validere Interpretation von Instruktionssensitivitätsindikatoren unterstützen.

Schließlich fehlt es an einem Bezugsmaßstab zur Beurteilung des Ausmaßes der Instruktionssensitivität von Test und Items. Zwar gibt es Wege zur Prüfung der statistischen Bedeutsamkeit der Sensitivität, inwiefern in der Praxis von einer geringen oder hohen Sensitivität zu sprechen ist, bleibt derzeit jedoch offen.

Konstruktion instruktionssensitiver Items

Während Studien regelmäßig die Instruktionssensitivität bestehender Testverfahren untersuchen, gibt es bislang nur sehr wenig Wissen über deren zielgerichtete Konstruktion. Für die zielgerichtete Konstruktion instruktionssensitiver Items konnten Ruiz-Primo und Kollegen (2012) in einem experimentellen Design zeigen, dass die globale

Instruktionssensitivität eines Items mit dessen Nähe zu den Inhalten und den Aktivitäten des implementierten Curriculums ansteigt. Nähe zum implementierten Curriculum bedeutet beispielsweise, wie stark sich die Aufgabenstellung eines Items mit Fragestellungen im Unterricht überschneidet, in welchem Ausmaß Schülerinnen und Schüler in den itemrelevanten Inhalten unterrichtet wurden oder inwiefern im Unterricht vermittelte Strategien hilfreich zur korrekten Lösung des Items sind (Ruiz-Primo et al., 2002). Items nahe am implementierten Curriculum sollten also prinzipiell die zentralen Konzepte, Strategien und Erklärungsmodelle innerhalb einer Unterrichtseinheit, also des intendierten Curriculums, erfassen. Detailliertes Wissen über das intendierte Curriculum ist daher Voraussetzung für die Konstruktion instruktionssensitiver Items (Ruiz-Primo et al., 2012).

In der Praxis stellt nicht jeder Unterricht eine perfekte Realisation des intendierten Curriculums dar. Außerhalb eines experimentellen Designs, das implementiertes sowie intendiertes Curriculum stark aufeinander abstimmt, bietet sich demnach keine bestimmte Art von Unterricht als Referenzpunkt für die Itemkonstruktion an. Tatsächlich kann es eine Bandbreite möglicher Implementationen des Curriculums geben, die alle gleichermaßen dem intendierten Curriculum entsprechen. Bei Anwendungen wie beispielsweise der Testkonstruktion in Large Scale Assessments oder Unterrichtsstudien bleibt nach wie vor die Frage nach der zielgerichteten Konstruktion instruktionssensitiver Items offen, sofern die Testwerte hinsichtlich des aufgrund des Unterrichts erreichten Lernfortschritts interpretiert oder mit Merkmalen von Schule und Unterricht in Beziehung gesetzt werden sollen.

Bedeutung für die Nutzung und Interpretation der Testwerte

Das Verhältnis von Test- und Itemebene spielt eine zentrale Rolle in der Nutzung und Interpretation von Testwerten. Ungeklärt ist die Frage, wie die Selektion bestimmter Items aufgrund ihrer Instruktionssensitivität die Interpretation der Testwerte des daraus konstruierten Tests beeinflusst. Einerseits tangiert das die bereits zuvor beschriebene Frage,

in welchem Maße ein Test instruktionssensitiv ist, der aus instruktionssensitiven Items besteht, und andererseits resultiert daraus die Frage, welche Konsequenzen für Rückschlüsse über Unterricht anhand der Testwerte sich aufgrund eines bestimmten Grades an Instruktionssensitivität der Items ergeben. Letztere Frage untersuchte van der Linden (1981) am Beispiel der Itemselektion rein anhand des PPDI und dessen Einflusses auf die Testwerte. Er entdeckte einerseits ein Ansteigen von Effektstärken, wenn Tests anhand eines möglichst hohen PPDI-Werts zusammengestellt werden. Andererseits sind diese Effektstärken nicht sinnvoll interpretierbar, da sie künstlich erzeugt und inhaltlich bedeutungslos scheinen. Im Hinblick auf die Zwischen-Gruppen-Varianzkomponente gibt es dazu bislang jedoch keine empirischen Befunde. van der Linden (1981) folgend scheint die reine Maximierung der absoluten Sensitivität nur bedingt erstrebenswert. Stattdessen bedarf es weiterer Studien zum Verhältnis von Konstrukt, Iteminhalt und Itemsensitivität.

Auch ist bisher wenig Wissen über den Einfluss einer heterogenen Instruktionssensitivität von Items auf Veränderungsmessungen in schulischen Kontexten verfügbar. In Veränderungsmessungen führt eine über Items hinweg variierende globale Sensitivität vermutlich zu Verletzungen von Messinvarianzannahmen über die Zeit (vgl. Naumann et al., 2014). Das heißt, die Rangfolge der Itemschwierigkeiten könnte sich vor und nach einem Unterricht unterscheiden und damit ließe sich keine gemeinsame Skala über die Messzeitpunkte hinweg erstellen. Theoretisch wäre eine mögliche Ursache in den Anforderungen und Inhalten des Unterrichts zu finden. Bestimmte Unterrichtsinhalte sind oftmals leichter zu erlernen als andere. Items zeigen also möglicherweise allein schon aufgrund ihres Inhalts einen variierenden Grad an Instruktionssensitivität. Ein variierender Grad an globaler Instruktionssensitivität stellt daher theoretisch ein immanentes Problem für Veränderungsmessungen in schulischen Kontexten dar. Inwiefern Veränderungsmessungen in

schulischen Kontexten in der Praxis mit diesem Problem konfrontiert sind, ist mangels empirischer Studien bislang weitgehend offen.

In ähnlicher Weise ist ein Einfluss eines variierenden Grades an differentieller Instruktionssensitivität von Items auf die Testwertinterpretation denkbar. Gibt es innerhalb eines Tests zwei oder mehr Sets von Items, die in unterschiedlicher Weise differentiell sensitiv sind, hat dies möglicherweise Implikationen für die Dimensionalität von Veränderungswerten auf der Gruppenebene. Kovariieren die klassenspezifischen Itemschwierigkeitsveränderungen für bestimmte Sets von Items untereinander stärker über die Zeit als für andere Itemsets, könnten die Veränderungswerte auf der Gruppenebene in der Folge mehrdimensional erscheinen. Im Gegensatz zu Fragebogenskalen wird die Mehrdimensionalität von Test- und Veränderungswerten auf der Gruppenebene in Item-Response-Modellen nur selten empirisch überprüft. In der Praxis könnte eine Mehrdimensionalität vorteilhaft sein, um beispielsweise eine detailliertere Rückmeldung über Lernfortschritte auf der Gruppenebene zu geben als es mit einem einzelnen Wert möglich wäre. Entsprechende Studien zur gültigen Interpretation und Nutzbarmachung des Konzepts der Instruktionssensitivität für die wissenschaftliche Praxis wären daher wünschenswert.

Nichtsdestotrotz lassen sich bereits auf Basis des aktuellen Kenntnisstandes zwei potentielle Wege benennen, um Instruktionssensitivität im Forschungsprozess sicherzustellen: (a) durch den Rückgriff auf bewährte Testinstrumente oder (b) mittels Pilotierungsstudien zur Untersuchung der Instruktionssensitivität neu erstellter Testinstrumente. Für bewährte Testinstrumente, die sich bereits in früheren Studien als instruktionssensitiv gezeigt haben, lässt sich auch für nachfolgende Studien vermuten, dass sie Effekte von Schule und Unterricht auffangen können. Gleichermaßen scheint diese Annahme ebenso für neu erstellte Testinstrumente plausibel, wenn diese sich in Pilotierungsstudien als instruktionssensitiv erweisen. Beide Vorgehensweisen setzen voraus, dass die vorherigen (Pilotierungs-)Studien

mit der geplanten neuen Studie weitgehend vergleichbar sind. Vergleichbar bedeutet, dass (a) sich die Evidenz für Instruktionssensitivität auf ähnliche Varianzquellen bezieht wie die für die neue Hauptstudie geplanten Analysen (vgl. Naumann et al., 2016) und (b) eine ähnliche Schülerstichprobe sowie (c) ein ähnlicher Unterricht zugrunde liegt. Ein Test, der sich beispielsweise in Anwendung in einer Jahrgangsstufe als instruktionssensitiv gezeigt hat, muss nicht notwendigerweise auch in einer anderen Jahrgangsstufe sensitiv sein, da sich die Fähigkeiten der Schülerinnen und Schüler sowie der Unterricht zwischen den Stufen mitunter stark unterscheiden. Insbesondere an Pilotierungen stellt dies enorm hohe Ansprüche, da sie im Idealfall (a) über ein zur Hauptstudie kompatibles Erhebungsdesign verfügen müssen mit (b) Stichproben von Schülerinnen und Schülern, die vergleichbare Lerngelegenheiten wie die Stichprobe der Hauptuntersuchung hatten, sowie (c) einer Bandbreite an Unterricht, die ähnlich der in der Hauptstudie zu erwartenden Variation ist (vgl. Aichele et al., eingereicht). Offen bleibt daher, ob sich solch eine Art Pilotierung ökonomisch sinnvoll umsetzen lässt. Letzten Endes gilt es in beiden Vorgehensweisen zu beachten, dass ein Test, der für eine spezifische Facette von Unterrichtsqualität oder für einen spezifischen Unterrichtsinhalt sensitiv ist, nicht notwendigerweise auch sensitiv für andere Facetten von Unterrichtsqualität oder Unterrichtsinhalte sein muss (vgl. Polikoff, 2016).

Schlussendlich muss Instruktionssensitivität immer dann sichergestellt sein, wenn Testergebnisse in Beschreibungs- oder Erklärungsmodellen als ein Erfolgskriterium für Schule und Unterricht herangezogen werden sollen (vgl. Klieme & Leutner, 2006). Ihre Relevanz hängt also nicht damit zusammen, *was* (z.B. Wissen oder Kompetenzen), *auf welche Art* (z.B. Small-Scale oder Large-Scale) oder *in welchem Kontext* (Bildungsmonitoring oder -forschung) getestet wird, sondern *zu welchem Zweck*. Steht die Identifikation allgemeiner Fähigkeiten von Schülerinnen und Schülern oder ihres Lernstands im Vordergrund, ist Instruktionssensitivität nachrangig. Steht der Lernerfolg oder der

Lernfortschritt aufgrund des Unterrichts im Vordergrund, ist Instruktionssensitivität zentral. Das heißt, um gültige Rückschlüsse zu ziehen ist es erforderlich, dass die eingesetzten Instrumente potentielle Effekte von Schule und Unterricht erfassen können. Entsprechend müssen Unterschiede und Veränderungen in den Inhalten und der Qualität des Unterrichts mit Veränderungen im Antwortverhalten und den Testwerten einhergehen (vgl. Burstein, 1989). Nur wenn dieser Zusammenhang vorab geklärt ist, kann die gemessene Leistung als ein gültiges Kriterium für den Erfolg oder Misserfolg eines Unterrichts dienen. Instruktionssensitivität von Tests oder Testitems ist also insbesondere dann von zentraler Bedeutung, wenn auf Basis der Testwerte der Schülerinnen und Schüler a) Schul- und Unterrichtsentwicklung oder b) Schul- und Unterrichtseffektivitätsforschung betrieben wird.

Zu beachten ist, dass Instruktionssensitivität nicht bedeutet, die direkte Beobachtung von Unterricht durch die Messung von Leistung ersetzen zu können. Leistungstestwerte als Produkt eines Lernprozesses erlauben keinen Einblick darin, wie die gemessene Leistung zustande gekommen ist, also auf welche Weise der Unterricht stattgefunden hat (vgl. Helmke, 2012). Dennoch ist es essentiell, Argumente für ihre valide Interpretation und Nutzung zu liefern (vgl. AERA et al., 2014). Für gültige Rückschlüsse über Schule und Unterricht heißt dies, die Instruktionssensitivität der Instrumente zu beachten.

Referenzen

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing.* Washington, DC.

Aichele, C., Naumann, A., Michaelis, C., & Hartig, J. (eingereicht). *Der Einfluss von Stichprobeneigenschaften auf die Itemselektion*. Manuskript eingereicht zur Publikation.

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

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

Anderson, L.W. (2002). Curricular alignment: a re-examination. *Theory Into Practice, 41*(4), 255–260.

Arnold, K.-H. (2005). Mehr Fairness im Bildungssystem. Fragen zu Standards und Vergleichsarbeiten. *Friedrich-Jahresheft, 23*, 25–27.

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

Baumert, J., Brunner, M., Lüdtke, O. & Trautwein, U. (2007). Was messen internationale Schulleistungsstudien? – Resultate kumulativer Wissenserwerbsprozesse. Eine Antwort auf Heiner Rindermann. *Psychologische Rundschau, 58*, 118–127.

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

Burstein, L. (1989). *Conceptual considerations in instructionally sensitive assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Chen, J. (2012). *Impact of instructional sensitivity on high-stakes achievement test items: A comparison of methods*. Lawrence, KS: University of Kansas.

Clauser, B. E., Nungester, R. J. & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33*(4), 453–464.

Cox, R. C., & Vargas, J. S. (1966). *A comparison of item-selection techniques for norm referenced and criterion referenced tests*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

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

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.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17*, 265–300.

Deutscher, V. & Winther, E. (2017). Instructional Sensitivity in Vocational Education. *Learning and Instruction*. Advance online publication. doi: 10.1016/j.learninstruc.2017.07.004

Drechsel, B., Prenzel, M., & Seidel, T. (2015). Nationale und internationale Schulleistungsstudien. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 353-380). Berlin, Heidelberg: Springer.

Fend, H. (2002). Mikro- und Makrofaktoren eines Angebot-Nutzungsmodells von Schulleistungen. Zum Stellenwert der Pädagogischen Psychologie bei der Erklärung von Schulleistungsunterschieden verschiedener Länder. *Zeitschrift für Pädagogische Psychologie, 16*(3/4), 141–149.

Fend, H. (2011). Die Wirksamkeit der Neuen Steuerung – theoretische und methodische Probleme ihrer Evaluation. *Zeitschrift für Bildungsforschung, 1*, 5-24.

Fischer, G. H. (1972). *Conditional maximum-likelihood estimations of item parameters for a linear logistic test model* (Research Bulletin 9). Vienna: University of Vienna, Psychological Institute.

Fischer, N., Sauerwein, M. N., Theis, D., & Wolgast, A. (2016). Vom Lesenlernen in der Ganztagsschule: Leisten Ganztagsangebote einen Beitrag zur Leseförderung am Beginn der Sekundarstufe I? *Zeitschrift Für Pädagogik, 62*(6), 780–796.

French, B. F., Finch, W. F., Randel, B., Hand, B., & Gotch, C. M. (2016). Measurement invariance techniques to enhance measurement sensitivity. *International Journal of Quantitative Research in Education, 3*, 79–93.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29*(1), 38–44.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Greer, E. A. (1995). *Examining the validity of a new large-scale reading assessment instrument from two perspectives*. Urbana.

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*, 293–303.

Grünkorn, J., Klieme, E., & Stanat, P. (im Druck). Bildungsmonitoring und Qualitätssicherung. In O. Köller, M. Hasselhorn, F. W. Hesse, K. Maaz, J. Schrader, C. K. Spieß, H. Solga, & K. Zimmer (Hrsg.), *Das Bildungswesen in Deutschland: Bestand und Potentiale.* Bad Heilbrunn: UTB/Klinkhardt.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Roid, G. H. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement, 18*, 39–53.

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion (2. Auflage, S. 143–171). Berlin, Heidelberg: Springer.

Hartig, J., Klieme, E. & Leutner, D. (Hrsg.) (2008). *Assessment of competencies in educational contexts*. Hogrefe & Huber Publishers.

Hascher, T. & Schmitz, B. (Hrsg.) (2010). *Pädagogische Interventionsforschung: Theoretische Grundlagen und empirisches Handlungswissen.* Weinheim: Juventa.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (4. überarb. Aufl.). Seelze: Klett-Kallmeyer.

Hochweber, J., Naumann, A., Hartig, J., Kleinbub, I. & Musow, S. (2017). *Using item properties to predict the instructional sensitivity of test items*. Paper presented at the 17th Biennial Conference of the European Association for Research on Learning and Instruction (EARLI), Tampere.

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

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Klieme, E. (2008). Systemmonitoring für den Sprachunterricht. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* (S. 1–10). Weinheim und Basel: Beltz.

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

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

Kosecoff, J. B. & Klein, S. P. (1974). *Instructional sensitivity statistics appropriate for objectives-based test items.* Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago.

Kultusministerkonferenz (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring.* München: Wolters Kluwer.

Li, H., Qin, Q., Lei, P.-W. (2016). An examination of the instructional sensitivity of the TIMSS math items: A hierarchical differential item functioning approach. *Educational Assessment, 22*(1), 1–17.

Li, M., Ruiz-Primo, M. A., & Wills, K. (2012). *Comparing methods to evaluate the instructional sensitivity of items.* Paper presented at the Annual Conference of the National Council on Measurement in Education, Vancouver.

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.

Lossen, K., Tillmann, K., Holtappels, H. G., Rollett, W., & Hannemann, J. (2016). Entwicklung der naturwissenschaftlichen Kompetenzen und des sachunterrichtsbezogenen Selbstkonzepts bei Schüler/innen in Ganztagsgrundschulen: Ergebnisse der Längsschnittstudie StEG-P zu Effekten der Schülerteilnahme und der Angebotsqualität. *Zeitschrift für Pädagogik, 62*(6), 760–779.

Maag Merki, K. (2010). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 145-169). Wiesbaden: Springer.

McClung, M. S. (1979). Competency testing programs: Legal and educational issues. *Fordham Law Review, 47*, 652–711.

Mehrens, W. A. & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357–370.

Millman, J. (1970). Reporting student progress: A case for a criterion-referenced marking system. *Phi Delta Kappan, 54*(4), 226–230.

Musow, S., Naumann, A., Hartig, J., & Hochweber, J. (2018). *Expertenratings – Ein Verfahrensvergleich zur Evaluation der Instruktionssensitivität von Testitems*. Vortrag bei der 6. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Basel.

Muthén, B. O. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika, 54*, 385–396.

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*, 1–22.

Nagy, G., Retelsdorf, J., Goldhammer, F., Schiepe-Tiska, A., & Lüdtke, O. (2017). Veränderungen der Lesekompetenz von der 9. zur 10. Klasse: Differenzielle Entwicklungen in Abhängigkeit der Schulform, des Geschlechts und des soziodemografischen Hintergrunds? *Zeitschrift für Erziehungswissenschaft, 2*(20), 177–203.

Naumann, A., Hartig, J., & Hochweber, J. (2017). Absolute and Relative Measures of Instructional Sensitivity. *Journal of Educational and Behavioral Statistics, 42*(6), 678–705.

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.

Naumann, A., Hochweber, J., & Hartig, J. (2015). *An Explanatory Longitudinal Multilevel IRT Approach to Instructional Sensitivity.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Chicago.

Naumann, A., Hochweber, J., & Klieme, E. (2016). A Psychometric Framework for the Evaluation of Instructional Sensitivity. *Educational Assessment, 21*(2), 1–13.

Pellegrino, J. W. (2002). Knowing what students know. *Issues in Science & Technology, 19*(2), 48–52.

Polikoff, M. S. (2010). Instructional Sensitivity as a Psychometric Property of Assessments. *Educational Measurement: Issues and Practice, 29*(4), 3–14.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*(4), 399-416.

Popham, W. J. (2007). Instructional Insensitivity of Tests: Accountability's Dire Drawback. *Phi Delta Kappan, 89*(2), 146–155.

Popham, J.W. & Ryan, J.M. (2012). *Determining a high-stakes test's instructional sensitivity*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Vancouver, BC.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3–14.

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* (S. 119–144). Wiesbaden: Springer.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? *Psychologische Rundschau, 57*(2), 69–86.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In D. Granzer, O. Köller, & A. Bremerich-Vos (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 42–106). Weinheim, Basel: Beltz.

Ruiz-Primo, M. A., Li, M., 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.

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.

Spoden, C. & Leutner, D. (2011). *Vergleichsarbeiten*. URN: urn:nbn:de:0111-pedocs-107492.

Stanat, P. & Pant, H.-A. (2016). Die IQB-Bildungstrends als zentrales Element des Bildungsmonitorings in Deutschland. In P. Stanat, K. Böhme, S. Schipolowski & N. Haag (Hrsg.), *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich.* (S. 13-19). Münster: Waxmann.

United States Court of Appeals, fifth Circuit. (1981). DEBRA P. v. Ralph D. TURLINGTON (Nr. No. 79-3074).

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.

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

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn, and equity: New standards examinations for the California Mathematics Renaissance*. Los Angeles, CA: Center for the Study of Evaluation.

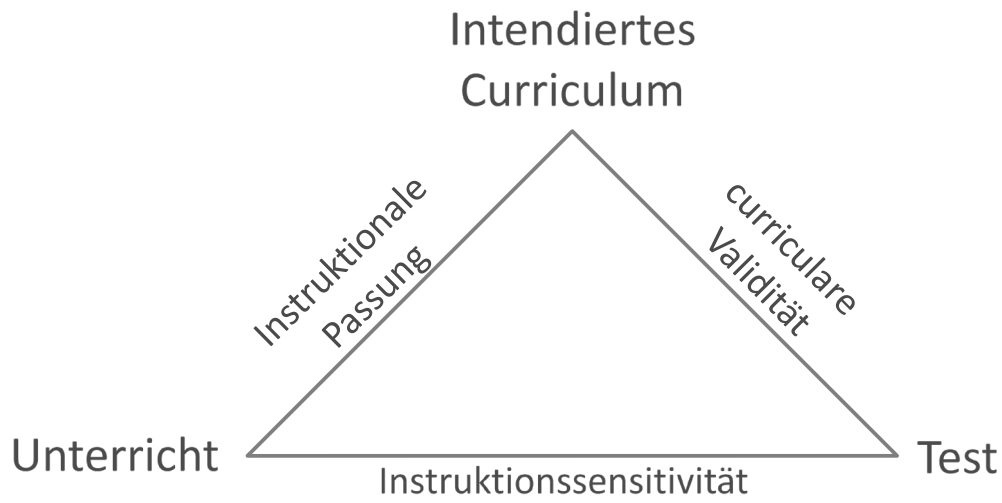

Abbildung 1. Verhältnis von intendiertem Curriculum, Unterricht (implementiertes Curriculum) und Test adaptiert nach Anderson (2002) und Pellegrino (2001).

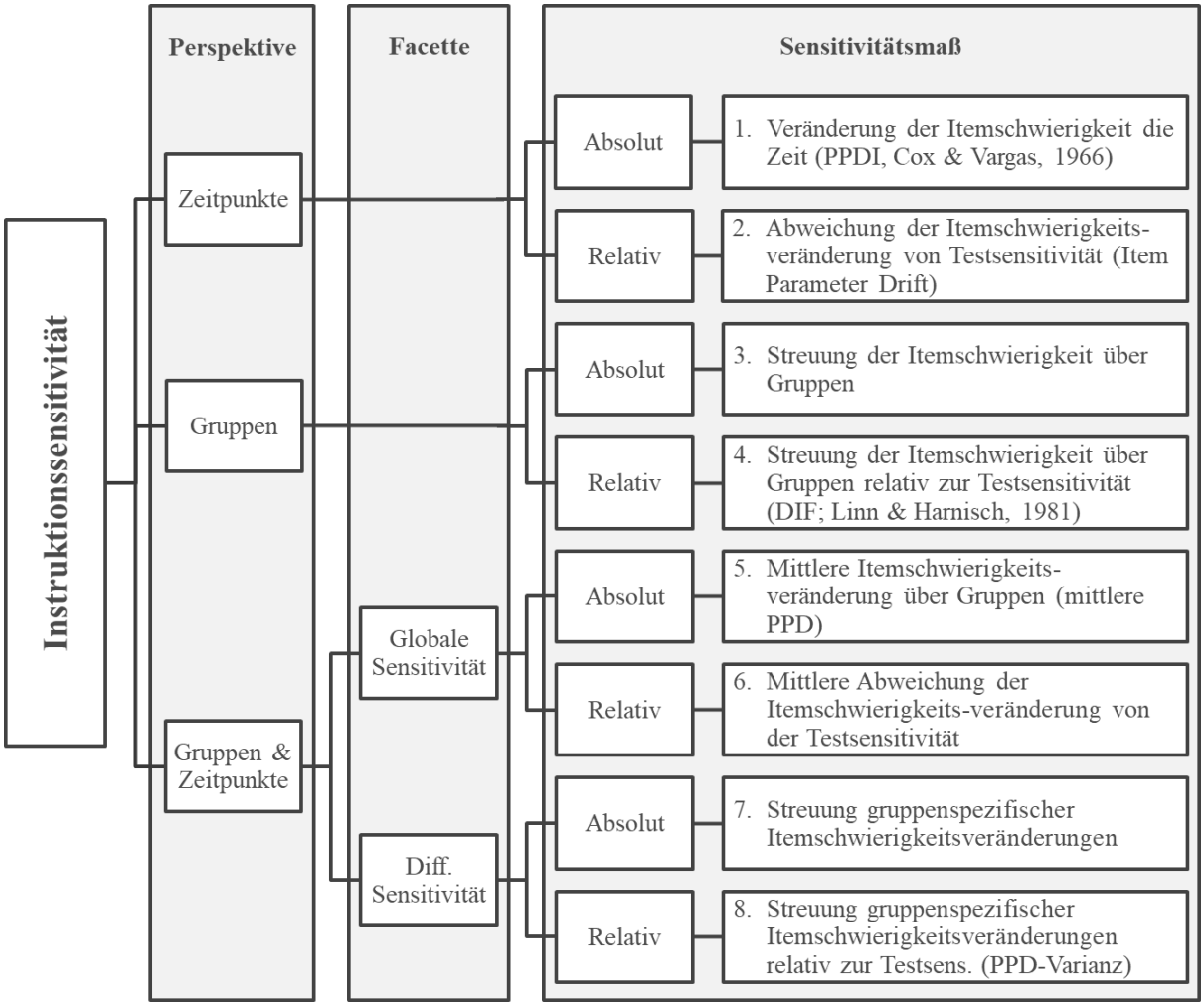

Abbildung 2. Psychometrischer Rahmen zur Messung von Instruktionssensitivität nach Naumann und Kollegen (2017).

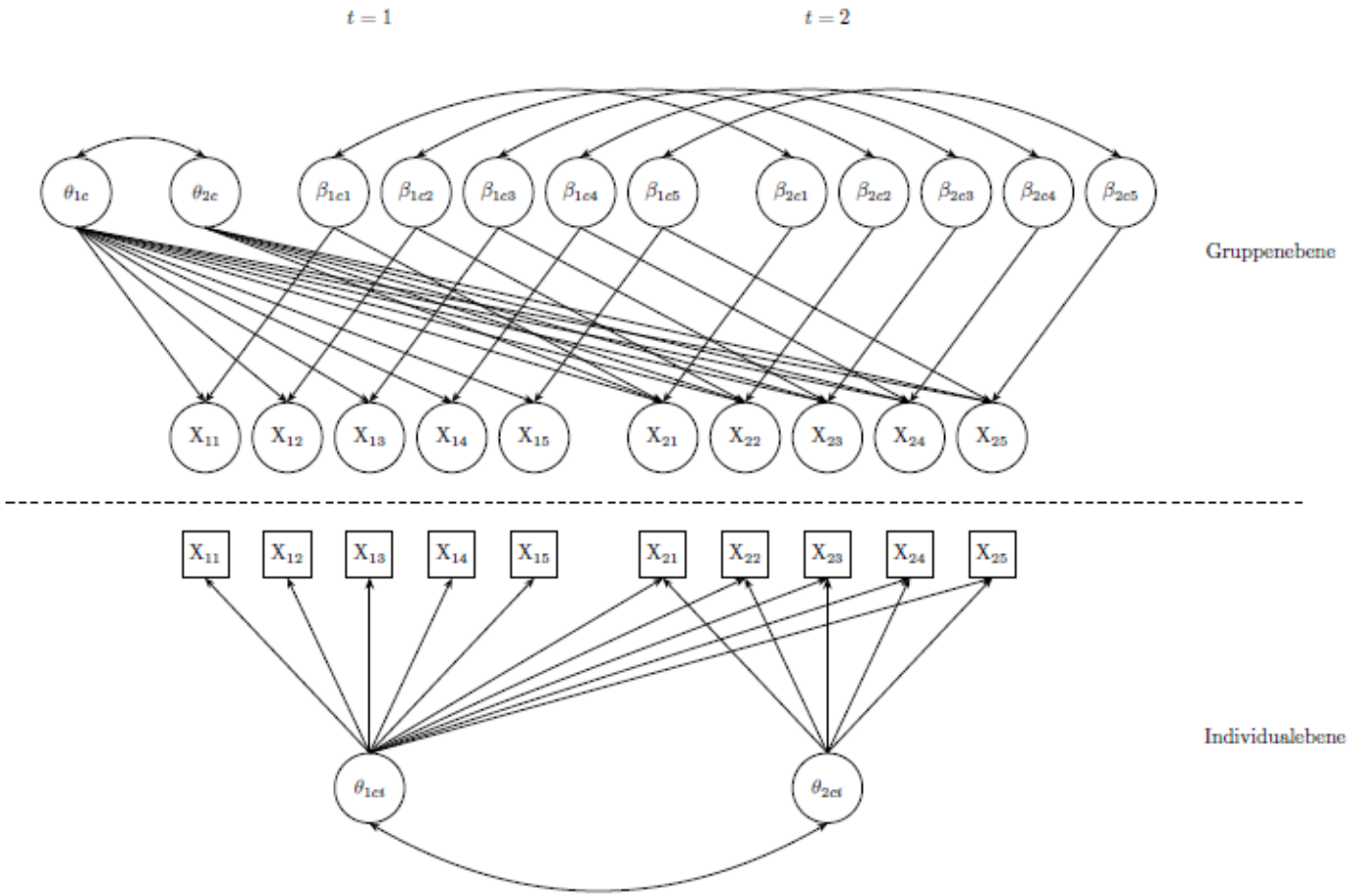

Abbildung 3. LMLIRT Modell für fünf Items zu zwei Messzeitpunkten. Die Gruppenebene ist oberhalb der Linie dargestellt, die Individualebene unterhalb. Parameter $\boldsymbol{\theta_{tc}}$ und $\boldsymbol{\theta_{tci}}$ beschreiben die Gruppen- und individuellen Fähigkeitskomponenten. Parameter $\boldsymbol{\beta_{tck}}$ ist der gruppen- und zeitpunktspezifische Schwierigkeitsparameter von Item *k*. Mittelwert und Varianz von $\boldsymbol{\beta_{tck}}$ dienen als Indikatoren für die globale und differentielle Sensitivität eines Items.

Tabelle 1

Relative Antworthäufigkeiten der Kategorien eines Items für eine Gesamtstichprobe zu zwei Zeitpunkten (A), für zwei Klassen A und B innerhalb der Stichprobe zu T2 (B) und für dieselben Klassen A und B zu zwei Zeitpunkten

A) Relative Antworthäufigkeiten Gesamtstichprobe

Antwortkategorie	T1 n=986	T2 n=991
Falsche Antwort (0)	86.1 %	55.2 %
Teilweise gelöst (1)	9.7 %	27 %
Richtige Antwort (2)	4.2 %	17.8 %

B) Relative Antworthäufigkeiten T2 für zwei Klassen

	Klasse A (n = 22)	Klasse B (n = 20)
Falsche Antwort (0)	13.6 %	15 %
Teilweise gelöst (1)	40.9 %	65 %
Richtige Antwort (2)	45.5 %	20 %

C) Veränderung der relativen Antworthäufigkeiten von T1 zu T2 für beide Klassen

	Klasse A (n = 22)	Klasse B (n = 20)
Falsche Antwort (0)	−81.9 %	−58.7 %
Teilweise gelöst (1)	+36.4 %	+43.9 %
Richtige Antwort (2)	+45.5 %	+14.7 %