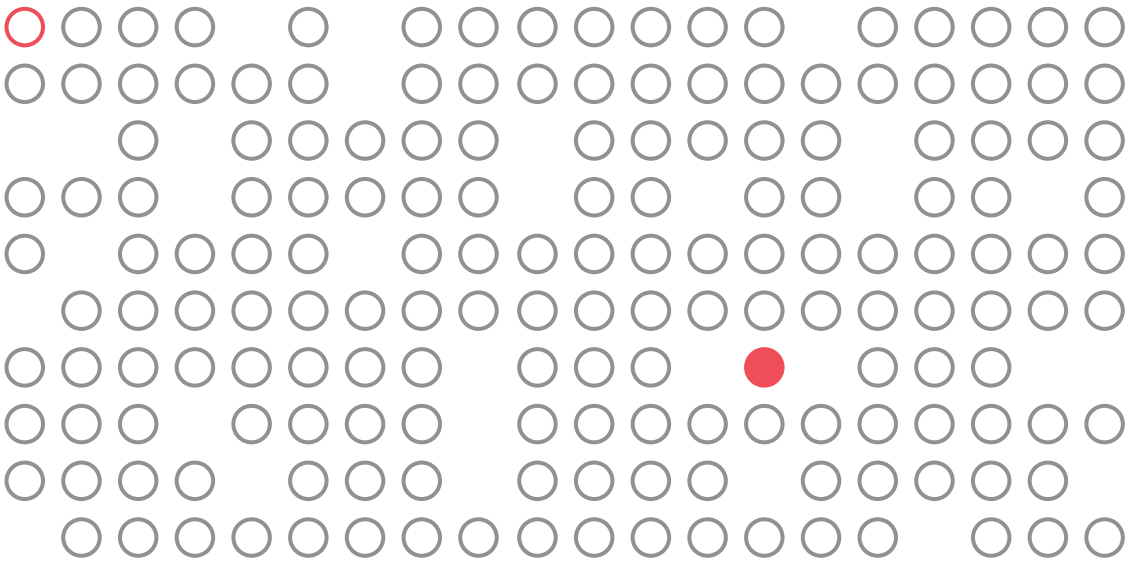
INAUGURAL DISSERTATION 2017

Lehrerurteile über Schülerleistungen

UNTERSUCHUNGEN ZUR DIAGNOSTISCHEN KOMPETENZ VON LEHRKRÄFTEN

Tobias Rausch



BAMBERG
GRADUATE SCHOOL
OF SOCIAL SCIENCES

Lehrerurteile über Schülerleistungen

Untersuchungen zur diagnostischen Kompetenz von Lehrkräften

Inaugural-Dissertation

in der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Tobias Rausch

aus Münchberg

Bamberg, den 21.12.2016

Tag der mündlichen Prüfung: 22.05.2017

Dekan: Prof. Dr. Stefan Hörmann

Erstgutachterin: Prof. Dr. Cordula Artelt

Zweitgutachter: Prof. Dr. Michael Hock

Danksagung

Der Abschluss einer akademischen Lebensphase, der sich in der vorliegenden Arbeit niederschlägt, ist für mich verbunden mit einem Rückblick auf diese Zeit und auf all diejenigen Menschen, die mich auf dem Weg zur Promotion professionell und privat begleitet haben.

Mein besonderer Dank gilt Prof. Dr. Cordula Artelt für ihre fachliche Unterstützung und Betreuung. Ihre stets fundierten und hilfreichen Hinweise und Anmerkungen haben zum Gelingen meines Dissertationsvorhabens wesentlich beigetragen. Zu ihrer guten Betreuung gehörte auch die stete Eröffnung von Möglichkeiten, mich persönlich, inhaltlich und akademisch weiterzuentwickeln.

Die Zeitschriftenbeiträge in dieser Arbeit sind im Austausch mit meinen Ko-Autorinnen und Ko-Autoren entstanden, denen ich ebenfalls herzlichen Dank sagen möchte. An Prof. Dr. Tobias Dörfler für die Begleitung meiner ersten akademischen Schritte und für den fachlichen Austausch. An Dr. Constance Karing für Diskussionen zur diagnostischen Kompetenz, die in die gemeinsame Arbeit eingeflossen sind. An Jacqueline Matthäi für ihre offenen Ohren und für die stets anregenden Bürodiskussionen um große Themen und kleine Satzteile.

Eine angenehme Zusammenarbeit und Arbeitsatmosphäre ist Grundlage für das Wohlfühlen und für die Produktivität am Arbeitsplatz. Danke daher an die Kolleginnen und Kollegen am Lehrstuhl für Empirische Bildungsforschung für ihre Unterstützung und für praktische Ratschläge zum akademischen Alltag.

Auch das weltläufige, produktive und entspannte Arbeitsumfeld an der Bamberg Graduate School of Social Sciences (BAGSS) hat meine Doktorandenzeit geprägt. Danke an die Kolleginnen und Kollegen der BAGSS für das Schaffen eines Umfelds, in dem sich auch und vor allem Gelegenheit zum Austausch bot, der mich über den Tellerrand des eigenen Fachbereichs schauen ließ und mir viele neue Perspektiven in Diskussionen, Kolloquien und Seminaren eröffnete.

Nicht zuletzt gilt mein ganz besonderer Dank meiner Familie, die immer an mich geglaubt hat und mich in all dem unterstützt hat, was ich mir in Studium und Promotionszeit vorgenommen und erreicht habe.

Bamberg im Dezember 2016

Lehrerurteile über Schülerleistungen

Untersuchungen zur diagnostischen Kompetenz von Lehrkräften

– Inhaltsverzeichnis –

1. Einleitung 3
2. Diagnostische Kompetenz von Lehrkräften 5
   2.1 Definition und begriffliche Abgrenzung 5
   2.2 Relevanz diagnostischer Kompetenz 7
   2.3 Diagnostische Kompetenz als Aspekt der Professionalität von Lehrkräften 9
   2.4 Ein heuristisches Modell der diagnostischen Urteilsbildung 10
3. Herangehensweisen an die Forschung zur diagnostischen Kompetenz 14
   3.1 Untersuchungen zur diagnostischen Kompetenz im Klassenraum 15
   3.2 Untersuchungen zur diagnostischen Kompetenz im Simulierten Klassenraum . 16
   3.3 Zum Verhältnis der beiden Herangehensweisen zueinander 18
4. Überblick über den Forschungsstand und Ableitung von Forschungsdesideraten 20
   4.1 Aspekte der diagnostischen Aufgabenstellung bei der Entstehung von Lehrerurteilen 21
   4.2 Aspekte der Informationsverarbeitung bei der Entstehung von Lehrerurteilen ... 23
   4.3 Aspekte des diagnostischen Handelns bei der Entstehung von Lehrerurteilen 25
5. Darstellung der zentralen Fragestellungen und Befunde der einzelnen Beiträge 27
   5.1 Beitrag 1: Personality similarity between teachers and their students influences teacher judgement of student achievement 27
   5.2: Beitrag 2: Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens 29
   5.3 Beitrag 3: Teacher judgment accuracy and assessment strategies in a Simulated Classroom 30
6. Diskussion 33
   6.1 Diskussion der Bewährung des heuristischen Modells 33
   6.2 Diskussion der Herangehensweisen an die Forschung zur diagnostischen Kompetenz 35
   6.3 Ausblick 37

Literaturangaben 38

Anhang (Verzeichnis der Originalbeiträge) 43

Lehrerurteile über Schülerleistungen

Untersuchungen zur diagnostischen Kompetenz von Lehrkräften

1. Einleitung

Lehrkräfte haben vielfältige Aufgaben im Schulsystem. Neben dem Unterrichten, Erziehen und Innovieren spielt das Beurteilen und Beraten eine zentrale Rolle (Deutscher Bildungsrat, 1970; Kultusministerkonferenz (KMK), 2014). In ihrer Funktion als Entscheider oder Berater bei Übergangsentscheidungen, aber auch bei der Vergabe von Noten und bei der formativen und summativen Beurteilung von Schülerleistungen fungieren Lehrkräfte als „Gatekeeper" im Schulsystem (Becker & Birkelbach, 2013). Damit zusammenhängende Entscheidungen und Urteile von Lehrkräften können für einzelne Schülerinnen und Schüler konkrete Lerngelegenheiten eröffnen, aber auch verschließen.

Generell erscheint es vor dem Hintergrund von Professionstheorien als funktional, Lehrkräften als professionell Lehrenden hinsichtlich ihrer Urteile zu vertrauen (vgl. Clement, 2012). Allerdings wurde in verschiedenen Untersuchungen festgestellt, dass an der Veridikalität (Hoge & Coladarci, 1989; Südkamp, Kaiser & Möller, 2012) und Reliabilität von Lehrkrafturteilen (Harlen, 2005) begründete Zweifel angemeldet werden können. Dies hat Auswirkungen auf die Verteilung von Lernchancen: Wenn aufgrund inakkurater oder verzerrter Urteile Schülerinnen und Schüler z.B. bei der Übergangsempfehlung einem weniger passenden Schultyp zugeordnet werden, beeinflusst dies trotz Korrekturmöglichkeiten im Schulsystem deren Bildungsbiografie. Auch auf der Mikroebene des Lehrens und Lernens können sich inakkurate Urteile über den aktuellen Lernstand von Schülerinnen und Schülern auf die Lernenden und ihren Lernfortschritt auswirken. Werden Verständnisprobleme von der Lehrkraft nicht bemerkt, Defizite falsch interpretiert oder falsch attribuiert, wird dem betreffenden Schüler oder der betreffenden Schülerin die Möglichkeit zum Weiterlernen verwehrt oder erschwert. Zahlreiche weitere Aspekte von professionellem Lehrerhandeln bauen auf der akkuraten Beurteilung von Situationen und Schülereigenschaften auf. Nur auf dieser Basis können informierte pädagogische und didaktische Entscheidungen getroffen werden.

Es existieren umfangreiche Befunde zur *Genauigkeit* von Lehrerurteilen in unterschiedlichen Kontexten (Hoge & Coladarci, 1989; Südkamp et al., 2012; Machts, Kaiser, Schmidt & Möller, 2016). Zur *Entstehung* dieser diagnostischen Urteile gibt es jedoch weiteren Forschungsbedarf.

Die vorliegende Arbeit[1] möchte zum besseren Verständnis des Zustandekommens von diagnostischen Lehrerurteilen beitragen und umfasst drei empirische Studien zur diagnostischen Kompetenz von Lehrkräften. Im Mittelpunkt stehen dabei Aspekte der Urteilsanforderungen und des Lehrerwissens sowie die Betrachtung des diagnostischen Handelns von (angehenden) Lehrkräften bei der Beurteilung von Schülerleistungen.

Ausgehend von einer Begriffsklärung und von Überlegungen zur Relevanz der diagnostischen Kompetenz als einem Aspekt der Professionalität von Lehrkräften wird ein heuristisches Modell der Urteilsbildung herausgearbeitet (Abschnitt 2). In den drei Studien werden zwei verschiedene Herangehensweisen an die Erforschung diagnostischer Kompetenz verwendet. Daher wird in Abschnitt 3 ein Überblick über diese Forschungsansätze gegeben und auf das Verhältnis der beiden Herangehensweisen zueinander eingegangen. Aus generellen theoretischen Überlegungen zur Entstehung von Lehrerurteilen bei der Beurteilung von Schülerleistungen werden Forschungsdesiderate für die durchgeführten empirischen Studien abgeleitet (Abschnitt 4) und die zentralen Fragestellungen und Befunde der drei Beiträge dargestellt (Abschnitt 5). Außerdem werden diese abschließend in einem Gesamtzusammenhang diskutiert (Abschnitt 6).

[1] Die vorliegende Arbeit wurde durch die Förderung der Bamberg Graduate School of Social Sciences (GSC 1024) im Rahmen der Exzellenzinitiative des Bundes und der Länder, sowie durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft im Rahmen der Bamberger Forschergruppe BiKS (Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vorschul- und Schulalter; FOR 543) ermöglicht.

2. Diagnostische Kompetenz von Lehrkräften

2.1 Definition und begriffliche Abgrenzung

Der Begriff der diagnostischen Kompetenz bezeichnet zunächst die Fähigkeit eines Menschen, Merkmale anderer Personen korrekt einzuschätzen (Schrader, 2010). Diese grundlegende Definition lässt sich auf vielfältige Kontexte anwenden, in denen soziale Interaktionen stattfinden. Dabei wird meist ein Bezug zu professionellen Kontexten insbesondere im medizinischen oder pädagogischen Bereich hergestellt. Aber auch jenseits professioneller pädagogischer Handlungsfelder wird der Anspruch auf eine akkurate Einschätzung von Merkmalen formuliert, beispielsweise gegenüber Elternteilen, die die Ängstlichkeit (Karing, Dörfler & Artelt, 2013) oder die Schulleistung (Frischknecht, Reimann, Gut, Ledermann & Grob, 2014) ihrer Kinder einschätzen, oder gegenüber Fremdeinschätzungen von Schülerselbstkonzepten durch Mitschülerinnen und Mitschüler (Pohlmann, Möller & Streblow, 2004).

Im Hinblick auf die *professionelle* Einschätzung von lern- bzw. leistungsrelevanten personenbezogenen Beurteilungsgegenständen im schulischen Bereich ist es nötig, den Begriff inhaltlich stärker zu spezifizieren, um unterrichtsbezogenen Urteilsanforderungen Rechnung tragen zu können. In schulischen Settings schätzen Lehrkräfte fortlaufend Merkmale ihrer Schülerinnen und Schüler ein. Der Fokus kann dabei u.a. auf dem Feststellen von Lernvoraussetzungen und Kompetenzen, auf der Überwachung des Lernfortschritts, auf der Abklärung von Lernschwierigkeiten (formatives assessment), oder auf der finalen Bewertung von Lernprozessen und -ergebnissen (summatives assessment) liegen (vgl. Schrader, 2011; Aufschnaiter et al., 2015). Zusammen mit der Zielsetzung des Urteils, die sich zwischen informeller Eindrucksbildung zur adaptiven Unterrichtsgestaltung einerseits und weitreichenden formellen Entscheidungen z.B. im Kontext der Schullaufbahnempfehlung andererseits bewegen kann, haben die genannten Zwecke auch Implikationen für die Vorgehensweisen und Methoden der Lehrkraft bei der Einschätzung von Schülermerkmalen.

Vor diesem Hintergrund soll in dieser Arbeit diagnostische Kompetenz verstanden werden als die Fähigkeit von Lehrkräften, Schülermerkmale akkurat einschätzen zu können und dafür Vorgehensweisen und Methoden auswählen zu können, die dem Zweck und dem Ziel der Diagnose angemessen sind. Verbunden ist damit auch der Anspruch, dass die so gewonnenen diagnostischen Einsichten handlungsleitend für darauf folgendes pädagogisches und didaktisches Handeln sind bzw. sein können (vgl. Helmke, 2012).

In der Literatur wird auf unterschiedliche Komponenten diagnostischer Kompetenz verwiesen: Neben der „diagnostischen Sensitivität im engeren Sinne“ (Schrader & Helmke, 1987, S. 33), die sich auf die Überprüfung der Übereinstimmung der Rangfolge von Schülerleistungen und Lehrerurteilen bezieht (Rangordnungskomponente), werden in einzelnen Untersuchungen zwei weitere Komponenten berichtet: die Niveaukomponente bezieht sich auf die Akkuratheit der Einschätzung des Leistungsniveaus in einer Klasse, während die Differenzierungskomponente auf die Genauigkeit der Einschätzung der Leistungsstreuung in der Klasse abzielt (zur Übersicht vgl. Karing, Matthäi & Artelt, 2011). Dies stellt eine inhaltlich sinnvolle Unterscheidung dar, die von Südkamp, Möller und Pohlmann (2008) durch ein globales Abweichungsmaß ergänzt wird. Damit wird die mittlere absolute Abweichung der Lehrereinschätzungen von den gezeigten Schülermerkmalen erfasst und das generelle Ausmaß der Verschätzung quantifiziert.

Der Begriff der diagnostischen *Kompetenz* impliziert, dass damit Aussagen über ein zeitlich stabiles und situationsunabhängiges Lehrermerkmal (trait) getroffen werden. Dieses Merkmal wird jedoch meist über die Urteilsgüte operationalisiert, was als ein zustands- und situationsabhängiges Merkmal (state) interpretiert werden kann (Artelt, 2016). Wenn eine Lehrkraft über hohe diagnostische Kompetenz verfügt, so wird diese – unter bestimmten theoretisch wie empirisch zu klärenden Umständen – bei relevanten Einschätzaufgaben eine hohe Urteilsgüte erzielen. Von der Urteilsgüte wird dann wiederum auf das dahinter liegende Konstrukt der diagnostischen Kompetenz, auf das „Bündel von Fähigkeiten“ geschlossen, das es der Lehrkraft ermöglicht, „den Kenntnisstand, die Lernfortschritte und die Leistungsprobleme der einzelnen Schüler … fortlaufend beurteilen zu können“ (Weinert, 2000, S. 19). Es zeigt sich allerdings, dass mit dem Begriff der diagnostischen Kompetenz nicht etwa eine generelle Fähigkeit gemeint sein kann, die sich empirisch entlang der genannten Komponenten in Kompetenzfacetten einteilen lässt (Spinath, 2005; vgl. Baumert & Kunter, 2006). Vielmehr wird damit ein Gedankenkonstrukt dargestellt, das Aussagen über die Akkuratheit von Lehrerurteilen auf den einzelnen Komponenten und bezogen auf die eingeschätzten Schülermerkmale differenziert zusammenfasst (vgl. Spinath, 2005).

Helmke (2012) grenzt den Begriff der diagnostischen Kompetenz im engeren Sinne bewusst von dem der diagnostischen Expertise ab. Damit soll über die bloße Urteilsgenauigkeit bzw. über die reine Übereinstimmung von Schülereigenschaft und Lehrerurteil hinaus ein umfassenderes Konzept verstanden werden, welches methodisches, prozedurales und konzeptuelles Wissen der Lehrkräfte mit einbezieht. Das Konzept

unterscheidet sich von der Definition Weinerts (2000) eher in den verwendeten Begrifflichkeiten als in den dahinter liegenden Ideen. So ist es für die fortlaufende Beurteilung von Kenntnisständen, Lernfortschritten und Leistungsproblemen ja nicht nur nötig, dass diese Urteile akkurat sind. Vielmehr kann angenommen werden, dass das von Helmke (2012) postulierte Wissen über Methoden und Vorgehensweisen sowie über Urteilsfehler und -tendenzen eine Grundlage für die Entstehung akkurater Urteile bildet. Die Urteilsgüte ist dabei lediglich als ein Indikator für die Ausprägung der diagnostischen Kompetenz bzw. Expertise zu sehen, der neben weiteren möglichen Indikatoren steht und u.a. durch diagnostisches Handeln und durch spezifische Wissensaspekte beeinflusst wird.

2.2 Relevanz diagnostischer Kompetenz

Diagnostisch kompetent zu handeln erscheint als unerlässlich für das erfolgreiche Unterrichten (z.B. Helmke, Hosenfeld & Schrader, 2004) und hat nachhaltige positive Auswirkungen auf die Individualisierung und Differenzierung im Unterricht (Weinert, 2000). Akkurate Urteile sind außerdem Grundlage für viele Lehrertätigkeiten, die Hattie (2009) als erfolgreiche Unterrichtsfaktoren für das Lernen identifiziert, z.B. für passendes Feedback an die Schüler, für die fortlaufende Überwachung des Lernerfolgs und für das erfolgreiche Durchführen adaptiven Unterrichts (vgl. Schrader, 2013). Auch über die hier nur beispielhaft genannten Tätigkeiten, die den Lernerfolg von Schülerinnen und Schülern positiv beeinflussen, hinaus ist es notwendig, dass Lehrkräfte informierte pädagogische und didaktische Entscheidungen treffen. Informiert sind diese Entscheidungen dann, wenn das Verhalten von Schülerinnen und Schülern korrekt erkannt und beurteilt wird (Funder, 1999) und nicht auf anderen, nicht direkt leistungs- oder performanzbezogenen Schülerinformationen beruht (vgl. Ready & Wright, 2011). Die empirische Befundlage bietet vielfältige Beispiele für potentiell verzerrende Informationen, wie u.a. für den Migrationsstatus des Schülers oder der Schülerin (z.B. Glock & Krolak-Schwerdt, 2013; Glock, Krolak-Schwerdt, Klapproth & Böhmer, 2013), den sozioökonomischen Status der Eltern (z.B. Alvidrez & Weinstein, 1999) oder für die Attraktivität von Schülerinnen und Schülern (z.B. Ritts, Patterson & Tubbs, 1992) gezeigt werden konnte. Zielsetzung von pädagogischen Diagnosen sollte es jedoch immer sein, Informationen zu identifizieren, die es erlauben, spezifische pädagogische und didaktische Entscheidungen und Handlungen abzuleiten (Trittel, Gerich & Schmitz, 2014). Edelenbos und Kubanek-German (2004) schließen diesen Aspekt einer angemessenen didaktischen Reaktion auf eine Diagnose mit in ihre Definition diagnostischer Kompetenz ein. Abs (2007) sowie Klug, Bruder, Kelava,

Spiel und Schmitz (2013) machen jedoch deutlich, dass zumindest die Urteilsgenauigkeit an sich zunächst keine Informationen darüber enthält, wie das sich anschließende pädagogische bzw. didaktische Lehrerhandeln optimal an den diagnostizierten Lernstand angepasst werden sollte (vgl. Hoth et al., 2016). Welche Entscheidungen basierend auf der Grundlage von Diagnosen getroffen werden und ob diese letztlich zielführend sind, geht über die Betrachtung der diagnostischen Kompetenz im Rahmen dieser Arbeit hinaus.

Die hohe Plausibilität eines Zusammenhangs zwischen den diagnostischen Fähigkeiten von Lehrkräften und der Leistungsentwicklung der jeweiligen Schülerinnen und Schüler wird häufig betont (z.B. Brunner, Anders, Hachfeld & Krauss, 2011), empirische Überprüfungen dieser Annahme zeichnen jedoch ein differenzierteres Bild. So ist der Lernerfolg von Schülerinnen und Schülern in Mathematik empirisch dann am größten, wenn Lehrkräfte mit einer hohen diagnostischen Kompetenz (bezogen auf die Rangordnungskomponente) gleichzeitig auch viele Strukturierungshilfen im Unterricht einsetzten (Schrader & Helmke, 1987). Karing, Pfost und Artelt (2011) fanden einen positiven Zusammenhang zwischen der Urteilsgüte bei der aufgabenspezifischen Einschätzung von Schülertestleistungen und der Entwicklung der Lesekompetenz von Schülerinnen und Schülern. Dieser Zusammenhang wurde jedoch von Unterrichtsvariablen wie Individualisierung und Einsatz von Strukturierungshilfen moderiert. Für die Rangordnungskomponente konnten in dieser Untersuchung jedoch keine positiven Zusammenhänge oder Wechselwirkungen mit der Leistungsentwicklung nachgewiesen werden. Weiterhin konnte gezeigt werden, dass unter Kontrolle von Kontextmerkmalen auf der Klassenebene die diagnostische Sensitivität der Lehrkräfte die Mathematikleistung der jeweiligen Klassen ein Schuljahr später positiv beeinflusst (Anders, Kunter, Brunner, Krauss & Baumert, 2010). Behrmann und Souvignier (2013) identifizierten eine Wechselwirkung von Feedback-Häufigkeit im Unterricht und hoher Urteilsgenauigkeit auf den Leistungszuwachs von Schülerinnen und Schülern in der Lesekompetenz. Diese differenzierten Befunde lassen vermuten, dass eine hohe Genauigkeit bei der Einschätzung der Schülerleistung alleine keinen direkten Leistungszuwachs bei Schülerinnen und Schülern zur Folge hat. Ein akkurates Erkennen von leistungsbezogenen Ausgangslagen (Vorwissen, Motivation, Strategien, Fähigkeiten etc.) der Schülerinnen und Schüler muss dazu erst noch von der Lehrkraft in konkrete pädagogische bzw. didaktische Handlungen überführt werden und sich (dadurch) in Unterrichtsprozessen niederschlagen (s.a. Schrader & Helmke, 1987; Klug et al., 2013). Neben diesen indirekten Effekten der diagnostischen Kompetenz auf die

Leistungsentwicklung der Schülerinnen und Schüler konnten auch Auswirkungen auf motivationale und emotionale Schülervariablen gefunden werden. Von ihren Lehrkräften unterschätzte Schülerinnen und Schüler schätzen sich selbst u.a. hinsichtlich ihres Fähigkeitsselbstkonzepts, ihrer Testängstlichkeit, und ihrer Lernzielorientierung ungünstiger ein als überschätzte Schülerinnen und Schüler (Urhahne, 2015).

2.3 Diagnostische Kompetenz als Aspekt der Professionalität von Lehrkräften

Insbesondere vor dem Hintergrund der beschriebenen zentralen Funktion der diagnostischen Kompetenz für das pädagogische und didaktische Lehrerhandeln verwundert es nicht, dass sich diese sowohl in Standards für die Lehrerbildung als auch in einschlägigen Kompetenzmodellen zum Lehrerberuf wiederfindet. In den Standards für Lehrerbildung der Kultusministerkonferenz (KMK, 2014) ist das Beurteilen einer von vier zentralen Kompetenzbereichen. Dabei geht es insbesondere um die Diagnose von Lernvoraussetzungen und Lernprozessen mit dem Ziel der Förderung und Beratung, sowie um die Erfassung von Schülerleistungen mit transparenten Beurteilungsmaßstäben. Weiterhin wird dort davon ausgegangen, dass für diese Anforderungen pädagogisch-psychologische und diagnostische Kompetenzen von Lehrkräften erforderlich sind.

Auch jenseits dieser Standards, die Ansprüche an erwünschtes, optimales Verhalten von Lehrkräften formulieren (Frey & Jung, 2011) und Qualitätsmerkmale beruflicher Kompetenzen markieren, stellen diagnostische Fähigkeiten neben der Klassenführungskompetenz, der fachwissenschaftlichen und der didaktischen Kompetenz eine der Schlüsselkompetenzen im Lehrerberuf dar (Weinert, Schrader & Helmke, 1990). In Kompetenzmodellen zum Lehrerberuf, die eher auf die diesen Standards zugrunde liegenden Fertigkeiten und Wissenselementen fokussieren (Frey & Jung, 2011) und das Benennen von Kompetenzkomponenten und Kompetenzstufen ermöglichen (Klieme et al., 2003) werden diese dementsprechend auch als eine Facette der professionellen Kompetenz von Lehrkräften (Brunner et al., 2011; Spinath, 2005) modelliert. Als zentrales Beispiel kann das COACTIV-Modell zur professionellen Kompetenz von Lehrkräften gelten. Hier werden diagnostische Fähigkeiten als mehrdimensionale Kompetenzfacette im Schnittbereich von fachdidaktischem Wissen und pädagogisch-psychologischem Wissen verortet (Brunner et al., 2011). Innerhalb des Modells erfordern diagnostische Fähigkeiten die Integration des Wissens über fachspezifische Kognitionen von Schülerinnen und Schülern und über kognitive Anforderungen von Aufgaben, sowie des fachunspezifischen Wissens um

Leistungsbeurteilung zum Zwecke der kognitiven Aktivierung des Unterrichts und des Aufbaus einer konstruktiv-unterstützenden Lernumgebung (Brunner et al., 2011).

2.4 Ein heuristisches Modell der diagnostischen Urteilsbildung

In der vorliegenden Arbeit wird von einem breiten Verständnis diagnostischer Kompetenz ausgegangen, welches sich nicht nur auf die Urteilsgüte als alleinigen Indikator für die Qualität des Urteils beschränkt, sondern sich auch auf den Prozess der Urteilsbildung und auf dabei wirksam werdende Einflussgrößen bezieht (Artelt & Rausch, 2014). Der Urteilsprozess umfasst dabei den Ablauf von der diagnostischen Aufgabenstellung bis zur Abgabe des Urteils. Er mündet schließlich in pädagogischem bzw. didaktischem Handeln der Lehrkraft, welches sich auf die vorher gesammelten Informationen und das daraus gebildete Urteil stützt. In Abbildung 1 wird ein im Folgenden näher beschriebenes heuristisches Modell der diagnostischen Urteilsbildung schematisch dargestellt.

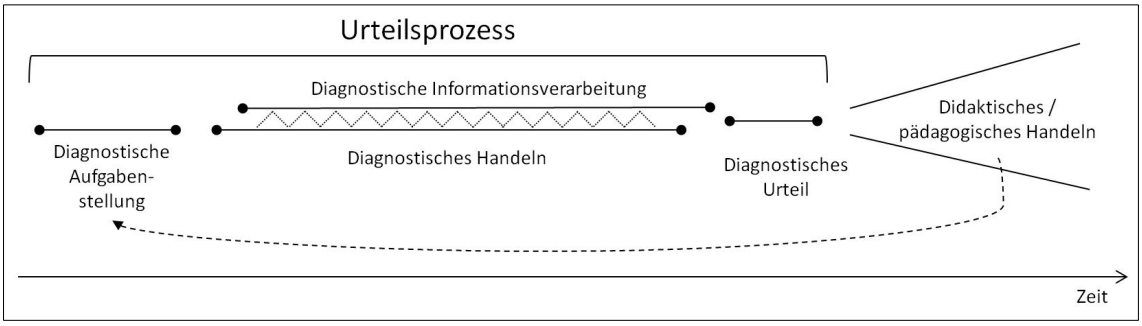


Abb. 1: Heuristisches Modell der diagnostischen Urteilsbildung

Die *diagnostische Aufgabenstellung* stellt den ersten Ausgangspunkt für alle weiteren Schritte im Urteilsprozess dar. Sie wird gespeist vom Zweck der Diagnose, d.h. in Abhängigkeit davon, welche Zielsetzungen mit dem didaktischen oder pädagogischen Handeln später verfolgt werden sollen. Beinhaltet die diagnostische Aufgabenstellung das Ziel, basierend auf einem zu treffenden Urteil eine unmittelbare mikrodidaktische Entscheidung zu treffen (z.B. welche Frage als nächstes an einen Schüler oder eine Schülerin gestellt werden soll), wird das diagnostische Handeln und die Informationsverarbeitung genauso wie das daraus resultierende Urteil anders beschaffen sein, als wenn der Urteilsprozess mit dem Ziel durchlaufen wird, aufbauend auf dem Urteil eine langfristig bedeutsame pädagogische Entscheidung zu treffen (z.B. eine Übergangsempfehlung), die mit einer höheren Verantwortlichkeit des Urteils einhergeht.

Beim *diagnostischen Handeln* – in der englischsprachigen Literatur auch als „classroom assessment practices“ (Randel & Clark, 2013, S. 145) bezeichnet – handelt es sich im Modell um das beobachtbare, zielgerichtete Vorgehen zur Erlangung von Informationen, um ein der diagnostischen Aufgabenstellung angemessenes funktionales Urteil abgeben zu können. Es bezieht sich auf die Planung und den adaptiven Einsatz von Methoden und Vorgehensweisen, mit denen Lernvoraussetzungen, Lernbedingungen und Lernergebnisse im Sinne einer diagnostischen Urteilsbildung ermittelt werden können. Zur Deskription des diagnostischen Handelns dienen meist Selbstberichte der Lehrkräfte zu ihren auf die Diagnostik bezogenen Vorgehensweisen (z.B. Jäger-Flor & Jäger, 2008) oder Auswertungen von Unterrichtsvideos (z.B. Gast, Herppich, Wittwer & Nückles, 2014). Außerdem erscheinen Beobachtungen, Interviews, Dokumentenanalysen oder Tests zur Erhebung geeignet (Randel & Clark, 2013), die z.T. auch die Arbeit mit Fallvignetten umfassen (z.B. Oser, Heinzer & Salzmann, 2010). Direkte handlungsnahe Erhebungen, mit denen das diagnostische Handeln von Lehrkräften direkt aufgezeichnet wird, existieren jedoch bislang kaum. Eine Ausnahme bildet hier die Untersuchung von Wylie und Lyon (2015), in der täglich erhobene Dokumentationen von diagnostischen Aktivitäten und Diskursen im Unterricht genutzt wurden, um Selbsteinschätzungen von Lehrkräften zu ihrem diagnostischen Handeln zu validieren. Außerdem wurde in zwei weiteren Untersuchungen das Informationssuchverhalten von Lehrkräften im Zusammenhang mit Übergangsempfehlungen beschrieben (Böhmer, Gräsel, Hörstermann & Krolak-Schwerdt, 2012; Böhmer, Hörstermann, Gräsel, Krolak-Schwerdt & Glock, 2015). Die Autoren nutzten dafür mit der Mouselab-Methode (Johnson, Payne, Schkade & Bettman, 1989) ein Computer-Programm, das die Analyse von Informationssuchprozessen bei Entscheidungs- und Urteilsaufgaben ermöglicht.

Die *diagnostische Informationsverarbeitung* bezieht sich auf die dem Urteil direkt vorauslaufenden Kognitionen der Lehrkräfte (Van Ophuysen & Lintorf, 2014), sowie auf die Verarbeitung der Informationen, die durch das diagnostische Handeln gesammelt wurden und weiteres diagnostisches Handeln leiten. Die Forschung nähert sich der Informationsverarbeitung bisher hauptsächlich über die Unterscheidung von Experten und Novizen. Die zentrale Annahme ist dabei, dass je nach Grad der Expertise inhaltliche und strukturelle Unterschiede im fachlichen, fachdidaktischen und diagnostisch-methodischen Wissen und in den Verarbeitungsprozessen existieren (z.B. van Ophuysen, 2006). In diesem Kontext konnte auch gezeigt werden, dass erfahrene Lehrkräfte flexibler und zielbezogener urteilen als Lehramtsstudierende (Dünnebier, Gräsel & Krolak-Schwerdt,

2009; Krolak-Schwerdt, Böhmer & Gräsel, 2009). Die Erfassung der diagnostischen Informationsverarbeitung geschieht dabei über die Handlungsebene, indem aus beobachtbarem Handeln kognitive Prozesse abgeleitet und als Informationsverarbeitungsstrategien interpretiert werden. Methodisch interessant ist aus dieser Perspektive der Ansatz, aus den von den Lehrkräften ausgewählten Informationen zu Schülerinnen und Schülern regelgeleitete bzw. informationsintegrierende Strategien der diagnostischen Informationsverarbeitung bei Übergangsempfehlungen abzuleiten (Böhmer et al., 2015). Offenbar gibt es jedoch über diese auf Schullaufbahnempfehlungen bezogenen Untersuchungen hinaus kaum Forschung zur diagnostischen Informationsverarbeitung von Lehrkräften bei der formativen und summativen Beurteilung von Schülerleistungen, die auch den Prozess der Urteilsbildung in den Blick nimmt.

Ein zentraler Punkt des heuristischen Modells der diagnostischen Urteilsbildung wird durch die fortlaufende Interaktion des diagnostischen Handelns mit der diagnostischen Informationsverarbeitung dargestellt. Die theoretische Grundlage dafür lässt sich aus der sozialen Kognitionsforschung ableiten (vgl. Bless, Fiedler & Strack, 2004). Die Initiierung diagnostischen Handelns ist der Informationsverarbeitung im Modell vorgelagert, da zunächst basierend auf der Aufgabenstellung ein zielgerichtetes Handeln ausgewählt werden muss. Dabei werden methodische, prozedurale und konzeptuelle Wissensbestände der Lehrkräfte relevant (Helmke, 2012). Die Informationen, die sich als sichtbares Ergebnis diagnostischen Handelns manifestieren (z.B. die Lösung einer Aufgabe durch einen Schüler), werden wahrgenommen, enkodiert, in vorhandene Wissensbestände eingeordnet und damit vor einem bestimmten Hintergrund interpretiert. Hier kommen domänenspezifische fachliche und fachdidaktische Wissensbestände zum Tragen, um in der Interaktion mit den Schülerinnen und Schülern bzw. in deren Arbeitsergebnissen relevante Hinweisreize erkennen und interpretieren zu können (Funder, 1995; Heritage, 2013). Auch Wissen über einzelne Schülerinnen und Schüler und fachübergreifendes diagnostisches Wissen, sowie Überzeugungen darüber, welche (ggf. weiteren) Hinweisreize Informationen über die Schülerleistung liefern können, sind Bestandteile des vorhandenen Wissens. Darauf baut die Interpretation der durch das diagnostische Handeln fortlaufend generierten und aktualisierten Informationen auf. Die interpretativ eingeordnete neue diagnostische Information wird abgespeichert und bildet gemeinsam mit dem so aktualisierten organisierten Wissen die Basis für die weitere Informationsverarbeitung und speist damit auch den nächsten Schritt des diagnostischen Handelns (vgl. Bless et al, 2004).

Alle weiteren diagnostischen Handlungen sollten sich adaptiv aus der wissens- und überzeugungsbasierten Informationsverarbeitung ergeben.

Das *diagnostische Urteil* steht als Ergebnis am Ende des Zusammenspiels aus diagnostischem Handeln und diagnostischer Informationsverarbeitung und ist auch abhängig von der diagnostischen Aufgabenstellung. Aus der Beschaffenheit des Urteils kann unter Umständen auf das diagnostische Handeln und auf die Informationsverarbeitung geschlossen werden (vgl. Bröder & Gaissmaier, 2007; Martignon & Hoffrage, 2002). Dabei kann angenommen werden, dass ein abgegebenes Urteil dann akkurat ist, wenn relevante und verfügbare Informationen von der Lehrkraft erkannt und für die Urteilsbildung genutzt wurden (Funder, 1995). Ein verzerrtes oder inakkurates Urteil kommt demnach dann zustande, wenn in das Urteil kriteriumsferne Aspekte eingeflossen sind (Ready & Wright, 2011). Das getroffene Urteil bildet dann die Grundlage für anschließendes didaktisches bzw. pädagogisches Handeln.

Innerhalb des Modells ist der Beurteilungsgegenstand zunächst unerheblich, genauso wie die Frage danach, ob einzelne Schülerinnen und Schüler oder eine Gruppe von Schülerinnen und Schülern eingeschätzt werden sollen. Diese Fragen beeinflussen jedoch alle Bestandteile des Modells (Aufgabenstellung, Handeln, Informationsverarbeitung, Zielsetzung des pädagogischen bzw. didaktischen Handelns). Weiterhin lassen sich neben formellen Diagnosen, bei denen das diagnostische Handeln reflektiert und methodisch kontrolliert in ein Urteil mündet, auch informelle Diagnosen mit dem Modell abbilden, die eher auf implizite und subjektive Urteile, Einschätzungen und Erwartungen basieren (Schrader & Helmke, 2001).

Basierend auf dem heuristischen Modell der Urteilsbildung werden in der vorliegenden Arbeit einzelne Aspekte des Urteilsprozesses schlaglichtartig betrachtet. Dabei wird neben der Betrachtung der Beschaffenheit der Aufgabenstellung zwischen der Qualität des diagnostischen Handelns, der Qualität der diagnostischen Informationsverarbeitung und der Qualität des diagnostischen Urteils unterschieden (vgl. Van Ophuysen & Lintorf, 2014). Der Begriff der Qualität soll dabei immer im Sinne der *Beschaffenheit* verstanden und verwendet werden und nicht als normative Ausprägung der *Güte* des Handelns, der Informationsverarbeitung oder des Urteils angesehen werden.

3. Herangehensweisen an die Forschung zur diagnostischen Kompetenz

Auf der Grundlage des eben beschriebenen heuristischen Modells erscheint es im Rahmen dieser Arbeit sinnvoll, die empirische Erforschung der diagnostischen Kompetenz von Lehrkräften bezogen auf die Bestandteile des Modells zu explizieren. Um den Einfluss potentiell erklärender Variablen auf die einzelnen Teile des Urteilsprozesses adäquat beschreiben zu können bedarf es unterschiedlicher Herangehensweisen, die die entsprechenden Aspekte jeweils gezielt abbilden können.

In den meisten bisherigen Studien wird die diagnostische Kompetenz von Lehrkräften in realen Schulklassen untersucht (z.B. Begeny, Krouse, Brown & Mann, 2011; Karing, 2011; Lorenz, 2011). Dazu wird die Fähigkeit von Schülerinnen und Schülern mit geeigneten Tests überprüft und Lehrkräfte werden gebeten, ihre Schülerinnen und Schüler hinsichtlich der mit dem Test gemessenen Fähigkeit einzuschätzen. Aus der Schülerperformanz im Test und den korrespondierenden Urteilen der jeweiligen Lehrkraft wird dann die Urteilsgüte ermittelt (siehe Abschnitt 2.1). Im Sinne des heuristischen Modells können damit Aussagen über die Beschaffenheit des diagnostischen Urteils, ggf. in Abhängigkeit von Urteilsanforderungen getroffen werden. Mit entsprechenden Forschungsdesigns kann die Auswirkung der Urteile auf die sich anschließenden pädagogischen und didaktischen Handlungen der Lehrkräfte analysiert werden. Die Effekte hoher bzw. niedriger diagnostischer Kompetenz auf die Entwicklung relevanter Schülereigenschaften und ihr Auftreten in Interaktion mit anderen Lehrer- oder Unterrichtsmerkmalen (siehe Abschnitt 2.3) können dabei ebenfalls im Fokus stehen.

Fragestellungen, die insbesondere auf den diagnostischen Urteilsprozess der Lehrkräfte bei der Urteilsbildung abzielen, und damit eine gezielte Betrachtung des diagnostischen Handelns oder der Informationsverarbeitung voraussetzen, benötigen jedoch unter Umständen andere empirische Ansatzpunkte. Gleiches gilt auch für Fragestellungen, die sich gezielt mit der Aufdeckung von Urteilsfehlern und Urteilsverzerrungen im Prozess beschäftigen.

Eine vielversprechende Herangehensweise an diese Fragestellungen ist die, Urteilsgüte und Urteilsprozesse in Abhängigkeit von bestimmten Bedingungen in einem Simulierten Klassenraum experimentell zu untersuchen (z.B. Fiedler, Walther, Freytag & Plessner, 2002; Fiedler, Freytag & Unkelbach, 2007; Südkamp, Möller & Pohlmann, 2008; Südkamp & Möller, 2009; Kaiser, Helm, Retelsdorf, Südkamp & Möller, 2012). Dazu werden Studienteilnehmer in einer Computersimulation jeweils in die Rolle einer Lehrkraft versetzt, die die Aufgabe hat, mit virtuellen Schülerinnen und Schülern so zu interagieren,

dass basierend auf den dargebotenen und abgerufenen leistungsbezogenen und nichtleistungsbezogenen Informationen ein Urteil über die Schülerinnen und Schüler getroffen werden kann (Kaiser & Möller, 2016). Solche Untersuchungen arbeiten zwar mit einer artifiziellen und komplexitätsreduzierten Untersuchungsumgebung, jedoch können dabei durch die im Unterricht nicht mögliche gezielte Variation von Schüler- und Klasseneigenschaften sowie der verfügbaren Hinweisreize und der diagnostischen Aufgabenstellung detaillierte Aussagen über das diagnostische Handeln im Urteilsprozess und über dessen Zusammenhang mit der Urteilsgüte getroffen werden.

Je nach den unmittelbar interessierenden Aspekten des Urteilsprozesses sollte die Herangehensweise an die Untersuchung diagnostischer Kompetenz gezielt ausgewählt und angewendet werden. Im Folgenden werden zunächst die beiden Herangehensweisen näher beschrieben und anschließend zueinander in einen empirischen, theoretischen und forschungspraktischen Bezug gesetzt.

3.1 Untersuchungen zur diagnostischen Kompetenz im Klassenraum

Die Erhebung von Schülerleistungen und korrespondierenden Lehrkrafturteilen im realen Klassenraum ist eine intuitiv nahe liegende Herangehensweise und verspricht aufgrund der Nähe zu den im Schulalltag gegebenen Kontextbedingungen ökologische Validität. Lehrkräfte haben im Unterricht zahlreiche Gelegenheiten, diagnostisch zu handeln und Erfahrungswissen über ihre Schülerinnen und Schüler aufzubauen, auf dessen Grundlage sie deren einzuschätzende Fähigkeiten beurteilen können. In den meisten dieser Untersuchungen werden Aussagen darüber getroffen, wie gut eine Lehrkraft zu einem gegebenen Zeitpunkt ihre Schülerinnen und Schüler anhand bestimmter Kriterien eingeschätzt hat. Es kann angenommen werden, dass aus der Urteilsgüte der Lehrkräfte in dieser konkreten Testsituation auf deren Urteilsgüte im Unterricht geschlossen werden kann.

Meta-Analysen zeigen übereinstimmend, dass die Urteilsgüte von Lehrkräften bei der Einschätzung der Reihenfolge ihrer Schülerinnen und Schüler hinsichtlich deren Leistung im mittleren Bereich liegt (Median-Korrelation von $r = .53$ bei Südkamp et al. 2012 und $r = .66$ bei Hoge und Coladarci, 1989). Auffällig sind hier jedoch die deutlichen interindividuellen Unterschiede zwischen den Lehrkräften, die sich in einer großen Bandbreite der Urteilsgüte innerhalb der einzelnen Studien niederschlägt. Lehrkräfte scheinen darüber hinaus das Leistungsniveau ihrer Schülerinnen und Schüler regelmäßig zu überschätzen (z.B. Bates & Nettelbeck, 2001; Helmke et al., 2004), während die

Befundlage für die Einschätzung der Streuung von Schülermerkmalen über die Klasse uneinheitlich erscheint (Helmke et al., 2004; aber auch Brunner, Anders, Hachfeld, & Krauss, 2011).

Eine detaillierte und schrittweise Beobachtung des diagnostischen Handelns im Unterricht ist in realen Klassenräumen direkt nur schwierig möglich. Annäherungen über detaillierte Selbstberichte (z.B. Jäger-Flor & Jäger, 2008; Wylie & Lyon, 2015) können jedoch kaum den Prozess der Urteilsbildung in der Interaktion mit Schülerinnen und Schülern sowie bezogen auf die Auswahl und Interpretation von Aufgabenmaterialien abbilden. Auch die diagnostische Informationsverarbeitung ist nur eingeschränkt im realen Klassenraum zu untersuchen, da die Informationsgrundlage für die Informationsverarbeitung bei den Lehrkräften vermutlich umfassender ist, als die in den Untersuchungen jeweils interessierenden und erhobenen Schüler- und Kontextmerkmale. Vor dem Hintergrund einer quantitativen wie qualitativen Ungleichverteilung der Lehrer-Schüler-Interaktionen im Unterricht (Lipowsky, Rakoczy, Pauli, Reusser & Klieme, 2007) kann auch die Menge und die Verfügbarkeit von Informationen über einzelne Schülerinnen und Schüler die Qualität der Einschätzung eines Schülermerkmals beeinflussen.

Für die systematische Bearbeitung bestimmter Fragestellungen insbesondere zum Prozess der diagnostischen Urteilsbildung zeichnen sich mit der Herangehensweise über Untersuchungen im realen Klassenraum also Grenzen ab, die sich mit der Verwendung des Simulierten Klassenraums als Forschungsinstrument zum Teil überwinden lassen.

3.2 Untersuchungen zur diagnostischen Kompetenz im Simulierten Klassenraum

Brown (1999) fasst potentielle Effekte neuer Technologien auf die Lehrerausbildung zusammen und diskutiert unter anderem auch den Einsatz von Simulationen zu instruktionalen Zwecken: Während in realen Schulklassen Lehrkrafturteile und die daraus gezogenen pädagogischen und didaktischen Konsequenzen unmittelbaren Einfluss auf die Schülerinnen und Schüler haben, und dabei gemachte Fehler eben Fehler bleiben, kann im Simulierten Klassenraum kein Schaden an realen Schülerinnen und Schülern angerichtet werden. Dies macht den Simulierten Klassenraum als Instrument zur Übung und Reflexion für Aus- und Fortbildungssituationen im Bereich des Lehramts interessant, bietet aber insbesondere auch Potential als Untersuchungsumgebung zur Erforschung der diagnostischen Kompetenz von Lehrkräften.

Dazu wird mit dem Simulierten Klassenraum eine komplexitätsreduzierte Beurteilungssituation geschaffen, in der die Teilnehmenden die Rolle einer Lehrkraft

einnehmen und mit den gegebenen Elementen der simulierten Umwelt (Schülerinnen und Schüler sowie Aufgaben, ggf. Vorinformationen über die Schülerinnen und Schüler) interagieren (Heinich, Molenda & Russell, 1993; Brown, 1999). Obwohl die Komplexität einer solchen Untersuchungsumgebung verglichen mit der realen Umwelt geringer ist und nicht alle Aspekte des unterrichtlichen Handelns abgebildet werden können, ist es dennoch möglich, Situationen zu schaffen, in denen eng definierte Aufgaben erfüllt werden müssen, die in der Praxis Teil des Lehrerhandelns sind. Abstrahiert von zusätzlichen Anforderungen, denen Lehrkräfte sonst im Unterricht begegnen, kann hier unter optimalen (nicht abgelenkten) Urteils- und Informationsverarbeitungsbedingungen überprüft werden, zu welchen Urteilsleistungen Lehrkräfte in der Lage sind, und welche Fehler ihnen dennoch unterlaufen (Artelt, Krolak-Schwerdt, Hörstermann & Rausch, 2015). Südkamp, Möller & Pohlmann (2008) berichten Befunde, wonach Lehramtsstudierende die Rangfolge der Schülerinnen und Schüler im Simulierten Klassenraum relativ akkurat einschätzen konnten. Das Leistungsniveau der simulierten Schülergruppe wurde jedoch tendenziell überschätzt, während die Streuung der Leistungen in der Gruppe unterschätzt wurde. Die Ergebnisse stehen in der Tendenz nicht im deutlichen Widerspruch zu Befunden von Untersuchungen im realen Klassenraum. Kaiser, Retelsdorf, Südkamp und Möller (2013) konnten darüber hinaus bei jeweils gleichartig strukturierten Beurteilungsaufgaben Hinweise darauf finden, dass Lehrereinschätzungen zu Schülerleistung und Motivation im Simulierten Klassenraum akkurater waren als im realen Klassenraum. Dies kann zum Teil auf die Komplexitätsreduktion zurückgeführt werden, wird doch in einem experimentellen Design der Fokus auf bestimmte interessierende Variablen gelegt, die dann in einer nichtnatürlichen Umgebung stärker zu Tage treten (Klauer, 1973/2005). Schülerleistung und Motivation (in der zitierten Studie operationalisiert über die Meldehäufigkeit der simulierten Schülerinnen und Schüler) sind durch die Komplexitätsreduktion im Simulierten Klassenraum unmittelbarer und direkter beobachtbar, während die zu beurteilenden Merkmale im realen Klassenraum nicht immer der direkten Beobachtbarkeit zugänglich sind (vgl. Kaiser et al., 2013) und von den Beurteilenden zum Teil anders erschlossen werden müssen.

Die Vorteile der Herangehensweise über den Simulierten Klassenraum liegen vor allem in der Möglichkeit der experimentellen Variation von Schüler- und Klasseneigenschaften. So können hier neben den Performanzparametern der Schülerinnen und Schüler auch andere Eigenschaften (z.B. deren Geschlecht, physische Attraktivität, sozioökonomischer Status, Herkunft, vorherige Leistungen, etc.) manipuliert werden, was in realen Klassenräumen so

nicht möglich ist. Auch die Zusammensetzung der Klasse hinsichtlich darstellbarer Schülermerkmale (z.B. Sitzordnung oder Geschlechterverteilung) ist beeinflussbar (Fiedler et al., 2002). Dies kann zur experimentellen Kontrolle von Ursache-Wirkungs-Zusammenhängen dienen, Erkenntnisse aus Studien absichern, die in realen Klassenräumen durchgeführt wurden (Südkamp, Kaiser & Möller, 2014), aber auch helfen, Quellen von Urteilsfehlern aufzudecken. Außerdem kann durch Reduktion und gezielte Variation der gegebenen bzw. abrufbaren Informationen über Schülerinnen und Schüler die Informationsbasis, auf der die Urteile aufbauen, kontrolliert untersucht werden.

Mit der Aufzeichnung von Verlaufsdaten im Simulierten Klassenraum können zudem detaillierte Daten über das diagnostische Handeln der Lehrkräfte erfasst werden. Ohne die sequenzielle Vorgehensweise bei der Informationssammlung detailliert zu betrachten, kann – wenn überhaupt – nur basierend auf dem gegebenen Urteil darauf geschlossen werden, auf welche Informationen sich die Urteile stützen (können), weil diese explizit betrachtet wurden. Betrachtet man jedoch alleine das getroffene Urteil, können durchaus unterschiedliche Strategien zum gleichen Urteilen geführt haben (Martignon & Hoffrage, 2002; s.a. Bröder & Gaissmaier, 2007). Durch die explizite Beobachtung und Aufzeichnung der Vorgehensweisen können Informationen darüber gewonnen werden, ob und wie das Vorgehen sich auch in der Urteilsgüte widerspiegelt.

3.3 Zum Verhältnis der beiden Herangehensweisen zueinander

Der Simulierte Klassenraum sollte explizit nicht als Ersatz für Untersuchungen in realen Klassenräumen betrachtet werden, sondern als eine Möglichkeit, mit der spezielle Situationen und diagnostische Aufgabenstellungen fokussiert betrachtet werden können. Er stellt daher eine vielversprechende Ergänzung zu Untersuchungen im realen Klassenraum dar (Brown, 1999; Schrader, 2010; Spinath, 2012). Ein Versuch, die beiden Untersuchungsansätze zusammenzufügen, wurde von Kaiser und Kollegen (2013) unternommen. Der Einfluss der Schülermotivation auf die Lehrereinschätzung im Leistungsbereich und der Einfluss der Schülerleistung auf die Einschätzung der Motivation wurde hier zunächst mit der Herangehensweise über die Untersuchung diagnostischer Kompetenz in realen Klassenräumen untersucht. Die dabei beobachteten Effekte konnten anschließend im Simulierten Klassenraum (jedoch nicht mit derselben Stichprobe) weitgehend repliziert werden. Offen bleibt jedoch weiterhin die Frage, ob diese Ergebnisse auch so beobachtbar sind, wenn dieselben Personen in beiden Untersuchungsumgebungen getestet werden (Kaiser et al., 2013). Die Studie bietet jedoch einen Anhaltspunkt dafür,

dass für die Bewältigung der Urteilsanforderungen im realen wie im Simulierten Klassenraum ähnliche – wenn nicht gar die gleichen – diagnostischen Kompetenzen nötig sind (Kaiser et al., 2013; Südkamp et al., 2014). Unter dieser Annahme kann die Auswertung und Systematisierung des diagnostischen Handelns und der Informationsverarbeitung im Simulierten Klassenraum dazu dienen, weitere Varianz in der Urteilsgüte aufzuklären und zusätzlich auch erste Ansatzpunkte für die Förderung diagnostisch kompetenten Lehrerhandelns liefern.

Sowohl im realen als auch im Simulierten Klassenraum sammeln Lehrkräfte innerhalb einer Zeitspanne Informationen über Schülerinnen und Schüler, die dabei je nach pädagogischer bzw. didaktischer Zielsetzung und diagnostischer Aufgabenstellung verarbeitet werden, um ein möglichst akkurates, zielführendes Urteil zu bilden. In der unterrichtlichen Praxis ist diese Zeit unbestimmter als im Simulierten Klassenraum, der Urteilsprozess kann daher nicht oder nur sehr schwierig als Ganzes beobachtet werden. Im realen Klassenraum kommt darüber hinaus mit dem unvermeidlich vorhandenen Vorwissen der Lehrkräfte über ihre Schülerinnen und Schüler eine weitere schwer zu kontrollierende Einflussgröße zum Tragen, welche das diagnostische Handeln und die Informationsverarbeitung und damit auch die Beschaffenheit des Urteils beeinflussen kann. Der gezielten Untersuchung von diagnostischem Handeln und diagnostischer Informationsverarbeitung sind daher im realen Klassenraum Grenzen gesetzt.

Im Simulierten Klassenraum hingegen erfolgt die Informationssammlung und Informationsverarbeitung innerhalb einer bestimmten, komplett beobachtbaren Zeitspanne unmittelbar vor der Abgabe konkreter Urteile. Die Lehrkräfte haben dabei entweder kein Vorwissen über die zu beurteilenden simulierten Schülerinnen und Schüler, oder aber die Vorinformationen werden in der Untersuchung gezielt manipuliert, um Auswirkungen dieser Informationen auf das diagnostische Handeln, die Informationsverarbeitung und die Urteilsgüte zu untersuchen (für Übergangsentscheidungen z.B. bei Glock et al., 2013). Eine gezieltere Betrachtung der Urteilsprozesse ist insbesondere auch wegen der Eliminierung zusätzlicher Aufgaben der Lehrkräfte (z.B. classroom management) und dem damit einhergehenden stärkeren Fokus auf die diagnostische Aufgabenstellung im Simulierten Klassenraum möglich. Dadurch können Situationen geschaffen werden, in denen diagnostische Entscheidungen bewusster und reflektierter getroffen werden können. Dies ermöglicht wiederum Aussagen darüber, zu welchen diagnostischen Leistungen Lehrkräfte unter optimalen, nicht abgelenkten Bedingungen in der Lage sind.

4. Überblick über den Forschungsstand und Ableitung von Forschungsdesideraten

Die Herleitung der Forschungsdesiderate für die vorliegende Arbeit nehmen die Teilbereiche der diagnostischen Aufgabenstellung, des diagnostischen Handelns und der diagnostischen Informationsverarbeitung in den Blick. Dabei werden die Fragestellungen jeweils auf die Beschaffenheit des diagnostischen Urteils bezogen. Dadurch wird der Bezug zu dem oben eingeführten heuristischen Modell der diagnostischen Urteilsbildung hergestellt. Auf der Basis der diagnostischen Aufgabenstellung muss je nach Ziel und Zweck der Diagnose eine Vorgehensweise ausgewählt werden, mit der relevante und verfügbare Hinweisreize zielführend gesammelt und aufgenommen werden können (Beschaffenheit des diagnostischen Handelns) (vgl. Funder, 1999). Im Rückgriff auf fachliches und fachdidaktisches Wissen, auf Überzeugungen darüber, welche Beobachtungen über Schülerfähigkeiten Auskunft geben können, sowie auf das evtl. verfügbare (bzw. sich im Urteilsprozess kumulierende) Wissen über einzelne Schülerinnen und Schüler werden dabei die gesammelten Hinweisreize verarbeitet (Beschaffenheit der diagnostischen Informationsverarbeitung) und das diagnostische Handeln gegebenenfalls angepasst. Dabei spielt auch das Wissen und Überzeugungen der Lehrkräfte über die einzuschätzende Schülerfähigkeit, sowie die Einschätzung der konkreten Aufgabenanforderungen eine Rolle. Dies mündet schließlich in ein Urteil über die einzuschätzende Fähigkeit der jeweiligen Schülerinnen und Schüler (Beschaffenheit des diagnostischen Urteils), welches dann für entsprechende didaktische oder pädagogische Entscheidungen genutzt werden kann.

Die drei in Abschnitt 5 beschriebenen Beiträge greifen jeweils unterschiedliche Aspekte des heuristischen Modells heraus. Zur Einordnung der später präsentierten Ergebnisse folgt nun ein Überblick über theoretische Annahmen und über den jeweiligen Forschungsstand zu Urteilsanforderungen als Derivat der diagnostischen Aufgabenstellung, zu Aspekten des Lehrerwissens als Teilbereich der diagnostischen Informationsverarbeitung und zum diagnostischen Handeln als informationelle Grundlage der Urteilsbildung. Daraus werden nun jeweils Forschungsdesiderate für die Untersuchungen abgeleitet.

4.1 Aspekte der diagnostischen Aufgabenstellung bei der Entstehung von Lehrerurteilen

Im Modell der diagnostischen Urteilsgenauigkeit von Lehrkräften (Südkamp, 2010; Südkamp et al., 2012) werden neben Schülermerkmalen, Lehrermerkmalen und Merkmalen des eingesetzten Tests die an die Lehrkraft gestellten Urteilsanforderungen als (potentieller) Moderator für die Beschaffenheit des diagnostischen Urteils gesehen. Südkamp, Kaiser und Möller (2012) unterscheiden hier zwischen informierten Urteilen, bei denen Lehrkräfte bei der Einschätzung die einzelnen Aufgaben des Tests, mithin also den konkreten Vergleichsmaßstab, kennen, und uninformierten Urteilen, bei denen Lehrkräfte die Schülerperformanz in Unkenntnis eine konkreten Vergleichsmaßstabs einschätzen. In ihrer Metaanalyse zeigte sich, dass der Zusammenhang zwischen Lehrerurteil und Schülerleistung enger war, wenn ein informiertes Urteil zu treffen war (Südkamp et al., 2012; vgl. a. Hoge & Coladarci, 1989). Eine speziellere Unterscheidung zwischen informierten und uninformierten Urteilsanforderungen findet sich bei Karing, Matthäi und Artelt (2011), die je nach Spezifität der Urteile zwischen einer globalen Urteilsdimension (Einschätzung eines in der Beurteilungsaufgabe nicht näher operational bestimmten globalen Schülermerkmals) und einer aufgabenspezifischen Urteilsdimension (Einschätzung von Schülerleistungen bei der Bearbeitung einer Anzahl von vorliegenden Aufgaben) unterscheiden (vgl. Artelt & Gräsel, 2009; Helmke et al., 2004). Diese unterschiedlich gefassten Urteilsanforderungen werden in der vorliegenden Arbeit als Teil der diagnostischen Aufgabenstellung im heuristischen Modell der der diagnostischen Urteilsbildung interpretiert. Damit in Verbindung stehen entsprechende Auswirkungen auf das diagnostische Handeln und die diagnostische Informationsverarbeitung, sowie in der Konsequenz auch auf die Beschaffenheit des diagnostischen Urteils.

Globale Urteilsanforderungen, bei denen Lehrkräfte in Unkenntnis des Vergleichsmaßstabs auf einer mehrstufigen Rating-Skala ein globales Urteil über Schülerleistung abgeben sollen (z.B.: „Der Schüler / die Schülerin ist im Vergleich zum Durchschnitt: sehr schwach … sehr gut in Arithmetik“; BiKS Forschergruppe, o.J.) zielen eher auf eine Eindrucksbildung ab. Diese pädagogische bzw. didaktische Zielsetzung impliziert eine entsprechende diagnostische Aufgabenstellung und beeinflusst in der Folge auch die weiteren Schritte im Urteilsprozess bis hin zur Urteilsgüte. Diese hängt dabei auch davon ab, was die Lehrkraft unter dem einzuschätzenden Konstrukt versteht und welche Hinweisreize im Rahmen des diagnostischen Handelns genutzt und verarbeitet werden (Karing, Matthäi & Artelt, 2011). Globale Urteilsanforderungen bieten der Lehrkraft

wenig Struktur für die diagnostische Informationsverarbeitung, so dass hier bspw. Heuristiken ins Spiel kommen können, in denen Hinweisreize verwendet werden, die nicht notwendigerweise mit der Schülerleistung in Verbindung stehen, aber für die einschätzende Lehrkraft vor dem Hintergrund einer wenig verbindlichen Zielsetzung im Urteilsprozess einfacher abrufbar sind (vgl. Kahneman, 2011).

Aufgabenspezifische Urteilsanforderungen hingegen, bei denen Lehrkräfte für einzelne Schülerinnen und Schüler Einschätzungen über deren Performanz bei konkreten Aufgaben aus dem Test abgeben sollen (z.B. „Er/Sie kann die Aufgabe lösen / nicht lösen"; BiKS Forschergruppe, o.J.), zielen hingegen eher auf die Integration von Aufgaben- und Personenwissen ab (Karing, Matthäi & Artelt, 2011). Solche aufgabenspezifischen Urteilsanforderungen sind als Implikationen aus konkreten didaktischen Zielsetzungen zu sehen, die über eine einfache Eindrucksbildung hinausgehen und ggf. mit einer höheren Verantwortlichkeit des Urteils verbunden sind. Die Kenntnis der Aufgabencharakteristika bei aufgabenspezifischen Urteilsanforderungen kann strukturierende Wirkung für den Urteilsprozess haben (Dipboye & Gaugler, 1993). Die bei der globalen bzw. uninformierten Urteilsanforderung mitschwingende Unsicherheit über das einzuschätzende Merkmal ist hier eliminiert. Die Einschätzung der Schülerfähigkeit kann vielmehr an konkreten Aufgabenmerkmalen festgemacht werden. Es kann angenommen werden, dass dadurch der Urteilsprozess hier weniger anfällig für die Einbeziehung leistungsferner Schülermerkmale ist.

Basierend auf diesen Überlegungen wird angenommen, dass unterschiedliche diagnostische Aufgabenstellungen, die sich aus unterschiedlichen didaktischen Zielsetzungen ableiten lassen, den weiteren Urteilsprozess und damit auch die Urteilsgüte differenziell beeinflussen. So kann angenommen werden, dass je nach Aufgabenstellung im weiteren Verlauf des diagnostischen Urteilsprozesses unterschiedliche Hinweisreize gesucht und verarbeitet werden. Dabei ergibt sich folgendes erstes Forschungsdesiderat:

Wird die Urteilsgüte je nach diagnostischer Aufgabenstellung (globale vs. aufgabenspezifische Urteile) unterschiedlich stark von nicht-leistungsbezogenen Schülerinformationen beeinflusst?

4.2 Aspekte der Informationsverarbeitung bei der Entstehung von Lehrerurteilen

Für die Entstehung von akkuraten Urteilen bei der Einschätzung von Schülerleistungen sind zunächst grundlegende methodische Wissensbestände der Lehrkraft nötig (Helmke et al., 2004). Insbesondere die Kenntnis und Beherrschung diagnostischer Methoden sollte sich demnach im Urteilsprozess niederschlagen. Generelles Wissen über Diagnostik und Urteilsfehler zeigt sich daher auch empirisch als substantieller Prädiktor der diagnostischen Kompetenz von Lehrkräften: Klug, Bruder und Schmitz (2015) konnten zeigen, dass Lehramtsstudierende und Lehrkräfte, die eine hohe Performanz in einem diagnostischen Wissenstest erzielten, auch höhere Leistungen in einem auf der Beurteilung von Vignetten basierenden Test der diagnostischen Kompetenz zeigten. Weiterhin spielt Wissen über die einzuschätzenden Personen und Personengruppen („knowledge of learners and their characteristics"; Shulman, 1987, S. 8), beispielsweise über individuelle Stärken und Schwächen, spezifische Lösungsstrategien bei der Bearbeitung von Aufgaben, aber auch über das generelle Leistungsniveau der unterrichteten Schulklasse eine Rolle für die Art und Weise, wie neue Informationen verarbeitet werden.

Ein zentraler Aspekt, der insbesondere für die diagnostische Informationsverarbeitung Relevanz besitzt, ist darüber hinaus das bereichsspezifische Wissen der Lehrkraft. Hinweise darauf bieten Untersuchungen, die zeigen, dass Lehrkräfte nicht über verschiedene Domänen hinweg konstant gleich gute Urteile abgeben (z.B. Eckert, Dunn, Codding, Begeny & Kleinmann, 2006; Hopkins, George & Williams, 1985; Lorenz & Artelt, 2009). Diagnostische Kompetenz von Lehrkräften erscheint daher als eine bereichsbezogene Fähigkeit, die in Abhängigkeit des einzuschätzenden Merkmals variiert (Spinath, 2005; Lorenz & Artelt, 2009; Schrader, 2010). Dementsprechend sollten auch bereichsspezifische Wissenskomponenten eine Rolle bei der Entstehung von Lehrerurteilen spielen. Dieses geht über methodische und schülerbezogene Wissensbestände hinaus und äußert sich z.B. im Wissen über die Schwierigkeit von Aufgaben bzw. des zu bearbeitenden Materials, über Anforderungen im jeweiligen Lerngebiet, förderliche und hinderliche Lösungsstrategien, oder über typische Fehler bei der Aufgabenbearbeitung (Helmke et al., 2004).

Basierend auf dem Realistic Accuracy Model (Funder, 1995; 1999) kann angenommen werden, dass Wissensaspekte insbesondere beim Erkennen und bei der Nutzung von Hinweisreizen eine Rolle spielen, und damit auf der Ebene der diagnostischen Informationsverarbeitung angesiedelt werden können. Nur wenn die Lehrkraft über entsprechendes fachliches und fachdidaktisches Wissen verfügt, können die im

Urteilsprozess verfügbaren Hinweise als relevant erkannt werden, entsprechend aufgenommen und eingeordnet werden sowie für die Urteilsbildung eingesetzt werden. Weiterhin kann in Anlehnung an das Tetraeder-Modell (Campione & Armbruster, 1985) argumentiert werden, dass Lehrkräfte, die die Fähigkeit von Schülerinnen und Schülern in einem Inhaltsbereich akkurat einschätzen wollen, Informationen aus verschiedenen Quellen integrieren müssen: die Leistung eines Schülers bzw. einer Schülerin bei der Bearbeitung einer bestimmten Aufgabe ist demnach abhängig von Merkmalen wie dem Vorwissen, der Intelligenz oder der Motivation, von den Aktivitäten, die die Schülerin bzw. der Schüler zur Lösung der Aufgabe unternimmt (z.B. Anwendung angemessener Strategien), von den Anforderungen der zu bearbeitenden Aufgabe und von der Beschaffenheit des Materials (vgl. Artelt, 2016).

Die Relevanz bereichsspezifischer Wissensaspekte bei der Entstehung von Lehrerurteilen lässt sich somit plausibel aus den skizzierten theoretischen Annahmen ableiten. Konkrete fachbezogene, deklarative, prozedurale und konditionale Wissensaspekte und ihr Zusammenhang mit der Urteilsgüte wurden bisher jedoch kaum empirisch betrachtet. Untersuchungen, die sich dem Zusammenhang zwischen fachlichen und fachdidaktischen Wissensaspekten von Lehrkräften und der Urteilsgüte widmen, nähern sich dem Wissen meist über formale Indikatoren. So konnten beispielsweise Johansson, Strietholt, Rosén und Myrberg (2014) zeigen, dass eine aus formalen Indikatoren zusammengesetzte Variable der Lehrerkompetenz (Relevante Ausbildung, Berufserfahrung in Jahren, Abschluss als Lehrkraft, Lesedidaktik als relevante Komponente der Ausbildung) einen positiven Einfluss auf die Genauigkeit des Lehrerurteils hat. Diese formalen Indikatoren stellen jedoch lediglich Annäherungen an das fachliche und fachdidaktische Wissen der Lehrkräfte dar.

Auch im Rahmen des heuristischen Modells der diagnostischen Urteilsbildung ist es wünschenswert, handlungsnahe Wissensbestände zu betrachten und diese in den Mittelpunkt der Analysen des Zusammenhangs zwischen fachlichem bzw. fachdidaktischem Wissen von Lehrkräften und ihrer Urteilsgüte zu stellen. Es wird angenommen, dass sich diese Wissensbestände insbesondere auf die diagnostische Informationsverarbeitung beziehen und dabei eine fortlaufende Einordnung der Ergebnisse des diagnostischen Handelns sowie einen unmittelbaren Rückbezug der daraus gewonnenen diagnostischen Erkenntnisse auf die weitere Informationsverarbeitung ermöglichen. Dadurch sollte – bei entsprechendem Wissen – auch die Urteilsgüte positiv beeinflusst werden. Aus diesen Annahmen ergibt sich folgendes Forschungsdesiderat:

Stehen direkt erhobene bereichsspezifische Wissensaspekte von Lehrkräften in einem direkten systematischen (linearen) Zusammenhang mit der Urteilsgüte bei der Einschätzung von Schülerleistungen?

4.3 Aspekte des diagnostischen Handelns bei der Entstehung von Lehrerurteilen

Gerade im Hinblick auf die Erklärung des Zustandekommens von akkuraten oder verzerrten Urteilen ist der Blick allein auf die Urteilsgüte nur unzureichend. So können zwar über den Versuch, mit bestimmten Schülereigenschaften Varianzanteile in den Lehrkrafturteilen aufzuklären, Rückschlüsse auf die Einbeziehung dieser mitunter nicht explizit auf den jeweiligen Gegenstandsbereich der Diagnostik bezogenen Eigenschaften in die Urteilsbildung gezogen werden. In Untersuchungen zum Einfluss sachfremder Merkmale auf die Einschätzung der Schülerleistung werden diese Merkmale wie bspw. der Bildungshintergrund der Eltern (z.B. Riek & Van Ophuysen, 2016) meist durch Selbstberichte von Eltern oder von Schülerinnen und Schülern erfasst. Dies entspricht jedoch nicht notwendigerweise der von der Lehrkraft vorgenommenen Einschätzung dieses Merkmals oder entzieht sich ggf. sogar ganz der Kenntnis der einschätzenden Lehrkraft. Dies könnte den gemessenen Einfluss dieser Merkmale auf die Urteilsgüte verzerren (Van Ophuysen & Lintorf, 2016).

Auf welche Beobachtungen sich Urteile zur Schülerfähigkeit tatsächlich stützen und auf welcher Informationsbasis Urteile letztlich getroffen werden, bleibt in den meisten Untersuchungen weitestgehend offen (vgl. Aufschnaiter et al., 2015). Lediglich für Entscheidungen zu Schullaufbahnempfehlungen gibt es wenige Studien, in denen versucht wird, mit Eyetracking-Verfahren (Glock, Hörstermann, Krolak-Schwerdt & Pit-ten Cate, 2014) oder im Mouselab (Böhmer et al., 2012; Böhmer et al., 2015) zu untersuchen, welche gegebenen Informationen auf welche Weise genutzt werden, um daraus Schlüsse zu ziehen, wie basierend auf diesen Informationen Urteile gebildet werden. Neben der Annahme einer multiplikativen Kette von relevanten und verfügbaren Hinweisreizen, die von der Lehrkraft im Urteilsprozess erkannt und genutzt werden müssen, damit ein akkurates Urteil zustande kommt (Funder, 1995), erscheint es plausibel, dass auch die Art und Weise, wie die verfügbaren Hinweisreize ausgewählt und genutzt werden, die Urteilsgüte beeinflusst. Dies berührt im heuristischen Modell der diagnostischen Urteilsbildung insbesondere die Aspekte des diagnostischen Handelns und der diagnostischen Informationsverarbeitung. Basierend auf der im Modell angenommenen Verquickung dieser beiden Bereiche kann die Interpretation von beobachteten

Handlungsmustern Hinweise auf die diagnostische Informationsverarbeitung liefern. Dies ist insbesondere auch interessant vor dem Hintergrund der Frage danach, welche kognitiven Prozesse bei der Urteilsbildung wirksam werden: „There is a lack of understanding of cognitive processes of teachers guiding their judgement“ (Philipp & Leuders, 2013, S. 426). Auch aus fachdidaktischer Perspektive wird die Frage aufgeworfen, wodurch sich diagnostisches Handeln derjenigen Lehrkräfte auszeichnet, denen eine hohe Diagnosegenauigkeit attestiert wird (Schmidt, 2015). Dabei ergibt sich als drittes Forschungsdesiderat die Frage:

Welcher Zusammenhang besteht zwischen dem diagnostischen Handeln und der Güte des diagnostischen Urteils über die Schülerleistung?

Mit der im Simulierten Klassenraum möglichen Verlaufsdatenanalyse kann diese Fragestellung bearbeitet werden, indem detaillierte Einblicke in die Vorgehensweisen von Lehrkräften bei der Informationssuche über Schülerinnen und Schüler gewonnen und auf die Beschaffenheit des diagnostischen Urteils bezogen werden können.

5. Darstellung der zentralen Fragestellungen und Befunde der einzelnen Beiträge

Die einzelnen Beiträge, deren zentrale Fragestellungen und Befunde im Folgenden dargestellt und anschließend diskutiert werden, sind im Rahmen der Bamberger Forschergruppe BiKS (Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter), sowie im Rahmen der Vorbereitung eines Kooperationsprojekts der Universitäten Bamberg und Luxemburg (DIKOMPAS; Artelt et al., 2015) entstanden. Sie greifen die in Abschnitt 4 formulierten Forschungsdesiderate jeweils auf und versuchen, zur Schließung der identifizierten Lücken beizutragen.

5.1 Beitrag 1: Personality similarity between teachers and their students influences teacher judgement of student achievement[2]

Im Beitrag wird für die Kompetenzbereiche Mathematik und Lesen der Frage nachgegangen, ob globale und aufgabenspezifische Urteilsanforderungen unterschiedlich stark von nicht-leistungsbezogenen Informationen über Schülerinnen und Schüler beeinflusst werden. Exemplarisch wird in diesem Beitrag die Persönlichkeitsähnlichkeit als eine nicht-leistungsbezogene, potentiell in das Urteil einfließende und damit ggf. auch das Urteil verzerrende Information verwendet.

Zur Beantwortung der Fragestellung wurde für Schülerinnen und Schüler aus der achten Klassenstufe basierend auf einem gleichzeitig bei Lehrkräften und Schülerinnen und Schülern eingesetzten Kurzfragebogen zur Selbsteinschätzung von Persönlichkeitseigenschaften ein Index berechnet, der für jeden Schüler und jede Schülerin die Ähnlichkeit zur jeweiligen Lehrkraft bei den Antworten auf den Items des Persönlichkeitsfragebogens quantifiziert. Dieser Ähnlichkeitsindex wird als ein Indikator für Sympathie zwischen Lehrkraft und Schüler angesehen, wobei davon ausgegangen wird, dass die Sympathie umso größer ist, je ähnlicher Schüler und Lehrkraft auf die Fragen geantwortet haben (Byrne, 1997). Die Schülerinnen und Schüler wurden hinsichtlich ihrer Fähigkeiten in Mathematik und Textverstehen getestet. Außerdem gaben die Lehrkräfte

[2] Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: AR301/6-1, AR301/6-2 und AR301/6-3) im Rahmen der Bamberger Forschergruppe BiKS (Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (FOR 543)).

globale und aufgabenspezifische Urteile zu den Fähigkeiten ihrer Schülerinnen und Schüler in diesen Bereichen ab.

Für jeden der beiden Bereiche wurden jeweils Regressionsmodelle berechnet, in denen die globalen und aufgabenspezifischen Lehrkrafturteile durch die Testleistung der Schülerinnen und Schüler vorhergesagt werden sollten. Die Schülerleistung hat im Lesen und in der Mathematik einen signifikanten Einfluss auf das globale und aufgabenspezifische Lehrkrafturteil, kann aber erwartungsgemäß nur einen Teil der Varianz erklären (vgl. Südkamp et al., 2012). Darüber hinaus zeigt sich in beiden Bereichen ein geringer, jedoch signifikanter inkrementeller Einfluss der Persönlichkeitsähnlichkeit auf das globale Urteil der Lehrkraft. Für aufgabenspezifische Urteile konnte dieser Effekt hingegen nicht beobachtet werden. Globale und aufgabenspezifische Urteilsanforderungen scheinen somit unterschiedlich stark von nicht-leistungsbezogenen Informationen über Schülerinnen und Schüler beeinflusst zu werden.

Im heuristischen Modell der Urteilsbildung wurde angenommen, dass die diagnostische Aufgabenstellung den weiteren Urteilsprozess beeinflusst. Besteht die Aufgabenstellung darin, globale Urteile über eine Anzahl von Schülerinnen und Schülern abzugeben, sollte – verglichen mit der Aufgabenstellung, spezifische und auf einzelne Aufgaben bezogene Urteile abzugeben – im Prozess der Informationsverarbeitung auf andere Hinweisreize zurückgegriffen werden. So kann argumentiert werden, dass die Informationsverarbeitung durch die Vorgabe von konkreten Aufgaben, anhand derer die Schülerleistung eingeschätzt werden soll, stärker vorstrukturiert ist. Unter diesen Vorzeichen erfolgt für jeden einzuschätzenden Schüler eher ein Abruf von Erinnerungen an Situationen, die für diesen Schüler jeweils Aufschluss über sein Fähigkeitsniveau und damit verbundene typische Fehler geben. Dadurch wird ein Rückbezug auf fachliche und fachwissenschaftliche Wissensgrundlagen bei der Einschätzung von Schülerleistungen deutlich. Durch die eher vage bleibende Vorgabe, Schülerinnen und Schüler hinsichtlich deren Fähigkeit auf einem nicht näher spezifizierten Konstrukt (z.B. Arithmetik oder Textverstehen) einzuschätzen, erfolgt nur eine geringere Vorstrukturierung der Einschätzaufgabe, was eher dazu führt, dass generelle Informationen über den einzuschätzenden Schüler aus dem Gedächtnis abgerufen werden, die schnell verfügbar sind. Die Wahrnehmung von Sympathie erscheint hier – auch empirisch – als eine solche schnell abrufbare Kategorie.

Dass das Befundmuster in beiden betrachteten Bereichen in ähnlicher Form beobachtet werden konnte, spricht trotz geringer zusätzlicher Varianzaufklärung durch die Sympathie über die tatsächliche Schülerleistung hinaus dafür, dass die diagnostische

Aufgabenstellung die Informationsverarbeitung beeinflusst. Es kann daraus abgeleitet werden, dass es im diagnostischen Prozess auch darum geht, bei der diagnostischen Aufgabenstellung die richtigen Fragen zu stellen, um zu möglichst unverzerrten Urteilen zu gelangen.

5.2 Beitrag 2: Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens[3]

Im Beitrag wird der Frage nachgegangen, ob direkt erhobene bereichsspezifische Wissenskomponenten von Deutschlehrkräften mit der globalen und aufgabenspezifischen Urteilsgüte bei der Einschätzung der Schülerfähigkeit im Textverstehen zusammenhängen. Zur Beantwortung der Fragestellung wurden zunächst Wissensgrundlagen von Deutschlehrkräften erhoben, die sich auf Text- und Aufgabenmerkmale, sowie auf den adäquaten Einsatz von Lesestrategien beziehen. Diese Teilbereiche des im Rahmen des BiKS-Projekts entwickelten Tests (Matthäi, in Vorbereitung) reflektieren die Annahme, dass individuelle Textverstehensleistungen insbesondere in Abhängigkeit von den gestellten Leseanforderungen, der Beschaffenheit des Texts, sowie der Merkmale und Aktivitäten des Lesenden variieren (vgl. Tetraeder-Modell: Campione & Armbruster, 1985; Artelt, 2016).

Zudem wurde untersucht, wie akkurat diese Lehrkräfte die Fähigkeiten ihrer Schülerinnen und Schüler der Klassenstufen 8 und 9 im Textverstehen einschätzen. Anschließend wurden Korrelationen zwischen den Ergebnissen des Wissenstests und der Urteilsgüte berechnet. Dabei wurde erwartet, dass Lehrkräfte, die über ein umfangreicheres gegenstandsbezogenes Wissen in den Teilbereichen des Tests verfügen, auch akkuratere Urteile über die Schülerleistung abgeben. Dies sollte – so der angenommene Wirkmechanismus – insbesondere darüber vermittelt sein, dass Lehrkräfte aufbauend auf diesen Wissensgrundlagen auch über Wissen über Schülerkognitionen beim Textverstehen verfügen, sowie Kenntnisse und subjektive Theorien darüber haben, welche Hinweisreize verlässliche Indikatoren dafür sind, dass ein Schüler eine gegebene Aufgabenanforderung beherrscht und wie das beobachtete Schülerverhalten in Bezug auf die einzuschätzende Schülerfähigkeit interpretiert werden sollte (vgl. National Research Council, 2001). Das Wissen sollte also vermittelt über das diagnostische Handeln und die diagnostische

[3] Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: AR301/6-1, AR301/6-2 und AR301/6-3) im Rahmen der Bamberger Forschergruppe BiKS (Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (FOR 543)).

Informationsverarbeitung positiv mit der Urteilsgüte zusammenhängen. Ebenso erscheint es plausibel, dass mit umfassenderem Wissen die an die Schülerinnen und Schüler gestellten Anforderungen im Test besser eingeschätzt werden können. Insbesondere bei aufgabenspezifischen Urteilen sollte dies der Fall sein, da den Lehrkräften hier die Items des eingesetzten Tests im Sinne eines informierten Urteils vorlagen. Entgegen diesen theoretisch hergeleiteten Erwartungen konnten jedoch keine substanziellen und signifikanten korrelativen Zusammenhänge zwischen den erhobenen Wissensgrundlagen der Lehrkräfte und der globalen bzw. aufgabenspezifischen Urteilsgüte gefunden werden.
Bezogen auf den Prozess der Urteilsbildung war die zentrale Annahme des Beitrags, dass bereichsspezifisches Wissen von Lehrkräften im Rahmen der diagnostischen Informationsverarbeitung wirksam wird und damit auch Auswirkungen auf die Urteilsgüte hat. Die im Beitrag angewandte korrelative Herangehensweise lässt jedoch keine direkten Aussagen darüber zu, in welcher Weise bei den Lehrkräften durchaus vorhandenes bereichsspezifisches fachliches und fachdidaktisches Wissen in die diagnostische Informationsverarbeitung einfließen. Während des diagnostischen Prozesses können jedoch auch andere als die im Test erhobenen Wissensbestände eine Wirkung entfalten. Womöglich waren die im Beitrag operationalisierten Wissensmerkmale bezogen auf die diagnostische Urteilsbildung nicht handlungsnah genug. Eine andere Herangehensweise an das Testen des fachlichen bzw. fachdidaktischen Wissens, die sich stärker an diagnostisch relevantes Fachwissen anlehnt, wäre für weitere Forschungsvorhaben zum Einfluss des Wissens im Urteilsprozess wünschenswert. Dabei wären zunächst Untersuchungen zielführend, die z.B. unter der Verwendung der Methode des stimulated recall über die Rückbeziehung beobachteter Hinweisreize auf fachliche und fachdidaktische Wissensbestände im Urteilsprozess Aufschluss geben können.

5.3 Beitrag 3: Teacher judgment accuracy and assessment strategies in a Simulated Classroom [4]

Der Beitrag geht der Frage nach, wie angehende Lehrkräfte leistungsbezogene Informationen über Schülerinnen und Schüler suchen, um daraus ein Urteil über deren mathematische Fähigkeiten treffen zu können. Weiterhin werden die im Simulierten Klassenraum gezeigten Vorgehensweisen mit der globalen Urteilsgüte bei anschließend abgegebenen Urteilen zur Schülerleistung in Verbindung gebracht. Darüber hinaus

[4] Die Durchführung dieser Untersuchung wurde durch eine Förderung der Bamberg Graduate School of Social Sciences (BAGSS) im Rahmen der Exzellenzinitiative des Bundes und der Länder ermöglicht (GSC1024).

interessierte auch, ob die Möglichkeit zur Einbeziehung von Aufgabenmerkmalen, also die Variation der Art der zu verwendenden Hinweisreize, die Vorgehensweise und die Urteilsgüte beeinflusst. Zur Beantwortung dieser Fragestellungen wurde der Simulierte Klassenraum genutzt, da hier mit der Analyse von aufgezeichneten Verlaufsdaten detaillierte Einblicke in den Urteilsprozess gewonnen werden können. Die Beschaffenheit des diagnostischen Handelns sowie ihr unmittelbarer Zusammenhang mit der Urteilsgüte konnte direkt untersucht werden, da den Versuchspersonen keinerlei Vorinformationen über die Schülerinnen und Schüler vorlagen und Urteile somit alleine basierend auf den von ihnen abgerufenen Informationen und deren Nutzung abgegeben wurden.

Zwei Gruppen von Lehramtsstudierenden erhielten die Aufgabe, leistungsbezogene Informationen über neun simulierte Schülerinnen und Schüler zu sammeln, um anschließend die mathematischen Fähigkeiten dieser Schülerinnen und Schüler beurteilen zu können. Dazu konnten vorgegebene Aufgaben an die Schülerinnen und Schüler gerichtet werden, die von diesen je nach simuliertem Fähigkeitsniveau richtig oder falsch beantwortet wurden. Die Vorgehensweise bei der Informationssammlung konnte frei gewählt werden. Anschließend gaben die Teilnehmenden globale Urteile über die mathematische Fähigkeit der einzelnen Schülerinnen und Schüler ab. Während eine Gruppe nur annähernd gleich schwierige, mittelschwere Mathematikaufgaben nutzen konnte, griff die andere Gruppe auf Aufgaben mit einer größeren Bandbreite von Schwierigkeitsgraden zurück.

Die gezeigten Muster diagnostischen Handelns unterscheiden sich überzufällig zwischen den beiden Gruppen: Bei den Teilnehmenden, die Aufgaben ähnlichen Schwierigkeitsgrades verwenden konnten, wurde mehrheitlich eine schülerbezogene Vorgehensweise beobachtet. Dabei wurde zunächst einer einzelnen Schülerin bzw. einem einzelnen Schüler hintereinander eine Reihe von Aufgaben gestellt, bevor zum nächsten Schüler bzw. zur nächsten Schülerin übergegangen wurde und diesem bzw. dieser Aufgaben gestellt wurden. Teilnehmende hingegen, die im diagnostischen Prozess aus unterschiedlich schwierigen Aufgaben auswählen konnten, gingen mehrheitlich aufgabenbezogen vor. Dabei wurde zunächst eine bestimmte Aufgabe ausgewählt und nacheinander an mehrere Schülerinnen und Schüler gestellt, bevor die nächste Aufgabe ausgewählt und gestellt wurde. In der ersten Gruppe ging eine schülerbezogene Vorgehensweise mit deutlich akkurateren Urteilen einher als ein aufgabenbezogenes Vorgehen, während in der zweiten Gruppe eine aufgabenbezogene Vorgehensweise zusammen mit einer deutlich akkurateren Urteilsgüte auftrat.

Für beide Gruppen wurde die diagnostische Aufgabenstellung konstant gehalten (globale Fähigkeitseinschätzung), während die verfügbaren Hinweisreize experimentell variiert wurden. Diese Variation ging einher mit unterschiedlichen Mustern diagnostischen Handelns, was sich in differierenden Wahrscheinlichkeiten des Auftretens von beobachteten Vorgehensweisen in den beiden Bedingungen widerspiegelt. Unterschiedliche Vorgehensweisen waren in Abhängigkeit von der Art der zugrunde liegenden und genutzten Hinweisreize unterschiedlich zielführend für ein akkurates Urteil. Dabei bleibt jedoch offen, ob und wie Lehrkräfte unterschiedliche Arten von Informationen bei der Urteilsbildung gewichten (diagnostische Informationsverarbeitung), und ob Lehrereigenschaften oder subjektive Theorien zum Unterricht mit unterschiedlichen Vorgehensweisen einhergehen. Im Sinne des heuristischen Modells der diagnostischen Urteilsbildung kann dieser Beitrag erste Hinweise darauf liefern, dass das diagnostische Handeln von (angehenden) Lehrkräften Auswirkungen auf die Urteilsgüte bei der Einschätzung von Schülerleistungen haben kann.

Die Interpretation der Schülerantworten durch die Lehrkräfte kann unter der zweiten Bedingung stärker auf fachliche bzw. fachdidaktische Wissensbestände aufsetzen, da mit der Unterscheidungsmöglichkeit zwischen leichteren und schwierigeren Aufgaben eine realistischere Fähigkeitsabstufung zwischen den Schülerinnen und Schülern vorgenommen werden kann. Zumindest bei Lehramtsstudierenden spielte jedoch das fachliche bzw. fachdidaktische Wissen – operationalisiert über die Güte der Einschätzung von Aufgabenschwierigkeiten – keine Rolle für die gewählte Vorgehensweise und für die Urteilsgüte.

Im Beitrag stehen Muster diagnostischen Handelns im Vordergrund, die Frage nach der Informationsverarbeitung, die mit diesem Handeln einhergeht, wird jedoch nicht direkt beantwortet. Auch hier könnten z.B. mit der Methode des stimulated recall Einblicke in das Einbeziehen der abgerufenen Hinweisreize in eine vorhandene Informationsbasis gelingen. Insbesondere vor dem Hintergrund der Interpretation von Schülerantworten basierend auf fachlichem und fachdidaktischen Wissen, sowie bezogen auf die schrittweise Planung des darauf aufbauenden weiteren diagnostischen Handelns erscheint dies aufschlussreich.

6. Diskussion

Basierend auf einem heuristischen Modell der diagnostischen Urteilsbildung wurden in der vorliegenden Arbeit drei Studien beschrieben, die sich auf die zentralen Bereiche des Modells zur diagnostischen Aufgabenstellung, zur diagnostischen Informationsverarbeitung und zum diagnostischen Handeln, sowie auf die Urteilsgüte als zentralem Indikator der diagnostischen Kompetenz von Lehrkräften beziehen. Die hier zusammengefassten empirischen Beiträge sollten zu einem besseren Verständnis der Entstehung von Lehrerurteilen beitragen. Im Folgenden wird zunächst basierend auf den drei Beiträgen eine erste Einschätzung der Bewährung des heuristischen Modells vorgenommen (Abschnitt 6.1). In den Beiträgen wurden außerdem zwei unterschiedliche Herangehensweisen an die Untersuchung diagnostischer Kompetenz von Lehrkräften eingesetzt, die für die Beantwortung der jeweiligen Forschungsfragen jeweils als zielführend erachtet wurden. Das Spannungsfeld, das sich für die Interpretation von Befunden aus den beiden Herangehensweisen ergibt, wird in Abschnitt 6.2 diskutiert.

6.1 Diskussion der Bewährung des heuristischen Modells

Die drei Beiträge der vorliegenden Arbeit beziehen sich auf das oben vorgeschlagene heuristische Modell der diagnostischen Urteilsbildung und zeigen schlaglichtartig Facetten auf, die jeweils in unterschiedlichen Teilen des Urteilsprozesses wirksam werden. Damit kann nun eine erste Bewertung dieses Modells vorgenommen werden.

Für die Untersuchung der diagnostischen Aufgabenstellung wurden im ersten Beitrag die Urteilsanforderungen bei Einschätzaufgaben variiert. Dies ging auf der Ebene der Informationsverarbeitung mit der mehr oder weniger starken Einbeziehung von nicht-leistungsbezogenen Aspekten einher. Gezeigt wurde dies mithilfe von Regressionsmodellen, in denen diese Aspekte in unterschiedlichen Größenordnungen zusätzliche Varianzaufklärung lieferten. Die Untersuchung gibt damit Hinweise auf den Einfluss der diagnostischen Aufgabenstellung auf die Informationsverarbeitung. Dabei ist jedoch klar, dass eine Untersuchung, in der das Lehrkrafturteil durch die in einem Test gemessene Schülerfähigkeit und weiterer nicht-leistungsbezogener Schülermerkmale vorhergesagt wird, keine detaillierte prozessbezogene Auskunft über die tatsächliche Informationsverarbeitung geben kann (vgl. Schrader & Helmke, 1990). Die methodische Annäherung über regressionsanalytische Verfahren scheint trotz dieser Kritik dennoch quantitativen Aufschluss über den differentiellen Einbezug unterschiedlicher

Informationen bei der Entstehung von Lehrerurteilen als Antwort auf zwei verschiedene Arten der diagnostischen Aufgabenstellungen zu geben.

Im zweiten Beitrag wurden bereichsspezifische Wissensaspekte von Lehrkräften im Rahmen der diagnostischen Informationsverarbeitung betrachtet. Dabei wurde angenommen, dass die Verarbeitung von Hinweisreizen insbesondere dann zu akkurateren Urteilen führt, wenn diese auf der Basis einer breiten fachlichen Wissensbasis eingeordnet und interpretiert, sowie für die Einleitung weiterer diagnostischer Handlungsschritte verwendet werden (Bless et al., 2004). Im Beitrag wurde nicht die direkte Einbeziehung dieser Wissensbestände in die Informationsverarbeitung im Prozess beobachtet, vielmehr wurde das Wissen vor dem Hintergrund dieser Annahmen korrelativ mit der Urteilsgüte in Bezug gesetzt. Die durchwegs nicht signifikanten Zusammenhänge zwischen dem Wissen und der Urteilsgüte sollten jedoch weniger darauf bezogen diskutiert werden, dass die Lehrkräfte nicht auf ihre jeweiligen Wissensbestände zurückgreifen, sondern eher vor dem Hintergrund, dass die mit dem Test gemessenen Wissensbereiche eine geringe Handlungsrelevanz für den diagnostischen Prozess haben könnten oder in der Beurteilungsaufgabe nicht eingesetzt wurden. Ähnlich können auch Befunde aus dem dritten Beitrag interpretiert werden, wonach die über die Güte von Aufgabenschwierigkeitseinschätzungen operationalisierte Aspekte fachdidaktischen Wissens („knowledge of content and students“; Shulman, 1986) ebenfalls nicht positiv mit der Urteilsgüte zusammenhingen. Im Bereich der diagnostischen Informationsverarbeitung und insbesondere bezogen auf die Zusammenhänge mit den anderen Teilen des heuristischen Modells ist daher weitere empirische Klärung nötig, die nicht nur Wissensaspekte korrelativ mit der Urteilsgüte in Zusammenhang bringt, sondern auch den Rückgriff auf die Wissensbestände im Urteilsprozess beleuchtet.

Das diagnostische Handeln wurde im dritten Beitrag explizit aufgegriffen und mit der Urteilsgüte in Verbindung gebracht. Mit Logfile-Analysen konnte detailliert aufgezeichnet werden, welche Vorgehensweisen angehende Lehrkräfte zeigen, unmittelbar bevor sie Urteile über die mathematischen Fähigkeiten von simulierten Schülerinnen und Schülern abgeben. Je nach zu verwendendem Aufgabenmaterial gingen unterschiedliche Vorgehensweisen mit akkurateren Urteilen einher. Auch wenn mit den vorgenommenen Kategorisierungen (prototypischer) Vorgehensweisen ein Beitrag zur Betrachtung diagnostischen Handelns geleistet werden konnte, erscheint weitere Forschung insbesondere zur Differenzierung dieser Vorgehensweisen nötig, die auch der im Modell

angenommenen engen Verquickung zwischen diagnostischem Handeln und diagnostischer Informationsverarbeitung angemessen Rechnung trägt.

Die Betrachtung der pädagogischen bzw. didaktischen Handlungen, die sich aus den getroffenen Urteilen ergeben, war explizit nicht Teil der Definition diagnostischer Kompetenz in der vorliegenden Arbeit. Dennoch könnte dieser Aspekt durchaus mit der Einbeziehung des Simulierten Klassenraums in die Lehrerbildung (vgl. Brown, 1999) stärker beleuchtet werden, indem ausgehend von der Bearbeitung von Aufgaben durch simulierte Schülerinnen und Schüler didaktische Fehleranalysen und entsprechende weitere Handlungsschritte erörtert werden können. Insgesamt erscheinen Untersuchungsdesigns wünschenswert, die möglichst viele Bereiche der diagnostischen Urteilsbildung vereinen und sich auf die Urteilsgüte und das sich ggf. anschließende didaktische Handeln beziehen.

Aus den Beiträgen wird zusammenfassend deutlich, dass die vielfältigen Einflussvariablen auf Lehrkrafturteile, die in der wissenschaftlichen Fachliteratur untersucht werden, stärker vor dem Hintergrund eines Prozessmodells der diagnostischen Kompetenz betrachtet werden sollten. Die diagnostische Kompetenz wird als eine zentrale Komponente im Gefüge der Lehrerkompetenzen angesehen (z.B. Weinert et al., 1990), jedoch erscheint die Urteilsgüte von Lehrkräften bei der Einschätzung von Schülermerkmalen weiterhin als verbesserungswürdig (z.B. Südkamp et al., 2012). Insbesondere mit einer Trainings- und Förderperspektive auf die diagnostischen Fähigkeiten von Lehrkräften sollte bei den zu untersuchenden Einflussfaktoren auf die Urteilsgüte deutlich gemacht werden, an welchem Bereich des diagnostischen Prozesses diese ansetzen. Damit können die teils widersprüchlichen Ergebnisse in der Forschung zur diagnostischen Kompetenz besser eingeordnet und zueinander in Bezug gesetzt werden.

6.2 Diskussion der Herangehensweisen an die Untersuchung diagnostischer Kompetenz

Die beiden in Abschnitt 3 geschilderten Herangehensweisen an die Untersuchung der diagnostischen Kompetenz von Lehrkräften werden in den Beiträgen der vorliegenden Arbeit vor dem Hintergrund der übergeordneten Fragestellung nach der Entstehung von Lehrkrafturteilen einander ergänzend eingesetzt. Dabei ist jedoch zu beachten, dass die Beurteilungssituationen, die jeweils beobachtet werden, nicht unmittelbar kongruent sind. Im Simulierten Klassenraum läuft der komplette Urteilsprozess von der Aufgabenstellung bis zur Abgabe des entsprechenden Urteils in einem Block ab und ist auch beobachtbar.

Bei Untersuchungen, die reale Schülerinnen und Schüler und deren Lehrkräfte im Blick haben, werden hingegen die zu betrachtenden und der Urteilsabgabe unmittelbar vorausgehenden Prozesse wohl etwas anders verlaufen als im Simulierten Klassenraum. So haben die Lehrkräfte in der Situation, in der sie ihre Urteile über einzelne reale Schülerinnen und Schüler abgeben, diese in der Regel nicht unmittelbar vor sich. Die Informationen über die Schülerinnen und Schüler wurden zu diesem Zeitpunkt von der Lehrkraft bereits vorher erhoben und verarbeitet. Das diagnostische Handeln ist dabei der Informationsverarbeitung und der Urteilsabgabe zeitlich z.T. deutlich vorgelagert. Der diagnostischen Aufgabenstellung, mit der sie konfrontiert werden, folgend (z.B. globale Urteilsanforderung: „Der Schüler / die Schülerin ist im Vergleich zum Durchschnitt: sehr schwach ... sehr gut in Arithmetik"; BiKS-Forschergruppe, o.J.), ziehen die Lehrkräfte Inferenzen über ihre einzuschätzenden Schülerinnen und Schüler aus dem Gedächtnis („inferences from memory": Gigerenzer & Todd; vgl. a. Bröder & Gaissmaier, 2007). Die Verarbeitung dieser inferierten Schülerinformationen zum Zweck der Urteilsbildung gerät damit in einer solchen Untersuchung stärker in den Vordergrund, während das diagnostische Handeln eine untergeordnete Rolle spielt. Der Simulierte Klassenraum nähert sich in der kompakten Betrachtung des Urteilsprozesses hingegen den „inferences from givens" (Gigerenzer & Todd, 1999). Dies ist bei der vergleichenden Interpretation von Untersuchungsergebnissen aus den beiden Herangehensweisen zu beachten. Die Betrachtung des diagnostischen Handelns („search for information", Bröder & Gaissmaier, 2007, S. 895) erscheint gar als unmöglich, wenn in Beurteilungsaufgaben Informationen aus dem Gedächtnis abgerufen werden müssen und die zu verarbeitenden Informationen nicht unmittelbar verfügbar sind. Vielmehr müsste hier – was in den meisten Untersuchungen zur diagnostischen Kompetenz von Lehrkräften argumentativ auch getan wird (z.B. Schrader & Helmke, 1990) – von der Beschaffenheit der Urteilsgüte auf die dahinterliegenden Strategien geschlossen werden. Möchte man die Vorgehensweisen und Strategien jedoch speziell in den Fokus der Analysen stellen, spricht dies deutlich für den Einsatz des Simulierten Klassenraums als ergänzendes Instrument, um den eigentlichen Urteilsprozess genauer nachzeichnen zu können.

6.3 Ausblick

Die Befundlage zur *Akkuratheit* von Lehrereinschätzungen ist in der wissenschaftlichen Literatur bereits relativ umfassend und systematisch aufgearbeitet und dokumentiert (Hoge & Coladarci, 1989; Südkamp et al., 2012; Machts et al., 2016). Diese häufig nur deskriptiv bleibenden Befunde (Wie akkurat sind Lehrerurteile über Schülerleistungen?), die zudem große interindividuelle Unterschiede zwischen Lehrkräften aufdecken, sind unter der Perspektive der *Entstehung* von Lehrerurteilen jedoch weiterhin erklärungsbedürftig. Die drei in der vorliegenden Dissertationsschrift beschriebenen Beiträge zielten darauf ab, einzelne Aspekte der Entstehung von Lehrerurteilen über Schülerleistungen in den Blick zu nehmen, um künftig besser erklären zu können, wann und unter welchen Umständen akkurate Lehrerurteile beobachtet werden können (Artelt & Rausch, 2014). Dazu bedarf es noch weiterer Forschungsaktivitäten um umfassenderes, empirisch gesichertes Wissen zu generieren. Die hier beschriebenen Arbeiten sollten dazu beitragen, eine differenzierte Betrachtung einzelner Aspekte im diagnostischen *Prozess* bei der Einschätzung von Schülerleistungen voranzutreiben. Sie sollen dazu ermuntern, insbesondere die Merkmale der Lehrkraft, des Tests, der Schülerinnen und Schüler und des Urteils (Südkamp et al., 2011) nicht nur auf die Urteilsgüte als Indikator für diagnostische Kompetenz zu beziehen. Vielmehr sollte deren Verbindung insbesondere mit Prozessmerkmalen als Teil einer weiten Definition diagnostischer Kompetenz, die neben der Urteilsgüte auch Prozessindikatoren beinhaltet stärker beleuchtet werden. Von einer größeren Handlungsnähe solcher prozessbezogenen Untersuchungen können auch (zukünftige) Lehrkräfte profitieren, indem damit Wissen und Bewusstsein über die Beschaffenheit des eigenen diagnostischen Handelns und der diagnostischen Informationsverarbeitung geschaffen werden kann. Aufbauend auf solchen bewussteren Urteilsprozessen können fundierte und informierte pädagogische und didaktische Entscheidungen getroffen werden, die den Schülerinnen und Schülern an ihre jeweils gezeigten Fähigkeiten angepasste Lerngelegenheiten eröffnen können.

Literaturangaben

Abs, H. J. (2007). Überlegungen zur Modellierung diagnostischer Kompetenz bei Lehrerinnen und Lehrern. In M. Lüders & J. Wissinger (Hrsg.), *Forschung zur Lehrerbildung* (S. 63-84). Münster: Waxmann.

Alvidrez, J. & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*, 731-746.

Anders, Y., Kunter, M., Brunner, M., Krauss, S. & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht, 57*, 175-193.

Artelt, C. (2016). Teacher judgments and their role in the educational process. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*.

Artelt, C. & Gräsel, C. (2009). Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*, 157-160.

Artelt, C., Krolak-Schwerdt, S., Hörstermann, T. & Rausch, T. (2015). *Diagnostische Kompetenz von Lehrkräften: Wie interagieren Aufgaben- und Schülermerkmale im Prozess der Leistungsbeurteilung?* : unveröffentlichter Forschungsantrag.

Artelt, C. & Rausch, T. (2014). Accuracy of teacher judgments. When and for what reasons? In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Hrsg.), *Teachers' professional development: Assessment, training, and learning* (S. 27-43). Rotterdam: Sense Publishers.

Aufschnaiter, C. v., Cappell, J., Dübbelde, G., Ennemoser, M., Mayer, J., Stiensmeier-Pelster, J.et al. (2015). Diagnostische Kompetenz. Theoretische Überlegungen zu einem zentralen Konstrukt der Lehrerbildung. *Zeitschrift für Pädagogik, 61*, 738-758.

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*, 177–187.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*, 469-520.

Becker, D. & Birkelbach, K. (2013). Lehrer als Gatekeeper? Eine theoriegeleitete Annäherung an Determinanten und Folgen prognostischer Lehrerurteile. In R. Becker & A. Schulze (Hrsg.), *Bildungskontexte. Strukturelle Voraussetzungen und Ursachen ungleicher Bildungschancen* (S. 207-237). Wiesbaden: Springer VS.

Begeny, J. C., Krouse, H. E., Brown, K. G. & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review, 40*, 23-38.

Behrmann, L. & Souvignier, E. (2013). The relation between teachers' diagnostic sensitivity, their instructional activities, and their students' achievement gains in reading. *Zeitschrift für Pädagogische Psychologie, 27*, 283-293.

BiKS Forschergruppe. (o.J.). *Codebuch zum Kindbezogenen Einschätzbogen Welle 7*. Verfügbar unter: https://www.iqb.hu-berlin.de/fdz/studies/BiKS_8-14/Kindbezogener_Ei_6.pdf [11.03.2016]

Bless, H., Fiedler, K. & Strack, F. (2004). *Social cognition: how individuals construct social reality*. Hove: Psychology Press.

Böhmer, I., Gräsel, C., Hörstermann, T. & Krolak-Schwerdt, S. (2012). Die Informationssuche bei der Erstellung der Übergangsempfehlung - Die Rolle von Fallkonsistenz und Expertise. *Unterrichtswissenschaft, 40*, 140-155.

Böhmer, I., Hörstermann, T., Gräsel, C., Krolak-Schwerdt, S. & Glock, S. (2015). Eine Analyse der Informationssuche bei der Erstellung der Übergangsempfehlung: Welcher Urteilsregel folgen Lehrkräfte? *Journal for Educational Research Online, 7*, 59-81.

Bröder, A. & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review, 14*, 895-900.

Brown, A. H. (1999). Simulated classrooms and artificial students: The potential effects of new technologies on teacher education. *Journal of Research on Computing in Education, 32*, 307-318.

Brunner, M., Anders, Y., Hachfeld, A. & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum & U. Klusmann (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 215-234). Münster: Waxmann.

Byrne, D. (1997). An overview (and underview) of research and theory within the attraction paradigm. *Journal of Social and Personal Relationships, 14*, 417-431.

Campione, J. C. & Armbruster, B. B. (1985). Acquiring information from texts: An analysis of four approaches. In J. W. Segal, S. F. Chipman & R. Glaser (Hrsg.), *Thinking and learning skills. Volume 1: Relating instruction to research* (S. 317-359). Hillsdale, NJ: Erlbaum.

Clement, U. (2012). Vertrauen in Lehrkräfte und Neue Verwaltungssteuerung. In H. Möller (Hrsg.), *Vertrauen in Organisationen. Riskante Vorleistung oder hoffnungsvolle Erwartung?* (S. 143-167). Wiesbaden: Springer VS.

Deutscher Bildungsrat. (1970). *Empfehlungen der Bildungskommission. Strukturplan für das Bildungswesen*. Stuttgart: Klett.

Dipboye, R. & Gaugler, B. B. (1993). Cognitive and behavioral processes in the selection interview. In N. Schmitt & W. C. Borman (Hrsg.), *Personnel selection in organizations* (S. 135-170). San Francisco: Jossey-Bass.

Dünnebier, K., Gräsel, C. & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung. Eine experimentelle Studie zu Ankereffekten. *Zeitschrift für Pädagogische Psychologie, 23*, 187-195.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C. & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*, 247-265.

Edelenbos, P. & Kubanek-German, A. (2004). Teacher assessment: the concept of 'diagnostic competence'. *Language Testing, 21*(3), 259-283.

Fiedler, K., Freytag, P. & Unkelbach, C. (2007). Pseudocontingencies in a simulated classroom. *Journal of Personality and Social Psychology, 92*, 665-677.

Fiedler, K., Walther, E., Freytag, P. & Plessner, H. (2002). Judgment biases in a simulated classroom - A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes, 88*, 527-561.

Frey, A. & Jung, C. (2011). Kompetenzmodelle und Standards in Lehrerbildung und Lehrerberuf. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 540-572). Münster: Waxmann.

Frischknecht, M.-C., Reimann, G., Gut, J., Ledermann, T. & Grob, A. (2014). Wie genau können Mütter die Mathematik- und Sprachleistungen ihrer Kinder einschätzen? Ein Vergleich zwischen Müttereinschätzung und Testleistungen bei Kindern im Alter von 6-10 Jahren. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46*, 67-78.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670.

Funder, D. C. (1999). *Personality judgment. A realistic approach to person perception*. San Diego: Academic Press.

Gast, A., Herppich, S., Wittwer, J. & Nückles, M. (2014). Lernprozessdiagnose im Dialog - Erkennen und Bewerten diagnostisch relevanter Interaktionsstrategien im Video, *2. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF)*. Frankfurt am Main.

Gigerenzer, G. & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd & the ABC Research Group (Hrsg.), *Simple heuristics that make us smart* (S. 3-34). New York: Oxford University Press.

Glock, S., Hörstermann, T., Krolak-Schwerdt, S. & Pit-ten Cate, I. (2014). Noten oder sozialer Hintergrund? Der erste Blick beeinflusst das Gedächtnis für Schülerinformationen und die Genauigkeit des Urteils, *2. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF)*. Frankfurt am Main.

Glock, S. & Krolak-Schwerdt, S. (2013). Does nationality matter? The impact of stereotypical expectations on student teachers' judgments. *Social Psychology of Education, 16*, 111-127.

Glock, S., Krolak-Schwerdt, S., Klapproth, F. & Böhmer, M. (2013). Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Social Psychology of Education, 16*, 555-573.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*, 245-270.

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

Heinich, R., Molenda, M. & Russell, J. D. (1993). *Instructional media and the new technologies of instruction* (4th ed.). New York: Macmillan.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze-Velber: 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.), *Schulmanagement und Schulentwicklung* (S. 119-144). Hohengehren: Schneider-Verlag.

Heritage, M. (2013). Gathering evidence of student understanding. In J. H. McMillan (Hrsg.), *SAGE Handbook of research on classroom assessment* (S. 179-195). Thousand Oaks: SAGE Publications.

Hoge, R. D. & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*, 297-313.

Hopkins, K. D., George, C. A. & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*, 177-182.

Hoth, J., Döhrmann, M., Kaiser, G., Busse, A., König, J. & Blömeke, S. (2016). Diagnostic competence of primary school mathematics teachers during classroom situations. *ZDM, The International Journal of Mathematics Education, online first.*

Jäger-Flor, D. & Jäger, R. S. (2008). *Bildungsbarometer zum Thema "Förderung im Bildungssystem". Ergebnisse, Bewertungen und Perspektiven.* Verfügbar unter: http://www.zepf.eu/fileadmin/user_upload/documents/Bildungsbarometer/Bildungsbarometer_2008_2.pdf [23.2.2016]

Johansson, S., Strietholt, R., Rosén, M. & Myrberg, E. (2014). Valid inferences of teachers' judgements of pupils' reading literacy: does formal teacher competence matter? *School Effectiveness and School Improvement, 25*, 394-407.

Johnson, E. J., Payne, J. W., Schkade, D. A. & Bettman, J. R. (1989). *Monitoring information processing and decisions: The Mouselab System*Verfügbar unter: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA205963 [02.11.2016

Kahneman, D. (2011). *Schnelles Denken, langsames Denken.* München: Siedler.

Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A. & Möller, J. (2012). Zum Zusammenhang von Intelligenz und Urteilsgenauigkeit bei der Beurteilung von Schülerleistungen im Simulierten Klassenraum. *Zeitschrift für Pädagogische Psychologie, 26*, 251-261.

Kaiser, J. & Möller, J. (2016). Diagnostische Kompetenz von Lehramtsstudierenden. In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals. Interdisziplinäre Betrachtungen, Befunde und Perspektiven* (S. 55-74). Wiesbaden: Springer.

Kaiser, J., Retelsdorf, J., Südkamp, A. & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73-84.

Karing, C. (2011). *Diagnostische Kompetenz von Lehrkräften in der Sekundarstufe I.* Kumulative Dissertation, Universität Bamberg.

Karing, C., Dörfler, T. & Artelt, C. (2013). How accurate are teacher and parent judgements of lower secundary school children's test anxiety? *Educational Psychology, 35*, 909-925.

Karing, C., Matthäi, J. & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I - Eine Frage der Spezifität? *Zeitschrift für Pädagogische Psychologie, 25*, 159-172.

Karing, C., Pfost, M. & Artelt, C. (2011). Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen? *Journal for Educational Research Online, 3*, 119-147.

Klauer, K. J. (1973/2005). *Das Experiment in der pädagogisch-psychologischen Forschung: eine Einführung.* Münster: Waxmann.

Klieme, E., Avenarius, H., Blum, W., Doebrich, P., Gruber, H., Prenzel, M.et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Bonn: Bundesministerium für Bildung und Forschung.

Klug, J., Bruder, S., Kelava, A., Spiel, C. & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*, 38-46.

Klug, J., Bruder, S. & Schmitz, B. (2015). Which variables predict teachers diagnostic competence when diagnosing students' learning behavior at different stages of a teachers' career? *Teachers and Teaching: Theory and Practice, online first.*

Krolak-Schwerdt, S., Böhmer, M. & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess. Der Lehrer als 'flexibler Denker'. *Zeitschrift für Pädagogische Psychologie, 23*, 175-186.

Kultusministerkonferenz (KMK). (2014). Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004 i.d.F. vom 12.06.2014. Bonn: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Lipowsky, F., Rakoczy, K., Pauli, C., Reusser, K. & Klieme, E. (2007). Gleicher Unterricht - gleiche Chancen für alle? Die Verteilung von Schülerbeiträgen im Klassenunterricht. *Unterrichtswissenschaft, 35*, 125-147.

Lorenz, C. (2011). *Diagnostische Kompetenz von Grundschullehrkräften. Strukturelle Aspekte und Bedingungen.* Bamberg: University of Bamberg Press.

Lorenz, C. & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie, 23*, 211-222.

Machts, N., Kaiser, J., Schmidt, F. T. C. & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review, 19*, 85-103.

Martignon, L. & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision, 52*, 29-71.

Matthäi, J. (in Vorbereitung). *Wissensgrundlagen diagnostischer Kompetenz im Bereich des Textverstehens.* Dissertation, Universität Bamberg, Bamberg.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Oser, F., Heinzer, S. & Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten: Chancen und grenzen des advokatorischen Ansatzes. *Unterrichtswissenschaft, 38*, 5-28.

Philipp, K. & Leuders, T. (2013). Diagnostic competences of mathematics teachers - what kind of knowledge do teachers use? In A. M. Lindmeier & A. Heinze (Hrsg.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education*. Kiel: PME.

Pohlmann, B., Möller, J. & Streblow, L. (2004). Zur Fremdeinschätzung von Schülerselbstkonzepten durch Lehrer und Mitschüler. *Zeitschrift für Pädagogische Psychologie, 18*, 157-169.

Randel, B. & Clark, T. (2013). Measuring classroom assessment practices. In J. H. McMillan (Hrsg.), *SAGE Handbook of Research on Classroom Assessment* (S. 145-163). Los Angeles: Sage.

Ready, D. D. & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*, 335-360.

Riek, K. & Van Ophuysen, S. (2016). Nicht immer zählt nur die Leistung - schulformabhängige Prädiktoren der Übergangsempfehlung. In K. Liebers, B. Landwehr, S. Reinhold, S. Riegler & R. Schmidt (Hrsg.), *Facetten grundschulpädagogischer und -didaktischer Forschung* (S. 13-18). Wiesbaden: VS Verlag für Sozialwissenschaften.

Ritts, V., Patterson, M. L. & Tubbs, M. E. (1992). Expectations, impressions, and judgments of physically attractive students - a review. *Review of Educational Research, 62*, 413-426.

Schmidt, F. (2015). Den diagnostischen Blick schärfen - Vorstellungen und Orientierungen von Deutschlehrerinnen und Deutschlehrern zur Diagnose von Lesekompetenz. In C. Bräuer & D. Wieser (Hrsg.), *Lehrende im Blick. Empirische Lehrerforschung in der Deutschdidaktik* (S. 89-109). Wiesbaden: Springer VS.

Schrader, F.-W. (2010). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 102-108). Göttingen: Hogrefe.

Schrader, F.-W. (2011). Lehrer als Diagnostiker. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 683-698). Münster: Waxmann.

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerbildung, 31*, 154-165.

Schrader, F.-W. & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik, 1*, 27-52.

Schrader, F.-W. & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 22*, 312-324.

Schrader, F.-W. & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.), *Leistungsmessung in Schulen* (S. 45-58). Weinheim: Beltz.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1-22.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 19*, 85-95.

Spinath, B. (2012). Beiträge der Pädagogischen Psychologie zur Professionalisierung von Lehrerinnen und Lehrern: Diskussion zum Themenschwerpunkt. *Zeitschrift für Pädagogische Psychologie, 26*, 307-312.

Südkamp, A. (2010). *Diagnostische Kompetenz: Zur Genauigkeit der Beurteilung von Schülerleistungen durch Lehrkräfte*. Unveröffentlichte Dissertation, Christian-Albrechts-Universität, Kiel.

Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743-762.

Südkamp, A., Kaiser, J. & Möller, J. (2014). Teachers' judgments of students' academic achievement: results from field and experimental studies. In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Hrsg.), *Teachers' professional development: assessment, training, and learning* (S. 5-26). Rotterdam: Sense Publishers.

Südkamp, A. & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum: direkte und indirekte Einschätzungen von Schülerleistungen. *Zeitschrift für Pädagogische Psychologie, 23*, 161-174.

Südkamp, A., Möller, J. & Pohlmann, B. (2008). Der Simulierte Klassenraum. Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 22*, 261-276.

Trittel, M., Gerich, M. & Schmitz, B. (2014). Training prospective teachers in educational diagnostics. In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Hrsg.), *Teachers' professional development: Assessment, training, and learning* (S. 63-78). Rotterdam: Sense Publishers.

Urhahne, D. (2015). Teacher behavior as a mediator of the relationship between teacher judgment and students' motivation and emotion. *Teaching and Teacher Education, 45*, 73-82.

Van Ophuysen, S. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 38*, 154-161.

Van Ophuysen, S. & Lintorf, K. (2014). Unterschiede in der diagnostischen Praxis - Eine Frage der pädagogischen Zielsetzung? *Empirische Pädagogik, 28*, 211-228.

Van Ophuysen, S. & Lintorf, K. (2016). Leistung ist nicht alles – Empfehlungskriterien bei sicheren vs. unsicheren Übergangsempfehlungen am Ende der Grundschulzeit,*Deutsche Gesellschaft für Psychologie (DPGs)*. Leipzig.

Weinert, F. E. (2000). Lehren und Lernen für die Zukunft - Ansprüche an das Lernen in der Schule. *Nachrichten der Gesellschaft zur Förderung Pädagogischer Forschung, 2*, 4-23.

Weinert, F. E., Schrader, F.-W. & Helmke, A. (1990). Educational expertise: Closing the gap between educational research and classroom practice. *School Psychology International, 11*, 163-180.

Wylie, E. C., Lyon, C. J. (2015). The fidelity of formative assessment implementation: issues of breadth and quality. *Assessment in Education: Principles, Policy & Practice, 22*, 140-160.

Anhang

Verzeichnis der Originalbeiträge

1. Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (2016). Personality similarity between teachers and their students influences teacher judgement of student achievement. *Educational Psychology, 36*, 863-878. doi: 10.1080/01443410.2014.998629

2. Rausch, T., Matthäi, J., & Artelt, C. (2015). Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*, 147-158. doi: 10.1026/0049-8637/a000124

3. Rausch, T., & Artelt, C. (eingereicht). Teacher judgment accuracy and assessment strategies in a Simulated Classroom.

Personality similarity between teachers and their students influences teacher judgement of student achievement[1]

Tobias Rausch[a] , Constance Karing[b], Tobias Dörfler[c] and Cordula Artelt[a]
[a]Department of Educational Research, University of Bamberg, Bamberg, Germany;
[b]Department of Psychology, University Jena, Jena, Germany;
[c] Institute of Psychology, University of Education Heidelberg, Heidelberg, Germany

Abstract. This study examined personality similarity between teachers and their students and its impact on teacher judgement of student achievement in the domains of reading comprehension and mathematics. Personality similarity was quantified through intraclass correlations between personality characteristics of 409 dyads of German teachers and their students. This similarity index was combined with teachers' global and task-specific judgements of student achievement. Personality similarity has a significant effect on global judgement in both domains under study. Students who are similar to their teacher are judged more positively than students who are dissimilar, even when students' test performance is controlled. This effect could not be verified for task-specific judgements. Results indicate that impact of potential sympathy bias in social judgements differs between different types of judgement. That is, global judgements are more likely to be biased than more specific judgements. Theoretical and educational relevance of the findings are discussed.

Keywords: teacher judgement accuracy; personality similarity; judgement bias

Teacher judgements about their students' achievement play an important role in everyday school context. Besides the aspects of selection and allocation, teacher judgements are bases for immediate educational decisions, such as adapting teaching to the class's, or individual students' needs. Feedback for students and their parents, as well as information for other teachers and educational decision-makers, is also based on (informal) teacher judgement. We also know that human social perception is usually prone to systematic errors, potentially leading to judgement bias (e.g. Kruglanski & Ajzen, 1983). If a person evaluates "two groups as differing on some criterion more or less than they really do differ" (Jussim, Eccles, & Madon, 1996, p. 329), and if the difference in the evaluation can be systematically belayed on one or more particular variables that are irrelevant to the criterion, the judgement is considered to be biased. This is not only a problem for laypersons; expert judgements can be biased, too, although it can be assumed that teachers with high diagnostic expertise are aware of judgement tendencies and biases (Helmke, 2007). This normative aim is challenged by several studies finding different biases in teacher

[1] This is an Accepted Manuscript of an article published by Taylor & Francis Group in Educational Psychology on 06/01/2015, available online: http://www.tandfonline.com/10.1080/01443410.2014.998629.

judgements (e.g. Hany, 1997; Wang, Treat, & Brownell, 2008). The present study takes up the particular discussion about the perception of sympathy or attraction as a biasing factor in teacher judgement (Hadley, 1954; Itskowitz, Navon, & Strauss, 1988). Whereas former studies use teachers' direct perceptions of their students' likeableness, it can also be assumed that attraction can, among others, be derived by similarity between two persons in terms of different characteristics (Byrne, 1997; Montoya & Horton, 2013). The present study focuses on actual similarity between students and teachers in terms of personality traits rather than on perceived sympathy. The central aim of our study is to investigate, if personality similarity between teachers and students can account for a biased teacher judgement in two types of judgements in the domains of reading comprehension and mathematics.

Theoretical framework

Teachers' ability to accurately judge student performance is regarded as an integral part of teachers' professional competence (Baumert & Kunter, 2006; Brunner, Anders, Hachfeld, & Krauss, 2011; Ready & Wright, 2011). Thus, it is demanded that teachers assess their students' achievement and ability in an accurate and unbiased way.

A review of the literature about teacher judgement accuracy in general and two meta-analyses in particular (Hoge & Coladarci, 1989; Südkamp, Kaiser, & Möller, 2012), reveals that teacher judgements do not always correspond highly with students' achievement in standardised assessment tests and that more than two-thirds of the variance in teacher judgements on student performance cannot be explained by student performance alone (Südkamp et al., 2012). Furthermore, there are huge inter-individual differences between teachers with respect to their judgement accuracy. These findings lead to the assumption that teacher judgement accuracy is affected by certain moderator variables. Based on empirical findings and theoretical considerations, they can be grouped into four categories (Südkamp et al., 2012):

Teacher characteristics such as, for example, job experience (Impara & Plake, 1998), intelligence (Kaiser, Helm, Retelsdorf, Südkamp, & Möller, 2012) and beliefs (Shavelson & Stern, 1981) are thought to influence judgement accuracy, but overall results are heterogeneous in this field and do not tend to find strong relations. Studies about the influence of *student characteristics* on teachers' judgement accuracy consider for example students' intelligence, motivational or sociodemographic variables (Schrader & Helmke, 1990), their ability level (Carr & Kurtz-Costes, 1994), gender (Karing, Matthäi, & Artelt,

2011), behaviour (Bennett, Gottesman, Rock, & Cerullo, 1993), engagement (Kaiser, Retelsdorf, Südkamp, & Möller, 2013) or disability status (Hurwitz, Elliott, & Braden, 2007). Combinations of teacher and student characteristics (e.g. same gender or ethnicity, similarities in socioeconomic background or in personality traits) may also influence judgement accuracy. Different *requirements of judgement tasks* can affect accuracy, depending on how teacher ratings are assessed. Teachers either have to estimate their students' ability on a very specific (e.g. estimation of the number of solved tasks) or on a more global level (e.g. judgement of students' ability in a certain domain). Judgement accuracy can also be influenced by *test characteristics*, such as features of the subject area, specific task sets, task difficulty or the congruence between teachers' judgement task and students' achievement test (cf. Demaray & Elliott, 1998).

Similarity between student and teacher and the impact on judgement accuracy

Classrooms are places where interactions between students and teachers permanently occur. Given the diversity of student characteristics in a classroom, it is quite obvious that teachers do not interact with each student in the same way. Teachers might act differently towards students, depending on student characteristics, such as e.g. gender, race, socioeconomic status (see Englehart, 2009) or physical attractiveness (Ritts, Patterson, & Tubbs, 1992). Alexander, Entwisle, and Thompson (1987) also shed light on this issue by showing that teachers acted differently towards students, dependent on the social distance between them: Low similarity lead to lower expectations and teacher disaffection. Low distance in socioeconomic status might be only an indicator variable, which covers issues of similarity in behaviour or habitus. It thus seems also likely that teacher–student interaction is influenced by similarity or dissimilarity between teachers and their respective students, in terms of attitudes, or personality traits.

It might be an intuitive hypothesis that a higher degree of similarity between teacher and student with respect to personality attributes leads to a more accurate estimation of the students' achievement, because it is easier to take over another persons' perspective if that person is more similar to oneself when compared to a less similar person. However, several studies show that the degree of personality similarity between two persons is rather connected with attraction or sympathy (for an overview see Montoya, Horton, & Kirchner, 2008), which leads to a more positive estimation of a person. There are two competing explanations for the phenomenon that individuals are attracted to other persons who are similar to them: the reinforcement model (e.g. Byrne, 1971) and the information processing

perspective (e.g. Ajzen, 1974). The latter perspective assumes that because one's own attributes are used as an anchor for evaluating another person's attributes, more similar persons are perceived as likable and therefore are more likely to be evaluated in a more positive manner. The reinforcement model argues as follows: The higher the degree of similarity the more likely it is, that one will obtain consensual validation for one's own personality traits from the other person. This is in return experienced as rewarding. It is more likely that the other person is perceived as likeable and thus is evaluated more positively. In an early work of Hadley (1954), it was shown that students who are estimated as likable by teachers receive better grades when compared to students considered to be less likable. Itskowitz et al. (1988) found an influence of perceived sympathy on teacher judgements about students' self-perceptions of their strengths and weaknesses in homework, self-evaluation and learning ability. Teachers tend to judge children to whom they feel attached more positively than children to whom they feel indifferent or children they reject. Although similarity in terms of personality traits cannot be seen as one single predictor for sympathy, we consider actual personality similarity as a proxy for it.

Judgement characteristics and the impact on judgement accuracy

Teacher judgements are either direct or indirect (Hoge & Coladarci, 1989). For direct judgements, teachers are aware of the characteristics and tasks of the underlying achievement test and they either have to estimate the number of a student's correctly solved tasks in a test (Coladarci, 1986), or they have to estimate for a number of items in a test, if a particular student solved these items or not (e.g. Karing et al., 2011). We will refer to that type of judgement as task-specific judgements. In school context, task-specific judgements are especially relevant for teachers' (formal) assessment of student achievement. For indirect judgements, teachers do not know about the characteristics and tasks of the underlying achievement test and they have to rate students' overall achievement in a certain domain on a rating scale (e.g. Hopkins, George, & Williams, 1985). We will refer to that type of judgement as global judgements. In school context, global judgements are more relevant for teachers' (informal) impression formation about students. Both types of judgement can be expected to be accurate and unbiased. Results concerning the accuracy of different types of judgement are heterogeneous. Karing et al. (2011) found secondary school teachers' global judgements of students' reading comprehension to be more accurate than task-specific judgements in the same domain. Demaray and Elliott (1998) and Feinberg and Shapiro (2009) found that primary school teachers' task-specific judgements of students' reading skills are more accurate than global judgements about students' general abilities

and their motivation. On a more aggregated level, however, meta-analyses (Hoge & Coladarci, 1989; Südkamp et al., 2012) found evidence that task-specific judgements yield more accurate judgements than global judgements. A teacher's global judgement is dependent on his or her own understanding of the domain in question, which is not necessarily congruent with the construct measured in the underlying student test. For task-specific judgements, however, teachers can refer to more given information about specific tasks. This also provides more structure to teachers for their judgement. Dipboye and Gaugler (1993) showed that unstructured judgement processes are more vulnerable to judgement bias than more structured processes. Thus, we expect that teacher judgement has to deal with different extents of bias concerning different types of judgement (see also Artelt & Rausch, 2014).

The two types of judgement also differ in the amount of teachers' knowledge that has to be called upon in order to deliver an accurate judgement. For a task-specific judgement, teachers have to estimate the difficulty of the given tasks and – based on that – estimate whether or not a student will answer each of the tasks correctly. For that, teachers need both, knowledge about characteristics that make a task difficult, as well as knowledge about the student's traits relevant for solving the particular task (e.g. previous knowledge, motivation, strategies). Teacher's global judgement, on the other hand, is probably less knowledge-driven and is rather influenced not only by the teacher's perception of student's performance, but also by other aspects irrelevant to the student's actual achievement. When evaluating student achievement in a certain domain on a rating scale, teachers might not have had access to information about the items or properties of the test (e.g. theoretical framework, test development). Principles underlying the test and the criteria teachers apply to their judgements may differ. This means that teachers might also draw upon more distal aspects for their judgements that are not directly related to students' achievement in a test. This can be, for example, the student's overall performance in school, the student's grade in the last test or perceptions of sympathy.

Expectations and hypotheses

We assume that the extent of similarity between teachers' and students' personality traits is a source for bias in teacher judgements. We further expect that global judgements are more likely to be affected by personality similarity between teachers and students than task-specific judgements are. Our hypotheses are as follows:

(1) Personality similarity between student and teacher influences teachers' global judgements of students' performance in the domains of reading comprehension and mathematics. Students with a higher degree of similarity to their teachers are judged more positively by their teachers than those with a lower degree of similarity.

(2) Personality similarity between student and teacher influences teachers' task-specific judgements of students' performance in the domains of reading comprehension and mathematics, if at all, to a lower degree.

Methods

Participants

Data collection took place in the context of the BiKS-study (Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor-und Grundschulalter, FOR 543 [educational processes, competence development and selection decisions in pre-school and elementary school age]). In our analyses, we used data from teachers (N = 94) and their students (N = 293) from 72 classrooms at the end of Grade 8 in German secondary schools (lower, middle and higher academic track). As language arts and mathematics cover a wide range of students' general academic achievement and because they are also important for other school subjects, we chose to focus on these two subjects. Combination of teacher and respective student data yielded 168 teacher–student dyads in German language classes and 241 teacher–student dyads in mathematics classes. Due to the design of the study, teachers did not estimate abilities for all students in their class, which led to a mean of 4.2 student ratings per teacher in the German language class sample and a mean of 4.5 student ratings per teacher in the mathematics class sample. Table 1 shows descriptive data for both the student and the teacher sample.

Table 1. Descriptive data for student and teacher sample.

	German language class sample					Mathematics class sample				
	N	Male	Female	Age		*N*	Male	Female	Age	
				M	SD				*M*	SD
Students	168	64	104	14.39	0.40	241	101	140	14.41	0.47
Teachers	40	13	27	39.95	11.79	54	32	22	47.34	13.43

Measures

Student achievement

The reading comprehension measures used in this study were constructed in close communication with reading experts from the German PISA consortium. Students' reading comprehension was assessed with a test consisting of three texts (one fictional and two non-fictional) and a total of 25 multiple-choice items. The students had to read the texts, search for relevant information and draw inferences from the text in order to answer the questions. Regarding criterion validity, a correlation of $r = .29$ between students' raw scores in the reading comprehension test and the overall grade the students received for the German language class was obtained[1]. Because the formation of grades in German language arts bases on manifold facets of the subject, such as students' ability in reading, writing, interpretation of prose and poetry, and the like, the quite low correlation has a lot to commend that the test measures a more specific ability than the grade in the subject. The internal consistency of the test was acceptable (Cronbach's $\alpha = .71$).

To assess the students' mathematical competence, we used a total of 29 items in a newly developed test covering arithmetic, written math problems and geometry. Correlation between raw scores in the mathematics test and the grade the students received for the mathematics class was $r = .55$[2]. The internal consistency of the mathematics test was sufficient (Cronbach's $\alpha = .80$). Descriptive data for student achievement are shown in Table 2.

Teacher judgements

Teacher judgement about their students' achievement was assessed using two different types of judgement. For global judgements about students' competences in reading comprehension and mathematics, teachers were asked to judge their students when compared to an average eighth-grade student along a 5-point rating scale with semantic anchors at 1 ('very weak'), 3 ('average') and 5 ('very good'). Teachers' global judgement on

[2] In German school system, grades range from 1 (outstanding performance) to 6 (insufficient performance). To facilitate interpretation, grades for mathematics and for German were recoded so that higher numbers represent better performance.

students' reading comprehension was based on a single item. Global judgement of students' mathematical competence was assessed with three items, each with a 5-point rating scale about students' abilities in arithmetic, written math problems and geometry. For global judgement in mathematics we used the mean of the three items. Internal consistency for this scale was Cronbach's $\alpha = .93$. When teachers delivered their global judgement, they were not informed about the items in the tests, and thus did neither know about the construction principles of the test nor about the operationalisation of reading comprehension and mathematical competence in the test.

To assess teachers' task-specific judgements about student achievement, seven questions about one of the texts from the reading comprehension test and seven tasks from the mathematics test, respectively, were selected and presented to the teachers. For each of their students in the sample they then had to estimate whether each of the students would solve each of the tasks or not. For our analyses these seven single judgements were summed up to one variable ranging from 0 (teacher expects that a student did not solve one single task) to 7 (teacher expects that a student solved all seven tasks). Descriptive data about teacher judgements can be found in Table 2.

Personality measures

Teachers' and students' personality traits were assessed by a German short version of the Big Five Inventory (BFI-S; Gerlitz & Schupp, 2005). The instrument consists of 15 items, to which teachers and students responded on a rating scale ranging from 1 ('do not at all agree') to 7 ('fully agree')[3]. Each of the 15 items is considered to represent a single measure of a person characteristic and so, we decided to use all 15 items as a basis for the computation of the personality similarity index.

Personality similarity

To investigate the impact of personality similarity in teacher–student dyads on teacher judgements, it is necessary to choose an appropriate measure of similarity. First insight in this topic can be gained with an article by Cronbach and Gleser (1953) about assessing similarity between profiles. According to the authors, there are three factors influencing the size of the similarity measurement (cf. Kenny, Kashy, & Cook, 2006): level, spread and shape of profiles. The most important aspect is the similarity of the ups and downs of two

[3] Examples for the items: 'I am an inventive person'; 'I am a person who works thoroughly'; 'I am a communicative and talkative person'; 'I am a friendly person who treats others with respect'; 'I am a person who worries a lot'.

profiles (shape). It is also a crucial aspect to comprise differences in average values of the profile across all items (level) and variability across items (spread).

Research literature offers different possibilities for measuring similarity. For example, subtracting one person's score on a particular dimension from the other person's score (Gaunt, 2006) yields an index of difference-score based similarity, when used as an absolute value (Gaunt, 2006; Luo & Klohnen, 2005). Using this method, quite a lot of information gets lost and only differences in profile levels are taken into account. More information can be gained by measuring similarity with profile scores (Gaunt, 2006; Luo & Klohnen, 2005). An individual's responses across all items in the given domain are correlated with another individual's responses in items of the same domain. Yet, while sensitivity to varying shapes is given, differences in level of the two profiles are not taken into account. As none of the presented methods consider all relevant aspects, we followed the advice of Kenny et al. (2006, p. 327): "The intraclass correlation ... should be used when shape, spread, and level are all relevant". Hence, for every teacher–student dyad we computed an intraclass correlation based on all 15 items of the instrument we used to measure students' and teachers' personality traits, resulting in one personality similarity value per student.

Results

Personality similarity

For the following examinations it is necessary that personality similarity values differ sufficiently, so that there are students who are more similar and students who are less similar to their respective teacher. For the German language class sample, the mean personality similarity value was $i = .21$ ($SD = .29$) with a range from $i_{min} = -.58$ to $i_{max} = .86$. For the mathematics class sample, the mean personality similarity value was $i = .17$ ($SD = .29$) with a range from $i_{min} = -.87$ to $i_{max} = .84$.

Judgement bias

In the present study, judgement bias is defined as the additional impact of task-irrelevant aspects with students' achievement level held under control. Therefore, we used stepwise multiple regression analysis in SPSS 20 to investigate the biasing impact of personality similarity on teacher judgements. In both domains under study, two models for each type of judgement (global and task-specific judgement) were set up: In the first model, teacher judgement was predicted by student achievement in test, in the second model, the personality similarity index was added as a second independent variable.

Impact of personality similarity on teacher judgements of reading comprehension

Global Judgements. For global judgements, personality similarity had a positive and significant impact on teacher judgement of reading comprehension, even with student achievement in test held constant. A higher degree of personality similarity between teacher and student went along with a more positive teacher global judgement. Student achievement (i.e. raw scores in the reading comprehension test) as a single predictor explained 10% of variance in teacher judgement (see Model 1 in Table 3). Both, student achievement and personality similarity accounted for 14% of variance in teacher judgement, indicating an incremental validity of 5% for personality similarity, which is a significant increase (see Model 2 in Table 3). Correlation between the two independent variables was $r = -.04$, so we can dismiss that they are confounded and that student achievement is affected by personality similarity. No matter how similar students were to their respective teacher, they did not differ in their achievement in test.

Task-specific Judgements. For task-specific judgements, teachers had to decide for seven selected tasks of the reading comprehension test, if their students will solve each of the tasks or not. Consequently, student performance as the independent variable that predicts teacher judgement was operationalised as the number of students' solved tasks out of the seven selected tasks. Student performance as a single predictor explained 5% of variance in teachers' task-specific judgements in the domain of reading comprehension (see Model 3 in Table 3). With student performance and personality similarity put in a multiple regression model simultaneously, personality similarity had no significant additional impact on teacher judgement of reading comprehension. Again, we can dismiss that the independent variables are confounded, as the correlation between them was $r = -.02$.

Table 2. Descriptive data and intercorrelations for student test scores, teacher judgements, and personality similarity.

		N	*M*	SD	Min	Max	RSM	STM	GJM	TJM	RSR	STR	GJR	TJR
Math	**Raw scores (RSM)**[a]	241	17.51	4.80	2.00	26.00								
	Solved tasks (STM)[b]	241	4.43	1.69	0.00	7.00	.73**							
	Global judgement (GJM)[c]	241	3.41	1.00	1.00	5.00	.55**	.45**						
	Task-specific judgement (TJM)	236	5.42	1.52	0.00	7.00	.49**	.41**	.75**					
	Personality similarity[d]	293	.17	.29	-.87	.84	.03	.07	.12#	.13#				
Reading	**Raw scores (RSR)**[a]	168	17.27	3.16	5.00	24.00	.26**	.34**	.18**	.13*				
	Solved tasks (STR)[b]	165	3.44	1.71	0.00	7.00	.19**	.23**	.13*	.12#	.77**			
	Global judgement (GJR)	168	3.67	1.00	1.00	5.00	.30**	.31**	.44**	.34**	.39**	.36**		
	Task-specific judgement (TJR)	165	4.98	1.53	1.00	7.00	.36**	.31**	.39**	.31**	.24**	.23**	.65**	
	Personality similarity[e]	197	.21	.29	-.58	.86	-.05	-.04	.21*	.11	-.04	-.02	.21**	.07

[a] Raw scores are the sum of all solved items in test.
[b] Solved tasks are the number of solved tasks out of the seven selected tasks.
[c] Mean value of three global judgements about student's ability in arithmetic, geometry, and written math problems.
[d] Personality similarity between student and the respective mathematics teacher.
[e] Personality similarity between student and the respective German language teacher.
#$p < .10$; *$p < .05$; **$p < .01$.

Table 3. Multiple regression analysis predicting teacher judgement of reading comprehension from student performance and personality similarity.

Reading comprehension	Global judgement		Task-specific judgement	
	Model 1 (n = 168)	Model 2 (n = 168)	Model 3 (n = 165)	Model 4 (n = 165)
	β	β	β	β
Student performance	.32** [a]	.33** [a]	.23** [b]	.23** [b]
Personality similarity		.22**		.07
R^2	.10	.15	.05	.06
Adjusted R^2	.10	.14	.05	.05
ΔR^2		.05**		.00

[a] Student performance as measured with raw scores (sum score, students achieved in the reading comprehension test).
[b] Student performance as measured with the number of solved tasks out of the seven selected tasks.
*p < .05; **p < .01.

Impact of personality similarity on teacher judgements of mathematics

Global Judgements. Student achievement in the mathematics test as a single predictor explained 30% of variance in the teachers' global judgements on students' abilities in mathematics (see Model 1 in Table 4). Multiple regression analysis on global judgements of students' achievement in mathematics showed that personality similarity had a significant additional impact on teacher judgement, when controlling for student achievement (see Model 2 in Table 4). As was the case for global judgements in reading comprehension, a higher degree of personality similarity came together with a more positive teacher judgement about the students' general ability in mathematics. However, incremental validity for personality similarity was significant but rather small for global judgements in mathematics (ΔR^2 = .01). Together, the two independent variables accounted for 31% of variance in the teachers' ratings. With correlation coefficient r = .03 for the independent variables, we can again assume that the independent variables are not confounded.

Task-specific Judgement. Student performance on the mathematics test (i.e. the number of students' solved tasks out of the seven selected items) as a single independent variable explained 18% of variance in teachers' task-specific judgements of mathematics (see Model 3 in Table 4). A multiple regression analysis with teachers' estimation of the number of solved tasks as a dependent variable and the number of students' solved tasks and personality similarity as independent variables was conducted. While controlling for students' performance, results show that personality similarity did not have a significant additional impact on teachers' task-specific judgement (see Model 4 in Table 4). Again, independent variables were not confounded (r = .07).

Table 4. Multiple regression analysis predicting teacher judgement of mathematics from student performance and personality similarity.

	Global judgement		Task-specific judgement	
Mathematical competence	Model 1	Model 2	Model 3	Model 4
	(n = 241)	(n = 241)	(n = 236)	(n = 236)
	β	β	β	β
Student performance	.55** [a]	.55** [a]	.43** [b]	.43** [b]
Personality similarity		.11*		.11
R^2	.30	.31	.18	.19
Adjusted R^2	.30	.31	.18	.19
ΔR^2		.01*		.01

[a] Student performance as measured with raw scores (sum score, students achieved in the mathematical competence test).
[b] Student performance as measured with the number of solved tasks out of the seven selected tasks.
*p < .05; **p < .01.

Discussion

In the present study, the accuracy of global and task-specific teacher judgements in the domains of mathematics and reading and the impact of personality similarity between students and their teacher on teacher judgements was examined. We found that student performance was a significant predictor of teacher judgement in both domains and in both types of judgement. Nevertheless, teacher judgements can be considered inaccurate and far from being perfect. This is in line with former research on teacher judgement accuracy (Hoge & Coladarci, 1989; Südkamp et al., 2012). For the domain of reading we found that personality similarity influenced teachers' global judgements (which was not the case for task-specific judgements), even with students' achievement level controlled. A smaller, but also significant effect was found in the domain of mathematics. Students who were more similar to their teacher were judged more positively than students who were more dissimilar. Thus, personality similarity, which was found to be uncorrelated with student achievement in the present study, systematically affects the (inaccurate) global judgement. In neither domain, however, was task-specific judgement significantly influenced by personality similarity between teacher and student over and above student performance. Global judgements are rather uninformed judgements, as teachers are asked to deliver a judgement on a student's ability on a rating scale without knowledge about the items in the tests. Without that knowledge, teachers have to bring together their own understanding of the domain with their memory of the particular student's achievement behaviour. Especially if the teacher has to deliver judgements about several students (as it was the case in our study), he or she might use pieces of information about the student to be judged, that are easier to access than information about relevant achievement behaviour (cf. Kahneman,

2011: heuristics of substituting questions). Such easily accessible information can be, for example, the student's gender, which is probably confounded with gender stereotypes, or physical attraction (Ritts et al., 1992). Similarly, feelings of likeableness towards a student are also easily to access and the present study shows that personality similarity between student and teacher contributes to a small amount to the explanation of variance in global judgement.

Task-specific judgements, however, are more informed judgements, as teachers are asked to deliver a judgement on student's ability relying on the estimated performance on a certain number of given tasks. To provide an accurate task-specific judgement about a particular student, teachers have to integrate knowledge about the student (his or her ability and strategies to solve the tasks in the particular domain) with knowledge or even facts about task characteristics (i.e. which aspects contribute to make the task difficult or easy to solve) (Karing et al., 2011). Compared to global judgement, the focus of task-specific judgement relies to a smaller degree on the student as a person, while aspects of (pedagogical) content knowledge become more important in this type of judgement: In order to deliver an accurate judgement, teachers also have to judge the difficulty of the questions depending on the particular student's ability. Given that, global judgement as a judgement that relies mainly on knowledge about the student is more likely to be affected by personality similarity than task-specific judgement, which is to a greater degree also dependent on the teacher's accurate judgement of task difficulty.

A noticeable result concerns the higher degree of explained variance in the domain of mathematics, as compared to the domain of reading comprehension. Although we found a similar pattern of results for both domains under study, the impact of personality similarity on global judgements was higher in the domain of reading (β = .22) than in the domain of mathematics (β = .11). Student performance, however, predicted global judgement to a higher degree in mathematics (β = .55) and to a lower degree in reading (β = .33). Estimating a student's ability in more commonly defined domains like arithmetic, geometry or written math problems, seems easier for the teachers than estimating a student's ability in reading comprehension, which is also more difficult to observe. In the present study, the global judgement in the domain of mathematics is a combined measure of teachers' global judgements of arithmetic, geometry and written math problems, whereas global judgement in reading consisted of only one item. Accordingly, mathematics teachers had at least some hints about the content of the underlying test that tapped into specific math knowledge and skills. This feature of the present study might also contribute to the explanation of the

higher beta value of student performance and the lower beta value of personality similarity in the domain of mathematics. Additionally, the items in the test are more comparable to tasks in every day lessons in mathematics, whereas the reading items are not that close to every day lessons in German language arts.

The finding that different types of judgement are differentially affected by a judgement bias also points to a rarely discussed point in research about teacher judgement accuracy. To date, only a few moderators of teacher judgement accuracy have been uncovered to explain the widely found great variability of teacher judgement accuracy in different domains and judgement contexts (Hoge & Coladarci, 1989; Südkamp et al., 2012). One reason that searching for moderator variables yielded no sound results to date might be that there is no real consent in literature about methodologies for examining judgement accuracy: "the inaccuracy of teachers' judgements may be grounded in the studies' methodologies rather than in the teachers' diagnostic competence" (Südkamp et al., 2012, p. 744). The present study adds further evidence that the moderator variables found for global judgements might not be the same for task-specific judgements.

Limitations and directions for further research

The results of the present study suggest that it is worth considering the interplay of teachers' and students' characteristics, when further examinations in the field of teacher judgement accuracy are planned and conducted. Nevertheless, some limitations have to be taken into account, when interpreting the results. It might also provide directions for future research. A higher degree of similarity between teachers and students in terms of their personality traits yielded a more positive global judgement of students' achievement in mathematical competence and reading comprehension, with student performance level controlled. Whether the pattern of results is also the same for teacher judgements about other student characteristics, e.g. about motivational or emotional variables, could not be clarified with this study. We have to assume that our findings are limited to judgements of student achievement and generalisations cannot be made for estimations of other student characteristics.

In order to examine if teacher judgements about student competences in reading comprehension and mathematics are accurate, we used students' performance on standardised tests as a criterion. With that, we are in line with other research in the field of judgement accuracy (e.g. Bates & Nettelbeck, 2001; Demaray & Elliott, 1998). It can be questioned if standardised tests assessment tests are valid indicators of student

competences, which is an important prerequisite for the analyses. Yet, internal consistency and measures of criterion validity seem to corroborate our assumption that the tests are validly measuring student achievement in mathematics and reading.

Another limitation can be seen in the fact that our analyses are based on teachers' and students' self-reports on the personality inventory. Neither did we ask students how similar they perceive themselves to be to their respective teachers, and vice versa. Nor did we use teachers' perceptions of students' personality traits to compute the similarity index. It can be hypothesised that perceived similarity, as compared to actual similarity has an even stronger impact on teacher judgement (Montoya et al., 2008; Strauss, Barrick, & Connerley, 2001). Yet we were not able to test this hypothesis.

Research concerning the similarity-attraction effect (Byrne, 1997) considers several aspects of similarity to be relevant for attraction, such as e.g. gender, sociodemographic variables or attitudes, whereas the present study concentrated on similarity of personality traits. A generalisation on similarity in other aspects cannot be drawn. Due to practical considerations, teachers' and students' personality traits were detected with a short version of the Big-Five-Inventory (BFI-S; Gerlitz & Schupp, 2005), and without considering the factor structure for the computation of the personality similarity index. More extended analyses allowing a more detailed look into which factors or facets are more or less important concerning the impact of similarity on teacher judgements were not possible with the used inventory.

Teachers differ in their accuracy of judgements and they might be more or less aware of their judgement being biased by sympathy or attraction. Thus, it might be of interest if regression coefficients show variation across classes (cf. Blien, Wiedenbeck, & Arminger, 1994). This could have been addressed by multilevel analyses. A prerequisite for hierarchical linear models is a sufficient number of both, level 1 and level 2 units. With an average of 4.2 students per classroom in the German language classroom sample, and with an average of 4.6 students in the mathematics classroom sample we do not reach the suggested necessary sample size of minimum 30 individuals in 30 classrooms (Kreft, 1996). However, the necessity of conducting multilevel analyses for our present study was not taken for granted. As we are interested in the marginal effects, that is, the individual effect of personality similarity on the teacher judgement as an individual outcome, multilevel modelling is not the appropriate approach: We consider the effect to be independent from the assignment to clusters. This enables us to conduct population-averaged analyses, which

do not take into account the cluster-structure of the data (cf. Graubard & Korn, 1994). The effects of the present study thus have to be interpreted also against this background. It therefore seems to be a direction for future research to investigate effects on teacher judgement accuracy in a hierarchical model. This could probably provide more insight in the role of teachers' awareness concerning judgement biases.

Conclusion

In summary, the present study adds some empirical evidence on the research about teacher judgement accuracy, as it allows basically three conclusions: First, similarity between student and teacher personality characteristics matters for the accuracy of teachers' judgements. Second, the type of judgement matters for the accuracy of teacher judgement. We found global judgements in two different domains to be affected by a similarity bias, whereas task-specific judgement was not. Third, test characteristics, that is, the subject domain, matter for the accuracy of teacher judgements, as differences in explained variance between the two domains of reading comprehension and mathematics can be observed for both types of judgement.

This work was supported by the German Research Foundation (DFG) [grant numbers AR301/6-1, AR301/6-2 & AR301/6-3].

References

Ajzen, I. (1974). Effects of information on interpersonal attraction: Similarity versus affective value. *Journal of Personality and Social Psychology, 29*, 374–380. doi:10.1037/h0036002

Alexander, K. L., Entwisle, D. R., & Thompson, M. S. (1987). School performance, status relations, and the structure of sentiment: Bringing the teacher back in. *American Sociological Review, 52*, 665–682.

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgements. When and for what reasons? In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training and learning* (pp. 27–43). Rotterdam: Sense Publishers.

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*, 177–187. doi:10.1080/01443410020043878

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften [Teachers' professional competence]. *Zeitschrift für Erziehungswissenschaft, 9*, 469–520. doi:10.1007/s11618-006-0165-2

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behaviour perceptions and gender on teacher's judgements of students' academic skill. *Journal of Educational Psychology, 85*, 347–356. doi:10.1037/0022-0663.85.2.347

Blien, U., Wiedenbeck, M., & Arminger, G. (1994). Reconciling macro and micro perspectives by multilevel models: An application to regional wage differences. In I. Borg & P. P. Mohler (Eds.), *Trends and perspectives in empirical social research* (pp. 266–282). Berlin: de Gruyter.

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften [Mathematics teachers' diagnostic skills]. In M. Kunter, J. Baumert, W. Blum, & U. Klusmann (Eds.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (pp. 215–234). Münster: Waxmann.

Byrne, D. (1971). *The attraction paradigm*. New York, NY: Academic Press.

Byrne, D. (1997). An overview (and underview) of research and theory within the attraction paradigm. *Journal of Social and Personal Relationships, 14*, 417–431. doi:10.1177/0265407597143008

Carr, M., & Kurtz-Costes, B. E. (1994). Is being smart everything? The influence of student achievement on teachers' perceptions. *British Journal of Educational Psychology, 64*, 263–276. doi:10.1111/j.2044-8279.1994.tb01101.x

Coladarci, T. (1986). Accuracy of teacher judgements of student responses to standardized test items. *Journal of Educational Psychology, 78*, 141–146. doi:10.1037/0022-0663.78.2.141

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456–473. doi:10.1037/h0057173

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgements of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*, 8–24. doi:10.1037/h0088969

Dipboye, R., & Gaugler, B. B. (1993). Cognitive and behavioral processes in the selection interview. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 135–170). San Francisco, CA: Jossey-Bass.

Englehart, J. M. (2009). Teacher–student interaction. In L. J. Saha & A. G. Dworkin (Eds.), *International handbook of research on teachers and teaching* (pp. 711–722). Boston, MA: Springer.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgements of students' reading with differing achievement levels. *The Journal of Educational Research, 102*, 453–462. doi:10.3200/JOER.102.6.453-462

Gaunt, R. (2006). Couple similarity and martial satisfaction: Are similar spouses happier? *Journal of Personality, 74*, 1401–1420. doi:10.1111/j.1467-6494.2006.00414.x

Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP [Assessment of big-five-based personality traits in the German socioeconomic panel study]. Berlin: DIW.

Graubard, B. I., & Korn, E. L. (1994). Regression analysis with clustered data. *Statistics in Medicine, 13*, 509–522. doi:10.1002/sim.4780130514

Hadley, S. T. (1954). A school mark – Fact or fancy? *Educational Administration and Supervision, 40*, 305–312.

Hany, E. A. (1997). Modeling teachers' judgement of giftedness: A methodological inquiry of biased judgement. *High Ability Studies, 8*, 159–178. doi:10.1080/1359813970080203

Helmke, A. (2007). Unterrichtsqualität erfassen, bewerten, verbessern [Assessment, evaluation, and improvement of instructional quality]. Stuttgart: Klett/Kallmeyer.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgements of academic achievement: A review of literature. *Review of Educational Research, 59*, 297–313. doi:10.3102/00346543059003297

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*, 177–182. doi:10.1111/j.1745-3984.1985.tb01056.x

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgements of students' test performance. *School Psychology Quarterly, 22*, 115–144. doi:10.1037/1045-3830.22.2.115

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81. doi:10.1111/j.1745-3984.1998.tb00528.x

Itskowitz, R., Navon, R., & Strauss, H. (1988). Teachers' accuracy in evaluating students' self-image: Effect of perceived closeness. *Journal of Educational Psychology, 80,* 337–341. doi:10.1037/0022-0663.80.3.337

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology, 29*, 281–388.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A., & Möller, J. (2012). Zum Zusammenhang von Intelligenz und Urteilsgenauigkeit bei der Beurteilung von Schülerleistungen im Simulierten Klassenraum [On the relation of intelligence and judgement accuracy in the process of assessing student achievement in the simulated classroom]. *Zeitschrift für Pädagogische Psychologie, 26*, 251–261. doi:10.1024/1010-0652/a000076

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgements. *Learning and Instruction, 28*, 73–84. doi:10.1016/j.learninstruc.2013.06.001

Karing, C., Matthäi, J., & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? [Lower secondary school teacher judgement accuracy of students' reading competence – A matter of specificity?]. *Zeitschrift für Pädagogische Psychologie, 25*, 159–172. doi:10.1024/1010-0652/a000041

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: The Guilford Press.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Los Angeles: California State University.

Kruglanski, A. W., & Ajzen, I. (1983). Bias and error in human judgement. *European Journal of Social Psychology, 13*, 1–44. doi:10.1002/ejsp.2420130102

Luo, S., & Klohnen, E. C. (2005). Assortative mating and martial quality in newlyweds: A couple-centered approach. *Journal of Personality and Social Psychology, 88*, 304–326. doi:10.1037/0022-3514.88.2.304

Montoya, R. M., & Horton, R. S. (2013). A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships, 30*, 64–94. doi:10.1177/0265407512452989

Montoya, R. M., Horton, R. S., & Kirchner, J. (2008). Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships, 25*, 889–922. doi:10.1177/0265407508096700

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*, 335–360. doi:10.3102/0002831210374874

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgements of physically attractive students: A review. *Review of Educational Research, 62*, 413–426. doi:10.3102/00346543062004413

Schrader, F.-W., & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile [Let teachers guide themselves by inappropriate factors when judging their students' abilities? Determinants of teacher judgements]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 22, 312–324.

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgements, decisions, and behavior. *Review of Educational Research, 51*, 455–498. doi:10.3102/00346543051004455

Strauss, J. P., Barrick, M. R., & Connerley, M. L. (2001). An investigation of personality similarity effects (relational and perceived) on peer and supervisor ratings and the role of familiarity and liking. *Journal of Occupational and Organizational Psychology, 74*, 637–657. doi:10.1348/096317901167569

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgements of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743–762. doi:10.1037/a0027627

Wang, S. S., Treat, T. A., & Brownell, K. D. (2008). Cognitive processing about classroom relevant contexts: Teachers' attention to and utilization of girls' body size, ethnicity, attractiveness, and facial affect. *Journal of Educational Psychology, 100*, 473–489. doi:10.1037/0022-0663.100.2.473

Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens[1]

Tobias Rausch, Jacqueline Matthäi und Cordula Artelt
Otto-Friedrich-Universität Bamberg

Zusammenfassung. Die vorliegende Studie untersucht, ob und inwieweit das Wissen im Bereich Textverstehen von Deutschlehrkräften der Sekundarstufe mit der Urteilsgüte bei der Einschätzung von Schülerleistungen in diesem Bereich zusammenhängt. Dazu wurde das Text-, Aufgaben- und Strategiewissen von 77 Lehrkräften erhoben und mit Indikatoren der globalen und aufgabenspezifischen Urteilsgüte bei der Einschätzung der Textverstehensleistung von Schülerinnen und Schülern der achten und neunten Klassenstufe korrelativ in Bezug gesetzt. Die Lehrkräfte schätzen ihre Schülerinnen und Schüler relativ akkurat ein und verfügen auch über umfassendes Wissen in den getesteten Bereichen. Es zeigen sich jedoch weder für die Güte von globalen Urteilen noch für die Güte von aufgabenspezifischen Urteilen substanzielle und signifikante Zusammenhänge mit dem Lehrerwissen.

Schlüsselwörter: Textverstehen, diagnostische Kompetenz, Urteilsgüte, Lehrerwissen

Bisherige Befunde zur diagnostischen Kompetenz von Lehrkräften haben gezeigt, dass hierbei nicht von einer bereichsübergreifenden Kompetenz ausgegangen werden kann, sondern dass diese fach- bzw. domänenspezifisch geprägt ist (Schrader, 2010; Spinath, 2005). Es erscheint daher plausibel, dass für die angemessene Einschätzung von Schülerleistungen in spezifischen (z.B. fachlichen) Bereichen nicht nur allgemeines diagnostisches und methodisches Wissen auf Seiten der Lehrkräfte notwendig ist, sondern auch deren Wissensgrundlagen hinsichtlich der fachlichen Anforderungen der zu beurteilenden Schülerleistung eine Rolle spielen. Entsprechend ist davon auszugehen, dass ein umfassendes fachliches und fachdidaktisches Wissen dazu beiträgt, dass Lehrkräfte die Leistungen ihrer Schülerinnen und Schüler im jeweiligen Bereich akkurat beurteilen (Artelt, 2009). Neben den vergleichsweise gut untersuchten Wissensgrundlagen von Lehrkräften im mathematischen und naturwissenschaftlichen Bereich, die vor allem auf ihre Auswirkungen auf die Leistungen von Schülerinnen und Schülern hin untersucht

[1] Diese Artikelfassung entspricht nicht vollständig dem in der Zeitschrift veröffentlichten Artikel. Dies ist nicht die Originalversion des Artikels und kann daher nicht zur Zitierung herangezogen werden. Der Originalbeitrag ist erschienen als: Rausch, T., Matthäi, J., & Artelt, C. (2015). Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47,* 147-158. doi: 10.1026/0049-8637/a000124 . Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47(3), © Hogrefe Verlag, Göttingen 2015.

wurden (z. B. Krauss, Baumert & Blum, 2008; Lange, Kleickmann, Tröbst & Möller, 2012) gibt es zu den Wissensgrundlagen von Lehrkräften im Bereich Lesen und Textverstehen bisher nur wenig empirische Evidenz. In der vorliegenden Untersuchung soll untersucht werden, wie akkurat Lehrkräfte Schülerfähigkeiten im Bereich Textverstehen einschätzen. Außerdem soll überprüft werden, wie die Wissensgrundlagen von Deutschlehrkräften im Bereich des Lesens und Textverstehens ausgeprägt sind und ob von einem Einfluss des Lehrerwissens auf die Urteilsgüte bei der Beurteilung von Schülerleistungen im Textverstehen ausgegangen werden kann.

Theoretischer Hintergrund

Das Beurteilen von Schülerleistungen ist zentraler Bestandteil der täglichen Arbeit von Lehrkräften. Die Genauigkeit dieser Urteile wird als ein wichtiger Aspekt der professionellen Handlungskompetenz von Lehrkräften angesehen (z. B. Baumert & Kunter, 2006; Brunner, Anders, Hachfeld & Krauss, 2011). Sowohl für eine faire Leistungsbeurteilung zum Zweck der Notengebung als auch für die Beurteilung des Leistungsstands und Leistungsfortschritts von einzelnen Schülerinnen und Schülern bzw. Schulklassen und zur didaktischen Planung sind möglichst akkurate Urteile über die Schülerinnen und Schüler nötig. In Untersuchungen zur Urteilsgenauigkeit werden diese oft als Ausdruck einer zugrunde liegenden Fähigkeit eines Urteilers angesehen, Personen zutreffend zu beurteilen (vgl. Schrader, 2010). Einer konkreten Beurteilungssituation gehen jedoch vielfältige Erfahrungen der Lehrkraft mit den jeweiligen Schülerinnen und Schülern voraus, worauf sich dann das in der Untersuchung abgegebene Urteil stützen sollte. Die Akkuratheit des Urteils ist dementsprechend auch ein Indikator für die Güte der Einschätzungen der Lehrkräfte im Unterricht. Bisherige Untersuchungen zeigen, dass Lehrkräfte die Leistungsrangfolge ihrer Schülerinnen und Schüler in unterschiedlichen Leistungsbereichen einigermaßen gut einschätzen können: Südkamp, Kaiser und Möller (2012) berichten auf der Basis einer Metaanalyse über verschiedene Leistungsbereiche hinweg eine mittlere Korrelation von $r = .63$ zwischen den Lehrerurteilen und der Testleistung ihrer Schüler (vgl. auch Hoge & Coladarci, 1989). Allerdings finden sich in der Regel große interindividuelle Unterschiede zwischen den Lehrkräften. Südkamp und Kollegen (2012) gehen davon aus, dass die Urteilsgüte nicht nur von Merkmalen der Lehrkraft, der Beurteilungsaufgabe oder des zugrundeliegenden Tests abhängt, sondern dass sie auch in Abhängigkeit von dem einzuschätzenden Schülermerkmal variiert. So hat sich z.B. gezeigt, dass die Urteilsgüte bei der Einschätzung leistungsbezogener Merkmale meist höher ist als bei emotional-motivationalen Merkmalen (Karing, 2009; Spinath, 2005).

Zudem scheint sich im Leistungsbereich die Urteilsgüte eher fachbezogen abzubilden: Eine Reihe von Studien konnte zeigen, dass die Urteilsgüte zwischen den eingeschätzten Domänen variiert, wenn Lehrkräfte ihre Schülerinnen und Schüler jeweils in unterschiedlichen Bereichen einschätzen (Eckert, Dunn, Codding, Begeny & Kleinmann, 2006; Hopkins, George & Williams, 1985; Lorenz & Artelt, 2009). Eine generelle bereichsübergreifende Beurteilungskompetenz, die unabhängig vom Urteilsgegenstand ist, scheint es bei Lehrkräften demnach nicht zu geben (vgl. auch Schrader, 2010). Vielmehr kann davon ausgegangen werden, dass es neben generellem und bereichsübergreifendem methodischem Wissen (z. B. Kenntnis und Beherrschung diagnostischer Methoden, Wissen über Urteilsfehler und -tendenzen) auch bereichsspezifische Komponenten gibt. Es erscheint daher lohnenswert, neben dem methodischen Wissen auch fach- und domänenspezifische Wissensgrundlagen in den Blick zu nehmen (Helmke, Hosenfeld & Schrader, 2004). Auch Helmke, Hosenfeld und Schrader (2004) sehen im gegenstandsbezogenen Wissen von Lehrkräften – bspw. über schwierigkeitsgenerierende Aufgabenmerkmale und über angemessene Lösungsstrategien im Umgang mit den gestellten Aufgaben – neben der Intelligenz der Lehrkräfte, der Kenntnis und Beherrschung diagnostischer Methoden, ihrem Wissen über Urteilsfehler und über die einzuschätzenden Schüler und Klassen die Grundlage für akkurate Urteile. Um die Leistungen von Schülerinnen und Schülern akkurat einschätzen zu können, müssen Lehrkräfte über Wissen über Schülerkognitionen und Lernprozesse im jeweiligen Bereich verfügen. Hierzu zählen etwa Kenntnisse bzw. Überzeugungen darüber, welche Beobachtungen darüber Auskunft geben, ob eine Schülerin oder ein Schüler eine bestimmte Aufgabenanforderung beherrscht und wie das jeweilige Schülerverhalten in Bezug auf die einzuschätzende Schülerfähigkeit interpretiert werden sollte (National Research Council, 2001).

Neben den verschiedenen Wissenskomponenten werden in einer spezifischen Beurteilungssituation natürlich auch bereits während des zurückliegenden Unterrichtsgeschehens gebildete Einschätzungen der Schülerfähigkeiten wirksam, die für die Urteilsbildung ggf. sogar dominanter sind als fachspezifische Wissenselemente.

Wissensgrundlagen diagnostischer Kompetenz im Bereich Textverstehen

Um zu untersuchen, wie sich das gegenstandsbezogene Wissen der Lehrkräfte in der Urteilsgüte widerspiegelt, ist es zunächst notwendig, zu spezifizieren, welche fach- bzw. domänenbezogenen Wissensgrundlagen für akkurate Einschätzungen von Schülerleistungen im jeweiligen Anforderungsbereich zentral sind. Für den Bereich des

Lesens bzw. des Textverstehens wird davon ausgegangen (Artelt, McElvany, Christmann, Richter, Groeben, Köster et al., 2005), dass die individuelle Textverstehensleistung variiert in Abhängigkeit von

- leserbezogenen Merkmalen (wie z. B. Vorwissen oder Motivation),
- den Aktivitäten des Lesers (z. B. anforderungsabhängiger Einsatz von Strategien),
- der Leseanforderung (z.B. verstehendes Lesen) und
- der Beschaffenheit des Textes (z. B. Textschwierigkeit).

Lehrkräfte, die die Leistungen ihrer Schülerinnen und Schüler im Textverstehen einschätzen wollen, müssen demnach bei der Beurteilung diese vier Varianzquellen berücksichtigen. Über je mehr Wissen sie in den jeweiligen Bereichen verfügen, desto akkurater sollte ihr Urteil ausfallen. Unter anderem auf Basis dieser Überlegungen hat Matthäi (in Vorb.) ein Testverfahren entwickelt, das insbesondere drei der oben genannten Bereiche abbildet: Das Wissen der Lehrkräfte über (1) schwierigkeitsgenerierende Merkmale von Texten und von (2) Lese(verstehens)anforderungen, sowie über (3) Strategien, die für die Lösung unterschiedlicher Verstehensanforderungen notwendig sind.

Konzeptualisierung der Wissensgrundlagen

Wissen über Textmerkmale

Bezüglich der Analyse und Beschreibung von Textschwierigkeit kann auf Vorarbeiten aus verschiedenen Disziplinen zurückgegriffen werden. Zu unterscheiden sind klassische Arbeiten in der Psychologie zur Textverständlichkeit von Groeben (1982) oder Langer, Schulz von Thun und Tausch (2006) sowie die auf der Basis kognitionspsychologischer Theorien (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983) ableitbaren schwierigkeitsgenerierenden Textmerkmale. Zudem existieren unterschiedliche Varianten von Lesbarkeitsindizes (LIX, Flesh etc.), die sich in der Regel allein auf die Wort- und Satzlänge in den Texten konzentrieren. Im Resultat stellen Texte mit kurzen Wörtern und kurzen Sätzen leicht lesbare Texte dar. Diese Annahme basiert auf der Beobachtung, dass der Zugriff auf das mentale Lexikon und damit die Effizienz der Worterkennung bei kürzeren (oft auch bei bekannteren) Wörtern rascher erfolgt. Zudem vereinfachen kurze Sätze u. a. auch deshalb Texte, weil damit in der Regel komplexe grammatikalische Strukturen vermieden werden. Insofern sind diese einfachen und z.T. automatisiert auswertbaren Indikatoren gute Proxys von Textschwierigkeit. Faktoren, die die Leserseite fokussieren oder über Wort- und Satzlänge hinausgehen, werden in dieser Art von Lesbarkeitsindizes jedoch nicht berücksichtigt. In der Literatur finden sich darüber hinaus aber auch Annahmen und z.T. auch Belege für makrostrukturelle Aspekte, die die

Lesbarkeit von Texten erhöhen (vgl. Groeben, 1982; Langer et al., 2006). Diese beziehen sich auf Merkmale wie den Aufbau des Textes, die Kürze und Prägnanz der Darstellung, sowie auf Elemente zusätzlicher Stimulanz, die u. a. auch an motivationalen Faktoren ansetzen. Ähnlich wie in der Linguistik wird in kognitionspsychologischen Arbeiten insbesondere auf das Merkmal der Textkohäsion gesetzt. Wenngleich ihre alleinige Wirkung in Abhängigkeit vom Vorwissensniveau und der Lesefähigkeit kontrovers diskutiert wird (vgl. Gilabert, Martinez & Vidal-Abarca, 2005; McNamara & Kintsch, 1996; O'Reilly & McNamara, 2007), lässt sich sehr allgemein sagen, dass in einem kohäsiveren Text der rote Faden des Textes expliziter zum Ausdruck kommt. Um dies zu erreichen, wird auf lokaler Ebene (d. h. auf Satzebene bzw. bei der Verbindung benachbarter Sätze) u. a. mit Kohäsionsmarkern wie z. B. Konnektoren und einfacheren grammatikalischen Strukturen gearbeitet. Auf globaler Ebene (d. h. bei der Verbindung über Textabsätze hinweg) werden dafür Kohäsionsmarker wie z. B. Zwischenüberschriften eingesetzt (vgl. Graesser, McNamara & Louwerse, 2003). Wie sich aber schon an der Diskussion um die widersprüchliche Wirkung von Textkohäsion in Abhängigkeit von Lesermerkmalen ausdrückt, ist die Frage nach eindeutig zu kennzeichnenden Textmerkmalen nicht leicht zu beantworten und oft nicht unabhängig vom Textgenre. So weisen literarische Texte etwa andere Anforderungen auf als expositorische Texte (u. a. Artelt & Schlagmüller, 2004). Weitere Merkmale, wie die Propositionsdichte (z. B. Turner & Greene, 1977) sind nur sehr aufwändig auszuwerten und deshalb bei der Konstruktion des Wissenstests nicht weiter berücksichtigt worden.

Wissen über Aufgabenmerkmale

Es ist vielfach belegt, dass auch die Leseabsicht einen Effekt auf die Qualität des Textverstehens hat. So variiert etwa die Menge und Art der gezogenen Inferenzen aus dem Text in Abhängigkeit von den Rezeptionszielen (Graesser, Singer & Trabasso, 1994). Fragen bzw. Aufgaben zum Text geben diese Rezeptionsziele vor und stellen unterschiedliche Anforderungen an den Leser. Die Aufgabenanforderungen können jedoch nicht unabhängig vom zugrunde liegenden Text bewertet werden. Kirsch und Mosenthal haben sich im Kontext von Reanalysen zu amerikanischen Leseuntersuchungen (Mosenthal & Kirsch, 1994) mit schwierigkeitsgenerierenden Faktoren von Aufgaben beschäftigt und eine Systematik entwickelt (vgl. a. Kirsch, 2001). Zentral ist demnach das Zusammenspiel notwendiger Bearbeitungsprozesse in Bezug auf den zugrunde liegenden Text und der jeweils dazu formulierten Leseaufgabe. Auch die Rolle von ablenkenden Informationen im Text für die korrekte Beantwortung der Leseaufgabe wurde berücksichtigt. Zudem

unterscheiden sich Aufgaben in ihrer Schwierigkeit entsprechend der Menge und Art der für eine korrekte Beantwortung der Aufgabe zu ziehenden Inferenzen. Schwierige Aufgaben unterscheiden zwischen dem Text der Aufgabe und des Textes, besitzen mehrere Distraktoren und können nur mit Hilfe schwer zu ziehender Inferenzen (bspw. wissensbasiert) beantwortet werden.

Wissen über Strategien

Die Aktivitäten des Lesers zum Verstehen des Textes sollten an die Leseabsicht und an den Lesestoff angepasst sein (Groeben, 1982), jedoch unterscheiden sich Schülerinnen und Schüler untereinander hinsichtlich des angemessenen Einsatzes von Strategien um gegebene Leseanforderungen zu erfüllen (Schneider & Pressley, 1997). Der Einsatz von Strategien, die an Leseabsicht, Textinhalt und -struktur angepasst sind, erhöht die Verstehens- und Behaltensleistung des Lesers (Grzesik, 1990). Während sich kognitive Strategien auf Wiederholung, Organisation und Elaboration beim Lesen beziehen, umfassen metakognitive Strategien die Planung, Überwachung und Regulation des Leseprozesses (Baumert & Köller, 1996; Weinstein & Mayer, 1986). Um einschätzen zu können, ob die von den Schülerinnen und Schülern eingesetzten Strategien vor dem Hintergrund der Verstehensanforderung zielführend sind, d. h. dazu führen, dass der Text verstanden wird, ist es notwendig, dass Lehrkräfte über Wissen über die Angemessenheit von Vorgehensweisen bei der Arbeit mit Texten verfügen.

Wissen über Personen

In der vorliegenden Arbeit wird allein das Wissen der Lehrkräfte über schwierigkeitsgenerierende Merkmale von Texten und Aufgaben sowie über angemessene Vorgehensweisen zur Lösung von Verstehensanforderungen gemessen und mit der Urteilsgüte der Lehrkräfte in Bezug gesetzt. Für die akkurate Einschätzung von Schülerinnen und Schülern ist jedoch auch Wissen über leserbezogene Merkmale notwendig, z. B. über die Motivation oder das inhaltliche Vorwissen der Schülerinnen und Schüler sowie über deren konkrete Stärken und Schwächen bzgl. der Leseleistung und der Leseprozesse (Helmke et al., 2004). Dieses personenbezogene urteilsrelevante Wissen lässt sich in einem Wissenstest jedoch nicht einfach abbilden, da es sich aus vielfältigen Gelegenheiten der Lehrkraft speist, in denen die Lehrkräfte im Unterricht Informationen über die Schülerinnen und Schüler generieren. Zusammengenommen stellt das erfolgreiche Zusammenbringen des verfügbaren Wissens über den einzuschätzenden Schüler und des Wissens über den einzuschätzenden Gegenstandsbereich die Grundlage für akkurate Urteile im jeweiligen Leistungsbereich dar (Karing, Matthäi & Artelt, 2011).

Urteilsgüte im Bereich Textverstehen

Es wird angenommen, dass die skizzierten Wissensgrundlagen als notwendige Voraussetzung für die akkurate Beurteilung von Schülerleistungen angesehen werden können. Mit zunehmendem Wissen – so die Annahme – sollten Lehrkräfte daher tendenziell genauere Urteile in Bezug auf konkrete Schülerleistungen im Textverstehen fällen.

Globale und aufgabenspezifische Einschätzung

In der Forschung zur Urteilsgüte wird oft zwischen globalen und aufgabenspezifischen Urteilen unterschieden (z.B. Karing, Matthäi & Artelt, 2011). Während die Lehrkräfte bei einem globalen Urteil eine eher unspezifische Einschätzung über einen Schüler oder eine Schülerin abgeben sollen (z.B. „Wie gut kann dieser Schüler lesen?"), schätzen Lehrkräfte bei aufgabenspezifischen Urteilen ein, ob der Schüler oder die Schülerin bestimmte Aufgaben in einem Leistungstest jeweils richtig gelöst hat oder nicht. Es ist nicht unwahrscheinlich, dass in Abhängigkeit von der Spezifität der Urteilsanforderung unterschiedliche Kriterien bei der Bildung des Urteils herangezogen werden, fachliches Wissen in unterschiedlichem Umfang genutzt wird und auch unterschiedliche Strategien der Informationsverarbeitung angewandt werden (Artelt & Rausch, 2014).

Das Ziel globaler Urteilsbildung kann eher als Eindrucksbildung verstanden werden. Dementsprechend lässt sich vermuten, dass Lehrkräfte hier stärker auf Schülerstereotypen oder soziale Kategorien zurückgreifen (Hofer, 1981; Hörstermann, Krolak-Schwerdt & Fischbach, 2010) und weniger auf der Grundlage ihres domänenspezifischen Wissens urteilen. Um jedoch ein korrektes aufgabenspezifisches Urteil abgeben zu können, müssen Lehrkräfte ihr Wissen über den einzuschätzenden Schüler (z. B. über die im Unterricht gezeigte Performanz bei ähnlichen Aufgaben, über seine typischerweise angewendeten Strategien, aber auch über seine Motivation oder seine Interessen im Zusammenhang mit dem im Test zu bearbeitenden Text) mit ihrem Wissen über die vom Text und von der konkreten Leseaufgabe gestellten Anforderungen (schwierigkeitsgenerierende Merkmale von Text und Aufgaben) und mit dem Wissen über dafür angemessene Bearbeitungsstrategien zusammenbringen (Karing, Matthäi & Artelt, 2011). Das domänenspezifische Wissen der Lehrkräfte sollte hier also eine stärkere Rolle für die Urteilsgüte spielen.

Fragestellungen und Hypothesen

(1) Wie akkurat schätzen Lehrkräfte die Leistung ihrer Schülerinnen und Schüler im Textverstehen global und aufgabenspezifisch ein? Es wird erwartet, dass die Urteilsgüte der Lehrkräfte im Einklang mit bisherigen in der Literatur berichteten Ergebnissen steht.

(2) Über welches Wissen verfügen Lehrkräfte im Hinblick auf schwierigkeitsgenerierende Merkmale von Texten und Aufgaben sowie hinsichtlich der Angemessenheit von Strategien zum Textverstehen? Aufgrund mangelnder bisheriger Befunde hierzu können keine Hypothesen formuliert werden. Daher wird diese Fragestellung explorativ untersucht.

(3) Hat das Wissen von Deutschlehrkräften im Bereich des Textverstehens einen Einfluss auf die globale und auf die aufgabenspezifische Urteilsgüte bei der Beurteilung von Schülerleistungen im Textverstehen? Es wird dabei erwartet, dass ein umfassenderes Wissen der Lehrkräfte im Bereich Textverstehen mit akkurateren Urteilen in diesem Bereich einhergeht. Allerdings lässt sich vermuten, dass der Zusammenhang mit der aufgabenspezifischen Urteilsgüte stärker ist als der Zusammenhang mit der globalen Urteilsgüte.

Methode

Stichprobe

Die Daten für den vorliegenden Beitrag wurden im Rahmen der Bamberger DFG-Forschergruppe BiKS (Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter) erhoben. Die Stichprobe der an der Untersuchung teilnehmenden Lehrkräfte ist definiert über die an der Längsschnittstudie BiKS 8–13 teilnehmenden Schülerinnen und Schüler. Bei 48 Deutschlehrkräften dieser Schülerinnen und Schüler in der Sekundarstufe I erfolgte die Erfassung des Lehrerwissens zunächst im Schuljahr 2010/11. Da nicht alle Lehrkräfte der 462 Schülerinnen und Schüler an der Testung teilnahmen, wurde die Testung ein Jahr später noch einmal durchgeführt, wodurch dann von weiteren 29 Lehrkräften Informationen zu ihrem Wissen im Bereich Textverstehen vorlagen. Insgesamt stehen also Daten aus dem Lehrerwissenstest von 77 Lehrkräften (59 % weiblich) zur Verfügung. Die Lehrkräfte waren im Mittel zum Zeitpunkt

der Befragung 37,5 Jahre alt (SD = 10,6) und verfügten über M = 8,9 Jahre Berufserfahrung (SD = 8,8)[2].

Zu beiden Erhebungszeitpunkten für das Lehrerwissen wurden die Lehrkräfte ebenfalls gebeten, die Leistungen im Textverstehen ihrer an der BiKS-Studie teilnehmenden Schülerinnen und Schüler einzeln einzuschätzen. Insgesamt lagen Leistungsdaten aus dem Lesekompetenztest von 303 Schülerinnen und Schülern und die Schülereinschätzungen von 86 Deutschlehrkräften vor. Für die Untersuchung wurden jedoch diejenigen Lehrkräfte ausgeschlossen, die weniger als drei Schülerinnen und Schüler eingeschätzt haben. Diese Einschränkung erfolgte v. a. aufgrund der Operationalisierung der globalen Urteilsgüte (s.u.), die auf einer Korrelation zwischen Lehrerurteil und Schülertestleistung basiert. Erst ab drei eingeschätzten Schülerinnen und Schülern ist es jedoch möglich aus diesen eine Rangreihe zu bilden, weswegen dieser Cutoff-Wert als Ausschlusskriterium festgelegt wurde. Die Analysestichprobe setzt sich schließlich zusammen aus 44 Deutschlehrkräften der Sekundarstufe I (65 % weiblich), die zum Zeitpunkt der Befragung im Mittel 37,2 Jahre alt waren (SD = 10,5), über 8,1 Jahre Berufserfahrung (SD = 8,2) verfügten. Diese Lehrkräfte schätzten insgesamt 233 Schülerinnen und Schüler (67 % weiblich) der achten bzw. neunten Jahrgangsstufe ein. 42 Schülerinnen und Schüler besuchten die Realschule und 191 das Gymnasium.

Instrumente

Lesekompetenz

Zur Erfassung der Lesekompetenz der Schülerinnen und Schüler wurden die im Rahmen der Forschergruppe BiKS für die Klassenstufen 8 und 9 entwickelten Lesekompetenztests eingesetzt (siehe auch Pfost & Artelt, 2013), die die Lesekompetenz als eindimensionales Fähigkeitskonstrukt messen und über ein Anker-Item Design miteinander verlinkt sind. In beiden Klassenstufen bestand der Test aus drei (zwei expositorische, ein narrativer) Texten und dazugehörigen Items, die jeweils mit rund 25 Minuten Bearbeitungszeit administriert wurden. In der achten Klasse variierte die Textlänge zwischen 443 und 560 Wörtern und umfasste 25 dazugehörige Multiple-Choice Items. In der neunten Klassenstufe bestand der Test aus drei Texten (455 bis 560 Wörter) und 29 auf diese Texte bezogenen Multiple-Choice Items. In beiden Testversionen konnten die Items unter Rückgriff auf den jeweiligen Text beantwortet werden. Inhaltliche Anforderungen an die Schüler bestanden neben dem Finden relevanter Informationen im Text primär darin, textbezogene Interpretationen zu generieren und Schlussfolgerungen zu ziehen. Die jeweilige

[2] Für zehn Lehrkräfte liegen keine demographischen Angaben vor.

Testleistung ergibt sich als Summenwert aus der Anzahl der richtig beantworteten Items ($M_{Klasse8}$ = 17,3; $SD_{Klasse8}$ = 2,9; $Min_{Klasse8}$ = 10; $Max_{Klasse8}$ = 24; $M_{Klasse9}$ = 17,3; $SD_{Klasse9}$ = 3,8; $Min_{Klasse9}$ = 8; $Max_{Klasse9}$ = 26). Die interne Konsistenz der eingesetzten Lesekompetenztests erwies sich als akzeptabel ($\alpha_{Klasse8}$ = .65; $\alpha_{Klasse9}$ = .71). Die Testleistung der Schülerinnen und Schüler korreliert zu r = -.30 (p < .01) mit der Deutschnote im Halbjahreszeugnis des jeweiligen Schuljahrs.

Wissensgrundlagen der Lehrkräfte

Der eingesetzte Test zur Erfassung der Wissensgrundlagen im Bereich Textverstehen wurde im Rahmen des BiKS-Teilprojektes zur Diagnostischen Kompetenz als Dissertationsprojekt (Matthäi, in Vorb.) entwickelt. Der Test gliedert sich in drei thematische Bereiche, die wiederum jeweils in Untertests gegliedert sind:

- Schwierigkeitsgenerierende Merkmale von Texten mit den Untertests
  - Texte einschätzen
  - Texte umschreiben
  - Texte anstreichen.
- Schwierigkeitsgenerierende Merkmale von Aufgaben mit den Untertests
  - Aufgaben einschätzen
  - Aufgabenpaare.
- Angemessene Strategien beim Bearbeiten von Textverstehensaufgaben mit vier Szenarien als Untertests.

Die einzelnen Untertests werden im Folgenden einzeln beschrieben (siehe dazu auch Tabelle 1).

Schwierigkeitsgenerierende Merkmale von Texten. Zur Erfassung des Wissens über schwierigkeitsgenerierende Merkmale von Texten wurden drei Untertests eingesetzt:
Für den Untertest *Texte einschätzen* sollten die Lehrkräfte fünf vorgegebene Texte (86 bis 109 Wörter pro Text) in eine Schwierigkeitsrangreihe bringen. Die von der Lehrkraft vorgenommene Reihung der Texte wurde dann verglichen mit einer Schwierigkeitsrangreihe, die anhand von zentralen, theoretisch hergeleiteten Kennwerten der Textschwierigkeit auf Wort-, Satz- und Textebene erstellt wurde (Wortlänge in Buchstaben, Anteil der Wörter mit mindestens sieben Buchstaben, Anteil der Wörter, die nicht im Findefix (Fackelmann, Müller, Patho & Patho, 2006) stehen, Anteil der Nomen, Satzlänge in Wörtern, Anteil an Nebensätzen, Anteil des erweiterten Infinitivs mit ‚zu', Type-Token-Relation der Nomen, Anteil der Pronomen, Anteil der Konnektoren). Hierzu wurde für jeden für einen Text vergebenen Rang ein Paarvergleich mit der Musterlösung vorgenommen. Wurde Text 2 bspw. als schwieriger als Text 1, aber leichter als die Texte 3,

4 und 5 bewertet, und entsprach das der Musterlösung, so ergab das vier Punkte für die Lehrkraft. Analog wurde für jeden anderen Text verfahren, so dass in diesem Untertest maximal 20 Punkte zu erzielen waren. Die interne Konsistenz der Skala Texte einschätzen ist akzeptabel (Cronbachs α = .70). Um die Summenwerte der einzelnen Untertests untereinander vergleichbar zu machen, wurden die von der Lehrkraft erzielten Punkte jeweils durch die Anzahl der maximal zu erreichenden Punkte dividiert und mit 100 multipliziert.

Im Untertest *Texte umschreiben* sollten zwei vorgegebene Texte (75 bzw. 101 Wörter) derart umgeschrieben werden, dass sie für leseschwache Schüler leichter verständlich werden. Die von den Lehrkräften produzierten Texte wurden in Bezug auf sieben Kennwerte der Textschwierigkeit auf Wort-, Satz- und Textebene mit den vorgegebenen Texten verglichen und auf dieser Basis bewertet: Geringere durchschnittliche Wort- bzw. Satzlängen, weniger Nebensätze, weniger Substantive und eine geringere Type-Token-Relation wurden als positiv für die Verständlichkeit des umgeschriebenen Textes angesehen und jeweils mit einem Punkt bewertet. Höhere Anteile an Pronomen und Konnektoren im Text tragen ebenfalls zur besseren Verständlichkeit bei und wurden entsprechend auch mit jeweils einem Punkt bewertet. Insgesamt konnten in diesem Untertest 14 Punkte erzielt werden. Analog zu den anderen Untertests wurden auch hier die erreichten Punkte in eine Prozentmetrik gebracht. Die interne Konsistenz der Skala Texte umschreiben ist befriedigend (Cronbachs α = .77).

Der Untertest *Texte anstreichen* beinhaltete die Aufgabe, in einem vorgegebenen Text (208 Wörter) diejenigen Stellen anzustreichen, die bei leseschwachen Schülerinnen und Schülern zu Verständnisproblemen führen könnten. Ausgewertet wurde hier, ob die Lehrkräfte im Text die unpassende Überschrift, schwierige Wörter, den langen Satz und doppelte Pronomen als schwierigkeitsgenerierende Merkmale erkannt haben. Für jedes erkannte schwierigkeitsgenerierende Merkmal erhielt die Lehrkraft einen Punkt, so dass maximal vier Punkte erreichbar waren. Analog zu den anderen Untertests wurden auch hier die erreichten Punkte in eine Prozentmetrik gebracht. Die interne Konsistenz der Skala Texte anstreichen ist allerdings alles andere als zufriedenstellend (Cronbachs α = .15).

Schwierigkeitsgenerierende Merkmale von Aufgaben. Zur Erfassung des Wissens über schwierigkeitsgenerierende Merkmale auf der Aufgabenebene bearbeiteten die Lehrkräfte zwei Untertests:

Die Anforderung im Untertest *Aufgaben einschätzen* bestand darin, für sieben Aufgaben zu einem vorgegebenen Text jeweils einzuschätzen, wie schwierig diese für Schüler der 6. bzw. 7. Klassenstufe sind („Bitte geben Sie … an, welcher Anteil der Schüler (in Prozent)

einer durchschnittlichen Klasse der jeweiligen Klassenstufe Ihrer Meinung nach die Fragen vermutlich richtig beantworten wird."). Für diese Aufgaben lagen die empirisch gefundenen Lösungshäufigkeiten aus den entsprechenden Erhebungen im Rahmen der BiKS-Studie vor. Für die 14 einzuschätzenden Aufgaben wurde jeweils der Betrag der Differenz zwischen der empirisch gefundenen Aufgabenschwierigkeit und der von der Lehrkraft eingeschätzten Schwierigkeit berechnet. Das so ermittelte globale Abweichungsmaß drückt aus, in welchem Umfang sich die Lehrkraft verschätzt hat. Die Richtung der Verschätzung bleibt dabei jedoch unberücksichtigt. Die Summe der Abweichungen in Prozent wird anschließend von 100 abgezogen, so dass auch hier Vergleichbarkeit mit den anderen Untertests auf der Prozentmetrik gegeben ist. Je genauer die Einschätzung über die Aufgaben hinweg ist, desto höher ist der erzielte Wert. Die interne Konsistenz der Skala ist gut (Cronbachs $\alpha = .87$).

Außerdem sollte im Untertest *Aufgabenpaare* bei neun Fragenpaaren eingeschätzt werden, welche der beiden Aufgaben jeweils die schwierigere ist. Die Konstruktion der einfachen bzw. schwierigen Aufgaben folgte dem Schema der schwierigkeitsgenerierenden Merkmale von Aufgaben (Kirsch, 2001). Zur Auswertung wurde die Lehrereinschätzung jeweils mit einem Expertenrating verglichen. Für jeden korrekten Vergleich wurde ein Punkt vergeben, so dass maximal neun Punkte erreichbar waren. Analog zu den anderen Untertests wurden auch hier die erreichten Punkte in eine Prozentmetrik gebracht. Die interne Konsistenz des Untertests Aufgabenpaare ist deutlich zu niedrig (Cronbachs $\alpha = .13$).

Strategien beim Bearbeiten von Textverstehensaufgaben. Zur Erfassung des Wissens über adäquate Strategien beim Textverstehen wurden im Untertest *Szenarien* vier Szenarien vorgegeben, in denen die Art des zu bearbeitenden Textes sowie die an die Schüler gerichteten Anforderungen bei der Arbeit mit dem Text genannt wurden. In Anlehnung an die Vorgehensweise des WLST (Schlagmüller & Schneider, 2007) und auf Basis weiterer Vorarbeiten (Artelt, Beinicke, Schlagmüller &Schneider, 2009; Neuenhaus, 2011) sollten die Lehrkräfte bei jedem der Szenarien fünf bzw. sechs vorgeschlagene Vorgehensweisen danach beurteilen, wie hilfreich diese jeweils sind, um die gestellten Anforderungen zu erfüllen (von (1) „sehr hilfreich" bis (6) „überhaupt nicht hilfreich"). Das Lehrerwissen in diesem Bereich wurde quantifiziert als Ausmaß der Übereinstimmung der Urteile der Lehrkräfte mit den Urteilen von Experten. Für den Abgleich wurden jeweils Paare von Strategievorschlägen gebildet, die immer einen guten und einen schlechten Vorschlag beinhalteten. Pro Paarvergleich wurde anschließend die Differenz aus der Ziffernbewertung der einzelnen Strategien gebildet. Wurde die bessere Strategiealternative auch als hilfreicher bewertet, erhielten die Befragten einen Punkt. Negative Differenzen

wurden nicht gewertet (0 Punkte). Insgesamt gingen 31 Paarvergleiche in die Bewertung ein. Die interne Konsistenz des Untertests Szenarien kann als befriedigend angesehen werden (Cronbachs $\alpha = .70$).

Für den Bereich Texte und den Bereich Aufgaben wurden Mittelwerte aus den drei bzw. zwei Untertests gebildet. Die erreichten Punkte bei den vier Szenarien wurden auf Basis der Prozentmetrik ebenfalls gemittelt zu einem Untertest zusammengefasst. Der Gesamtscore des Wissenstest ist das ungewichtete arithmetische Mittel der sechs Untertests.

Globale Urteilsgüte

Die globale Einschätzung, wie gut eine Schülerin oder ein Schüler im Textverstehen jeweils im Vergleich zu einem durchschnittlichen Schüler der jeweiligen Klassenstufe ist, wurde von den Lehrkräften mittels einer fünfstufigen Ratingskala von (1) „sehr schwach im Textverstehen“ über (3) „durchschnittlich“ bis (5) „sehr gut im Textverstehen“ vorgenommen. Anhand dieser globalen Einschätzung wurde die Rangkomponente der Urteilsgenauigkeit (Schrader, 1989) berechnet. Hierfür wurden für jede Lehrkraft auf Klassenebene Korrelationskoeffizienten nach Pearson zwischen den Lehrerurteilen und den Schülertestleistungen berechnet. Eine perfekte Einschätzung der Reihenfolge der einzuschätzenden Schülerinnen und Schüler geht mit einem Koeffizienten von $r = 1.0$ einher. Um mit dieser Lehrervariablen weitere Berechnungen durchführen zu können, wurden die Koeffizienten mittels der Fisher-Z-Transformation in intervallskalierte Werte überführt. Allerdings ergaben sich für drei Lehrkräfte Werte von $r = 1.0$ bzw. $r = -1.0$. Diese Werte sind per definitionem nicht Fisher-Z-transformierbar. Um diese Lehrkräfte für die späteren Berechnungen nicht ausschließen zu müssen, wurden die Extremwerte in moderatere Werte umcodiert. In Anlehnung an die von Lipsey und Wilson (2001) vorgeschlagene Vorgehensweise wurde zunächst der αMittelwert und die Standardabweichung der Fischer-Z-transformierten Werte berechnet, ohne die nicht transformierbaren Werte von $|r| = 1$ einzubeziehen ($Z_r = .46$; $SD = 0{,}69$). Die nicht transformierbaren Werte wurden anschließend basierend auf dieser Berechnung umcodiert, so dass sie Werte von ±3 Standardabweichungen über bzw. unter dem Mittelwert annehmen. Für $r = 1.0$ ergibt sich somit ein Wert von $Z_r = 2.53$ und für $r = -1.0$ ein Wert von $Z_r = -1.61$. Das globale Lehrerurteil korreliert zu $r = -.58$ ($p < .01$) mit der von der Lehrkraft vergebenen Deutschnote im Halbjahreszeugnis der jeweiligen Klassenstufe.

Aufgabenspezifische Urteilsgüte

Für die aufgabenspezifischen Einschätzungen wurde den Lehrkräften ein expositorischer Text aus dem bei den Schülerinnen und Schülern administrierten Lesekompetenztest zusammen mit den sieben Fragen, die die Schülerinnen und Schülern zu diesem Text zu bearbeiten hatten, vorgelegt. Die Lehrkräfte sollten für jede dieser Aufgaben – ohne die Antworten ihrer Schülerinnen und Schüler zu kennen – einschätzen, ob der jeweilige Schüler bzw. die Schülerin diese Aufgabe im Test richtig gelöst hat oder nicht („Wir möchten Sie bitten, einzuschätzen, ob dieser Schüler/diese Schülerin die Fragen des Ihnen vorliegenden Lesetests aller Wahrscheinlichkeit nach korrekt beantworten konnte."). Zur Auswertung der aufgabenspezifischen Urteilsgüte wurden die dichotom erfassten Lehrerurteile (ja/nein) zu den einzelnen Aufgaben mit den jeweiligen dichotomisierten Schülerantworten im Test (korrekt beantwortet/falsch beantwortet) verglichen. Daraus wurden für jede Lehrkraft drei Kennwerte der Urteilsgenauigkeit gebildet: Der aufgabenspezifische Treffer bezieht sich auf die Anzahl der exakten Übereinstimmungen zwischen Schülerleistung (gelöst bzw. nicht gelöst) und Lehrerurteil (gelöst bzw. nicht gelöst). Der aufgabenspezifische Nichttreffer (Überschätzung) bezieht sich auf die Anzahl der Aufgaben, die von den Schülerinnen und Schülern falsch beantwortet wurden, jedoch von der Lehrkraft als richtig gelöst eingeschätzt wurden. Der aufgabenspezifische Nichttreffer (Unterschätzung) bezieht sich auf die Anzahl der Aufgaben, die von den Schülerinnen und Schülern richtig gelöst wurden, jedoch von der Lehrkraft als falsch gelöst eingeschätzt wurden (vgl. Karing, Matthäi, & Artelt, 2011). Um zu vergleichbaren Aussagen über die Urteilsgüte der einzelnen Lehrkräfte zu gelangen, wurde der jeweilige Anteil der Treffer und Nichttreffer jeder Lehrkraft an den von diesen jeweils insgesamt eingeschätzten Aufgaben ermittelt. Bei einer Lehrkraft, die bei allen von ihr eingeschätzten Schülerinnen und Schülern alle Aufgaben korrekt eingeschätzt hat, liegt die aufgabenspezifische Trefferquote entsprechend bei 1 und die Quote der überschätzten bzw. unterschätzten Aufgaben jeweils bei 0.

Ergebnisse

Fragestellung 1: Urteilsgüte von Deutschlehrkräften im Bereich Textverstehen

Globale Urteilsgüte

Für die Rangkomponente der globalen Urteilsgüte auf der Klassenebene ergab sich eine durchschnittliche Korrelation von $\bar{r}$= .50 (95 % KI [.24; .76]). Allerdings zeigten sich große interindividuelle Unterschiede zwischen den Lehrkräften (SD = 0,86). Bei sieben Lehrkräften war eine negative Korrelation zu beobachten.

Aufgabenspezifische Urteilsgüte

Die durchschnittliche Trefferquote bei den aufgabenspezifischen Einschätzungen lag bei M = 0,53 (SD = 0,11; 95 % KI [0,50; 0,57]). Die Lehrkräfte schätzten im Durchschnitt also 53 % der Schülerantworten korrekt ein. Bei den mit der Trefferquote korrespondierenden Nichttreffern zeigte sich, dass der Anteil der überschätzten Aufgaben (M = 0,35; SD = 0,11; 95 % KI [0,31; 0,38]) signifikant größer (t = 8,78; df = 40; p < .01) war als der Anteil der unterschätzten Aufgaben (M = 0,12; SD = 0,08; 95 % KI [0,09; 0,14]).

Fragestellung 2: Wissen der Lehrkräfte im Bereich Textverstehen[3]

Die Lehrkräfte erzielten im gesamten Wissenstest durchschnittlich mehr als die Hälfte der erreichbaren Punkte (M = 60,5; SD = 7,6). Für einen Überblick über die Ergebnisse des Wissenstests und über die Interkorrelationen zwischen den Bereichen des Tests siehe Tabellen 1 und 2. Im Bereich der Textschwierigkeit wurden insgesamt M = 50,7 % (SD = 10,8) der Anforderungen richtig bearbeitet. Die Anforderung im Untertest Texte einschätzen bestand darin, Texte in eine Schwierigkeitsreihenfolge zu bringen. Im Schnitt lagen die Lehrkräfte bei M = 74,5 % (SD = 12,9) der Vergleiche richtig. Im Untertest Texte umschreiben ging die Anzahl der Parameter, an denen die Lehrkraft den vorgegebenen Text in eine leichtere Richtung verändert hat, in die Bewertung ein (z. B. eine geringere Anzahl von Nebensätzen im umgeschriebenen Text). Die Lehrkräfte nahmen im Schnitt 42,3 % der intendierten Veränderungen vor (SD = 14,6). Im Untertest Texte anstreichen konnten die Lehrkräfte durchschnittlich M = 33,6 % (SD = 21,8) der schwierigkeitsgenerierenden Textmerkmale identifizieren.

Im Bereich der Aufgabenschwierigkeit wurden insgesamt M = 61,0 % (SD = 8,7) der Aufgaben richtig gelöst. Durchschnittlich wichen die Lehrereinschätzungen der Aufgabenschwierigkeit um 17,4 % (SD = 8,7) von den empirisch gefundenen Aufgabenschwierigkeiten ab. Dies entspricht einer Erfolgsquote im Untertest Aufgaben

[3] Den hier berichteten Ergebnissen zum Lehrerwissen im Bereich Textverstehen liegen die Daten aller getesteten Lehrkräfte zugrunde (N = 77).

einschätzen von $M = 82{,}6$ % ($SD = 8{,}7$). Im Schnitt schätzten die Lehrkräfte 3,6 der neun Paarvergleiche im Untertest Aufgabenpaare richtig ein, dies entspricht $M = 40{,}0$ % ($SD = 15{,}6$) der dort zu erzielenden Punkte.

Bei der Bearbeitung der vier Szenarien im Bereich Strategien lagen die Lehrkräfte in $M = 69{,}4$ % ($SD = 14{,}9$) der Fälle richtig.

Tabelle 1. Deskriptive Statistik für den Gesamtscore sowie für die Bereiche und Untertests im Test zu Wissensgrundlagen diagnostischer Kompetenz von Lehrkräften im Bereich Textverstehen

Wissensbereich	***M***	**95% KI**	***SD***	***Min***	***Max***
Texte gesamt ($N = 74$)	50,7	[48,2; 53,2]	10,8	21,4	76,4
Texte einschätzen ($N = 76$)	74,5	[71,5; 77,4]	12,9	40,0	100,0
Texte umschreiben ($N = 76$)	42,3	[39,0; 45,6]	14,6	7,1	64,3
Texte anstreichen ($N = 76$)	33,6	[28,6; 38,5]	21,8	0,0	75,0
Aufgaben gesamt ($N = 74$)	61,0	[59,0; 63,0]	8,7	43,3	80,8
Aufgaben einschätzen ($N = 74$)	82,6	[80,6; 84,6]	8,7	46,6	92,8
Aufgabenpaare ($N = 75$)	40,0	[36,4; 43,6]	15,6	0,0	77,8
Strategien ($N = 72$)	69,4	[65,9; 72,9]	14,9	24,7	94,1
Gesamtscore ($N = 69$)	60,5	[58,7; 62,3]	7,6	44,6	76,6

Anmerkungen: KI = Konfidenzintervall:[Obergrenze; Untergrenze]

Tabelle 2. Interkorrelationen der Untertests, Bereiche und des Gesamtscores im Test zu Wissensgrundlagen diagnostischer Kompetenz von Lehrkräften im Bereich Textverstehen

	Wissensbereich	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	**Texte gesamt** ($N = 74$)								
2	**Texte einschätzen** ($N = 76$)	.47**							
3	**Texte umschreiben** ($N = 76$)	,68**	.20#						
4	**Texte anstreichen** ($N = 76$)	.78**	.08	.24*					
5	**Aufgaben gesamt** ($N = 74$)	.07	.24*	.00	.01				
6	**Aufgaben einschätzen** ($N = 74$)	.07	.21#	.12	.08	.44**			
7	**Aufgabenpaare** ($N = 75$)	.04	.13	-.08	-.03	.87**	-.07		
8	**Strategien** ($N = 72$)	.06	.07	.19	-.07	.24*	.19	.17	
	Gesamtscore ($N = 69$)	.55**	.35**	.47**	.33**	.57**	.31*	.48**	.78**

Anmerkungen: ** $p < .01$, * $p < .05$, # $p < .10$

Fragestellung 3: Zusammenhang des Lehrerwissens mit der Urteilsgüte

Zur Beantwortung der Fragestellungen zum Zusammenhang des Lehrerwissens mit der Urteilsgüte wurden die Wissensgrundlagen der Lehrkräfte jeweils mit den Indikatoren der globalen und der aufgabenspezifischen Urteilsgüte korreliert. Als Signifikanzniveau wurde $\alpha = 5$ % gewählt. Der mittels Bonferroni-Korrektur ermittelte α'-Wert betrug $\alpha' = .001$.

Wissensgrundlagen und globale Urteilsgüte

Es zeigen sich keine signifikanten und substanziellen Zusammenhänge zwischen dem Lehrerwissen im Bereich Textverstehen und der Güte der globalen Lehrerurteile. Weder bezogen auf den Gesamtscore ($r = -.26$; n.s.), noch bezogen auf die einzelnen Bereiche des Text-, Aufgaben- und Strategiewissens sind die Zusammenhänge signifikant, der Tendenz nach jedoch eher negativ (vgl. Tabelle 3).

Wissensgrundlagen und aufgabenspezifische Urteilsgüte

Auch die Güte der aufgabenspezifischen Urteile hängt nicht mit den im Test diagnostizierten Wissenskomponenten zusammen. Weder für den Gesamtscore des Wissenstests($r = .08$; n.s.), noch für die einzelnen Bereiche und Untertests zeigen sich substanzielle und signifikante Zusammenhänge mit der Trefferquote (vgl. Tabelle 3). Die Betrachtung des Zusammenhangs zwischen den beiden Indikatoren des aufgabenspezifischen Nichttreffers und dem Lehrerwissen ergibt ein uneinheitliches Bild. Es wurde erwartet, dass mit umfassenderem Wissen akkuratere Urteile getroffen werden. Je höher das Wissen, umso geringer wäre dann die Unter- bzw. Überschätzung. Auch hier finden sich jedoch keine substanziellen und signifikanten Zusammenhänge zwischen der Urteilsgüte und dem Gesamtscore des Wissenstests sowie den einzelnen Wissensbereichen.

Tabelle 3. Korrelationen zwischen den Wissensgrundlagen und den Indikatoren der Urteilsgüte ($N = 44$ Lehrkräfte)

Wissensbereich	Globale Urteilsgüte	Aufgaben-spezifische Trefferquote	Quote der überschätzten Aufgaben	Quote der unterschätzten Aufgaben
Texte gesamt ($N = 74$)	-.13	.10	-.25	.24
Texte einschätzen ($N = 76$)	.04	.20	-.14	-.04
Texte umschreiben ($N = 76$)	.05	-.02	-.14	.22
Texte anstreichen ($N = 76$)	-.16	-.01	-.06	.12
Aufgaben gesamt ($N = 74$)	-.17	-.05	.05	-.02
Aufgaben einschätzen ($N = 74$)	.08	.05	.29	-.46
Aufgabenpaare ($N = 75$)	-.24	-.07	-.13	.24
Strategien ($N = 72$)	-.15	.08	-.13	.08
Gesamtscore ($N = 69$)	-.26	.08	-.24	.24

Anmerkungen: Nach vorgenommener Bonferroni-Korrektur ist keine der Korrelationen auf dem korrigierten 5%-Niveau ($\alpha' = .001$) signifikant.

Diskussion

In der vorliegenden Arbeit wurde untersucht, wie akkurat Deutschlehrkräfte der Sekundarstufe Schülerfähigkeiten im Bereich Textverstehen auf globaler und aufgabenspezifischer Ebene einschätzen, über welches Wissen sie im Bereich Textverstehen verfügen und wie die Urteilsgüte der Lehrkräfte mit diesem Wissen zusammenhängt. Die Ergebnisse sind hinsichtlich der Urteilsgüte und der gefundenen interindividuellen Unterschiede zwischen den Lehrkräften insgesamt vergleichbar mit den in der Literatur berichteten Ergebnissen zur Güte globaler Einschätzungen (siehe z. B. Südkamp et al., 2012), sowie mit den weniger häufig berichteten Ergebnissen zur Urteilsgüte von aufgabenbezogenen Einschätzungen (Artelt & Rausch, 2014; Karing, Matthäi & Artelt, 2011). Aufgrund von (inhaltlichen) Restriktionen hinsichtlich der Mindestanzahl der von jeder Lehrkraft eingeschätzten Schülerinnen und Schüler verkleinerte sich die Stichprobe sehr stark. Vor dem Hintergrund der plausiblen Annahme, dass die Präzision der Urteilsgüte mit der Anzahl der eingeschätzten Schülerinnen und Schüler variiert, kann hierin eine Einschränkung gesehen werden, die auch auf die Zusammenhänge zwischen Urteilsgüte und Lehrerwissen Einfluss haben kann.

Die Lehrkräfte bearbeiteten im gesamten Wissenstest zwischen 45 % und 77 % der Anforderungen korrekt. In der vorliegenden Stichprobe wird somit bezogen auf den Gesamtscore nur ein Teil der Bandbreite des Tests abgebildet. Allerdings relativiert sich dieses Bild mit Blick auf die einzelnen Bereiche und Untertests, die durchaus stärker zwischen den Lehrkräften differenzieren (siehe Tabelle 1). Ein Vergleich der Testergebnisse von Lehrkräften einer gesonderten Stichprobe, die von den jeweiligen Schulleitern als Experten benannt wurden, mit den Testergebnissen von Lehrkräften und

Lehramtsstudierenden zeigt darüber hinaus, dass hier teilweise deutliche Unterschiede zwischen diesen Gruppen zu verzeichnen sind (Matthäi, in Vorb.).

Zentrales Anliegen der vorliegenden Arbeit war die Beantwortung der Frage, ob und ggf. wie das Wissen von Lehrkräften im Bereich Textverstehen mit der Urteilsgüte bei der Einschätzung von Schülerinnen und Schülern hinsichtlich deren Leistung im Textverstehen zusammenhängt. Ausgehend von den aus der Theorie hergeleiteten Annahmen zu (notwendigem) Lehrerwissen im Bereich des Textverstehens wurde erwartet, dass umfassendes Wissen der Lehrkräfte mit akkurateren Urteilen über Schülerleistungen einhergeht. Dies sollte insbesondere für die aufgabenbezogenen Urteile gelten, da hier davon ausgegangen wird, dass die Einschätzung, ob ein Schüler oder eine Schülerin eine bestimmte Aufgabe löst, auch abhängig ist von der (korrekten) Einschätzung der jeweiligen Aufgabenschwierigkeit durch die Lehrkraft (Karing, Matthäi & Artelt, 2011). Insofern muss hier stärker auf fachspezifisches Wissen zurückgegriffen werden, um zu einem akkuraten Urteil zu gelangen, als es bei globalen Einschätzungen der Fall ist. Weder für den Gesamtscore des Wissenstests noch für einzelne Bereiche konnten jedoch substanzielle Zusammenhänge mit den Indikatoren der globalen und aufgabenspezifischen Urteilsgüte gefunden werden. Für dieses Befundmuster kann es sicher mehrere Gründe geben. Die Ergebnisse scheinen zunächst nahezulegen dass es keinen Automatismus zwischen dem Wissen und seiner Anwendung gibt. Dies kann mit mangelnder Motivation der Lehrkräfte zu tun haben, Urteile zu jedem einzelnen Schüler und jeder einzelnen Schülerin in Bezug auf sehr konkrete Anforderungen zu fällen, ohne hierfür irgendeine Rückmeldung zu bekommen.

Außerdem wissen wir nur wenig über die Beurteilungssituation, in der die Lehrkräfte in der vorliegenden Untersuchung ihre Urteile über ihre Schülerinnen und Schüler getroffen haben, sowie über die Strategien der Lehrkräfte zur Eindrucks- und Urteilsbildung in der der Beurteilungssituation vorausgehenden Zeit im Unterricht. Die hohe Korrelation zwischen der Deutschnote und dem globalen Urteil der Lehrkräfte kann dahingehend interpretiert werden, dass die Lehrkräfte sich bei der Beurteilung eher an distalen Merkmalen wie bspw. der Deutschnote orientieren. Jedoch repräsentiert der als Außenkriterium für die Urteilsgüte dienende Lesekompetenztest nur einen speziellen Teil des Deutschunterrichts, was u. a. auch an der geringeren Korrelation zwischen der Testleistung und der Deutschnote abgelesen werden kann. Um akkurate aufgabenspezifische Urteile abgeben zu können, müssen im Urteilsprozess die jeweiligen Textverstehensanforderungen der Items und Texte vor dem Hintergrund der erwarteten Schülerfähigkeit berücksichtigt werden. Wenn hier von den Lehrkräften Abkürzungen bzw.

Heuristiken verwendet werden, ist eine substanzielle Korrelation auch hier nicht sehr wahrscheinlich (vgl. Krolak-Schwerdt, Böhmer & Gräsel, 2009). Dies ist insbesondere vor dem Hintergrund der Forschung zur sozialen Urteilsbildung relevant, die gezeigt hat, dass z.B. die Beurteilungsziele (Krolak-Schwerdt, Böhmer & Gräsel, 2012) oder auch die von der Lehrkraft empfundene Verantwortlichkeit für das Urteil (Krolak-Schwerdt, Böhmer & Gräsel, 2013) die Vorgehensweise bei der Urteilsbildung und die Güte der Urteile beeinflussen.

Aber auch die Art der Operationalisierung des Wissenstests kann mit den gefundenen Ergebnissen zusammenhängen. Die Konstruktion des Tests hatte insbesondere das Ziel, aufbauend auf vielfältige und interdisziplinäre Theorien zu überprüfen, über welches Wissen im Bereich Textverstehen Lehrkräfte verfügen. Es könnte sein, dass die dadurch bedingte Auswahl der Inhaltsbereiche und Anforderungen im Test insofern einseitig war, als dass die überprüften Wissensbereiche in der konkreten Diagnosesituation für die Lehrkräfte nicht lösungsrelevant waren und insofern nicht genutzt werden konnten. Die Interkorrelationen zwischen den Untertests und Bereichen des Wissenstests und auch die internen Konsistenzen legen weiterhin nahe, dass es sich beim erfassten Wissen nicht um ein einheitliches Konstrukt handelt, sondern die gestellten Anforderungen durchaus spezifisch sind. Es kann also nicht von einem Indikator der Wissensgrundlagen gesprochen werden, der als Grundlage akkurater diagnostischer Urteile in diesem Bereich gesehen werden kann. Vielmehr muss wohl von einer Reihe von teils sehr spezifischen Unteranforderungen ausgegangen werden, die in Bezug auf die konkrete Aufgabe, die Leistungen der Schülerinnen und Schüler einzuschätzen von nachgeordneter Relevanz war.

Weiterer Forschungsbedarf kann vor allem hinsichtlich der Frage der konkreten Vorgehensweise der Lehrkräfte bei der Urteilsbildung festgestellt werden. Welche Wissenselemente und Heuristiken in der konkreten Beurteilungssituation – auch in Abhängigkeit von den an die Lehrkräfte gestellten Urteilsanforderungen – genutzt werden, und wie diagnostische Prozesse bei der Selektion und Integration der Informationen über Schülerinnen und Schüler aussehen, sollte in weiteren Untersuchungen verstärkt thematisiert werden.

Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: AR301/6-1, AR301/6-2 und AR301/6-3) im Bamberger Forschungsprojekt BiKS (Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (FOR 543)).

Literatur

Artelt, C. (2009). Diagnostische Urteile von Lehrkräften im Bereich der Lesekompetenz. In A. Bertschi-Kaufmann & C. Rosebrock (Hrsg.), *Literalität. Bildungsaufgabe und Forschungsfeld* (S. 125–136). Weinheim: Juventa.

Artelt, C., Beinicke, A., Schlagmüller, M. & Schneider, W. (2009). Diagnose von Strategiewissen beim Textverstehen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 41*, 96–103.

Artelt, C., McElvany, N., Christmann, U., Richter, T., Groeben, N., Köster, J. et al. (2005). Expertise – Förderung von Lesekompetenz. Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF).

Artelt, C. & Rausch, T. (2014). Accuracy of teacher judgments. When and for what reasons? In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Hrsg.), *Teachers' professional development: Assessment, training, and learning* (S. 27–43). Rotterdam: Sense Publishers.

Artelt, C. & Schlagmüller, M. (2004). Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz – Vertiefende Analysen im Rahmen von PISA 2000* (S. 169–196). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J. & Köller, O. (1996). Lernstrategien und schulische Leistungen. In J. Möller & O. Köller (Hrsg.), *Emotionen, Kognitionen und Schulleistung* (S. 137–154). Weinheim: Psychologie Verlags Union.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*, 469–520.

Brunner, M., Anders, Y., Hachfeld, A. & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum & U. Klusmann (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 215–234). Münster: Waxmann.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C. & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*, 247–265.

Fackelmann, J., Müller, R., Patho, K. & Patho, S. (2006). *Findefix. Wörterbuch für die Grundschule*. München: Oldenbourg.

Gilabert, R., Martinez, G. & Vidal-Abarca, E. (2005). Some good texts are always better: Text revision to foster inferences of readers with high and low prior background knowledge. *Learning and Instruction, 15*, 45–68.

Graesser, A. C., McNamara, D. S. & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. Sweet & C. E. Snow (Hrsg.), *Rethinking reading comprehension* (S. 82 –98). New York: Guilford.

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

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

Grzesik, J. (1990). *Textverstehen lernen und lehren. Geistige Operationen im Prozess des Textverstehens und typische Methoden für die Schulung zum kompetenten Leser*. Stuttgart: Klett.

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

Hofer, M. (1981). Schülergruppierungen in Urteil und Verhalten des Lehrers. In M. Hofer (Hrsg.), *Informationsverarbeitung und Entscheidungsverhalten von Lehrern. Beiträge zu einer Handlungstheorie des Unterrichtens* (S. 192–222). München: Urban & Schwarzenberg.

Hoge, R. D. & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*, 297–313.

Hopkins, K. D., George, C. A. & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*, 177–182.

Hörstermann, T., Krolak-Schwerdt, S. & Fischbach, A. (2010). Die kognitive Repräsentation von Schülertypen bei angehenden Lehrkräften. Eine typologische Analyse. *Schweizerische Zeitschrift für Bildungswissenschaften, 32*, 143–158.

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Pädagogische Psychologie, 23*, 197–209.

Karing, C., Matthäi, J. & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? *Zeitschrift für Pädagogische Psychologie, 25*, 159– 172.

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

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

Krauss, S., Baumert, J. & Blum, W. (2008). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *The International Journal on Mathematics Education, 40*, 873– 892.

Krolak-Schwerdt, S., Böhmer, M. & Gräsel, C. (2013). The impact of accountability on teachers' assessments of student performance: A social cognitive analysis. *Social Psychology of Education, 16*, 215– 239.

Krolak-Schwerdt, S., Böhmer, M. & Gräsel, C. (2012). Leistungsbeurteilungen von Schulkindern. Welche Rolle spielen Ziele und Expertise der Lehrkraft? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44*, 111– 122.

Krolak-Schwerdt, S., Böhmer, M. & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess. Der Lehrer als 'flexibler Denker'. *Zeitschrift für Pädagogische Psychologie, 23*, 175–186.

Lange, K., Kleickmann, T., Tröbst, S. & Möller, K. (2012). Fachdidaktisches Wissen von Lehrkräften und multiple Ziele im naturwissenschaftlichen Sachunterricht. *Zeitschrift für Erziehungswissenschaft, 15*, 55–75.

Langer, I., Schulz von Thun, F. & Tausch, R. (2006). *Sich verständlich ausdrücken*. München: Reinhard.

Lipsey,M. W. &Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Lorenz, C. & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie, 23*, 211–222.

Matthäi, J. (in Vorbereitung). *Wissensgrundlagen diagnostischer Kompetenz im Bereich des Textverstehens*. Dissertation, Universität Bamberg, Bamberg.

McNamara, D. S. & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247–288.

Mosenthal, P. & Kirsch, I. (1994). *Defining the proficiency standards of adult literacy in the U.S.: A profile approach*. Paper presented at the National Reading Conference, San Diego.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Neuenhaus, N. (2011). *Metakognition und Leistung. Eine Längsschnittuntersuchung in den Bereichen Lesen und Englisch bei Schülerinnen und Schülern der fünften und sechsten Jahrgangsstufe*. Dissertation, Universität Bamberg.

O'Reilly, T. & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, highknowledge readers. *Discourse Processes, 43*, 121–152.

Pfost, M. & Artelt, C. (2013). Reading literacy development in secondary school and the effect of differential institutional learning environments. In M. Pfost, C. Artelt & S. Weinert (Hrsg.), *The development of reading literacy from early childhood to adolescence. Empirical findings from the Bamberg BiKS longitudinal studies*. Bamberg: University of Bamberg Press.

Schlagmüller, M. & Schneider, W. (2007). *WLST 7–12. Würzburger Lesestrategie-Wissenstest für die Klassen 7–12*. Göttingen: Hogrefe.

Schneider, W. & Pressley, M. (1997). *Memory development between two and twenty*. Mahwah, NJ: Erlbaum.

Schrader, F.-W. (2010). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 102–108). Göttingen: Hogrefe.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Frankfurt am Main: Peter Lang.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 19*, 85 –95.

Südkamp, A., Kaiser, J. &Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743–762.

Turner, A. & Greene, E. (1977). *The construction and use of a propositional text base*. Colorado: Institute for the Study of Intellectual Behavior.

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

Weinstein, C. E. & Mayer, R. E. (1986). The teaching of learning strategies. In M. C. Wittrock (Hrsg.), *Handbook of research on teaching* (S. 315–327). New York: Macmillan Publishing Company.

Teacher judgment accuracy and assessment strategies in a Simulated Classroom

Tobias Rausch und Cordula Artelt
Otto-Friedrich-Universität Bamberg

Abstract. Interrelations between preservice teachers' assessment strategies, their knowledge about task difficulties and their judgment accuracy was investigated in an experimental study using the Simulated Classroom approach. 57 participants were provided with tasks varying in their diagnostic relevance between two conditions, for assessing simulated students' mathematical abilities. Preservice teachers' assessment strategies differentially affected judgment accuracy, with a task-focused strategy being related to higher accuracy outcomes when a broad variety of tasks difficulties could be used for the assessment, while a student-focused strategy was related to higher accuracy when only similar tasks of medium difficulty could be used. Participants' knowledge about task difficulty did not have an effect on judgment accuracy over and above the strategy use.

Keywords: Judgment Accuracy; Decision Making; Preservice Teachers; Simulated Classroom

Zusammenfassung. Im Beitrag wird untersucht, wie angehende Lehrkräfte bei der Informationssammlung für die anschließende Beurteilung von mathematischen Schülerfähigkeiten vorgehen und welche der gezeigten Vorgehensweisen unter welchen Umständen mit akkurateren Urteilen einhergehen. Dazu wurde in einem experimentellen Zwei-Gruppen-Design die Möglichkeit der Berücksichtigung von diagnostisch sinnvoll nutzbaren Aufgabenmerkmalen variiert, das Wissen von Lehramtsstudierenden über Aufgabenschwierigkeiten erhoben und deren diagnostische Vorgehensweisen im Simulierten Klassenraum erfasst und kategorisiert. Besteht die Möglichkeit zur Verwendung unterschiedlich schwieriger Aufgaben, geht eine aufgabenbezogene Vorgehensweise mit akkurateren Leistungsurteilen einher. Können lediglich gleich schwierige Aufgaben verwendet werden, kommen Teilnehmer insbesondere dann zu akkurateren Urteilen, wenn sie eine eher schülerbezogene Vorgehensweise anwenden. Über die Vorgehensweisen hinaus hat das Wissen der Studierenden über Aufgabenschwierigkeiten keinen Effekt auf die Urteilsgüte.

Schlüsselwörter: Urteilsgüte; Urteilsprozess; diagnostische Kompetenz; Lehramtsstudierende; Simulierter Klassenraum

Accurately estimating their students' ability is a central task for teachers and an important aspect of teacher professionalism (Ready & Wright, 2011). Teachers use formative assessments for making didactical decisions and summative assessments in order to grade students, both based on manifold information collected in the classroom. Giving subject-specific tasks to students plays a central role for ability assessment in the everyday school context (Niegemann & Stadler, 2001). Yet, not all of these tasks contribute equally to giving a teacher access to student thinking. Hence, tasks can differ in their potential diagnostic value for formative and summative ability judgment. Judgment formation based on a student's performance on a selection of tasks with varying levels of difficulty (high

diagnostic value) enables teachers to draw inferences on student's current abilities and their additional educational needs (Wang, 1980).

In contrast, judgment formation based on tasks with low diagnostic value (i.e., a bunch of tasks with similar task requirements and thus similar task difficulties) cannot provide an equally differentiated notion of student thinking and does not allow deriving appropriate didactical consequences. However, for being able to constructively use tasks with diagnostic potential for ability assessment, teachers need to have specific knowledge (Heritage, 2013). They further need to apply assessment strategies enabling them to appropriately collect and process information on student performance in order to form an accurate judgment.

By now, most studies on teacher judgment accuracy focus on accuracy outcomes and try to relate teacher traits to these outcomes. Studies covering teachers' assessment behavior and their content-specific knowledge about task difficulties and task requirements are rare, when it comes to research on why and how (accurate) judgments develop. The present study takes up this desideratum by investigating teachers' assessment strategies, their content-specific knowledge, as well as the role of task material being used, and relates that to accuracy outcomes in a classroom assessment task.

The study is conducted within a Simulated Classroom (e.g., Fiedler, Walther, Freytag, & Plessner, 2002; Südkamp, Möller, & Pohlmann, 2008; Kaiser, Retelsdorf, Südkamp, & Möller, 2013). This instrument is considered to be a promising tool for studying aspects of teacher judgment accuracy (e.g., Brown, 1999; Schrader, 2010; Spinath, 2012), offering the possibility of closely observing judgment phenomena by means of experimental variations of student and task characteristics. Another advantage is the availability of detailed log-files, covering exactly what student information is collected by the teachers during the assessment process.

Teacher judgment accuracy and diagnostic competence

Accurate estimations of student and task characteristics are necessary for successfully conducting adaptive lessons (Helmke & Schrader, 1987; Wiliam, 2007) and for providing students with meaningful and helpful feedback (Hattie, 2009). However, studies show that teacher judgments are far from being accurate in a range of content areas. In their meta-analysis, Südkamp, Kaiser, and Möller (2012) report an overall mean effect size of $r = .63$ for the correlation between teacher judgments and student achievement. The repeatedly reported great inter-individual differences between teachers cannot be explained in a comprehensive way, yet. Characteristics of teachers, students, judgments and tests are

considered to be the key variables influencing teacher judgment accuracy (Südkamp et al., 2012).

Broader concepts of teacher decision making (e.g., Schrader, 2011; Van Ophuysen, 2010) – also referred to as diagnostic competence – claim, that teachers' knowledge about task difficulties and task requirements, as well as their knowledge about students' cognitive processes and typical learning strategies and aspects of the process quality of educational decision making, such as the collection and processing of available data are important predictors or even inherent elements of diagnostic competence (National Research Council, 2001). Accurately judging student abilities is rather seen as a domain-related than an overarching, domain-transcending competence. This view is also supported by the findings that teachers do not necessarily yield accurate judgments in different domains (e.g., Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006; Hopkins, George, & Williams, 1985; Lorenz & Artelt, 2009).

Formation of diagnostic judgments on student achievement

Ideally, teacher judgments about student characteristics are based on information providing relevant cues about the attributes to be assessed. In the everyday school context, such cues can be derived from student performance in class, e.g., from answers to given tasks. Amount, quality and variety of these cues can also be influenced by teachers' assessment behavior. They further have to detect these cues and make interpretative use of them in order to form appropriate ability judgments (Funder, 1995). Yet, dependent on its purpose, teachers might also form their judgment based on information that is not directly related to student performance. They might also use 'cognitive shortcuts' for their judgments of student ability, which is underpinned by studies finding that teacher judgments of student achievement can be influenced by student characteristics, such as e.g., students' social background (Krolak-Schwerdt, Böhmer, & Gräsel, 2012) or their engagement in classroom (Kaiser et al., 2013). Studies conducted in the Mouselab paradigm – a computer program, closely related to the framework of adaptive decision making (Payne, Bettman, & Johnson, 1993), that enables researchers to record and trace information acquisition processes in various areas of application – offer some evidence on which information teachers use for school tracking decisions (e.g., Böhmer, Hörstermann, Gräsel, Krolak-Schwerdt, & Glock, 2015; Gräsel & Böhmer, 2013). Teachers used information on student performance first, but backed up this information with students' social behavior and social background afterwards (Böhmer et al., 2015). Yet, it remains unclear, which of the available information on student performance teachers explicitly use for ability judgments in classroom, and how

this information is collected and processed, if there is no additional information available on the students over and above their performance on given tasks. Assessment strategies teachers apply during information acquisition might depend on the goal (impression formation vs. high stakes decision) (Gräsel & Böhmer, 2013), but also on the nature of available information. Having the possibility of using tasks with high diagnostic potential (i.e., by enabling teachers to choose from tasks with a wide range of difficulties) should lead to different assessment strategies and to different accuracy outcomes, as compared to diagnostic situations where only a narrower choice of tasks with lower diagnostic value is used.

The role of teacher knowledge in the judgment processes

Depending on the nature of available information and depending on how well teachers make use of this information, different assessment strategies and different accuracy outcomes can be expected (Funder, 1995). Given a high level of diagnostic potential in available tasks, teachers should choose those tasks from which they expect to reveal relevant cues about a student's ability. Appropriate task choice requires accurate estimations of their requirements and difficulty in relation to a certain age group. These task-specific estimations and the detection and interpretation of cues derived from student behavior for the purpose of judgment formation are broadly grounded on teachers' pedagogical content knowledge (pck) (Shulman, 1986): Besides knowledge on useful representations of ideas in a particular domain, knowledge on aspects making tasks in a given domain difficult for students of a certain age group or with a certain background is also explicitly covered within pck. Choosing appropriate tasks for students with different ability levels, detecting errors and interpret misconceptions from students' answers or solutions can be considered a basis for successful assessment strategies. Hill, Ball, and Schilling (2008) refine Shulman's conceptualization, introducing teachers' knowledge of content and students (kcs) as one strand among others associated with pck. It combines teachers' content knowledge and their knowledge of students' thinking, knowing, and learning of a specific content, as well as knowledge about common student errors, students' understanding of content, typical developmental sequences, and common strategies (Hill et al., 2008). These aspects come into action when teachers are required to estimate the task difficulty for a certain group of students in order to adaptively choose the right tasks for the assessment of their students' abilities. Teachers' estimations of task difficulty can be seen as a proxy to kcs.

Earlier studies claim teachers' accuracy of estimating task difficulties to be on a rather mediocre level (e.g., $\bar{r} = .50$ for learning materials that include instructional pictures)

(McElvany et al., 2009) and find teachers with different experience levels differentially overestimating the solution rates of mathematical tasks (Ostermann, Leuders, & Nückles, 2015). Yet, it is not really clear to date, how this teacher knowledge about task requirements and task difficulties is empirically related to assessment behavior and to accuracy of ability judgments. There is, however, first evidence from the field of text comprehension, suggesting low interrelations between teachers' content-specific knowledge and their judgment accuracy (Rausch, Matthäi, & Artelt, 2015). In order to determine a student's ability and to detect where he or she still has room for improvement, teachers should adaptively choose diagnostically relevant tasks based on their knowledge-informed estimations of task requirements and task difficulty.

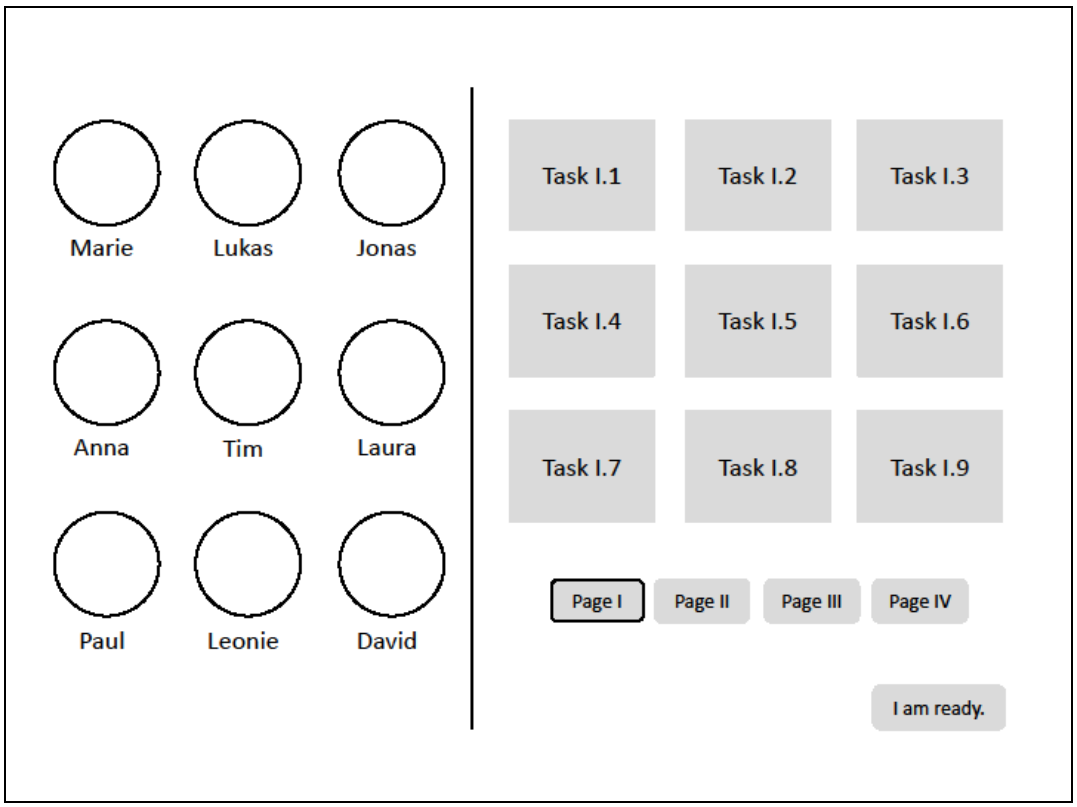


Figure 1: Simplified representation of the Simulated Classroom used in the present study (translation; original in German).

The present study

Of central interest in this experimental study is the formation of teacher judgments and its relation to accuracy outcomes and knowledge aspects, dependent on the nature of tasks participants were able to use for the assessment. As the process of judgment formation is rather difficult to observe in field studies, we chose to investigate this by means of a Simulated Classroom, a computerized research environment, developed in order to study (educational) decision making (Fiedler et al., 2002; Kaiser et al., 2013; Südkamp et al., 2008). To get an impression of the instrument used in the present study, Figure 1 provides a simplified representation of the Simulated Classroom.

Research questions

Judgment accuracy

How accurate are preservice teachers' judgments of students' mathematical abilities? It is expected that judgments are more accurate when participants have the possibility to use a range of diagnostically relevant task characteristics to form their judgment.

Assessment strategies

How do preservice teachers select student information for their judgment formation? As there is no previous evidence on which assessment strategies teachers use for ability judgment, this remains an explorative question. Nevertheless, it is expected that assessment strategies differ depending on the possibility of using a range of diagnostically relevant task characteristics for judgment formation.

Relating assessment strategies to accuracy outcomes

How are preservice teachers' assessment strategies related to accuracy outcomes of their judgments on students' mathematical ability? It is expected that strategies are differentially related to accuracy outcomes, dependent on the nature of the tasks participants can use for their assessment.

Knowledge about task difficulties and task requirements

How accurate are preservice teachers in estimating the difficulty of mathematical tasks? Based on prior studies, it is expected that preservice teachers' accuracy, as a proxy for their knowledge about task difficulties and task requirements, lies in a medium range.

Relating knowledge to assessment strategies and accuracy outcomes

How is preservice teachers' knowledge about task difficulties and task requirements related to assessment strategies and accuracy outcomes of their judgments on students' mathematical ability? It is expected that higher knowledge is related to more accurate judgments, when diagnostically relevant tasks can be used for the assessment. It is further expected that higher knowledge is related to more accurate judgments, when appropriate assessment strategies are applied.

Method

Experimental Setup

An experiment with two conditions was set up in a Simulated Classroom. Initially, particpants were instructed that they have the task to find out about students' mathematical ability in a Simulated Classroom, and to deliver a judgment on each of the students' ability afterwards. Then, the mode of operation of the Simulated Classroom was explained. The two experimental conditions differ in the tasks participants could use to assess the student ability: Participants in the first condition (in the remainder referred to as *SC*) used 34 mathematics tasks that were considered to be of similar, medium difficulty. Participants in the second condition (in the remainder referred to as *SC+D*) used 34 mathematics tasks with a broad range of difficulties. They additionally had to estimate the difficulty of each task before they used these tasks to assess student abilities, while participants in SC had the same amount of time to get familiar with the tasks. Participants were individually tested in computer laboratories at university either alone or with up to two other participants in the room at the same time. Duration of the whole test session was about one hour. Each of the participants received EUR 15 for participation.

Sample

The sample consists of 57 preservice primary school teachers from a German university in the state of Bavaria. They were recruited in lectures, via notice boards and mailing lists. 91.2% of the students were female, which roughly meets the common gender ratio (86.7% female teachers) in German primary school teachers (Malecki, Schneider, Vogel, & Wolters, 2014). Mean age of participants was 22.4 years ($SD = 3.1$), on average they were in their third semester ($M = 3.3$, $SD = 2.3$). For all participants, mathematics education was one academic subject within their university training to become a primary school teacher. Participants were randomly assigned to the two conditions. According to Mann-Whitney-U tests, the groups did not differ significantly in age ($Z = 0.50$; n.s.) nor in the number of semesters ($Z = 0.50$; n.s.). See Table 1 for sample description.

Table 1: Sample description

	N	female participants	age		number of semesters	
			M	*SD*	*M*	*SD*
whole sample	57	91.2%	22.4	3.1	3.3	2.3
SC	28	96.4%	21.9	2.1	3.1	1.9
SC+D	29	86.2%	23.0	3.7	3.6	2.6

The Simulated Classroom

After a brief general description of the Simulated Classroom, configurations for the two conditions will be described in detail. On the left side of the screen (see Figure 1), nine students (four female and five male) were presented. Drawn avatars were used particularly to avoid effects of visible student characteristics that might impact the judgment. Students' names were chosen from the most popular children's names in Germany in 2007, which was the presumed year of birth for the simulated students being in the fifth grade at the time of the assessment (cf. Kaiser et al., 2013). Each avatar was sitting on the same place for all participants. Yet, students' achievement parameters were rotated for every participant in order to avoid effects of seating arrangements.

On the right side of the screen, 34 mathematics tasks were presented in four blocks in a random order. Participants had to select a task and subsequently choose one student to answer the task. The student's answer was either correct or incorrect, according to his or her ability parameter, and was displayed without any time lag. After that, the next task could be chosen and posed to a student. A student's answer did not affect subsequent answers of other students. Participants did not get any direct information on whether a student's answer was correct or not; this information had to be gathered by the student's answer. To prevent participants from the time consuming need of solving all 34 tasks on their own, they were provided with correct answers for each task beforehand and they were able to use this information in the Simulated Classroom. There was no time limit for collecting information about the students, nor was there a limit of tasks per student that could have been chosen. Participants had to decide on their own when they have collected enough information in order to be able to estimate each of the students' ability. Participants indicated the completion of information collection by pressing a button which directed them to a page where they had to deliver their judgments about each of the students' ability. Participants were not allowed to take notes. The assessment process was saved event-based as log-files in a database.

For being able to investigate our research questions, a different configuration of the Simulated Classroom was deployed in each of the two groups (*SC* and *SC+D*). They differ in the task material participants could use and in the operationalization of student ability.

Simulated Classroom without consideration of task difficulty (SC). Participants in the first condition assessed student ability based on a number of tasks with presumably medium task difficulties and low variance in the degrees of difficulty. Student ability was operationalized by the proportion of a student's correct answers out of all tasks the student answered during the assessment. Thirty-four tasks of different types (addition, subtraction, multiplication, number ray tasks, short word problems) were taken over from a study by Südkamp et al. (2008). They selected tasks with medium empirical item difficulties from a common German curriculum-based and standardized mathematics screening test (DEMAT3+; Roick, Goelitz, & Hasselhorn, 2004) and constructed similar tasks based on that selection.

Each of the nine students was assigned to an ability stage, represented by the proportion of each student's correct answers. To approximate a normal distribution within the class, five ability stages were simulated: One student answers 18% (6 tasks), two students answer 32% (11), three students answer 50% (17), two students answer 65% (22) and one student answers 82% (28) of all possible tasks correctly (see also Table 2).

Simulated Classroom with consideration of task difficulty (SC+D). Participants in the second condition were enabled to use tasks with varying degrees of difficulty. Student ability was operationalized by the level of task difficulty reached by the respective student. Achievement behavior is fully determined by the item and person parameters, i.e., for each student there is a specific value on the ζ -scale, from which on all tasks are answered correctly by the respective student. The probability that a certain task is solved by a certain student changes from 1 to 0, when the task difficulty is above the particular student's ability (cf., deterministic latent-trait-model; Guttman, 1950).

Thirty-four tasks of different types (addition, subtraction, multiplication, short word problems) with a range of task difficulties were derived from the curriculum-based mathematics section of the German longitudinal study BiKS-8-14 (Weinert, Artelt, & Roßbach, in preparation). Data were re-analyzed in order to re-determine item parameters for the tasks particularly used in the present study. Mathematical performance data from N = 1,579 grade 5 students were analyzed using the eRm package in R (Mair, Hatzinger, & Maier, 2012), which resulted in item parameters that were used for simulating student abilities. Five ability groups were simulated: One student solves all tasks easier than

β = -1.32 (7 tasks), two students solve all tasks easier than β = -.90 (10), three students solve all tasks easier than β = 0.12 (18), two students solve all tasks easier than β = .90 (24), and one student solves all tasks easier than β = 1.99 (30) (see Table 2).

Table 2: Simulated student ability and achievement behavior in *SC* and *SC+D*

Simulated Classroom (*SC*) (*n* = 252)		**Simulated Classroom (*SC+D*) (*n* = 261)**		
simulated student ability[a]	**real achievement behavior[c] *M* (*SD*)**	**simulated student ability (logits)[b]**	**simulated student ability[a]**	**real achievement behavior[c] *M* (*SD*)**
0.18	0.22 (0.10)	-1.32	0.21	0.22 (0.05)
0.32	0.34 (0.12)	-0.90	0.29	0.33 (0.10)
0.50	0.51 (0.15)	0.12	0.53	0.53 (0.09)
0.65	0.66 (0.10)	0.90	0.71	0.68 (0.11)
0.82	0.81 (0.08)	1.99	0.88	0.84 (0.09)

Note: [a] Estimated share of solved tasks; [b] Students solve all tasks that are easier than this logit value; [c] empirical share of solved tasks

Measures

Student ability. Under the condition *SC*, student ability was a fixed proportion of correctly answered tasks. Because participants usually did not pose all questions to all students, it occurred that students' actual performance did not always reflect their simulated ability. Yet, on average, the implementation of the ability worked quite well (see Table 2). Nevertheless, the actual performance was used as a measure of comparison for calculating the participants' judgment accuracy. Under the condition *SC+D*, student ability was a five-step variable of the simulated ability, where ability is operationalized by the most difficult task a student answered correctly. Although the number of solved items is also a function of the ability in this condition, students' ability parameters θ (derived from the respective β described above) were used as a measure of comparison for calculating participants' judgment accuracy.

Knowledge about task difficulties and task requirements (only in SC+D). The 34 tasks, together with the respective correct answers were presented to the participants in *SC+D* on four sheets of paper. They had to make themselves familiar with the tasks and estimate the percentage of correctly answering students of an average fifth grade in an open ended question.

Judgment of student ability. After having collected information about the student ability, participants had to deliver judgments on each of the students' mathematical ability based on students' performance on a scale from 0 ("very low abilities in mathematics") to 10 ("very high abilities in mathematics").

Statistical analyses for judgment accuracy and teacher knowledge

Judgment accuracy of student achievement. As a central indicator for judgment accuracy, the rank-order component (Schrader & Helmke, 1987) was used in the present study. It is a measure for teachers' accuracy in ranking students according to their ability. For each participant, student achievement and preservice teachers' judgments on the respective student were correlated using Pearson correlation. A higher value indicates a more accurate judgment on the rank-order component. Note that in *SC*, the correlation was computed between teacher judgment and student performance (i.e., the students' individual real achievement behavior: share of solved tasks), whereas in *SC+D*, the correlation was computed between teacher judgment and the simulated student ability (i.e., the level of task difficulty reached by the respective student). In order to be able to compute a non-distorted mean value for the correlations and to be able to compare the rank-order component between groups, coefficients were transformed using Olkin & Pratt's (1958) G approximation[1].

Knowledge about task difficulties and task requirements (only in SC+D). Pearson correlations were computed between participants' ratings of task difficulty and empirical task difficulties, i.e. the percentage of N = 1,579 fifth-grade students in the BiKS-8-14 study who solved the particular item. A higher value indicates a more accurate judgment of the task ranking, and thus a higher level of knowledge about task difficulties and task requirements. Here, too, coefficients were transformed using Olkin and Pratt's (1958) G approximation for computing group means of participants' knowledge.

[1] We decided to use this approach based on simulation studies by Schulze (2004), which provided evidence that G approximation is a better estimator for averaging correlations than the commonly deployed Fisher-Z-transformations (cf. Eid, Gollwitzer, & Schmitt, 2015).

Measuring assessment strategies in the Simulated Classroom

For categorizing teachers' assessment strategies, participants' actions in the Simulated Classroom were recorded in log-files, so that it is known for every participant which task was posed to which student in which order and at which time in the assessment process. Choosing a task and posing this task to a student is considered one interaction. The transitions between interactions were focused for categorizing participants' strategies. Four different types of transitions between interactions are possible:

Type-1: task x is posed to student a before task y is posed to student a ($t_xs_a \rightarrow t_ys_a$).
Type-2: task x is posed to student a before task x is posed to student b ($t_xs_a \rightarrow t_xs_b$).
Type-3: task x is posed to student a before task y is posed to student b ($t_xs_a \rightarrow t_ys_b$).
Type-4: task x is posed to student a before task x is again posed to student a ($t_xs_a \rightarrow t_xs_a$).

For each participant, the type of transition was coded for each transition between two interactions, and the percentage of each type was calculated. For a participant's strategy to be considered as mainly *student-focused*, his or her share of type-1-transitions had to be at least 50%, while all other shares count less than 30%. The share of type-1-transitions in the present sample ranged from 54 to 96%, with the other transitions counting less than 27%, indicating that student-focused strategy users could be distinctly identified. The same criteria were used for type-2-transitions, leading to a mainly *task-focused* strategy when the participant's share of type-2-transitions was at least 50% while all other shares were smaller than 30%. The share of type-2-transitions in the present sample ranged from 54% to 90%, with the other transitions counting less than 28%, indicating that task-focused strategy users could also be distinctly identified. A participant was classified as a mixed information seeker, when his or her share of type-3-transitions was at least 50% while all other shares were smaller than 30%. Participants were also classified as mixed information seeker, when one share of transitions was higher than 50% and another type of transitions was higher than 30%. Likewise, a participant was classified as a *mixed* information seeker, if no type of transitions occurred in more than 50% of the transitions (cf. Cafferty, DeNisi, & Williams, 1986). Type-4-transitions occurred only rarely in the present sample and thus are merged into the other shares of transition when type-1- or type-2-transitions, respectively, were dominant.

Results

Judgment accuracy

Participants were quite accurate in estimating the rank-order of the students in the Simulated Classroom. Mean correlation between student performance and teacher judgment in *SC* was $G_{SC} = .75$ ($SD = .29$). Mean correlation between student ability and teacher judgment in *SC+D* was $G_{SC+D} = .79$ ($SD = .23$). An independent samples t-test indicated no significant difference between the two groups ($t(55) = 0.54$; n.s.).

Assessment strategies

In *SC*, eighteen participants were classified as predominantly using a student-focused strategy, indicating that they mainly focused on one student, posing a certain number of questions to him or her, before focusing on the next student. Six participants were classified as predominantly using a task-focused strategy, indicating that they focused on a specific task and posed this task to a certain number of students before focusing on the next task. Four persons predominantly used a mixed strategy.

In *SC+D*, nine participants were classified as predominantly using a student-focused strategy, sixteen participants were classified as predominantly using a task-focused strategy, and four persons predominantly used a mixed strategy. The proportions of transition types differed significantly between the two groups (Cramer's V = .363; $p < .05$) with student-focused strategies being more typical for *SC* and task-focused strategies being more typical for the *SC+D* condition.

Relating assessment strategies to accuracy outcomes

Levene's test for homogeneity of variance yielded a significant result ($F(5, 51) = 2.91$; $p < .05$). Thus, non-parametric ANOVA-type statistics (Brunner, Dette, & Munk, 1997) with the between-subject factors »condition« and »strategy« were conducted in order to relate assessment strategies to accuracy outcomes. Main effects for »condition« ($F(1, 14.34) = 0.55$; n.s.) and for »strategy« ($F(1.67, 14.34) = 2.01$; n.s.) were not significant. However, results showed a significant interaction effect ($F(1.67, 14.34) = 7.40$; $p < .01$) for »condition × strategy«. In *SC*, participants applying a student-focused strategy achieved more accurate judgments ($G_{SC} = .79$; $SD = .21$), as compared to participants applying a task-focused strategy ($G_{SC} = .56$; $SD = .49$). Participants applying a mixed strategy achieved slightly more accurate judgments ($G_{SC} = .84$; $SD = .07$). In *SC+D*, participants applying a task-focused strategy achieved more accurate judgments ($G_{SC+D} = .89$; $SD = .15$), as compared to participants applying a student-focused strategy ($G_{SC+D} = .59$; $SD = .26$). Judgment accuracy of participants applying a mixed strategy occurred to be similar to the accuracy of

participants applying a task-focused strategy (G_{SC+D} = .82; SD = .15). Descriptive results for the rank-order component are depicted in Figure 2.

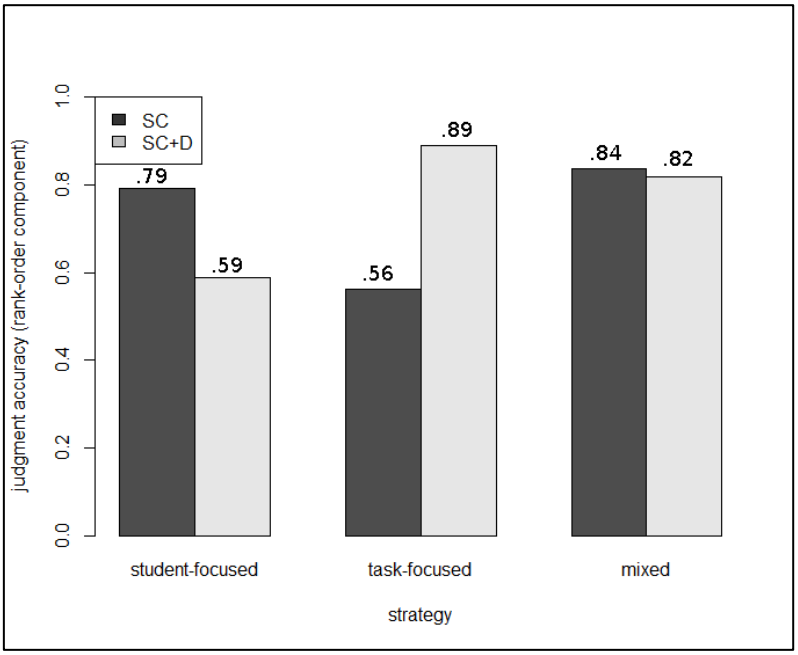


Figure 2: Judgment accuracy by strategy and condition

Knowledge about task difficulties and task requirements (only in SC+D)

G approximated correlations between participants' task ratings and empirical task difficulties served as a proxy for participants' knowledge about task difficulties and task requirements. Participants' accuracy of task difficulty estimations differed widely (G_{min} = .01; G_{max} = .77) with a mean correlation of G = .41 (SD = .18; Mdn = .41).

Relating knowledge and assessment strategies to judgment outcomes (only in SC+D)

For the following analyses, two groups of high vs. low level of knowledge about task difficulties and task requirements were created with a median split. All four participants applying a mixed strategy had a high level of knowledge and were therefore suspended from the analysis. Levene's test for homogeneity of variance yielded a significant result ($F(4, 24)$ = 3.89; $p < .05$). Thus, non-parametric ANOVA-type statistics (Brunner et al., 1997) with the between-subject factors »strategy« and »knowledge level« were conducted for the subsample *SC+D* in order to relate participants' knowledge level and assessment strategies to their accuracy outcomes. Results showed a significant main effect for »strategy« ($F(1, 10.11)$ = 22.20; $p < .01$). The main effect for »knowledge level« ($F(1, 10.11)$ = 0.01; n.s.) and the interaction effect for »strategy × knowledge level« ($F(1, 10.11)$ = 1.15; n.s.) were not significant. Participants applying a task-focused strategy yielded quite accurate ability judgments, independent of their knowledge level (G_{low} = .93; SD_{low} = .04; G_{high} = .83, SD_{high} =

.21). Participants applying a student-focused strategy yielded clearly less accurate judgments, but also independent of their knowledge level (G_{low} = .53, SD_{low} = .28; G_{high} = .70, SD_{high} = .20). Irrespectively of their strategy, participants with a high level of knowledge (G_{high} = .79; SD_{high} = .21) did not significantly differ from participants with a low level of knowledge (G_{low} = .77, SD_{low} = .27) ($t(27)$ = .30; n.s.). Descriptive results are depicted in Figure 3. Over and above the finding that strategies play a role for accuracy outcomes in the Simulated Classroom, an incremental influence of preservice teachers' knowledge about task difficulties and task requirements cannot be detected.

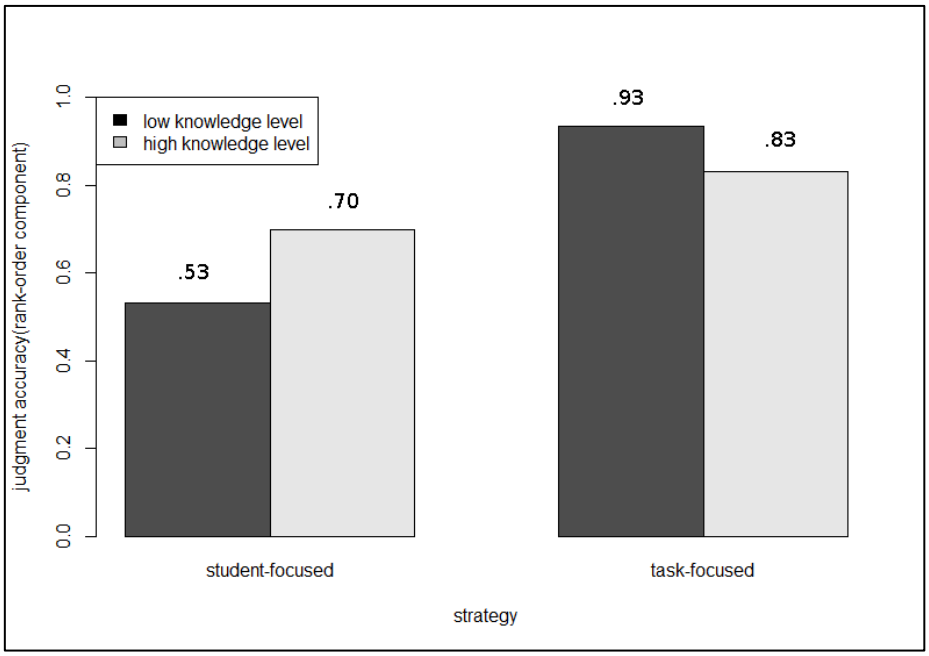


Figure 3: Judgment accuracy in *SC+D* by strategy and level of knowledge about task difficulties and task requirements

Discussion

Summary and discussion of the results

In order to detect, if preservice teachers differ in their assessment strategies and accuracy outcomes depending on the properties of the tasks they could use for ability assessment, their judgment accuracy and assessment strategies were experimentally investigated in a Simulated Classroom. One group of participants had, and one group did not have the possibility to use diagnostically relevant task characteristics to form their judgment.

Preservice teachers' judgment accuracy was higher than in previous Simulated Classroom studies, where correlations between teacher judgment and student achievement ranging from $\bar{r}$ = .60 to $\bar{r}$ = .70 were found for a similar group of preservice teachers (Südkamp &

Möller, 2009; Südkamp et al., 2008). Better performance could have occurred due to the fact that information about students' engagement was missing in the present study. However, this was implemented in previous studies (operationalized by a student's frequency of volunteering to answer a question). As teachers seem to use student engagement as a proxy for estimating student achievement in these studies (e.g., Kaiser et al., 2013), this might be misleading and thus distracting information yielding biased ability judgments.

In the present study, judgment accuracy was not affected by the mere possibility of using diagnostically relevant tasks for assessment. A different picture appears, however, when assessment strategies are taken into account. When participants were only able to use tasks of similar, medium difficulty, they were most likely to apply a student-focused strategy. This strategy can be considered to reflect a participant's tendency to count the number of a student's correct answers on a number of tasks before turning to the next student. Following Cafferty et al. (1986), such a pattern can facilitate the formation of general impressions, which appears to be a promising strategy in absence of any differentiable task characteristics to be used. Applying a task-focused strategy yielded lower judgment accuracy under this condition. This strategy might guide a teacher's attention more on task characteristics and on comparing students' performance on a certain task before turning to the next task and compare students' performance on it.

When participants were able to use tasks of varying difficulty, they were most likely to apply a task-focused strategy. This can be considered to facilitate a comparison between students' performance on a specific task before comparing their performance on the next task etc. The overall performance of each student seems to be estimated at the end of the assessment, based on these comparisons. This delay of overall assessment might also reduce teachers' susceptibility to halo error (cf. Cafferty et al., 1986), which might contribute to more accurate judgments for task-focused strategy users in that condition.

A phenomenon requiring further in-depth investigation is that participants applying a mixed strategy yielded relatively accurate judgments of student ability independent of the task material they could use for assessment. Compared to student-focused and task-focused strategy users, accurate judgments occurred with a much lower effort in terms of the number of teacher student interactions. A mixed strategy enables quick insights in different students' performance on different tasks for a first impression on their ability. However, the lack of any organizational scheme could also weaken the retrieval of student information from memory when delivering the judgment (Cafferty et al., 1986), especially with a higher number of students and tasks used for the assessment. Analyses combining

log-files with think-aloud protocols could reveal a closer understanding of these participants' strategy application and their rationale behind it.

Thus, assessment strategies occurred to be a predictor of judgment accuracy for preservice teachers in a Simulated Classroom. Their knowledge about task difficulties and task requirements, however, did not play an incremental role in predicting judgment accuracy. This does not imply that participants did not utilize task information at all. Yet, given their status as preservice teachers, they might use less sophisticated heuristics for perceiving task characteristics and for estimating task difficulty during the assessment. Knowledge of content and students, as described by Hill et al. (2008), is considered to be not directly covered in the beginning of the first phase of teacher education at German universities. It is rather acquired during practical experience in schools. Other sources for the estimation of task characteristics and task difficulty might have played a greater role in our sample, such as the type of task or participants' memories about how difficult they perceived the types of tasks when being a student themselves.

Limitations and directions for further research

Experimental approaches in artificial computerized settings as the one used in the present study have some obvious restrictions. The Simulated Classroom is a non-naturalistic research environment, where conditions can be created, that are not exactly found in everyday school context in an equal manner. Yet, the setting allows for detailed analyses of judgment processes by offering the possibility of experimentally manipulating amount and nature of available information on students and tasks. In the present study (almost) no additional student information was offered explicitly over and above their performance on tasks. This might not add to ecological validity, but serves as a baseline for further studies investigating if patterns of interaction between teacher and student change, when more information on students or tasks is offered.

Given that participants in previous studies in the Simulated Classroom were also preservice teachers (e.g., Südkamp & Möller, 2009), or even university students without any teaching background at all (e.g., Fiedler et al., 2002), having preservice teachers as participants of this study allows comparability to prior studies. However, the present findings cannot be generalized to more experienced in-service teachers. As differences between experienced and preservice teachers can be expected in the sources of information they rely on (Krolak-Schwerdt & Rummer, 2005), future research should investigate differences between experts and novices regarding their strategies and consideration of task characteristics during ability assessment.

As participants were not allowed to take notes during the assessment process, the less-than-perfect judgments can to some extent also be due to memory effects. Students better in memorizing might achieve more accurate judgments (Payne et al., 1993). This might be even more pronounced for participants counting correct answers rather than for participants considering task difficulty in the assessment. Further studies in the Simulated Classroom should therefore use memory measures in order to control for that.

Conclusion

The present study pioneered in implementing different task difficulties in the Simulated Classroom for enabling participants to apply assessment strategies relying on task characteristics and thus on teachers' pck rather than merely on counting students' right answers on a number of equally difficult questions (cf. Kaiser et al., 2013). It offers first hints that accuracy of teacher judgments is connected to judgment processes, as different strategies lead to different accuracy outcomes depending on the task material that could be used for assessment. With enabling a closer look into the black box of teachers' judgment behavior by analyzing assessment strategies using log-file data, it further opened the way for fine grain investigations of judgment processes using a Simulated Classroom approach.

The authors like to thank Dr. Sabine Krolak-Schwerdt (University of Luxemburg) and Dr. Thomas Hörstermann (University of Luxemburg) for discussions of theoretical and practical issues and for kindly supplying the initial version of the Simulated Classroom. Thanks to Peter Kuntner (University of Bamberg) for technical support in the development of the final version of the Simulated Classroom and during data collection, Jakob Neundorfer, who provided the avatar drawings, and to all participants in this study. This study was generously supported by the Bamberg Graduate School of Social Sciences, which is funded by the German Research Foundation (DFG) under the German Excellence Initiative (GSC1024).

References

Böhmer, I., Hörstermann, T., Gräsel, C., Krolak-Schwerdt, S., & Glock, S. (2015). Eine Analyse der Informationssuche bei der Erstellung der Übergangsempfehlung: Welcher Urteilsregel folgen Lehrkräfte? [An analysis of information search in the process of making school tracking decisions: Which judgment rule do teachers apply?]. *Journal for Educational Research Online, 7*, 59-81.

Brown, A. H. (1999). Simulated classrooms and artificial students: The potential effects of new technologies on teacher education. *Journal of Research on Computing in Education, 32*, 307-318.

Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association, 92*, 1494-1502. doi: 10.1080/01621459.1997.10473671

Cafferty, T. P., DeNisi, A. S., & Williams, K. J. (1986). Search and retrieval patterns for performance information: Effects on evaluations of multiple targets. *Journal of Personality and Social Psychology, 50*, 676-683. doi: 10.1037/0022-3514.50.4.676

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*, 247-265. doi: 10.1002/pits.20147

Eid, M., Gollwitzer, M., & Schmitt, M. (2015). *Statistik und Forschungsmethoden [Statistics and research methods].* Weinheim: Beltz.

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom - A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes, 88*, 527-561. doi: 10.1006/obhd.2001.2981

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652-670.

Gräsel, C., & Böhmer, I. (2013). Die Übergangsempfehlung nach der Grundschule: Welche Informationen nutzen Lehrerinnen und Lehrer für die Entscheidung? [School tracking decisions after primary school: Which information do teachers use for their decision?]. In N. McElvany & H. G. Holtappels (Eds.), *Empirische Bildungsforschung: Theorien, Methoden, Befunde und Perspektiven* (pp. 235-248). Münster: Waxmann.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer (Ed.), *The American soldier. Studies in social psychology in World War II*. Princeton: Princeton University Press.

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

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education, 3*, 91-98. doi: 10.1016/0742-051X(87)90010-2

Heritage, M. (2013). Gathering evidence of student understanding. In J. H. McMillan (Ed.), *SAGE Handbook of research on classroom assessment* (pp. 179-195). Thousand Oaks: SAGE Publications.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*, 372-400.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*, 177-182. doi: 10.1111/j.1745-3984.1985.tb01056.x

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73-84. doi: 10.1016/j.learninstruc.2013.06.001

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2012). Leistungsbeurteilungen von Schulkindern. Welche Rolle spielen Ziele und Expertise der Lehrkraft? [Students' achievement judgments: The role of teachers' goals and expertise]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44*, 111-122. doi: 10.1026/0049-8637/a000062

Krolak-Schwerdt, S., & Rummer, R. (2005). Der Einfluss von Expertise auf den Prozess der schulischen Leistungsbeurteilung [The influence of expertise on the process of judging academic achievement]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 37*, 205-213. doi: 10.1026/0049-8637.37.4.205

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik [Domain specificity and stability of diagnostic competence among primary school teachers in the school subjects of German and mathematics]. *Zeitschrift für Pädagogische Psychologie, 23*, 211-222. doi: 10.1024/1010-0652.23.34.211

Mair, P., Hatzinger, R., & Maier, M. J. (2012). eRm: Extended Rasch Modeling. R package version 0.15-1. http://CRAN.R-project.org/package=eRm.

Malecki, A., Schneider, C., Vogel, S., & Wolters, M. (2014). Schulen auf einen Blick. Ausgabe 2014 [Schools at a glance 2014]. Retrieved from https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/BroschuereSchulenBlick0110018149004.pdf?__blob=publicationFile

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., . . . Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern [Teachers' diagnostic skills to judge student performance and task difficulty when learning materials include instructional pictures]. *Zeitschrift für Pädagogische Psychologie, 23*(3-4), 223-235. doi: 10.1024/1010-0652.23.34.223

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Niegemann, H., & Stadler, S. (2001). Hat noch jemand eine Frage? Systematische Unterrichtsbeobachtung zu Häufigkeit und kognitivem Niveau von Fragen im Unterricht [Is there any question? Systematic observation of classroom behavior concerning questioning]. *Unterrichtswissenschaft, 29*, 171-192.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29*, 201-211.

Ostermann, A., Leuders, T., & Nückles, M. (2015). Wissen, was Schülerinnen und Schülern schwer fällt. Welche Faktoren beeinflussen die Schwierigkeitseinschätzung von Mathematikaufgaben? [Knowing what students know. Which factors influence teachers' estimation of task difficulty?]. *Journal für Mathematik-Didaktik, 36*, 45-76. doi: 10.1007/s13138-015-0073-1

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Rausch, T., Matthäi, J., & Artelt, C. (2015). Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens [Teacher knowledge and judgment accuracy in the domain of text comprehension]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*, 147-158. doi: 10.1026/0049-8637/a000124

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*, 335-360. doi: 10.3102/0002831210374874

Roick, T., Goelitz, D., & Hasselhorn, M. (2004). *Deutscher Mathematiktest für dritte Klassen DEMAT 3+ (Vol. 4) [German Mathematics Test for Third Graders DEMAT 3+]*. Göttingen: Hogrefe.

Schrader, F.-W. (2010). Diagnostische Kompetenz von Eltern und Lehrern [Teachers' and parents' diagnostic competence]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (pp. 102-108). Göttingen: Hogrefe.

Schrader, F.-W. (2011). Lehrer als Diagnostiker [Teachers as diagnosticians]. In E. Terhart, H. Bennewitz & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (pp. 683-698). Münster: Waxmann.

Schrader, F.-W., & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen [Diagnostic competence of teachers: Components and Effects]. *Empirische Pädagogik, 1*, 27-52.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, MA: Hogrefe & Huber.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Spinath, B. (2012). Beiträge der Pädagogischen Psychologie zur Professionalisierung von Lehrerinnen und Lehrern: Diskussion zum Themenschwerpunkt [Educational Psychology's contributions to professional teacher development: Discussion of the special issue]. *Zeitschrift für Pädagogische Psychologie, 26*, 307-312. doi: 10.1024/1010-0652/a000082

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743-762. doi: 10.1037/a0027627

Südkamp, A., & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum: direkte und indirekte Einschätzungen von Schülerleistungen [Reference-group effects in a Simulated Classroom: Direct and indirect judgments]. *Zeitschrift für Pädagogische Psychologie, 23*, 161-174. doi: 10.1024/1010-0652.23.34.161

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum. Eine experimentelle Untersuchung zur diagnostischen Kompetenz [Simulated classroom. An experimental approach to diagnostic competence]. *Zeitschrift für Pädagogische Psychologie, 22*, 261-276. doi: 10.1024/1010-0652.22.34.261

Van Ophuysen, S. (2010). Professionelle pädagogisch-diagnostische Kompetenz - eine theoretische und empirische Annäherung [Professional pedagogical-diagnostic competence – A theoretical and empirical approach]. In N. Berkemeyer, W. Bos, H. G. Holtappels, N. McElvany & R. Schulz-Zander (Eds.), *Jahrbuch der Schulentwicklung* (Vol. 16, pp. 203-234). Weinheim: Juventa.

Wang, M. C. (1980). Adaptive instruction: Building on diversity. *Theory into Practice, 19*, 122-128.

Weinert, S., Artelt, C., & Roßbach, H.-G. (in preparation). *Die Forschergruppe BiKS - Anlage, Methoden und Instrumente der beiden Längsschnittstudien [The Bamberg research group BiKS - Layout, methods and instruments of the two BiKS longitudinal studies]*. Bamberg: University of Bamberg Press.

Wiliam, D. (2007). Content then process: Teacher learning communities in the service of formative assessment. In D. B. Reeves (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 183-204). Bloomington: Solution Tree.

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorgelegte Dissertation selbständig angefertigt, dabei keine anderen Hilfsmittel als die im Quellen- und Literaturverzeichnis genannten benutzt, alle aus Quellen und Literatur, einschließlich des Internets, wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht und auch die Fundstellen einzeln nachgewiesen habe.

Bamberg, 21.12.2016