Lehrkräfte als Diagnostikerinnen und Diagnostiker. Untersuchungen zu ausgewählten Tätigkeiten von Grundschullehrerinnen und Grundschullehrern im Bereich der pädagogisch-psychologischen Diagnostik

Dissertation zur Erlangung des akademischen Grades
Doctor philosophiae (Dr. phil.)
im Fachgebiet Psychologie

eingereicht bei der

Humanwissenschaftlichen Fakultät
der Universität Potsdam

von Dipl.-Psych. Lars Hoffmann

Betreuung:
Prof. Dr. Katrin Böhme

Jahr der Einreichung: 2017

Gutachter/Gutachterin:
1. Prof. Dr. Katrin Böhme
2. Prof. Dr. Martin Brunner

Tag der Verteidigung:
9. Oktober 2017

Online veröffentlicht auf dem
Publikationsserver der Universität Potsdam:
URN urn:nbn:de:kobv:517-opus4-404584
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-404584

Danksagung

Bei der Arbeit an dieser Dissertationsschrift wurde ich von zahlreichen Personen unterstützt, denen ich an dieser Stelle meinen Dank aussprechen möchte.

Zuvorderst gilt mein Dank Prof. Katrin Böhme, die die Fertigstellung dieser Arbeit mit vielen hilfreichen Anmerkungen, wertvollen Hinweisen und pragmatischen Ratschlägen begleitet hat. Danken möchte ich ihr auch dafür, dass sie trotz ihres Wechsels an die Universität Potsdam an der Betreuung meiner Arbeit festgehalten hat und diese auch begutachten wird. Ebenso danke ich Prof. Martin Brunner für die Bereitschaft, die vorliegende Arbeit zu begutachten. Meinen besonderen Dank möchte ich auch Prof. Petra Stanat aussprechen – für das sorgfältige Lesen meiner Manuskripte, die vielen guten Hinweise zu deren Optimierung und dafür, dass ich am IQB in einer überaus motivierenden und lehrreichen Atmosphäre arbeiten und promovieren durfte. Jutta Hoffmann und Christian Seydel danke ich herzlich für das Lektorat der Dissertationsschrift.

Am IQB hatte ich die Möglichkeit, in mehreren Projekten, sowohl im Bereich der Sekundarstufe I als auch im Abitur, mitzuwirken. Den Kolleginnen und Kollegen, mit denen ich dabei zusammenarbeiten durfte, und insbesondere meinen beiden Projektkoordinatorinnen, Gabriele Gippner und Susanne Hunger, danke ich für das Verständnis und die Bereitschaft, mir die für die Promotion benötigten zeitlichen Freiräume zur Verfügung zu stellen.

Last but not least will ich mit Ingo Fleischer, Dr. Malte Jansen, Barbara Mühlbacher, Xenia Steigerwald und Dr. Sebastian Wurster den wundervollen Menschen danken, die die Fertigstellung dieser Arbeit durch ihre Freundschaft, ihren Zuspruch, ihr Interesse, ihren emotionalen Beistand, ihre Ermutigung und ihre Appelle unterstützt und mein Leben in den vergangenen Jahren in unschätzbarer Weise bereichert haben.

Abstrakt

Im Fokus der kumulativen Dissertationsschrift stehen Grundschullehrerinnen und Grundschullehrer in ihrer Rolle als Diagnostikerinnen und Diagnostiker. Exemplarisch werden mit der Einschätzung von Aufgabenschwierigkeiten und der Feststellung von Sprachförderbedarf zwei diagnostische Herausforderungen untersucht, die Lehrkräfte an Grundschulen bewältigen müssen.

Die vorliegende Arbeit umfasst drei empirische Teilstudien, die in einem Rahmentext integriert sind, in dem zunächst die theoretischen Hintergründe und die empirische Befundlage erläutert werden. Hierbei wird auf die diagnostische Kompetenz von Lehrkräften bzw. die Genauigkeit von Lehrerurteilen eingegangen. Ferner wird der Einsatz standardisierter Testinstrumente als wichtiger Bestandteil des diagnostischen Aufgabenfeldes von Lehrkräften charakterisiert. Außerdem werden zentrale Aspekte der Sprachdiagnostik in der Grundschule dargestellt.

In Teilstudie 1 (Hoffmann & Böhme, 2014b) wird der Frage nachgegangen, wie akkurat Grundschullehrerinnen und -lehrer die Schwierigkeit von Deutsch- und Mathematikaufgaben einschätzen. Darüber hinaus wird untersucht, welche Faktoren mit der Über- oder Unterschätzung der Schwierigkeit von Aufgaben zusammenhängen.

In Teilstudie 2 (Hoffmann & Böhme, 2017) wird geprüft, inwiefern die Klassifikationsgüte von Entscheidungen zum sprachlichen Förderbedarf mit den hierfür genutzten diagnostischen Informationsquellen kovariiert. Der Fokus liegt hierbei vor allem auf der Untersuchung von Effekten des Einsatzes von sprachdiagnostischen Verfahren.

Teilstudie 3 (Hoffmann, Böhme & Stanat, 2017) untersucht schließlich die Frage, welche diagnostischen Verfahren gegenwärtig bundesweit an den Grundschulen zur Feststellung sprachlichen Förderbedarfs genutzt werden und ob diese Verfahren etwa testtheoretischen Gütekriterien genügen.

Die zentralen Ergebnisse der Teilstudien werden im Rahmentext zusammengefasst und studienübergreifend diskutiert. Hierbei wird auch auf methodische Stärken und Schwächen der drei Beiträge sowie auf Implikationen für die zukünftige Forschung und die schulische Praxis hingewiesen.

Abstract

The cumulative doctoral thesis focusses on elementary school teachers and their role as diagnosticians. Two diagnostic challenges that teachers in elementary schools typically have to manage are investigated: judging task difficulty and identifying children with a particular need for language support.

The present thesis consists of three empirical articles that are embedded in a framing text which at first outlines theoretical backgrounds and empirical findings and then elaborates on teachers' diagnostic skills and on the accuracy of teachers' judgments. Furthermore, the application of standardized tests is characterized as an important element of teachers' field of work. Additionally, central aspects of language diagnostics in elementary schools are illustrated.

The first article (Hoffmann & Böhme, 2014b) investigates the accuracy of elementary school teachers' difficulty judgments of German language and mathematics tasks. The article furthermore addresses the question of what factors are correlated with the over- and underestimation of task difficulty.

The second article (Hoffmann & Böhme, 2017) examines whether the classification accuracy of decisions on the need for language support covaries with the use of specific diagnostic information sources. The article puts a special focus on effects of the administration of particular diagnostic instruments (e. g., tests).

The third article (Hoffmann, Böhme & Stanat, 2017) finally addresses the question what diagnostic instruments are currently used in elementary schools throughout Germany to identify a need for language support and whether these instruments (for example) match test theoretical quality criteria.

Finally, the central results of the three articles are summarized and reflected in an overall discussion that outlines methodical strengths and weaknesses of each study as well as implications for future research and practical implications for schools.

Inhaltsverzeichnis

1

Einleitung

1 Einleitung

Diagnostische Fragestellungen fallen in unterschiedlichen Bereichen und Anwendungsfeldern an. Im Alltag begegnen sie uns vor allem, wenn wir gesundheitliche Beschwerden haben und Ärztinnen und Ärzte aufsuchen, um mit deren Hilfe möglichst schnell wieder gesund zu werden. Diese werden zunächst bestrebt sein, eine Diagnose zu stellen. Sie werden versuchen zu bestimmen, ob wir tatsächlich erkrankt sind, und, wenn dem so ist, zu klassifizieren, an welcher konkreten Krankheit wir leiden. Hierfür werden sie, insbesondere mithilfe entsprechender medizinischer Untersuchungsmethoden und im Rahmen des Patientengesprächs, diagnostische Daten erheben und prüfen, inwieweit diese Daten den Symptomen von in Frage kommenden Krankheitsbildern entsprechen. Aus ihrer Diagnose werden sie erste Indikationen ableiten sowie Prognosen zu Erfolgswahrscheinlichkeiten stellen und auf dieser Basis die aus ihrer Sicht am besten geeigneten medizinischen Behandlungsmaßnahmen auswählen. Die anschließende Behandlung wird dann durch weitere diagnostische Prozesse begleitet werden, die vor allem dazu dienen, den Heilungsverlauf und die Effektivität der gewählten medizinischen Maßnahmen zu kontrollieren und zu beurteilen, ob die Behandlung erfolgreich abgeschlossen ist, weiter fortgeführt werden sollte oder einer Modifikation bedarf (Wittchen, 2011; Wittchen & Hoyer, 2011).

Diagnostische Fragestellungen werden jedoch nicht nur von Ärztinnen und Ärzten bearbeitet, sondern sind auch Gegenstand anderer Berufsfelder. Eine herausgehobene Rolle spielen sie in der praktischen Tätigkeit vieler Psychologinnen und Psychologen – und das nicht nur bei denjenigen, die als Psychotherapeutinnen und -therapeuten in einem der Medizin inhaltlich nahem Gebiet tätig sind, sondern auch (und in einem oftmals noch viel größerem Maße) bei denen, die etwa in den Bereichen der Pädagogischen Psychologie, der Arbeits- und Organisationspsychologie, der Forensischen Psychologie oder der Verkehrspsychologie arbeiten (Roth & Herzberg, 2008; Schmidt-Atzert & Amelang, 2012).

Der *Psychologischen Diagnostik* bzw. *Psychodiagnostik* kommt ein dementsprechend hoher Stellenwert in der Psychologie als wissenschaftliche Disziplin und im Rahmen der universitären Ausbildung von Psychologinnen und Psychologen zu. Darüber hinaus werden das Know-how der Psychodiagnostik, ihre theoretischen Grundlagen und Methoden, auch in anderen Fachbereichen und Berufen genutzt. Sie finden sich

beispielsweise in den Curricula kaufmännischer Studiengänge mit einer Spezialisierung auf Personalwesen und bilden unter anderem eine wichtige konzeptuelle Grundlage der von Ingenieurinnen und Ingenieuren abgenommenen Fahrerlaubnisprüfungen (Sturzbecher, Mörl & Rüdel, 2013).

Auch Lehrkräfte können als Diagnostikerinnen bzw. Diagnostiker gelten (vgl. Praetorius, Greb, Lipowsky & Gollwitzer, 2010). Ihre diagnostische Kompetenz ist unter anderem bei der Kontrolle des Lernerfolgs ihrer Schülerinnen und Schüler, bei der Vergabe von Schulnoten, bei der Identifikation von Kindern und Jugendlichen mit bestimmten Begabungen, Lernschwächen und Förderbedarfen oder im Zusammenhang mit Versetzungsentscheidungen und Schullaufbahnempfehlungen gefordert. Darüber hinaus sind sie in ihrem Beruf immerfort mit diagnostischen Aufgaben konfrontiert, die häufig eher beiläufig im Zusammenhang mit den methodisch-didaktischen Entscheidungen anfallen, die sie bei der Planung und Durchführung ihres Unterrichts treffen müssen (Artelt & Gräsel, 2009; Brunner, Anders, Hachfeld & Krauss, 2011). (1) So stehen Lehrkräfte etwa vor der Herausforderung, die abstrakten, in den Bildungsstandards und Rahmenlehrplänen festgelegten Kompetenzziele für ihren eigenen Unterricht zu präzisieren und thematisch einzubetten (vgl. Helmke, 2010; Mager, 1978; Schott, Neeb & Wieberg, 1981; Tyler, 1971). Hierbei ist es erforderlich, auch das Vorwissen ihrer Schülerinnen und Schüler zu berücksichtigen. Um dies in angemessener Art und Weise tun zu können, müssen sie in der Lage sein, akkurat zu beurteilen, welche Kenntnisse und Fähigkeiten bereits erworben wurden. (2) Außerdem müssen Lehrkräfte entscheiden, anhand welcher konkreter Lernaufgaben die jeweils zu erwerbenden Kompetenzen angeeignet, gefestigt, automatisiert und gegebenenfalls auf andere Zusammenhänge übertragen werden sollen (vgl. Köster, 2008). Bei der Formulierung und Auswahl geeigneter Lernaufgaben müssen sie ebenfalls das Vorwissen ihrer Schülerinnen und Schüler berücksichtigen – etwa auch, um zu entscheiden, in welchem Umfang sie Unterstützungshilfen für die Bearbeitung der Lernaufgaben bereitstellen müssen. Zum selben Zweck müssen sie ferner in der Lage sein, das Schwierigkeitsniveau von Lernaufgaben bzw. die bei der Aufgabenbearbeitung jeweils zu bewältigenden Anforderungen akkurat zu beurteilen. (3) In ganz ähnlicher Weise sind das Vorwissen der Schülerinnen und Schüler und die jeweils zu bewältigenden Anforderungen auch bei methodisch-didaktischen Entscheidungen zu berücksichtigen, die das Arrangieren der

Lernumgebung betreffen (z. B. Auswahl von Lehr- und Lernmethoden sowie Lehrmaterialien und -medien, zeitliche Planungen).

Auch das Unterrichten selbst erfordert stetige (formative) Diagnoseleistungen (z. B. Brunner et al., 2011), bei denen neben den Lernvoraussetzungen und dem Lernfortschritt weitere lern- und leistungsrelevante Faktoren (z. B. Motivation, Aktivierung, Disziplin) in den Blick genommen werden. Die dabei gewonnenen Informationen dienen Lehrkräften als Ad-hoc-Entscheidungsgrundlage für ihr weiteres Vorgehen (Klauer, 1985; Kunz & Schott, 1987; Reusser, 2009). So lassen sich anhand eines Vergleichs des eingeschätzten Ist-Zustandes mit dem Soll-Zustand bzw. den jeweiligen Zielvorstellungen Maßnahmen zur Anpassung des Unterrichts an die Lernvoraussetzungen und den Lernfortschritt der Schülerinnen und Schüler ableiten, die sowohl unmittelbar während des Unterrichts (Mikroadaptation) als auch längerfristig auf größere Unterrichtseinheiten bezogen (Makroadaptation) erfolgen können (Schrader, 2013; Schrader & Helmke, 2001).

Ganz ähnlich wie bei Ärztinnen und Ärzten ist das Diagnostizieren also auch bei Lehrerinnen und Lehrern fester Bestandteil der beruflichen Tätigkeit. Im schulischen Alltag fallen die *Diagnosen* jedoch zumeist weniger explizit aus als in den Arztpraxen von Medizinerinnen und Medizinern. Gegenstand der Diagnosen von Lehrkräften sind Urteile bzw. Einschätzungen zu lern- und leistungsrelevanten Merkmalen von Schülerinnen und Schülern und zu den jeweils wesentlichen Lern- und Aufgabenanforderungen (Artelt & Gräsel, 2009; Brunner et al., 2011; Schrader, 2009). Lehrkräfte müssen diese Diagnosen vornehmen, um die vielfältigen Entscheidungen treffen zu können, vor die sie täglich gestellt werden. Und die Güte dieser Entscheidungen, so die Annahme, sollte umso höher sein, je akkurater die zuvor gestellten Diagnosen sind (Hoge & Coladarci, 1989).

Die im Lehrerberuf erforderliche diagnostische Kompetenz wird zunächst im Rahmen einer Ausbildung erworben und später im Arbeitsleben, gegebenenfalls auch mithilfe von Fortbildungsangeboten, weiterentwickelt. Die von Lehrkräften zu beherrschenden diagnostischen Tätigkeiten sind Gegenstand der *Pädagogischen Diagnostik*, die aufgrund ihrer großen inhaltlichen und methodischen Schnittmenge mit der psychologischen Diagnostik oftmals auch als *Pädagogisch-psychologische Diagnostik* bezeichnet wird (vgl. Hesse & Latzko, 2009; Langfeldt & Trolldenier, 1993b; Lukesch, 1998; Wilhelm & Kunina, 2009).

Der hohen Relevanz der pädagogischen Diagnostik für den Lehrerberuf wurde in der Vergangenheit nicht immer im hinreichenden Maße Rechnung getragen. Nachdem sie zunächst ab Ende der 1960er Jahre deutlich an Bedeutung gewann, folgte etwa ab Anfang der 1980er Jahre eine Phase der Stagnation und des Rückschritts. Diese war durch eine ausgeprägte Skepsis in der pädagogischen Praxis gegenüber diagnostischen Vorgehensweisen und Methoden geprägt und fand insbesondere in einer starken Ablehnung von standardisierten Tests ihren Ausdruck (vgl. Ingenkamp, 1989).[1] Eine dementsprechend geringe Rolle spielte die pädagogische Diagnostik auch in der Lehrerausbildung und in der Bildungsforschung der damaligen Zeit (Ingenkamp & Lissmann, 2008; Langfeldt & Trolldenier, 1993a).

Im Zuge der neuen „empirischen Wende“ (Bos, Klieme & Köller, 2010, S. 7) in den Bildungswissenschaften, deren ungefährer Beginn zumeist auf die Milleniumswende datiert wird und die zeitlich mit den Beteiligungen Deutschlands an internationalen Schulleistungsstudien wie TIMSS[2] oder PISA[3] (Baumert, Klieme, Neubrand, Prenzel & Schiefele, 2001) zusammenfiel, erlebte auch die pädagogische Diagnostik eine Renaissance (Hesse & Latzko, 2009; Wilhelm & Kunina, 2009). Ursächlich verantwortlich hierfür waren nicht zuletzt die auch von einer breiten Öffentlichkeit rezipierten Ergebnisse der ersten PISA-Studie. Das unerwartet schlechte Abschneiden der Schülerinnen und Schüler in Deutschland gab Anstoß für umfangreiche, konzentrierte Vorhaben zur Qualitätsentwicklung und -sicherung im Bildungswesen. Diese Bemühungen bezogen sich unter anderem auf Verbesserungen von Struktur und Wirksamkeit der Aus-, Fort- und Weiterbildung von Lehrkräften (von Aufschnaiter et al., 2015). Hierzu wurden im Jahr 2004 unter der Herausgeberschaft der Kultusministerkonferenz (KMK) „Standards für die Lehrerbildung“ veröffentlicht (KMK, 2004), in denen der *diagnostischen Kompetenz* von Lehrkräften eine prominente Rolle zukommt. So wird unter anderem hervorgehoben, dass Lehrerinnen und Lehrer zur kompetenten, gerechten und verantwortungsbewussten Ausübung ihrer „Beurteilungs- und Beratungsaufgabe im Unterricht […] hohe diagnostische Kompetenzen [benötigen]“ (KMK, 2004, S. 4). Eine Stärkung der diagnostischen Kompetenz war bereits im

[1] Diese (skizzenhaften) historischen Ausführungen zur pädagogischen Diagnostik beziehen sich auf die Entwicklung in der Bundesrepublik Deutschland.

[2] *T*rends in *I*nternational *M*athematics and *S*cience *S*tudy.

[3] *P*rogramme for *I*nternational *S*tudent *A*ssessment.

unmittelbaren Nachgang der PISA-Studie des Jahres 2000 durch die KMK in den Blick genommen und als eines von sieben vorrangigen Handlungsfeldern benannt worden (KMK, 2002). Einen wichtigen Impuls für diesen bildungspolitischen Bedeutungsgewinn bildete nicht zuletzt der bei PISA-2000 ermittelte Befund, nach dem Deutschlehrkräfte an Hauptschulen oftmals nicht in der Lage waren, die besonders schwachen Leserinnen und Leser in den von ihnen unterrichteten Klassen korrekt zu identifizieren (Artelt, Stanat, Schneider & Schiefele, 2001; Lorenz, 2011).

Nicht nur in den genannten Veröffentlichungen der KMK, sondern auch in anderen Publikationen nimmt die diagnostische Kompetenz von Lehrkräften einen zentralen Stellenwert ein. Beispielsweise führt Weinert (2000) sie als eine von vier Basiskompetenzen von Lehrerinnen und Lehrern an. Nach Baumert und Kunter (2006) stellt sie eine zentrale Facette professioneller Kompetenz von Lehrkräften dar (vgl. auch Anders, Kunter, Brunner, Krauss & Baumert, 2010). Für Artelt und Gräsel (2009) gilt sie als Schlüsselkompetenz in Lehr- und Lernkontexten. Mittlerweile ist die pädagogische Diagnostik zu einem obligatorischen, unverzichtbaren Bestandteil der Curricula für die Lehrerbildung in Deutschland geworden. Gleichzeitig wurden in den vergangenen Jahren zahlreiche Professuren mit pädagogisch-diagnostischem Inhaltsschwerpunkt geschaffen (Artelt & Gräsel, 2009).

Im Zuge des Bedeutungsgewinns der pädagogischen Diagnostik in der Bildungspolitik und in der Aus-, Fort- und Weiterbildung von Lehrkräften rückte die diagnostische Kompetenz auch in den Fokus der ab der Jahrtausendwende erstarkenden empirischen Bildungsforschung. In seinen in den Jahren 2009 und 2013 erschienenen Überblicksartikeln nahm Schrader eine Strukturierung der daraus resultierenden Forschungstätigkeiten vor und stellte zwei zentrale Forschungsbereiche heraus: (1) Untersuchungen, die sich mit der Genauigkeit von Lehrerurteilen befassen, bezeichnete Schrader als *genauigkeitsorientierte Ansätze*. Im Zentrum dieser Ansätze stehen deskriptive Analysen zur Akkuratheit von Lehrerurteilen, aber auch Untersuchungen zur Stabilität und Binnenstruktur der Diagnosegenauigkeit, zu ihren Bedingungsfaktoren und Korrelaten. (2) In Abgrenzung dazu fasste Schrader in einem zweiten Bereich *prozessorientiere Forschungsansätze* zusammen, also Forschungsarbeiten, in denen schwerpunktmäßig die Lehrerurteilen zugrunde liegenden Mechanismen und somit ein breites Spektrum an diagnostischen Fähigkeiten und diagnostischem Wissen betrachtet werden (Schrader, 2009, 2013). Hierbei wird beispielsweise untersucht, auf welchen diagnostischen

Informationen die Urteile von Lehrkräften basieren, mit welchen Methoden sie die Informationen erheben und wie sie diese im Prozess der Urteilsbildung weiterverarbeiten.

Zur Erfassung von diagnostischen Informationen können Lehrkräfte potenziell eine große Bandbreite an Methoden und Verfahren nutzen. Als Informationsquellen können, je nach Zielstellung, zum Beispiel die Ergebnisse von Leistungskontrollen und Klassenarbeiten, eigene Beobachtungen während des Unterrichts, Hinweise von anderen pädagogischen Fachkräften wie Erzieherinnen und Erziehern, Elterngespräche, Gespräche mit den betreffenden Schülerinnen und Schülern oder die Ergebnisse schulärztlicher Untersuchungen dienen (vgl. Geist & Voet Cornelli, 2015; Stanat, Weirich & Radmann, 2012). Einer wachsenden Popularität erfreuen sich ferner andere, dem Zeitgeist des kompetenzorientierten Unterrichtens entsprechende Methoden wie Portfolios und Lerntagebücher (z. B. Gläser-Zikuda & Hascher, 2007). Eine weitere potenzielle Informationsquelle können formelle diagnostische Verfahren und insbesondere standardisierte Testinstrumente darstellen. Allerdings scheinen nicht wenige Schulen und Lehrkräfte noch immer skeptisch gegenüber dem Einsatz von Testverfahren zu sein. Dies zeigt sich besonders pointiert an der Kritik gegenüber den flächendeckend in der dritten und achten Jahrgangsstufe durchgeführten Vergleichsarbeiten.[4] Eine ausgeprägte Skepsis gegenüber Testinstrumenten findet man aber auch in anderen Zusammenhängen, wie etwa beim Einsatz von sprachdiagnostischen Verfahren in Kindertagesstätten (Kitas) und Grundschulen zur Identifikation von Kindern mit besonderem Sprachförderbedarf. Hier trifft man etwa auf das Vorurteil, dass die Verfahren lediglich die Beobachtungen bestätigten, die pädagogische Fachkräfte ohnehin bereits im Alltag machen, und daher überflüssig seien.[5]

Im Fokus der vorliegenden kumulativen Dissertation stehen Lehrkräfte in ihrer Rolle als Diagnostikerinnen und Diagnostiker sowie ausgewählte diagnostische Tätigkeiten des Lehrerberufs. Die *erste Teilstudie* beschäftigt sich mit der Genauigkeit von Lehrerurteilen. Schwerpunktmäßig werden die Akkuratheit von Einschätzungen zu absoluten

[4] Zum Beispiel: „Lehrer schimpfen über Vergleichstests“ (ohne Autor, Spiegel Online, 5.5.2014), verfügbar unter: http://www.spiegel.de/lebenundlernen/schule/vergleichsarbeiten-vera-an-schulen-lehrer-wettern-gegen-tests-a-967642.html [30.11.2016].

[5] Zum Beispiel: „Delfin 4 in der Kritik: Kitas sehen Chance“ (Sebastian Ritscher, RuhrNachrichten.de, 7.1.2014) http://www.ruhrnachrichten.de/staedte/bochum/Sprachstandserhebung-Delfin-4-in-der-Kritik-Kitas-sehen-Chance;art932,2239061 [30.11.2016].

Merkmalsausprägungen sowie die dabei auftretenden Fehlurteile untersucht. In den anderen beiden Teilstudien wird der Einsatz von sprachdiagnostischen Verfahren zur Identifikation von Schülerinnen und Schülern mit besonderem sprachlichem Förderbedarf betrachtet. Vor dem Hintergrund der in der pädagogischen Praxis diesbezüglich offenbar bestehenden Skepsis wird in *Teilstudie 2* empirisch geprüft, inwiefern sich die Güte von diagnostischen Entscheidungen zum Sprachförderbedarf erhöht, wenn zu dessen Feststellung sprachdiagnostische Verfahren zum Einsatz kommen. *Teilstudie 3* geht der Frage nach, welche konkreten diagnostischen Verfahren an den Grundschulen zur Feststellung von Sprachförderbedarf genutzt werden und in welchem Umfang sie verbreitet sind.

Die Dissertationsschrift ist wie folgt aufgebaut: Zur Einbettung der Teilstudien in das Forschungsfeld wird in Kapitel 2 ein Überblick zum theoretischen Hintergrund und zu zentralen empirischen Befunden gegeben. Da die Genauigkeit von Lehrerurteilen im Fokus von Teilstudie 1 steht, erfolgt zunächst, in *Teilkapitel 2.1*, eine Darstellung der genauigkeitsorientierten Perspektive auf die diagnostische Kompetenz von Lehrkräften. Nach einer ersten Begriffsbestimmung wird die Bedeutung von Lehrerurteilen und ihrer Genauigkeit für die schulische Praxis und die angewandte Forschung herausgearbeitet. Im Anschluss daran wird dargelegt, weshalb Schulnoten sich weniger gut als die in empirischen Untersuchungen erhobenen Leistungseinschätzungen eignen, um die Diagnosegenauigkeit von Lehrkräften zu untersuchen. Danach wird erläutert, wie in empirischen Forschungsarbeiten die Genauigkeit von Lehrerurteilen bestimmt wird. Das Teilkapitel schließt mit einer Darstellung des Forschungsstands zur Akkuratheit von Lehrerurteilen sowie zu Bedingungen und Kovariaten der Urteilsgenauigkeit.

Teilkapitel 2.2 dient der thematischen Überleitung zu dem in den Teilstudien 2 und 3 betrachteten Einsatz von sprachdiagnostischen Verfahren in der Grundschule. Ausgehend von einer Kritik an der genauigkeitsorientierten Definition diagnostischer Kompetenz wird unter Bezugnahme auf das Konzept der *Assessment Literacy* begründet, dass die Durchführung von Leistungsmessungen in der Schule bzw. der Einsatz von Tests ein wichtiger Bestandteil des diagnostischen Aufgabenfelds von Lehrkräften ist. Gleichzeitig wird erläutert, dass die Skepsis gegenüber Tests an Schulen in Deutschland noch immer verbreitet ist.

In *Teilkapitel 2.3* wird schließlich spezifiziert, was Gegenstand der Sprachdiagnostik in der Grundschule ist. Des Weiteren werden Anforderungen beschrieben, denen

sprachdiagnostische Verfahren genügen sollten. Außerdem werden die in anderen Forschungsarbeiten gewonnenen Erkenntnisse zum aktuellen Stand der Sprachdiagnostik im Elementar- und Primarbereich skizziert. Zudem erfolgt eine konzise Darstellung des Konzepts der sprachdiagnostischen Kompetenz.

Auf der Grundlage des theoretischen Rahmens werden in *Kapitel 3* die Fragestellungen der drei Teilstudien und ihre inhaltliche Klammer vorgestellt. Die Teilstudien selbst finden sich in den darauffolgenden *Kapiteln 4 bis 6*. Ihre zentralen Ergebnisse werden in *Kapitel 7* zunächst studienübergreifend zusammengefasst, diskutiert und in die Forschungsliteratur eingeordnet. Daran anschließend erfolgt eine kritische Reflexion der methodischen Stärken und Schwächen der einzelnen Teilstudien. Zum Abschluss werden Implikationen der Befunde für Forschung und Praxis dargestellt.

2

Theoretischer Rahmen und empirische Befundlage

2 Theoretischer Rahmen und Empirische Befundlage

2.1 Diagnostische Kompetenz als Genauigkeit von Lehrerurteilen

2.1.1 Begriffsbestimmung

Die überwiegende Mehrheit der (vornehmlich) in den letzten anderthalb Jahrzehnten publizierten Forschungsarbeiten zur diagnostischen Kompetenz von Lehrkräften ist den *genauigkeitsorientieren Ansätzen* zuzuordnen (Schrader, 2009, 2013). Wie bereits skizziert, steht im Fokus dieser Ansätze die Genauigkeit bzw. Akkuratheit von Lehrerurteilen im Sinne der Übereinstimmung der Urteile mit der Realität (auch als „Veridikalität“ bezeichnet) (Schrader & Helmke, 1987). Konzeptionell dienen die in diesen Arbeiten ermittelten Genauigkeitskennwerte als Indikatoren für die diagnostische Kompetenz der jeweils untersuchten Lehrkräfte. Somit erfolgt de facto eine Gleichsetzung von diagnostischer Kompetenz und Urteilsgenauigkeit (Artelt & Gräsel, 2009; Geist, 2014; Lehmann & Hoffmann, 2009; Schrader, 2013). Folgerichtig wird im Kontext der genauigkeitsorientierten Ansätze die diagnostische Kompetenz von Lehrkräften als Fähigkeit definiert, „Personen oder Personengruppen (z. B. Schüler oder Schulklassen) zutreffend zu beurteilen bzw. genaue diagnostische Urteile abzugeben“ (Helmke, 2010, S. 121; vgl. auch Meyer, 2004; Schrader, 2006). Von einigen Autorinnen und Autoren wird diese genauigkeitsorientierte Konzeption der diagnostischen Kompetenz von Lehrkräften definitorisch noch um die Facette erweitert, bestimmte Aufgabenmerkmale bzw. -anforderungen adäquat beurteilen und insbesondere die Schwierigkeit von Aufgaben zutreffend einschätzen zu können (Anders et al., 2010; Artelt & Gräsel, 2009; McElvany et al., 2009; Südkamp, Möller & Pohlmann, 2008). In der angloamerikanischen Forschung zur Genauigkeit von Lehrerurteilen ist der Begriff der *diagnostic competence* praktisch nicht eingeführt. Hier finden sich stattdessen Termini wie *teacher judgment accuracy* (z. B. Hoge & Coladarci, 1989; Südkamp, Kaiser & Möller, 2012).

In den nachfolgenden Abschnitten dieses Teilkapitels (2.1) sollen theoretische Hintergründe und empirische Befunde zur genauigkeitsorientierten Konzeption der diagnostischen Kompetenz von Lehrkräften dargestellt werden. Der Begriff der diagnostischen Kompetenz wird dabei (wenn nicht anders vermerkt) jeweils entsprechend der skizzierten Gleichsetzung im Sinne von „Urteilsgenauigkeit“ verwendet. Zuvor sei noch darauf hingewiesen, dass in den Forschungsarbeiten, die den genauigkeitsorien-

tierten Ansätzen zuzurechnen sind, überwiegend Lehrerurteile zu Merkmalen untersucht werden, die dem *Leistungsbereich* zuzuordnen sind. Auch die in der vorliegenden kumulativen Dissertation integrierten Teilstudien und die Ausführungen in den nachfolgenden Abschnitten fokussieren jeweils auf Diagnosen, Urteile oder Entscheidungen, die (zumindest mittelbar) auf ein Leistungskriterium bezogen sind (z. B. Lehrerurteile zur Schwierigkeit von Deutsch- und Mathematikaufgaben in *Teilstudie 1* unter Kapitel 4). Des Weiteren finden sich in der Forschungsliteratur auch Forschungsarbeiten, in denen die Akkuratheit von Lehrerurteilen zu Merkmalen aus anderen Bereichen betrachtet wurde. Untersucht wurden hierbei unter anderem Einschätzungen zum allgemeinen Fachinteresse von Schülerinnen und Schülern (z. B. Karing, 2009), zur Lernfreude im Fach Mathematik, zur Aufmerksamkeit und subjektiv erlebten Unterforderung (z. B. Hosenfeld, Helmke & Schrader, 2002), zur Lernbereitschaft (Brunner et al., 2011), zur schulbezogenen Leistungsängstlichkeit, zur Lern- und Leistungsmotivation (z. B. Givvin, Stipek, Salmon & MacGyvers, 2001; Spinath, 2005; Urhahne, Timm, Zhu & Tang, 2013; Urhahne et al., 2010) sowie zu ausgewählten Persönlichkeitseigenschaften wie *sociability* oder *troublesomeness* (ter Laak, De Goede & Brugman, 2001). Stark verkürzt lassen sich die Ergebnisse dieser Studien so zusammenfassen, dass Lehrkräfte im Durchschnitt kaum dazu in der Lage sind, die Ausprägung von Merkmalen aus den genannten Bereichen akkurat zu beurteilen.

2.1.2 Bedeutung von Lehrerurteilen und ihrer Genauigkeit für die schulische Praxis und die angewandte Forschung

Entscheidungen zur Planung und Durchführung des Unterrichts sind in der Regel mit dem Ziel verbunden, für möglichst viele Schülerinnen und Schüler die bestmöglichen Lernbedingungen zu schaffen, damit diese die an sie gestellten Lernziele erreichen können. Eine entscheidende Bedingung hierfür ist, dass Lehrkräfte die unterrichtlichen Angebote, also ihre Instruktionen und ihr unterrichtliches Handeln, optimal an den Lernvoraussetzungen ihrer Schülerinnen und Schüler ausrichten (Corno & Snow, 1986; McElvany et al., 2009; Rogalla & Vogt, 2008). Solche Anpassungsprozesse können sowohl auf die gesamte Schulklasse als auch auf einzelne Schülerinnen und Schüler bezogen sein. Adaptationen auf der *Klassenebene* finden sich zum Beispiel beim Frontalunterricht. Dieser ist vor allem dann effektiv, wenn Aufgaben und Materialien ausgewählt werden, die für das Leistungsniveau der Schulklasse angemessen sind (z. B. Brunner et al., 2011). Adaptationen auf der *Individualebene* erfolgen insbesondere

bei einer individualisierten, differenzierenden Unterrichtsgestaltung. In deren Fokus steht die an den individuellen Lernständen und Lernbedarfen ausgerichtete Unterstützung, Begleitung und Förderung von Lernprozessen einzelner Schülerinnen und Schüler (vgl. Racherbäumer & Kühn, 2013; Vogt & Rogalla, 2009).

Vielfach wird der diagnostischen Kompetenz von Lehrkräften eine zentrale Bedeutung für die skizzierten Anpassungsprozesse zugeschrieben (z. B. Artelt & Gräsel, 2009; Begeny & Buchanan, 2010; Begeny, Eckert, Montarello & Storie, 2008; Begeny, Krouse, Brown & Mann, 2011; Eckert & Arbolino, 2007; Hurwitz, Elliott & Braden, 2007; McElvany et al., 2009; Meisels, Bickel, Nicholson, Xue & Atkins-Burnett, 2001). Dies wird damit begründet, dass die Adaptationen jeweils auf Lehrerurteilen basieren, die sich zum einen auf die Lernvoraussetzungen der Schülerinnen und Schüler beziehen (z. B. Einschätzungen zu Leistungsniveau, Leistungsheterogenität und Leistungsbereitschaft der Klasse, zu Lernausgangslage, Lernstand und Lernbedarf einzelner Schülerinnen und Schüler) und zum anderen auf die Schwierigkeit und die Anforderungen von Aufgaben bezogen sind, die im Unterricht bearbeitet werden. Dabei wird angenommen, dass die Adaptation umso besser gelingt, je höher die Genauigkeit der zugrunde liegenden Lehrerurteile ist.

In ganz ähnlicher Art und Weise argumentieren Brunner und Kollegen (2011) mit Blick auf das *Potenzial der kognitiven Aktivierung*, das in theoretischen Modellen als eine von drei Basisdimensionen der Unterrichtsqualität angeführt wird (Klieme, Pauli & Reusser, 2009; Kunter & Voss, 2011; Pianta & Hamre, 2009). Als kognitiv aktivierende Aufgaben gelten dabei insbesondere solche, die einerseits am Vorwissen der Schülerinnen und Schüler ansetzen, andererseits die vorhandenen Präkonzepte in Frage stellen und somit zu einer aktiven Auseinandersetzung mit den jeweiligen Inhalten anregen. Um solche Aufgaben formulieren bzw. identifizieren zu können, ist es notwendig, das Vorwissen der Schülerinnen und Schüler sowie die kognitiven Anforderungen bestimmter Aufgabenmerkmale zutreffend einzuschätzen (Brunner et al., 2011).

Die Genauigkeit von Lehrerurteilen ist außerdem im Zusammenhang mit der *konstruktiven (individuellen) Unterstützung* von Schülerinnen und Schülern (z. B. bei der Bearbeitung kognitiv aktivierender Aufgaben) von Bedeutung, die gleichfalls als Basisdimension der Unterrichtsqualität gilt (Klieme et al., 2009; Kunter & Voss, 2011; Pianta & Hamre, 2009). Eine solche Unterstützung erfordert ebenfalls, dass die Anforderungen und Schwierigkeiten von Aufgaben sowie das Vorwissen von

Schülerinnen und Schülern adäquat eingeschätzt werden. Darüber hinaus müssen Lehrkräfte in der Lage sein zu erkennen, ob ihre Schülerinnen und Schüler Verständnisprobleme haben (Brunner et al., 2011).

Inwiefern sich die Bedeutung der diagnostischen Kompetenz von Lehrkräften für die letztgenannte Basisdimension der Unterrichtsqualität empirisch belegen lässt, wurde in einer Studie von Westphal, Gronostaj, Vock, Emmrich und Harych (2016) geprüft, die an Gymnasien des Landes Brandenburg in der achten Jahrgangsstufe durchgeführt wurde. Hierbei konnte zumindest für den Mathematikunterricht ermittelt werden, dass Lehrerinnen und Lehrer, die die Leistungsstände ihrer Schülerinnen und Schüler akkurat einschätzen konnten, aus Schülersicht in höherem Maße differenziert unterrichteten als Lehrkräfte, deren Leistungsurteile durch eine geringe Genauigkeit gekennzeichnet waren.

In einer Reihe weiterer Forschungsarbeiten wurde der Frage nachgegangen, ob sich Zusammenhänge zwischen den diagnostischen Kompetenzen von Lehrkräften und den Leistungen ihrer Schülerinnen und Schüler nachweisen lassen. In einer 1987 publizierten Studie konnten Helmke und Schrader zwar keinen direkten Effekt der diagnostischen Kompetenz von Lehrkräften auf die Mathematikleistungen von Schülerinnen und Schülern der 5. und 6. Jahrgangsstufe feststellen, jedoch fanden sie Interaktionseffekte, die den Schluss nahe legten, dass die Genauigkeit von Lehrerurteilen die Effektivität bestimmter Instruktionstechniken erhöhen kann. So hatten Strukturierungsmaßnahmen (z. B. Maßnahmen zur Regulierung und Fokussierung der Aufmerksamkeit, Hervorhebung wichtiger Informationen, Hinweise zu geeigneten Methoden für die Bearbeitung von Arbeitsaufträgen) und Unterstützungsmaßnahmen (z. B. Beantworten von Fragen und Geben von Lösungshinweisen bei Stillarbeiten) vor allem dann einen statistisch bedeutsamen Effekt auf die Schülerleistungen, wenn sich Lehrkräfte außerdem durch eine hohe Urteilsgenauigkeit auszeichneten. Ein ganz ähnliches Ergebnis berichteten Behrmann und Souvignier (2013), die in einer Längsschnittstudie nur dann einen positiven Effekt der Häufigkeit von Leistungsfeedbacks durch Lehrkräfte auf den Kompetenzerwerb von Schülerinnen und Schülern im Lesen feststellen konnten, wenn diese gleichzeitig in der Lage waren, Schülerleistungen akkurat zu beurteilen.

Statistisch signifikante Zusammenhänge zwischen der diagnostischen Kompetenz von Lehrkräften einerseits und den Mathematikleistungen ihrer Schülerinnen und Schüler

andererseits wurden in einer anhand von Daten des COACTIV[6]-Projekts durchgeführten Untersuchung von Anders und Kollegen (2010) ermittelt. Die Schülerleistungen waren dabei umso höher, je besser die Lehrkräfte in der Lage waren, die Schwierigkeit von Mathematikaufgaben einzuschätzen. Darüber hinaus korrelierte die Urteilsgenauigkeit der Lehrkräfte positiv mit dem kognitiven Aktivierungspotenzial der von den untersuchten Lehrkräften erstellten Klassenarbeiten. Die Befundlage wurde von den Autorinnen und Autoren der Studie als Indiz dafür interpretiert, dass der positive Effekt der diagnostischen Kompetenz auf die Schülerleistung über ein höheres kognitives Aktivierungspotenzial der von den Lehrkräften gestellten Aufgaben und mithin durch eine höhere Unterrichtsqualität (vgl. Klieme et al., 2009; Kunter & Voss, 2011; Pianta & Hamre, 2009) vermittelt wird (vgl. auch Brunner et al., 2011). Statistisch signifikante Zusammenhänge zwischen der Urteilsgenauigkeit und den Schülerleistungen wurden auch in einigen weiteren Studien festgestellt (z. B. Fisher et al., 1978), wobei jedoch die ermittelten Effekte zumeist gering ausfielen (z. B. Lehmann et al., 2000; Weinert, F. E., Schrader & Helmke, 1990).

Die diagnostische Kompetenz von Lehrkräften bildet nicht nur eine wesentliche Grundlage für die Planung und Durchführung des Unterrichts. Auch bei der Konstruktion und Gestaltung von Leistungskontrollen sind genaue Urteile und insbesondere akkurate Einschätzungen der Schwierigkeit der in Frage kommenden Leistungsaufgaben von Bedeutung. So beinhaltet etwa eine didaktische Heuristik, dass, sofern möglich und inhaltlich sinnvoll, leichte Aufgaben eher zu Beginn und schwierige Aufgaben eher zum Ende von Leistungskontrollen gestellt werden sollten (vgl. Hascher, 2008). Auch sollte darauf geachtet werden, dass die verwendeten Aufgaben insgesamt schwierigkeitsadäquat sind, also im Mittel weder zu leicht noch zu schwer ausfallen.

Des Weiteren können bestimmte Lehrerurteile (und deren Genauigkeit) auch weitreichende Konsequenzen für die schulischen Laufbahnen und die späteren Lebenswege von Schülerinnen und Schülern haben (Begeny et al., 2008; Feinberg & Shapiro, 2003; Trautwein & Baeriswyl, 2007). Eine wichtige Rolle kommt den Einschätzungen von Lehrkräften beispielsweise bei der Wahl der weiterführenden Schule am Ende der Grundschulzeit zu. Grundschullehrkräfte haben hierbei die Aufgabe, eine Übergangsempfehlung auszusprechen, also die Schulart zu benennen, die ihrem Urteil

[6] *Co*gnitive *Activ*ation in the Classroom.

nach für die jeweilige Schülerin bzw. den jeweiligen Schüler am besten geeignet ist. Zumeist ist diese Übergangsempfehlung wesentlich für die Entscheidung über die zukünftige Sekundarschule: In vielen Fällen wird sie von den Eltern der Kinder übernommen (Stubbe & Bos, 2008), in einigen Bundesländern ist sie sogar bindend (Gräsel, Krolak-Schwerdt, Nölle & Hörstermann, 2010).

Zu den Pflichten von Lehrkräften gehört auch, frühzeitig darauf hinzuweisen, wenn sie bei einer Schülerin oder einem Schüler besondere Begabungen oder Lernbeeinträchtigung vermuten, die Eltern und Kinder in diesem Fall adäquat zu beraten und Empfehlungen für angemessene Fördermaßnahmen auszusprechen (z. B. Begeny et al., 2011; Beswick, Willms & Sloat, 2005; Feinberg & Shapiro, 2009; Graney, 2008; Gresham, MacMillan & Bocian, 1997; Hoge, 1983; Hoge & Butcher, 1984; Madelaine & Wheldall, 2005; Neber & Heller, 2004; Ready & Wright, 2011; Teisl, Mazzocco & Myers, 2001; Ysseldyke, Vanderwood & Shriner, 1997). Auch hierbei ist die diagnostische Kompetenz relevant: Je akkurater Lehrerinnen und Lehrer die Leistungen und Kompetenzausprägungen von Schülerinnen und Schüler beurteilen können, umso besser sollten sie in der Lage sein, Hinweise auf besondere Förderbedarfe wahrzunehmen.

Die hohe Bedeutung, die der Genauigkeit von Lehrerurteilen für die schulische Praxis zukommt, ergibt sich nicht nur aus den verschiedenen diagnostischen Aufgaben, die Lehrkräfte in ihrem Beruf bewältigen müssen. Lehrerurteile können darüber hinaus auch Auswirkungen auf weitere Faktoren haben, die im Zusammenhang mit dem Lernerfolg in der Schule stehen. So beeinflussen sie unter anderem die Erwartungen, die Lehrkräfte an die zukünftigen Leistungen ihrer Schülerinnen und Schüler haben (Begeny et al., 2008; Brophy & Good, 1986). Diese Leistungserwartungen können dann wiederum einen Effekt auf die tatsächliche Leistungsentwicklung von Schülerinnen und Schülern haben. Das kann insbesondere dann problematisch sein, wenn auf der Grundlage ungenauer Urteile inadäquate, negative Erwartungen an die Leistungsentwicklung von Schülerinnen und Schülern entstehen, die dann im Sinne selbsterfüllender Prophezeiungen Wirklichkeit werden (z. B. de Boer, Bosker & van der Werf, 2010; Jussim & Eccles, 1992; Jussim & Harber, 2005; Rosenthal & Rubin, 1978). Ein solcher Erwartungseffekt könnte etwa dadurch vermittelt sein, dass Lehrkräfte ihre Schülerinnen und Schüler in Abhängigkeit von ihren Erwartungen in jeweils unterschiedlichem Maße unterstützen, fördern und fordern, so dass in der Folge hoch eingeschätzte bessere Möglichkeiten zur Entwicklung und Festigung ihrer Kompetenzen erhalten als niedrig

eingeschätzte (Brophy, 1983; Jussim & Harber, 2005). Schülerseitig kann darüber hinaus die Wahrnehmung, anders als andere Schülerinnen und Schüler behandelt zu werden, das Lernverhalten beeinflussen (Brattesani, Weinstein & Marshall, 1984). Darüber hinaus wurde wiederholt gezeigt, dass auch Lehrerurteile bzw. Leistungsrückmeldungen an sich einen Effekt auf leistungsrelevante Faktoren wie Lernmotivation, akademisches Selbstkonzept, Selbstwirksamkeitserwartung und Anstrengungsbereitschaft von Schülerinnen und Schüler haben können (Baumert, Trautwein & Artelt, 2003; Brookhart, 1997; Rodriguez, 2004; Trautwein, Lüdtke, Köller, Marsh & Baumert, 2006).

Schließlich sei angemerkt, dass Lehrerurteile nicht nur in der schulischen Praxis allgegenwärtig, sondern auch für die angewandte Forschung bedeutsam sind. So werden Leistungsurteile von Lehrkräften nicht selten herangezogen, um Kennwerte für die kriteriale Validität (konvergente, äußere Kriteriumsvalidität) von standardisierten Leistungstests zu bestimmen, die etwa auf die Erfassung der mathematischen oder sprachlichen Kompetenzen von Kindern und Jugendlichen zielen (z. B. Gölitz, Roick & Hasselhorn, 2006; May, 2002). Auch in der Evaluationsforschung wird auf Einschätzungen von Lehrkräften zurückgegriffen – zum Beispiel bei der Bewertung der Wirksamkeit spezieller Maßnahmebündel, die auf eine Stärkung bestimmter Schülerkompetenzen zielen (Hoge, 1983).

2.1.3 Diagnostische Güte von Schulnoten

Bislang nicht thematisiert wurde die vielleicht augenfälligste Beurteilungsaufgabe von Lehrkräften – die Vergabe von Schulnoten. Diese können weitreichende Folgen für die Bildungskarrieren von Schülerinnen und Schülern haben, da ihnen insbesondere die Funktion eines zentralen Selektions- und Allokationsinstruments zukommt (Krampen, 1984; Ziegenspeck, 1999): Zum Beispiel sind gute Noten (zumindest in mehrgliedrigen Schulsystemen) Voraussetzung für die Zuweisung zu bestimmten Schularten, schlechte Noten können zur Folge haben, dass eine Klasse wiederholt werden muss. Im Abschlusszeugnis behalten Schulnoten diese Funktion auch über das Ende der Schullaufbahn hinaus: Gute Noten erlauben etwa den Zugang zu bestimmten Ausbildungswegen und Arbeitsplätzen, schlechte Noten können ihn verhindern (z. B. Dünnebier, Gräsel & Krolak-Schwerdt, 2009; Krolak-Schwerdt, Böhmer & Gräsel, 2009).

Allerdings dienen Schulnoten nicht allein der Selektion, sondern häufig auch noch anderen Zwecken. Ihnen kommt beispielsweise eine Informationsfunktion zu, indem sie

Schülerinnen und Schülern sowie deren Eltern, aber auch anderen schulischen Akteuren wie Erzieherinnen und Erziehern oder Schulpsychologinnen und Schulpsychologen Rückmeldungen zum Leistungsstand und zur Leistungsentwicklung geben (vgl. Hoge & Coladarci, 1989; Ziegenspeck, 1999). Darüber hinaus können mit der Notenvergabe bestimmte pädagogische Absichten, wie etwa eine zusätzliche Motivierung von Schülerinnen und Schülern, intendiert sein (vgl. Brookhart, 1993; Jäger & Petermann, 1999). Auch kann die Notengebung von dem Motiv beeinflusst sein, Aspekte wie positive Leistungsentwicklungen (z. B. Krampen, 1984; Schröder, 2000), die wahrgenommene Selbststeuerung bzw. die Arbeitsweise, die Selbstdisziplin und die Anstrengungsbereitschaft von Schülerinnen und Schülern (z. B. Kuhl & Hannover, 2012) besonders belohnen zu wollen. Allerdings sind eine Berücksichtigung der letztgenannten Aspekte bei der Notengebung und das damit verbundene Abweichen vom reinen Leistungsprinzip nicht vollkommen unproblematisch. Es erscheint zwar durchaus legitim, bei der Notenvergabe auch pädagogische Motive einzubeziehen und zum Beispiel positive Leistungsentwicklungen zu honorieren. Dies sollte jedoch in einem angemessenen Maße geschehen und nicht dazu führen, dass etwa Schülerinnen und Schüler, die Leistungen auf einem konstant hohem Niveau zeigen und sich daher kaum noch verbessern können, relativ benachteiligt werden (z. B. Dünnebier et al., 2009).

Über viele Jahrzehnte hinweg haben sich empirische Untersuchungen zu den diagnostischen Fähigkeiten von Lehrkräften vor allem auf die Notengebung konzentriert. Die Forschung zur diagnostischen Güte von Schulnoten reicht dabei bis ins 19. Jahrhundert zurück (vgl. Kronig, 2015). Die Befundlage hierzu lässt sich unter dem knappen Resümee zusammenfassen, dass Schulnoten vielfach nicht die zentralen Testgütekriterien erfüllen, also nicht objektiv, reliabel und valide sind. Unter anderem konnte gezeigt werden, dass sie oftmals nicht kriterial, das heißt nicht bezogen auf ein absolutes Kriterium, wie zum Beispiel auf das Erreichen vorher festgelegter Kompetenzziele, erteilt werden. Stattdessen ist die Notenvergabe häufig an einer sozialen Bezugsnorm orientiert; sie erfolgt also zum Beispiel in Relation zu den Leistungen der anderen Schülerinnen und Schüler in der Klasse (Davis, 1966; Ingenkamp, 1995; Ingenkamp & Lissmann, 2008; Lintorf, 2012; Rheinberg, 2001). Wenn Schulnoten sozialnormorientiert vergeben werden, dann können sie zwar prinzipiell noch immer akkurate Urteile zur Leistungsverteilung in der jeweiligen Referenzgruppe darstellen. Sie sind jedoch auf das Leistungs-

niveau der Referenzgruppe bezogen und somit als absolute Einschätzungen von Schülerleistungen per se ungenau.

Neben der Sozialnormorientierung sind in der Forschungsliteratur zudem zahlreiche Befunde zu leistungsfernen Faktoren dokumentiert, für die in empirischen Studien ein (jeweils in Richtung der hierzu bestehenden Stereotype) verzerrender Effekt auf die Notengebung festgestellt wurde. Zu diesen Faktoren zählen unter anderem das Geschlecht der Schülerinnen und Schüler (Lintorf, 2012; Schreiner, Breit & Haider, 2008), der sozioökonomische Status ihrer Eltern (Cicmanec, Johanson & Howley, 2001), die Sympathie der Lehrkraft für eine Schülerin oder einen Schüler (Hadley, 1954, zit. nach Hadley, 1995), die Länge von Schulaufsätzen (Birkel & Birkel, 2002) oder die Sauberkeit der Handschrift (Osnes, 1995).

Des Weiteren wird die Fragwürdigkeit der Zensurengebung (z. B. Ingenkamp, 1995) durch eine Reihe von Untersuchungen illustriert, in denen Schulnoten mit den in standardisierten Leistungstests erzielten Leistungen verglichen wurden. Ingenkamp (1995) fand zum Beispiel, dass die Mathematiknoten von Schülerinnen und Schülern, die in einem Rechentest identisch abschnitten, erheblich variierten. Konsistent dazu wurde später auch in internationalen Schulleistungsstudien sowohl für den Primar- (PIRLS[7]/IGLU[8]) als auch für den Sekundarbereich (PISA) ermittelt, dass Schülerinnen und Schüler, für die auf Basis der gezeigten Testleistungen die jeweils gleiche Kompetenzstufe bestimmt wurde, zum Teil sehr unterschiedliche Zeugnisnoten in den betreffenden Schulfächern erhielten (Baumert et al., 2003; Bos et al., 2003; vgl. auch Lorenz & Artelt, 2009).

Welche Schlussfolgerungen lassen sich aus der skizzierten Befundlage zur diagnostischen Güte von Schulnoten für die diagnostischen Kompetenzen von Lehrkräften ziehen? An verschiedener Stelle wird bezweifelt, dass die Ergebnisse von Untersuchungen zu Schulnoten fundierte Rückschlüsse auf die Genauigkeit von Lehrerurteilen erlauben (z. B. Lintorf, 2012; Lorenz & Artelt, 2009). Hierbei wird argumentiert, dass die Notengebung, wie bereits weiter oben dargestellt, auch mit bestimmten pädagogischen Motiven verbunden ist und daher im Notenurteil neben der Schülerleistung vielfach noch weitere Aspekte Berücksichtigung finden. Folglich bilden

[7] *P*rogress in *I*nternational *R*eading *L*iteracy *S*tudy.

[8] *I*nternationale *G*rundschul-*L*ese-*U*ntersuchung.

Schulnoten häufig keine reinen Leistungsurteile ab und erscheinen dementsprechend wenig geeignet, um auf ihrer Basis zu untersuchen, wie akkurat Lehrkräfte die Ausprägung von bestimmten Leistungsmerkmalen bei ihren Schülerinnen und Schülern beurteilen können (Lorenz & Artelt, 2009).

2.1.4 Operationalisierung und Erfassung der Genauigkeit von Lehrerurteilen

Aus den skizzierten Gründen wird die Genauigkeit von Lehrerurteilen mittlerweile nur noch selten anhand von Schulnoten untersucht. Stattdessen werden Lehrerurteile zu direkteren, spezifischeren Leistungskriterien betrachtet. Paradigmatischen Charakter hat dabei folgendes Vorgehen: Die an einer Studie teilnehmenden Lehrerinnen und Lehrer werden in einem ersten Schritt aufgefordert, ihre Schülerinnen und Schüler hinsichtlich bestimmter Leistungskriterien zu beurteilen, wobei sie zumeist darum gebeten werden, das Abschneiden in einem standardisierten Leistungstest oder bei bestimmten Testaufgaben zu prognostizieren. Die somit erfassten Lehrerurteile werden in einem zweiten Schritt mit den tatsächlichen Ergebnissen der Schülerinnen und Schüler in dem betreffenden Test bzw. bei den jeweiligen Aufgaben verglichen. Hierbei wird von dem Grad an Übereinstimmung bzw. anhand der Korrespondenz von prognostizierten und tatsächlichen Leistungen auf die Urteilsgenauigkeit resp. auf die diagnostische Kompetenz einer Lehrkraft geschlossen (Hoge & Coladarci, 1989; Schrader & Helmke, 1987; Südkamp et al., 2012).

Ein analoges Vorgehen findet sich in Forschungsarbeiten, in denen statt der Genauigkeit von Lehrerurteilen zu Schülerleistungen die Akkuratheit von Einschätzungen zu Aufgabenschwierigkeiten untersucht wird (z. B. Anders et al., 2010; Brunner et al., 2011; McElvany et al., 2009): Hierbei werden Lehrkräfte um Beurteilungen dazu gebeten, ob bestimmte Aufgaben für die Schülerinnen und Schüler ihre Klasse eher leicht oder eher schwer zu lösen sein werden. Ihre Einschätzungen werden dann mit den empirisch ermittelten Aufgabenschwierigkeiten verglichen, die zumeist, im Sinne der klassischen Testtheorie (z. B. Lienert & Raatz, 1998), als Lösungshäufigkeiten operationalisiert werden.

Auch wenn sich die meisten Studien zur Urteilsgenauigkeit von Lehrerinnen und Lehrern in dem skizzierten Vorgehen ähneln, weisen sie im Detail methodische Unterschiede auf. Viele dieser Unterschiede betreffen die *Art der jeweils erhobenen Lehrerurteile*. Je nach Forschungsarbeit differieren diese beispielsweise hinsichtlich des

Grads an Direktheit bzw. Informiertheit, das heißt bezüglich des Ausmaßes, mit dem Lehrkräfte über den Vergleichsmaßstab informiert sind, zu dem ihre Urteile ins Verhältnis gesetzt werden. Informierte Urteile liegen etwa dann vor, wenn Lehrkräfte explizit angeben sollen, wie viele Items ihre Schülerinnen und Schüler in einem Leistungstest korrekt lösen können werden. Demgegenüber weisen zum Beispiel Urteile, bei denen Lehrerinnen und Lehrer die Ergebnisse ihrer Schülerinnen und Schüler in einem Test anhand einer Rating-Skala beurteilen sollen, einen geringeren Grad an Informiertheit auf (Hoge & Coladarci, 1989; Südkamp et al., 2012).

Eng verbunden mit der Informiertheit ist der Aspekt der Urteilsspezifität. Prognosen zu Testleistungen, die auf einer verbalen Rating-Skala vorgenommen werden, aber auch Leistungseinschätzungen, bei denen Schülerinnen und Schüler in eine Rangreihe gebracht werden, gelten als wenig spezifisch. Eine hohe Spezifität weisen hingegen Lehrerurteile auf, bei denen für einzelne Schülerinnen und Schüler eingeschätzt wird, ob diese in der Lage sein werden, eine bestimmte Aufgabe korrekt zu lösen oder nicht (Hoge & Coladarci, 1989; Südkamp et al., 2012). Lehrerurteile zur Aufgabenschwierigkeit, bei denen die Lösungshäufigkeiten für einzelne Aufgaben eingeschätzt werden, sind also nach dieser Logik durch eine recht hohe Spezifität gekennzeichnet.

Die in Untersuchungen zur diagnostischen Kompetenz von Lehrkräften betrachteten Urteile variieren außerdem hinsichtlich der sogenannten Domänenspezifität. Dieser Aspekt betrifft die Frage, ob Lehrkräfte entweder aufgefordert sind, eine eher allgemeine Einschätzung von Schülerleistungen vorzunehmen (z. B. Leistung in einem bestimmten Unterrichtsfach wie Mathematik), oder ob sie Leistungen in einer spezifischen Domäne (z. B. arithmetischen Fähigkeiten) beurteilen sollen (Südkamp et al., 2012).

Forschungsarbeiten zur diagnostischen Kompetenz von Lehrkräften differieren darüber hinaus in der Art und Weise, mit der die *Korrespondenz* zwischen den Lehrerurteilen einerseits und den empirisch gefundenen Schülerleistungen andererseits ermittelt wird. In der überwiegenden Mehrzahl der Studien (vor allem aus dem angloamerikanischen Sprachraum) wird diese Korrespondenz mithilfe von Korrelationskoeffizienten berechnet (Hoge & Coladarci, 1989; Südkamp et al., 2012). Eine solche korrelative Bestimmung der Korrespondenz bedingt, dass die ermittelten Genauigkeitskennwerte relative Maße sind: Sie zeigen an, inwiefern die Rangreihe der von einer Lehrerin oder einem Lehrer vorgenommenen Leistungseinschätzungen mit der Rangreihe der Einzelleistungen übereinstimmt, die von einer Gruppe aus Schülerinnen

und Schülern, beispielsweise in einem Leistungstest, erreicht werden. Den Grad an absoluter Übereinstimmung von Lehrerurteilen und Schülerleistungen können korrelative Genauigkeitskennwerte allerdings nicht abbilden. So können hohe Korrelationskoeffizienten beispielsweise auch dann resultieren, wenn die Leistungen von Schülerinnen und Schülern systematisch über- oder unterschätzt werden (Coladarci, 1986; Eckert & Arbolino, 2007; Feinberg & Shapiro, 2003; Flynn & Rahbar, 1998; Graney, 2008; Kenny & Chekaluk, 1993; Salvesen & Undheim, 1994). Oder anders ausgedrückt: Wenn die Genauigkeit von Lehrerurteilen ausschließlich anhand von Korrelationen untersucht wird, dann besteht die Gefahr, die diagnostische Kompetenz von Lehrkräften zu überschätzen (Begeny & Buchanan, 2010; Begeny et al., 2008; Begeny et al., 2011).

In der deutschsprachigen Forschungsliteratur zur Genauigkeit von Lehrerurteilen hat sich die Unterscheidung von drei Genauigkeitsfacetten bzw. -komponenten etabliert. Eingeführt wurde diese Differenzierung von Schrader und Helmke (1987). Sie geht allerdings auf Konzepte von Cronbach (1955) zurück, der zeigen konnte, dass verschiedene Urteilsfehler miteinander vermischt werden, wenn die Urteilsgenauigkeit allein anhand der Abweichungen ermittelt wird, die zwischen den prognostizierten und den tatsächlichen Testleistungen bestehen.

Eine der drei Genauigkeitsfacetten ist die sogenannte *Rangkomponente*, die in einigen Forschungsarbeiten auch als diagnostische Sensitivität bezeichnet wird (z. B. Anders et al., 2010; Brunner et al., 2011). Sie bildet die korrelative Übereinstimmung zwischen Lehrerurteilen und Schülerleistungen in der bereits beschriebenen Weise ab (Schrader & Helmke, 1987).

Eine zweite Genauigkeitsfacette in der Konzeption von Schrader und Helmke (1987) stellt die *Niveaukomponente* dar. Im Unterschied zur Rangkomponente bildet diese nicht die relative, sondern die absolute Übereinstimmung zwischen den Leistungsurteilen von Lehrkräften und den tatsächlichen Leistungen ihrer Schülerinnen und Schüler ab. Als Kennwerte für die Niveaukomponente werden in der Forschungsliteratur oftmals der Urteilsfehler oder die Urteilstendenz bestimmt, die beide Differenzmaße sind. Der Urteilsfehler ist als Mittelwert der Absolutbeträge der Differenzen von Lehrerurteilen und Schülerleistungen definiert und gibt somit das mittlere Ausmaß an Fehleinschätzungen in einer Lehrerstichprobe an. Die Urteilstendenz wird als Mittelwert der einfachen Differenzen von Lehrerurteilen und Schülerleistungen berechnet. Sie zeigt folglich auch an, ob Lehrkräfte im Mittel eher zur Über- oder Unterschätzung der Leistungen ihrer

Schülerinnen und Schüler in einem bestimmten Leistungskriterium neigen (z. B. Brunner et al., 2011; McElvany et al., 2009).

Als eine weitere Variante der Niveaukomponente können prozentuale Übereinstimmungsmaße gelten, die sich bestimmen lassen, wenn Lehrerurteile oder Schülerleistungen nicht auf metrischen, sondern auf kategorialen Skalen vorliegen oder in Kategorien überführt werden. Solche Übereinstimmungsmaße können zum Beispiel abbilden, wie viel Prozent der untersuchten Lehrkräfte die Leistungen ihrer Schülerinnen und Schüler jeweils korrekt eingeschätzt bzw. über- oder unterschätzt haben (Begeny & Buchanan, 2010; Begeny et al., 2008; Begeny et al., 2011; Eckert & Arbolino, 2007; Leinhardt, 1983). Kritisch anzumerken ist, dass eine Transformation von metrischen zu kategorialen Daten stets mit einem Informationsverlust verbunden ist. Außerdem ist zu vermuten, dass es nicht immer leicht sein dürfte, theoretisch und methodisch begründet zu entscheiden, welche Abweichungen zwischen Urteil und Empirie noch toleriert werden können oder aber als Über- oder Unterschätzung zu kategorisieren sind (vgl. Coladarci, 1986; Südkamp et al., 2012).

Die dritte und letzte Facette der Urteilsgenauigkeit in der Konzeption von Schrader und Helmke (1987) ist die sogenannte *Differenzierungskomponente*. Sie bildet ab, zu welchem Grad Lehrkräfte in der Lage sind, Leistungsstreuungen akkurat zu beurteilen. Bestimmt wird die Differenzierungskomponente zumeist über den Quotienten der Streuungen (in der Regel Standardabweichungen) von Lehrerurteilen und Schülerleistungen. Wenn dieser Quotient größer als 1 ist, zeigt dies eine Überschätzung der Leistungsstreuung durch die Lehrkraft an. Ein Quotient kleiner 1 indiziert hingegen eine Unterschätzung.

Angemerkt sei, dass sich in der Forschungsliteratur auch Studien finden, in denen der Frage nachgegangen wurde, inwiefern die Unterscheidung von Rang-, Niveau- und Differenzierungskomponente empirisch belegt werden kann. In diesen Studien wurden konsistent geringe Interkorrelationen zwischen den drei Genauigkeitsfacetten ermittelt (z.B. Anders et al., 2010; McElvany et al., 2009; Schrader, 1989; Spinath, 2005), was als empirischer Beleg für die Validität der Unterscheidung anzusehen ist (vgl. Rjosk, McElvany, Anders & Becker, 2011).

2.1.5 Empirische Ergebnisse zur Genauigkeit von Lehrerurteilen

Einen systematischen Überblick über die Forschungsarbeiten, in denen die Ausprägung der *Rangkomponente* der Genauigkeit von Lehrerurteilen untersucht wurde, liefern die Metaanalysen von Hoge und Coladarci (1989) und von Südkamp und Kollegen (2012). In der erstgenannten Publikation fanden insgesamt 16 Primärstudien Berücksichtigung, die in den Jahren von 1971 und 1988 publiziert wurden. Auf Basis der in diesen Studien festgestellten Ergebnisse berechneten Hoge und Coladarci (1989) zwischen den Leistungsurteilen von Lehrkräften einerseits und den empirisch gefundenen Schülerleistungen andererseits eine mittlere Korrelation (Median) von 0.66. Außerdem ermittelten sie, dass die in den einzelnen Studien berichteten Korrelationskoeffizienten zum Teil erheblich differieren, nämlich eine Spannweite von 0.64 bei einem Minimum von 0.28 und einem Maximum von 0.92 aufweisen.

Durch eine sehr viel breitere Datenbasis zeichnet sich die Metaanalyse von Südkamp und Kollegen (2012) aus, die auf den Ergebnissen von insgesamt 75 Primärstudien basiert und in der neben Forschungsarbeiten, in denen explizit die Genauigkeit von Lehrerurteilen untersucht wurde, auch Studien Berücksichtigung fanden, in denen Lehrerurteile lediglich herangezogen wurden, um die Kriteriumsvalidität standardisierter Testverfahren zu prüfen. Über die in die Metaanalyse einbezogenen Untersuchungen hinweg wurde eine mittlere Korrelation zwischen Lehrerurteilen und Schülerleistungen von 0.66 (arithmetisches Mittel) bzw. 0.53 (Median) berechnet. Analog zu den Ergebnissen von Hoge und Coladarci (1989) wiesen diese Kennwerte auch in der Metaanalyse von Südkamp und Kollegen (2012) eine hohe Varianz (Spannweite: 0.87, Minimum: –0.03, Maximum: 0.84) auf.

In Forschungsarbeiten zur Ausprägung der Rangkomponente der Genauigkeit von Lehrerurteilen wurden außerdem konsistent hohe interindividuelle Unterschiede bei den für die einzelnen Lehrkräfte ermittelten Korrelationskoeffizienten festgestellt (z. B. Helmke & Schrader, 1987; Hopkins, George & Williams, 1985; Hosenfeld et al., 2002; Karing, Matthäi & Artelt, 2011). Diese Unterschiede verdeutlichen, dass es durchaus Lehrerinnen und Lehrer gibt, deren Rangurteile sehr genau ausfallen. Gleichzeitig legen sie den Schluss nahe, dass es aber auch vielen Lehrkräften nicht zu gelingen scheint, Leistungsrangreihen ihrer Schülerinnen und Schüler akkurat zu beurteilen.

In der Forschungsliteratur finden sich auch einige Studien, in denen explizit die *Rangkomponente* der Genauigkeit von Lehrerurteilen zur *Schwierigkeit von Aufgaben* untersucht wurde. Tendenziell fallen die in diesen Untersuchungen gefundenen Korrelationen etwas geringer aus, als die gemittelten Koeffizienten, die in den Metaanalysen zur Genauigkeit von Leistungsurteilen zu Schülermerkmalen (Hoge & Coladarci, 1989; Südkamp et al., 2012) bestimmt wurden, liegen aber im Bereich der dort berichteten Spannweiten. Für Schwierigkeitseinschätzungen von Aufgaben zu Texten mit instruktionalen Bildern ermittelten McElvany und Kollegen (2009) eine Rangkomponente von $r = 0.50$. Hosenfeld und Kollegen (2002) fanden, ebenfalls für Mathematikaufgaben, eine Rangkorrelation von 0.56 zwischen den Schwierigkeitseinschätzungen von Lehrkräften und den empirischen Lösungshäufigkeiten der beurteilten Aufgaben. Für Schwierigkeitsurteile zu Leseaufgaben konnten Karing und Kollegen (2011) eine Rangkomponente in Höhe von $r = 0.20$ bestimmen.

In Forschungsarbeiten zur *Niveaukomponente* der Urteilsgenauigkeit wurde mehrheitlich ermittelt, dass Lehrerinnen und Lehrer die Leistungen ihrer Schülerinnen und Schüler im Durchschnitt eher überschätzen (z. B. Eckert, Dunn, Codding, Begeny & Kleinmann, 2006; Feinberg & Shapiro, 2009; Hachfeld, Anders, Schroeder, Stanat & Kunter, 2010; Hamilton & Shinn, 2003; Martin & Shapiro, 2011; Rjosk et al., 2011; Südkamp & Möller, 2009). Auch zeigte sich, dass es Lehrkräften besonders schwer zu fallen scheint, leistungsschwache Schülerinnen und Schüler korrekt zu identifizieren (z. B. Artelt et al., 2001; Begeny et al., 2008; Begeny et al., 2011; Gresham et al., 1997; Madelaine & Wheldall, 2005). Allerdings liegen auch einige wenige Studien vor, in denen eine mittlere Tendenz zur Unterschätzung von Schülerleistungen festgestellt wurde (z. B. Doherty, J. & Conolly, 1985; McElvany et al., 2009).

Kennwerte für die *Niveaukomponente* der Genauigkeit von Urteilen, bei denen Lehrkräfte explizit die *Schwierigkeit einzelner Aufgaben einschätzen* sollten, wurden zum Beispiel anhand von Daten aus der COACTIV-Studie (s. o.) von Anders und Kollegen (2010) sowie von Brunner et al. (2011) bestimmt. Hierbei tendierten die untersuchten Lehrkräfte im Mittel zu einer Unterschätzung der Aufgabenschwierigkeiten (bzw. zu einer Überschätzung der Lösungshäufigkeiten in der eigenen Klasse). Gleichzeitig wurden aber auch große interindividuelle Unterschiede in den Urteilstendenzen festgestellt, wobei einige wenige Untersuchungsteilnehmerinnen und -teilnehmer sogar zu einer Überschätzung der Schwierigkeit neigten. Auch Hosenfeld und Kollegen (2002)

sowie Lehmann und Kollegen (2000) fanden jeweils eine mittlere Tendenz zur Unterschätzung von Aufgabenschwierigkeiten (bei ebenfalls großen Streuungen). Demgegenüber wurde in den Studien von McElvany und Kollegen (2009) und von Lintorf (2012), in denen Lehrkräfte die Schwierigkeit von Aufgaben zu Texten mit instruktionalen Bildern beurteilen sollten, eine Überschätzung von Aufgabenschwierigkeiten ermittelt.

Insgesamt konnten für die Niveaukomponente der Urteilsgenauigkeit empirisch zumeist deutlich geringere Korrespondenzen zwischen Lehrerurteilen und Schülerleistungen gefunden werden als bei einer Betrachtung der Rangkomponente. In Übereinstimmung mit den Ergebnissen von Untersuchungen zur Vergabe von Schulnoten argumentieren Feinberg und Shapiro (2009) in diesem Zusammenhang, dass es Lehrkräfte aus der schulischen Praxis gewohnt seien, ihre Leistungsurteile stark an der relativen Position der Schülerin bzw. des Schülers innerhalb der Leistungsrangreihe ihrer Klasse – also eher an einer sozialen als an einer kriterialen Bezugsnorm – zu orientieren. Aus diesem Grund würden ihnen Rangurteile (Rangkomponente) etwas leichter fallen als absolute Leistungsurteile (Niveaukomponente).

Kennwerte für die *Differenzierungskomponente* der Genauigkeit von Lehrerurteilen wurden bislang nur selten bestimmt. Insgesamt fallen die Ergebnisse der wenigen hierzu veröffentlichten Forschungsarbeiten uneinheitlich aus (vgl. Schrader, 2013; van Ophuysen, 2010). Eine Tendenz zur Überschätzung der Streuung von Schülerleistungen wurde in den Studien von Seeber (2009) sowie von Urhahne und Kollegen (2010) ermittelt. Hingegen berichteten Südkamp und Möller (2009), die die Genauigkeit von Lehrerurteilen experimentell mithilfe einer Computersimulation, dem so genannten virtuellen Klassenzimmer untersuchten, eine Tendenz zur Unterschätzung der Variabilität von Schülerleistungen. In den Studien von Karing und Kollegen (2011) sowie von Brunner und Kollegen (2011) waren die Lehrkräfte hingegen recht gut in der Lage, die Streuung von Schülerleistungen zu prognostizieren.

Kennwerte für die *Differenzierungskomponente* der Genauigkeit von Lehrerurteilen, bei denen explizit die *Schwierigkeit einzelner Aufgaben* beurteilt werden sollte, wurden in der Untersuchung von Lintorf (2012) bestimmt. Im Mittel wurde hierbei eine Tendenz zur Unterschätzung der Streuung von Aufgabenschwierigkeiten festgestellt.

Eine bezogen auf die Systematik von Rang-, Niveau- und Differenzierungskomponente besondere Rolle nimmt die Untersuchung von Graney (2008) ein. Im Fokus dieser Studie standen Lehrerurteile, die auf einer individuellen Bezugsnorm beruhten. Die untersuchten Lehrerinnen und Lehrer hatten die Aufgabe, den Lernfortschritt von ausgewählten Schülerinnen und Schülern ihrer Klassen im Lesen für einen Zeitraum von sechs Wochen und über mehrere Messzeitpunkte hinweg zu prognostizieren. Nur wenige der untersuchten Lehrkräfte waren in der Lage, die Lernfortschritte ihrer Schülerinnen und Schüler akkurat vorherzusagen, wobei im Mittel eine Tendenz zur Überschätzung der Leistungsentwicklung festgestellt wurde.

Die skizzierte Befundlage, nach der Lehrerurteile nicht selten ungenau sind und es bei der Beurteilung von Schülerleistungen und Aufgabenschwierigkeiten häufiger zu Fehleinschätzungen kommt, wird zumeist als Beleg dafür interpretiert, dass die diagnostischen Fähigkeiten von Lehrkräften vielfach eher gering sind und daher im Rahmen der Aus- und Fortbildung von Lehrkräften stärker gefördert werden sollten. Als ein zusätzlicher Erklärungsansatz wird zudem häufig die Vermutung formuliert, dass Lehrkräfte eher kompetenz- als performanzorientiert urteilen, also Faktoren, die die Leistungen ihrer Schülerinnen und Schüler in einem Test mindern können (z. B. Aufregung, Stress oder Zeitdruck bei der Testbearbeitung, Flüchtigkeitsfehler, Vergessen des bereits behandelten Lernstoffs, mangelnde Anstrengung, Leistungsangst), bei ihren Einschätzungen nicht hinreichend berücksichtigen (z .B. Hosenfeld et al., 2002; Karing et al., 2011; Schrader, 1989; Urhahne et al., 2013).

2.1.6 Bedingungen und Kovariaten der Genauigkeit von Lehrerurteilen

2.1.6.1 Modell der Genauigkeit von Lehrerurteilen

Im vorherigen Abschnitt wurde skizziert, dass ein zentraler Befund der Forschung zur diagnostischen Kompetenz von Lehrkräften darin besteht, dass die Genauigkeit von Lehrerurteilen interindividuell stark variiert (z. B. Anders et al., 2010; Karing et al., 2011). Zudem verdeutlichen die Ergebnisse der angeführten Metaanalysen, dass auch zwischen verschiedenen Forschungsarbeiten deutliche Unterschiede in der Ausprägung der jeweils berichteten Genauigkeitskennwerte bestehen (vgl. Hoge & Coladarci, 1989; Südkamp et al., 2012). Vor dem Hintergrund dieser Heterogenität findet sich in der Forschungsliteratur eine wachsende Anzahl an Studien, die nicht nur auf eine Bestimmung der Genauigkeit von Lehrerurteilen beschränkt sind, sondern auch auf die

Suche nach Faktoren zielen, die mit der Urteilsgenauigkeit kovariieren und diese gegebenenfalls moderieren oder sogar bedingen (z. B. Feinberg & Shapiro, 2009). Postuliert werden solche Faktoren zum Beispiel im *Modell der Genauigkeit von Lehrerurteilen*[9], das in dem Forschungsartikel zur Metanalyse von Südkamp und Kollegen (2012) vorgestellt wird und in Abbildung 2.1 grafisch dargestellt ist.

Im Kern des Modells steht die Urteilsgenauigkeit von Lehrkräften, die als Korrespondenz von Lehrerurteilen und Schülerleistungen konzeptualisiert ist. Dem Modell nach kann die Urteilsgenauigkeit durch vier Merkmalsgruppen beeinflusst sein: durch Merkmale der Lehrkraft (z. B. Berufserfahrung), durch Charakteristika der zu treffenden Urteile (z. B. Informiertheit, Urteilsspezifität), durch Schülermerkmale (z. B. Leistungsniveau und -heterogenität in der Klasse) und durch Merkmale des von den Schülerinnen und Schülern bearbeiteten diagnostischen Verfahrens (z. B. ein Leistungstest), mit dessen Ergebnissen die Lehrerurteile verglichen werden (z. B. getesteter Kompetenzbereich, Merkmale einzelner Aufgaben).

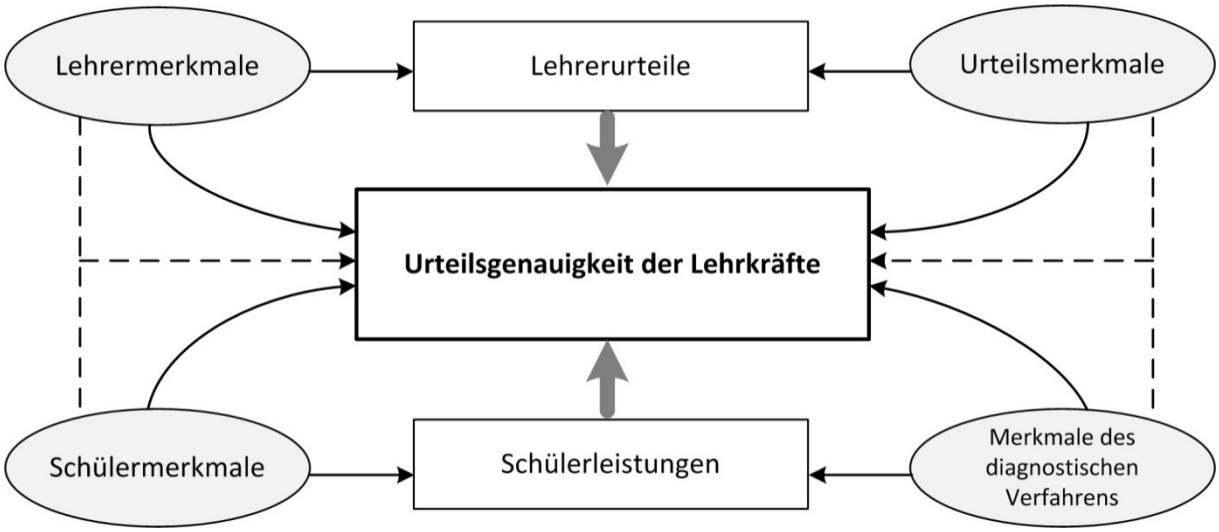

Abbildung 2.1: Modell der Genauigkeit von Lehrerurteilen (nach Südkamp et al., 2012)

Das Modell selbst ist empirisch nicht geprüft, es basiert insbesondere auf theoretischen Annahmen, Einzelbefunden empirischer Studien und sachlogischen Erwägungen. Folglich hat es vor allem heuristischen Charakter (vgl. Kaiser, Möller, Helm & Kunter, 2015). Es kann also beispielsweise dazu dienen, Forschungsfragen zu generieren oder Forschungsergebnisse zu systematisieren. Entsprechend der letztgenannten Funktion werden im Folgenden die wesentlichen Befunde aus der Forschungsliteratur zu den

[9] Im Original: Model of teacher judgement accuracy.

Kovariaten der Genauigkeit von Lehrerurteilen, gruppiert nach den vier im Modell genannten Merkmalsgruppen, dargestellt.

2.1.6.2 Empirische Befunde zu Zusammenhängen zwischen Urteilsgenauigkeit und Lehrermerkmalen

In verschiedenen Studien wurde der Frage nachgegangen, inwiefern die *Berufserfahrung* von Lehrkräften mit der Genauigkeit von Lehrerurteilen kovariiert. Vor allem in der Praxis wird häufig angenommen, dass zwischen der Anzahl der Berufsjahre und der Urteilsgenauigkeit positive korrelative Zusammenhänge bestehen. Begründet wird diese Annahme vor allem damit, dass Lehrkräfte mit einer längeren Berufspraxis in der Vergangenheit eine deutlich höhere Anzahl an Lerngelegenheiten (z. B. Unterrichtserfahrungen, Austausch im Kollegium, Fortbildungen) zum Erwerb von diagnostischen Fähigkeiten hatten als Lehrkräfte, die noch am Anfang ihres Berufslebens stehen (vgl. Coladarci, 1986; McElvany et al., 2009).

Zumeist wurden in empirischen Untersuchungen jedoch nur geringe und nicht in jedem Fall statistisch signifikante Zusammenhänge zwischen der Urteilsgenauigkeit und der Berufserfahrung festgestellt (z. B. Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003; Leinhardt, 1983; Wild & Rost, 1995). Einzig Freeman (1993) konnte einen deutlichen Effekt der Berufserfahrung finden. Insgesamt ist dieses Befundmuster konsistent zu den Ergebnissen von Studien aus der psychologischen Diagnostik, in denen die Vorhersagegüte von Urteilen untersucht wurde, die entweder statistisch bzw. mechanisch, also nach standardisierten Regeln (z. B. auf der Basis von Testergebnissen), oder aber klinisch, also nicht nach festen, expliziten Regeln (z. B. Urteile von Experten, die etwa auf Erfahrungen oder Intuition beruhen) gebildet wurden. Hierbei zeigt sich, dass Vorhersagen, die auf standardisierten Regeln basieren, im Mittel eine höhere prognostische Güte aufweisen (z. B. Grove, Zald, Lebow, Snitz & Nelson, 2000; Meehl, 1954). Bemerkenswert ist dabei, dass die Höhe des relativen Vorteils der statistischen gegenüber der klinischen Urteilsbildung weitgehend unbeeinflusst von der beruflichen Erfahrung der Diagnostikerinnen und Diagnostiker zu sein scheint, die die klinischen Beurteilungen vornehmen (z. B. Garb, 1989).

In der Forschungsliteratur zur diagnostischen Kompetenz von Lehrkräften wird mit Blick auf die geringen Effekte, die für die Berufserfahrung gefunden wurden, unter anderem argumentiert, dass weniger die Berufsjahre einer Lehrkraft als vielmehr ihre

fachdidaktischen Kompetenzen entscheidend für die Genauigkeit ihrer Urteile seien (vgl. z. B. Brunner et al., 2011; McElvany et al., 2009). In diesem Zusammenhang wird angenommen, dass fachdidaktische Kompetenzen vor allem im Studium und durch Fort- und Weiterbildungen erworben werden und berufliche Erfahrungen zwar eine notwendige, aber keine hinreichende Bedingung für deren Entwicklung darstellen. Demnach könnten Lehrkräfte, die auf viele Berufsjahre zurückblicken, mutmaßlich nicht mehr auf dem aktuellen Stand der Fachdidaktik sein und mithin Schwierigkeiten haben, bestimmte Aufgabenanforderungen zu erkennen und akkurat zu beurteilen. Bisherige Befunde stützen jedoch auch diese Annahme nur bedingt (vgl. McElvany et al., 2009). Allerdings deuten die Ergebnisse einiger prozessorientierter Studien zur diagnostischen Kompetenz von Lehrkräften darauf hin, dass zwischen erfahrenen Lehrkräften und „Lehrernovizen“ (z. B. Lehramtsstudierende) qualitative Unterschiede im Urteilsprozess bestehen. So sind erfahrene Lehrkräfte offenbar weniger anfällig für Urteilsverzerrungen und im Vergleich zu ihren unerfahreneren Kolleginnen und Kollegen besser in der Lage, je nach den Erfordernissen des Urteilskontexts zwischen verschiedenen Modi der Informationsverarbeitung zu wechseln (vgl. Abschnitt 2.2.1 sowie Dünnebier et al., 2009; Krolak-Schwerdt et al., 2009; Krolak-Schwerdt, Böhmer & Gräsel, 2012; Krolak-Schwerdt & Rummer, 2005; van Ophuysen, 2006).

In weiteren Studien wurde untersucht, ob die *Kontaktdauer*, also der Zeitraum, in dem eine Lehrkraft eine Schulklasse bereits unterrichtet, die Akkuratheit ihrer Urteile erhöht. Hintergrund dieser Forschungsfrage ist die Annahme, dass Lehrerinnen und Lehrer, die eine Klasse bereits seit einigen Jahren unterrichten, über mehr leistungsbezogene Informationen zu einzelnen Schülerinnen und Schülern verfügen dürften als bei einer eher kurzen Kontaktdauer. Diese zusätzlichen Informationen sollten es ihnen ermöglichen, Schülerleistungen und Aufgabenschwierigkeiten genauer zu beurteilen. Allerdings fielen die bislang in empirischen Studien festgestellten korrelativen Zusammenhänge zwischen Kontaktdauer und Urteilsgenauigkeit ebenfalls gering und zumeist nicht statistisch signifikant aus (Oerke, McElvany, Ohle, Ullrich & Horz, 2015; Wild & Rost, 1995).

Inwiefern sich *Grundschul-* und *Gymnasiallehrkräfte* hinsichtlich ihrer diagnostischen Kompetenz unterscheiden, wurde in einer Studie von Karing (2009) für die Rangkomponente der Urteilsgenauigkeit geprüft. In dieser Untersuchung konnten die untersuchten Grundschullehrkräfte die Leistungen ihrer Schülerinnen und Schüler in

Deutsch (Wortschatz, Textverstehen) und Mathematik (Arithmetik) etwas genauer einschätzen als Lehrerinnen und Lehrer, die am Gymnasium unterrichteten. Dieses Ergebnis wurde, empirisch gestützt durch weitere Befunde der Studie, mit Unterschieden in der Zusammensetzung der Schülerschaft erklärt. So weisen Schulklassen an Grundschulen, bedingt durch die institutionelle Trennung der Bildungswege zu Beginn der Sekundarstufe I, in der Regel eine höhere Leistungsstreuung auf als Schulklassen an Gymnasien. Aufgrund dieser höheren Leistungsheterogenität ist es vermutlich leichter möglich, Kinder bezüglich ihres Leistungsvermögens zu differenzieren (Schrader, 1989). Dementsprechend scheint es auch besser zu gelingen, die Schülerinnen und Schüler einer Klasse zum Beispiel in eine korrekte Leistungsrangreihe zu bringen. Als eine weitere Erklärung wurde das an Grundschulen vorherrschende Klassenlehrerprinzip[10] diskutiert, das durch eine längere Kontaktdauer mit der jeweiligen Schulklasse und einer engeren Lehrer-Schüler-Beziehung gekennzeichnet ist (s. o.). Ferner wurde auf Unterschiede in der Ausbildung von Grundschul- und Gymnasiallehrkräften hingewiesen: Während sich das Studium für Lehramt an Gymnasien verstärkt auf fachwissenschaftliche Inhalte konzentriert, weist das Studium für Lehramt an Grundschulen einen höheren Anteil an pädagogisch-psychologischen und fachdidaktischen Veranstaltungen auf, die unter anderem auf den Erwerb wichtiger Grundlagen zur Bewältigung der diagnostischen Anforderungen des Lehrerberufs abzielen (Karing, 2009, vgl. auch Brunner, 2011).

Schließlich wurde auch vermutet, dass *normative Leistungserwartungen* von Lehrkräften einen verzerrenden Einfluss auf die Genauigkeit von Lehrerurteilen haben können (vgl. Feinberg & Shapiro, 2009; Schrader, 1989). Zum Beispiel könnte eine Lehrkraft deswegen davon überzeugt sein, dass eine bestimmte Aufgabe von einem großen Teil der Schülerinnen und Schüler ihrer Klasse korrekt gelöst werden wird, weil der hierfür benötigte Lernstoff bereits im vorangegangenen Schuljahr Gegenstand des Lehrplans war und dementsprechend im Unterricht behandelt wurde. Dieser bislang empirisch nicht untersuchten Hypothese wird (u. a.) im Rahmen von *Teilstudie 1* der vorliegenden Dissertationsschrift (Kapitel 4) nachgegangen.

[10] Beim Klassenlehrerprinzip der Grundschule werden die Schülerinnen und Schüler einer Schulklasse zumindest in den meisten Fächern von derselben Lehrkraft unterrichtet. Beim Fachlehrerprinzip, das an den meisten weiterführenden Schulen anzutreffen ist, erfolgt der Unterricht in den einzelnen Fächern durch jeweils andere, (idealiter) hierfür speziell ausgebildete Lehrkräfte (vgl. Hammel, 2011).

2.1.6.3 Empirische Befunde zu Zusammenhängen zwischen Urteilsgenauigkeit und Merkmalen der zu treffenden Urteile

Bereits weiter oben wurde skizziert, dass die verschiedenen Forschungsarbeiten zur Genauigkeit von Lehrerurteilen unter anderem hinsichtlich der Art der erhobenen Urteile differieren (vgl. Abschnitt 2.1.4). Ein wesentliches Unterscheidungsmerkmal stellt hierbei der Grad dar, zu dem die untersuchten Lehrkräfte über den Vergleichsmaßstab informiert sind, mit dem ihre Urteile in Beziehung gesetzt werden (Informiertheit). Die von Lehrkräften geforderten Leistungseinschätzungen können zudem hinsichtlich ihrer Spezifität differieren (Urteilsspezifität) und entweder sehr allgemein sein (Leistung in einem bestimmten Fach) oder auf eine ganz bestimmte (Teil-)Kompetenz zielen (Domänenspezifität).

Ein positiver korrelativer Zusammenhang zwischen dem Grad an *Informiertheit* von Lehrerurteilen einerseits und der Urteilsgenauigkeit andererseits konnte vielfach, auch metanalytisch, belegt werden (z. B. Feinberg & Shapiro, 2003; Feinberg & Shapiro, 2009; Hoge & Coladarci, 1989; Südkamp et al., 2012). Leistungsurteile von Lehrkräften fallen also im Durchschnitt akkurater aus, wenn Lehrkräfte zum Beispiel die Items des Tests kennen, mit dessen Ergebnis ihre Einschätzungen verglichen werden.

Für den Aspekt der *Urteilsspezifität* wurde in der Metaanalyse von Südkamp und Kollegen (2012) kein statistisch signifikanter Zusammenhang mit der Urteilsgenauigkeit gefunden. Ein differenzierteres Befundmuster zeigen die Ergebnisse der Metanalyse von Hoge und Coladarci (1989). Hier wurde festgestellt, dass in den Primärstudien, in denen mit Ratingskalen die Art von Urteilen untersucht wurde, die als am wenigsten spezifisch gilt (vgl. auch Südkamp et al., 2012)[11], deutlich geringere Korrelationen zwischen Lehrereinschätzungen und Schülerleistungen ermittelt wurden als in Untersuchungen zu Lehrerurteilen mit höherer Urteilsspezifität. Demgegenüber differierten die mittleren Genauigkeitskennwerte aus Primärstudien, in denen Lehrerurteile der anderen vier Spezifitätsgrade erfasst wurden, nur unwesentlich.

[11] In der Forschungsliteratur werden folgende Spezifitätsgrade unterschieden, aufsteigend von geringer zu hoher Spezifität: (1) Einschätzen von Leistungen auf einer Ratingskala, (2) Bilden einer Leistungsrangreihe, (3) Beurteilen von Leistungen auf einer Skala mit Zensuren äquivalenter Metrik, (4) Angabe der Anzahl korrekt gelöster Testaufgaben, (5) Angaben zum Abschneiden einzelner Schülerinnen und Schüler bei einzelnen Testaufgaben.

Der Aspekt der *Domänenspezifität* allein scheint keinen moderierenden Effekt auf die Urteilsgenauigkeit zu haben (Begeny et al., 2011; Südkamp et al., 2012). Von Bedeutung ist jedoch, ob das Lehrerurteil und das Kriterium, zu dem es in Beziehung gesetzt wird (z. B. das Konstrukt, das in dem Test erfasst wird, für den die Schülerleistungen zu prognostizieren sind) hinsichtlich der Domänenspezifität kongruent zueinander sind. Wenn dies nicht der Fall ist, also etwa sehr allgemeine Beurteilungen mit den Leistungen in einem Test zu einer sehr spezifischen Fähigkeit in Beziehung gesetzt werden, fallen die gefundenen Zusammenhänge deutlich geringer aus (Südkamp et al., 2012).

2.1.6.4 Empirische Befunde zu Zusammenhängen zwischen Urteilsgenauigkeit und Schülermerkmalen

Für einige Schülermerkmale ist bereits in den vorangegangenen Abschnitten skizziert worden, dass sie einen moderierenden Effekt auf die Akkuratheit von Lehrerurteilen haben können. Unter anderem wurde darauf hingewiesen, dass die Urteilsgenauigkeit je nach Kompetenzniveau der eingeschätzten Schülerinnen und Schüler variieren kann. So legen die Ergebnisse einer größeren Anzahl von Studien den Schluss nahe, dass Lehrerurteile zu *Schülerinnen und Schülern aus dem unteren Leistungsspektrum* oftmals wenig akkurat sind (u. a. Begeny et al., 2008; Begeny et al., 2011; Gresham et al., 1997; Madelaine & Wheldall, 2005). In einer Zusatzuntersuchung zur PISA-Studie des Jahres 2000 war zum Beispiel eine Stichprobe aus Lehrkräften an Hauptschulen nicht in der Lage, einen Großteil der Schülerinnen und Schüler, die aufgrund ihrer Testergebnisse als schwache Leserinnen bzw. Leser eingestuft wurden, korrekt zu identifizieren (Artelt et al., 2001). Eine geringere Urteilsgenauigkeit bei der Identifikation leistungsschwacher Schülerinnen und Schüler ist vor allem deswegen problematisch, weil gerade diese Schülergruppe einer besonderen Unterstützung bedarf, um erfolgreich lernen zu können. Wenn eine Lehrkraft nicht in der Lage ist, diesen Bedarf akkurat festzustellen, wird sie auch keine Maßnahmen zur spezifischen Unterstützung der betreffenden Schülerinnen und Schüler ergreifen.

In der Forschungsliteratur finden sich zudem einige wenige Befunde, die vermuten lassen, dass Lehrkräfte ebenfalls Schwierigkeiten haben, besonders *leistungsstarke Schülerinnen und Schüler* akkurat zu beurteilen (z. B. Eckert et al., 2006). Eine geringe Urteilsgenauigkeit im oberen Leistungsspektrum birgt zum Beispiel das Risiko, dass

Schülerinnen und Schülern mit besonderen Begabungen übersehen werden und nicht die Förderung erhalten, die sie benötigen, um ihre Begabungen entfalten zu können.

Bereits im Zusammenhang mit den von Karing (2009) festgestellten Diskrepanzen in der Urteilsgenauigkeit von Grundschul- und Gymnasiallehrkräften wurde der Erklärungsansatz skizziert, dass diese Unterschiede unter anderem dadurch bedingt sein könnten, dass die *Leistungsheterogenität* in Grundschulklassen deutlich höher als an Gymnasien ist. Tatsächlich konnten auch andere Autorinnen und Autoren empirisch zeigen, dass es Lehrkräften leichter fällt, eine Rangordnung ihrer Schülerinnen und Schüler entsprechend ihrer Leistungen vorzunehmen, wenn in der betreffenden Schulklasse deutliche Leistungskontraste bestehen (z. B. Rjosk et al., 2011; Schrader, 1989; Weinert, F. E. et al., 1990). Demgegenüber scheint die Größe einer Klasse nicht (z. B. Wild & Rost, 1995) oder nur in geringem Maße (z. B. Anders et al., 2010) für die Urteilsgenauigkeit von Bedeutung zu sein.

Bereits in den Darstellungen zur diagnostischen Güte der Notengebung (siehe Abschnitt 2.1.3) wurde erläutert, dass Schulnoten nicht selten durch *Referenzgruppeneffekte* verzerrt sind (z. B. Davis, 1966; Ingenkamp, 1995; Ingenkamp & Lissmann, 2008; Lintorf, 2012; Rheinberg, 2001). Ähnliche Ergebnisse konnten auch für Übergangsempfehlungen ermittelt werden (z. B. Trautwein & Baeriswyl, 2007). Südkamp und Möller (2009) untersuchten, inwiefern auch andere Formen von Lehrerurteilen vom mittleren Leistungsniveau einer Schulklasse beeinflusst sind. Ein Referenzgruppeneffekt wurde hierbei nur für Lehrerurteile mit einem eher geringen Grad an Informiertheit und Spezifität gefunden. Wurden die Untersuchungsteilnehmerinnen und -teilnehmer um Angaben zur Anzahl der von einzelnen Schülerinnen und Schülern korrekt gelösten Aufgaben gebeten (hoher Grad an Informiertheit, hohe Urteilsspezifität), waren ihre Urteile hingegen nicht in Richtung der mittleren Leistungsstärke der Schulklasse verzerrt.

Ebenfalls bereits im Zusammenhang mit der diagnostischen Güte von Schulnoten wurden die Ergebnisse von Studien skizziert, in denen ein *verzerrender Effekt leistungsferner Faktoren* wie Geschlecht (z. B. Schreiner et al., 2008) oder sozioökonomischer Status der Eltern (z. B. Cicmanec et al., 2001) auf die Notenvergabe festgestellt wurde. Darüber hinaus findet sich in der Forschungsliteratur auch eine größere Anzahl an Studien, in denen der Frage nachgegangen wurde, inwiefern leistungsferne Schülermerkmale die Genauigkeit von Lehrerurteilen beeinflussen, bei denen keine Noten vergeben werden. Als potenzielle schülerseitige Kovariaten der Urteilsgenauigkeit

wurden hierbei unter anderem das Geschlecht (z. B. Karing et al., 2011; Madelaine & Wheldall, 2005; Madon et al., 1998; Shinn, Tindal & Spira, 1987; Tiedemann, 2002), der soziökonomische Hintergrund (z. B. Karing et al., 2011; Madon et al., 1998; Ready & Wright, 2011), der ethnische Hintergrund (z. B. Darling-Hammond, 1995; Madon et al., 1998; Ready & Wright, 2011; Shinn et al., 1987), Mehrsprachigkeit (z. B. Rjosk et al., 2011), Intelligenz (z. B. Hoge & Butcher, 1984; Schrader & Helmke, 1990) und Verhaltensprobleme (z. B. Bennett, Gottesman, Rock & Cerullo, 1993; Hecht & Greenfield, 2002) untersucht. Insgesamt ist die Befundlage zum Einfluss der genannten Hintergrundmerkmale sehr gemischt. Wenn die Leistungsurteile von Lehrkräften mit bestimmten Schülermerkmalen korrelieren, also für Schülerinnen und Schüler mit einer bestimmten Merkmalsausprägungen deutlich positivere Testleistungen prognostiziert werden als für andere, dann korrespondieren diese Disparitäten oftmals auch mit den tatsächlich erreichten Testleistungen (z. B. Ready & Wright, 2011).

Allerdings zeigen die Ergebnisse einiger Forschungsarbeiten, dass die Genauigkeit von Lehrerurteilen offenbar durch sogenannte Halo-Effekte gemindert sein kann. Diese bezeichnen eine kognitive Urteilsverzerrungen, bei der Personen dazu neigen, von einem ihnen bekannten Merkmal auf die Ausprägung anderer Merkmale zu schießen, die mit diesem nur wenig korreliert sind (Thorndike, 1920). So beeinflussen etwa Informationen zu den Schülerleistungen in einem Fach, wie Lehrkräfte die Leistungen der betreffenden Schülerinnen und Schüler in einem anderen Fach einschätzen (Dompnier, Pansu & Bressoux, 2006). Außerdem konnte empirisch gezeigt werden, dass auch Informationen zur Wiederholung einer Klassenstufe in der Vergangenheit oder die wahrgenommene Lernmotivation von Schülerinnen und Schülern einen verzerrenden Einfluss auf die von Lehrkräften vorgenommenen Urteile im Sinne eines Halo-Effekts haben können (vgl. Dompnier et al., 2006; Kaiser, Retelsdorf, Südkamp & Möller, 2013; Urhahne et al., 2010).

2.1.6.5 Empirische Befunde zu Zusammenhängen zwischen Urteilsgenauigkeit und Merkmalen des standardisierten Verfahrens

In einigen Primärstudien wurde untersucht, ob die Genauigkeit von Lehrerurteilen je nach *Schulfach* variiert, also ob Lehrkräfte beispielsweise die Ausprägung sprachlicher Kompetenzen akkurater einschätzen können als Kompetenzen in Mathematik oder in den naturwissenschaftlichen Fächern. Auch die in diesen Studien gefundenen Ergebnisse sind

insgesamt uneinheitlich (z. B. Coladarci, 1986; Demaray & Elliot, 1998; Eckert et al., 2006). Dementsprechend wurde auch in der Metaanalyse von Südkamp und Kollegen (2012) kein signifikanter Effekt des Schulfachs auf die Urteilsgenauigkeit gefunden.

Hervorzuheben ist allerdings die Studie von Lorenz und Artelt (2009), in der die Urteile von Klassenlehrerinnen und Klassenlehrern an Grundschulen zu den Kompetenzen ihrer Schülerinnen und Schüler im sprachlichen und mathematischen Bereich untersucht wurden (sprachlich: Wortschatz, Textverstehen; mathematisch: Arithmetik). Die mittlere Urteilsgenauigkeit differierte dabei nur in geringem Maße zwischen den eingeschätzten Kompetenzbereichen. Allerdings wurden zwischen den Genauigkeitskennwerten, die für die Lehrerurteile zu den beiden sprachlichen Kompetenzen ermittelt wurden, deutlich höhere korrelative Zusammenhänge gefunden als zwischen den Kennwerten für die beiden sprachlichen Kompetenzen einerseits und den Kennwerten für die arithmetische Kompetenz andererseits. Untersuchungsteilnehmerinnen und -teilnehmer, deren Urteile zu Schülerkompetenzen im Textverstehen einen hohen Grad an Genauigkeit aufwiesen, waren also oftmals ebenfalls gute Diagnostikerinnen bzw. Diagnostiker im Bereich Wortschatz, allerdings nicht unbedingt für den Bereich Arithmetik. Aus diesen Ergebnissen wurde geschlossen, dass es zwar kein Schulfach gibt, in dem die Genauigkeit von Lehrerurteilen im Allgemeinen besonders gering oder besonders hoch ausfällt, jedoch die diagnostischen Fähigkeiten individueller Lehrerinnen und Lehrer je nach Schulfach sehr unterschiedlich ausgeprägt sein können (vgl. auch Lorenz, 2011).

In mehreren Publikationen zur diagnostischen Kompetenz von Lehrkräften wurde die Vermutung formuliert, dass eine größere Zahl von Lehrerinnen und Lehrern möglicherweise deswegen Schwierigkeiten hat, die Leistung ihrer Schülerinnen und Schüler in einem standardisierten Leistungstest akkurat zu beurteilen, weil die in den betreffenden Untersuchungen eingesetzten Tests nur bedingt repräsentativ für den im Unterricht behandelten Lernstoff sind. Vor diesem Hintergrund wird von einigen Autorinnen und Autoren angenommen, dass es Lehrkräften im Vergleich dazu bedeutend leichter fallen dürfte, Schülerleistungen in einem *curriculum based measurement* (CBM) einzuschätzen (z. B. Feinberg & Shapiro, 2009; Madelaine & Wheldall, 2005). Unter diesem Begriff wird eine ganze Reihe von zumeist standardisierten Testverfahren zusammengefasst, die dazu dienen, Lehrkräften Rückmeldungen zum Lernfortschritt ihrer Schülerinnen und Schüler zu geben. Von standardisierten Leistungstests zu

schulrelevanten Kompetenzen (wie etwa dem Rechtschreibtest Hamburger Schreibprobe (u. a. May, 2002)) heben sich CBMs dadurch ab, dass sie eine hohe Kontentvalidität in Bezug auf die Inhalte von Lehrplan und Unterricht haben. Zudem zeichnen sich CBMs in der Regel dadurch aus, dass sie speziell dafür entwickelt wurden, von Lehrkräften im Unterricht eingesetzt zu werden (d. h. geringer zeitlicher Aufwand, einfache Durchführung und Auswertung usw.) (vgl. Deno, 1985). Die Ergebnisse der Metaanalyse von Südkamp und Kollegen (2012) können allerdings die Hypothese, dass Lehrereinschätzungen zu Schülerleistungen in CBMs genauer ausfallen als Urteile zu den in standardisierten Leistungstests erreichten Ergebnissen, nicht empirisch unterstützen.

Zusammenhänge zwischen der Genauigkeit von Lehrerurteilen zur Schwierigkeit einzelner Aufgaben einerseits und bestimmten *Aufgabenmerkmalen* andererseits sind bislang verhältnismäßig selten untersucht worden. Lintorf und Kollegen (2011) fanden in einer exploratorischen Faktorenanalyse der mittleren Urteilsfehler, die aus den Schwierigkeitseinschätzungen von Lehrkräften für mehrere Testitems bestimmt wurden, dass sich psychometrisch leichte und psychometrisch schwere Testitems zu jeweils einem Faktor gruppierten. Dieser Befund lässt vermuten, dass die Genauigkeit von Schwierigkeitseinschätzungen mit der psychometrischen Schwierigkeit der beurteilten Aufgaben variiert. In Übereinstimmung mit dieser Interpretation ermittelten McElvany und Kollegen (2009), dass die in ihrer Studie untersuchten Lehrkräfte die Schwierigkeit eines Sets aus schwierigkeitshomogenen, psychometrisch insgesamt leichten Aufgaben akkurater beurteilen konnten als ein schwierigkeitsheterogenes, psychometrisch insgesamt mittelschweres Aufgabenset.

2.1.7 Zusammenfassung

Die große Mehrzahl der Studien zur diagnostischen Kompetenz von Lehrkräften ist den genauigkeitsorientierten Forschungsansätzen zuzuordnen, bei denen auf die Akkuratheit von Lehrerurteilen fokussiert wird. Die Fähigkeit von Lehrkräften, die Leistungen ihrer Schülerinnen und Schüler sowie die Schwierigkeit der im Unterricht und bei Lernerfolgskontrollen verwendeten Aufgaben möglichst genau beurteilen zu können, ist von hoher schulpraktischer Bedeutung und bildet beispielsweise eine wichtige Grundlage des adaptiven Unterrichtens und des Gestaltens von Lernerfolgskontrollen.

In empirischen Forschungsarbeiten wird die Urteilsgenauigkeit zumeist als Höhe der Korrespondenz zwischen den Urteilen von Lehrkräften einerseits und den Leistungen von

Schülerinnen und Schülern in einem bestimmten Kriterium (wie etwa der Leistung in einem standardisierten Test) andererseits operationalisiert. Schulnoten sind weniger gut geeignet, um die Genauigkeit von Lehrerurteilen zu untersuchen, da bei der Notenvergabe neben der Leistung oftmals auch andere Aspekte berücksichtigt werden (v. a. pädagogische Motive). Bei der Bestimmung der Korrespondenz von Lehrerurteilen und Schülerleistungen können drei unterschiedliche Facetten betrachtet werden, auf denen die vor allem in der deutschsprachigen Forschungsliteratur eingeführte Unterscheidung von Rang-, Niveau- und Differenzierungskomponente basiert.

Insgesamt zeigen insbesondere die empirischen Ergebnisse zur Niveaukomponente, dass die Urteile von Lehrkräften im Durchschnitt durch eine geringe Genauigkeit gekennzeichnet sind. Dabei scheinen Lehrkräfte im Mittel dazu zu neigen, ihre Schülerinnen und Schüler zu überschätzen sowie die Schwierigkeit von Aufgaben zu unterschätzen. Der Forschungsstand verdeutlicht darüber hinaus, dass die Akkuratheit von Lehrerurteilen interindividuell stark variiert, wobei sich neben einigen Lehrerinnen und Lehrern, deren Urteile durch eine hohe Akkuratheit gekennzeichnet sind, auch eine größere Zahl von Lehrkräften findet, deren Einschätzungen vielfach ungenau ausfallen.

An verschiedener Stelle wurde untersucht, welche Faktoren einen moderierenden Effekt auf die Urteilsgenauigkeit haben können. Im vorliegenden Teilkapitel wurden die wesentlichen Befunde zu dieser Forschungsfrage anhand der im Modell der Genauigkeit von Lehrerurteilen (Südkamp et al., 2012) genannten Merkmalsgruppen dargestellt. Insgesamt fokussieren die bislang zu den Bedingungen und Kovariaten der Urteilsgenauigkeit durchgeführten Studien nahezu ausschließlich die Akkuratheit von Leistungseinschätzungen; Lehrerurteile zu Aufgabenschwierigkeiten wurden hingegen nur selten untersucht. Als stark verknapptes Fazit lässt sich festhalten, dass die Urteilsgenauigkeit insbesondere mit dem Grad variiert, mit dem Lehrkräfte über den Vergleichsmaßstab informiert sind, zu dem ihre Einschätzungen in Beziehung gesetzt werden. Auch bestimmte Schülermerkmale haben einen moderierenden Effekt auf die Akkuratheit von Lehrerurteilen. Merkmale von Lehrkräften wie die Berufserfahrung oder die Kontakdauer mit der zu beurteilenden Klasse scheinen hingegen von geringerer Bedeutung zu sein. Jedoch ist zum Beispiel offen, ob und mit welchem Effekt für die Urteilsgenauigkeit Lehrkräfte bei ihren Einschätzungen zu Aufgabenschwierigkeiten auch Informationen dazu berücksichtigen, wann (d. h. in welcher Klassenstufe) der für die Lösung der betreffenden Aufgabe erforderliche Lernstoff Gegenstand von Lehrplan und

Unterricht war oder wie häufig ähnliche Aufgaben von ihren Schülerinnen und Schülern geübt wurden. Hieran wird Teilstudie 1 anschließen, die in Kapitel 4 der vorliegenden Arbeit dargestellt ist.

2.2 Der Einsatz standardisierter Testinstrumente in der Schule als wichtiger Bestandteil des diagnostischen Aufgabenfeldes von Lehrerinnen und Lehrern

2.2.1 Kritik der genauigkeitsorientierten Definition diagnostischer Kompetenz

Zu Beginn des vorherigen Kapitels wurde dargestellt, dass die diagnostische Kompetenz von Lehrkräften insbesondere im Rahmen der Forschungsarbeiten, die im Sinne Schraders (2009, 2013) den genauigkeitsorientierten Ansätzen zuzuordnen sind, als Fähigkeit definiert wird, die Leistung von Schülerinnen und Schülern oder auch die Schwierigkeit von Aufgaben akkurat einschätzen zu können. Dieses Begriffsverständnis von diagnostischer Kompetenz ist an verschiedener Stelle problematisiert worden (z. B. Helmke, 2010; von Aufschnaiter et al., 2015). Hauptsächlich bezieht sich die hierzu in der Forschungsliteratur (mehr oder weniger explizit) geäußerte Kritik darauf, dass es sich bei der Gleichsetzung von diagnostischer Kompetenz und Urteilsgenauigkeit de facto um eine *operationale Definition* handelt, die lediglich angibt, wie sich die diagnostische Kompetenz von Lehrkräften empirisch messen lässt (vgl. Döring & Bortz, 2016), die jedoch nicht beschreibt, welche Bestandteile diese Kompetenz umfasst. Es wird also nicht spezifiziert, welche Kenntnisse und Fähigkeiten die diagnostische Kompetenz einer Lehrkraft konstituieren und eine akkurate Beurteilung von Schülerleistungen oder Aufgabenschwierigkeiten ermöglichen.

Mit Blick auf die Elemente der diagnostischen Kompetenz von Lehrkräften heben etwa Brunner und Kollegen (2011) die Bedeutung fachdidaktischen und pädagogisch-psychologischen Wissens für die Genauigkeit von Lehrerurteilen hervor, nutzen aber, in bewusster Abgrenzung zum tradierten Begriffsverständnis von diagnostischer Kompetenz, bewusst den Terminus der *diagnostischen Fähigkeiten*. Aus ganz ähnlichen Gründen verwenden Weinert und Kollegen (1990) für ihr – methodische Kenntnisse und Fähigkeiten sowie konzeptuelles Wissen umfassendes – Konzept den Begriff der *diagnostischen Expertise*. Hingegen modelliert van Ophuysen (2010) explizit die *diagnostische Kompetenz* von Lehrkräften als Verknüpfung von methodischen Fähigkeiten und methodischem Wissen einerseits und didaktischem, pädagogisch-psychologischem Wissen andererseits (vgl. auch Praetorius, Lipowsky & Karst, 2012).

Des Weiteren legen die Themen und Ergebnisse von Forschungsarbeiten, die als *prozessorientiert* einzustufen sind (vgl. Schrader, 2009; Schrader, 2013), die Vermutung nahe, dass eine operationale Definition diagnostischer Kompetenz, die ausschließlich die Urteilsgenauigkeit fokussiert, unzureichend ist. In diesen Arbeiten wird der diagnostische Prozess als Ganzes betrachtet (vgl. hierzu auch das Prozessmodell diagnostischer Kompetenz bei Klug, Bruder, Kelava, Spiel & Schmitz, 2013). Neben dem eigentlichen Lehrerurteil werden Vorgänge untersucht, die diesem zeitlich vorangehen (z. B. Nutzung von bestimmten kognitiven Verarbeitungsstrategien, Nutzung und Gewichtung der verfügbaren diagnostischen Informationen) (u. a. Cooksey, Freebody & Davidson, 1986; Dünnebier et al., 2009; Kishor, 1994; Krolak-Schwerdt et al., 2009, 2012; Krolak-Schwerdt & Rummer, 2005; van Ophuysen, 2006). Vereinzelt finden sich auch Studien, in denen Zusammenhänge zwischen Urteilsgenauigkeit und Handlungen untersucht werden, die dem Lehrerurteil zeitlich nachgestellt sind bzw. auf dessen Grundlage erfolgen. Hierzu zählen etwa das Geben von Feedback (z. B. Behrmann & Souvignier, 2013), die Nutzung von Methoden zur Binnendifferenzierung im Unterricht (z. B. Westphal et al., 2016) oder die Beratung von Schülerinnen und Schülern und ihren Eltern (z. B. Klug, Bruder, Keller & Schmitz, 2012).

Die Ergebnisse aus prozessorientierten Forschungsarbeiten verdeutlichen, dass sich Lehrkräfte auch hinsichtlich der in diesen Studien untersuchten Aspekte zum Teil deutlich unterscheiden. So konnte etwa bei empirischen Vergleichen zwischen Lehrernovizen (Lehramtsstudierende) und Lehrerexperten (erfahrene Lehrkräfte) festgestellt werden, dass letztere eine geringere Anfälligkeit für bestimmte Urteilsverzerrungen (z. B. Bestätigungstendenz[12] (van Ophuysen, 2006), Ankereffekt[13] (Dünnebier et al., 2009)) aufweisen und vergleichsweise besser in der Lage sind, die kognitive Strategie, mit der sie diagnostische Informationen verarbeiten, an die kontextuellen Bedingungen der von ihnen geforderten Urteile anzupassen (Krolak-Schwerdt et al., 2009, 2012; Krolak-Schwerdt & Rummer, 2005). Diese Befundmuster lassen sich auch als ein Fingerzeig darauf deuten, dass sich eine operationale Definition bzw. eine Erfassung diagnostischer

[12] Diese Urteilsverzerrung bezeichnet die Tendenz von Personen, ein einmal gefasstes Urteil trotz gegensätzlicher oder widersprüchlicher Informationen beizubehalten (z. B. Nickerson, 1998).

[13] Ankereffekte sind das Ergebnis der Anwendung der Ankerheuristik, bei der numerische Urteile an einem vorher gesetzten Wert orientiert werden bzw. in dessen Richtung verzerrt sind (vgl. Tversky & Kahneman, 1974).

Kompetenz von Lehrkräften nicht nur auf die Genauigkeit von Urteilen beschränken, sondern beispielsweise auch auf die Adäquanz der zugrundeliegenden Informationsverarbeitungsprozesse beziehen sollte (z. B. Schrader, 2013).

Gegen eine Definition der diagnostischen Kompetenz von Lehrkräften als Fähigkeit, Schülerleistungen und Aufgabenschwierigkeiten akkurat einschätzen zu können, ist ferner einzuwenden, dass das heutige diagnostische Aufgabenfeld von Lehrerinnen und Lehrern nicht mehr nur darauf beschränkt ist, summative Urteile vorzunehmen, sondern insbesondere in den letzten Jahren um zusätzliche Tätigkeiten erweitert wurde. Wichtiger Motor dieser Erweiterung waren nicht zuletzt die Veränderungen in der Steuerungslogik im Bildungswesen, die in Deutschland nach der Jahrtausendwende durch Bildungsreformen vorgenommen wurden und eine stärkere Akzentuierung der Lernergebnisse von Schülerinnen und Schülern (Outputsteuerung) zum Gegenstand hatten (z. B. Halbheer & Reusser, 2008; Oelkers & Reusser, 2008). Ein zentrales Instrument des reformierten Steuerungsmodells stellen die bereits in der Einleitung genannten Vergleichsarbeiten (VERA)[14] dar (z. B. Bach, Wurster, Thillmann, Pant & Thiel, 2014; KMK, 2012). Dies sind Kompetenztests für Schülerinnen und Schüler der dritten und achten Jahrgangsstufe, die auf der Grundlage der Bildungsstandards der KMK für den Primarbereich entwickelt werden und untersuchen sollen, welche Kompetenzen Schülerinnen und Schüler zu diesen Zeitpunkten jeweils erreicht haben. Die VERA-Tests werden flächendeckend durchgeführt, das heißt in allen allgemein bildenden Schulen und Klassen in Deutschland (IQB, 2014). Sie sind als Diagnoseinstrument im Sinne eines *low-stakes* Verfahrens konzipiert: Ihre Ergebnisse sind also weder für die beteiligten Individuen noch für die involvierten Schulen unmittelbar mit hoch bedeutsamen Konsequenzen wie etwa Versetzungs- oder Schullaufbahnentscheidungen für Schülerinnen und Schüler, Entscheidungen zu Kündigung bzw. Weiterbeschäftigung von Lehrkräften oder Beschlüssen zur Finanzierung von Schulen verbunden. Vielmehr sollen die VERA-Tests dazu dienen, die Schul- und Unterrichtsentwicklung zu unterstützen (IQB, 2014; KMK, 2012).

[14] In einigen Bundesländern sind die VERA-Tests abweichend benannt: In Hessen und Nordrhein-Westfalen tragen sie den Namen „Lernstandserhebungen“ (LSA, 2015; QUA-LIS NRW, 2016), in Hamburg heißen sie „KERMIT – Kompetenzen ermitteln“ (IfBQ, 2016), in Sachsen und Thüringen werden sie als „Kompetenztests“ (Sächsisches Bildungsinstitut, Nachtigal, 2016; 2012) bezeichnet (vgl. auch IQB, 2014).

Die Lehrerinnen und Lehrer, deren Klassen an den VERA-Testungen teilnehmen, müssen verschiedene Tätigkeiten übernehmen. Zunächst sind sie für die Durchführung der Tests sowie für die Bewertung der Schülerantworten anhand der ihnen zur Verfügung gestellten Bewertungsvorgaben zuständig.[15] In der Regel übernehmen sie außerdem die Eingabe der Testdaten in eine Online-Eingabemaske. Die von ihnen eingetragenen Daten werden in einem nächsten Schritt von der hierfür zuständigen wissenschaftlichen Einrichtung statistisch analysiert. Anschließend werden die Ergebnisse dieser Auswertungen aufbereitet (z. B. in Form eines Berichtes, mit Tabellen und Grafiken) und wieder an die betreffenden Lehrkräfte zurückgemeldet (z. B. Sächsisches Bildungsinstitut, 2012). Diese stehen dann vor der Herausforderung, die berichteten statistischen Kennwerte für die Evaluation und Weiterentwicklung ihres Unterrichts zu nutzen. Grundvoraussetzung für die Bewältigung dieser Aufgabe ist, dass die Lehrkräfte in der Lage sind, die für ihre Klasse(n) ermittelten Kennwerte korrekt zu interpretieren und angemessene Schlussfolgerungen zur Einschätzung ihres Unterrichts und für die Verbesserung der Unterrichtsqualität zu ziehen. In der Fachliteratur werden die hierfür erforderlichen Kompetenzen insbesondere im Konzept der *Assessment Literacy* spezifiziert (Richter, ohne Datum).

2.2.2 Das Konzept der Assessment Literacy

Assessment literacy wird definiert als „an individual‘s understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions" (Popham, 2011, S. 267). Formal in die Forschungsliteratur eingeführt wurde der Begriff von Stiggins (1991), Reflexionen über die von Lehrkräften benötigten Kenntnisse und Fähigkeiten im Bereich der Leistungsmessung[16] finden sich allerdings bereits in früheren Publikationen; eine Übersicht hierzu geben zum Beispiel Gulliksen (1986) oder Schafer und Lissitz (1987). Sowohl in den damaligen Arbeiten als auch in heutigen Publikationen wird darauf hingewiesen, dass viele Lehrkräfte in schulische Entscheidungsprozesse involviert sind, die auf Daten aus schulischen Leistungsmessungen basieren, ohne jedoch über das hierfür erforderliche Hintergrundwissen im Bereich der Leistungsmessung zu verfügen oder entsprechende Fortbildungsmaßnahmen besucht zu haben (vgl. auch DeLuca, LaPointe-McEwan & Luhanga, 2016; Xu & Brown, 2016).

[15] Eine Ausnahme bildet das Land Hamburg, in dem nicht Lehrerinnen und Lehrer, sondern geschulte Hilfskräfte die Korrektur bzw. Kodierung der Schülerantworten übernehmen.

[16] Assessment wird hier übersetzt als Leistungsmessung.

Seit dem Beginn der 1990er Jahre sind in verschiedenen Ländern (z. B. USA, Vereinigtes Königreich, Australien), zumeist unter der Herausgeberschaft von Behörden und Verbänden, mehrere Konzeptionen mit Beschreibungen der Elemente von Assessment Literacy veröffentlicht worden. Diese Beschreibungen erfolgten jeweils in Form von Standards; es werden also normative Erwartungen an die Kompetenzen von Lehrerinnen und Lehrern formuliert. Ein systematischer Überblick über diese Standards und darüber, welche Entwicklungsprozesse seit den 1990er Jahren in Bezug auf die den Standards zugrundeliegenden konzeptionellen Vorstellungen zu Assessment Literacy zu identifizieren sind, findet sich bei DeLuca und Kollegen (2016). Im Ergebnis einer Analyse von 15 verschiedenen Dokumenten mit Standards zu den für Lehrkräfte im Bereich der Leistungsmessung als notwendig erachteten Kenntnissen und Fähigkeiten, arbeitete dieses Autorenteam insgesamt acht Schwerpunktthemen heraus:

(1) Standards zu *assessment purposes* zielen darauf, dass Lehrkräfte in der Lage sein sollten, in Abhängigkeit von zuvor klar definierten Zielstellungen, die jeweils angemessene Art der Leistungsmessung (z. B. formative vs. summative Assessments, standardisierte Leistungstests vs. informelle, d. h. von der Lehrkraft selbst entwickelte, idealiter lernzielorientierte Tests) auszuwählen.

(2) Standards zum *assessment process* betreffen die Konstruktion, Administration und Auswertung von Tests sowie die Interpretation von Testergebnissen in Hinblick auf schulische Entscheidungsprozesse. Dieser Themenbereich umfasst also auch einige der Teilkompetenzen, die Lehrkräfte für die erfolgreiche Bewältigung der mit den VERA-Testungen verbundenen Anforderungen benötigen (siehe oben).

(3) Standards zu *communication of assessment results* fokussieren auf Fähigkeiten bezüglich der Kommunikation der Zielstellungen von Leistungsmessungen, der realisierten Testprozeduren und der ermittelten Testergebnisse an verschiedene Interessengruppen („stakeholder“).

(4) Standards zu *assessment fairness* akzentuieren die Beachtung und Etablierung fairer Testbedingungen für alle Teilnehmerinnen und Teilnehmer und heben die besonderen Herausforderungen hervor, die sich für die Leistungsmessung in stark heterogenen Lerngruppen oder bei sehr lernschwachen Schülerinnen und Schülern ergeben.

(5) Standards zu *assessment ethics* betreffen vor allem den Datenschutz bzw. den sensiblen und verantwortungsvollen Umgang mit den erhobenen Leistungsdaten von Schülerinnen und Schülern.

(6) Standards zu *measurement theory* adressieren das Verständnis der psychometrischen Eigenschaften von Verfahren der Leistungsmessung (z. B. Gütekriterien wie Reliabilität und Validität).

(7) Standards zu *assessment for learning* zielen darauf, dass Lehrkräfte in der Lage sein sollten, die Resultate formativer Leistungsmessungen für die Optimierung von Lehr- und Lernprozessen zu nutzen (z. B. durch effektives Feedback). Auch dieser Themenbereich umfasst also Teilkompetenzen, die im Rahmen von VERA relevant sind, da Lehrkräfte aufgefordert sind, die VERA-Ergebnisse ihrer Schulklassen für die eigene Unterrichtsentwicklung zu nutzen.

(8) Die Standards zu *education support for teachers* fallen etwas aus dem Rahmen dieser Systematik, da sie weniger verschiedene Teilkompetenzen von Assessment Literacy als vielmehr deren Erwerb thematisieren. So werden unter anderem Anforderungen an die Aus- und Fortbildung von Lehrkräften formuliert (z. B. Bereitstellung geeigneter Lernmöglichkeiten für die Aneignung von Assessment Literacy, vgl. DeLuca et al., 2016). Auch wird auf die Bedeutung von On-the-Job-Training hingewiesen (z. B. Xu & Brown, 2016).

Die acht im Überblicksbericht von DeLuca und Kollegen (2016) herausgearbeiteten Themenbereiche werden in den hierfür analysierten 15 Publikationen zu Standards für Lehrerkompetenzen im Bereich der Leistungsmessung in unterschiedlichem Maße berücksichtigt und akzentuiert. Als Blaupause für die Formulierung von Kompetenzerwartungen zu Assessment Literacy gelten die im Jahr 1990 vorgelegten *Standards for Teacher Competence in Educational Assessment of Students* (AFT, NCME & NEA, 1990; vgl. auch Helmke, 2010), die allerdings nur Angaben zu den ersten vier von DeLuca und Kollegen (2016) differenzierten Themenbereichen umfassen – mit einem deutlichen Fokus auf *assessment process*. Standards zu diesen vier Bereichen finden sich auch in aktuelleren Publikationen (z. B. Brookhart, 2011; Gardner, Harlen, Hayward & Stobart, 2008; Klinger et al., 2015). Im Unterschied zu den Standards von AFT, NCME & NEA (1990) heben diese (wiederum mit unterschiedlichem Umfang und Gewicht) jedoch noch andere Kompetenzfacetten hervor, die den übrigen vier Themenbereichen zuzuordnen

sind. Besonderer Wert wird dabei in zunehmenden Maße auf Kenntnisse und Fähigkeiten im Bereich *assessment for learning* gelegt (vgl. DeLuca et al., 2016). Eine Ausnahme von diesem Trend stellt das 2012 von der europäischen Association of Educational Assessment (AEA) vorgelegte European Framework of Standards for Educational Assessment (AEA, 2012) dar, das in seiner inhaltlichen Ausrichtung stark an den Standards von AFT, NCME & NEA (1990) orientiert ist und etwa den Themenbereich *assessment for learning* nicht berücksichtigt.

Zusammenfassend ist hervorzuheben, dass die im Überblicksbericht von DeLuca und Kollegen (2016) erfassten Publikationen zu Standards für Lehrerkompetenzen im Bereich der Leistungsmessung, mit Ausnahme des *framework* der AEA (2012), jeweils aus Ländern stammen, in denen erstens der Einsatz von Tests in schulischen Kontexten bereits seit vielen Jahrzehnten tradiert ist (z. B. Ingenkamp, 1989; Langfeldt & Trolldenier, 1993b) und die zweitens als Vorreiter einer am Output orientierten, auf (u. a. auch mittels schulischer Leistungstests erhobenen) Daten gestützten Steuerung des Bildungssystems gelten können (vgl. Wurster, 2016). Es existieren allerdings auch spezifisch auf die Lehrerbildung in Deutschland fokussierte Systematiken, die auch die Frage berühren, über welche Kenntnisse und Fähigkeiten Lehrkräfte im Bereich der Leistungsmessung verfügen sollten. Im Mittelpunkt dieser Systematiken stehen vor allem Kompetenzen von Lehrkräften im Einschätzen und Beurteilen, darüber hinaus werden jedoch (zum Teil auch eher indirekt) Aspekte von Assessment Literacy thematisiert.

Zuvorderst sind hier die *Standards für die Lehrerbildung* zu nennen, die im Jahr 2004 von der Kultusministerkonferenz (KMK) veröffentlicht wurden (KMK, 2004). In diesem Dokument werden „Diagnostik, Beurteilung und Beratung“ als ein separater Kompetenzbereich gefasst. Spezifiziert wird dieser Kompetenzbereich durch Beschreibungen der diagnostischen Aufgaben von Lehrkräften. Diese lauten wie folgt: (1) „Lehrerinnen und Lehrer diagnostizieren Lernvoraussetzungen und Lernprozesse von Schülerinnen und Schülern; sie fördern Schülerinnen und Schüler gezielt und beraten Lernende und deren Eltern.“ sowie (2) „Lehrerinnen und Lehrer erfassen Leistungen von Schülerinnen und Schülern auf der Grundlage transparenter Beurteilungsmaßstäbe.“ (S. 11). Die hierfür erforderlich (Teil-)Kompetenzen werden in Form von Standards ausgeführt, die sich zum einen auf die universitäre Ausbildung von Lehrkräften und zum anderen auf den Vorbereitungsdienst (das Referendariat) beziehen. Mit diesen Standards wird spezifiziert, wozu Studierende des Lehramts jeweils am Ende der beiden Ausbildungsphasen in der

Lage sein sollten. Zum Beispiel wird die Erwartung formuliert, dass die Absolventinnen und Absolventen von Lehramtsstudiengängen unterschiedliche Formen der Leistungsbeurteilung, deren Funktionen sowie deren Vor- und Nachteile kennen. Des Weiteren werden auch Kenntnisse zu verschiedenen Bezugssystemen der Leistungsbeurteilung eingefordert. Thematisiert werden ferner solche Aspekte wie die Berücksichtigung unterschiedlicher Lernvoraussetzungen, das Erkennen von Entwicklungsständen, Lernhindernissen und Lernfortschritten, die Konzeption von Aufgabenstellungen, die Anwendung von Bewertungsmodellen und -maßstäben, die adressatengerechte Begründung von Beurteilungen, Hoch- und Sonderbegabung, Lernprozessdiagnostik, Beratung, Förderung oder Leistungsrückmeldungen.

Detaillierte Beschreibungen der diagnostischen Kenntnisse und Fähigkeiten, die Lehrerinnen und Lehrer im Studium erwerben sollten, werden auch in curricularen Dokumenten wie dem *Rahmencurriculum zu Psychologie in Lehramtsstudiengängen* vorgenommen, das im Jahr 2008 von einer Kommission der DGPs vorgeschlagen wurde (DGPs, 2008). In dem Curriculum findet sich ein separater Bereich zu pädagogisch-psychologischer Diagnostik und Evaluation. Thematisiert werden unter anderem grundlegende methodische Kenntnisse (etwa zu Testtheorien und Gütekriterien), Leistungs- und Verhaltensbeurteilung, Wissen zu Test-, Befragungs- und Beobachtungsverfahren sowie die Evaluation von Unterricht und Methoden der Qualitätssicherung. Auch werden verschiedene diagnostische Aufgabenbereiche von Lehrkräften benannt bzw. differenziert: Lern- und Instruktionsdiagnostik, Entwicklungs- und Erziehungsdiagnostik, Schullaufbahndiagnostik sowie Diagnostik bei Lern- und Verhaltensschwierigkeiten.

Eine weitere, im Jahr 2015 in der Zeitschrift für Pädagogik zur Diskussion gestellte Systematik von Kompetenzbeschreibungen basiert auf der Unterscheidung von vier Arten der Diagnostik – Statusdiagnostik, Prozessdiagnostik, Veränderungsdiagnostik und Verlaufsdiagnostik (von Aufschnaiter et al., 2015). Dabei wird angenommen, dass alle diese Arten der Diagnostik für die Lehrertätigkeit relevant sind und die in ihrem Zusammenhang anfallenden Aufgaben bzw. die damit verbundenen Anforderungen eine hohe Bandbreite an diagnostischen Fähigkeiten als notwendig erscheinen lassen. Diese Fähigkeiten werden als Standards formuliert und drei verschiedenen Kompetenzfacetten zugeorndet: (Fachspezifische) Diagnostik, Befundlagen und Theorien zu (fachspezifischen) kognitiven Kompetenzen und Kompetenzentwicklungen sowie Befundlagen und

Theorien zu (fachspezifischen) motivational-emotionalen Zuständen. In den Standards wird unter anderem auf Wissen über Gütekriterien, auf Kenntnisse zu diagnostischen Verfahren, auf verschiedene methodische Fähigkeiten (v. a. Beobachtung, Dokumentation), auf Kenntnisse zu Urteilsfehlern und -verzerrungen oder auf Wissen zu Kennzeichen von Hoch- und Minderbegabung sowie zu fachspezifisch typischen Kompetenzveränderungen und -entwicklungen referiert. Thematisiert werden beispielsweise auch die Gestaltung von Statusdiagnostik mithilfe verschiedener Aufgabenformate sowie Kenntnisse von Zusammenhängen zwischen verschiedenen schülerseitigen Merkmalen (z. B. Selbstkonzept, Motivation, Interesse) und dem schulischen Lernerfolg.

2.2.3 Die Testskepsis an deutschen Schulen

Mit Blick auf die drei im letzten Abschnitt skizzierten, auf Deutschland bezogenen Systematiken (Standards für die Lehrerbildung der KMK, Rahmencurriculum zu Psychologie in Lehramtsstudiengängen der DGPs, die von Aufschnaiter und Kollegen (2015) vorgeschlagenen Standards für die diagnostische Kompetenz von Lehrkräften) ist insgesamt festzuhalten, dass diese hinsichtlich ihrer Inhalte und Schwerpunkte nicht unwesentlich differieren. Allen drei Konzeptionen ist jedoch gemein, dass sie jeweils auch Kenntnisse und Fähigkeiten im Bereich der Leistungsmessung berühren (z. B. Lernprozessdiagnostik, Gütekriterien, Kenntnis von Testverfahren). Dies geschieht allerdings in einem deutlich geringeren Umfang und sehr viel weniger explizit als in den etwa bei DeLuca und Kollegen (2016) berücksichtigten Standards zu Assessment Literacy. Als Hauptgrund hierfür ist zu vermuten, dass Deutschland, im Vergleich zu den Ländern, aus denen die Asssessment-Literacy-Standards stammen, lange Zeit ein „Entwicklungsland“ für die Nutzung von Testverfahren in schulischen Kontexten war, dem beispielsweise Ingenkamp (1989) eine ausgesprochene „Testaversion“ attestierte. Auch heute noch werden Testverfahren an deutschen Schulen in deutlich geringerem Maße genutzt als beispielsweise in den USA. Dennoch belegen nicht zuletzt die flächendeckend durchgeführten VERA-Erhebungen, dass der Einsatz von Tests auch für die Arbeit von Lehrerinnen und Lehrern in Deutschland an Bedeutung gewonnen hat.

Allerdings zeigt sich gerade am Beispiel von VERA, dass hierzulande zumindest Teile der pädagogischen Praxis dem Einsatz von Testverfahren in Schulen noch immer ausgesprochen skeptisch gegenüberstehen. Zumindest wird dieser Eindruck durch ein im

Jahr 2014 veröffentlichtes und von GEW (Gewerkschaft Erziehung und Wissenschaft), BLLV (Bayerischen Lehrer- und LehrerInnenverband), Grundschulverband und VBE (Verband Bildung und Erziehung) unterzeichnetes „Manifest“ [17] suggeriert, in dem die „Testeritits“ an den Schulen im Allgemeinen und VERA im Speziellen vehement kritisiert werden. In diesem Zusammenhang verdeutlichen auch die Ergebnisse empirischer Forschungsarbeiten, dass der Nutzen und die Diagnosegüte von VERA als vergleichsweise gering eingeschätzt werden (Wurster, Richter, Schliesing & Pant, 2013).[18]

Die Skepsis von Pädagoginnen und Pädagogen gegenüber Testverfahren manifestiert sich allerdings nicht nur in den Einstellungen gegenüber VERA, sondern beispielsweise auch in Bezug auf den Einsatz von Sprachtests, die den Sprachstand von Schülerinnen und Schülern erfassen und dabei helfen sollen, Kinder mit einem besonderen sprachlichen Förderbedarf valider zu identifizieren. Auch hier finden sich Statements der GEW, in denen einer negativen Einstellung gegenüber der Nutzung sprachidagnostischer Verfahren in schulischen und vorschulischen Kontexten Ausdruck verliehen wird. So heißt es in einer Pressemitteilung vom 22. Oktober 2008: „Sprachkompetenz verbessert man nicht durch Tests, sondern durch bewusst erlebte Kommunikation“.[19] Als ein weiteres Indiz für eine Skepsis gegenüber Tests lassen sich die im Berichtsband der IQB-Ländervergleichsstudie 2011 im Primarbereich zur Sprachdiagnostik dargestellten Ergebnisse interpretieren, die verdeutlichen, dass es noch immer einen großen Anteil an Grundschulen gibt, an denen gänzlich auf den Einsatz von sprachdiagnostischen Verfahren verzichtet wird (Stanat, Weirich, et al., 2012).

Dabei gibt es gute Argumente dafür, zur Identifikation von Schülerinnen und Schülern mit geringen sprachlichen Kompetenzausprägungen (zum Beispiel) nicht allein auf die Urteile von Lehrkräften zu vertrauen, sondern auch auf sprachdiagnostische Verfahren zurückzugreifen. So verdeutlichen die unter 2.1.5 skizzierten empirischen Befunde aus genauigkeitskeitsorientierten Forschungsarbeiten zur diagnostischen

[17] Verfügbar unter: https://www.gew-berlin.de/public/media/Beschluss_11.pdf [25.11.2016]

[18] Anhand der Ergebnisse von Wurster und Kollegen (2013) ist diese Schlussfolgerung insbesondere für Lehrkräfte im Land Berlin zu ziehen. Lehrkräfte aus Brandenburg, die in dieser Studie ebenfalls betrachtet wurden, schätzten den Nutzen und die Diagnosegüte von VERA hingegen etwas positiver ein.

[19] Verfügbar unter: https://www.gew.de/aktuelles/detailseite/neuigkeiten/bildungsgipfel-qualitaet-fruehkindlicher-bildung-staerken/ [25.11.2016]

Kompetenz, dass viele Lehrkräfte das Leistungsniveau ihrer Schülerinnen und Schüler nur sehr ungenau beurteilen können und oftmals überschätzen. Hier bieten diagnostische Verfahren (sofern sie elementaren Gütekriterien genügen) den Vorteil, dass sie weniger anfällig für Beurteilungsfehler und -verzerrungen sind. Ein weiterer Vorzug diagnostischer Verfahren besteht darin, dass sie den zumeist auf die eigene Klasse fokussierten Bezugsrahmen von Lehrerurteilen erweitern – etwa durch Bezugnahme auf kriteriale Normen oder auf die Ergebnisse von Normstichproben. Sie erlauben es ferner, Messfehler zu quantifizieren und somit bei der Auswertung und Beurteilung explizit zu berücksichtigen. Darüber hinaus erfassen gute diagnostische Verfahren Merkmale theoretisch fundiert, wohingegen Lehrkräfte nicht selten ein breiteres, von theoretischen Konzeptionen abweichendes und interindividuell differierendes Verständnis der von ihnen zu beurteilenden Schülermerkmale haben (vgl. Rost, Sparfeldt & Schilling, 2006).

Ungeachtet der genannten Vorteile steht der empirische Nachweis, dass der Einsatz von sprachdiagnostischen Verfahren in Grundschulen dabei hilft, die Identifikation von Kindern mit sprachlichem Förderbedarf zu verbessern, derzeit noch aus. Wie bereits in der Einleitung zu dieser Dissertationsschrift erwähnt, findet sich in der Praxis sogar das Vorurteil, dass der Informationsgehalt von sprachdiagnostischen Verfahren nicht über das hinausgeht, was Pädagoginnen und Pädagogen ohnehin aus ihren alltäglichen Beobachtungen erschließen. An diesem Punkt setzt die *Teilstudie 2* (Kapitel 5) an, die der Frage nachgeht wird, inwiefern die Güte von diagnostischen Entscheidungen zum sprachlichen Bedarf mit der Nutzung verschiedener Informationsquellen (z. B. Beobachtungen von Lehrkräften, sprachdiagnostische Verfahren) kovariiert. Daran anknüpfend wird in *Teilstudie 3* (Kapitel 6) untersucht, welche konkreten sprachdiagnostischen Verfahren an den deutschen Grundschulen zum Einsatz kommen.

Die Abschnitte des nachfolgenden Teilkapitels 2.3 sollen dazu dienen, inhaltlich in das Thema *Sprachdiagnostik in der Grundschule* einzuführen. Hierbei wird zunächst zwischen Sprachentwicklungsstörungen und umgebungsbedingten Sprachauffälligkeiten unterschieden (2.3.1). Im Anschluss daran werden Erklärungsansätze für die Entstehung von Sprachauffälligkeiten skizziert und es wird erläutert, welche Schülergruppen im Fokus der von Pädagoginnen und Pädagogen im Primarbereich durchgeführten Sprachdiagnostik liegen (2.3.2). Des Weiteren wird auf die auch für die Sprachdiagnostik bedeutsame Unterscheidung zwischen integrierter und additiver Sprachförderung hingewiesen (2.3.3). Da der Einsatz von sprachdiagnostischen Verfahren jeweils im

Mittelpunkt der Teilstudien 2 und 3 steht, werden zudem Anforderungen an diese Verfahren herausgearbeitet (2.3.4). Danach wird überblicksartig skizziert, was bislang über die Nutzung von sprachdiagnostischen Verfahren im Primarbereich bekannt ist, wobei auch eine konzise Darstellung der Situation im Elementarbereich erfolgt (2.3.5). Abschließend soll noch das Konzept der sprachdiagnostischen Kompetenz umrissen werden (2.3.6).

2.3 Sprachdiagnostik in der Grundschule

2.3.1 Sprachentwicklungsstörungen und umgebungsbedingte Sprachauffälligkeiten

Sprachliche Kompetenzen sind Schlüsselqualifikationen. Sie stellen eine wesentliche Voraussetzung des schulischen Lernens dar, sie bilden eine wichtige Grundlage für eine erfolgreiche Bildungskarriere und sind mitentscheidend für den Schulerfolg und die späteren Berufschancen (z. B. Holler, 2007; Lüdtke & Kallmeyer, 2007; Voet Cornelli, Geist, Grimm & Schulz, 2012). Nicht zuletzt kommen ihnen auch soziale und psychologische Funktionen zu, da sie die gesellschaftliche Teilhabe ermöglichen und in bedeutendem Maße zur Persönlichkeitsentwicklung beitragen (vgl. Ehlich, 2005b; Furnham, 1990). Dementsprechend nachteilig kann es sich auswirken, wenn bei Kindern sprachliche Beeinträchtigungen bzw. *Sprachauffälligkeiten* vorliegen (Voet Cornelli et al., 2012).

Sprachauffälligkeiten sind dadurch gekennzeichnet, dass die sprachlichen Kompetenzen eines Kindes deutlich geringer ausgeprägt sind als bei Gleichaltrigen und somit unter dem Kompetenzniveau liegen, das für das jeweilige Alter zu erwarten ist (Ehlich, 2005b; Lüdtke & Kallmeyer, 2007). Die Ursachen hierfür lassen sich vor allem zwei Gruppen von Faktoren zuordnen:

(1) *Pathologisch* bedingte Sprachauffälligkeiten (vgl. Lüdtke & Kallmeyer, 2007) werden als Sprachentwicklungsstörungen bezeichnet. Diese lassen sich wiederum nach primären beziehungsweise *spezifischen Sprachentwicklungsstörungen* (SSES)[20] und nach *sekundären Sprachentwicklungsstörungen* (SES) differenzieren. Letztere treten sekundär als Folge oder im Rahmen von anderen Primärerkrankungen auf (z. B. Hörschädigungen, kognitive Einschränkungen). SSES betreffen hingegen ausschließlich den sprachlichen

[20] SSES werden synonym auch als umschriebene Sprachentwicklungsstörungen (USES) bezeichnet (z. B. Neumann & Euler, 2013).

Bereich, alle anderen Fähigkeiten eines Kindes sind altersgemäß entwickelt (z. B. Kannengieser, 2012; Rothweiler, 2013; Rupp, 2013). Als Ursachen von SSES werden vor allem genetische Faktoren vermutet (Rothweiler, 2013). Die Prävalenzrate, also der Anteil eines Jahrgangs, der von einer SSES betroffen ist, liegt in Deutschland bei 5 bis 8 Prozent. Es wird zwischen einem synchronen und asynchronen Störungsprofil unterschieden. Bei ersterem ist die sprachliche Entwicklung in allen sprachlichen Ebenen[21], d. h. in der *phonetisch-phonologischen* (incl. suprasegmentaler und prosodischer Elemente), der *morphologisch-syntaktischen*, der *semantisch-lexikalischen* sowie der *pragmatisch-kommunikativen* Sprachebene (z. B. Kany & Schöler, 2014) in vergleichbarem Ausmaß verzögert. Hingegen sind bei SSES mit asynchronem Profil nur einzelne sprachliche Ebenen betroffen (Rupp, 2013). Bei mehrsprachig aufwachsenden Kindern wirkt sich die Störung nicht selektiv auf eine Sprache aus, sondern betrifft alle Sprachen, die das Kind erwirbt (Rothweiler, 2013). Kinder, die eine SSES haben, bedürfen einer speziellen *Sprachtherapie*, die in der Regel von Logopäden durchgeführt wird. Angeordnet werden Sprachtherapien durch Kinderärzte. Ihnen obliegt auch die Aufgabe, SSES zu diagnostizieren (Geist & Voet Cornelli, 2015; Voet Cornelli et al., 2012).

(2) *Umgebung*sbedingte Sprachauffälligkeiten sind nicht durch pathologische Faktoren bedingt, sondern auf ungünstige Bedingungen beim Spracherwerb zurückzuführen; von einigen Autorinnen und Autoren werden sie auch als „Sprachdefizite[] ohne medizinischen Störungswert“ (z. B. Neumann & Euler, 2013, S. 177) beschrieben. Hierzu zählen etwa anregungsarme sprachliche Umgebungen, ein später Erwerbsbeginn, mangelnde Lerngelegenheiten bzw. eine insgesamt zu kurze Kontaktdauer mit der deutschen Sprache oder falsche Sprachvorbilder (z. B. Geist, 2014; Lüdtke & Kallmeyer, 2007; Neumann & Euler, 2013; Voet Cornelli et al., 2012). Bei umgebungsbedingten Sprachauffälligkeiten ist keine Sprachtherapie angezeigt. Vielmehr soll eine *Sprachförderung* helfen, die Sprachkompetenz der betroffenen Kinder zu stärken. Sprachförderung kann dabei als „Sammelbegriff für eine Interventionsform bei Sprachdefiziten ohne medizinischen Störungswert [verstanden werden]“ (Neumann & Euler, 2013, S. 177). Entsprechende Sprachfördermaßnahmen werden in der Regel von Heilpädagoginnen und -pädagogen, Erzieherinnen und Erziehern oder Lehrerinnen und

[21] Zum Teil findet sich auch die Bezeichnung „sprachliche *Komponenten*“ (z. B. Weinert, S. & Grimm, 2008).

Lehrern durchgeführt. Diese Berufsgruppen sind häufig auch an der Diagnostik sprachlichen Förderbedarfs beteiligt (Voet Cornelli et al., 2012).

2.3.2 Fokus der Sprachdiagnostik im Primarbereich

Vielfach wird betont, dass eine Förderung sprachlicher Kompetenzen möglichst frühzeitig, also bereits im Elementarbereich, einsetzen sollte, um allen Schülerinnen und Schülern angemessene Entwicklungsmöglichkeiten zu garantieren und somit zu verhindern, dass einzelne Kinder bereits zu Beginn ihrer Bildungskarriere unaufholbar „abgehängt" werden. Im Bildungssystem ist die Sprachdiagnostik zur Identifikation von Kindern mit sprachlichem Förderbedarf dementsprechend vor allem auf den Elementarbereich fokussiert (Autorengruppe Bildungsberichterstattung, 2012; Becker-Mrotzek et al., 2013; z. B. Bund-Länderinitiative zur Sprachförderung, 2012; Lengyel, 2012; Lisker, 2013; Paetsch, Wolf, Stanat & Darsow, 2014; Redder et al., 2011). So finden mittlerweile in vielen Bundesländern für alle Kinder verpflichtende, vorschulische Sprachstandserhebungen statt, bei denen der sprachliche Entwicklungsstand systematisch erfasst wird (Becker-Mrotzek et al., 2013). Im Primarbereich ist hingegen die Teilnahme an Sprachstandserhebungen in nur wenigen Bundesländern und auch nur für bestimmte Schülergruppen, vor allem für Kinder mit nichtdeutscher Herkunftssprache, obligatorisch (vgl. Redder et al., 2011).

Auch die Forschungsliteratur zum sprachlichen Förderbedarf im Primarbereich fokussiert oftmals vor allem die Gruppe von *Schülerinnen und Schülern mit nichtdeutscher Herkunftssprache* (z. B. Jeuk, 2009). Hintergrund dieser Schwerpunktsetzung ist die (durch empirische Befunde gestützte) Annahme, dass insbesondere sukzessiv bilinguale Kinder, die Deutsch erst ab dem zweiten Lebensjahr, also beispielsweise erst mit dem Besuch der Kindertagesstätte (Kita) erwerben (Rothweiler, 2007; Rothweiler & Ruberg, 2011), auch noch in der Grundschule einen Förderbedarf hinsichtlich ihrer sprachlichen Kompetenzen aufweisen, dessen Identifizierung eine systematische Erfassung des Sprachstands auf den oben genannten sprachlichen Ebenen erfordert (Ehlich, 2005b; Knapp, 1999). Darüber hinaus ist dokumentiert, dass bei sukzessiv bilingualen Kindern ein erhöhtes Risiko kinderärztlicher Fehldiagnosen besteht, die schlussendlich auf Schwierigkeiten bei der Differenzierung von SSES und umgebungsbedingten Sprachauffälligkeiten zurückgehen. Hierbei wird zwischen zwei Arten von Fehldiagnosen unterschieden: Bei einer *missed identity* wird das Vorliegen

einer SSES *übersehen*, etwa, weil die beobachteten sprachlichen Auffälligkeiten fälschlicherweise als umgebungsbedingt klassifiziert werden. Bei einer *mistaken identity* wird *fälschlicherweise* eine SSES *diagnostiziert*, etwa, weil nicht erkannt wird, dass die beobachteten sprachlichen Auffälligkeiten nicht auf pathologische, sondern auf umgebungsbedingte Faktoren zurückzuführen sind (Genesee, Paradis & Crago, 2004; Paradis, 2005). Für die betroffenen Kinder haben diese Fehldiagnosen zur Folge, dass sie die jeweils adäquaten Förder- oder Therapiemaßnahmen nicht erhalten.

Im Zuge der Diskussion über Ursachen für den bei Kindern mit nichtdeutscher Herkunftssprache nicht selten auch noch im Primarbereich vorhandenen sprachlichen Förderbedarf wird in der Forschungsliteratur häufig auf die Besonderheiten des schulspezifischen Gebrauchs der deutschen Sprache hingewiesen. Dieser Gebrauch hebt sich deutlich von der Alltagssprache ab, da er durch ein sprachliches Register geprägt ist, das eigene formale Anforderungen hat und unter anderem durch konzeptionelle Schriftlichkeit (z. B. hohe Informationsdichte und Elaboriertheit sowie geringe kontextuelle bzw. situative Einbettung von sprachlichen Äußerungen, vgl. Koch, P. und Oesterreicher (1985)), einen anspruchsvolleren Wortschatz und eine komplexere Grammatik gekennzeichnet ist (vgl. Berendes, Dragon, Weinert, Heppt & Stanat, 2013; Heppt, 2016). Für diese „Schulsprache“ finden sich in der Forschungsliteratur mehrere, konzeptuell sehr ähnliche bzw. nur im Detail differierende Konstrukte. Hierbei ist das Konzept der *Bildungssprache* (z. B. Feilke, 2012; Gogolin, 2009; Gogolin & Lange, 2011) im deutschsprachigen Raum mittlerweile etabliert. Ähnliche Konstrukte sind zum Beispiel die *Alltägliche Wissenschaftssprache* (z. B. Ehlich, 1995), die *schriftförmige Rede* (z. B. Dehn, 2011), *language of schooling*, *academic language*, *register of schooling* oder *genres of schooling* (vgl. Schleppegrell, 2001, 2004; Schleppegrell, 2012; Snow, 2010). Vielfach wird den bildungssprachlichen Fähigkeiten eine hohe Bedeutung für den schulischen Lernerfolg beigemessen (z. B. Berendes et al., 2013; Heppt, 2016), da sie Voraussetzung dafür sind, dass Schülerinnen und Schüler am kommunikativen Unterrichtsgeschehen teilhaben und das im Unterricht vermittelte Wissen erschließen und nutzen können.

Konzeptuell geht die Differenzierung von Alltags- und Bildungssprache insbesondere auf Bernstein (1962, 1964) zurück, der zwischen einem restringierten Code (z. B. einfaches Vokabular, geringe Komplexität der konzeptuellen Hierarchie sprachlicher Äußerungen) und einem elaborierten Code (z. B. differenziertes Vokabular, höhere

Komplexität sprachlich-konzeptueller Strukturen) unterscheidet. Für die Zweitsprach-erwerbsforschung aufgegriffen wurde diese Konzeption von Cummins (1984), der zwei unterschiedliche Gruppen sprachlicher Fähigkeiten und Fertigkeiten differenziert: *BICS* (*basic interpersonal communication skills*), die vor allem sprachliche Äußerungen betreffen, die kognitiv wenig anspruchsvoll und kontextuell stark eingebettet sind, sowie *CALP* (*cognitive academic language proficiency*), die vor allem auf den Umgang mit sprachlichen Äußerungen zielen, die kognitiv anspruchsvoll, dekontextualisiert und konzeptionell schriftlich sind (vgl. auch Berendes et al., 2013). Cummins (1984) nahm an, dass bilinguale Kinder durch die alltägliche Kommunikation in der Zweitsprache die Fertigkeiten zur alltäglichen Konversation (BICS) oftmals relativ schnell erlernen, während der Erwerb der kognitiven, akademischen (quasi bildungssprachlichen) Sprachfähigkeiten (CALP) deutlich mehr Zeit benötigt. Ein zweisprachig aufwachsendes Kind kann also einerseits gut in der Lage sein, im Alltag beziehungsweise in Bezug auf konzeptionell mündliche sprachliche Äußerungen zu kommunizieren, und gleichzeitig Schwierigkeiten damit haben, die Anforderungen der bildungssprachlich geprägten Schulsprache zu bewältigen. Für die Sprachdiagnostik besteht in solchen Fällen also die besondere Herausforderung, dass ein hinsichtlich bildungssprachlicher Fähigkeiten vorhandener Förderbedarf übersehen werden kann, weil ein Kind ansonsten über gute (unauffällige) alltagssprachliche Fähigkeiten verfügt (vgl. Junk-Deppenmeier, 2009; Koch, K., 2012).

Wie bereits erwähnt, ist die Differenzierung von BICS und CALP als zwei unterschiedliche Aspekte der Sprachbeherrschung vor allem als theoretische Erklärung für geringere Bildungserfolge von Zweitsprachlernenden entwickelt worden (vgl. Cummins, 1984). Eine geringe Ausprägung bildungssprachlicher Fähigkeiten ist allerdings nicht nur häufig bei Kindern festzustellen, die spät mit dem Zweitspracherwerb begonnen haben, sondern kann auch einsprachig aufwachsende Kinder betreffen. Als Ursache hierfür hatte bereits Bernstein (1962, 1964) auf die Bedeutung der schicht-spezifischen sprachlichen Sozialisation hingewiesen. Seinen Annahmen zufolge sind Schülerinnen und Schüler aus der *lower working class* gegenüber Kindern aus der *middle class* hinsichtlich ihrer Bildungschancen benachteiligt, da sie unter den Sozialisations-bedingungen, unter denen sie aufwachsen, lediglich den restringierten, nicht aber den als Voraussetzung für den schulischen Lernerfolg bedeutsamen elaborierten sprachlichen Code erwerben (s. o., vgl. auch Bourdieu, 1977). Auch in der heutigen Forschungslitera-

tur wird eine geringe Ausprägung von bildungssprachlichen Fähigkeiten bei monolingualen Kindern auf soziale bzw. familiale Faktoren wie das Aufwachsen in einer bildungsfernen, einkommensschwachen Familie oder ein ungünstiger sprachlicher Input bzw. falsche Sprachvorbilder (vgl. auch die unter 2.3.1 skizzierten Ursachen zum Entstehen umgebungsbedingter Sprachauffälligkeiten) zurückgeführt (vgl. Berendes et al., 2013; siehe z. B. auch Chall, Jacobs & Baldwin, 1990).

Trotz der Bedeutung der Bildungssprache für den schulischen Lernerfolg und der Tatsache, dass sowohl Schülerinnen und Schüler mit nichtdeutscher Herkunftssprache als auch einsprachig aufwachende Kinder einen bildungssprachlichen Förderbedarf haben können, ist derzeit noch ein Mangel an standardisierten Verfahren zu konstatieren, mit deren Hilfe reliabel und valide geprüft werden kann, ob Schülerinnen und Schüler, die Anforderungen des unterrichtsspezifischen Gebrauchs der deutschen Sprache bewältigen können (z. B. Heppt, Stanat, Dragon, Berendes & Weinert, 2014; Schuth, Heppt, Köhne, Weinert & Stanat, 2015; Uesseler, Runge & Redder, 2013). Gleichsam ist festzustellen, dass gegenwärtig für den Primarbereich (und insbesondere für einsprachig aufwachsende Kinder) nur eine geringe Anzahl an diagnostischen Verfahren zur Verfügung steht, die explizit mündliche Sprachkompetenzen[22] fokussieren (z. B. Geist, 2014; Schulz, 2013; Voet Cornelli et al., 2012). Diese sind vor allem in der Grundschule von besonderer Bedeutung, da die schulische Wissensvermittlung schwerpunktmäßig im Medium der Mündlichkeit erfolgt (vgl. Böhme, 2012). Um schulische Lernangebote angemessen nutzen zu können, müssen Schülerinnen und Schüler daher nicht zuletzt auch in der Lage sein, spezifische mündliche Anforderungen zu bewältigen. Diese Anforderungen sind unter anderem dadurch gekennzeichnet, dass sich die Unterrichtskommunikation von der Alltagskommunikation durch bestimmte formale und informale Regeln bzw. durch die Art der Interaktion von der Alltagskommunikation unterscheidet. So verbringen Schülerinnen und Schüler einen großen Teil der Unterrichtszeit damit zuzuhören (Belgrad, Eriksson, Pabst-Weinschenk & Vogt, 2008). Dies erfordert, dass sie ihr Zuhörverhalten entsprechend steuern können müssen. Ein weiterer Unterschied zur Alltagskommunikation betrifft das Rederecht, das bei der Unterrichtskommunikation vollkommen anders geregelt ist als in nichtschulischen Kontexten (vgl. Böhme, 2012).

[22] Im Zeitschriftenbeitrag zu Teilstudie 3 wird synonym zu „mündliche Sprachkompetenzen“ der Terminus „lautsprachliche Kompetenzen“ verwendet.

Die Ursache dafür, dass für den Primarbereich praktisch keine (oder bestenfalls lediglich auf Kinder mit nichtdeutscher Herkunftssprache abzielende) sprachdiagnostischen Verfahren (wie z. B. Tests, Beobachtungsverfahren) existieren, ist unter anderem darin zu suchen, dass das Interesse der Spracherwerbsforschung an der Aneignung von mündlichen Sprachkompetenzen insbesondere bei einsprachigen Kindern ab dem Grundschulalter deutlich nachlässt (Ehlich, 2005b). Stark verknappt ausgedrückt, gilt der Erwerb mündlicher Sprachkompetenzen zu diesem Zeitpunkt als bereits sehr weit vorangeschritten, sodass die weitere Entwicklung der betreffenden sprachlichen Ebenen „nur noch“ durch Prozesse der Stabilisation, Erweiterung, Ausdifferenzierung und Verfeinerung gekennzeichnet ist. Als Folge des dergestalt verteilten Forschungsinteresses liegen für den Primarbereich (und darüber hinaus) kaum Forschungsergebnisse zur normgerechten Erstsprachentwicklung mündlicher Sprachkompetenzen vor, auf denen etwa eine Konstruktion entsprechender diagnostischer Instrumente zur Feststellung dieser Kompetenzen basieren könnte (Ehlich, 2005b). Wenn Forschungsarbeiten den Spracherwerb ab dem Grundschulalter betrachten, dann wird vielmehr zumeist auf die Aneignung *schriftsprachlicher Kompetenzen* fokussiert (z. B. Owens, 2015). Dementsprechend ist auch die Sprachdiagnostik im Primarbereich vor allem von dem Ziel geprägt, die Ausprägung der schriftsprachlichen Kompetenzen von Schülerinnen und Schülern zu untersuchen.

2.3.3 Integrierte und additive Sprachförderung

Insbesondere an den Grundschulen unterrichten Lehrkräfte nicht selten in Schulklassen, die hinsichtlich der jeweiligen sozialen und herkunftsbezogenen Bedingungen des Erwerbs der deutschen Sprache sehr heterogen sind. Als Resultat der im vorangehenden Abschnitt beschriebenen Mechanismen stehen sie mithin vor der Herausforderung, dass die (bildungs-)sprachlichen Kompetenzen ihrer Schülerinnen und Schüler häufig sehr unterschiedlich ausfallen und für einige Kinder eine besondere Unterstützung erforderlich ist, um die für den schulischen Lernerfolg notwendigen sprachlichen Fähigkeiten aufzubauen. Im Konzept der durchgängigen Sprachbildung ist diese Unterstützung im Regelunterricht integriert. Hierbei werden Lehrkräfte unter anderem dazu angehalten, sich die sprachlichen Anforderungen ihres Unterrichts bewusst zu machen und als Sprachvorbilder zu fungieren. Dies schließt zum Beispiel ein, dass sie ihre Sprache *nicht* zu stark vereinfachen (also im Unterricht etwa nur Alltagssprache verwenden), sondern vielmehr einen reichhaltigen, variationsreichen sprachlichen Input sicherstellen

(vgl. FörMig-Transfer Berlin, 2009; Gogolin et al., 2011). Ferner sollten Lehrkräfte ihren Schülerinnen und Schülern in einem kooperativen Lernprozess entsprechende Brücken bzw. Lerngerüste bauen, durch die es diesen möglich wird, sprachliche Anforderung zu bewältigen, die eigentlich noch über ihrem aktuellen sprachlichen Entwicklungsniveau liegen (*Scaffolding*, z. B. Donato (1994); Gibbons (2002); Wood, Bruner und Ross (1976)). Intendiert ist weiterhin, dass der Unterricht sprachintensiv gestaltet wird, also etwa hohe Sprechanteile für die einzelnen Schülerinnen und Schüler vorgesehen werden. Darüber hinaus soll das Sprachlernen nicht nur auf den Deutschunterricht beschränkt sein, sondern auch explizit in anderen Fächern erfolgen (vgl. FörMig-Transfer Berlin, 2009; Gogolin et al., 2011).

Allerdings ist anzunehmen, dass die im Regelunterricht implementierten Unterstützungsmaßnahmen der durchgängigen Sprachbildung für einige Schülerinnen und Schüler mit besonders ausgeprägten Schwächen nicht hinreichend sind, um ein für ein erfolgreiches Lernen erforderliches bildungssprachliches Mindestniveau zu erreichen. Für diese Kinder erscheint eine Förderung notwendig, die *zusätzlich* zum Regelunterricht in einer kleineren Lerngruppe stattfindet. Damit solche *additiven* Fördermaßnahmen gezielt eingesetzt und gegebenenfalls kompensatorisch wirksam werden können, sollten sie vor allem mit denjenigen Schülerinnen und Schülern durchgeführt werden, die auch tatsächlich einen Sprachförderbedarf aufweisen. Additive Fördermaßnahmen erfordern also Selektionsentscheidungen, bei denen bestimmt wird, welche Kinder an den Maßnahmen teilnehmen sollten und welche nicht. Diese sollten natürlich nicht zufällig erfolgen, sondern idealiter erst nach einer prognostisch validen Sprachdiagnostik.

Bei sprachdiagnostischen Selektionsentscheidung zum Förderbedarf besteht allerdings die Herausforderung, dass gegenwärtig weder in der schulischen Praxis noch in der Forschungsliteratur allgemein gültige, konsensual geteilte Kriterien existieren, die spezifizieren, ab wann ein Kind eine additive Sprachförderung erhalten sollte. Dies zeigt sich etwa bei den (zumeist verpflichtenden) Sprachstandserhebungen im Elementarbereich. Hierbei werden in den einzelnen Bundesländern zum einen jeweils andere Verfahren verwendet und zum anderen unterschiedliche kriteriale Definitionen von (additivem) Sprachförderbedarf angelegt (Autorengruppe Bildungsberichterstattung, 2012; Becker-Mrotzek et al., 2013). Ob ein Kind an einer speziellen Sprachfördermaßnahme teilnimmt oder nicht ist also unter anderem davon abhängig, in welchem Bundesland es aufwächst (vgl. hierzu auch Abschnitt 2.3.5). In der Forschungsliteratur

wird vorgeschlagen, bei einsprachig aufwachsenden Kindern dann förderbedürftige Sprachauffälligkeiten zu diagnostizieren, wenn deren Leistung in einem validen sprachdiagnostischen Verfahren ein bis anderthalb Standardabweichungen unter dem Altersdurchschnitt liegt (Neumann & Euler, 2013). Im Kontrast dazu findet sich die Empfehlung, die Diagnose einer SSES bzw. eines Sprachtherapiebedarfs ab einer Abweichung von anderthalb (AWMF, 2011) bis zwei Standardabweichungen (WHO, 1992) von der Altersnorm zu vergeben (vgl. auch Law, Boyle, Harris, Harkness & Nye, 2000). An verschiedener Stelle wurde versucht, geeignete Cut-Off-Werte für die diagnostische Differenzierung zwischen sprachunauffälligen, pädagogisch sprachförderbedürftigen und therapiebedürftigen Kindern auf Basis empirischer Ergebnisse zu bestimmen (z. B. Neumann & Euler, 2013; Sachse, Anke & von Suchodoletz, 2007). Methodisch werden dabei die aus Cut-Off-Verschiebungen resultierenden Verteilungen mit einer validen Referenzklassifikation der jeweils untersuchten Kinder in diese drei Kategorien verglichen. Hierbei besteht allerdings die Herausforderung, dass erstens kein Sprachtest mit „Goldstandardqualität“ vorliegt, auf denen eine valide Referenzklassifikation basieren könnte, und zweitens auch Expertenurteile aufgrund oftmals geringer Urteilerübereinstimmungswerte keine idealen Vergleichsmaßstäbe bieten (Euler et al., 2010).

2.3.4 Anforderung an Verfahren für die Sprachdiagnostik

In der vergangenen Dekade wurden in mehreren Publikationen Kriterienkataloge vorgelegt, in denen Anforderungen formuliert werden, denen sprachdiagnostische Verfahren als Voraussetzung für eine adäquate Erfassung sprachlicher Kompetenzen genügen sollten (Becker-Mrotzek et al., 2013; Lüdtke & Kallmeyer, 2007). Als Folie hierfür diente vielmals eine vom Bundesministerium für Bildung und Forschung in Auftrag gegebene, durch ein Konsortium um Konrad Ehlich verfasste Expertise zu „Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Sprachförderung von Kindern mit und ohne Migrationshintergrund“ (Ehlich, 2005a, 2005b). Im Folgenden sollen die wesentlichen der in der Forschungsliteratur genannten Anforderungen an sprachdiagnostische Verfahren, die sich insbesondere aus der psychologischen Testtheorie oder aus Theorien der Sprachwissenschaft bzw. Befunden der Spracherwerbsforschung ableiten lassen (vgl. Voet Cornelli et al., 2012), knapp dargestellt werden. Darüber hinaus soll die Unterscheidung von Selektions- und Förderdiagnostik skizziert werden. Abschließend wird ein Bewertungs-

rahmen vorgestellt, der als Grundlage für eine Evaluation von sprachdiagnostischen Verfahren entwickelt wurde.

Testtheoretische Anforderungen

Testtheoretische Anforderungen gelten nicht speziell für die Sprachdiagnostik, sondern für diagnostische Verfahren und deren Einsatz im Allgemeinen. Hierzu wird in der Forschungsliteratur eine ganze Reihe von (Test-) Gütekriterien formuliert, die oftmals nach Haupt- und Nebengütekriterien differenziert werden (z. B. Lienert & Raatz, 1998). Als Hauptgütekriterien gelten die klassischen Gütemerkmale der Objektivität, Reliabilität und Validität und ihre Facetten (z. B. Paralleltest- vs. Retest-Reliabilität, konvergente vs. diskriminante Validität). Als Nebengütekriterien haben sich derzeit sieben weitere Kriterien etabliert, die die Skalierung beziehungsweise Skalierbarkeit, die Normierung (auch: Eichung), die Ökonomie, die Nützlichkeit, die Zumutbarkeit, die Unverfälschbarkeit und die Fairness eines Testverfahrens betreffen (z. B. Moosbrugger & Kelava, 2012). Ferner finden sich Qualitätsrichtlinien, wie etwa die *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), die auf den genannten Haupt- und Nebengütekriterien (sowie auf weiteren Qualitätsaspekten) basieren und etwa dabei helfen können, die Güte von Testverfahren zu beurteilen, für die jeweiligen diagnostischen Zielstellungen geeigneten Verfahren auszuwählen oder eine Testung in Übereinstimmung mit bestimmten Qualitätsmaßstäben durchzuführen.

Sprachwissenschaftliche Anforderungen

In sprachwissenschaftlichen Konzeptionen zu Sprachstandserhebungen im *Elementarbereich* wird die Anforderung formuliert, dass alle sprachlichen Ebenen bzw. Subsysteme erfasst werden müssten, um den sprachlichen Entwicklungsstand eines Kindes diagnostizieren zu können (z. B. Lüdtke & Kallmeyer, 2007).[23,24] Dies wird unter

[23] Bei Ehlich (2005b) werden anstelle von sprachlichen Ebenen bzw. Komponenten sogenannte sprachliche Basisqualifikationen unterschieden. Diese Konzeption betont zum einen die Entwicklungsdynamik des Spracherwerbs und zum anderen die enge Verzahnung einzelner sprachlicher Teilkompetenzen: In der Summe bilden die Basisqualifikationen einen Qualifikationsfächer, der angeeignet und systematisch aufgebaut wird und zum sprachlichen Handeln in konkreten Situationen befähigt. Die Abfolge der Basisqualifikationen ist an der Erwerbsdominanz im Verlauf der Sprachaneignung orientiert (vgl. auch Redder, 2013; Redder & Weinert, 2013). Auch Ehlich (2005b) fordert, dass in Sprachstandserhebungen der Qualifikationsfächer vollständig untersucht werden sollte.

anderem damit begründet, dass von den sprachlichen Fähigkeiten in einem Bereich nicht auf das Kompetenzniveau in einem anderen Bereich geschlossen werden kann (z. B. Voet Cornelli et al., 2012). Zudem wird argumentiert, dass die verschiedenen sprachlichen Ebenen in komplexer Weise miteinander interagieren – und zwar sowohl im Verlauf der sprachlichen Entwicklung als auch im konkreten sprachlichen Handeln (vgl. Ehlich, 2005b; Owens, 2015; Redder, 2013). Auf den Primarbereich ist diese Anforderung insofern übertragbar, als dass auch dort, wie bereits unter 2.3.2 dargestellt, nicht allein Schriftsprachkompetenzen, sondern auch mündliche Kompetenzen untersucht werden sollten. Zudem wird empfohlen, bei der Sprachdiagnostik sowohl rezeptive (Lesen, Zuhören) als auch produktive (Schreiben, Sprechen) sprachliche Kompetenzen zu berücksichtigen (z. B. Schulz, Tracy & Wenzel, 2008; Voet Cornelli et al., 2012).

Andere Autorinnen und Autoren postulieren, dass die Sprachdiagnostik sprach-erwerbstheoretisch begründet sein sollte (vgl. Geist, 2014; Voet Cornelli et al., 2012). Dies bedeutet, dass die für die sprachliche Entwicklung in einem bestimmten Alter relevanten Teilkompetenzen fokussiert und dabei insbesondere diejenigen sprachlichen Phänomene betrachtet werden, die empirisch gut untersucht sind (vgl. auch Ehlich, 2005b). Zudem wird argumentiert, dass einmalige Feststellungen des Sprachstands aufgrund der Dynamik des Sprachaneignungsprozesses lediglich Momentaufnahmen mit begrenzter Aussagekraft seien. Stattdessen wird ein prozessbegleitendes Vorgehen mit Mehrfacherhebungen empfohlen (z. B. Bredel, 2005; Ehlich, 2005b; Lengyel, 2012).

Weitere Anforderungen an die Sprachdiagnostik werden aus der Forschung zum Zweitspracherwerb abgeleitet. Diese beinhalten zum Beispiel, dass bei Kindern mit nichtdeutscher Herkunftssprache auch Informationen zu den jeweiligen Erwerbs-bedingungen erhoben werden (z. B. Beginn des Zweitspracherwerbs, Kontaktdauer mit der Zweitsprache und Qualität des sprachlichen Inputs), um etwa den jeweiligen Zweitspracherwerbstyp (z. B. simultan vs. sukzessiv bilingual) zu ermitteln (z. B. Voet Cornelli et al., 2012). Darüber hinaus wird in verschiedenen Beiträgen zur Sprach-diagnostik bei Mehrsprachigkeit gefordert, sowohl den Sprachstand in der Erst- als auch

[24] Die Forderung, dass bei der Sprachdiagnostik alle sprachlichen Ebenen erfasst werden sollten, ist nicht unumstritten (z. B. Geist, 2014), auch hängt ihre Relevanz vom jeweiligen Ziel der Diagnostik ab. Zum Beispiel werden in Sprachscreenings ganz bewusst nur einige wenige Sprachausschnitte (mit einem möglichst hohen Vorhersagewert) untersucht, um den Sprachstand einer größeren Anzahl von Kindern möglichst zeitökonomisch erheben zu können (Becker-Mrotzek et al., 2013; Lengyel, 2012).

in der Zweitsprache zu erfassen (z. B. Reich, 2003), um SSES bei Kindern mit nichtdeutscher Herkunftssprache zu identifizieren und das Risiko zu reduzieren, durch ungünstige Erwerbsbedingungen begründete Auffälligkeiten im Zweitspracherwerb fälschlicherweise als SSES zu interpretieren. Kontrovers diskutiert wird, inwiefern separate Normen für Zweitsprachlernende bereitgestellt werden sollten (vgl. Lengyel, 2012; McNamara, 2005; Reich, 2003; Schulz, 2013; Schulz & Tracy, 2011).

Selektionsdiagnostik und Förderdiagnostik

In der pädagogischen Literatur findet sich die (allerdings recht künstliche beziehungsweise stark zugespitzte) Unterscheidung zwischen einer *Selektions- und* einer *Förderdiagnostik* (z. B. Ingenkamp & Lissmann, 2008). Die Selektionsdiagnostik zielt vorrangig darauf, Leistungen zu klassifizieren und Platzierungsentscheidungen (z. B. Zuweisung zu einer bestimmten Fördermaßnahme) vorzubereiten. Demgegenüber fokussiert die Förderdiagnostik auf die Erfassung von spezifischen Informationen dazu, worin der individuelle Förderbedarf besteht, sodass auf Basis dieser Daten gezielte Fördermaßnahmen geplant und durchgeführt werden können. Als Ideal wird formuliert, dass sprachdiagnostische Verfahren beides leisten sollten, also zum einen dabei helfen, Kinder mit sprachlichem Förderbedarf zu identifizieren und zum anderen zumindest erste Hinweise auf geeignete Fördermaßnahmen geben (vgl. Geist, 2014; Lengyel, 2012).

Bewertungsrahmen zur Evaluation sprachdiagnostischer Verfahren

Viele der hier skizzierten Anforderung an sprachdiagnostische Verfahren finden sich in einem Bewertungsrahmen wieder, der im Jahr 2013 von einer vom Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache geleiteten Expertenkommission vorgelegt (Becker-Mrotzek et al., 2013) und zur Evaluation der in den Kitas mutmaßlich am häufigsten genutzten Sprachstandsverfahren entwickelt wurde (Neugebauer & Becker-Mrotzek, 2013). Der Bewertungsrahmen erhebt den Anspruch, auf alle Formen von Sprachstandsverfahren anwendbar zu sein, das heißt „auf Tests, Screenings, Einschätzverfahren und Beobachtungsbögen, die Aufschluss darüber geben sollen, ob ein Kind zusätzliche Sprachförderung benötigt oder nicht“ (Becker-Mrotzek et al., 2013, S. 9). Bei der Evaluation werden insgesamt zehn „Handlungsfelder“ betrachtet, die jeweils durch mehrere Qualitätsmerkmale unterlegt sind und auf einer vierstufigen Skala kriterial bewertet werden. Die zehn Handlungsfelder beinhalten zum einen testtheoretische Anforderungen (s. o.) wie *Objektivität*, *Reliabilität*, *Validität* und *Normierung* der

Verfahren, *zeitliche Anforderungen* (i. S. v. Testökonomie) sowie das der Validität zuzuordnende Kriterium der *Fehlerquote* (i. S. v. Sensitivität bzw. Spezifität), das durch seine separate Nennung ein besonderes Gewicht erhält. Zum anderen betreffen die Handlungsfelder sprachwissenschaftliche und spracherwerbstheoretische Anforderungen. Hierzu zählen der Umfang der *Berücksichtigung der sprachlichen Basisqualifikationen* nach Ehlich (2005b), der Umgang mit *Mehrsprachigkeit*, Angaben zur *Qualifizierung der pädagogischen Fachkräfte*, die das jeweilige Sprachstandsverfahren anwenden sollen, und die *Spezifität der Diagnostik* im Sinne der Ableitbarkeit konkreter Sprachfördermaßnahmen aus den mit dem jeweiligen Instrument ermittelten Ergebnissen (Becker-Mrotzek et al., 2013).

Erstmalig angewendet wurde der Bewertungsrahmen in einer ebenfalls im Jahr 2013 von Neugebauer und Becker-Mrotzek vorgelegten Studie, bei der die nach Angaben der zuständigen Ministerien zu diesem Zeitpunkt in den einzelnen Ländern bei Sprachstandserhebungen im Elementarbereich eingesetzten sprachdiagnostischen Verfahren[25] evaluiert wurden. Hierbei zeigte sich, dass viele Verfahren Mängel aufwiesen und nur wenige die meisten der formulierten Kriterien erfüllten (Neugebauer & Becker-Mrotzek, 2013). Allerdings wurden verschiedene Aspekte der Untersuchung kritisch diskutiert. Zum Beispiel wurde darauf hingewiesen, dass der Evaluationsbericht nur wenige Informationen zum methodischen Vorgehen bei der Analyse und Bewertung der betrachteten Sprachstandserhebungsverfahren enthielt. Hinterfragt wurde auch, dass im Bericht die Qualitätsmerkmale bzw. Evaluationskriterien vollkommen ungewichtet und ohne Berücksichtigung von Art und Zielstellung der einzelnen Verfahren (z. B. Tests vs. Screenings vs. Beobachtungsverfahren) angelegt wurden (vgl. Hoffmann & Böhme, 2014a; Maihack, 2014; Maihack, Grimm, Schöler, Becker-Mrotzek & Neugebauer, 2014).

2.3.5 Aktueller Stand der Sprachdiagnostik im Elementar- und Primarbereich

In den vergangenen Jahren haben die Länder ihre Aktivitäten in der vorschulischen und schulischen Sprachförderung deutlich intensiviert (Becker-Mrotzek et al., 2013; Neugebauer & Becker-Mrotzek, 2013; Redder et al., 2011). Um Bildungsdisparitäten entgegenzuwirken, wurden von der Kultusministerkonferenz unter anderem die „Verbesserung der Sprachkompetenz bereits im vorschulischen Bereich“ (KMK, 2002,

[25] Insgesamt wurden 21 Verfahren bewertet.

S. 6) und die „durchgängige Verbesserung der Lesekompetenz “ (S. 7) als wichtige Ziele definiert. Ihren Ausdruck finden diese kurz nach der Jahrtausendwende formulierten Absichtserklärungen heute zum Beispiel in der verstärkten Förderung von Forschungs- und Entwicklungsprogrammen zur Sprach- und Leseförderung (z. B. Bund-Länderinitiative zur Sprachförderung, 2012) sowie insbesondere in der Erprobung und Implementation verschiedener Sprachfördermaßnahmen (vgl. Autorengruppe Bildungsberichterstattung, 2012). Wie bereits weiter oben erwähnt, haben mittlerweile fast alle Bundesländer im Elementarbereich verpflichtende Sprachstandserhebungen eingeführt, bei denen mithilfe einer breiten Vielfalt an sprachdiagnostischen Verfahren wie Tests, Screenings, Einschätzverfahren oder Beobachtungsbögen die sprachlichen Kompetenzen von Kindern erhoben und Sprachförderbedarf frühzeitig und valide diagnostiziert werden soll (vgl. Autorengruppe Bildungsberichterstattung, 2012; Lisker, 2013).

In der Forschungsliteratur finden sich mehrere Berichte, die einen Überblick über die Verfahren geben, die in den einzelnen Bundesländern bei Sprachstandserhebungen im Elementarbereich zum Einsatz kommen. Exemplarisch sei hier die vom Deutschen Jugendinstitut (DJI) in Auftrag gegebene, 2010 veröffentlichte und 2013 nochmals aktualisierte Expertise „Sprachstandsfeststellung und Sprachförderung im Kindergarten sowie beim Übergang in die Schule“ (Lisker, 2013) genannt, die auch (in einer aktualisierten Fassung) die Grundlage für die im Nationalen Bildungsbericht 2012 dargestellten Statistiken zu „[s]prachliche[n] Kompetenzen der Kinder vor der Einschulung“ bildet (Autorengruppe Bildungsberichterstattung, 2012, S. 62).

Insgesamt verdeutlichten die Expertise des DJI und die im Nationalen Bildungsbericht 2012 referierten Statistiken, dass die Bundesländer bei der Erfassung des Sprachstandes von Kindern im Elementarbereich uneinheitlich vorgehen und zumeist jeweils andere Instrumente einsetzen. So dokumentiert die DJI-Expertise die Angaben von 14 Ländern, in denen insgesamt 18 verschiedene sprachdiagnostische Verfahren verwendet werden. Dabei nutzen einige Länder bereits etablierte, standardisierte Verfahren, wie den im Land Bremen eingesetzten Sprachtest des niederländischen Cito-Instituts (Citogroep, 2004). Andere, wie etwa das Land Hamburg mit HAVAS 5 (Reich & Roth, 2003), haben die Entwicklung eigener Instrumente in Auftrag gegeben. Ferner differieren die in den Ländern eingesetzten Verfahren im Hinblick auf Erhebungsziel und Methodik. Neben relativ breit angelegten, umfassenden Sprachtests werden auch

zeitsparende Screenings eingesetzt, die nicht direkt auf die Diagnose von Förderbedarf zielen, sondern die Notwendigkeit weiterer (differential-) diagnostischer Schritte ermitteln sollen. Ferner kommen Beobachtungsverfahren mit geringerem Standardisierungsgrad zum Einsatz, wie Seldak (Ulich & Mayr, 2006) und Sismik (Ulich & Mayr, 2003). Darüber hinaus wurde gezeigt, dass sich die Sprachstandserhebungen der Länder zudem dahingehend unterscheiden, in welchem Alter und wie häufig der Sprachstand der Kinder überprüft wird. Dementsprechend wurde im Nationalen Bildungsbericht 2012 resümiert, dass die von den Ländern bei den Sprachstands-erhebungen eingesetzten Verfahren „nur eingeschränkt vergleichbar sind, da sie nicht das Gleiche erheben“ (Autorengruppe Bildungsberichterstattung, 2012, S. 62). Empirisch untermauert wird diese These zum Beispiel durch eine Studie von Settinieri (2010), in der die konvergente Validität der in Sprachstandserhebungen eingesetzten Verfahren Sismik und Delfin 4 untersucht wurde. Die Autorin fand nur geringe bis mittlere Korrelationen zwischen den Skalen dieser Instrumente. In Übereinstimmung mit den Autorinnen und Autoren anderer Publikationen forderte sie daher eine systematische Überprüfung der Messgüte der in den Sprachstandserhebungen eingesetzten Verfahren (Becker-Mrotzek et al., 2013; Neugebauer & Becker-Mrotzek, 2013; Redder et al., 2011; Settinieri, 2010).

Hervorzuheben sind ferner die Angaben im Nationalen Bildungsbericht 2012 zur Größe des jeweiligen Anteils an Kindern, die in den Ländern als förderbedürftig eingestuft werden. Dass dieser Anteil nicht unerheblich zwischen den Ländern variiert (zwischen 13 % im Saarland und 42 % im Land Bremen), wird von einigen Autorinnen und Autoren unter anderem auf Diskrepanzen in der Definition des sprachlichen Förderbedarfs zwischen den Ländern und auf das unterschiedliche Vorgehen bei den Sprachstandserhebungen zurückgeführt (Neugebauer, Becker-Mrotzek & Stanat, 2014). Allerdings ist bei der Interpretation dieser hohen Spannweite zu berücksichtigen, dass identische Förderquoten gar nicht zu erwarten sind, da sich die einzelnen Länder hinsichtlich ihrer Bevölkerungszusammensetzung (z. B. Anteil von Kindern mit Zuwanderungshintergrund) deutlich unterscheiden.

Weitere Überblicksberichte verdeutlichen, dass die in den einzelnen Ländern zur Sprachstandserhebung im Elementarbereich eingesetzten Verfahren dem Zweit-spracherwerb von Kindern mit nichtdeutscher Herkunftssprache und den damit verbundenen Besonderheiten in unterschiedlichem Maße und in unterschiedlicher Weise Rechnung tragen (z. B. Lengyel, 2012; Lüdtke & Kallmeyer, 2007). Teilweise wird

Mehrsprachigkeit nicht gesondert berücksichtigt, teilweise werden sprachbiografische Hintergrundinformationen erfasst, beispielsweise zum Beginn des Zweitspracherwerbs oder zur Intensität und Qualität des Kontakts mit der Zweitsprache (z. B. Fit in Deutsch, MK NI, 2003). Mit einigen wenigen Verfahren (z. B. HAVAS 5, Cito-Sprachtest) kann zusätzlich der Sprachstand in der Erstsprache festgestellt werden (vgl. Böhme, 2012). In einigen Fällen werden (für den Sprachstand im Deutschen) sogar separate Normen für Kinder mit nichtdeutscher Herkunftssprache bereitgestellt (z. B. Cito-Sprachtest, Lise-DaZ) (Citogroep, 2004; Schulz & Tracy, 2011).

Methodisch basierten die in der DJI-Expertise und in weiteren Berichten dokumentierten Übersichten zu den im Elementarbereich eingesetzten sprachdiagnostischen Verfahren vor allem auf den Ergebnissen von Befragungen der jeweils zuständigen Ländereinrichtungen. Eine Ausnahme hiervon bildet ein Bericht des Hamburger Zentrums zur Unterstützung der wissenschaftlichen Begleitung und Erforschung schulischer Entwicklungsprozesse (ZUSE, Redder et al., 2011). In diesem Bericht werden nicht nur die Ergebnisse einer Länderbefragung zu den bei Sprachstandserhebungen eingesetzten sprachdiagnostischen Verfahren dargestellt, sondern auch Instrumente angeführt, die in einer Literaturrecherche zu weiteren publizierten sprachdiagnostischen Verfahren ermittelt wurden. Der im ZUSE-Bericht dokumentierte Überblick umfasst also sowohl Angaben zu den in Sprachstands-erhebungen der Länder verwendeten Verfahren als auch zu Instrumenten, die von den einzelnen Kitas über die landesweiten Sprachstandserhebungen hinaus potenziell eingesetzt werden könnten. Des Weiteren ist der ZUSE-Bericht nicht nur auf den Elementarbereich beschränkt. Auch für den Primarbereich wurden Ländervorgaben zu Sprachstandserhebungen erfragt. Hierbei wurde ermittelt, dass im Vergleich zum Elementarbereich in deutlich weniger Ländern konkrete Vorgaben zu den zu verwendenden Verfahren existieren. Diese Vorgaben haben zudem häufig nur Empfehlungscharakter und fokussieren bestimmte Gruppen von Schülerinnen und Schülern (z. B. Kinder mit nichtdeutscher Herkunftssprache). Darüber hinaus umfasst der im ZUSE-Bericht vorgenommene Überblick auch die in der Literaturrecherche zusätzlich gefundenen sprachdiagnostischen Verfahren, die für Kinder im Grundschulalter vorliegen (Redder et al., 2011).

Insgesamt finden sich im ZUSE-Bericht steckbriefartig aufbereitete Angaben zu 58 verschiedenen sprachdiagnostischen Verfahren. Hiervon zielen 30 Instrumente auf den

Elementarbereich und 21 Verfahren auf Kinder im Grundschulalter. Darüber hinaus dokumentiert der Bericht, dass viele der in anderen Publikationen für den Elementarbereich resümierten Herausforderungen und Optimierungsbedarfe auch für den Primarbereich gelten: So steht auch hier eine größere Anzahl an verschiedenen Verfahren zur Verfügung, die hinsichtlich ihrer theoretischen und empirischen Fundierung, der erfassten Kompetenzen, ihrer Methodik und testtheoretischen Güte sowie der Berücksichtigung von Mehrsprachigkeit differieren (Redder et al., 2011). Allerdings basiert der im ZUSE-Bericht dokumentierte Überblick, wie bereits erläutert, auf den Angaben zuständiger Ländereinrichtungen und den Ergebnisse einer Literaturrecherche. Der Bericht bildet somit zum einen ab, welche sprachdiagnostischen Verfahren die einzelnen Länder für Sprachstandserhebungen im Primarbereich vorgeben bzw. empfehlen. Zum anderen wird berücksichtigt, welche sprachdiagnostischen Instrumente darüber hinaus noch potenziell zur Verfügung stehen und somit an den Grundschulen zum Einsatz kommen könnten. Der Bericht erlaubt jedoch nur bedingt Rückschlüsse auf die Frage, welche Verfahren an den Grundschulen tatsächlich genutzt werden und wie häufig sie zum Einsatz kommen. Um Aussagen hierzu treffen zu können, wäre eine direkte Befragung an Grundschulen erforderlich.

Eine solche direkte Befragung von Grundschulen zu der von ihnen umgesetzten Sprachdiagnostik war bislang nur Gegenstand einer einzigen Untersuchung. In dieser im Jahr 2014 in der Dissertation von Geist veröffentlichten Studie gaben die an den Grundschulen jeweils für die Sprachdiagnostik zuständigen Sprachförderkräfte an, oftmals selbstentwickelte, hausinterne Verfahren wie Frage- oder Beobachtungsbögen einzusetzen und eher selten auf publizierte Instrumente zurückzugreifen. Allerdings wurde die Studie ausschließlich an hessischen Grundschulen durchgeführt. Außerdem bezog sich die Befragung exklusiv auf die Sprachdiagnostik bei der Schulanmeldung und der Sprachförderung in vorschulischen Vorlaufkursen (Geist, 2014). Da die Studie somit de facto auf den Elementarbereich fokussierte und zudem auf Hessen beschränkt war, kann auch sie keinen Aufschluss über die bundesweit an Grundschulen eingesetzten sprachdiagnostischen Verfahren geben.

2.3.6 Sprachdiagnostische Kompetenz von Lehrkräften

Im Zusammenhang mit der Feststellung von Sprachförderbedarf bestehen nicht nur Anforderungen an die hierbei zum Einsatz kommenden sprachdiagnostischen Verfahren.

Auch die Lehrkräfte, die an der Sprachdiagnostik beteiligt sind, müssen besondere Anforderungen bewältigen. So fordert etwa Bredel (2005), dass in der Ausbildung von Lehramtsstudierenden Kenntnisse und Fähigkeiten zu folgenden Themenbereichen vermittelt werden sollten: kriteriengeleitete Auswahl von Erhebungsinstrumenten beziehungsweise Bekanntmachen mit den standardmäßig anzuwendenden Instrumenten, handlungspraktische Kompetenzen bei der Durchführung von Sprachstandserhebungen, Grundlagen der qualitativen und quantitativen Auswertung von Sprachstandserhebungen sowie kriteriengeleitete Prüfung der Erhebungsresultate (S. 109).

Hopp, Thoma und Tracy (2010) haben ein sprachwissenschaftliches Modell zur Sprachförderkompetenz pädagogischer Fachkräfte entwickelt, in dem die Sprachdiagnostik als ein separater Bereich aufgeführt wird, der sich in die Teilbereiche Methoden, Instrumente und Konsequenzen gliedert. Unterlegt sind die Teilbereiche mit (hier nicht so genannten) Standards, die unter anderem auf die Differenzierung von SSES und umgebungsbedingten Sprachauffälligkeiten, auf die Kenntnis grundlegender Methoden und Verfahren, auf die Auswahl von jeweils geeigneten sprachdiagnostischen Verfahren, auf die Nutzung diagnostischer Informationen für die Sprachförderung und auf eine Reflexion der Grenzen sprachdiagnostischer Verfahren (z. B. Messfehler) zielen (vgl. auch Geist, 2014).

An anderer Stelle formuliert Wildemann (2010) die Erwartung, dass insbesondere Deutschlehrkräfte sprachdiagnostische Expertise aufbauen sollten. Als Bestandteile dieser Expertise werden genannt: sprachliches und curriculares Wissen (vor allem zur Aneignung schriftsprachlicher Kompetenzen), Kompetenzen in Bezug auf die Prognose zu erwartender Lernerfolge und Kenntnisse einschlägiger sprachdiagnostischer Verfahren. Diese Bestandteile sollen zu einer angemessenen Auswahl sprachdiagnostischer Verfahren, zur Vermeidung von Bezugsgruppeneffekten (vgl. sozialnormorientiere Bewertung), zu einer kriterienorientierten Auswertung der Ergebnisse und zu einer adäquaten Dokumentation der Lernentwicklung befähigen. Darüber hinaus hebt Wildemann (2010) die Notwendigkeit von Selbstreflexion und Evaluation des eigenen sprachdiagnostischen Vorgehens hervor.

Eine weitere Konzeption *zur sprachdiagnostischen Kompetenz* findet sich bei Geist (2014), die in Anlehnung an das Prozessmodell diagnostischer Kompetenz von Klug und Kollegen (2013) ein drei Phasen umfassendes Modell formuliert (siehe Abbildung 2.2), das die jeweils zu bewältigenden diagnostischen Anforderungen spezifiziert. Diese

Phasen bzw. Anforderungen lauten: (1) Auswahl, Vorbereitung, Durchführung und Auswertung der Sprachstandserhebung, (2) Beurteilung der sprachlichen Fähigkeiten und der Sprachentwicklung der untersuchten Kinder (inkl. einer Entscheidung dazu, ob eine Förderbedarf besteht oder nicht) sowie (3) Ableitung des individuellen Förderbedarfs und Formulierung von Förderzielen (Geist, 2014, S. 68).

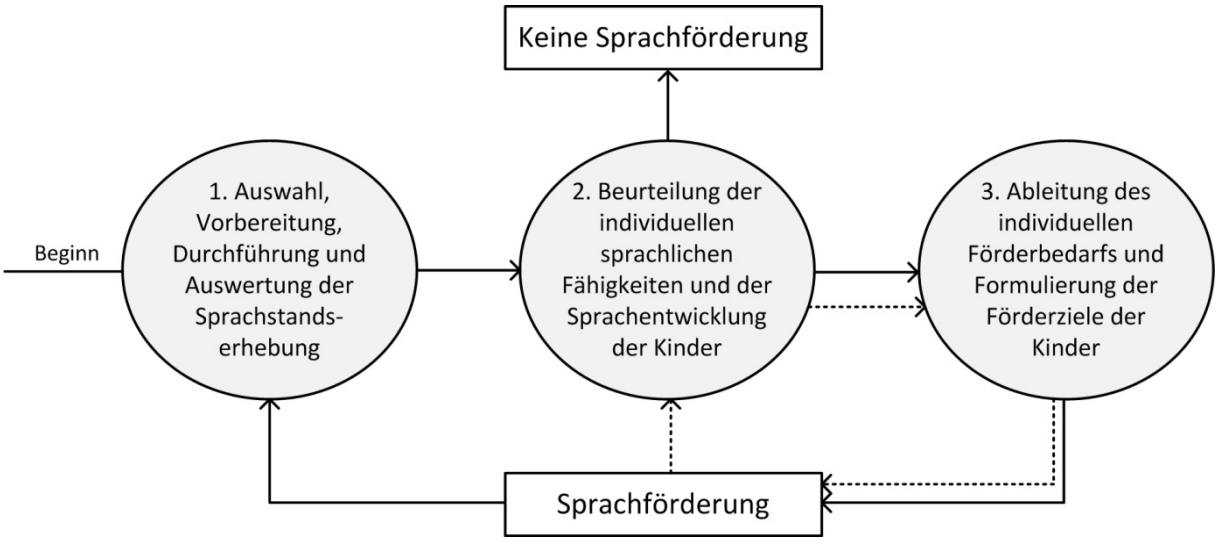

Abbildung 2.2: Prozessmodell zu sprachdiagnostischen Fähigkeiten von Sprachförderkräften von Geist (2014)

Insgesamt kann also festgehalten werden, dass in theoretischen Überlegungen und Konzeptionen zu den Anforderungen, die Lehrkräfte bei der Sprachdiagnostik bewältigen müssen, nicht zuletzt auch die Bedeutung von Kompetenzen hervorgehoben wird, die bereits im Abschnitt 2.2.2 als Bestandteile von Assessment Literacy beschrieben wurden.

2.3.7 Zusammenfassung

Lehrkräfte stehen in ihrem Beruf nicht nur vor der Herausforderung, die Leistungen von Schülerinnen und Schülern oder die Schwierigkeit von Aufgaben möglichst akkurat zu beurteilen, sondern sind noch mit weiteren diagnostischen Tätigkeiten und Anforderungen konfrontiert. Zusätzliche diagnostische Aufgaben erwachsen beispielsweise aus der Durchführung von schulischen Leistungsmessungen – etwa im Zuge von VERA oder im Rahmen der Sprachdiagnostik. Hierbei haben Lehrkräfte beispielsweise die Aufgabe, die Testergebnisse ihrer Schülerinnen und Schüler zu interpretieren und angemessene Schlussfolgerungen daraus zu ziehen. Unter Umständen müssen sie (z. B. im Falle der Sprachdiagnostik) sogar entscheiden, welche diagnostischen Verfahren sie einsetzen. Die

hierfür erforderlichen Kompetenzen werden zum Beispiel im Konzept der *Assessment Literacy* beschrieben.

Die Erfassung der sprachlichen Kompetenzen von Schülerinnen und Schülern im Rahmen einer Selektionsdiagnostik zielt überwiegend darauf ab, Kinder zu identifizieren, die umgebungsbedingte Sprachauffälligkeiten in einem Maße aufweisen, das eine zusätzliche Förderung ihrer sprachlichen Kompetenzen außerhalb des Regelunterrichts notwendig erscheinen lässt. Zum Einsatz von sprachdiagnostischen Verfahren (und insbesondere von Tests) bei der Feststellung sprachlichen Förderbedarfs in der Grundschule finden sich in der Praxis (wie zu den VERA-Tests auch) noch immer kritische Stimmen, die etwa bezweifeln, dass Sprachtests Informationen liefern, die über die Alltagsbeobachtungen von Lehrkräften hinausgehen. Vor dem Hintergrund dieser noch immer verbreiteten Auffassung soll in *Teilstudie 2* (Kapitel 5) untersucht werden, inwiefern die Güte von Entscheidungen zum Sprachförderbedarf von Kindern mit der Nutzung von diagnostischen Informationsquellen wie sprachdiagnostische Verfahren oder Beobachtungen von Lehrkräften kovariiert.

Bislang fehlt es an Informationen dazu, welche konkreten sprachdiagnostischen Verfahren bereits heute (und trotz bestehender Vorbehalte) an den Grundschulen zum Einsatz kommen. Dieses Desiderat wird in *Teilstudie 3* (Kapitel 6) aufgegriffen, in der die Angaben von Schulleitungen zu den an ihren Grundschulen verwendeten sprachdiagnostischen Instrumenten ausgewertet werden. Hierbei erfolgt auch eine differenzierte Betrachtung der jeweils genannten Verfahren vor dem Hintergrund der unter 2.3.4 skizzierten Anforderungen an sprachdiagnostische Verfahren.

3

Überblick über die Teilstudien der vorliegenden Arbeit

3 Überblick über die Teilstudien der vorliegenden Arbeit

Die Beurteilung der Schwierigkeit von im Unterricht oder bei Lernerfolgskontrollen verwendeten Aufgaben zählt zu den zentralen diagnostischen Anforderungen, die Lehrkräfte in ihrem Beruf bewältigen müssen. Akkurate Schwierigkeitsurteile sind insbesondere dann von Relevanz, wenn schwierigkeitsadäquate Aufgaben ausgewählt werden sollen, um das unterrichtliche Lernangebot an die Lernvoraussetzungen der Schulklasse oder einzelner Schülerinnen und Schüler anzupassen (s. Kap. 2.1.2, vgl. auch Anders und Kollegen, 2010, sowie Brunner und Kollegen, 2011). Aufgrund dieser Funktion für das zielorientierte, adaptive Unterrichten wurden die in genauigkeitsorientierten Untersuchungsansätzen formulierten Definitionen der diagnostischen Kompetenz von Lehrkräften in nicht wenigen Publikationen um den Aspekt der Genauigkeit von Schwierigkeitseinschätzungen erweitert (Anders et al., 2010; Artelt & Gräsel, 2009; McElvany et al., 2009; Südkamp et al., 2008). Allerdings wurde die Genauigkeit von Schwierigkeitseinschätzungen bislang deutlich seltener untersucht als die Akkuratheit von Lehrerurteilen zu Schülerleistungen (s. Kap. 2.1.5, vgl. McElvany et al., 2009). Diese ungleiche Verteilung des Forschungsinteresses ist in besonderem Maße in Bezug auf die Frage zu konstatieren, welche Faktoren die Urteilsgenauigkeit bedingen oder moderieren. Die hierzu durchgeführten Studien fokussieren mit wenigen Ausnahmen auf Urteile von Lehrerinnen und Lehrern zu Schülerleistungen. Die Bedingungen und Kovariaten der Genauigkeit von Lehrereinschätzungen zur Schwierigkeit von Aufgaben wurden hingegen nur in wenigen Studien (vgl. Lintorf, 2012; McElvany et al., 2009) explizit betrachtet (s. auch Abschnitt 2.1.6).

Die in Kapitel 4 dargestellte Teilstudie 1 (Titel: *Wie gut können Grundschullehrkräfte die Schwierigkeit von Deutsch- und Mathematikaufgaben beurteilen? Eine Untersuchung zur Genauigkeit aufgabenbezogener Lehrerurteile auf Klassenebene*.) soll den bisherigen Forschungsstand zur Genauigkeit von Lehrereinschätzungen zur Aufgabenschwierigkeit erweitern. Die Untersuchung basiert auf den Daten einer Normierungsstudie des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin, an der im Jahr 2007 mit dem Ziel durchgeführt wurde, Kompetenzstufenmodelle zu den Bildungsstandards in den Fächern Deutsch und Mathematik im Primarbereich (KMK, 2005a, 2005b) sowie Instrumente für eine bundesweite Testung zu entwickeln. In der Normierungsstudie wurde auch ein Fragebogen eingesetzt, in dem diejenigen Lehrkräfte (Deutschlehrkräfte:

N = 239, Mathematiklehrkräfte: N = 133), deren Schülerinnen und Schüler die bildungsstandardbasierten Testaufgaben zu den Fächern Deutsch und Mathematik bearbeiteten, unter anderem beurteilten sollten, wie schwierig ausgewählte Aufgaben für die Kinder ihrer Klasse sind. Anhand dieser Datengrundlage wird zunächst der Forschungsfrage nachgegangen, wie genau Grundschullehrkräfte beurteilen können, wie schwierig einzelne Aufgaben für die Schülerinnen und Schüler ihrer Klasse sind. Hierzu werden die Rang-, Niveau- und Differenzierungskomponente der Genauigkeit der Schwierigkeitseinschätzungen ermittelt. Darüber hinaus wird mithilfe von Mehrebenenanalysen untersucht, welche Faktoren mit der Über- oder Unterschätzung der Schwierigkeit von Aufgaben zusammenhängen. Als mögliche Kovariaten der Urteilsgenauigkeit werden hierbei sowohl Lehrermerkmale (Berufserfahrung, Kontaktdauer mit der Klasse) als auch Aufgabenmerkmale (psychometrische Itemschwierigkeit) berücksichtigt. Zusätzlich werden Zusammenhänge mit weiteren, von anderen empirischen Forschungsarbeiten bislang nicht betrachteten Aspekten untersucht. Diese Aspekte beziehen sich auf weitere Daten, die Lehrkräfte möglicherweise nutzen könnten, um die Schwierigkeit von Aufgaben zu bestimmen. Diese Daten umfassen Angaben dazu, wann (d. h. in welchem Schuljahr) und wie intensiv (d. h. wie häufig) die jeweils zur Aufgabenlösung erforderlichen Teilkompetenzen im Unterricht vermittelt und trainiert wurden.

Insgesamt lässt sich aus den Ergebnissen genauigkeitsorientierter Forschungsarbeiten zur diagnostischen Kompetenz von Lehrkräften resümieren, dass die Genauigkeit von Lehrerurteilen vielfach gering ist (z. B. Schrader, 2013). Problematisch ist eine geringe Urteilsgenauigkeit insbesondere dann, wenn Fehlurteile mit unmittelbar negativen Konsequenzen für die betroffenen Schülerinnen und Schüler verbunden sind. Dies ist etwa bei Übergangsempfehlungen für die Sekundarstufe I der Fall, da sie über die weitere schulische und berufliche Karriere von Kindern mitentscheiden (z. B. Stubbe & Bos, 2008). Auch bei der Feststellung sprachlichen Förderbedarfs sind genaue Urteile und valide Entscheidungen wesentlich. Wenn ein Kind zwar sprachlichen Förderbedarf hat, dieser aber nicht erkannt wird, dann wird es nicht die benötigte kompensatorische Unterstützung durch eine zusätzliche Sprachförderung erhalten und als Konsequenz in der Schule möglicherweise weniger erfolgreich sein. Unerwünschte Folgen sind auch dann zu vermuten, wenn ein Kind als förderbedürftig diagnostiziert wird, obwohl es tatsächlich keinen Förderbedarf hat: Zum einen bindet die Förderung eines jeden Kindes, das ohne

tatsächlichen Bedarf eine zusätzliche Sprachförderung erhält, Ressourcen (z. B. zeitliche, personelle, finanzielle Mittel), die dann für Schülerinnen und Schüler mit tatsächlichem Bedarf nicht mehr zur Verfügung stehen. Zum anderen ist denkbar, dass sich eine nicht erforderliche Teilnahme an einer zusätzlichen Sprachförderungen auch negativ auf die Lernmotivation und die Schülerleistungen auswirken können, etwa als Folgen des Erlebens von Unterforderung und Langeweile (z. B. Pekrun, Hall, Goetz & Perry, 2014) oder als Ergebnis von Erwartungseffekten (z. B. Gasteiger-Klicpera, Klicpera & Schabmann, 2001; Murphy, Campbell & Garavan, 1999).

Angesichts der Befunde zur Genauigkeit von Lehrerurteilen ist zu vermuten, dass die diagnostische Güte von Entscheidungen zum sprachlichen Förderbedarf eher gering ausfallen dürfte, wenn Grundschulen den Förderbedarf von Schülerinnen und Schülern ausschließlich basierend auf Beobachtungen bzw. Einschätzungen von Lehrkräften ermitteln. Folglich ist empfehlenswert, dass Grundschulen bei der Sprachdiagnostik auf weitere diagnostische Informationsquellen zurückgreifen. Hierbei stellt sich jedoch die Frage, welche Informationsquellen dazu geeignet sind, die Güte von Entscheidungen zum sprachlichen Förderbedarf zu erhöhen. Diese Frage soll in Teilstudie 2 empirisch untersucht werden (Titel: *Wird sprachlicher Förderbedarf in der Grundschule sicher erkannt? Zur Klassifikationsgüte von diagnostischen Entscheidungen*, Kap. 5). Der Fokus der Studie liegt dabei vor allem auf den Effekten des Einsatzes von sprachdiagnostischen Verfahren, denen in der Praxis zum Teil noch immer mit einer gewissen Skepsis begegnet wird.

Im Zeitschriftenbeitrag zur Teilstudie 2 wird zunächst herausgearbeitet, dass Sprachförderentscheidungen dichotom sind und sich mithin bei der Klassifikation von Schülerinnen und Schülern als förderbedürftig oder nicht förderbedürftig zwei Arten von Fehlentscheidungen differenzieren lassen. Dieser Unterscheidung wird Rechnung getragen, indem die Güte der Klassifikation anhand von zwei (interdependenten) Indikatoren (Sensitivität und Spezifität) bestimmt wird. Die Untersuchung basiert auf Daten der IQB-Ländervergleichsstudie 2011 in der Primarstufe (Stanat, Pant, Böhme & Richter, 2012), wobei die dort festgestellten sprachlichen Kompetenzen von Schülerinnen und Schülern (N = 12808) mit den Angaben der befragten Schulleitungen zu der an ihren Grundschulen umgesetzten Sprachdiagnostik verknüpft werden. Anhand dieser Daten wird untersucht, inwiefern die Nutzung von Ergebnissen aus diagnostischen Verfahren sowie die Berücksichtigung weiterer diagnostischer Informationen zu den sprachlichen

Kompetenzen von Grundschulkindern, wie Beobachtungen von Lehrkräften oder Angaben von Eltern, zu einer Steigerung der Sensitivität und Spezifität von Sprachförderentscheidungen beitragen können.

Bislang ist wenig dazu bekannt, welche konkreten sprachdiagnostischen Verfahren bereits heute an Grundschulen eingesetzt werden. Zwar findet sich in dem unter 2.3.5 beschriebenen Bericht des ZUSE-Instituts auch für den Primarbereich eine Übersicht zu sprachdiagnostischen Verfahren (Redder et al., 2011). Da diese Übersicht jedoch auf den Angaben der jeweils zuständigen Ländereinrichtungen sowie auf ergänzenden Literaturrecherchen zu weiteren einschlägigen Verfahren und nicht auf einer direkten Befragung von Grundschulen basiert, kann sie nicht abbilden, welche Instrumente tatsächlich verwendet werden und wie hoch der Verbreitungsgrad einzelner Verfahren ist. Es ist folglich weitgehend offen, welche Verfahren in welchem Umfang bei der Feststellung von sprachlichem Förderbedarf an den Grundschulen zum Einsatz kommen. Diese offene Frage wird in Teilstudie 3 bearbeitet (Titel: *Mit welchen diagnostischen Verfahren wird in Grundschulen Sprachförderbedarf festgestellt? Eine bundesweite Bestandsaufnahme*, Kap. 6), die ebenfalls auf Daten aus der IQB-Ländervergleichsstudie 2011 bzw. auf den darin erfassten Angaben von Schulleiterinnen und Schulleitern ($N = 1227$) zu den an ihren Schulen als diagnostische Informationsquelle genutzten sprachdiagnostischen Verfahren basiert und dementsprechend auch der inhaltlichen Ergänzung von Teilstudie 2 dient.

Der Einsatz von sprachdiagnostischen Verfahren stellt bestimmte Anforderungen an die Lehrkräfte, die an der Sprachdiagnostik beteiligt sind. Insbesondere sind Kompetenzen gefordert, die Bestandteil des Konzepts der Assessment Literacy sind (DeLuca et al., 2016). Lehrkräfte können etwa vor der Aufgabe stehen, Testergebnisse interpretieren oder die Eignung von Verfahren für die Diagnostik sprachlichen Förderbedarfs beurteilen zu müssen. Auch in Teilstudie 3 wird untersucht, ob die Verfahren, die laut den Angaben der Schulleitungen an den Grundschulen eingesetzt werden, basalen Anforderungen (vgl. Abschnitt 2.3.4) genügen. Hierbei wird etwa geprüft, ob die genannten Verfahren tatsächlich auf die Erfassung sprachlicher Kompetenzen von Kindern im Grundschulalter zielen, inwiefern sie Kinder mit nichtdeutscher Herkunftssprache spezifisch berücksichtigen, ob sie auf mündliche oder auf schriftliche Sprachkompetenzen fokussieren und wie sie in Bezug auf testtheoretische Gütekriterien einzuschätzen sind.

4

Teilstudie 1

4 Teilstudie 1 (Originalarbeit)

Wie gut können Grundschullehrkräfte die Schwierigkeit von Deutsch- und Mathematikaufgaben beurteilen? Eine Untersuchung zur Genauigkeit aufgabenbezogener Lehrerurteile auf Klassenebene

How Elementary School Teachers Judge the Difficulty Levels of German Language and Mathematics Tasks: A Study on the Accuracy of Teacher Judgements

Lars Hoffmann[1], Dr. Katrin Böhme[1]

[1]Humboldt Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen (IQB)

Die Teilstudie ist als Zeitschriftenbeitrag veröffentlicht und wie folgt zugänglich:

Hoffmann, L. & Böhme, K. (2014). Wie gut können Grundschullehrkräfte die Schwierigkeit von Deutsch- und Mathematikaufgaben beurteilen? *Psychologie in Erziehung und Unterricht, 1*, 42–55. doi: 10.2378/peu2014.art05d ©2014 by Reinhardt

Zusammenfassung: In diesem Beitrag wird der Frage nachgegangen, inwieweit Lehrkräfte die Schwierigkeit von Aufgaben akkurat einschätzen. Die Ergebnisse basieren auf einer Stichprobe von 239 Deutsch- und 133 Mathematiklehrkräften aus 212 Grundschulen in ganz Deutschland. Diese sollten einschätzen, wie schwer bestimmte Aufgaben der Fächer Deutsch und Mathematik für die Kinder in ihren Klassen sind. Für die Rangkomponente der Schwierigkeitsurteile wurden im Mittel Koeffizienten in moderater Höhe identifiziert. Die Ausprägung der Differenzierungskomponente lässt darauf schließen, dass die Schwierigkeitseinschätzungen der Lehrkräfte durch eine Tendenz zur Mitte gekennzeichnet waren. Zwischen der Neigung zur Über- bzw. Unterschätzung der Schwierigkeit einerseits und der Erfahrung der Lehrkräfte (Dauer der Lehrtätigkeit, Kontaktdauer mit der jeweiligen Klasse) andererseits fanden sich nur geringe Zusammenhänge. Wurden ähnliche Aufgaben zu einem frühen Zeitpunkt in der Primarstufe behandelt, zeigte sich eine Tendenz zur Unterschätzung; eine Thematisierung ähnlicher Aufgaben zu einem späteren Zeitpunkt erhöhte hingegen die Wahrscheinlichkeit zur Überschätzung der Schwierigkeit.

Schlüsselbegriffe: Diagnostische Kompetenz, Diagnostische Fähigkeiten, Urteilsgenauigkeit, aufgabenbezogene Lehrerurteile

Summary: This study investigates to what extent teachers are able to judge the difficulty of tasks accurately. A sample of 239 German language teachers and 133 mathematics teachers from 212 elementary schools throughout Germany were asked to judge the difficulty of German language and mathematics tasks for the pupils of their classes. Our results on average show moderately high coefficients for the rank component of the assessment of difficulty. The characteristics of the differentiation component suggest that the difficulty judgments of the teachers are characterized by an error of central tendency. We found only weak relationships between the tendencies to over- or underestimate the task difficulty and the experiences of the teachers. If similar tasks were addressed early in elementary school, results showed a tendency to underestimate. However, a thematization of similar tasks at a later date increased the probability to overestimate task difficulty.

Keywords: Diagnostic competence, diagnostic skills, teacher judgment accuracy, teacher judgments of tasks

Einleitung

Ein zentrales Handlungsfeld gegenwärtiger Reformen zur Qualitätssicherung im Bildungswesen ist die Optimierung der Lehreraus-, -fort- und -weiterbildung. Ein bedeutsamer Aspekt ist dabei die diagnostische Kompetenz von Lehrkräften, die als Basiskompetenz für guten Unterricht (Weinert, 2000) bzw. als wichtiger Bestandteil von Lehrerexpertise (Baumert & Kunter, 2006) gilt. Entsprechend wird ihre Stärkung auch in den Standards für die Lehrerbildung der Kultusministerkonferenz (KMK) betont (KMK, 2004). Einhergehend mit der hohen Relevanz, die der diagnostischen Kompetenz somit beigemessen wird, findet sich in jüngerer Zeit ein deutliches Interesse der Bildungsforschung an diesem Thema. Diagnostische Kompetenz wird dabei zumeist als die Fähigkeit definiert, „Personen oder Personengruppen (z. B. Schüler oder Schulklassen) zutreffend zu beurteilen bzw. genaue diagnostische Urteile abzugeben“ (Helmke, 2010, S.121). Ein erweitertes Verständnis des Konzepts diagnostischer Kompetenz umfasst außerdem die Fähigkeit, Aufgabenmerkmale (d. h. insbesondere deren Schwierigkeit) korrekt einzuschätzen (Lorenz & Artelt, 2009). Dieser Fähigkeit, die auch Gegenstand der vorliegenden Forschungsarbeit ist, kommt eine hohe unterrichtspraktische Bedeutung zu: Ein zentrales Element zielorientierten, adaptiven Unterrichtens besteht darin, Aufgaben hinsichtlich ihrer Schwierigkeit mit dem Leistungsniveau der Schülerinnen und Schüler abzustimmen (Anders et al., 2010). Die Aufgaben sollten nicht zu leicht sein, um eine Unterforderung der Schülerschaft zu vermeiden (McElvany et al., 2009); sie sollten vielmehr am Vorwissen der Kinder anknüpfen und somit zu einem kognitiv aktivierenden Unterricht beitragen (Brunner et al., 2011). Sehr schwere Aufgaben sollten nur dann im Unterricht verwendet werden, wenn die für ihre Lösung notwendigen Kenntnisse und Fähigkeiten in der „Zone der nächsten Entwicklung“ der Kinder liegen und ein geeignetes „Gerüst“ (scaffolding) an Hilfestellungen bereitgestellt wird (Vygotskij, 1986).

Die Auswahl von Aufgaben im optimalen Schwierigkeitsbereich erfordert, dass Lehrkräfte in der Lage sind, das Leistungsniveau ihrer Schülerinnen und Schüler hinreichend genau einzuschätzen. Dabei sollte die Aufgabenauswahl sowohl im Lernprozess als auch in der Phase der Leistungsüberprüfung idealiter an den *individuellen* Voraussetzungen bzw. Kompetenzständen der einzelnen Schülerinnen und Schüler orientiert sein (z. B. Helmke, 2010). Da aber Lernsituationen regelmäßig im Klassen-

verband gestaltet werden, und auch die Leistungsüberprüfung, insbesondere in schriftlicher Form, für alle Kinder einer Klasse im Normalfall dieselben Aufgaben vorsieht und somit nicht zwischen unterschiedlichen Kompetenzniveaus der Schülerschaft differenziert, ist außerdem bedeutsam, dass Lehrkräfte auch das *mittlere Leistungsniveau* ihrer Klasse präzise beurteilen können.

Urteilsgenauigkeit von Schwierigkeitseinschätzungen

Um die Genauigkeit von Lehrerurteilen zu bestimmen, bedarf es eines geeigneten Vergleichskriteriums, zu dem sie in Beziehung gesetzt werden können. Bei Urteilen zu Schülermerkmalen werden hierfür meist die Schülerleistungen in einem standardisierten Test herangezogen. Im Fall von Urteilen zur Schwierigkeit von Aufgaben dient ihre empirisch ermittelte Schwierigkeit als Vergleichskriterium. Diese wird oft als relative Lösungshäufigkeit angegeben und berechnet sich aus dem Anteil an Schülerinnen und Schülern einer größeren Schülergruppe (z. B. der eigenen Klasse), die eine Aufgabe korrekt lösen konnte (vgl. McElvany et al., 2009).

Beim Vergleich zwischen Lehrerurteil und Kriterium können unterschiedliche Facetten der Urteilsgenauigkeit fokussiert werden. Diese Facetten finden ihre Entsprechung in der vor allem in der deutschsprachigen Literatur etablierten Unterscheidung zwischen Rang-, Niveau- und Differenzierungskomponente (Schrader & Helmke, 1987): Die *Rangkomponente* beschreibt die Fähigkeit von Lehrkräften, Fähigkeits- bzw. Leistungsabstufungen zwischen Schülerinnen und Schülern akkurat einzuschätzen. Sie wird über die Korrelation zwischen Lehrerurteil und Kriterium ermittelt. Die *Niveaukomponente* bezieht sich auf die absolute Einschätzung der Ausprägung einer Schülerfähigkeit oder der Schülerleistung bei einer Aufgabe. Zu ihrer Bestimmung wird entweder der *Urteilsfehler* (absoluter Betrag der Abweichung zwischen Lehrerurteil und Kriterium) oder die *Urteilstendenz* (Grad der Über- oder Unterschätzung des Kriteriums) ermittelt. Die *Differenzierungskomponente* fokussiert auf den Vergleich zwischen der Streuung von Fähigkeitsausprägungen bzw. Schülerleistungen und den entsprechenden Angaben der Lehrkräfte.

Während die Genauigkeit von Lehrerurteilen zu Schülermerkmalen bereits verhältnismäßig häufig untersucht wurde (vgl. Südkamp, Kaiser & Möller, 2012), waren Lehrerurteile zur Schwierigkeit von Aufgaben für die Schülerinnen und Schüler der eigenen Klasse bislang selten Gegenstand von Forschungsarbeiten. In den wenigen

Studien hierzu, denen, im Unterschied zur vorliegenden Untersuchung zur aufgabenbezogenen Urteilsgenauigkeit von Grundschullehrkräften, überwiegend Daten aus der Sekundarstufe I zugrunde lagen (Ausnahmen: z. B. Lorenz, 2011), wurden meist Schwierigkeitsurteile zu strukturell ähnlichen Aufgaben (z. B. Aufgaben zu Texten mit instruktionalen Bildern, McElvany et al., 2009) betrachtet, oder es wurden Aufgaben in den Blick genommen, die sich auf die gleiche Kompetenz bezogen (z. B. Mathematikaufgaben: Anders et al., 2010; Hosenfeld, Helmke & Schrader, 2002; Lehmann et al., 2000). Dabei wurden für die Rangkomponente der Genauigkeit von Schwierigkeitseinschätzungen mehrheitlich geringe bis moderate Korrelationskoeffizienten im Bereich zwischen $.35 < r < .55$ berichtet. In Untersuchungen, in denen zusätzlich die Niveaukomponente bestimmt wurde, fand sich mehrheitlich ein Trend zur Überschätzung der Leistung – und somit eine Tendenz zur Unterschätzung der Aufgabenschwierigkeit (z. B. Anders et al., 2010; Hosenfeld et al., 2002; Lehmann et al., 2000). Vereinzelt wurden jedoch auch gegenteilige Befunde, d. h. eine Tendenz zur Überschätzung der Aufgabenschwierigkeit, berichtet (z. B. McElvany et al., 2009; Lintorf et al., 2011). In den seltenen Fällen, in denen die Differenzierungskomponente bestimmt wurde, neigten die Lehrkräfte zu einer Unterschätzung der Streuung der Aufgabenschwierigkeit (Lintorf et al., 2011).

Kovariaten der Genauigkeiten von Schwierigkeitseinschätzungen

In Studien zur diagnostischen Kompetenz finden sich konsistent große interindividuelle Unterschiede in der Urteilsgenauigkeit (z. B. Hosenfeld et al., 2002). Demnach scheint es Lehrkräften zum Beispiel unterschiedlich gut zu gelingen, zu beurteilen, wie schwierig bestimmte Aufgaben für die Schülerinnen und Schüler der eigenen Klasse sind. Zur Erklärung dieser Unterschiede wurden Merkmale der Lehrkräfte bzw. der Aufgaben in den Blick genommen, für die aus theoretischen Modellen zum Prozess der Leistungsbeurteilung (vgl. Schrader & Helmke, 2001) ein Einfluss auf die Urteilsgenauigkeit abgeleitet werden kann. In Untersuchungen zur Bedeutung von *Aufgabenmerkmalen* wurden u.a. Zusammenhänge zwischen der Urteilsgenauigkeit und der psychometrisch bestimmten Schwierigkeit der Aufgaben gefunden: So waren beispielsweise Lehrkräfte, welche die Schwierigkeit psychometrisch schwerer Aufgaben akkurat beurteilen konnten, weniger gut darin, die Schwierigkeit psychometrisch leichter Aufgaben einzuschätzen (Lintorf et al., 2011).

Ein *Lehrermerkmal,* für das Zusammenhänge mit der Genauigkeit von Schwierigkeitseinschätzungen untersucht wurden, ist die Berufserfahrung der Lehrkräfte. Unter Bezugnahme auf Erkenntnisse aus der Expertiseforschung wurde vermutet, dass Lehrkräfte mit größerer Berufserfahrung genauere diagnostische Urteile fällen können (z. B. Coladarci, 1986). Allerdings fanden sich meist nur geringe und nicht in jedem Fall statistisch signifikante Korrelationen zwischen der Dauer der Tätigkeit im Lehrerberuf und der Genauigkeit von Lehrerurteilen. Auch für die Dauer der Lehrtätigkeit in einer bestimmten Klasse (Kontaktdauer) und für die Ausprägung des fachdidaktischen Wissens der Lehrkräfte wurden nur geringe Zusammenhänge ermittelt (vgl. McElvany et al., 2009; Anders et al., 2010).

Weitere Ansätze zur Erklärung interindividueller Unterschiede in der Urteilsgenauigkeit lassen sich aus Systematiken ableiten, die im Rahmen der Modellierung von Urteilsprozessen in der Psychologischen Diagnostik Verwendung finden: Im Linsenmodell von Brunswick (1956) wird u.a. postuliert, dass ein distales, nicht beobachtbares Merkmal, auf das ein Urteil zielt, durch proximale, beobachtbare Merkmale abgebildet wird. Wenn etwa Lehrkräfte prognostizieren sollen, wie schwierig eine Aufgabe für die Schülerinnen und Schüler ihrer Klasse ist (distales Merkmal), dann werden sie ihr Urteil auf proximalen Merkmalen gründen, also zum Beispiel auf Informationen zu Schülerleistungen bei ähnlichen Aufgaben, auf bestimmte Charakteristika der einzuschätzenden Aufgabe oder auf Angaben aus Lehrplänen und Curricula. Interindividuelle Unterschiede in der Urteilsgüte können dabei zum einen aus Unterschieden in der Verarbeitung und Kombination dieser Informationen durch die Lehrkräfte, zum anderen aus Diskrepanzen in der Validität und in der bloßen Verfügbarkeit der proximalen Merkmale resultieren. Wurden zum Beispiel im Unterricht häufig Aufgaben behandelt, für deren Bearbeitung die gleichen Teilkompetenzen erforderlich sind wie für die Bewältigung der zu beurteilenden Aufgaben, sollten genauere Schwierigkeitsurteile möglich sein, als wenn diese Teilkompetenzen nicht oder nur selten thematisiert wurden. Eine hohe Urteilsgüte sollte außerdem dann erreicht werden können, wenn die Thematisierung relevanter Teilkompetenzen im Unterricht zeitnah zur Schwierigkeitseinschätzung erfolgte, sodass Informationen zur Leistung der eigenen Schülerinnen und Schüler bei der Bearbeitung ähnlicher Aufgaben verhältnismäßig aktuell und kognitiv leicht verfügbar sind.

Fragestellungen und Hypothesen

Die vorliegende Studie zielt darauf, den gegenwärtig noch unbefriedigenden Erkenntnisstand zur Genauigkeit von Lehrerurteilen zur Aufgabenschwierigkeit um weitere Forschungsergebnisse zu bereichern. Hierfür werden Schwierigkeitsurteile zu Aufgaben aus unterschiedlichen Kompetenzbereichen der Fächer Deutsch (Lesen; Orthografie; Sprache und Sprachgebrauch untersuchen) und Mathematik (Daten; Häufigkeit und Wahrscheinlichkeit; Raum und Form; Zahlen und Operationen) betrachtet, die an einer großen, repräsentativen Lehrerstichprobe erhoben wurden. Darüber hinaus soll untersucht werden, welche Faktoren mit der Genauigkeit von Schwierigkeitseinschätzungen zusammenhängen. Neben der psychometrischen Schwierigkeit der Aufgaben werden dabei insbesondere Lehrermerkmale und Angaben zur Thematisierung relevanter Teilkompetenzen im Unterricht in den Blick genommen. Dabei wird folgenden Forschungsfragen nachgegangen:

Wie genau können Lehrkräfte beurteilen, wie schwierig einzelne Aufgaben für die Schülerinnen und Schüler ihrer Klasse sind?

Im Einklang mit den oben skizzierten Ergebnissen anderer Untersuchungen, werden für die Rangkomponente geringe bis moderate Korrelationen zwischen den Schwierigkeitseinschätzungen der Lehrkräfte einerseits und der psychometrischen Schwierigkeit der Aufgaben andererseits erwartet. Gleichzeitig werden große interindividuelle Unterschiede in der Höhe der berechneten Korrelationen angenommen. Mit Blick auf die Niveaukomponente wird eine Tendenz zur Unterschätzung der psychometrischen Aufgabenschwierigkeit vermutet. Für die Differenzierungskomponente wird erwartet, dass die Lehrkräfte zur Unterschätzung der Varianz der psychometrischen Aufgabenschwierigkeit neigen.

Welche Faktoren stehen im Zusammenhang mit der Über- bzw. Unterschätzung der Schwierigkeit der Aufgaben?

Korrespondierend mit den Befunden anderer Studien, wird ein Zusammenhang zwischen der psychometrischen Schwierigkeit der Aufgaben und der Über- oder Unterschätzung der Aufgabenschwierigkeit angenommen. Dabei ist angesichts der Vorhersage zur Ausprägung der Differenzierungskomponente zu vermuten, dass die Schwierigkeit psychometrisch leichter Aufgaben – im Sinne einer Tendenz zur Mitte – eher über- und

die Schwierigkeit psychometrisch schwerer Aufgaben eher unterschätzt wird. Es wird ferner erwartet, dass die Dauer der beruflichen Tätigkeit und die Kontaktdauer mit der Klasse in einem geringen, negativen Zusammenhang mit der Urteilstendenz stehen. Lehrkräfte mit einer längeren Berufserfahrung und längerem Kontakt zu ihrer Klasse sollten demnach in geringerem Umfang zu Verschätzungen der Aufgabenschwierigkeit neigen. Zudem werden Zusammenhänge zwischen der Urteilstendenz einerseits und Angaben zur Thematisierung (Häufigkeit, Zeitpunkt) relevanter Teilkompetenzen im Unterricht der jeweiligen Schulklasse andererseits vermutet. Dabei wird erwartet, dass eine häufige Thematisierung aufgabenrelevanter Inhalte im Unterricht zu einer geringeren Verschätzung der Aufgabenschwierigkeit führt. Ebenso sollte eine vor kurzer Zeit im Unterricht erfolgte Thematisierung relevanter Teilkompetenzen zu präziseren Schwierigkeitseinschätzung führen.

Methode

Stichprobe

Die vorliegende Untersuchung beruht auf Sekundäranalysen von Daten der Normierungsstudie zu den Bildungsstandards in den Fächern Deutsch und Mathematik für den Primarbereich (KMK, 2005a, 2005b), die im Frühjahr 2007 vom Institut zur Qualitätsentwicklung im Bildungswesen (IQB) in den Ländern der Bundesrepublik Deutschland durchgeführt wurde (vgl. Böhme et al., 2012). Primäres Anliegen dieser Normierung, bei der Schülerinnen und Schüler der dritten und vierten Klassenstufe getestet wurden, war die Entwicklung von Kompetenzstufenmodellen und die Konstruktion von länderübergreifend gültigen Kompetenzskalen. Neben den bildungsstandardbasierten Testinstrumenten, die von den Schülerinnen und Schülern bearbeitet wurden, umfasste die Studie auch den zeitgleichen Einsatz von Lehrer-fragebögen. In diesen wurden die Deutsch- und Mathematiklehrkräfte der an der Studie teilnehmenden Schülerinnen und Schüler u.a. gebeten, die Schwierigkeit einiger Testaufgaben zu beurteilen, die in den Testheften der Kinder zur Erhebung der Kompetenzstände in den Fächern Deutsch und Mathematik eingesetzt wurden.

Die hier präsentierten Befunde basieren auf den Angaben einer Teilstichprobe von 239 Deutsch- und 133 Mathematiklehrkräften aus 212 Schulen in ganz Deutschland. Hierbei werden nur die Urteile derjenigen Lehrkräfte berücksichtigt, für deren Schulklassen (aufgrund des Testdesigns) eine Lösungswahrscheinlichkeit der jeweiligen

Aufgaben ermittelt werden konnte. Ein Vergleich der Stichprobenmerkmale der hier untersuchten Teilstichprobe mit der Gesamtstichprobe der Normierungsstudie zeigt allerdings, dass die Repräsentativität durch die Selektion nicht eingeschränkt wird (vgl. Tabelle 4.6 im Anhang).

Die Deutschlehrkräfte gaben im Mittel an, ihr Fach seit 18.51 Jahren (*SD*=11.64) zu unterrichten. Die Mathematiklehrkräfte berichteten eine durchschnittliche Lehrerfahrung von 19.03 Jahren (*SD*=11.98). Sowohl die Deutsch- als auch die Mathematiklehrkräfte waren mehrheitlich weiblich (87% bzw. 85%).

Instrumente

Für die Ermittlung der Kompetenzstände der Schülerinnen und Schüler wurde in der Normierungserhebung eine Vielzahl von Aufgaben zu allen relevanten Kompetenzbereichen aus dem geschützten Aufgabenpool des IQB eingesetzt. Da diese Aufgaben für Trendaussagen in späteren Ländervergleichsstudien (vgl. Böhme et al., 2012) benötigt werden, ist ein vertraulicher Umgang mit den Testinstrumenten unerlässlich. Aus diesem Grund konnten für die Lehrkräftebefragung nur wenige Aufgaben verwendet werden. Die Auswahl dieser wenigen Aufgaben erfolgte vornehmlich auf Grundlage inhaltlicher Erwägungen, die aus den Erfordernissen der Normierungsstudie des Jahres 2007 abgeleitet wurden, und nicht vorrangig mit Blick auf die in diesem Artikel untersuchten Fragestellungen zur Genauigkeit von Schwierigkeitseinschätzungen: Hierdurch unterscheiden sich die ausgewählten Aufgaben z.T. erheblich hinsichtlich der bei ihrer Bearbeitung zu bewältigenden Anforderungen; auch zeigen die Ergebnisse psychometrischer Analysen, dass die ausgewählten Aufgaben nicht gleichmäßig über das gesamte Schwierigkeitsspektrum streuen, sondern mehrheitlich eher leicht waren (vgl. Tabelle 4.1). Insgesamt wurden im Lehrkräftefragebogen für das Fach Deutsch Schwierigkeitsurteile zu fünf Aufgaben aus den Kompetenzbereichen *Lesen (L)*, *Orthografie (O)* sowie *Sprache und Sprachgebrauch untersuchen (SG)* erhoben, für das Fach Mathematik handelte es sich um vier Aufgaben zu den Kompetenzbereichen *Daten, Häufigkeit und Wahrscheinlichkeit (DHW)*, *Raum und Form (RF)* sowie *Zahlen und Operationen (ZO)*. Alle neun Aufgaben sind im Anhang dieses Berichts abgebildet.

Zur Beurteilung der Aufgabenschwierigkeit stand im Lehrerfragebogen eine sechsstufige Ratingskala zur Verfügung (1=„sehr leicht“ bis 6=„sehr schwer“). Die Lehrkräfte wurden gebeten, ihre Einschätzung auf die Kompetenzstände der von ihnen

unterrichteten und in der Normierungsstudie getesteten Klasse zu beziehen. Der Zeitpunkt der Thematisierung relevanter Teilkompetenzen wurde wie folgt erfragt: „Bitte kreuzen Sie an, ob der Stoff behandelt wurde oder behandelt wird, der die Grundlage dafür bildet, dass folgende oder ähnliche Aufgaben korrekt gelöst werden können. Der Stoff wird/wurde behandelt in der Klassenstufe…“ (1=„2 oder früher“ bis 3=„4“). Zur Erfassung der Kontaktdauer mit den Schülerinnen und Schülern wurden die Lehrkräfte gefragt, seit wann sie Deutsch bzw. Mathematik in der betreffenden Klasse unterrichten (1=„seit der 1. Klasse“ bis 4=„seit der 4. Klasse“). Der Lehrerfragebogen enthielt außerdem eine Liste mit kompetenzbezogenen Tätigkeiten. Für jede dieser Tätigkeiten sollte angegeben werden, wie häufig sie im laufenden Schuljahr Gegenstand des Unterrichts waren (1=„nie“ bis 6=„fast jede Stunde“). Im Fach Deutsch enthielt diese Liste u.a. die Tätigkeit „Sprachliche Begriffe und Strukturen anwenden“, die der Aufgabe $\text{Deutsch}_{\text{SG1}}$ zugeordnet werden kann. Im Fach Mathematik umfasste die Liste u.a. die Tätigkeit „Mathematische Strukturen in Alltagskontexten erkennen“, auf die sich Aufgabe $\text{Mathe}_{\text{DHW2}}$ bezieht. Zu den weiteren sieben Aufgaben liegen keine Informationen über die Häufigkeit der Thematisierung korrespondierender Tätigkeiten im Unterricht vor.

Statistische Analysen

Um die Vorzüge eines Multi-Matrix-Designs nutzen zu können, bearbeiteten die an der Normierungsstudie teilnehmenden Kinder unterschiedliche Testhefte. Diese Testheftrotation erfolgte nicht nur zwischen, sondern teilweise auch innerhalb von Klassen (z. B. Winkelmann & Böhme, 2009). Entsprechend der Designvorgaben beinhalteten nur einige der eingesetzten Testhefte die neun Aufgaben, deren Schwierigkeit von den Lehrkräften eingeschätzt werden sollte. Daher war es nicht möglich, die Genauigkeit der Lehrerurteile anhand der relativen Lösungshäufigkeiten der Aufgaben in den Klassen zu überprüfen. Als Vergleichskriterium dienten stattdessen Lösungswahrscheinlichkeiten, für deren Berechnung zunächst eine Raschskalierung aller in den Testheften enthaltenen Aufgaben mit Hilfe der Software ACER ConQuest 2.0 erfolgte (Wu, Adams, Wilson & Haldane, 2007). Anschließend wurden, unter Verwendung der geschätzten Itemparameter (σ_i), der Personenparameter (Weighted-Maximum-Likelihood-Estimates/WLE; Warm, 1989) und der Modellgleichung des Rasch-Modells, für jede Aufgabe (i) und für alle Schülerinnen und Schüler (j) separate Lösungswahrscheinlichkeiten p_{ij} bestimmt. Diese wurden für jede Schulklasse (k) zu einer

klassenbezogenen Lösungswahrscheinlichkeit p_{ik} gemittelt. Um die Lösungswahrscheinlichkeiten sinnvoll mit den sechsstufigen Schwierigkeitsurteilen der Lehrkräfte in Beziehung setzen zu können, erfolgte eine Segmentierung von p_{ik} in sechs gleich große Abschnitte (C). Aus der Kategorisierung der einzelnen p_{ik} in diese Abschnitte resultierte die sechsstufige (und zusätzlich umgepolte) Skala C_{ik}.

Den Ausgangspunkt für die Berechnung der Rangkomponente bildete ein separat für jede Lehrkraft durchgeführter Vergleich der Rangfolge der Schwierigkeitseinschätzungen zu den Deutsch- bzw. Mathematikaufgaben mit der Rangfolge der auf C_{ik} abgebildeten Lösungswahrscheinlichkeiten. Die dabei gefundenen Korrelationskoeffizienten wurden mithilfe einer Fisher's Z-Transformation über alle Lehrkräfte gemittelt. Zur deskriptiven Abbildung der Verteilung der Koeffizienten wurden Quartile berechnet. Die Bestimmung der Urteilstendenz (Niveaukomponente) erfolgte auf Grundlage eines Vergleichs der Lehrerurteile mit den kategorisierten Lösungswahrscheinlichkeiten (C_{ik}). Da aus den verbalen Verankerungen der Ratingskala nicht direkt ablesbar war, welche Lösungswahrscheinlichkeiten die Lehrkräfte mit den einzelnen Skalenstufen verbanden, wurde, für die Abbildung der Ratings auf C_{ik}, ein Toleranzbereich definiert: Eine Über- bzw. Unterschätzung der Schwierigkeit wurde nur dann angenommen, wenn die Ratings der Lehrkräfte um mehr als eine Kategorie von C_{ik} abwichen. Die Berechnung der Differenzkomponente erfolgte über den Quotienten aus der gepoolten Varianz der Schwierigkeitseinschätzungen und der Varianz der psychometrischen Aufgabenschwierigkeiten.

Die Untersuchung von Zusammenhängen zwischen der Urteilstendenz mit Aufgaben- und Lehrermerkmalen (s.o.) erfolgte mithilfe multinomial-logistischer Regressionsanalysen, die eine Berechnung von Regressionsmodellen mit polytomen Kriteriumsvariablen erlauben (hier: „Überschätzung", „Unterschätzung" und „akkurate Einschätzung der Aufgabenschwierigkeit" als Referenzkategorie). Für die Fächer Deutsch und Mathematik wurden separate Regressionsanalysen durchgeführt, deren Datenbasis jeweils Lehrerurteile zur Schwierigkeit aller Deutsch- bzw. Mathematikaufgaben umfasste. Da diese Datenbasis eine hierarchische Struktur aufweist (Schwierigkeitsurteile geschachtelt in Lehrkräften), wurden die Regressionsanalysen mithilfe der Software Mplus 6 (Muthén & Muthén, 1998-2011) als Mehrebenenmodelle spezifiziert (Random-Intercept-Modelle). Dabei beinhaltete Ebene 1 die Prädiktoren „psychometrischer Itemparameter" und „Zeitpunkt der Thematisierung relevanter Teilkompetenzen". Ebene 2 umfasste die

Prädiktoren „Dauer Tätigkeit als Deutsch- bzw. Mathematiklehrkraft“ und „Kontaktdauer“. Für die Aufgaben Deutsch_{SG1} und Mathe_{DHW2} wurden außerdem separate Regressionsanalysen (d. h. einzeln für jede Aufgabe und daher ohne Modellierung der Mehrebenenstruktur) durchgeführt, in denen als zusätzlicher Prädiktor die Häufigkeit der Thematisierung der jeweils relevanten Teilkompetenzen im Unterricht berücksichtig wurde.

Ergebnisse

Wie genau können Lehrkräfte beurteilen, wie schwierig einzelne Aufgaben für die Schülerinnen und Schüler ihrer Klasse sind?

Für die Rangkomponente der Genauigkeit der Lehrereinschätzungen zu den Aufgaben für das Fach Deutsch wurde eine mittlere Korrelation von r=.42 ermittelt. Beim Blick auf die Quartile (Q) der für die Lehrkräfte jeweils separat berechneten Koeffizienten ergab sich eine erhebliche Streuung der Kennwerte ($Q_{0.25}$=.09, $Q_{0.50}$=.31, $Q_{0.75}$=.69). Für die Genauigkeit der Schwierigkeitseinschätzungen zu den Aufgaben des Fachs Mathematik wurde eine Rangkomponente von r=.51 gefunden. Die berechneten Koeffizienten variierten auch hier stark über die einzelnen Lehrkräfte ($Q_{0.25}$=.18, $Q_{0.50}$=.34, $Q_{0.75}$=.57). Die Niveaukomponente der Genauigkeit der Lehrerurteile wurde separat für jede Aufgabe als prozentuale Anteile der Über- und Unterschätzung der empirischen Aufgabenschwierigkeit für die eigene Klasse bestimmt. Die entsprechenden Ergebnisse sind in Tabelle 4.1 dargestellt, in der sich außerdem Angaben zur jeweiligen Höhe des psychometrischen Itemparameters σ_i finden.

Tabelle 4.1: Prozentuale Anteile der Über- und Unterschätzung der Schwierigkeit der einzelnen Aufgaben durch die Lehrkräfte und Ausprägung der Itemparameter σi auf der Logit-Skala

Aufgabe i	Unterschätzung (in %)	Überschätzung (in %)	σ_i
$Deutsch_{O1}$	17.9	6.0	-1.04
$Deutsch_{O2}$	1.2	41.2	-2.56
$Deutsch_{SG1}$	9.7	6.8	-0.81
$Deutsch_{SG2}$	0.6	33.1	-2.01
$Deutsch_{L1}$	66.9	-	1.37
$Mathe_{DHW1}$	3.1	11.5	-0.39
$Mathe_{DHW2}$	10.8	6.2	-0.49
$Mathe_{RF1}$	-	69.2	-5.16
$Mathe_{ZO1}$	19.5	4.1	0.67

Insgesamt variieren Höhe und Art der Fehleinschätzungen erheblich über die Aufgaben. Für die Aufgaben $Deutsch_{SG1}$, $Mathe_{DHW1}$ und $Mathe_{DHW2}$ wurden relativ wenige Fehlurteile ermittelt. Eine verhältnismäßig hohe Anzahl an Fehleinschätzungen fand sich bei den Aufgaben $Deutsch_{O2}$, $Deutsch_{L1}$ und $Mathe_{RF1}$. Die zur Schätzung der Differenzierungskomponente ermittelten Quotienten (Deutsch: .89; Mathematik: .61) zeigen, dass die Varianz der empirischen Aufgabenschwierigkeiten im Durchschnitt unterschätzt wurde.

Welche Faktoren stehen im Zusammenhang mit der Über- bzw. Unterschätzung der Schwierigkeit der Aufgaben?

Die Ergebnisse der multinomial-logistischen Mehrebenenanalysen sind in Tabelle 4.2 und Tabelle 4.3 dargestellt. Zur besseren Interpretation der Ergebnisse sei darauf hingewiesen, dass aufgrund der dreistufigen Kriteriumsvariablen für jeden Prädiktor simultan zwei Regressionskoeffizienten berechnet wurden (d. h. ein Koeffizient für die Überschätzung und ein Koeffizient für die Unterschätzung).

Tabelle 4.2: Multinomial-logistische Mehrebenenanalyse für die Lehrerurteile zur Schwierigkeit der Deutschaufgaben für die Schülerinnen und Schüler ihrer Klassen

	Unterschätzung		*Überschätzung*	
Ebene 1 (Urteile)	*B*	*SE (B)*	*B*	*SE (B)*
Höhe d. Itemparameters σ_i	1.29**	0.11	-1.45**	0.16
Zeitpunkt Thematisierung im Unterricht	-0.15	0.08	0.15*	0.06
Ebene 2 (Lehrkräfte)				
Kontaktdauer	-0.16	0.11	0.11	0.12
Dauer Lehrertätigkeit (Jahre)	-0.02	0.01	0.01	0.01
$R^2_{Nagelkerke}$=.51				

Anmerkungen: *p≤.05, **p<.01

Tabelle 4.3: Multinomial-logistische Mehrebenenanalyse für die Lehrerurteile zur Schwierigkeit der Mathematikaufgaben für die Schülerinnen und Schüler ihrer Klassen für das Fach Mathematik

	Unterschätzung		*Überschätzung*	
Ebene 1 (Urteile)	*B*	*SE (B)*	*B*	*SE (B)*
Höhe d. Itemparameters σ_i	1.38**	0.34	-0.74**	0.07
Zeitpunkt Thematisierung im Unterricht	-0.98**	0.28	0.89**	0.26
Ebene 2 (Lehrkräfte)				
Kontaktdauer	-0.06	0.19	-0.19	0.13
Dauer Lehrertätigkeit (Jahre)	<0.01	0.02	-0.03*	0.02
$R^2_{Nagelkerke}$=.50				

Anmerkungen: *p≤.05, **p<.01

Für den Prädiktor „Höhe des psychometrischen Itemparameters σ_i“ zeigten die Ergebnisse der Mehrebenenanalysen ein analoges Muster für die Aufgaben beider Fächer. Eine höhere Ausprägung von σ_i ging jeweils mit einer höheren Wahrscheinlichkeit zur Unterschätzung der empirischen Aufgabenschwierigkeit für die eigene Klasse einher

($b_{Deutsch}$=1.29, p<.01; b_{Mathe}=1.38, p<.01). Korrespondierend dazu zeigte sich bei geringerer Ausprägung von σ_i eine höhere Wahrscheinlichkeit zur Überschätzung der Schwierigkeit ($b_{Deutsch}$=-1.45, p<.01; b_{Mathe}=-.74, p<.01). Die Thematisierung relevanter Teilkompetenzen zu einem frühen Zeitpunkt in der Schullaufbahn ging tendenziell mit einer höheren Wahrscheinlichkeit zur Unterschätzung der Schwierigkeit einher ($b_{Deutsch}$=-.15, p=.06; b_{Mathe}=-.98, p<.01). Erfolgte die Thematisierung erst zu einem späteren Zeitpunkt, fand sich eine höhere Wahrscheinlichkeit zur Überschätzung der Schwierigkeit ($b_{Deutsch}$=.15, p<.05; b_{Mathe}=.89, p<.01). Im Mittel korrelierten die Angaben zum Zeitpunkt der Thematisierung nur gering mit den Lehrerurteilen zur Schwierigkeit der Aufgaben für die Schülerinnen und Schüler ihrer Klasse (r=.26). Hinsichtlich der auf Ebene 2 der Regressionsmodelle betrachteten Merkmale wurde lediglich für Aufgaben zum Fach Mathematik und nur für die Dauer der Lehrtätigkeit ein statistisch signifikanter Prädiktor gefunden: Demnach sank die Wahrscheinlichkeit zur Überschätzung der empirischen Aufgabenschwierigkeit mit wachsender Berufserfahrung der Mathematiklehrkräfte (b_{Mathe}=-.03, p<.05).

Die Ergebnisse der multinomial-logistischen Regressionsanalyse zur Aufgabe Deutsch_{SG1} sind in Tabelle 4.4 dargestellt: Lehrkräfte mit größerer beruflicher Erfahrung neigten seltener zur Unterschätzung der Schwierigkeit dieser Aufgabe für die Schülerinnen und Schüler ihrer Klasse (b=-.08, p<.01). Lediglich marginal signifikant zeigte sich, dass Lehrkräfte mit einer geringen Kontaktdauer die Schwierigkeit der Aufgabe mit höherer Wahrscheinlichkeit überschätzten (b=.52, p=.07). Für die Häufigkeit, mit der „Sprachliche Begriffe und Strukturen anwenden“ im laufenden Schuljahr thematisiert wurde, konnten keine statistisch bedeutsamen Regressionskoeffizienten ermittelt werden.

Tabelle 4.4: Multinomial-logistische Regressionsanalyse für die Lehrerurteile zur Schwierigkeit der Aufgabe $Deutsch_{SG1}$ für die Schülerinnen und Schüler ihrer Klassen

	Unterschätzung		*Überschätzung*	
	B	*SE (B)*	*B*	*SE (B)*
Sprachliche Begriffe und Strukturen anwenden	-0.36	0.26	-0.47	0.29
Kontaktdauer	0.06	0.27	0.56	0.31
Dauer Lehrertätigkeit (Jahre)	-0.08**	0.03	-0.02	0.03
$R^2_{Nagelkerke}$=.16				

Anmerkungen: *p≤.05, **p<.01

In Tabelle 4.5 finden sich die Ergebnisse der multinomial-logistischen Regressionsanalyse zur Aufgabe $Mathe_{DHW2}$. Auch hier neigten Lehrkräfte mit einer geringeren Kontaktdauer stärker zur Überschätzung der Schwierigkeit (b=1.24, p<.05). Demgegenüber ging eine häufigere Thematisierung der Teilkompetenz „Mathematische Strukturen in Alltagskontexten erkennen“ im Unterricht mit einer geringeren Neigung zur Überschätzung der Schwierigkeit einher (b=-1.23, p<.05).

Tabelle 4.5: Multinomial-logistische Regressionsanalyse für die Lehrerurteile zur Schwierigkeit der Aufgabe $Mathe_{DHW2}$ für die Schülerinnen und Schüler ihrer Klassen

	Unterschätzung		*Überschätzung*	
	B	*SE (B)*	*B*	*SE (B)*
Mathematische Strukturen in Alltagskontexten erkennen	0.21	0.54	-1.23*	0.02
Kontaktdauer	0.06	0.80	1.24*	0.04
Dauer Lehrertätigkeit (Jahre)	-0.01	0.86	0.04	0.29
$R^2_{Nagelkerke}$=.14				

Anmerkungen: *p≤.05, **p<.01

Diskussion

Zusammenfassung

Im Fokus der vorliegenden Untersuchung stand zunächst die Frage, wie genau Grundschullehrkräfte die Schwierigkeit von Aufgaben der Fächer Deutsch und Mathematik beurteilen können. Für die Rangkomponente der Genauigkeit der Schwierigkeitseinschätzungen wurden, hypothesenkonform und vergleichbar mit den Befunden anderer Studien (z. B. Lehmann et al., 2000), in beiden Fächern Korrelationskoeffizienten in moderater Höhe ermittelt. Außerdem wurde, ebenfalls ähnlich wie in anderen Studien, eine hohe interindividuelle Varianz in der Genauigkeit der Lehrerurteile gefunden (z. B. Hosenfeld et al., 2002).

Die Ergebnisse zur Niveaukomponente verdeutlichen, dass die Schwierigkeit einiger Aufgaben im Durchschnitt relativ genau eingeschätzt wurde, während für andere Aufgaben z.T. recht hohe prozentuelle Anteile von Fehleinschätzungen zu verzeichnen waren. Ferner zeigte sich sowohl deskriptiv als auch in den Ergebnissen der Mehrebenenanalysen der Trend, dass die Schwierigkeit psychometrisch leichter Aufgaben verstärkt überschätzt und die Schwierigkeit psychometrisch schwerer Aufgaben vermehrt unterschätzt wurde. Gleichzeitig bilden die als Maß für die Differenzierungskomponente berechneten Quotienten hypothesenkonform eine tendenzielle Unterschätzung der Varianz der empirischen Schwierigkeiten ab (vgl. Lintorf et al., 2011). Zusammengenommen legen diese Befunde den Schluss nahe, dass die Schwierigkeitseinschätzungen der Lehrkräfte durch eine Vermeidung extremer Urteile (bzw. extremer Antwortkategorien) bzw. durch eine Tendenz zur Mitte gekennzeichnet waren.

In der vorliegenden Studie wurde weiterhin untersucht, ob bestimmte Lehrermerkmale mit der Genauigkeit von Schwierigkeitseinschätzungen zusammenhängen. Diese Lehrermerkmale umfassten die Dauer der Tätigkeit im Lehrerberuf und die Kontaktdauer mit der jeweiligen Klasse. Für beide Merkmale wurden nur vereinzelt statistisch signifikante Effekte auf die Über- oder Unterschätzung der Schwierigkeit einer Aufgabe für die Schülerinnen und Schüler der eigenen Klassen gefunden. Dieses Ergebnis erscheint zwar kontraintuitiv – für Lehrkräfte mit einer größeren Erfahrung ist eine höhere Expertise zu vermuten – stimmt jedoch mit den Befunden anderer Studien überein (z. B. McElvany et al., 2009).

Die Arbeit widmete sich ferner der Untersuchung von Zusammenhängen zwischen der Genauigkeit von Schwierigkeitseinschätzungen einerseits und Merkmalen bzw. Inhalten des Unterrichts andererseits, auf die sich die Schwierigkeitsurteile bezogen. Unter anderem wurde vermutet, dass eine häufigere Thematisierung von Teilkompetenzen, die für die Bewältigung der zu beurteilenden Aufgaben erforderlich sind, Schwierigkeitseinschätzungen erleichtern könnte. Allerdings fand sich lediglich für eine Aufgabe des Fachs Mathematik der Befund, dass Lehrkräfte weniger zur Unterschätzung der Schwierigkeit neigten, wenn die betreffende Teilkompetenz im laufenden Schuljahr häufig Gegenstand ihres Unterrichts war. Hingegen weisen die ermittelten Ergebnisse darauf hin, dass die Schwierigkeitseinschätzungen der Lehrkräfte dadurch beeinflusst wurden, in welcher Klassenstufe Aufgaben mit ähnlichen Anforderungen im Unterricht behandelt worden sind. Allerdings scheint die Urteilsgüte nicht, wie vermutet, mit zunehmender zeitlicher Nähe zwischen der Thematisierung im Unterricht und dem Zeitpunkt der Schwierigkeitseinschätzungen zu wachsen. Zwar ging eine Thematisierung ähnlicher Aufgaben zu einem frühen Zeitpunkt in der Primarstufe mit einer stärkeren Tendenz zur Unterschätzung der Schwierigkeit einher, gleichzeitig zeigte sich aber, dass eine Behandlung ähnlicher Aufgaben zu einem späteren Zeitpunkt die Wahrscheinlichkeit zur Überschätzung der Schwierigkeit erhöhte.

Stärken und Grenzen der vorliegenden Untersuchung

Die dargestellten Ergebnisse basieren auf Urteilen einer großen, repräsentativen Stichprobe von Grundschullehrkräften zur Schwierigkeit von Aufgaben aus verschiedenen Fächern und Kompetenzbereichen. Im Unterschied zu vielen anderen Studien wurden dabei sowohl die Rang- und Niveaukomponente als auch die Differenzierungskomponente der Urteilsgenauigkeit betrachtet. Eine weitere Besonderheit der vorliegenden Arbeit besteht in der Untersuchung von Zusammenhängen zwischen der Genauigkeit von Schwierigkeitseinschätzungen und bestimmten Merkmalen bzw. Inhalten des Unterrichts. Ungeachtet der genannten Stärken, weist die vorliegende Arbeit auch einige Beschränkungen auf, die in weiterführenden Untersuchungen berücksichtigt werden sollten. Beispielsweise ist anzumerken, dass die hier gewählte Form der Schwierigkeitseinschätzungen optimierbar ist, zum Beispiel durch eine direkte Angabe der Lösungswahrscheinlichkeiten für die Klassen. Außerdem wäre es wünschenswert gewesen, die dargestellten Analysen zu Kovariaten der Genauigkeit von Schwierigkeitseinschätzungen für eine deutlich größere Zahl von Aufgaben

durchzuführen. Die Aufgaben sollten hinsichtlich der bei ihrer Bearbeitung zu bewältigenden Anforderungen nach Möglichkeit homogener als die hier verwendeten sein und eine Fokussierung auf eine überschaubare Menge relevanter Teilkompetenzen gestatten. Eine derartig zusammengesetzte Aufgabenstichprobe wäre Voraussetzung für eine zusätzliche inhaltliche Auseinandersetzung mit den zu beurteilenden Aufgaben. Dabei könnten insbesondere mögliche Zusammenhänge zwischen bestimmten Merkmalen der Aufgaben einerseits und den Schwierigkeitseinschätzungen der Lehrkräfte andererseits fokussiert werden.

Schlussfolgerungen für Forschung und Praxis

Trotz der genannten Beschränkungen ist die vorliegende Arbeit eine wichtige Bereicherung des bislang eher unbefriedigenden Erkenntnisstands zur Genauigkeit von Lehrerurteilen. Gleichwohl verweisen die Ergebnisse auf zusätzlichen Forschungsbedarf, wobei insbesondere Untersuchungen zur Bedeutung weiterer Faktoren für die Genauigkeit der Schwierigkeitseinschätzungen lohnenswert erscheinen. Wie bereits skizziert, erscheinen hierbei insbesondere Analysen zur Rolle bestimmter Aufgabenmerkmale (z. B. schwierigkeitsbestimmende Merkmale, Hartig, 2007) zielführend. Mit Blick auf die Praxis lässt sich aus den dargestellten Ergebnissen – korrespondierend mit den Empfehlungen anderer Studien (vgl. Anders et al., 2010) – ein erhöhter Bedarf an Aus- und Fortbildungsmaßnahmen zur Optimierung der diagnostischen Fähigkeiten von Lehrkräften schlussfolgern. Im Sinne einer kompetenzorientierten, kognitiv anregenden Unterrichtsgestaltung ist es entscheidend, Übungs- und Lernaufgaben in einem angemessenen Schwierigkeitsbereich einzusetzen und auch die Überprüfung der Kompetenzstände der Schülerinnen und Schüler mit Aufgaben adäquater Schwierigkeit durchzuführen. Hierfür müssen Lehrkräfte zunächst in die Lage versetzt werden, nicht nur die Kompetenzstände ihrer Schülerinnen und Schüler zutreffend und differenziert einzuschätzen, sondern ferner möglichst genaue Vorhersagen zur Schwierigkeit von möglichen Aufgaben zu treffen. Einen wirkungsvollen Mechanismus zur Verbesserung der Urteilsgenauigkeit könnten dabei Reflexionsprozesse darstellen. Diese können beispielsweise angeregt werden, indem Lehrkräfte durch entsprechende Rückmeldungen die Möglichkeit erhalten, Vergleiche zwischen den eigenen Einschätzungen und den empirischen Aufgabengabenschwierigkeiten bzw. den tatsächlichen Schülerleistungen vorzunehmen (vgl. Helmke, Hosenfeld & Schrader, 2002).

Literaturverzeichnis

Anders, Y., Kunter, M., Brunner, M., Krauss, S. & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht, 3*, 175–193.

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

Böhme, K., Richter, D., Stanat, P., Pant, H. A. & Köller, O. (2012). Die länderübergreifenden Bildungsstandards in Deutschland. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011* (S. 11–18). Münster: Waxmann.

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

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*, 141–146.

Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze-Velber: Klett/Kallmeyer.

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

Hartig, J. (2007). Skalierung und Kompetenzniveaus. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI Ergebnisse Band 1*. Weinheim: Beltz.

KMK (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12. 2004*. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf [01.03.2012].

KMK (2005a). *Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004*. München: Luchterhand.

KMK (2005b). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004*. München: Luchterhand.

Lehmann, R. H., Peek, R., Gänsfuß, R., Lutkat, S., Mücke, S. & Barth, I. (1999). *QuaSUM. Qualitätsuntersuchungen an Schulen zum Unterricht in Mathematik. Ergebnisse einer repräsentativen Untersuchung im Land Brandenburg* (Reihe Schulforschung in Brandenburg, Heft 1). Potsdam: Ministerium für Bildung, Jugend und Sport im Land Brandenburg.

Lintorf, K., McElvany, N., Rjosk, C., Schroeder, S., Baumert, J., Schnotz, W., Horz, H. & Ullrich, M. (2011). Zuverlässigkeit von diagnostischen Lehrerurteilen – Reliabilität verschiedener Urteilsmaße bei der Einschätzung von Aufgabenschwierigkeiten. *Unterrichtswissenschaft, 39(2)*, 102–120.

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(3–4)*, 211–222.

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., Horz, H. & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern. *Zeitschrift für Pädagogische Psychologie, 23(3-4)*, 223–235.

Muthén, L. K. & Muthén, B. O. (1998-2011). *Mplus User's Guide* (6. Aufl.). Los Angeles, CA: Muthén & Muthén.

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

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

Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 4(3)*, 743–762.

Vygotskij (Wygotski), L. (1987). Ausgewählte Schriften, Bd. II: Arbeiten zur psychischen Entwicklung der Persönlichkeit. Berlin: Volk und Wissen.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54(3)*, 427–450.

Weinert, F. E. (2000). *Lehren und Lernen für die Zukunft – Ansprüche an das Lernen in der Schule*. Verfügbar unter http://www2.ibw.uni-heidelberg.de/~gerstner/WeinertLehren&Lernen.pdf [01.03.2012]

Winkelmann, H. & Böhme, K. (2009). Anlage und Durchführung der Pilotierung der Bildungsstandards. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den

Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 31–-41). Weinheim: Beltz.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER Conquest Version 2.0*. Mulgrave: ACER Press.

Anhang I

Tabelle 4.6: Charakteristika der Gesamtstichprobe der Kinder, die an der Normierungserhebung teilnahmen, sowie Merkmale der Schülerstichproben, auf denen die Berechnung der klassenbezogenen Lösungswahrscheinlichkeiten p_{ik} basierte (Anhang zu Teilstudie 1)

	Teilstichprobe Mathematik	*Teilstichprobe Kompetenzbereich Lesen*	*Teilstichprobe Kompetenzbereich Sprache und Sprachgebrauch untersuchen*	*Teilstichprobe Kompetenzbereich Orthografie*	*Gesamtstichprobe*
	$N_{Schüler}$ = 3268	$N_{Schüler}$ = 3600	$N_{Schüler}$ = 2385	$N_{Schüler}$ = 1846	$N_{Schüler}$ = 11396
Geschlecht der Kinder (männl., weibl.)	50.4%, 49.6%	51.3%, 48.7%	49.6%, 50.4%	50.8%, 49.2%	50.5%, 49.5%
Deutsch-Note im Halbjahr (*M*, *SD*)	2.69 (0.90)	2.66 (0.88)	2.67 (0.90)	2.67 (0.87)	2.71 (0.90)
Mathematik-Note im Halbjahr (*M*, *SD*)	2.71 (0.97)	2.60 (0.94)	2.62 (0.95)	2.60 (0.94)	2.66 (0.97)
WLE Orthografie	0.22 (1.31)	-0.50 (1.28)	-0.13 (1.41)	-0.52 (1.37)	-0.52 (1.37)
WLE Sprache und Sprachgebrauch untersuchen	0.03 (1.50)	-0.32 (1.55)	-0.16 (1.45)	0.23 (1.39)	-0.16 (1.45)
WLE Lesen	-0.02 (1.37)	-0.28 (1.36)	-0.31 (1.56)	-0.16 (1.21)	-0.28 (1.36)
WLE Mathematik	0.01 (1.17)	0.12 (1.28)	0.17 (1.24)	0.48 (1.04)	0.01 (1.17)

Anhang II

[Den Schülerinnen und Schülern wurde der Text „Alarm, wenn der Kuckuck ruft" (nach Hans Bahmer) vorgelegt. Im Anschluss mussten sie u.a. die nachfolgende Frage beantworten.]

Ist der junge Kuckuck selbst erwachsen, legt er sein Ei in die Nester der Singvogelarten, von denen er selbst aufgezogen wurde.

Schreibe alles auf, was er sich dafür merken musste!

1) _____

2) _____

3) _____

Abbildung 4.1: Aufgabe $Deutsch_{L1}$ (Anhang zu Teilstudie 1)

Instruktion:

Die Sätze in dieser Aufgabe werden dir gleich vollständig vorgelesen.
Lies bitte in deinem Heft mit! In den Sätzen im Heft fehlt immer ein Wort.

Setze die fehlenden Wörter ein!

Lena bekommt zum_____ drei Geschenke.

Lösung: Geburtstag

Abbildung 4.2: Aufgabe $Deutsch_{O1}$ (Anhang zu Teilstudie 1)

Instruktion:

Die Sätze in dieser Aufgabe werden dir gleich vollständig vorgelesen.
Lies bitte in deinem Heft mit! In den Sätzen im Heft fehlt immer ein Wort.

Setze die fehlenden Wörter ein!

Wenn man in die Schule geht,
sollte man schon mit Löffel und _____ essen können.

Lösung: Gabel

Abbildung 4.3: Aufgabe $Deutsch_{O2}$ (Anhang zu Teilstudie 1)

(Den Schülern wurde ein Text präsentiert)

Im Text findest du einige Verben. Suche 4 unterschiedliche Verben heraus und schreibe sie auf!

1) _____

2) _____

3) _____

4) _____

Abbildung 4.4: Aufgabe $Deutsch_{SG1}$ (Anhang zu Teilstudie 1)

Finde ein Wort, das mit dem unterstrichenen Wort ein Reimpaar bildet!

Schildkröten interessieren uns sehr,

sie leben an Land, in Sumpfgebieten oder im _____.

Abbildung 4.5: Aufgabe $Deutsch_{SG2}$ (Anhang zu Teilstudie 1)

Hier siehst du den Fahrplan von Köln mit dem Intercity IC 800 nach Hamburg.

Bahnhof	an	ab
Köln Hbf		10:09
Düsseldorf Hbf	10:30	10:32
Duisburg Hbf	10:44	10:46
Essen Hbf	10:57	10.59
Bochum Hbf	11:07	11:09
Dortmund Hbf	11:20	11:24
Münster (Westf) Hbf	11:53	11:55
Osnabrück Hbf	12:18	12:20
Bremen Hbf	13:13	13:15
Hamburg-Harburg	13:59	14:01
Hamburg Hbf	14:09	

Wie lange braucht der Zug von Köln bis Hamburg Hbf?

Abbildung 4.6: Aufgabe $Mathe_{DHW1}$ (Anhang zu Teilstudie 1)

Die Klasse 3 plant einen Ausflug. Jedes Kind hat sich für ein Ziel entschieden.

	Spielplatz	Freibad	ZOO
Jungen	𝍸	𝍸 ||	|||
Mädchen	||||	||	|||

Wie viele Mädchen möchten nicht mit ins Freibad?

Abbildung 4.7: Aufgabe $\text{Mathe}_{\text{DHW2}}$ (Anhang zu Teilstudie 1)

Spiegele an der Geraden g.

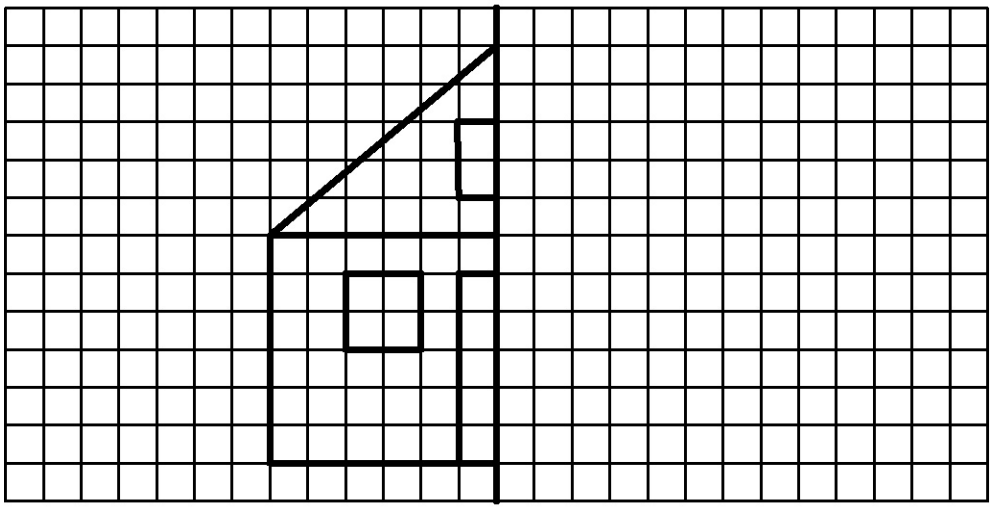

Abbildung 4.8: Aufgabe $\text{Mathe}_{\text{RF1}}$ (Anhang zu Teilstudie 1)

Bestimme die fehlenden Zahlen! Gleiche Form bedeutet gleiche Zahl.

○ + ◇ = 7
○ − ◇ = 5

○ = _____

◇ = _____

Abbildung 4.9: Aufgabe Mathe (Anhang zu Teilstudie 1)

5

Teilstudie 2

5 Teilstudie 2 (Originalarbeit)

Wird sprachlicher Förderbedarf in der Grundschule sicher erkannt?
Zur Klassifikationsgüte von diagnostischen Entscheidungen

Is Need for Language Support in Elementary School Identified Reliably?
A Study of the Classification Accuracy in Diagnostic Decisions

Lars Hoffmann[1], Dr. Katrin Böhme[1]

[1]Humboldt Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen (IQB)

Die Teilstudie ist als Zeitschriftenbeitrag veröffentlicht und wie folgt zugänglich:

Hoffmann, L. & Böhme, K. (2017). Wird sprachlicher Förderbedarf in der Grundschule sicher erkannt? Zur Klassifikationsgüte von diagnostischen Entscheidungen. *Zeitschrift für Pädagogische Psychologie, 31*(2), 137–147. doi: 10.1024/1010-0652/a000203

Zusammenfassung: Additive Sprachfördermaßnahmen basieren auf einer vorgelagerten Entscheidung, ob bei einem Kind ein sprachlicher Förderbedarf besteht. Im vorliegenden Beitrag, in dem die Sprachförderdiagnostik in Grundschulen anhand von Daten der IQB-Ländervergleichsstudie 2011 betrachtet wird, gehen wir der Frage nach, welche diagnostische Güte solche Entscheidungen aufweisen. Darüber hinaus wird untersucht, inwieweit die Entscheidungsgüte mit der Nutzung bestimmter diagnostischer Informationsquellen (u. a. Beobachtungen durch Lehrkräfte, Schulnoten, sprachdiagnostische Verfahren) kovariiert. Unsere Ergebnisse zeigen, dass in den Schulen nur ein geringer Anteil der Kinder mit sprachlichem Förderbedarf korrekt identifiziert wird (geringe Sensitivität). Hingegen werden Kinder ohne Förderbedarf zumeist zuverlässig erkannt (hohe Spezifität). Positive Zusammenhänge mit der Diagnosegüte wurden lediglich für den Einsatz sprachdiagnostischer Verfahren (z. B. Tests) gefunden, wobei eine maßgebliche Verbesserung der Diagnosegüte offenbar nur dann erzielt wird, wenn sprachliche Kompetenzen nicht nur einmalig, sondern mehrfach erfasst werden.

Schlüsselbegriffe: Sprachdiagnostik, Sprachförderung, sprachdiagnostische Verfahren, Sensitivität, Spezifität

Abstract: Additive language support programs are based on a preceding decision on whether a child needs language support. In this article, which considers the language diagnostics in German elementary schools based on data from the German National Assessment Study (IQB-Ländervergleichsstudie) 2011, we determine the diagnostic accuracy of such decisions. Furthermore, we examine if the accuracy of the decisions covaries with the use of specific diagnostic information sources (e. g., observations of teachers, school grades, and diagnostic instruments). Our results show that only a rather small proportion of the children with a need for language support is identified correctly in schools (low sensitivity). By contrast, children without a need for language support are mostly recognized reliably (high specificity). Positive relationships with diagnostic accuracy were only found for the use of diagnostic instruments (e. g., tests), although substantial improvements of diagnostic accuracy seem to occur only if language proficiency is assessed more than once.

Keywords: language diagnostics, language support, diagnostic instruments for language assessment, sensitivity, specificity

Problem- und Zielstellung

Um in vollem Umfang von schulischen Lerngelegenheiten profitieren und erfolgreich lernen zu können, müssen Schülerinnen und Schüler die deutsche Sprache in ihrem unterrichtsspezifischen Gebrauch sicher und kompetent beherrschen (Baumert & Schlümer, 2001). Dabei wird angenommen, dass dieser Gebrauch durch ein formelles sprachliches Register – die sogenannte „Bildungssprache“ – geprägt ist, das sich an schriftsprachlichen Normen orientiert und dementsprechend deutlich von der Alltagssprache abhebt.

Die Anforderungen der Bildungssprache können für Schülerinnen und Schüler maßgebliche Hürden darstellen, die es ihnen in allen Fächern erschweren, dem Unterricht zu folgen, und in der Konsequenz einer erfolgreichen Schullaufbahn im Wege stehen. Hiervon besonders häufig betroffen sind Schülerinnen und Schüler mit Zuwanderungshintergrund, die Deutsch als Zweitsprache erwerben, aber auch Kinder und Jugendliche aus einkommensschwachen, bildungsfernen Familien, die einsprachig Deutsch aufwachsen (Eckhardt, 2008; Gogolin & Lange, 2011; Heppt, Stanat, Dragon, Berendes & Weinert, 2014). Diese Schülerinnen und Schüler benötigen besondere Unterstützung, um schul- und bildungsrelevante sprachliche Fähigkeiten aufzubauen und zu festigen. Eine solche Unterstützung kann, wie etwa beim Konzept der durchgängigen Sprachbildung, in den Regelunterricht integriert sein und zum Beispiel auf Methoden des sprachbewussten oder sprachsensitiven Unterrichts und des Scaffolding basieren (Gogolin et al., 2011). Bei besonders stark ausgeprägten sprachlichen Defiziten können aber auch ergänzende Sprachfördermaßnahmen erforderlich sein, die an Schulen zusätzlich zum Regelunterricht – etwa in der Kleingruppe – durchgeführt werden.

Solche *additiven Sprachfördermaßnahmen* richten sich vor allem an Schülerinnen und Schüler mit sogenannten umgebungsbedingten Sprachauffälligkeiten, die auf ungünstige Spracherwerbsbedingungen (z. B. Anregungsarmut des sprachlichen Inputs, falsche Sprachvorbilder) zurückzuführen sind. Sie sind abzugrenzen von Maßnahmen der Sprachtherapie, die außerhalb des schulischen Kontexts (z. B. von Logopädinnen und Logopäden) durchgeführt werden. Diese dienen ausschließlich der angemessenen Behandlung von Kindern, die sprachliche Auffälligkeiten mit medizinischem Störungswert zeigen – also beispielsweise die diagnostischen Kriterien einer Spezifischen

Sprachentwicklungsstörung (SSES) erfüllen. (Mußmann, 2012; Neumann & Euler, 2013; Rothweiler, 2013).

Additive Sprachfördermaßnahmen setzen Sprachförderentscheidungen voraus, in deren Rahmen für jede Schülerin und jeden Schüler einzeln festgestellt wird, inwiefern bestimmte Sprachauffälligkeiten vorliegen, wie stark diese gegebenenfalls ausgeprägt sind und ob sie die Teilnahme an einer zusätzlichen Sprachfördermaßnahme erfordern oder nicht. Die Güte solcher Förderentscheidungen wird in diesem Beitrag untersucht, wobei der Fokus auf Sprachförderentscheidungen in der Grundschule liegt.

Bestimmung der Güte diagnostischer Entscheidungen

Gegenstand von Sprachförderentscheidungen ist eine Zuordnung von Kindern zu einer der beiden Klassen „sprachlicher Förderbedarf“ ($F+$) und „kein sprachlicher Förderbedarf“ ($F–$). Dieser Zuordnung liegen *Entscheidungsstrategien* zugrunde, die unterschiedlich explizit sein können und sowohl beinhalten, welche diagnostischen Informationen als Prädiktoren für den Sprachförderbedarf herangezogen werden, als auch, auf welche Weise die einzelnen Prädiktoren zu einer Entscheidung kombiniert werden.

Förderentscheidungen auf Basis diagnostischer Informationen haben den Charakter von Vorhersagen, die den sprachlichen Förderbedarf eines Kindes nicht mit hundertprozentiger Sicherheit, sondern nur mit einer bestimmten Wahrscheinlichkeit abbilden. Dementsprechend ist von der vorhergesagten Klassenzuordnung die tatsächliche beziehungsweise „wahre“ Zugehörigkeit zu den beiden Klassen $F+$ und $F–$ zu unterscheiden. Kombiniert man die Vorhersage (V_{F+} bzw. $V_{F–}$) mit der „wahren“ Klassenzugehörigkeit (W_{F+} bzw. $W_{F–}$), ergibt sich ein aus der Signalentdeckungstheorie (Swets, 1964) bekanntes und in Tabelle 5.1 dargestelltes Vier-Felder-Schema. Dieses Schema beinhaltet zum einen zwei Arten richtiger Zuordnungen: die Kategorie der „Wahr-positiven“ (*WP*), in die (im Falle von Sprachförderentscheidungen) alle Kinder fallen, die tatsächlich einen Sprachförderbedarf haben (W_{F+}) und als förderbedürftig diagnostiziert wurden (V_{F+}), sowie die Kategorie der „Wahr-negativen“ (*WN*), die alle Kinder umfasst, die keinen Förderbedarf aufweisen ($W_{F–}$) und als nicht förderbedürftig klassifiziert wurden ($V_{F–}$). Zum anderen werden in dem Schema zwei Arten von Fehlentscheidungen differenziert: die Gruppe der „Falsch-positiven“ (*FP*, auch bekannt als Fehler erster Art oder α-Fehler), zu der all jene Fälle zählen, in denen ein Kind keinen Förderbedarf hat ($W_{F–}$), jedoch fälschlicherweise als förderbedürftig diagnostiziert wurde

(V_{F+}), und die Gruppe der „Falsch-negativen“ (*FN*, auch Fehler zweiter Art oder *β*-Fehler), die alle Fälle umfasst, bei denen Kinder eigentlich einen Förderbedarf aufweisen (W_{F+}), dieser aber nicht erkannt wurde (V_{F-}) (vgl. Noack & Petermann, 1995; Schmidt-Atzert & Amelang, 2012).

Tabelle 5.1: Klassifikation von Förderentscheidungen

		Vorhersage / Förderentscheidung	
		Vorhersage „Förderbedarf“ (V_{F+})	Vorhersage „kein Förderbedarf“ (V_{F-})
Tatsächlicher Förderbedarf	„Förderbedarf“ (W_{F+})	*WP* (Wahr-positiv)	*FN* (Falsch-negativ, Fehler zweiter Art, *β*-Fehler)
	„kein Förderbedarf“ (W_{F-})	*FP* (Falsch-positiv, Fehler erster Art, *α*-Fehler)	*WN* (Wahr-negativ)

Ordnet man die für eine Gruppe von Kindern (z. B. für eine Schulklasse) getroffenen Sprachförderentscheidungen den Kategorien des Vier-Felder-Schemas zu, ergibt sich eine Häufigkeitsverteilung. Auf deren Basis lässt sich die Güte der Entscheidungsstrategie bestimmen, die der Vorhersage des Förderbedarfs zugrunde liegt (vgl. Noack & Petermann, 1995; Schmidt-Atzert & Amelang, 2012). Als entsprechende Gütekennwerte können zum Beispiel die Sensitivität und die Spezifität herangezogen werden. Die Sensitivität bildet den Anteil der Wahr-positiven an allen Kindern mit tatsächlichem Förderbedarf, die Spezifität den Anteil der Wahr-negativen an allen Kindern ohne Förderbedarf ab.

Sensitivität und Spezifität haben den Vorteil, dass sie unabhängig von der Basisrate, also vom Anteil der Kinder mit Förderbedarf (W_{F+}) an allen jeweils betrachteten Kindern, bestimmt werden. Sie können daher herangezogen werden, wenn mehrere Entscheidungs-

strategien hinsichtlich ihrer Klassifikationsgüte miteinander verglichen werden sollen, die Gruppen mit differierenden oder unbekannten Basisraten betreffen (z. B. Vergleiche zwischen Schulen).

Allerdings stehen Sensitivität und Spezifität in einem komplementären Zusammenhang, der durch die Selektionsrate (Anteil der als förderbedürftig diagnostizierten Kinder (V_{F+}) an allen betrachteten Kindern) – oder genauer durch die Festlegung des Cut-Off-Werts – moderiert wird (vgl. Schmidt-Atzert & Amelang, 2012). Dieser Grenzwert spezifiziert den „kritischen" Wert, ab dem – auf Basis bestimmter diagnostischer Informationen – eine positive beziehungsweise negative Förderentscheidung erfolgt. Wählt man diesen kritischen Wert eher liberal, hat dies einerseits eine hohe Anzahl an wahr-positiven Entscheidungen und somit eine hohe Sensitivität zur Folge (d. h. viele Kinder mit tatsächlichem Förderbedarf werden auch als förderbedürftig diagnostiziert). Ein solches Vorgehen führt andererseits aber auch zu einer hohen Anzahl an falsch-positiven Entscheidungen und somit zu einer geringen Spezifität (d. h. es werden viele Kinder als förderbedürftig eingestuft, die eigentlich gar keinen Förderbedarf haben). Analog dazu resultiert die Wahl eines eher konservativen Cut-Offs sowohl in einer geringen Anzahl an wahr-positiven Entscheidungen als auch in einer geringen Anzahl an falsch-positiven Entscheidungen – und somit in einer hohen Spezifität bei gleichzeitig geringer Sensitivität.

Aufgrund des skizzierten Zusammenhangs ist es erforderlich, bei vergleichenden Analysen zur Klassifikationsgüte von Entscheidungsstrategien Sensitivität und Spezifität simultan zu betrachten. Nur so kann ermittelt werden, inwieweit die jeweils untersuchten Entscheidungsstrategien unterschiedlich gut klassifizieren oder ob sie lediglich hinsichtlich der gewählten Cut-Off-Werte differieren.

Diagnose von Sprachförderbedarf in der Grundschule

Wie bereits erläutert, sind additiven Sprachfördermaßnahmen diagnostische Entscheidungen zum Förderbedarf vorangestellt, die nicht zufällig, sondern auf der Basis bereits vorliegender oder speziell erhobener diagnostischer Informationen getroffen werden. Als Informationsquelle hierfür können unter anderem sprachdiagnostische Verfahren dienen. Unter diesem Begriff fassen wir im vorliegenden Beitrag alle Erhebungsinstrumente zusammen, mit deren Hilfe sprachliche Kompetenzen von Kindern in systematischer und geplanter Weise erhoben werden können. Hierunter fallen

Instrumente mit ganz unterschiedlichen Standardisierungsgraden wie Tests, Screenings, Einschätzverfahren sowie Beobachtungsbögen und Beobachtungsverfahren (Becker-Mrotzek et al., 2013). Tatsächlich kommt an deutschen Grundschulen gegenwärtig eine Vielzahl verschiedener sprachdiagnostischer Verfahren zum Einsatz. Entsprechend uneinheitlich erfolgt die Bestimmung des Sprachförderbedarfs, wobei die genutzten Instrumente teilweise erheblich hinsichtlich ihrer psychometrischen Qualität und der jeweils erfassten Konstrukte differieren (Hoffmann, Böhme & Stanat, 2016).

Spracherwerbstheoretisch begründete und sprachdidaktisch elaborierte Konzeptionen zu Anforderungen an sprachdiagnostische Verfahren, die zur Sprachförderdiagnostik genutzt werden, sind bislang ausschließlich für den Einsatz im Elementarbereich formuliert worden (vgl. Bredel, 2005). Einige dieser Anforderungen können jedoch für den Primarbereich adaptiert werden. Wesentlich ist dabei die Empfehlung, dass die sprachlichen Kompetenzen von Schülerinnen und Schülern nicht nur einmalig, sondern mehrfach und zwar in regelmäßigen Abständen erfasst werden sollten. Diese Empfehlung basiert auf der empirisch belegten Annahme, dass Spracherwerbsprozesse nicht immer linear, sondern zum Teil diskontinuierlich und inter-individuell heterogen verlaufen (Bredel, 2005).

Zumindest für einige der auch in Grundschulen eingesetzten sprachdiagnostischen Verfahren (z. B. ETS 4-8, KiSS, SET, SSV) liegen Angaben zu Sensitivität und Spezifität vor. Analog zu dem oben skizzierten Vorgehen wurden diese Kennwerte in empirischen Studien bestimmt, in denen jeweils erfasst wurde, wie ein Verfahren Kinder klassifiziert, die entweder sprachlich regelgerecht entwickelt sind oder besondere (zuvor mit anderen Verfahren oder durch Expertenurteile diagnostizierte) sprachliche Auffälligkeiten aufweisen. Die Höhe der jeweils ermittelten Sensitivitäts- und Spezifitätsraten variierte dabei deutlich je nach Zusammensetzung der Gruppe der sprachlich auffälligen Kinder. Ausgesprochen hohe Werte (> 80 %) für beide Gütemerkmale wurden insbesondere dann ermittelt, wenn – wie es in der überwiegenden Zahl der Studien der Fall war – Kinder mit einer diagnostizierten SSES betrachtet wurden (Kiese-Himmel & Rosenfeld, 2012; Rißling, Waldmann & Petermann, 2013). Ebenfalls recht hohe Sensitivitäts- und Spezifitätsraten wurden gefunden, wenn zum einen sprachlich regelgerecht entwickelte Kinder und zum anderen Kinder mit ausgeprägten sprachlichen Auffälligkeiten untersucht wurden – *und* dabei keine zusätzliche Differenzierung zwischen umgebungsbedingten Defiziten und Auffälligkeiten mit medizinischem Störungswert erfolgte. Wenn hingegen

eine solche Differenzierung vorgenommen wurde, fielen insbesondere die für die Gruppe der Kinder mit umgebungsbedingten Sprachauffälligkeiten ermittelten Sensitivitätsraten eher gering aus (< 25 %) (Neumann & Euler, 2013).

Im schulischen Alltag werden auch diagnostische Informationen aus anderen Informationsquellen herangezogen, um den sprachlichen Förderbedarf von Schülerinnen und Schülern zu bestimmen. Besonders häufig genutzt werden die Beobachtungen von Lehrerinnen und Lehrern sowie Schulnoten als spezifische Form des Lehrerurteils, die Beobachtungen weiterer pädagogischer Fachkräfte, schulärztliche Diagnosen und Informationen von Eltern (Stanat, Weirich & Radmann, 2012). Gegenüber dem Einsatz sprachdiagnostischer Verfahren hat die Verwendung von Informationen aus diesen Quellen den Vorteil, dass sie den Schulen zumeist bereits vorliegen, oder zumindest leicht zugänglich sind, und nicht aufwändig erhoben werden müssen. Untersuchungen zu der Frage, ob für die Nutzung dieser Informationsquellen ein positiver Zusammenhang mit der Sensitivität und Spezifität von Sprachförderentscheidungen auch empirisch belegt werden kann, stehen bislang allerdings noch aus.

Forschungsanliegen

Im vorliegenden Beitrag soll untersucht werden, welche Güte diagnostische Entscheidungen zur Bestimmung eines sprachlichen Förderbedarfs von Schülerinnen und Schülern in der Grundschule aufweisen. Hierfür soll zunächst ermittelt werden, wie hoch die Sensitivität und die Spezifität von Sprachförderentscheidungen in der Grundschule allgemein ausfallen. Im Anschluss soll der Frage nachgegangen werden, inwieweit die Klassifikationsgüte von Förderentscheidungen von der Verwendung bestimmter Informationsquellen bei der Sprachförderdiagnostik abhängt. Erwartet wird, dass die Nutzung sprachdiagnostischer Verfahren positiv mit der Klassifikationsgüte zusammenhängt (Hypothese A) und sich dieser Zusammenhang insbesondere bei einer mehrmaligen Erfassung der sprachlichen Kompetenzen mithilfe solcher Verfahren zeigt (Hypothese B). Zusätzlich erfolgt eine Betrachtung weiterer Informationsquellen (Beobachtungen der Lehrkraft, Informationen anderer pädagogischer Fachkräfte wie Erzieherinnen und Erzieher, schulärztliche Informationen, Elterninformationen, Schulnoten). Da diese Informationsquellen mit der Intention genutzt werden, möglichst valide Sprachförderentscheidungen zu treffen, soll auch hier geprüft werden, ob ein positiver Zusammenhang mit der Klassifikationsgüte besteht (Hypothese C).

Methode

Stichprobe

Der vorliegende Beitrag basiert auf Daten der IQB-Ländervergleichsstudie in der Primarstufe aus dem Jahr 2011 (Stanat, Pant, Böhme & Richter, 2012). Das Hauptziel dieser deutschlandweit durchgeführten Untersuchung bestand darin, auf Ebene der Länder zu überprüfen, inwieweit die bundesweit verbindlichen Bildungsstandards in den Fächern Deutsch und Mathematik für die Primarstufe (KMK, 2005a, 2005b) erreicht werden.

An der Ländervergleichsstudie 2011 nahmen Schülerinnen und Schüler der vierten Jahrgangsstufe aus insgesamt 1349 zufällig ausgewählten Grund- und Förderschulen teil. Nicht alle diese Schulen wurden in die hier berichteten statistischen Analysen einbezogen. Ausgeschlossen wurden Schulen, in denen keine systematische Sprachförderung zusätzlich zum Regelunterricht erfolgte. Auch Förderschulen wurden nicht berücksichtigt, da sich diese von den Grundschulen hinsichtlich der umgesetzten Sprachdiagnostik und -förderung teils deutlich unterscheiden. Im Ergebnis liegt diesem Beitrag eine Schulstichprobe von 651 Grundschulen zugrunde, an denen jeweils eine Schulklasse untersucht wurde. Die resultierende Schülerstichprobe umfasste 12808 Kinder. Die Kinder in dieser Stichprobe waren im Durchschnitt 10.4 Jahre alt (SD = 0.5); 49 Prozent von ihnen waren weiblich. Von diesen Kindern sprachen eigenen Angaben zufolge zu Hause 83 Prozent „immer“ oder „fast immer Deutsch“, 16 Prozent sprachen zu Hause „manchmal Deutsch und manchmal eine andere Sprache“ und ein Prozent gab an, zu Hause niemals Deutsch zu sprechen.

Instrumente

Vorhergesagter Förderbedarf (V_{F+} bzw. $V_{F–}$)

Informationen zur Sprachförderung wurden in der Ländervergleichsstudie 2011 in einer Schülerteilnahmeliste festgehalten, in der ein Schulkoordinator unter anderem vermerkte, ob eine Schülerin oder ein Schüler „zusätzliche Sprach- und Leseförderung außerhalb des Unterrichts innerhalb der Schule“ erhält. Im vorliegenden Beitrag bilden diese Angaben die Grundlage für die Bestimmung des vorhergesagten Förderbedarfs. Hierbei wurde für alle additiv geförderten Kinder angenommen, dass sie an ihren Schulen als förderbedürftig eingestuft wurden (V_{F+}). Für alle anderen Kinder wurde entsprechend

angenommen, dass sie als nicht förderbedürftig eingeschätzt wurden. Sie wurden der Kategorie V_{F-} zugeordnet.

Tatsächlicher Förderbedarf (W_{F+} bzw. W_{F-})

Für die Kompetenzmessung in der Ländervergleichsstudie 2011 wurden Testinstrumente eingesetzt, deren Aufgaben zuvor unter fachdidaktischer Anleitung entwickelt, empirisch erprobt und an einer bundesweit repräsentativen Schülerstichprobe normiert wurden. Im Fach Deutsch operationalisierten die Aufgaben zentrale Bildungsstandards aus den beiden Kompetenzbereichen „Lesen“ und „Zuhören“.[26] Aus den Testergebnissen der Schülerinnen und Schüler wurden ihre jeweiligen Kompetenzstände geschätzt, die auf einer Skala mit einem Mittelwert von 500 Punkten und einer Standardabweichung von 100 Punkten abgebildet wurden. Diese Skalenwerte wurden in einem weiteren Schritt so transformiert, dass erkennbar wurde, welche von fünf möglichen Kompetenzstufen ein Kind erreichte (Richter et al., 2012).

Die Kompetenzstufen sind als kriteriale Bezugsgrößen konzipiert und die pro Stufe relevanten Anforderungen sind in Kompetenzstufenmodellen beschrieben (Bremerich-Vos, Böhme, Krelle, Weirich & Köller, 2012). Dabei markiert Stufe II das Erreichen des Mindeststandards, der ein Minimum an Kompetenzen definiert, das am Ende der Primartstufe erreicht sein sollte, damit – bei entsprechender Unterstützung – eine erfolgreiche Integration in die Sekundarstufe I gelingen kann. Stark verkürzt dargestellt, gelingt es Schülerinnen und Schüler auf dieser Stufe, benachbarte Informationen in altersgemäßen Hör- und Lesetexten miteinander zu verknüpfen und weniger prominent platzierte Einzelinformationen korrekt wiederzugeben. Stufe III bezeichnet das Erreichen der Regelstandards, die festlegen, über welche Kompetenzen Schülerinnen und Schülern am Ende der Primarstufe im Durchschnitt verfügen sollten. Auf dieser Stufe gelingt es Kindern unter anderem, über einen Hör- oder Lesetext verstreute Informationen miteinander zu verknüpfen und ansatzweise ein globales Textverständnis zu erzielen.

Im vorliegenden Beitrag wurden die jeweils erreichten Kompetenzstufen zur Bestimmung des sprachlichen Förderbedarfs herangezogen. Grundlage hierfür bildete die Annahme, dass Schülerinnen und Schüler, deren Kompetenzen im Lesen und im Zuhören nicht den länderübergreifend definierten Mindeststandards entsprechen, einer

[26] Beispielaufgaben können unter folgendem Link eingesehen werden: https://www.iqb.hu-berlin.de/laendervergleich/LV2011/Beispielaufgaben [07.07.2016]

zusätzlichen Unterstützung beim Aufbau dieser Kompetenzen bedürfen. In diesem Zusammenhang sei auf die hohe Relevanz dieser beiden rezeptiven Sprachkompetenzen für das schulische Lernen hingewiesen: In der Grundschulzeit kommt der Zuhörkompetenz ein besonderer Stellenwert zu, da ein Großteil der Lernangebote mündlich durch die Lehrkraft vermittelt wird (Belgrad, Eriksson, Pabst-Weinschenk & Vogt, 2008). Um diese Lernangebote nutzen zu können, müssen Schülerinnen und Schüler in der Lage sein, die bildungssprachlich geprägte Unterrichtskommunikation zu verstehen und ihr Zuhörverhalten selbständig und kompetent zu steuern. Im Verlauf der Schulzeit und vor allem ab der Sekundarstufe I erfolgt die Wissensaneignung in immer stärkerem Maße durch Lesen. Die Grundlagen der hierfür notwendigen Lesekompetenz werden in der Grundschule erworben.

Für die Ermittlung der Klassifikationsgüte mussten die für Lesen und Zuhören separat vorliegenden Kompetenzstufen miteinander kombiniert werden, da im Indikator für den vorhergesagten Förderbedarf nicht nach diesen beiden Bereichen differenziert wurde. Hierbei wurden, in Anlehnung an das Vorgehen von Stanat, Weirich und Kollegen (2011), zwei Kompetenzstufengruppen gebildet. Der ersten Gruppe wurden alle Schülerinnen und Schüler zugeordnet, die in mindestens einem der beiden Kompetenzbereiche nur Stufe I (= alle Kinder, die die Mindeststandards verfehlen) und im jeweils anderen Bereich höchstens Stufe II erreichten. Für alle Kinder dieser Gruppe wurde das Vorliegen tatsächlichen Förderbedarfs angenommen (W_{F+}). Die zweite Gruppe umfasste alle Schülerinnen und Schüler, die sowohl im Lesen als auch im Zuhören mindestens die Regelstandards erreichten. Für alle Kinder dieser Gruppe wurde angenommen, dass sie keinen sprachlichen Förderbedarf haben (W_{F-}).[27]

Anzumerken ist, dass die im vorliegenden Beitrag verwendeten Kompetenzstufen in der Ländervergleichsstudie 2011 auf Basis von *plausible values* (*PV*) ermittelt wurden. In dem auf der Methode der multiplen Imputation (vgl. Rubin, 1987) beruhenden *PV*-Ansatz wird bei einer IRT-Skalierung von Testdaten statt eines einzelnen Fähigkeitswertes für jede Person eine individuelle Wahrscheinlichkeitsverteilung des Fähigkeitswerts modelliert. Aus dieser Verteilung wird zufällig eine festgelegte Anzahl an *PV* gezogen

[27] Hinweis: Die vorgenommene Klassifikation umfasst nicht alle Schülerinnen und Schüler. Ausgenommen sind diejenigen, die in beiden Bereichen zwar die Mindest-, nicht jedoch die Regelstandards erreichen, da in diesen Fällen nicht eindeutig entscheidbar ist, ob eine additive Förderung erforderlich ist oder nicht. Angaben zu den prozentualen Anteilen von W_{F+} und W_{F-} finden sich in Tabelle 5.2.

(Mislevy, Beaton, Kaplan & Sheehan, 1992). Verwendung findet der Ansatz insbesondere in großen Schulleistungsstudien, da er im Vergleich zu Punktschätzern der individuellen Fähigkeit den Vorteil hat, bei statistischen Analysen auf der Gruppenebene messfehlerbereinigte Kennwerte zu liefern.

In der Ländervergleichsstudie 2011 wurden für jedes Kind pro Kompetenzbereich 15 *PV* gezogen und zu jeweils 15 Angaben zum Erreichen einer bestimmten Kompetenzstufe transformiert. Der eingesetzte Leistungstest war zufriedenstellend reliabel, die *EAP/PV*-Reliabilität variierte je nach Bundesland und Kompetenzbereich zwischen .78 und .88.

Sprachförderung und genutzte Informationsquellen

Weitere für diesen Beitrag wesentliche Hintergrundinformationen wurden mit einem Schulleiterfragebogen erhoben. Zur Erfassung der umgesetzten Sprachförderung diente die Frage: „Findet an Ihrer Schule eine systematische Sprachförderung im Regelunterricht oder zusätzlich zum Regelunterricht statt?“. Angaben zu den Informationsquellen wurden mit folgender Frage erhoben: „Wie wird bestimmt, ob eine Schülerin/ein Schüler Ihrer Schule einen sprachlichen Förderbedarf hat?“. Als Auswahlantworten standen zur Verfügung: „eigene Beobachtung (der Lehrkraft)“, „Informationen anderer Pädagogen (Hort, Kolleginnen und Kollegen, Erzieherinnen und Erzieher)“, „Informationen aus schulärztlicher Untersuchung“, „Informationen der Eltern“, „schwache Noten“, „standardisierte Tests (z. B. Hamburger Schreibprobe, MSVK, HASE, BISC)“ und „unstandardisierte Verfahren (z. B. Beobachtungsverfahren SISMIK, SELDAK, Lernausgangsanalysen)“ (Mehrfachantworten waren möglich). Zusätzlich konnte in zwei weiteren Items mit offenem Antwortformat angegeben werden, welche konkreten „standardisierten“ und „unstandardisierten“ Verfahren an den Schulen zum Einsatz kommen. Ergänzend hierzu wurde gefragt, wie häufig der Sprachstand während der Grundschulzeit erfasst wird („nie“, „einmal“, „mehrmals“).

Der Anteil fehlender Werte variierte über diese Fragebogenitems hinweg, fiel jedoch insgesamt mit maximal acht Prozent relativ gering aus. Die Imputation fehlender Werte erfolgte mit dem Verfahren *Multivariate Imputation by Chained Equations* unter Verwendung des Pakets *mice* (van Buuren & Groothuis-Oudshoorn, 2011) in der Statistiksoftware R.

Statistische Auswertungsmethoden

Da der Bestimmung des tatsächlichen Förderbedarfs *PV* zugrunde lagen, erfolgte die statistische Auswertung auf der *Schulebene*. Alle Analysen wurden zunächst separat für jede der 225 Kombinationen durchgeführt, die aus der Verknüpfung von jeweils 15 Kompetenzstufenwerten im Lesen und Zuhören resultierten, und im Anschluss gepoolt (vgl. Marshall, Altman, Holder & Royston, 2009).

Zur Berechnung der Sensitivität von Sprachförderentscheidungen wurde zunächst in jeder Schule die Gruppe derjenigen Schülerinnen und Schüler bestimmt, die (nach unserer Definition) einen Sprachförderbedarf haben. Im Anschluss daran wurde die Sensitivitätsrate als prozentualer Anteil der Kinder in dieser Gruppe ermittelt, denen eine additive Sprach- oder Leseförderung zu Teil wird. Die Berechnung der Spezifitätsrate von Förderentscheidungen erfolgte analog dazu, in dem an jeder Schule für die Gruppe der Schülerinnen und Schüler ohne Förderbedarf der Anteil der Kinder ermittelt wurde, die keine zusätzliche Förderung erhalten.

Die oben genannten Hypothesen zu Zusammenhängen zwischen der Klassifikationsgüte und der Nutzung bestimmter diagnostischer Informationsquellen wurden mit multiplen Regressionsanalysen untersucht. Als Kriteriumsvariablen dienten die zuvor berechneten Sensitivitäts- bzw. Spezifitätsraten der einzelnen Schulen. Als Prädiktorvariablen wurden die im Schulleiterfragebogen erfassten Angaben über die zur Feststellung sprachlichen Förderbedarfs herangezogenen Informationsquellen in die Modelle integriert. „Standardisierte Tests“ und „unstandardisierte Verfahren“ wurden dabei zu einer Kategorie („sprachdiagnostische Verfahren“) zusammengefasst, da eine inhaltliche Auswertung der offenen Items zu den eingesetzten Verfahren ergab, dass die Schulleiterinnen und Schulleiter Schwierigkeiten mit der Differenzierung von „standardisiert“ und „unstandardisiert“ hatten.

Für die Prädiktorvariable „sprachdiagnostische Verfahren“ wurde zusätzlich differenziert, ob die Verfahren an den Schulen einmalig oder mehrmals zum Einsatz kamen. Da alle Prädiktorvariablen dummy codiert wurden, können die ermittelten Regressionskoeffizienten als die Veränderung an Sensitivität und Spezifität in Prozentpunkten interpretiert werden, die mit der jeweiligen Nutzung einer Informationsquelle (unter statistischer Kontrolle der Angaben zur Nutzung der weiteren Informationsquellen) einhergeht.

Ergebnisse

Deskriptive Angaben zum Anteil der Schülerinnen und Schüler, die zusätzlich zum Regelunterricht eine Sprach- oder Leseförderung in der Schule erhalten (V_{F+}), sowie zum Anteil der Kinder, die nach der von uns verwendeten Definition einen sprachlichen Förderbedarf haben (W_{F+}) beziehungsweise nicht haben ($W_{F–}$), sind in Tabelle 5.2 dargestellt.

Tabelle 5.2: Arithmetischer Mittelwert (*MW*) und Streuung (Quartilsgrenzen Q_x) der Anteile von V_{F+}, W_{F+} und $W_{F–}$ an den untersuchten Schulen

	M	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$
Anteil der Kinder, die eine additive Sprach- oder Leseförderung erhalten (V_{F+})	13.3 %	0 %	7.7 %	21.1 %
Anteil der Kinder mit sprachlichem Förderbedarf (W_{F+})	13.3 %	4.2 %	9.6 %	18.7 %
Anteil der Kinder ohne sprachliche Förderbedarf ($W_{F–}$)	57.1 %	43.8 %	59.0 %	71.7 %

Im Mittel aller Schulen wurde für rund 13 Prozent der Kinder vermerkt, dass sie eine additive Sprach- oder Leseförderung erhalten. Ein Blick auf die ebenfalls in Tabelle 5.2 dargestellten Quartilsgrenzen verdeutlicht, dass dieser Anteil zwischen den untersuchten Schulen stark variiert. In rund 40 Prozent der betrachteten Schulklassen erhielt kein einziges Kind eine additive Sprach- und Leseförderung. Der Anteil der Kinder, für die ein sprachlicher Förderbedarf ermittelt wurde, lag im Schulmittel ebenfalls bei rund 13 Prozent. Auch dieser Anteil variierte deutlich zwischen den Schulklassen. Der Anteil der Kinder, die mindestens die Regelstandards in den beiden Bereichen Lesen und Zuhören erreichten, betrug insgesamt rund 57 Prozent und war ebenfalls durch eine hohe Varianz gekennzeichnet.

Deskriptive Statistiken der auf Schulebene berechneten Sensitivitäts- und Spezifitätsraten finden sich in Tabelle 5.3. Insgesamt wurde eine mittlere gepoolte Sensitivitätsrate von rund 33 Prozent ermittelt. Auch diese Rate variierte deutlich zwischen den Schulklassen (SD = 0.38). Des Weiteren wurde eine überaus hohe mittlere

Spezifitätsrate von rund 93 Prozent errechnet, die eine weniger starke Streuung aufwies (*SD* = 0.15).

In Tabelle 5.3 sind außerdem die mittleren Sensitivitäts- und Spezifitätsraten separiert nach der Nutzung der unterschiedlichen diagnostischen Informationsquellen (deren Häufigkeit in der Schulstichprobe ebenfalls in Tabelle 5.3 zu finden sind) aufgeführt. Deutliche Differenzen zeigten sich deskriptiv bei den mittleren Sensitivitätsraten, wobei die größten Unterschiede im Hinblick auf die Nutzung von sprachdiagnostischen Verfahren (mehrmalige Nutzung: 39 %, einmalige Nutzung: 32 %, keine Verfahren: 27 %) und Beobachtungen der Lehrkraft (43 % vs. 32 %) als diagnostische Informationsquellen festgestellt wurden. Demgegenüber wurden für die Spezifitätsraten nur sehr geringe Unterschiede festgestellt.

Tabelle 5.3: Mittlere gepoolte Sensitivitäts- und Spezifitätsraten, separiert nach der Nutzung der unterschiedlichen Informationsquellen bei der Feststellung sprachlichen Förderbedarfs

	Sensitivität	Spezifität	Häufigkeit in Schulstichprobe
Alle Schulen	32.8 %	93.3 %	100 %
Eigene Beobachtung (der Lehrkraft)			
Nein	42.8 %	94.5 %	3.7 %
Ja	32.4 %	94.0 %	96.3 %
Informationen anderer Pädagogen			
Nein	33.0 %	94.2 %	18.4 %
Ja	32.8 %	93.4 %	81.6 %
Schulärztlicher Untersuchung			
Nein	36.4 %	94.5 %	21.7 %
Ja	31.8 %	93.2 %	78.3 %

Informationen der Eltern			
Nein	35.1 %	93.1 %	31.8 %
Ja	31.8 %	93.7 %	68.2 %
Schwache Noten			
Nein	33.3 %	93.4 %	52.5 %
Ja	32.3 %	93.7 %	47.5 %
Sprachdiagnostische Verfahren			
Nein	26.7 %	94.0 %	35.3 %
Ja	35.8 %	93.3 %	64.7 %
Einmalig	32.2 %	93.6 %	29.7 %
Mehrmals	38.8 %	93.1 %	35.0 %

Den Regressionsanalysen wurde eine Kollinearitätsanalyse nach dem Ansatz des *variance inflation factor* (*vif*) vorangestellt. Der *vif* einer Prädiktorvariable wird aus dem Bestimmtheitsmaß (R^2) für die Regression dieses Prädiktors auf die anderen Prädiktorvariablen berechnet. Ein *vif* größer Zehn wird häufig als Hinweis auf starke Multikollinearität interpretiert (O'Brien, 2007). In unseren Analysen lagen die für die einzelnen Prädiktorvariablen bestimmten *vif* deutlich unter diesem Grenzwert ($vif_{max} < 1.4$). Somit kann davon ausgegangen werden, dass keine Multikollinearitätsprobleme vorlagen.

In Tabelle 5.4 sind die gepoolten Ergebnisse der Regressionsanalysen dargestellt, in denen die zuvor (auf Schul- bzw. Klassenebene) berechneten Sensitivitätsraten mit den Variablen zur Nutzung bestimmter Informationsquellen bei der Feststellung sprachlichen Förderbedarfs prädiziert werden. Das Befundmuster stützt die Hypothesen A und B: So wurde in Modell 1 ein positiver, statistisch signifikanter Zusammenhang zwischen der Nutzung sprachdiagnostischer Verfahren und der Sensitivitätsrate ermittelt ($b_p = .09$, $p < .05$). Bei einer zusätzlichen Differenzierung nach der Einsatzhäufigkeit (Modell 2) blieb dieser Effekt nur für den mehrmaligen Einsatz der Verfahren bestehen ($b_p = .11$, $p < .01$). Keine Bestätigung bieten die Ergebnisse für Hypothese C: Trotz der deskriptiv

gefundenen Unterschiede wurden für die weiteren Informationsquellen keine signifikanten Zusammenhänge mit der Sensitivitätsrate festgestellt. Dabei fällt beim Blick auf den Prädiktor „Eigene Beobachtung (der Lehrkraft)“ der hohe Standardfehler auf, der statistisch darauf zurückgeführt werden kann, dass diese Informationsquelle in einem sehr großen Teil der untersuchten Schulen herangezogen wird und die Sensitivitätsraten gleichzeitig stark zwischen den Schulen variieren.

Tabelle 5.4: (Gepoolte) Ergebnisse der Regressionsanalysen zum Zusammenhang der Nutzung bestimmter Informationsquellen bei der Feststellung sprachlichen Förderbedarfs mit der Sensitivität von Sprachförderentscheidungen

Prädiktorvariablen (Kriterium: Sensitivitätsrate)	Modell 1		Modell 2	
	b_p	*SE*	b_p	*SE*
Eigene Beobachtung (der Lehrkraft)	-0.10	0.10	-0.09	0.10
Informationen anderer Pädagogen	0.03	0.05	0.03	0.05
Informationen aus schulärztlicher Untersuchung	-0.03	0.05	-0.03	0.05
Informationen der Eltern	-0.02	0.05	-0.02	0.05
Schwache Noten	-0.01	0.04	-0.01	0.04
Sprachdiagnostische Verfahren (Referenz: kein Einsatz)				
Einsatz	0.09*	0.04		
einmalig			0.05	0.05
mehrmals			0.11**	0.04

Anmerkung: $^{*}p < .05$, $^{**}p < .01$, $R^2 = .04$ (für beide Modelle)

Keine statistisch signifikanten Effekte und somit auch keine weitere empirische Bestätigung für die zuvor aufgestellten Hypothesen fanden sich in (hier aus Platzgründen nicht dargestellten) Regressionsmodellen, in denen die Spezifitätsraten auf die Angaben zur Nutzung diagnostischer Informationsquellen regrediert wurden.

Diskussion

Im vorliegenden Beitrag wurde untersucht, inwiefern sich die Klassifikationsgüte von Sprachförderentscheidungen verändert, wenn hierfür bestimmte diagnostische Informationsquellen herangezogen werden. Erwartungskonform mit Hypothese A wurde für die Nutzung sprachdiagnostischer Verfahren ein statistisch signifikanter Zusammenhang mit der Sensitivitätsrate ermittelt. Bei einer zusätzlichen Differenzierung nach der Einsatzhäufigkeit wurde dieser Effekt, konsistent zu Hypothese B, nur für die mehrmalige, nicht aber für die einmalige Nutzung von sprachdiagnostischen Verfahren als Informationsquelle gefunden. Eine signifikante Verringerung der Spezifitätsrate wurde hingegen *nicht* festgestellt. Dies lässt darauf schließen, dass die gefundene Verbesserung in der Sensitivitätsrate nicht durch eine Verschiebung des Cut-Off-Wertes hin zu einer liberaleren Klassifikation bedingt ist, sondern vielmehr eine tatsächliche Erhöhung der Klassifikationsgüte abbildet. Für die weiteren betrachteten diagnostischen Informationsquellen (Beobachtungen der Lehrkraft, Informationen anderer pädagogischer Fachkräfte wie Erzieherinnen und Erzieher, schulärztliche Informationen, Elterninformationen, Schulnoten) wurden keine signifikanten Zusammenhänge mit der Sensitivitäts- oder Spezifitätsrate ermittelt. Hypothese C, in der auch für die Nutzung dieser Informationsquellen eine Erhöhung der Klassifikationsgüte erwartet wurde, konnte folglich nicht bestätigt werden.

Insgesamt verdeutlichen die Ergebnisse, dass Sprachförderentscheidungen in Grundschulen im Mittel zwar eine ausgesprochen hohe Spezifität aufweisen, gleichzeitig aber durch eine niedrige Sensitivität gekennzeichnet sind. Dieses Befundmuster und auch die Höhe der von uns gefundenen Sensitivitäts- und Spezifitätsraten ähneln den Ergebnissen von Studien, in denen die Klassifikationsgüte einzelner sprachdiagnostischer Verfahren bei der Identifikation von förderbedürftigen Kindern mit umgebungsbedingten Sprachauffälligkeiten untersucht wurde (vgl. Neumann & Euler, 2013).

Für den mehrfachen Einsatz sprachdiagnostischer Verfahren konnte zwar ein signifikanter, positiver Zusammenhang mit der Sensitivitätsrate belegt werden, dennoch fällt die Sensitivität auch hier relativ gering aus. Dass keine stärkeren Veränderungen gefunden wurden, könnte durch die Heterogenität der an den Grundschulen eingesetzten Verfahren bedingt sein. So ist dokumentiert, dass die psychometrische Qualität der gegenwärtig genutzten Instrumente deutlich variiert und zum Teil so gering ausfällt, dass

ein positiver Beitrag zur Klassifikationsgüte zu bezweifeln ist (vgl. Hoffmann et al., 2016). Demnach erscheint für zukünftige Untersuchungen lohnenswert, sprachdiagnostische Verfahren nicht nur (wie hier) mit einer einzigen Variable abzubilden, sondern stärker zu differenzieren und zu untersuchen, welchen Einfluss insbesondere der Einsatz von Instrumenten, die essentiellen Gütekriterien (z. B. Objektivität, Reliabilität, Validität) genügen, auf die Klassifikationsgüte von Sprachförderentscheidungen hat. In der vorliegenden Studie war eine solche Differenzierung nicht möglich, da nur für einen kleinen Teil der untersuchten Schulen bekannt war, welche Verfahren konkret eingesetzt werden, und mithin die einzelnen Fallzahlen zu gering waren, um aussagekräftige statistische Analysen vornehmen zu können. Anzumerken sei, dass die Qualitätsbedenken nicht per se mit dem Standardisierungsgrad der Verfahren korrespondieren. So erfordert etwa die Erfassung pragmatischer Kompetenzen eine weniger standardisierte Erhebungssituation, da hierfür spontansprachliches Handeln elizitiert werden muss. Vielmehr beziehen sich die Bedenken auf die Auswertungs- bzw. Interpretationsobjektivität oder auf das Fehlen von empirischen Reliabilitäts- und Validitätsbelegen (vgl. Hoffmann et al., 2016).

Das Befundmuster zur Nutzung der weiteren Informationsquellen (vgl. Hypothese C) bedarf einer vertieften Diskussion. Die Ergebnisse zu den *Beobachtungen von Lehrkräften* sind durch einen hohen Standardfehler gekennzeichnet, der sich als Hinweis darauf interpretieren lässt, dass die Sensitivitätsrate in den betreffenden Grundschulen überaus unterschiedlich ausfällt, so dass vielfach auch durchaus zufriedenstellende Werte erreicht werden. Diese Interpretation ist konsistent mit Befunden zur diagnostischen Kompetenz von Lehrkräften, nach denen die Genauigkeit von Lehrerurteilen zu Schülerkompetenzen interindividuell stark variiert, sich also sowohl Lehrkräfte finden, die in der Lage sind, überaus akkurate Einschätzungen vorzunehmen, als auch Lehrpersonen, deren Beurteilungen sehr ungenau ausfallen (Südkamp, Kaiser & Möller, 2012). Die Tatsache, dass für die Beobachtungen von Lehrkräften keine Zusammenhänge mit der Klassifikationsgüte festgestellt wurden, lässt sich außerdem mit einem weiteren Befund aus der Forschung zur diagnostischen Kompetenz erklären. Es gilt als empirisch belegt, dass Lehrkräfte zu einer systematischen Überschätzung des Leistungsniveaus von Schülerinnen und Schülern tendieren (vgl. bspw. Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2009). Dies birgt die Gefahr, dass Kinder mit besonderen Schwächen übersehen werden (z. B. Artelt, Stanat, Schneider, Schiefele & Lehmann, 2004) und bestimmte

Förderbedarfe unerkannt bleiben. Darüber hinaus konnte für Schulnoten, die eine spezielle Form von Lehrerurteilen darstellen, gezeigt werden, dass diese oftmals vor dem Hintergrund eines klasseninternen Bezugsrahmens und somit nur bedingt kriterial vergeben werden (z. B. Ingenkamp, 1989).

Die Befunde zur Akkuratheit von *Elternauskünften* ähneln den Ergebnissen zur diagnostischen Kompetenz von Lehrkräften. Auch Eltern gelingt es unterschiedlich gut, die Kompetenzen ihrer Kinder zu beurteilen. Analog zu Lehrkräften neigen sie dazu, das Leistungsniveau ihrer Kinder zu überschätzen (Deimann, Kastner-Koller, Benka, Kainz & Schmidt, 2005; Frischknecht, Reimann, Gut, Ledermann & Grob, 2014).

Leistungseinschätzungen von Erzieherinnen und Erziehern weisen eine noch geringere Diagnosegenauigkeit als Lehrerurteile auf (Dollinger, 2013). Bezüglich *schulärztlicher Informationen* ist zu vermuten, dass diese weniger die Feststellung sprachlichen Förderbedarfs als vielmehr klinische Diagnosen, etwa zum Vorliegen einer SSES, zum Gegenstand haben.

Diese Darstellungen lassen darauf schließen, dass die bei Hypothese C betrachteten Informationsquellen nicht selten eine eingeschränkte Reliabilität und Validität aufweisen und daher im vorliegenden Beitrag nur geringe Zusammenhänge mit der Klassifikationsgüte ermittelt wurden. Doch noch ein weiterer Erklärungsansatz ist zu beachten: In diesem Beitrag basierte die Bestimmung des Anteils der Schülerinnen und Schüler mit tatsächlichem Förderbedarf auf den Ergebnissen eines Kompetenztests. Gleichzeitig wurden nur für die Nutzung sprachdiagnostischer Verfahren, nicht aber für die anderen Informationsquellen, statistisch signifikante Zusammenhänge mit der Klassifikationsgüte festgestellt. Folglich könnte vermutet werden, dass das Befundmuster einen Methodeneffekt abbildet. Uns erscheint diese Alternativerklärung aus folgenden Gründen unwahrscheinlich: (1) Unter den sprachdiagnostischen Verfahren waren in diesem Beitrag nicht nur reine Leistungstests, sondern auch andere Instrumente wie Einschätz- und Beobachtungsverfahren zusammengefasst, die sich methodisch deutlich vom Kompetenztest der Ländervergleichsstudie 2011 unterscheiden. (2) Die Besonderheit des in der Ländervergleichsstudie 2011 eingesetzten Testinstruments besteht in seiner theoretischen Grundlage. Der Test operationalisiert die in den Bildungsstandards formulierten Kompetenzerwartungen und dient (unter anderem) dazu, den Anteil der Kinder zu ermitteln, deren sprachliche Kompetenzen ein bestimmtes, für das schulische Lernen relevantes Mindestniveau nicht erreichen. Somit unterscheidet er sich von

sprachdiagnostischen Verfahren, die nicht an den Bildungsstandards orientiert sind, sondern (zumindest idealiter) auf Theorien und empirischen Befunden zum Spracherwerb basieren.

Der vorliegende Beitrag ist mit einigen Einschränkungen verbunden, die vor allem aus einem Mangel an Verfügbarkeit bestimmter Angaben resultieren. Zum Beispiel lagen keine Hintergrundinformationen zum Zeitpunkt der Diagnose und zur Dauer der jeweiligen Förderung vor. Denkbar ist jedoch, dass einige der untersuchten Kinder bereits vor längerer Zeit zutreffend als förderbedürftig eingestuft wurden, seither eine zusätzliche Förderung erhielten und diese Förderung so erfolgreich war, dass die Kinder zum Testzeitpunkt die Regelstandards in den Bereichen Lesen und Zuhören erreichen konnten. In der vorliegenden Studie wären solche Fälle als Falsch-positiv klassifiziert worden, obwohl vormals ein tatsächlicher Förderbedarf bestand und man die betreffenden Kinder der Kategorie Wahr-positiv hätte zuordnen müssen. Demnach ist mit Blick auf die von uns ermittelten Gütekennwerte darauf hinzuweisen, dass sie die tatsächlichen Sensitivitäts- und Spezifitätsraten möglicherweise leicht unterschätzen. Darüber hinaus wären differenziertere Analysen möglich gewesen, wenn zusätzlich zu der Information, dass der Sprachstand „mehrmals“ erfasst wird, Angaben zum jeweiligen Zeitintervall vorgelegen hätten.

In diesem Beitrag wurde die Klassifikationsgüte explizit auf Schulebene untersucht, eine nach individuellen Schülermerkmalen differenzierte Betrachtung der Klassifikationsgüte wurde nicht vorgenommen. Analysen zum Zusammenhang zwischen Schülermerkmalen und Förderquoten wurden bereits von Stanat, Weirich und Kollegen (2012) durchgeführt und sind im Berichtsband zur Ländervergleichsstudie 2011 dokumentiert. Es konnte gezeigt werden, dass Kinder mit Zuwanderungshintergrund bei gleichem Kompetenzniveau insgesamt häufiger eine additive Förderung erhalten (vgl. auch Böhme, Felbrich, Weirich & Stanat, 2013). Für den vorliegenden Beitrag ist darauf hinzuweisen, dass das berichtete Befundmuster bestehen bleibt, wenn zusätzlich für den Anteil von Kindern mit nichtdeutscher Muttersprache an der Schule statistisch kontrolliert wird.

Die hier dargestellten Ergebnisse und Erklärungsansätze legen weiterführende Forschungsfragen nahe, die mit den Daten der Ländervergleichsstudie 2011 nicht bearbeitet werden können und daher als Forschungsdesiderata festzuhalten sind: Lohnenswert erscheint insbesondere eine detailliertere Untersuchung der an den Schulen

umgesetzten diagnostischen Strategien und Prozesse, wobei betrachtet werden sollte, wie die erfassten diagnostischen Informationen zum sprachlichen Förderbedarf jeweils weiter verarbeitet, gewichtet und zu einem Urteil kombiniert werden. Es ist zu vermuten, dass den in diesem diagnostischen Prozess getroffenen Entscheidungen eine hohe Relevanz für die Klassifikationsgüte zukommt. Insbesondere bezüglich der bei Hypothese C betrachteten Informationsquellen sollte differenziert erfasst werden, welche konkreten diagnostischen Informationen jeweils erhoben und für die Förderentscheidungen herangezogen werden. So ist bekannt, dass die Qualität von Elternauskünften je nach Art der Information und je nach Befragungstechnik variiert (vgl. Deimann et al., 2005).

Für die schulische Praxis legen die hier dargestellten Ergebnisse den Schluss nahe, dass der Einsatz sprachdiagnostischer Verfahren einen wesentlichen Schlüssel zur Optimierung der Klassifikationsgüte von Sprachförderentscheidungen darstellt. Hierbei ist eine einmalige Erhebung des Sprachstands offenbar nicht ausreichend, um maßgebliche Verbesserungen der Klassifikationsgüte zu erzielen. Angezeigt erscheint vielmehr, die sprachlichen Kompetenzen mehrfach im Laufe der Grundschulzeit mithilfe sprachdiagnostischer Verfahren zu überprüfen, um Kinder mit ausgeprägtem Unterstützungsbedarf im sprachlichen Bereich möglichst zuverlässig zu identifizieren und anschließend in geeigneter Weise fördern zu können. Aufgrund der Tatsache, dass im vorliegenden Beitrag keine Zusammenhänge zwischen der Klassifikationsgüte und der Nutzung der übrigen Informationsquellen gefunden wurden, sollte aus unserer Sicht nicht gefolgert werden, dass auf diese zu verzichten ist. Auch Daten aus diesen Informationsquellen können potenziell wichtige Hinweise auf Förderbedarfe liefern und haben außerdem den Vorteil, dass sie nicht aufwändig erhoben werden müssen. Jedoch sollte die schulische Praxis stärker für Reliabilitäts- und Validitätsaspekte dieser Daten sensibilisiert werden. Außerdem erscheint es sinnvoll, die diagnostische Kompetenz von Lehrkräften, etwa durch entsprechende Fortbildungsmaßnahmen, systematisch zu stärken (vgl. Hosenfeld, Helmke & Schrader, 2002).

Literaturverzeichnis

Artelt, C., Stanat, P., Schneider, W., Schiefele, U. & Lehmann, R. (2004). Die PISA-Studie zur Lesekompetenz: Überblick, differenzierte Analysen und Einordnung der Befunde. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000*. (S. 139–168). Wiesbaden: VS Verl. für Sozialwissenschaften.

Bates, C. & Nettelbeck, T. (2001). Primary School Teachers' Judgements of Reading Achievement. *Educational Psychology, 21*(2), 177–187.

Baumert, J. & Schlümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In D. PISA-Konsortium (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 323–410). Opladen: Leske+Budrich.

Becker-Mrotzek, M., Ehlich, K., Füssenich, I., Günther, H., Hasselhorn, M., Hopf, M., Jeuk, S., Lengyel, D., Neugebauer, U., Panagiotopoulou, A., Stanat, P. & Wilbert, J. (2013). *Qualitätsmerkmale für Sprachstandsverfahren im Elementarbereich. Ein Bewertungsrahmen für fundierte Sprachdiagnostik in der Kita*. Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache.

Belgrad, J., Eriksson, B., Pabst-Weinschenk, M. & Vogt, R. (2008). Die Evaluation von Mündlichkeit. Kompetenzen in den Bereichen Sprechen, Zuhören und szenisch Spielen. *Didaktik Deutsch (Sonderheft zum 16. Symposium Deutschdidaktik Kompetenzen im Deutschunterricht), 14*, 20–45.

Böhme, K., Felbrich, A., Weirich, S. & Stanat, P. (2013). Sprachliche Kompetenzen von Schülern mit Zuwanderungshintergrund am Ende der 4. Jahrgangsstufe. *Die Deutsche Schule, 105*, 128–143.

Bredel, U. (2005). Sprachstandsmessung – eine verlassene Landschaft. In K. Ehlich (Hrsg.), *Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Förderung von Kindern mit und ohne Migrationshintergrund* (S. 77–119). Berlin: BMBF.

Bremerich-Vos, A., Böhme, K., Krelle, M., Weirich, S. & Köller, O. (2012). Kompetenzstufenmodelle im Fach Deutsch. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 56–71). Münster: Waxmann.

Deimann, P., Kastner-Koller, U., Benka, M., Kainz, S. & Schmidt, H. (2005). Mütter als Entwicklungsdiagnostikerinnen. Zeitschrift für Entwicklungspsychologie und *Pädagogische Psychologie, 37*(3), 122–134.

Dollinger, S. (2013). *Diagnosegenauigkeit von ErzieherInnen und LehrerInnen. Einschätzung schulrelevanter Kompetenzen in der Übergangsphase*. Wiesbaden: Springer VS.

Eckhardt, A. G. (2008). *Sprache als Barriere für den schulischen Erfolg. Potentielle Schwierigkeiten beim Erwerb schulbezogener Sprache für Kinder mit Migrationshintergrund*. Münster: Waxmann.

Feinberg, A. B. & Shapiro, E. S. (2009). Teacher Accuracy: An Examination of Teacher-Based Judgments of Students' Reading With Differing Achievement Levels. *The Journal of Educational Research, 102*(6), 453–462.

Frischknecht, M.-C., Reimann, G., Gut, J., Ledermann, T. & Grob, A. (2014). Wie genau können Mütter die Mathematik- und Sprachleistungen ihrer Kinder einschätzen? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46*(2), 67–78.

Gogolin, I. & Lange, I. (2011). Bildungssprache und Durchgängige Sprachbildung. In S. Fürstenau & M. Gomolla (Hrsg.), *Migration und schulischer Wandel: Mehrsprachigkeit* (S. 107–128). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gogolin, I., Lange, I., Hawighorst, B., Bainski, C., Heintze, A., Rutten, S. & Saalmann, W. (2011). *Durchgängige Sprachbildung. Qualitätsmerkmale für den Unterricht*. Münster: Waxmann.

Heppt, B., Stanat, P., Dragon, N., Berendes, K. & Weinert, S. (2014). Bildungssprachliche Anforderungen und Hörverstehen bei Kindern mit deutscher und nicht-deutscher Familiensprache. *Zeitschrift für Pädagogische Psychologie, 28*(3), 139–149.

Hoffmann, L., Böhme, K. & Stanat, P. (im Druck). Mit welchen diagnostischen Verfahren wird in Grundschulen Sprachförderbedarf festgestellt? Eine bundesweite Bestandsaufnahme. *Frühe Bildung*.

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

Ingenkamp, K. (Hrsg.). (1989). *Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte*. Weinheim: Beltz.

Kiese-Himmel, C. & Rosenfeld, J. (2012). Analyse aktueller Untersuchungsinstrumente zur Früherkennung von Auffälligkeiten in Sprechen und Sprache in der pädiatrischen Vorsorgeuntersuchung U8. *Gesundheitswesen, 74*, 661–672.

KMK. (2005a). *Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004*. München: Luchterhand.

KMK. (2005b). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004*. München: Luchterhand.

Marshall, A., Altman, D. G., Holder, R. L. & Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology, 9*, 9–57.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating Population Characteristics from Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mußmann, J. (2012). *Inklusive Sprachförderung in der Grundschule*. München: Reinhardt.

Neumann, K. & Euler, H. A. (2013). Kann ein Sprachstandsscreening zwischen dem Bedarf für Sprachförderung und für Sprachtherapie trennen? In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven. 1.* (S. 174-198). Münster; New York; München; Berlin: Waxmann.

Noack, H. & Petermann, F. (1995). Entscheidungstheorie. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (S. 295–310). Weinheim: PVU.

O'Brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity, 41*(5), 673–690.

Richter, D., Engelbert, M., Böhme, K., Haag, N., Hannighofer, J., Reimers, H., Roppelt, A., Weirich, S., Pant, H. A. & Stanat, P. (2012). Anlage und Durchführung des Ländervergleichs. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 85–102). Münster: Waxmann.

Rißling, J.-K., Waldmann, H.-C. & Petermann, F. (2013). Sprachstandserhebung im Grundschulalter - Sensitivität und Spezifität des SET 5-10. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 61*(2), 121–125.

Rothweiler, M. (2013). Spezifische Sprachentwicklungsstörungen bei mehrsprachigen Kindern. *Sprache Stimme Gehör, 37*(4), 186-190.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schmidt-Atzert, L. & Amelang, M. (2012). Zuordnungs- und Klassifikationsstrategien. In L. Schmidt-Atzert & M. Amelang (Hrsg.), *Psychologische Diagnostik* (5. Aufl., S. 409–428). Berlin: Springer.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

Stanat, P., Weirich, S. & Radmann, S. (2012). Sprach- und Leseförderung. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 251–276). Münster: Waxmann.

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.

Swets, J. A. (1964). *Signal detection and recognition by human observers: contemporary readings*. New York: Wiley.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations. *Journal of Statistical Software, 45*, 1–67.

6

Teilstudie 3

6 Teilstudie 3 (Originalarbeit)

Mit welchen diagnostischen Verfahren wird in Grundschulen Sprachförderbedarf festgestellt? Eine bundesweite Bestandsaufnahme

Which Diagnostic Instruments Are Used to Determine Need for Language Support in Elementary Schools? A Nationwide Survey

Lars Hoffmann[1], Katrin Böhme[1], Petra Stanat[1]

[1]Humboldt Universität zu Berlin,
Institut zur Qualitätsentwicklung im Bildungswesen (IQB)

Die Teilstudie ist als Zeitschriftenbeitrag veröffentlicht und wie folgt zugänglich:

Hoffmann, L., Böhme, K. & Stanat, P. (2017). Mit welchen diagnostischen Verfahren wird in Grundschulen Sprachförderbedarf festgestellt? *Frühe Bildung, 6*(3), 116–123. doi: 10.1026/2191-9186/a000313

Zusammenfassung: Basierend auf Daten der IQB-Ländervergleichsstudie 2011 (Stanat, Pant, Böhme & Richter, 2012) wird der Frage nachgegangen, welche diagnostischen Verfahren gegenwärtig bundesweit an den Grundschulen zur Feststellung sprachlichen Förderbedarfs genutzt werden. Die Ergebnisse zeigen, dass die Schulen hierbei sehr unterschiedlich vorgehen und eine Vielzahl verschiedener Verfahren einsetzen. Zusätzlich wird betrachtet, inwiefern die genutzten Verfahren (1) für das Grundschulalter entwickelt wurden, (2) auf Kinder mit nichtdeutscher Herkunftssprache fokussieren, (3) mündliche oder schriftliche Sprachkompetenzen erfassen und (4) testtheoretischen Gütekriterien genügen. Diese Differenzierung verdeutlicht etwa, dass an vielen Grundschulen ausschließlich Verfahren eingesetzt werden, die auf die Erfassung schriftsprachlicher Kompetenzen zielen. Abschließend werden die zentralen Befunde des Beitrags mit Blick auf die Forschungsliteratur diskutiert.

Schlüsselbegriffe: Sprachdiagnostik, Sprachförderung, Sprachstandsfeststellung, Primarbereich

Summary: Based on data from the IQB National Assessment Study 2011 (Stanat, Pant, Böhme & Richter, 2012), we determine which diagnostic instruments are currently used in elementary schools throughout Germany to identify need for language support. Our results show that schools proceed very differently while using a variety of different instruments. Furthermore, we consider whether the instruments (1) were developed for elementary school age, (2) are focused on children with a non-German mother tongue, (3) assess oral or written language proficiency and (4) meet quality criteria of test theory. This differentiation illustrates (e. g.) that many schools are exclusively using instruments addressing written language proficiency. Eventually, central results of the article are discussed in reference to the research literature.

Keywords: language diagnostics, language support, language assessment, primary education

Für das schulische Lernen kommt sprachlichen Kompetenzen ein zentraler Stellenwert zu, da sie Voraussetzung dafür sind, dass Schülerinnen und Schüler am kommunikativen Unterrichtsgeschehen teilhaben und das im Unterricht vermittelte Wissen erschließen können (Baumert & Schümer, 2001). An Grundschulen sollen Sprachfördermaßnahmen sicherstellen, dass sich möglichst alle Kinder auf einem für ein erfolgreiches Lernen erforderlichen sprachlichen Kompetenzniveau bewegen. Diese Maßnahmen können im Regelunterricht implementiert sein und somit auf alle Schülerinnen und Schüler zielen (vgl. Durchgängige Sprachbildung, z. B. Gogolin & Lange, 2011). Für Kinder mit ausgeprägtem Sprachförderbedarf können zudem additive Fördermaßnahmen erforderlich sein, die in einer zusätzlichen Lernzeit außerhalb des Regelunterrichts stattfinden.

Vor der Durchführung einer additiven Förderung ist jeweils zu entscheiden, inwiefern ein Kind sprachlichen Förderbedarf hat und ob dieser so stark ausgeprägt ist, dass eine zusätzliche Förderung erfolgen sollte. Im diagnostischen Prozess ist dieser Entscheidung eine Beurteilung der sprachlichen Kompetenzen des betreffenden Kindes vorangestellt (vgl. Geist, 2014). Empirische Befunde zeigen allerdings, dass Lehrerurteile zur Ausprägung (schrift-) sprachlicher Kompetenzen oftmals ungenau sind und vor allem Schülerinnen und Schüler, die leistungsschwach sind und mithin einen Förderbedarf haben, nicht sicher identifiziert werden (z. B. Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003). Daher wird in Konzeptionen zur (sprach-) diagnostischen Kompetenz die Erwartung formuliert, dass Lehrkräfte „Werkzeuge“ nutzen können sollten, mit deren Hilfe sich das Risiko von Fehlurteilen reduzieren lässt (vgl. Geist, 2014; von Aufschnaiter et al., 2015). Hierzu zählen unter anderem diagnostische Verfahren wie Tests, Screenings, Einschätzverfahren sowie Beobachtungsbögen. Diese Instrumente stehen im Fokus des vorliegenden Beitrags. Insbesondere soll betrachtet werden, welche konkreten diagnostischen Verfahren bundesweit an Grundschulen zur Feststellung sprachlichen Förderbedarfs eingesetzt werden.

Bisherige Bestandsaufnahmen

Wissenschaftliche Bestandsaufnahmen zum Einsatz sprachdiagnostischer Verfahren liegen in Deutschland bislang vor allem für den *Elementarbereich* vor (z. B. Autorengruppe Bildungsberichterstattung, 2012; Lisker, 2013). Methodisch basieren diese Berichte jeweils auf Informationen, die in schriftlichen Befragungen der zuständigen Landesministerien und mit ergänzenden Internetrecherchen erhoben wurden.

Sie bilden vor allem Empfehlungen und Vorgaben der Länder ab, die sich darauf beziehen, welche Verfahren bei den Sprachstandsfeststellungen eingesetzt werden sollen, die (zumeist) verpflichtend für die Kinder eines bestimmten Alters erfolgen. Die Berichte geben hingegen keine Auskunft dazu, ob an den Kindertagesstätten (Kitas) ergänzend zu den landesweit durchgeführten Sprachstandsfeststellungen noch weitere Verfahren genutzt werden. Diesbezügliche Angaben finden sich aber in einer Studie von Geist und Voet Cornelli (2015), in der in einer direkten Befragung an Kita in Hessen ermittelt wurde, dass im Elementarbereich eine große Vielfalt an Verfahren zum Einsatz kommt und häufig Instrumente genutzt werden, die „hausintern" entwickelt wurden.

Ein bundesweiter Überblick zum Einsatz sprachdiagnostischer Verfahren im *Primarbereich* findet sich in einem Bericht des Hamburger Zentrums zur Unterstützung der wissenschaftlichen Begleitung und Erforschung schulischer Entwicklungsprozesse (ZUSE) (Redder et al., 2011). Dieser Bericht umfasst unter anderem Steckbriefe zu 21 Verfahren, die als für die Primarstufe geeignet klassifiziert wurden. Methodisch basiert auch dieser Bericht auf einer Befragung der jeweils zuständigen Administrationen der Länder und ergänzenden Literaturrecherchen. Er bildet also primär ab, welche sprachdiagnostischen Verfahren die zuständigen Behörden der Länder empfehlen oder vorgeben, nicht jedoch, welche Instrumente tatsächlich an den Schulen genutzt werden.

Eine direkte Befragung an Grundschulen zu den eingesetzten sprachdiagnostischen Verfahren erfolgte in einer Studie von Geist (2014). Die dort ermittelten Ergebnisse dokumentieren, dass auch an Grundschulen eine große Vielfalt an Instrumenten genutzt wird und ebenfalls oftmals hausinterne Verfahren zum Einsatz kommen. Angemerkt sei, dass die Befragung nur auf das Bundesland Hessen beschränkt war und sich de facto auf den Elementarbereich bezog, da ausschließlich der Zeitpunkt der Schulanmeldung und die (in diesem Land im Verantwortungsbereich der Grundschulen liegende) vorschulische Sprachförderung in sogenannten Vorlaufkursen betrachtet wurden.

Anforderungen an die Diagnostik

Vor allem für den Elementarbereich sind mehrere Kriterienkataloge mit Anforderungen an Sprachstandsfeststellungen formuliert worden (z. B. Ehlich, 2005; Lengyel, 2012), die sich in vielen Aspekten für den Primarbereich adaptieren lassen. Die Anforderungen beinhalten unter anderem, dass die eingesetzten diagnostischen Verfahren testtheoretischen Haupt- und Nebengütekriterien genügen sollten (u. a. Objektivität,

Reliabilität, Validität, Normierung, vgl. Moosbrugger & Kelava, 2012). Zentral ist ferner die Forderung, das System Sprache differenziert zu betrachten, also die verschiedenen sprachlichen Ebenen (Phonetik, Phonologie, Morphologie, Syntax, Lexikon, Semantik, Pragmatik) zu berücksichtigen und dabei außerdem zwischen den Aspekten der Produktion und Rezeptionen zu unterscheiden. Für die Diagnostik im Primarbereich ist diese Differenzierung noch um schriftsprachliche Kompetenzen zu erweitern, die normativ spätestens ab dem Schuleintritt erworben werden.

Eine Betrachtung der im Bericht des ZUSE-Instituts aufgeführten diagnostischen Verfahren legt allerdings die Vermutung nahe, dass die Sprachdiagnostik im Primarbereich vor allem auf schriftliche Sprachkompetenzen fokussiert, während mündliche Sprachkompetenzen nur in wenigen der dort genannten Verfahren betrachtet werden (Redder et al., 2011). Tatsächlich wird in der Forschungsliteratur eine Diagnostik mündlicher Sprachkompetenzen im Primarbereich zumeist nur für bestimmte Schülergruppen, für Kinder mit sonderpädagogischem Förderbedarf (z. B. Braun, 2005) oder mit nichtdeutscher Herkunftssprache (z. B. Gogolin et al., 2011; Jeuk, 2009), thematisiert. Die Fokussierung auf Schülerinnen und Schüler mit nichtdeutscher Herkunftssprache wird damit begründet, dass der Zweitspracherwerb mehrsprachiger Kinder von dem einsprachiger Kinder differiert und dieser Unterschied maßgeblich durch sprachbiografische Faktoren (Alter bei Erwerbsbeginn, Kontaktdauer mit der deutschen Sprache) bestimmt wird (Schulz, 2013). Vor diesem Hintergrund werden in der Literatur spezifische Anforderungen an die Diagnostik mehrsprachiger Kinder formuliert. Diese betreffen etwa die Berücksichtigung der Sprachbiografie bzw. die Differenzierung von daraus abgeleiteten Spracherwerbstypen oder der Kompetenzen in der Erstsprache (Lengyel, 2012; Lüdtke & Kallmeyer, 2007), zum Teil werden auch separate Normen gefordert (Schulz, 2013).

Fragestellungen

Das Ziel des vorliegenden Beitrags besteht darin, den aktuellen Kenntnisstand zu den an Grundschulen für die Feststellung sprachlichen Förderbedarfs eingesetzten diagnostischen Verfahren zu erweitern. Während bisherige Arbeiten nur auf Befragungen von Landesbehörden und ergänzenden Literaturrecherchen basierten (vgl. Redder et al., 2011) oder nur ein einzelnes Bundesland betrachteten (vgl. Geist, 2014), wird hierbei auf

Fragebogendaten zurückgegriffen, die *bundesweit* direkt an Grundschulen erhoben wurden. Im Beitrag werden folgende Fragstellungen bearbeitet:

Welche diagnostischen Verfahren werden bundesweit an den Grundschulen eingesetzt, um Sprachförderbedarf festzustellen?

Wie häufig bzw. mit welchem Verbreitungsgrad werden diese Instrumente genutzt?

Um für Bildungspolitik und -forschung Desiderata aufzeigen und kritische Entwicklungen abbilden zu können, soll ferner untersucht werden, inwiefern die ermittelten Verfahren

- für Kinder im Grundschulalter entwickelt wurden,
- für Kinder mit nichtdeutscher Herkunftssprache konzipiert wurden und die diesbezüglichen diagnostischen Anforderungen berücksichtigen,
- mündliche oder schriftliche Sprachkompetenzen erfassen,
- testtheoretischen Gütekriterien genügen.

Methode

Die Grundlage der Analysen bilden Daten der IQB-Ländervergleichsstudie in der Primarstufe (Stanat, Pant, Böhme & Richter, 2012). Diese im Jahr 2011 deutschlandweit durchgeführte Untersuchung zielte primär darauf ab, das Erreichen der für alle Länder verbindlichen Bildungsstandards in den Fächern Deutsch und Mathematik für die Primarstufe (KMK, 2005a, 2005b) zu überprüfen und Bereiche mit besonderem Steuerungsbedarf zu identifizieren.

Das Kernstück der IQB-Ländervergleichsstudie 2011 bildeten auf den Bildungsstandards basierende Kompetenztests, die Aufgaben zu den Fächern Deutsch und Mathematik umfassten. Weiterhin wurden mehrere Fragebögen eingesetzt, die der Erfassung relevanter Hintergrundinformationen dienten und den an der Studie teilnehmenden Kindern, ihren Eltern, den Lehrkräften und den Schulleitungen vorgelegt wurden. Zur Beantwortung der oben genannten Forschungsfragen wurden insbesondere die Angaben der Schulleiterinnen und Schulleiter herangezogen.

Zur Auswahl der zu untersuchenden Schulen wurde für jedes Bundesland eine Zufallsstichprobe aus der Gesamtheit aller in den Schulverzeichnissen der zuständigen Ministerien gelisteten Schulen gezogen. Insgesamt wurde eine Schulstichprobe von 1349 Grund- und Förderschulen realisiert. Den Schulleiterfragebogen bearbeiteten 1272

Leiterinnen und Leiter dieser Schulen, was einer Teilnahmequote von 94.3 % entspricht. Da sich Grund- und Förderschulen hinsichtlich der Häufigkeit und Spezifik von Lernbeeinträchtigungen ihrer Schülerinnen und Schüler einerseits und den daraus resultierenden Anforderungen an die Sprachdiagnostik und -förderung andererseits erheblich voneinander unterscheiden, wurden Förderschulen aus den Analysen ausgeschlossen. Die nach diesem Ausschluss resultierende Stichprobe umfasste 1227 Schulleiterinnen und Schulleiter. Diese waren zu 68.2 % weiblich, im Durchschnitt 52.9 Jahre alt und seit 9.3 Jahren in ihrer Position tätig. Ihre Schulen wurden im Mittel von 259 Schülerinnen und Schülern besucht (SD = 150, MIN = 12, MAX = 1752), die zu 50.6 % weiblich und im Durchschnitt 10.5 Jahre alt waren.

Angaben zur Feststellung des sprachlichen Förderbedarfs und den dabei genutzten Informationen und Verfahren wurden mit einem Schulleiterfragebogen erhoben. Hierzu diente die Frage „Wie wird bestimmt, ob eine Schülerin/ein Schüler Ihrer Schule einen sprachlichen Förderbedarf hat?“. Zur Beantwortung standen folgende Alternativen zur Auswahl, wobei Mehrfachantworten möglich waren:

- „eigene Beobachtung (der Lehrkraft)“
- „Informationen anderer Pädagogen (Hort, Kolleginnen und Kollegen, Erzieherinnen und Erzieher)“
- „Informationen aus schulärztlicher Untersuchung“
- „Informationen der Eltern“
- „schwache Noten“
- „standardisierte Tests (z. B. Hamburger Schreibprobe, MSVK, HASE, BISC)“ und
- „unstandardisierte Verfahren (z. B. Beobachtungsverfahren SISMIK, SELDAK, Lernausgangsanalysen)“

Ergänzend konnten die Schulleiterinnen und Schulleiter in zwei weiteren Items mit offenem Antwortformat angeben, welche konkreten „standardisierten“ und „unstandardisierten“ Verfahren an ihren Schulen zum Einsatz kommen.

Entsprechend der in diesem Beitrag bearbeiteten Fragestellungen wurden zunächst die beiden geschlossenen Items zum Einsatz „standardisierter Tests“ und „unstandardisierter Verfahren“ deskriptiv ausgewertet. Zur anschließenden Analyse der beiden offenen Items erfolgte eine Aufbereitung der schriftlichen Angaben der Schulleitungen zu den eingesetzten Instrumenten. Hierbei wurde jede einzelne Antwort gesichtet und um Hintergrundinformationen zum jeweiligen Verfahren ergänzt, die mithilfe von Recherchen im Fachinformationssystem (FIS) Bildung des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) und in der Fachdatenbank PSYNDEX des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID) erhoben wurden. Die Hauptfunktion dieser Recherche bestand darin, Informationen zu folgenden Fragen zu erfassen:

- Ist das Verfahren publiziert, ist seine Existenz in der Literatur und/oder im Internet dokumentiert?
- Ist das Verfahren korrekt als standardisiert oder unstandardisiert eingeordnet worden?
- Dient das Verfahren der Erfassung sprachlicher Kompetenzen?
- Wurde das Verfahren für Kinder im Grundschulalter entwickelt?
- Wurde das Verfahren auch für Kinder mit nichtdeutscher Herkunftssprache konzipiert?
- Erfasst das Verfahren mündliche oder schriftliche Sprachkompetenzen?

Die weitere Datenauswertung orientierte sich an den von Mayring (1993) beschriebenen Schritten einer Häufigkeits- bzw. Frequenzanalyse, die ein Spezialfall der quantitativen Inhaltsanalyse ist. Am Anfang der Analyse stand die Aufstellung eines Kategoriensystems vor dem Hintergrund der oben genannten Fragen. Dieses Kategoriensystem umfasste die Hauptkategorien „Verfahren“, „Dokumentation“, „Standardisierung“, „Erfassung sprachlicher Kompetenzen“, „Grundschulalter“, „Fokus auf Kinder nichtdeutscher Herkunftssprache“ und „mündliche oder schriftliche Sprachkompetenzen“, die jeweils mit Unterkategorien untersetzt waren. Die Unterkategorien bildeten die Analyseeinheiten für die Kodierung der Angaben der Schulleitungen zu den eingesetzten Verfahren. Im Anschluss an die Kodierung wurde die

Häufigkeit ermittelt, mit der die einzelnen Unterkategorien auftraten. Die daraus resultierenden Verteilungen wurden abschließend deskriptiv ausgewertet. In einer Fortführung und Vertiefung der skizzierten Datenbankrecherche wurden für die fünf jeweils am häufigsten genannten standardisierten und unstandardisierten Verfahren zusätzliche Angaben zu testtheoretischen Gütekriterien ermittelt.

Ergebnisse

Insgesamt wurde von 734 Schulleitungen angegeben, dass zur Feststellung des Sprachförderbedarfs an ihren Schulen standardisierte Verfahren genutzt werden. Das offene Item zu den konkret eingesetzten standardisierten Instrumenten wurde in 335 Fällen, also mit einer Bearbeitungsquote von 45.6 % beantwortet. Die Verwendung unstandardisierter Verfahren wurde von 311 Schulleiterinnen und Schulleitern berichtet; 167 von ihnen beantworteten das zugehörige offene Item (Bearbeitungsquote: 53.7 %). Eine detaillierte Übersicht zur realisierten Schulstichprobe findet sich in Tabelle 6.1 (Elektronisches Supplement 1).

Ausführliche Ergebnisse zu den an den Grundschulen eingesetzten *standardisierten* Verfahren können der Tabelle 6.2 (Elektronisches Supplement 1) entnommen werden. Insgesamt wurden 50 verschiedene standardisierte Instrumente angegeben. Viele dieser Verfahren wurden an nur sehr wenigen Schulen genutzt, lediglich 15 Instrumente wurden von jeweils fünf oder mehr Schulleitungen genannt. Mit Abstand am häufigsten fand die Hamburger Schreibprobe (HSP) Erwähnung, deren Nutzung rund zwei Drittel (66.9 %) der Schulleiterinnen und Schulleiter berichteten, die das offene Item zu den eingesetzten Instrumenten bearbeiteten. Hervorzuheben ist, dass dieser Rechtschreibtest an mehr als der Hälfte dieser Schulen (117 von 224) das einzige Verfahren ist, das laut Angabe der Schulleitungen zur Identifikation von Schülerinnen und Schülern mit sprachlichem Förderbedarf eingesetzt wurde. Ebenfalls relativ häufig wurde der Stolperwörter-Lesetest (12.8 %) genannt. Vergleichsweise oft wurden außerdem HAVAS 5 (7.4 %), der Cito-Sprachtest (5.7 %), die Diagnostischen Bilderlisten (DBL) und der Deutsche Rechtschreibtest (DRT) (jeweils 5.1 %) angegeben.

Ausführliche Ergebnisse zu den an den Grundschulen eingesetzten *unstandardisierten* Verfahren können der Tabelle 6.3 (Elektronisches Supplement 1) entnommen werden. Der höchste Anteil von Nennungen fällt hier mit 26.4 % auf die Kategorie „nicht näher spezifizierte Lernausgangsanalysen“, in der alle Antworten

zusammengefasst sind, bei denen zwar die Durchführung von Lernausgangsanalysen angegeben, aber kein konkretes Verfahren genannt wurde. In ähnlicher Weise umfasst die Kategorie „Eigenentwicklung“ (12.6 %) alle Antworten, die eine Nutzung selbst konstruierter Instrumente umfassten. Von den verbleibenden 17 unstandardisierten Verfahren wurde der Beobachtungsbogen Sismik am häufigsten genannt (22.2 %). Relativ oft wurde außerdem die Nutzung von LauBe (12.0 %), ILeA (11.4 %), Mirola (9.6 %) und Seldak (3.0 %) angegeben.

Die Schulleitungen berichteten ferner den Einsatz von 13 weiteren standardisierten und unstandardisierten Verfahren, die nicht explizit für die Diagnose sprachlichen Förderbedarfs entwickelt wurden. Hierunter fallen unter anderem Intelligenztests und Verfahren zur Erfassung mathematischer Kompetenzen. Eine Verwendung solcher Verfahren im Rahmen der Sprachstandsdiagnostik wurde allerdings sehr selten (in 16 Fällen) berichtet. Darüber hinaus wurden 20 weitere Verfahren angegeben, zu denen keine Hintergrundinformationen recherchiert werden konnten und die deswegen als „nicht kodierbar“ klassifiziert wurden.

Anhand von Tabelle 6.2 und Tabelle 6.3 ist außerdem ersichtlich, dass mehrheitlich Verfahren eingesetzt werden, die entweder für Kinder im *Grundschulalter* entwickelt wurden oder zumindest eine Altersspanne anvisieren, die neben dem Elementarbereich auch die Zeit der ersten Grundschulphase umfasst. Allerdings finden sich selbst unter den häufiger genannten Verfahren auch Instrumente, die vor allem für den Elementarbereich bzw. für die Vorschulphase konzipiert sind (z. B. BISC, HASE, Sismik, Seldak, auch HAVAS 5 und Cito-Sprachtest).

Eine explizite Fokussierung auf *Kinder mit nichtdeutscher Herkunftssprache* erfolgt nur in wenigen Fällen, wobei die hierzu genannten Verfahren oftmals nicht für das Grundschulalter entwickelt wurden und bestenfalls ganz zu Beginn der Grundschulzeit eingesetzt werden können (z. B. HAVAS 5 und Cito-Sprachtest, vgl. Tab. 2 und 3). Oftmals beschränken sich die Verfahren auf die Erfassung sprachbiografischer Informationen (z. B. Fit in Deutsch), eine Feststellung von Kompetenzen in der Erst- und Zweitsprache erfolgt bei HAVAS 5 und im Cito-Sprachtest. Das letztgenannte Verfahren beinhaltet zudem separate Normwerte für Kinder mit nichtdeutscher Herkunftssprache, wobei jedoch nicht nach Spracherwerbstypen differenziert wird.

Ein Großteil der genannten Verfahren zielt auf die *Erfassung von schriftlichen Sprachkompetenzen* (z. B. DRT) oder deren Vorläuferfähigkeiten (z. B. BISC, vgl. Tab. 2 und 3). Auch die Einsatzhäufigkeit der einzelnen Verfahren belegt die starke Konzentration auf die Schriftsprache (z. B. HSP, Stolperwörter-Lesetest). Die Verfahren, die mündliche Sprachkompetenzen erfassen, sind entweder nicht für das Grundschulalter, sondern für den Elementarbereich konzipiert (z. B. Delfin), oder fokussieren insbesondere auf Mehrsprachigkeit (z. B. SFD, Tulpenbeet) oder Sprachentwicklungsstörungen (z. B. WWT).

Angaben dazu, inwieweit die am häufigsten genannten Verfahren hinsichtlich elementarer *testtheoretischer Gütekriterien* genügen, sind im Detail in Tabelle 6.4 und Tabelle 6.5 im Elektronischen Supplement 2 zu finden und lassen sich wie folgt zusammenfassen: Für HAVAS 5 wird aufgrund der komplexen Auswertung und zum Teil uneindeutiger Anweisungen trotz vorgeschriebener Schulungen eine geringe Auswertungsobjektivität vermutet (Neugebauer & Becker-Mrotzek, 2013), wobei hierzu keine Studien vorliegen. Zufriedenstellende Werte für die Beobachterübereinstimmung werden für Seldak (und Sismik) berichtet. Zu den anderen hier als unstandardisiert klassifizierten Verfahren (LauBe, Mirola, ILeA) finden sich keine Angaben zur Objektivität – und auch nicht zu den Testgütekriterien der Reliabilität und Validität. In vielen Fällen wird nur die interne Konsistenz als Reliabilitätsschätzer angegeben, der für HAVAS 5 als Gesamtverfahren gering ausfällt. Der für eine längsschnittliche Diagnostik wichtige Retestparameter wird nur in Ausnahmefällen berichtet (HSP, Stolperwörter). Die Validität wird zumeist durch Angabe von Untersuchungsergebnissen zur Kriteriumsvalidität (HSP, Stolperwörter-Lesetest, Cito-Sprachtest, DBL, Seldak, Sismik) und in Einzelfällen zur strukturellen Validität belegt (Cito-Sprachtest, HAVAS 5). Normen liegen für die meisten Verfahren vor. Beim Stolperwörter-Lesetest stammen diese jedoch von einer kleinen, nicht repräsentativen Ad-hoc-Stichprobe. Die Normen der Diagnostischen Bilderlisten sind stark veraltet (aus 1993). Bei ILeA werden lediglich zu erwartende Durchschnittsleistungen dargestellt. Keine Angaben zu Normen finden sich für LauBe und Mirola.

Diskussion

Belege dafür, dass an Grundschulen eine Vielzahl verschiedener diagnostischer Verfahren zum Einsatz kommt, konnten bereits in einer 2014 von Geist veröffentlichten Studie

vorgelegt werden, die allerdings nur Schulen in Hessen berücksichtigte und zudem den Zeitpunkt der Schulanmeldung sowie vor Schuleintritt angebotene Vorlaufkurse betrachtete. Im vorliegenden Beitrag wurde die große Heterogenität der eingesetzten Verfahren nun auch anhand einer bundesweiten Befragung von Schulleitungen und mit Fokus auf die Feststellung sprachlichen Förderbedarfs bei Grundschülerinnen und Grundschülern aufgezeigt. Hierbei konnte ebenfalls belegt werden, dass häufig auch hausinterne Verfahren genutzt werden. Zudem wurde der höchste Verbreitungsgrad für zwei standardisierte Verfahren (HSP und Stolperwörter-Lesetest) ermittelt, die laut den Angaben bei Redder und Kollegen (2011) nicht oder nur vereinzelt (im Fall der HSP nur in Baden-Württemberg) Bestandteil von Vorgaben und Empfehlungen der Länder sind. Nachfolgend sollen die weiteren Ergebnisse der Studie im Hinblick auf die weiter oben genannten Aspekte der Eignung für das Grundschulalter, der Fokussierung auf Kinder nichtdeutscher Herkunftssprache, der Erfassung von mündlichen und schriftlichen Sprachkompetenzen und der testtheoretischen Güte resümiert werden.

Nicht alle der von den Schulleitungen genannten Verfahren sind für das *Grundschulalter* konzipiert, sondern wurden explizit für den Elementarbereich entwickelt. Es bleibt allerdings unklar, ob diese Verfahren tatsächlich in dem hier festgestellten Umfang genutzt werden, um den sprachlichen Förderbedarf von Grundschulkindern zu ermitteln. Möglicherweise beziehen sich die Angaben der betreffenden Schulleitungen zum Teil auch auf die in vorschulischen Untersuchungen eingesetzten Verfahren, die mancherorts in der Verantwortung der Grundschulen liegen.

Angesichts eines in der IQB-Ländervergleichsstudie 2011 für die 4. Jahrgangsstufe bundesweit ermittelten Anteils von Schülerinnen und Schülern mit Zuwanderungshintergrund von rund 25 % (Haag, Böhme & Stanat, 2012) ist plausibel, dass die befragten Grundschulen von einer Vielzahl von *Kindern mit nichtdeutscher Herkunftssprache* besucht werden. Die Ergebnisse dieses Beitrags lassen jedoch vermuten, dass speziell für diese Schülergruppe im Grundschulalter insgesamt nur wenige Verfahren zur Verfügung stehen. Möglicherweise erklärt dieser Mangel auch die relativ häufige Nennung des Cito-Sprachtests und von HAVAS 5, die eigentlich nicht für das Grundschulalter konzipiert sind und sich bestenfalls für den Beginn der Grundschulzeit eignen. Bereits an anderer Stelle konnten wir zeigen, dass Schulen mit einem hohen Anteil mehrsprachiger Kinder tendenziell häufiger Verfahren nutzen, die auf Kinder mit nichtdeutscher Herkunftssprache zielen (Hoffmann & Böhme, im Druck).

Die Ergebnisse des Beitrags verdeutlichen außerdem die starke Fokussierung der Sprachdiagnostik in der Grundschule auf *schriftliche Sprachkompetenzen*. Hierfür lassen sich mehrere Gründe vermuten:

(1) Die Fokussierung spiegelt wider, dass die Vermittlung schriftsprachlicher Kompetenzen als originärer Auftrag der Beschulung im Primarbereich wahrgenommen wird. Der Förderung mündlicher Kompetenzen wird hingegen, ungeachtet der Verankerung des Kompetenzbereichs „Sprechen und Zuhören" in den Bildungsstandards (KMK, 2005a), weniger Aufmerksamkeit geschenkt (vgl. Böhme, 2012).

(2) Es wird offenbar davon ausgegangen, dass bei Schuleintritt alle Kinder über ein hinreichendes Kompetenzniveau verfügen, sodass eine Diagnostik mündlicher Sprachkompetenzen als nicht notwendig erachtet wird. Diese Annahme ist jedoch nicht haltbar: Zum einen ist zu bedenken, dass einige Schülerinnen und Schüler mit nichtdeutscher Herkunftssprache aufgrund eines späten Erwerbsbeginns und einer geringen Kontaktdauer mit der deutschen Sprache noch nicht über die erforderlichen Kompetenzen verfügen könnten. Zum anderen ist auf die Besonderheiten des unterrichtsspezifischen Gebrauchs der deutschen Sprache hinzuweisen, der durch ein formelles sprachliches Register – die sogenannte Bildungssprache – gekennzeichnet ist, das sich deutlich von der Alltagssprache abhebt (vgl. Gogolin & Lange, 2011). Neben einem späten Erwerbsbeginn und einer geringen Kontaktdauer können weitere, sowohl ein- als auch mehrsprachige Kinder betreffende Faktoren (z. B. bildungsferne Herkunft, schlechte Sprachvorbilder) dazu führen, dass die für den schulischen Lernerfolg relevanten bildungssprachlichen Fähigkeiten nicht im ausreichenden Maße entwickelt sind (Chall, Jacobs & Baldwin, 1990) und eine entsprechende Förderung notwendig wird. Im Übrigen können diesbezüglich bestehende Risiken bei den im Elementarbereich verankerten Sprachstandsfestellungen vielfach nicht aufgedeckt werden, da diese zumeist auf alltagssprachliche Fähigkeiten zielen.

(3) Es mangelt an geeigneten Verfahren zur Diagnostik lautsprachlicher und bildungssprachlicher Kompetenzen im Primarbereich (vgl. Uesseler, Runge & Redder, 2013).

Die Recherchen zur testtheoretischen Güte der am häufigsten eingesetzten Verfahren ergaben, dass an den Grundschulen zum Teil Instrumente genutzt werden, deren

Objektivität, Reliabilität und Validität gering oder unbekannt sind. Auch finden sich unter den häufigsten Nennungen Instrumente, zu denen nur stark veraltete, unbrauchbare oder überhaupt keine Normen vorliegen. Dieses Ergebnis kann auch als Indiz interpretiert werden, dass bei den Lehrkräften an Grundschulen Unsicherheiten darüber bestehen, wie die Qualität und Eignung einzelner Verfahren einzuschätzen ist. Hieraus lässt sich mit Blick auf die Aus- und Fortbildung von Lehrkräften ableiten, dass eine Stärkung der (sprach-) diagnostischen Kompetenzen wünschenswert ist (z. B. Geist, 2014; Wildemann, 2010).

Die vorliegende Studie ist mit in einer Reihe von Einschränkungen verbunden. So ist zu vermuten, dass die Formulierung der beiden offenen Items aus dem Schulleiterfragebogen zu den genutzten diagnostischen Verfahren nicht immer im intendierten Sinne verstanden wurde: Im Zuge der Auswertung dieser Items zeigte sich, dass die Schulleiterinnen und Schulleiter zum Teil Schwierigkeiten mit der Differenzierung von „standardisiert“ und „unstandardisiert“ hatten. Zudem ist eine eindeutige dichotome Zuordnung bei einigen Grenzfällen kaum möglich. Daher wird mit Blick auf zukünftige Erhebungen empfohlen, auf eine nach Standardisierungsgrad differenzierte Erfassung der Verfahren zu verzichten. Darüber hinaus ist darauf hinzuweisen, dass im Itemstamm des offenen Items zu den eingesetzten standardisierten Verfahren die HSP exemplarisch aufgeführt wurde. Folglich ist nicht auszuschließen, dass der hohe Verbreitungsgrad, der für diesen Test registriert wurde, zum Teil aus der gewählten Itemformulierung resultiert, da die HSP als mögliches Verfahren in besonderer Weise ins Bewusstsein gerückt wurde. Dies gilt allerdings auch für die weiteren beispielhaft im Itemstamm genannten Verfahren, für die jedoch keine gehäuften Nennungen zu verzeichnen waren. Denkbar ist schließlich, dass Schulleitungen die Praxis der Sprachdiagnostik an ihren Schulen nicht in jedem Fall vollumfänglich überblicken. In zukünftigen Studien erscheint es lohnenswerter, stattdessen die direkt beteiligten Akteure zu befragen (z. B. verantwortliche Lehrkräfte, Sprachförder-AGs).

Angemerkt sei, dass an Grundschulen neben den hier fokussierten diagnostischen Verfahren in der Regel noch weitere Informationsquellen genutzt werden, um den sprachlichen Förderbedarf von Schülerinnen und Schülern zu bestimmen. So zeigen weitere, bei Stanat, Weirich und Radmann (2012) dokumentierte Ergebnisse der IQB-Ländervergleichsstudie 2011, dass hierbei insbesondere auch auf die eigenen Beobachtungen von Lehrkräften und Erzieherinnen und Erziehern zurückgegriffen wird.

Abschließend möchten wir darauf hinweisen, dass im vorliegenden Beitrag betrachtet wurde, welche Verfahren an den Grundschulen zur Feststellung des Sprachförderbedarfs eingesetzt werden, nicht jedoch, wie die jeweils erhobenen diagnostischen Informationen zu Förderentscheidungen weiterverarbeitet oder für die additiven Sprachförderung genutzt werden. Tatsächlich zielen nicht alle der von den Schulleitungen genannten Verfahren auf eine Selektionsdiagnostik. Beispielsweise versteht sich HAVAS 5 als ein Instrument für die Förderdiagnostik, das jedoch in der Praxis entgegen dieser Intention auch zur Selektionsdiagnostik eingesetzt wird (vgl. Lüdtke & Kallmeyer, 2007).

Literaturverzeichnis

Autorengruppe Bildungsberichterstattung. (2012). *Bildung in Deutschland 2012. Ein indikatorengestützter Bericht mit einer Analyse zur kulturellen Bildung im Lebenslauf*. Bielefeld: Bertelsmann.

Bates, C. & Nettelbeck, T. (2001). Primary School Teachers' Judgements of Reading Achievement. *Educational Psychology, 21*(2), 177–187.

Baumert, J. & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 323–410). Opladen: Leske + Budrich.

Böhme, K. (2012). *Methodische und didaktische Überlegungen sowie empirische Befunde zur Erfassung sprachlicher Kompetenzen im Deutschen. Analysen zu den Bildungsstandards im Fach Deutsch für den Primarbereich*. Humboldt Universität zu Berlin, Philosophische Fakultät IV, Berlin.

Braun, O. (2005). *Sprachstörungen bei Kindern und Jugendlichen: Diagnostik-Therapie-Förderung*. Stuttgart: Kohlhammer.

Chall, J. S., Jacobs, V. A. & Baldwin, L. E. (1990). *The Reading Crisis: Why Poor Children Fall Behind*. Cambridge, Massachusetts: Harvard University Press.

Ehlich, K. (Hrsg.). (2005). *Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Förderung von Kindern mit und ohne Migrationshintergrund*. Berlin: BMBF.

Feinberg, A. B. & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18*(1), 52–65.

Geist, B. (2014). *Sprachdiagnostische Kompetenz von Sprachförderkräften*. Berlin: De Gruyter.

Geist, B. & Voet Cornelli, B. (2015). *Sprachdiagnostik mehrsprachiger Kinder im Elementarbereich*. In M. Urban, M. Schulz, K. Meser & S. Thoms (Hrsg.), Inklusion und Übergang. Perspektiven der Vernetzung von

Kindertageseinrichtungen und Grundschulen (S. 248–270). Bad Heilbrunn: Julius Klinkhardt.

Gogolin, I., Dirim, I., Klinger, T., Lange, I., Lengyel, D., Michel, U., Neumann, U., Reich, H. H., Roth, H.-J. & Schwippert, K. (2011). *Förderung von Kindern und Jugendlichen mit Migrationshintergrund FörMig. Bilanz und Perspektiven eines Modellprogramms*. Münster: Waxmann.

Gogolin, I. & Lange, I. (2011). Bildungssprache und Durchgängige Sprachbildung. In S. Fürstenau & M. Gomolla (Hrsg.), *Migration und schulischer Wandel: Mehrsprachigkeit* (S. 107–128). Wiesbaden: VS Verlag für Sozialwissenschaften.

Haag, N., Böhme, K. & Stanat, P. (2012). Zuwanderungsbezogene Disparitäten. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 209–236). Münster: Waxmann.

Hoffmann, L. & Böhme, K. (im Druck). Wird sprachlicher Förderbedarf in der Grundschule sicher erkannt? Zur Klassifikationsgüte von diagnostischen Entscheidungen. *Zeitschrift für Pädagogische Psychologie*.

Jeuk, S. (2009). Aktuelle Verfahren zur Einschätzung des Stands der Sprachaneignung bei mehrsprachigen Kindern im Grundschulalter. In S. Jeuk & I. Schmid-Barkow (Hrsg.), *Differenzen diagnostizieren und Kompetenzen fördern im Deutschunterricht* (S. 61–82). Freiburg: Fillibach.

KMK. (2005a). *Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004*. München: Luchterhand.

KMK. (2005b). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004*. München: Luchterhand.

Lengyel, D. (2012). *Sprachstandsfeststellung bei mehrsprachigen Kindern im Elementarbereich. Eine Expertise der Weiterbildungsinitiative Frühpädagogischer Fachkräfte* (WiFF). München: DJI.

Lisker, A. (2013). *Sprachstandserhebung und Sprachförderung vor der Einschulung. Eine Bestandsaufnahme in den Bundesländern. Expertise im Auftrag des Deutschen Jugendinstituts. Aktualisierung der Expertise von 2010*. München: DJI.

Lüdtke, U. M. & Kallmeyer, K. (2007). Kritische Analyse ausgewählter Sprachstandserhebungsverfahren für Kinder vor Schuleintritt aus Sicht der Linguistik, Diagnostik und Mehrsprachigkeitsforschung. *Die Sprachheilarbeit*, 2007(6), 261–278.

Mayring, P. (1993). *Qualitative Inhaltsanalyse – Grundlagen und Techniken* (4., erweiterte Aufl.). Weinheim: Deutscher Studien Verlag.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Berlin: Springer.

Neugebauer, U. & Becker-Mrotzek, M. (2013). *Die Qualität von Sprachstandsverfahren im Elementarbereich. Eine Analyse und Bewertung*. Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache.

Redder, A., Schwippert, K., Hasselhorn, M., Forschner, S., Fickermann, D. & Ehrlich, K. (2011). *Bilanz und Konzeptualisierung von strukturierter Forschung zu "Sprachdiagnostik und Sprachförderung"*. Hamburg: ZUSE.

Schulz, P. (2013). Sprachdiagnostik bei mehrsprachigen *Kindern. Sprache,·Stimme, Gehör, 37*(4), 191–195.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

Stanat, P., Weirich, S. & Radmann, S. (2012). Sprach- und Leseförderung. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 251–276). Münster: Waxmann.

Uesseler, S., Runge, A. & Redder, A. (2013). "Bildungssprache" diagnostizieren. Entwicklung eines Instruments zur Erfassung von bildungssprachlichen Fähigkeiten bei Viert- und Fünftklässlern. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven* (S. 42–67). Münster: Waxmann.

von Aufschnaiter, C., Cappell, J., Dübbelde, G., Ennemoser, M., Mayer, J., Stiensmeier-Pelster, J., Sträßer, R. & Wolgast, A. (2015). Diagnostische Kompetenz: Theoretische Überlegungen zu einem zentralen Konstrukt der Lehrerbildung. *Zeitschrift für Pädagogik, 61*(5), 738–758.

Wildemann, A. (2010). Sprachdiagnostisches Wissen angehender Deutschlehrkräfte – Annäherungen zwischen Utopie und Wirklichkeit. In J. König & B. Hofmann (Hrsg.), Lehrerprofessionalität. *Was sollen Lehrkräfte im Lese- und Schreibunterricht wissen und können?* (S. 178–194). DGLS: Berlin.

Elektronische Supplemente 1

Tabelle 6.1: Schulstichprobe nach Bundesländern

Bundesland	Anzahl der Schulen in Stichprobe der IQB-Ländervergleichs-studie 2011	Anzahl der Schulen, in denen standardisierte Verfahren eingesetzt wurden	Anzahl der Schulen, in der konkrete Angaben zu den eingesetzten standardisierten Verfahren gemacht wurden	Anzahl der Schulen, in denen unstandardisierte Verfahren eingesetzt wurden	Anzahl der Schulen, in der konkrete Angaben zu den eingesetzten unstandardisierten Verfahren gemacht wurden
Baden Württemberg	70	56	24	7	0
Bayern	74	40	13	33	18
Berlin	110	96	44	38	27
Brandenburg	75	43	17	22	16
Bremen	88	73	34	25	15
Hamburg	89	89	53	21	16
Hessen	72	54	24	12	6
Mecklenburg-VP	69	22	12	10	2
Niedersachsen	72	31	20	20	9
Nordrhein-Westfalen	71	55	26	20	10
Rheinland-Pfalz	70	22	10	14	6

Bundesland	Anzahl der Schulen in Stichprobe der IQB-Ländervergleichs-studie 2011	Anzahl der Schulen, in denen standardisierte Verfahren eingesetzt wurden	Anzahl der Schulen, in der konkrete Angaben zu den eingesetzten standardisierten Verfahren gemacht wurden	Anzahl der Schulen, in denen unstandardisierte Verfahren eingesetzt wurden	Anzahl der Schulen, in der konkrete Angaben zu den eingesetzten unstandardisierten Verfahren gemacht wurden
Saarland	75	38	15	4	1
Sachsen	68	27	7	15	7
Sachsen-Anhalt	73	33	13	20	8
Schleswig-Holstein	74	30	11	30	16
Thüringen	77	25	12	20	10
Gesamt	1227	734	335	311	167

Tabelle 6.2: Von den Schulleitungen genannte standardisierte sprachdiagnostische Verfahren (incl. der Häufigkeit der Nennung) (*Nges* = 335)[1]

#	Name des Verfahrens	Häufigkeit (in %)	N	Grund-schulalter[2]	Nicht-deutsche Herkunfts-sprache	Schrift-sprachliche Kompetenzen[3]
1	HSP (Hamburger Schreib-Probe)	66.9	224	X		X
2	Stolperwörter-Lesetest	12.8	43	X		X
3	HAVAS 5 (Hamburger Verfahren zur Analyse des Sprachstandes bei Fünf-Jährigen)	7.4	25		X	
4	Cito-Sprachtest	5.7	19		X	
5	DBL (Diagnostische Bilderlisten)	5.1	17	X		X
6	DRT (Deutscher Rechtschreibtest)	5.1	17	X		X
7	MÜSC (Münsteraner Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten)	3.6	12	(X)		(X)
8	KEKS-Test (Kompetenzerfassung in Kindergarten und Schule)	2.7	9	X	X	X[5]
9	BISC (Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten)	2.1	7			(X)
10	HASE (Heidelberger Auditives Screening in der Einschulungsuntersuchung)	2.1	7			(X)

#	Name des Verfahrens	Häufigkeit (in %)	N	Grund-schulalter[2]	Nicht-deutsche Herkunfts-sprache	Schrift-sprachliche Kompetenzen[3]
11	MSVK (Marburger Sprachverständnistest für Kinder)	2.1	7	(X)		
12	ELFE (Leseverständnistest für Erst- bis Sechstklässler)	1.8	6	X		X
13	Bremer Rechtschreibtest	1.5	5	X		X
14	HAMLET (Hamburger Lesetest)	1.5	5	X		X
15	SFD (Sprachstandsüberprüfung und Förderdiagnostik für Ausländer- und Aussiedlerkinder)	1.5	5	X	X	
16	Bärenstark	1.2	4		X	
17	Fit in Deutsch	1.2	4		X	
18	Salzburger Lesescreening	0.9	3	X		X
19	Salzburger Lese- und Rechtschreibtest	0.9	3	X		X
20	Differenzierungsprobe/Breuer-Weuffen-Test	0.9	3	(X)		

#	Name des Verfahrens	Häufigkeit (in %)	N	Grund-schulalter[2]	Nicht-deutsche Herkunfts-sprache	Schrift-sprachliche Kompetenzen[3]
21	Dysgrammatiker Prüfmaterial	0.6	2	(X)		
22	PB-LRS (Phonologische Bewusstheit bei Kindergartenkindern und Schulanfängern)	0.6	2	(X)		(X)
23	WLLP-R (Würzburger Leiseleseprobe)	0.6	2	X		X
24	HLP (Hamburger Leseprobe)	0.6	2	X		X
25	ARS (Anlaute hören, Reime finden, Silben klatschen)	0.6	2	(X)	X	(X)
26	LPB (Ravensburger Lautprüfbogen)	0.6	2			
27	Deutsch Plus	0.6	2		X	
28	Delfin (Diagnostik, Elternarbeit und Förderung der Sprachkompetenz)	0.6	2			
29	Rechtschreibwerkstatt von Stumpenhorst (incl. Diagnosetest)	0.6	2	X		X
30	BLT (Bremer Lesetest)	0.3	1	X		X

#	Name des Verfahrens	Häufigkeit (in %)	N	Grund-schulalter[2]	Nicht-deutsche Herkunfts-sprache	Schrift-sprachliche Kompetenzen[3]
31	Alfons Diagnostik	0.3	1	X		X
32	H-LAD (Heidelberger Lautdifferenzierungstest)	0.3	1	X		
33	Schulleistungsbatterie für Lernbehinderte	0.3	1	X		
34	ETS 4-8 (Entwicklungstest Sprache 4 bis 8 Jahre)	0.3	1	(X)		
35	HSET (Heidelberger Sprachentwicklungstest)	0.3	1	(X)		
36	Kieler Schreibprobe	0.3	1	X		X
37	C-Test	0.3	1	X	(X)[4]	X
38	Profilanalyse nach Grieshaber	0.3	1	X	X	
39	WET (Wiener Entwicklungstest)	0.3	1			
40	KISTE (Kindersprachtest für das Vorschulalter)	0.3	1			

#	Name des Verfahrens	Häufigkeit (in %)	N	Grund-schulalter[2]	Nicht-deutsche Herkunfts-sprache	Schrift-sprachliche Kompetenzen[3]
41	P-ITPA (Potsdam-Illinois Test für Psycholinguistische Fähigkeiten)	0.3	1	X		
42	Mottier-Test	0.3	1	X	X	
43	WRT (Weingartener Grundwortschatz Rechtschreib-Test)	0.3	1	X		X
44	WWT (Wortschatz- und Wortfindungstest für 6- bis 10-Jährige)	0.3	1	X		
45	Deutsch für den Schulstart	0.3	1	(X)	X	
46	Materialien aus dem Dudenverlag	0.3	1	X		X
47	GSS 8 (Göppinger sprachfreier Schuleignungstest)	0.3	1			
48	PLGT (Potsdamer Lesegeschwindigkeitstest)	0.3	1	X		X
49	Lesefitness	0.3	1	X		X
50	VERA 3	0.3	1	X		X[5]

[1] N_{ges} = Anzahl der Schulleiterinnen und Schulleiter, die das offene Item zu den eingesetzten standardisierten Verfahren bearbeitet haben.

[2] Geklammert sind Verfahren, die auf den Schulbeginn fokussieren oder neben dem Elementarbereich auch zu Beginn des Primarbereichs eingesetzt werden können.

[3] Geklammert sind Verfahren, die explizit auf die Erfassung von Vorläuferfähigkeiten schriftsprachlicher Kompetenz zielen.

[4] Der C-Test wurde ursprünglich für andere Anwendungsbereiche entwickelt (z. B. erwachsene Fremdsprachlernende), wird aber auch bei Kindern mir nichtdeutscher Herkunftssprache eingesetzt (z. B. Scholten-Akoun, D. & Baur, R. S. (2012). Der C-Test als ein Instrument zur Messung der Schriftsprachkompetenzen von Lehramtsstudierenden (auch) mit Migrationshintergrund – eine Studie. In B. Ahrenholz & W. Knapp (Hrsg.), *Sprachstand erheben – Spracherwerb erforschen. Beiträge aus dem 6. Workshop „Kinder mit Migrationshintergrund“* (S. 307–330). Freiburg: Fillibach.

[5] Erfassung mündlicher und schriftlicher Sprachkompetenzen.

Tabelle 6.3: Von den Schulleitungen genannte unstandardisierte sprachdiagnostische Verfahren (incl. der Häufigkeit der Nennung) (N_{ges} = 167)[1]

#	Name des Verfahrens	Häufigkeit (in %)	N	Grund-schulalter[2]	Nicht-deutsche Herkunfts-sprache	Schrift-sprachliche Kompetenzen[3]
1	(nicht näher spezifizierte) Lernausgangsanalysen	26	44	?	?	?
2	Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kindertageseinrichtungen (Sismik)	22	37		X	
3	Eigenentwicklungen	13	21	?	?	?
4	Lernausgangslage Berlin (LauBe)	12	20	X		X
5	Individuelle Lernstandsanalysen in der Grundschule (ILeA)	11	19	X		X
6	Mit Mirola durch den Zauberwald	9.6	16	(X)		
7	Sprachentwicklung und Literacy bei deutschsprachig aufwachsenden Kindern (Seldak)	3	5			
8	Das leere Blatt	1.8	3	(X)		
9	Kieler Einschulungsverfahren	1.8	3			
10	Informelle Schulleistungsdiagnostik IV (SLD IV)	1.2	2	X		X

#	Name des Verfahrens	Häufigkeit (in %)	N	Grund-schulalter[2]	Nicht-deutsche Herkunfts-sprache	Schrift-sprachliche Kompetenzen[3]
1	(nicht näher spezifizierte) Lernausgangsanalysen	26	44	?	?	?
2	Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kindertageseinrichtungen (Sismik)	22	37		X	
3	Eigenentwicklungen	13	21	?	?	?
4	Lernausgangslage Berlin (LauBe)	12	20	X		X
5	Individuelle Lernstandsanalysen in der Grundschule (ILeA)	11	19	X		X
6	Mit Mirola durch den Zauberwald	9.6	16	(X)		
7	Sprachentwicklung und Literacy bei deutschsprachig aufwachsenden Kindern (Seldak)	3	5			
8	Das leere Blatt	1.8	3	(X)		
9	Kieler Einschulungsverfahren	1.8	3			
10	Informelle Schulleistungsdiagnostik IV (SLD IV)	1.2	2	X		X

#	Name des Verfahrens	Häufigkeit (in %)	N	Grund-schulalter[2]	Nicht-deutsche Herkunfts-sprache	Schrift-sprachliche Kompetenzen[3]
11	WESPE (Wir Erzieherinnen schätzen den Sprachstand ein)	1.2	2			
12	Diagnosebogen zur Erfassung von Sprachstörungen (Grunwald)	0.6	1			
13	Unser Schulspiel (Burscheidt)	0.6	1	(X)		
14	Förder-Diagnose-Box (Schroedel)	0.6	1	X		X
15	Lernvoraussetzungen am Schulanfang (Ostermann)	0.6	1	(X)		
16	Beobachtungsverfahren nach Kleinmann	0.6	1	X		
17	Könnerheft	0.6	1	X		X
18	Sprachlerntagebuch	0.6	1			
19	Das Tulpenbeet (FörMig)	0.6	1	X	X	

[1] *Nges* = Anzahl der Schulleiterinnen und Schulleiter, die das offene Item zu den eingesetzten standardisierten Verfahren bearbeitet haben.

[2] Geklammert sind Verfahren, die auf den Schulbeginn fokussieren oder neben dem Elementarbereich auch zu Beginn des Primarbereichs eingesetzt werden können.

[3] Geklammert sind Verfahren, die explizit auf die Erfassung von Vorläuferfähigkeiten schriftsprachlicher Kompetenz zielen.

Elektronische Supplemente 2

Tabelle 6.4: Übersicht der am häufigsten genannten *standardisierten* Verfahren

	HSP[1]	**Stolperwörter-Lesetest**[2]	**HAVAS 5**[3]	**Cito-Sprachtest**[4]	**Diagnostische Bilderlisten**[5]
Diagnostische Zielsetzung	Erfassung von Rechtschreibfähigkeiten	Erfassung von Lesefähigkeiten	Profilanalyse zur kindlichen Sprachproduktion in Deutsch und in der Herkunftssprache	Sprachtest, Fokus auf Zweisprachigkeit	Früherkennung von Schwierigkeiten im Lese-lernprozess; Screening
Alter/Ziel-gruppe	von der Mitte der 1. bis zur 10. Jahrgangsstufe	1. bis 4. Klasse	5 bis 7 Jahre	4 bis 6 Jahre	Ende der 1. Klasse bis zum Beginn der 2. Klasse
Skalen/ Subtests	Zahl richtig geschriebener Wörter und Grapheme; „Lupenstellen“ als Indikator für Grad der Beherrschung untersch. Rechtschreibstrategien; überflüssige orthografische Elemente; Oberzeichenfehler	Lesegeschwindigkeit und -sicherheit auf der Satzebene	Auswertung eines „Gesprächs“ über eine Bildergeschichte z. B. nach Aufgabenbewältigung, Bewältigung der Kommunikationssituation, Wortschatz, Verbstellung, Satzverbindungen	Test mit 4 Komponenten: Passiver Wortschatz, Kognitive Begriffe, Phonologisches Bewusstsein, Textverständnis	Schriftliche Benennung von Bildern durch Kinder, Bestimmung der Anzahl der Fehler, Ermittlung von Fehlerkategorien
Mündliche vs. schriftsprachliche Kompetenzen	Schriftlich	Schriftlich	Mündlich	Mündlich	Schriftlich
Rezeptive vs. produktive Sprachfähigkeiten	Produktiv	Rezeptiv	Fokus auf Produktion, aber (mittelbar) auch rezeptive Fähigkeiten, da Gespräch als Erhebungsmethode	Rezeptiv	Produktion
Durchführung	< 30 min, Einzel- oder Gruppentest	5 bis 8 min, Gruppentest	10 bis 15 min, Einzeltest	25 min, computerbasierter Einzeltest	20 bis 30 min, Gruppentest

	HSP[1]	Stolperwörter-Lesetest[2]	HAVAS 5[3]	Cito-Sprachtest[4]	Diagnostische Bilderlisten[5]
Objektivität	Standardisierung von Durchführung, Auswertung, Interpretation	Keine Angaben, Standardisierung durch Testformat	Komplexe Auswertung, daher geringe Auswertungs-objektivität zu vermuten	Standardisierung von Durchführung, Auswertung, Interpretation	Standardisierung der Durch-führung, keine Überprüfung d. Interrater-Reliabilität
Reliabilität	Interne Konsistenz: α = .92 bis .99; Retest: r_{tt} = .52 bis .93	Reliabilitätsbestimmung im Zuge der LUST-1-Studie[6], $\alpha > .88$, $r_{tt} > .80$	α = .63 (Gesamtverfahren); hohe Heterogenität der Einzelindizes	α (separat für Testkom-ponenten und Sprachen) von .66 bis .91	α = .93
Validität	Untersuchungen zur Krite-riumsvalidität (z. B. Recht-schreibnoten, andere Recht-schreibtests, Leseleistung)	Untersuchungen zur Kriteriumsvalidität im Zuge der LUST-1-Studie[6] (Lesenoten, Lehrerurteile, andere Verfahren)	Nur Angaben zu Ergebnissen einer explorativen Faktorenanalyse	Nur Angaben zu Korrelation der Testkomponenten und zu Testergebnissen deutscher und türkischer Kinder	Untersuchungen zur Kriteriumsvalidität (u. a. Machsprechtest und Intelligenztest)
Normierung	Bundesweite Neunormierung aus dem Jahr 2012	Normierung 2005; nicht repräsentative Ad-hoc Stichprobe	Angaben zu Durchschnitts-werten einer Normierungs-stichprobe aus Hamburg	(u. a.) Normierungsstudien in Duisburg 2003 und Bremen 2012	Normierung aus dem Jahr 1993
Steckbrief im ZUSE-Bericht	Ja	Nein	Ja	Ja	Nein
PSYNDEX-Review	Ja	Nein	Nein	Nein	Ja

[1] May, P. (2013). *HSP 1 – 10. Hamburger Schreib-Probe. Manual/Handbuch: Diagnose orthografischer Kompetenz*. Dortmund: vpm.

[2] Metze, W. (2005). *Stolperwörter-Lesetest*. Verfügbar unter: http://wilfriedmetze.de/Handanweisung_2009.pdf [14.05.2014]

[3] Reich, H. H. & Roth, H.-J. (2003). *Hamburger Verfahren zur Analyse des Sprachstands Fünfjähriger – HAVAS 5*. Landesinstitut für Lehrerbildung und Schulentwicklung Hamburg.

[4] Citogroep. (2004). *CITO. Test Zweisprachigkeit*. Arnheim: National Institute for Educational Measurement.

[5] Dummer-Smoch, L. (1993). *Die Diagnostischen Bilderlisten. Siebungsverfahren zur Früherkennung von Leselernschwierigkeiten im Leselernprozess*. Kiel: Veris.

[6] Brügelmann, H. (2003*). Lese-Untersuchung mit dem Stolperwörter-Test. Abschlussbericht des Projekts LUST-1*. Verfügbar unter: http://www2.agprim.uni-siegen.de/lust/stolper%5B1%5D.03.bericht.schlussfassung.12-20.pdf [07.11.2014]

Tabelle 6.5: Übersicht der am häufigsten genannten unstandardisierten Verfahren

	SISMIK[7]	LauBe[8]	ILeA[9]	Mirola[10]	SELDAK[11]
Diagnostische Zielsetzung	Systematische Beobachtung und Dokumentation der Sprachentwicklung, Erfassung von Fördervorausset-zungen; Fokus auf Kinder mit Zuwanderungsgeschichte	Erfassung der Lernaus-gangslage zu Schulbeginn	Individuelle Lernstands-analyse, Entwicklung individueller Lernpläne	Erfassung der individuellen Lernausgangslage; Ableitung von Schlussfolgerungen für Unterricht, Förderung und Elternarbeit	Systematische Beobachtung und Dokumentation der Sprachentwicklung (vgl. SISMIK, aber mit Fokus auf Kinder mit Deutsch als Muttersprache)
Alter/Ziel-gruppe	ab 3.5 Jahre bis zur Einschulung	Einsatz in den ersten Wochen nach Schuleintritt	Versionen für alle Jahrgangs-stufen in der Grundschule	zu Beginn der 1. Klasse	ab 4 Jahre bis zur Einschulung
Skalen/ Subtests	6 Skalen zu Sprachverhalten, sprachlichem Interesse und Sprachkompetenz; Erfassung von Kontextbedingungen (Familiensituation)	Schülerheft mit Aufgaben im Bereich Sprachentwicklung; Verknüpfung mit Portfolioverfahren „Lerndokumentation Sprache“	Fokus der Materialien für das Fach Deutsch liegen auf Lesen u. Rechtschreibung; Kombination aus Test- und Beobachtungselementen	Erfasst werden Grob- und Feinmotorik, Wahrnehmung, Pränumerische Kompetenz, Lateralität, Sprachkompetenz, Artikulation, Phonologische Kompetenz, Merkfähigkeit sowie Arbeits- und Sozialemotionales Verhalten	10 Skalen zu sprachlichen Kompetenzen (z. B. Wortschatz, Zuhören/Sinn-verstehen, Sätze nachsprechen, Grammatik, Phonologie, Schreiben/ Schrift)
Mündliche vs. schriftsprachliche Kompetenzen	Mündlich	Mündlich und schriftlich	Schriftlich	Mündlich	Mündlich
Rezeptive vs. produktive Sprachfähigkeiten	Rezeptiv und produktiv	Rezeptiv und produktiv	Rezeptiv und produktiv	Rezeptiv und produktiv	Rezeptiv und produktiv
Durchführung	Keine Angaben, Einzelverfahren	Keine Angaben, Bearbeitung einzeln oder in Gruppen	Keine Angaben, Einsatz in Schulklasse, Kleingruppen od. einzeln	2 Schulstunden, Gruppen-spiel für 6-8 Kinder	Keine Angaben, Einzelverfahren

	SISMIK[7]	LauBe[8]	ILeA[9]	Mirola[10]	SELDAK[11]
Objektivität	Keine Angaben (s. SELDAK)	Keine Angaben	Standardisierte Durchführung, Auswertung und Interpretation der Testelemente	Strukturiertes Gruppenbeobachtungsverfahren, Einsatz von drei Beobachtern	Sehr hohe Beobachterübereinstimmung[12]
Reliabilität	α = .88 bis .95	Keine Angaben	Keine Angaben	Keine Angaben	α = .82 bis .94[12]
Validität	Keine Angaben (s. SELDAK)	Keine Angaben	Keine Angaben	Keine Angaben	Untersuchung zur Kriteriumsvalidität (z. B. andere Verfahren, Kinder mit und ohne Förderung) [12]
Normierung	Bundesweite Stichprobe aus Migrantenkindern und Erzieher/innen (Stand: 2003)	Keine Angaben	Kriterial: Vorgabe zu erwartender Durchschnittsleistungen bei den Testaufgaben	Keine Angaben	Bundesweite Normierungsstichprobe (Stand: 2006)
Steckbrief im ZUSE-Bericht	Ja	Nein	Nein	Nein	Ja
PSYNDEX-Review	Ja	Nein	Nein	Nein	Nein

[7] Ulich, M. & Mayr, T. (2006). *Seldak. Sprachentwicklung und Literacy bei deutschsprachig aufwachsenden Kindern* (Beobachtungsbogen und Begleitheft). Freiburg: Herder.

[8] http://bildungsserver.berlin-brandenburg.de/lernausgangslage_laube.html [12.06.2014]

[9] LISUM (Landesinstitut für Schule und Medien Berlin - Brandenburg) (2012). *ILeA I. Individuelle Lernstandsanalysen Deutsch Lesen/Rechtschreiben. Erprobungsfassung 2012/2013*. Ludwigsfelde: LISUM.

[10] Hirschfeld, C. & Lassek, M. (2008). *Mit Mirola durch den Zauberwald. Beobachtungsverfahren für den Schulanfang zum Erfassen der Lernvoraussetzungen im Rahmen einer Gruppenbeobachtung*. Oberursel: Finken-Verlag..

[11] Ulich, M. & Mayr, T. (2003). *Sismik. Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kindertageseinrichtungen* (Beobachtungsbogen und Begleitheft). Freiburg: Herder.

[12] Mayr, T. & Gsottschneider, J. (2013). *Zur Objektivität und Validität des Einschätzbogens „Sprache und Literacy bei deutschsprachig aufwachsenden Kindern – SELDAK“. Frühe Bildung, 2*(4), 203-211.

7

Gesamtdiskussion

7 Gesamtdiskussion

Zusätzlich zu den spezifischen Diskussionen der Befunde der einzelnen Teilstudien, die bereits in den jeweiligen Zeitschriftbeiträgen (Kapitel 4 bis 6) vorgenommen wurden, soll an dieser Stelle eine Gesamtdiskussion erfolgen, in der die Ergebnisse der drei Teilstudien nicht isoliert, sondern integriert und in Bezug auf den in Kapitel 2 dargestellten, inhaltlich breit gefächerten theoretischen Hintergrund betrachtet werden. Hierbei werden zunächst die zentralen Ergebnisse der drei Teilstudien dargestellt (7.1), diskutiert und in die bisherige empirische Befundlage eingeordnet (siehe 7.2). Dem schließt sich eine methodische Bewertung der Arbeit und eine Diskussion ihrer Limitationen an (siehe 7.3). Abschließend werden Implikationen und Desiderata für die zukünftige Forschung und die schulische Praxis (siehe 7.4) aufgezeigt.

7.1 Zusammenfassung der Ergebnisse der drei Teilstudien

In der Einleitung zu dieser Dissertationsschrift wurde argumentiert, dass Lehrerinnen und Lehrer unter anderem auch Diagnostikerinnen und Diagnostiker seien. Als Beleg hierfür wurde die Breite an diagnostischen Tätigkeiten und Aufgaben skizziert, die sie in ihrem Beruf bewältigen müssen. Von dieser Fülle an diagnostischen Tätigkeiten wurde in der ersten Teilstudie dieser Dissertationsschrift mit der Einschätzung zur Schwierigkeit von einzelnen Aufgaben ein Aspekt betrachtet, der in der schulischen Praxis insbesondere bei der Konzeption und Gestaltung von Lernerfolgskontrollen sowie im Zusammenhang mit der Adaptation der unterrichtlichen Lernangebote an die Lernvoraussetzungen der Schülerinnen und Schüler von hoher Bedeutung ist (z. B. Brunner et al., 2011). Im Fokus der Teilstudien 2 und 3 stand die Sprachdiagnostik in der Grundschule, an der oftmals auch Lehrkräfte beteiligt sind und die somit ebenfalls Bestandteil des diagnostischen Aufgabenbereichs vieler Grundschullehrerinnen und -lehrer ist.

In Teilstudie 1 wurde den Fragen nachgegangen, (1) wie genau Lehrkräfte an Grundschulen beurteilen können, ob einzelne Deutsch- und Mathematikaufgaben für die Schülerinnen und Schüler ihrer Klassen eher schwierig oder eher einfach sind und (2) welche Faktoren im Zusammenhang mit der Über- oder Unterschätzung der Schwierigkeit von Aufgaben stehen.

Zu (1): Die Genauigkeit der Lehrerurteile wurde bezogen auf die drei vor allem in der deutschsprachigen Forschungsliteratur differenzierten Genauigkeitsfacetten (Rang-, Niveau- und Differenzierungskomponente, Helmke & Schrader, 1987) bestimmt, wobei

die Niveaukomponente nicht auf der Grundlage von Differenzen zwischen eingeschätzter und empirischer Aufgabenschwierigkeit (also etwa wie bei Brunner et al., 2011; McElvany et al., 2009), sondern anhand von prozentuellen Übereinstimmungsmaßen (vgl. z. B. Begeny & Buchanan, 2010; Begeny et al., 2008) ermittelt wurde. Für die Rangkomponente wurden sowohl bei den Deutsch- als auch bei den Mathematikaufgaben Korrelationskoeffizienten gefunden, deren Höhe nach der Klassifikation von Cohen (1988) als moderat bis groß einzustufen ist und die weitestgehend in dem Bereich der Werte liegen, die in anderen Studien berichtet wurden, in denen eine Untersuchung der Akkuratheit der Einschätzungen von Lehrkräften zur Aufgabenschwierigkeit erfolgte (vgl. Anders et al., 2010; Hosenfeld et al., 2002; Karing et al., 2011; McElvany et al., 2009).

Anders als in anderen Forschungsarbeiten wurde in Teilstudie 1 für die Niveaukomponente keine einheitliche Tendenz zur Unterschätzung der Aufgabenschwierigkeit ermittelt (z. B. Anders et al., 2010). Vielmehr wurde festgestellt, dass die untersuchten Lehrkräfte je nach Aufgabe und offenbar kovariierend mit der jeweiligen psychometrischen Aufgabenschwierigkeit entweder relativ gut in der Lage waren, die Schwierigkeit von Aufgaben akkurat zu beurteilen oder aber (zum Teil auch recht häufig) zu einer Über- oder Unterschätzung neigten. Bei der Differenzierungskomponente wurde, wie bei Lintorf und Kollegen (2011), eine Unterschätzung der Streuung von Aufgabenschwierigkeiten festgestellt. Die Lehrkräfte zeigten bei ihren Einschätzungen also offenbar eine Tendenz zur Mitte bzw. zur Vermeidung extremer Urteile.

Zu (2): Bei der Frage zu möglichen Kovariaten der Über- oder Unterschätzung der Aufgabenschwierigkeit wurde unter anderem, als Proxyvariable für die Berufserfahrung, die Anzahl der Jahre untersucht, die die Untersuchungsteilnehmerinnen und -teilnehmer bereits als Lehrerinnen oder Lehrer unterrichten. Im Einklang mit den Ergebnissen anderer Studien (z. B. Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003; Schrader, 1989) wurden hierbei nur Regressionskoeffizienten in geringer Höhe ermittelt, die zudem nicht konsistent statistisch signifikant waren. Gleiches gilt auch für die ebenfalls analysierte Kontaktdauer mit der Klasse, für die weder in Teilstudie 1 noch in anderen Forschungsarbeiten substanzielle Zusammenhänge mit der Urteilsgenauigkeit (bzw. mit der Über- oder Unterschätzung) festgestellt wurden (z. B. Wild & Rost, 1995).

Des Weiteren wurde untersucht, ob zwischen der Genauigkeit der Schwierigkeitsurteile von Lehrkräften und weiteren Hintergrundinformationen (a) zum Unterricht in der

betreffenden Klasse, für die die Aufgabenschwierigkeit jeweils anzugeben war, und (b) zum Curriculum der Primarstufe statistisch bedeutsame Zusammenhänge bestehen. Hierbei wurde zum einen (a) erwartet, dass eine häufige Einübung oder Thematisierung derjenigen Teilkompetenzen im Unterricht des letzten Schuljahres, die für eine erfolgreiche Bewältigung der zu beurteilenden Aufgaben erforderlich sind, Einschätzungen zur Aufgabenschwierigkeit erleichtern könnte, da Lehrkräfte aus ihren Unterrichtsbeobachtungen diagnostische Informationen dazu erhalten, wie gut die betreffenden Teilkompetenzen bereits erworben wurden. Aus diesen diagnostischen Informationen sollten sie wiederum schließen können, wie gut die betreffenden Aufgaben von den Schülerinnen und Schülern ihrer Klasse beherrscht werden. Zum anderen (b) wurde vermutet, dass Lehrkräfte bei ihren Beurteilungen unter anderem auch Informationen dazu heranziehen, zu welchem Zeitpunkt in der Primarstufe Aufgaben mit ähnlichen Anforderungen behandelt worden sind. Grundlage dieser Vermutung war die Annahme, dass Lehrkräfte diese curricularen Informationen als Orientierungshilfe nutzen könnten, um einzuschätzen, inwiefern bestimmte Aufgaben von ihren Schülerinnen und Schülern bereits beherrscht werden oder nicht. Konkret wurde erwartet, dass Lehrkräfte präzisere Einschätzungen der Schwierigkeit von Aufgaben vornehmen können, wenn sie Aufgaben gleichen Typs noch vor nicht allzu langer Zeit (d. h. in Teilstudie 1: nicht schon am Anfang der Grundschulzeit) in ihrem Unterricht behandelt haben. Ihre im Unterricht gewonnenen Eindrücke, so die Annahme, sollten ihnen die Schwierigkeitsbeurteilung erleichtern und umso besser erinnert werden, je größer deren zeitliche Nähe zum Urteil ist.

Die Ergebnisse von Teilstudie 1 können die skizzierten Annahmen nur teilweise belegen: (a) So konnte für eine Mathematikaufgabe zum Kompetenzbereich (bzw. zur sogenannten Leitidee) „Daten, Häufigkeit und Wahrscheinlichkeit“ (vgl. KMK, 2005b) gezeigt werden, dass Lehrkräfte umso weniger zur Überschätzung der Schwierigkeit dieser Aufgabe tendierten, je häufiger die betreffende Teilkompetenz (Mathematische Strukturen in Alltagskontexten erkennen) Gegenstand ihres Unterrichts war. Außerdem (b) wurden sowohl bei den Deutsch- als auch bei den Mathematikaufgaben statistisch signifikante Zusammenhänge zwischen der Urteilsgenauigkeit einerseits und dem Zeitpunkt bzw. der Jahrgangsstufe andererseits festgestellt, in der Aufgaben ähnlichen Typs behandelt wurden. Diese Zusammenhänge bildeten allerdings ab, dass die Lehrkräfte zwar (hypothesenkonform) stärker zur Unterschätzung der Aufgaben-

schwierigkeit neigten, wenn sie angaben, ähnliche Aufgaben zu einem frühen Zeitpunkt in der Primarstufe behandelt zu haben (geringere zeitliche Nähe zum Urteil). Gleichzeitig ging jedoch eine Behandlung ähnlicher Aufgaben zu einem späteren Zeitpunkt (höhere zeitliche Nähe zum Urteil) mit einer Überschätzung der Aufgabenschwierigkeit einher (hypothesenkonträr).

Unterschiede in der Urteilsgenauigkeit zwischen einzelnen Lehrkräften wurden in Teilstudie 1 für die Rangkomponente analysiert. Hierbei wurde, ähnlich wie in anderen Studien (z. B. Hosenfeld et al., 2002), eine hohe interindividuelle Varianz der lehrerspezifischen Urteilsgenauigkeit festgestellt: Die Urteile einiger Lehrkräfte fielen also überaus genau aus, während andere nicht oder kaum in der Lage waren, die Schwierigkeit der ihnen vorgelegten Aufgaben akkurat einzuschätzen. Darüber hinaus zeigen auch die Ergebnisse zur Niveaukomponente, dass die Schwierigkeit zwar in vielen Fällen korrekt eingeschätzt wurde, es aber gleichzeitig auch nicht wenige Lehrkräfte gab, die das Schwierigkeitsniveau der ihnen vorgelegten Aufgaben über- oder unterschätzten. Diese Ergebnisse zur interindividuellen Varianz der Urteilsgenauigkeit ähneln den Befundmustern von Studien, in denen die Genauigkeit von Lehrerurteilen zu den Leistungen von Schülerinnen und Schülern untersucht wurde (Hoge & Coladarci, 1989; Südkamp et al., 2012).

In Kapitel 3 und 4 wurde als ein Spezifikum von Teilstudie 1 hervorgehoben, dass statt der Genauigkeit von Lehrerurteilen zu Schülerleistungen die Akkuratheit von Einschätzungen zu Aufgabenschwierigkeiten untersucht wurde, die in Bezug auf die Adaptation der unterrichtlichen Lernangebote an die Lernvoraussetzungen der Schülerinnen und Schüler sowie bei der Gestaltung von Lernerfolgskontrollen von hoher Bedeutung ist (z. B. Brunner et al., 2011; Hascher, 2008). Tatsächlich sind Lehrerurteile zu Schülerleistungen und zu Aufgabenschwierigkeiten inhaltlich nicht grundverschieden, sondern eng miteinander verwoben. Wenn Lehrkräfte prognostizieren sollen, wie gut ihre Schülerinnen und Schüler etwa in einem standardisierten Leistungstest abschneiden, dann werden sie, zumindest im Falle von informierten Urteilen (vgl. 2.1.4 sowie Hoge & Coladarci, 1989; Südkamp et al., 2012), gleichzeitig auch einschätzen, ob die Aufgaben des Tests für die Schülerinnen und Schüler eher leicht oder eher schwer zu lösen sein werden. Wenn Lehrkräfte die Schwierigkeit von Aufgaben beurteilen sollen, werden sie ihre Einschätzungen in der Regel auf eine bestimmte Schülergruppe beziehen – zum Beispiel (wie in Teilstudie 1) auf ihre Schulklasse. Sie werden ihre Schwierigkeitsurteile

also etwa daran festmachen, wie viele Schülerinnen und Schüler die Aufgaben ihrer Einschätzung nach lösen können werden.

Aufgrund der skizzierten Interrelationen zwischen leistungs- und aufgabenbezogenen Urteilen erlauben die Befunde von Teilstudie 1 also nicht nur Schlussfolgerungen über die Akkuratheit von Lehrereinschätzungen zur Schwierigkeit von Aufgaben, sondern sie lassen auch Rückschlüsse auf die Genauigkeit von Lehrerurteilen zu den Leistungen von Schülerinnen und Schülern zu. So zeigen die Ergebnisse von Teilstudie 1 nicht nur, dass die Genauigkeit der von Lehrkräften vorgenommenen Einschätzungen zur Schwierigkeit von Aufgaben in vielen Fällen gering ausfällt. Vielmehr deuten sie außerdem darauf hin, dass viele Lehrkräfte offenbar nicht oder nur bedingt in der Lage sind, die Leistungen ihrer Schülerinnen und Schüler akkurat zu beurteilen. Diese Schlussfolgerung korrespondiert mit den in Teilkapitel 2.1.5 dargestellten Ergebnissen zahlreicher Untersuchungen zur Genauigkeit von Lehrerurteilen zu Schülerleistungen, in denen ebenfalls resümiert wird, dass Lehrkräfte die Leistungen ihrer Schülerinnen und Schüler oftmals nur ungenau einschätzen können (z. B. Feinberg & Shapiro, 2009; Hoge & Coladarci, 1989; Südkamp et al., 2012).

Für die schulische Praxis impliziert dieser zentrale Befund der genauigkeitsorientierten Forschung zur diagnostischen Kompetenz von Lehrkräften unter anderem, dass wichtige diagnostische Entscheidungen, die für einzelne Schülerinnen und Schüler mit weitreichenden Konsequenzen verbunden sein können, nicht ausschließlich auf den Urteilen von Lehrerinnen und Lehrern, sondern (wenn möglich) auch noch auf weiteren diagnostischen Informationsquellen basieren sollten. Wie in Kapitel 3 skizziert, wirft diese Feststellung allerdings die Frage auf, welche weiteren diagnostischen Informationsquellen dazu geeignet sind, die Güte diagnostischer Entscheidungen zu erhöhen. Vor diesem Hintergrund wurde in Teilstudie 2 am Beispiel von Entscheidungen zur Teilnahme an additiven Sprachfördermaßnahmen in der Grundschule unter anderem untersucht, inwiefern die Klassifikationsgüte von Sprachförderentscheidungen mit den jeweils zugrundeliegenden diagnostischen Informationsquellen kovariiert. Als Informationsquellen wurden die Nutzung sprachdiagnostischer Verfahren, Beobachtungen der Lehrkraft, Informationen anderer pädagogischer Fachkräfte wie Erzieherinnen und Erzieher, schulärztliche Informationen, Elterninformationen sowie Schulnoten betrachtet. Ein signifikant positiver Zusammenhang mit der Klassifikationsgüte konnte ausschließlich für die Nutzung sprachdiagnostischer Verfahren und nur für den Indikator

der Sensitivität festgestellt werden. Ein ausgeprägter Förderbedarf wurde also in Grundschulen, in denen bei der Sprachdiagnostik auch sprachdiagnostische Instrumente zum Einsatz kommen, häufiger erkannt als in Grundschulen, die auf eine Verwendung solcher Verfahren verzichten. Darüber hinaus wurde anhand einer zusätzlichen Differenzierung nach der Einsatzhäufigkeit der Instrumente festgestellt, dass sich dieser Effekt nur für die mehrmalige, nicht jedoch für die einmalige Verwendung sprachdiagnostischer Instrumente findet. Im Zeitschriftenbeitrag zu Teilstudie 2 wurde erläutert, dass dieses Befundmuster mit Anforderungen an die Sprachdiagnostik korrespondiert (vgl. auch 2.3.4), die unter Bezugnahme auf Ergebnisse der Spracherwerbsforschung formuliert wurden und eine mehrmalige (regelmäßige) Erfassung sprachlicher Kompetenzen mittels sprachdiagnostischer Verfahren empfehlen (z. B. Bredel, 2005).

Insgesamt zeigt die in Teilstudie 2 ermittelte Klassifikationsgüte, dass die Sensitivität von Sprachförderentscheidungen selbst in den Grundschulen eher gering ausfiel, in denen die sprachlichen Kompetenzen der Kinder mehrfach mittels sprachdiagnostischer Verfahren überprüft werden. Hierzu ist anzumerken, dass in der Prädiktorvariable „sprachdiagnostische Verfahren“, analog zu dem Begriffsverständnis in anderen Publikationen (z. B. Becker-Mrotzek et al., 2013; Redder et al., 2011), sehr unterschiedliche Methoden wie Tests, Screenings, Einschätzverfahren sowie Beobachtungsbögen und Beobachtungsverfahren zusammengefasst waren. Weiteren Aufschluss über die verschiedenen Instrumente, die an den in Teilstudie 2 untersuchten Grundschulen genutzt wurden, geben die Ergebnisse von Teilstudie 3, bei der eine bundesweite Bestandsaufnahme zum Einsatz sprachdiagnostischer Verfahren vorgenommen wurde.

Die Ergebnisse von Teilstudie 3 verdeutlichen, dass an den Grundschulen in Deutschland viele verschiedene Verfahren eingesetzt werden, die nicht zuletzt hinsichtlich der Frage, in welchem Umfang sie die unter 2.3.4 skizzierten testtheoretischen Anforderungen erfüllen, zum Teil deutlich differieren. Zudem scheint die Nutzung von selbstentwickelten Instrumenten verbreitet zu sein, deren psychometrische Güte in der Regel nicht untersucht sein dürfte. Darüber hinaus dokumentieren die Befunde von Teilstudie 3, dass in den Grundschulen Verfahren dominieren, die auf eine Erfassung schriftsprachlicher Kompetenzen (insbesondere im

Lesen und in der Rechtschreibung) zielen. Vielfach wird als einziges Verfahren auch nur ein Rechtschreibtest (die Hamburger Schreibprobe) verwendet.

Als mögliche Erklärung dafür, dass in Teilstudie 2 auch beim Einsatz sprachdiagnostischer Verfahren verhältnismäßig geringe Sensitivitätswerte festgestellt wurden, lässt sich daher festhalten, dass in Grundschulen offenbar nicht selten Instrumente zum Einsatz kommen, die aufgrund ihrer geringen psychometrischen Güte keine objektive, reliable und valide Erfassung sprachlicher Kompetenzen erlauben und somit auch keinen Beitrag zur Erhöhung der Klassifikationsgüte leisten können. Als eine weitere mögliche Ursache könnte eine mangelhafte Passung zwischen Sprachdiagnostik und -förderung vermutet werden. Passungsprobleme lägen etwa dann vor, wenn die an einer Grundschule eingesetzten Verfahren ausschließlich auf die Erfassung von Lese- oder Rechtschreibkompetenzen zielen, die Förderung jedoch nicht nur auf schriftsprachliche Kompetenzen, sondern auch auf mündliche Kompetenzen bezogen erfolgt. Des Weiteren könnten Passungsprobleme in Bezug auf die Parallelität von Diagnostik und Förderung produktiver und rezeptiver Sprachkompetenzen bestehen: So zeigen die Ergebnisse von Teilstudie 3, dass viele der in Teilstudie 2 untersuchten Grundschulen einzig die Hamburger Schreibprobe einsetzen, die zwar eine für die Sprachentwicklung im Primarbereich relevante sprachliche Teilfähigkeit erfasst (orthografische Kompetenz), aber mit den produktiven schriftsprachlichen Fähigkeiten nur eine sprachliche Facette abbildet. Demgegenüber wurde jedoch die Teilnahme an additiven Sprachfördermaßnahmen für rezeptive schriftsprachliche Fähigkeiten und mündliche Kompetenzen erhoben („Sprach- und Leseförderung“) und der Sprachförderbedarf wurde anhand der im Lesen (rezeptiv-schriftsprachlich) und im Zuhören (rezeptiv-mündlich) festgestellten Kompetenzen operationalisiert. Folglich wäre denkbar, dass in Teilstudie 2 auch deswegen kein größerer Effekt des Einsatzes sprachdiagnostischer Verfahren gefunden werden konnte, weil die mit den verwendeten Verfahren erhobenen Informationen nur für Lesen und Zuhören, nicht aber für den Förderbedarf bei anderen sprachlichen Kompetenzen einen hohen inhaltlichen Vorhersagewert hatten.

In Teilstudie 2 konnte der skizzierten Vermutung leider nicht empirisch nachgegangen werden, da in der IQB-Ländervergleichsstudie keine Daten zu den inhaltlichen Schwerpunkten der von den Schülerinnen und Schülern besuchten Sprachfördermaßnahmen erfasst wurden. Das skizzierte Passungsproblem wird nochmals unter 7.3

(Methodische Bewertung und Grenzen der Arbeit) aufgegriffen. Zuvor soll zunächst eine vertiefte Diskussion der drei Teilstudien und ihrer zentralen Befunde erfolgen.

7.2 Diskussion der Ergebnisse der drei Teilstudien

7.2.1 Diskussion der Ergebnisse zur Urteilsgenauigkeit von Lehrkräften (Teilstudie 1)

Wie bereits erläutert, stand die Genauigkeit von Lehrerurteilen zur Schwierigkeit ausgewählter Deutsch- und Mathematikaufgaben im Mittelpunkt von Teilstudie 1. In der von Schrader (2009, 2013) vorgenommenen Systematisierung der empirischen Forschung zur diagnostischen Kompetenz von Lehrkräften ist Teilstudie 1 demgemäß primär den genauigkeitsorientierten Ansätzen zuzuordnen. Wie in Teilkapitel 2.1 ausgeführt, ist das Forschungsfeld dieser Ansätze von Forschungsarbeiten geprägt, in denen die Genauigkeit der Urteile von Lehrkräften zu Leistungsmerkmalen von Schülerinnen und Schülern analysiert wird. Bislang deutlich seltener wurde die eng damit verbundene Frage (s. o.) untersucht, wie akkurat Lehrkräfte die Schwierigkeit von Aufgaben einschätzen können, wobei insbesondere die Niveau- und Differenzierungskomponente der Genauigkeit aufgabenbezogener Lehrerurteile nur wenige Male betrachtet wurden. Folglich trägt die Teilstudie 1 dazu bei, den derzeitigen Forschungsstand zur diagnostischen Kompetenz von Lehrkräften zu schärfen, indem sie diesen um weitere Ergebnisse bereichert, die die bisherigen, noch wenig umfangreichen Erkenntnisse zur Akkuratheit von Schwierigkeitseinschätzungen im Wesentlichen bestätigen.

Eine Besonderheit von Teilstudie 1 ist, dass nicht nur die Akkuratheit von Schwierigkeitseinschätzungen ermittelt wurde, sondern zusätzlich auch eine Betrachtung möglicher Kovariaten der Urteilsgenauigkeit erfolgte. Dabei wurde erstmals untersucht, inwiefern der Zeitpunkt, zu dem Aufgaben ähnlichen Typs im Unterricht behandelt wurden sind, oder die Häufigkeit, mit der aufgabenrelevante Teilkompetenzen im Unterricht des letzten Jahres thematisiert wurden, mit dem Ausmaß an Über- oder Unterschätzung von Aufgabenschwierigkeiten kovariieren. Konzeptuell war die Untersuchung dieser beiden Aspekte durch die Überlegung motiviert, dass Lehrkräften der eigene Unterricht als diagnostische Informationsquelle dienen könnte, wenn sie einschätzen sollen, wie schwierig bestimmte Aufgaben für die Schülerinnen und Schüler ihrer Klasse sind. Hierbei war ursprünglich vermutet worden, dass die für Schwierigkeitsurteile relevanten diagnostischen Informationen umso besser verfügbar sind (d. h. erinnert werden), je weniger sie zeitlich zurückliegen (= Zeitpunkt der Behandlung ähnlicher

Aufgaben im Unterricht) und je häufiger sie vorkommen (= Häufigkeit, mit der relevante Teilkompetenzen Gegenstand des Unterrichts waren). Eine bessere Verfügbarkeit relevanter diagnostischer Informationen, so die Erwartung, sollte es Lehrkräften ermöglichen, akkuratere Einschätzungen vorzunehmen.

Allerdings belegen die Ergebnisse von Teilstudie 1 zwar, dass Lehrkräfte ihre Schwierigkeitseinschätzungen unter anderem daran orientieren, zu welchem Zeitpunkt in der Primarstufe sie Aufgaben ähnlichen Typs im Unterricht behandeln. Die Annahme, dass eine größere zeitliche Nähe zur Abgabe der Schwierigkeitseinschätzungen zu akkurateren Urteilen führt, konnte jedoch nicht bestätigen. Stattdessen spricht das Befundmuster dafür, dass Lehrkräften das Wissen um den Zeitpunkt der Thematisierung von ähnlichen Aufgaben im Unterricht als Urteilsheuristik (vgl. Tversky & Kahneman, 1974) dient. Diese Heuristik scheint auf der sachlogisch nachvollziehbaren Annahme zu basieren, dass Aufgaben, die zum Beginn der Primarstufe behandelt werden, leichter sein dürften als Aufgaben, die erst zu einem späteren Zeitpunkt Gegenstand des Unterrichts in der Primarstufe sind. In diesem Sinne reflektiert die Heuristik also normative Erwartungen von Lehrkräften an die Schwierigkeit von Aufgaben bzw. an den kumulativen Kompetenzerwerb und die Leistung ihrer Schülerinnen und Schüler. In der vorliegenden Dissertationsschrift waren mögliche Effekte solcher Erwartungen auf die Urteilsgenauigkeit von Lehrkräften bereits in den Erläuterungen zu dem unter 2.1.6 (siehe Abbildung 2.1) vorgestellten Modell von Südkamp und Kollegen (2012) thematisiert und als ein Desiderat für die Forschung hervorgehoben worden (vgl. 2.1.6.2).

Die Ergebnisse von Teilstudie 1 lassen vermuten, dass Lehrkräfte bei der Anwendung der skizzierten Heuristik nicht oder nur ungenügend berücksichtigen, dass bestimmte Aufgaben (vom Ende der Primarstufe) gerade deswegen besonders gut (bzw. besser als erwartet) gelöst werden können, weil ähnliche Aufgaben erst kürzlich im Unterricht geübt wurden. Ähnliches scheint auch für Aufgaben zu gelten, die in der Regel eher zum Beginn der Primarstufe behandelt werden. Bei deren Beurteilung wird offenbar nicht oder nur ungenügend berücksichtigt, dass sie trotz ihres eher geringen Anforderungsniveaus deshalb von einigen Schülerinnen und Schülern am Ende der Primarstufe nicht erfolgreich gelöst werden können, weil sie bereits seit längerer Zeit nicht mehr Gegenstand des Unterrichts waren.

Die verzerrende Wirkung der skizzierten Urteilsheuristik bzw. der ihr zugrundeliegenden Prozesse lässt sich auf der theoretischen Basis der *social judgment theory*

(z. B. Doherty, M. E. & Kurz, 1996) bzw. des *Linsenmodells* (Brunswik, 1955) erläutern. Im Linsenmodell wird unter anderem die Annahme formuliert, dass ein Kriterium (z. B. die Schülerleistung in einem Test) nicht direkt, sondern nur anhand von bestimmten Hinweisreizen (*cues*) beurteilt werden kann. Diese Hinweisreize unterscheiden sich in ihrer ökologischen Validität, also in dem Vorhersagewert, den sie für das Kriterium haben. Auch schränkt die Validität der verfügbaren Hinweisreize die Vorhersagbarkeit des Kriteriums bzw. die maximal mögliche Urteilsgenauigkeit ein. Bei der Beurteilung selbst wird eine Gewichtung der Hinweisreize vorgenommen. Je besser die Gewichtung den ökologischen Validitäten der Hinweisreize entspricht, umso mehr nähert sich die Urteilsgenauigkeit der maximal möglichen Genauigkeit an. In Teilstudie 1 diente den teilnehmenden Lehrkräften offenbar unter anderem das Wissen um den Zeitpunkt, zu dem Aufgaben ähnlichen Typs in der Primarstufe behandelt werden, als Hinweisreiz für die Aufgabenschwierigkeit. Die festgestellte Urteilsverzerrung resultierte, weil dieser Hinweisreiz nur einen geringen Vorhersagewert für das Kriterium (Aufgabenschwierigkeit) hatte und offenbar dennoch von vielen der an der Studie teilnehmenden Lehrkräfte bei der Schwierigkeitseinschätzung ein hohes Gewicht erhielt (vgl. auch Cooksey et al., 1986).

In der gemeinsamen Zusammenfassung der Ergebnisse der drei Teilstudien in Teilkapitel 7.1 fokussierten die Darstellungen zur Genauigkeit der Schwierigkeitseinschätzungen von Lehrkräften auf den festgestellten hohen interindividuellen Unterschieden in der Urteilsakkuratheit und auf die Frage, wie sich die im Mittel für die Rang-, Niveau- und Differenzierungskomponente gefundenen Genauigkeitskennwerte in die in der Forschungsliteratur dokumentierte Befundlage einordnen. Nicht diskutiert wurde hingegen, wie etwa eine mittlere Rangkomponente von $r = .51$[28] inhaltlich zu bewerten ist. Lässt dieses Ergebnis den Schluss zu, dass Rangurteile von Lehrkräften eine geringe Akkuratheit haben? Oder zeigt der Befund vielmehr, dass die Genauigkeit von Rangurteilen im Mittel durchaus akzeptabel ist?

Wie bereits in Teilkapitel 7.1 und im Zeitschriftenbeitrag beschrieben, korrespondieren die in Teilstudie 1 ermittelten Genauigkeitskennwerte weitestgehend mit dem Befundmuster aus anderen Studien (z. B. Lehmann et al., 2000; Lintorf et al., 2011). In

[28] Dieser Kennwert wurde in Teilstudie 1 für die Rangkomponente der Genauigkeit der Lehrerurteile zur Schwierigkeit der Mathematikaufgaben ermittelt.

der Forschungsliteratur wird dieses Befundmuster oftmals als Beleg dafür interpretiert, dass Lehrerurteile zu absoluten Merkmalsausprägungen (d. h. Niveauurteile) vielfach ungenau seien, Lehrkräfte jedoch offenbar relativ gut in der Lage zu sein scheinen, Schülerinnen und Schüler hinsichtlich ihrer Leistungen in eine korrekte Rangreihe zu bringen (z. B. Bates & Nettelbeck, 2001; Südkamp & Möller, 2009). Allerdings fehlt es an Maßstäben, ab welcher Ausprägung etwa die Rangkomponente der Urteilsgenauigkeit als zufriedenstellend gelten kann (z. B. Helmke, 2010; Lorenz & Artelt, 2009). Ein Ansatz zur inhaltlichen Bewertung könnte darin bestehen, den in der Rangkomponente abgebildeten statistischen Zusammenhang zwischen Lehrerurteilen einerseits und der Kriteriumsvariablen (d. h. tatsächliche Merkmalsausprägungen wie Testergebnisse von Schülerinnen und Schülern oder empirische Aufgabenschwierigkeiten) andererseits im Sinne der Logik einer Multi-Trait-Multi-Method-Matrix (Campbell & Fiske, 1959; Döring & Bortz, 2016) den Ergebnissen weiterer Messungen gegenüberzustellen. Hierbei könnte etwa geprüft werden, wie gut andere Methoden (z. B. ein anderer Mathematiktest) das Kriterium (z. B. die Testergebnisse in Mathematik) vorhersagen und inwieweit die dabei ermittelte Korrelation die Ausprägung der Rangkomponente übersteigt. Auch könnten Vergleiche mit den Ergebnissen von Verfahren erfolgen, die ein anderes Konstrukt bzw. Trait erfassen. In diesem Sinne lassen etwa die Ergebnisse einer Studie von Eaves, Williams, Winchester und Darch (1994) darauf schließen, dass Intelligenztests die Ergebnisse von Schülerinnen und Schülern in Lese- und Mathematiktests offenbar besser vorhersagen ($r_{Intelligenztest,\ Lesetest} = .71$ und $r_{Intelligenztest,\ Mathematiktest} = .85$) als die von Lehrkräften vorgenommenen Leistungseinschätzungen, wobei die in dieser Untersuchung über Subtests gemittelte Rangkomponente $r = .47$ (Mathematik) bzw. $r = .46$ (Lesen) betrug.

Die Ergebnisse der Studie von Eaves und Kollegen (1994) deuten folglich darauf hin, dass die Urteile von Lehrkräften weniger gut mit den von Schülerinnen und Schülern in einem bestimmten Kompetenzbereich erzielten Leistungen übereinstimmen als die Ergebnisse von Testverfahren, die eigentlich ein anderes Konstrukt (in diesem Fall: Intelligenz) erfassen. Folgte man dieser Interpretation, müssten Rangkomponenten in der Ausprägung, wie sie bei Eaves und Kollegen (1994) und somit auch in Teilstudie 1 gefunden wurden, als Indiz dafür gelten, dass Lehrerurteile im Mittel wenig akkurat sind.

Ein differenzierteres Bild ergibt sich hingegen, wenn man Überlegungen dazu anstellt, ob die Schwierigkeit von Aufgaben überhaupt vollkommen akkurat beurteilt

werden kann bzw. welchen Grad an Urteilsgenauigkeit Lehrkräfte überhaupt realistisch betrachtet erreichen können. Den Ausgangspunkt dieser in den folgenden Absätzen skizzierten Überlegungen bildet das bereits weiter oben skizzierte Linsenmodell (Brunswik, 1955). Folgte man der Logik dieses Modells, dann wäre die von Lehrkräften erzielte Urteilsgenauigkeit jeweils daran zu relativieren, a) welche Hinweisreize ihnen bei ihren Einschätzungen überhaupt zur Verfügung standen und b) welchen Vorhersagewert diese Hinweisreize für das Kriterium haben. Als mögliche Hinweisreize, auf denen eine Beurteilung basieren könnte, wurden in Teilstudie 1 zusätzlich Hintergrundinformationen zu Inhalten des Unterrichts berücksichtigt. Allerdings lassen die Ergebnisse von Teilstudie 1 zumindest im Hinblick auf Informationen zum Zeitpunkt der Thematisierung ähnlicher Aufgaben im Unterricht auf einen eher geringen Vorhersagewert schließen.

Als weitere (in Teilstudie 1 nicht berücksichtigte) Hinweisreize könnten im Falle aufgabenbezogener Urteile bestimmte Merkmale der einzuschätzenden Aufgaben dienen, die Anhaltspunkte für die Aufgabenschwierigkeit liefern können. Solche Anhaltspunkte oder auch „schwierigkeitsbestimmende Aufgabenmerkmale“ sind etwa das Vokabular und die grammatikalische Struktur von Lesetexten, die Sprecheranzahl und Hintergrundgeräusche in Hörtexten, die Plausibilität von Distraktorantworten in Multiple-Choice-Items oder die Komplexität von Aufgabenstellungen (z. B. Böhme, Robitzsch & Busè, 2010; Buck & Tatsuoka, 1998; Freedle & Kostin, 1993; Hartig & Frey, 2012; Isaac & Hochweber, 2011; Leucht, Harsch, Pant & Köller, 2012).

In der Forschungsliteratur finden sich mehrere Studien, die untersucht haben, wie gut die empirische Schwierigkeit von Aufgaben durch schwierigkeitsbestimmende Aufgabenmerkmale statistisch vorhergesagt werden kann. Hartig und Frey (2012) konnten anhand von Daten aus der DESI-Studie für Testitems zum Leseverstehen in Englisch eine Varianzaufklärung von $R^2 = .49$ ermitteln. Leucht und Kollegen (2012) fanden für Englischitems aus einer Studie des IQB zur Normierung der Bildungsstandards in der ersten Fremdsprache Englisch, dass die von ihnen betrachteten Aufgabenmerkmale 31 Prozent (für Items zum Leseverstehen) bzw. 50 Prozent (für Items zum Hörverstehen) der Varianz der im Raschmodell geschätzten Itemschwierigkeiten aufklären konnten. Isaac und Hochweber (2011) ermittelten für Aufgaben aus einem VERA-3-Test im Fach Deutsch für den Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“ einen Anteil von 70 Prozent aufgeklärter Varianz. Schließlich berechneten Böhme und Kollegen (2010), anhand von Daten aus Pilotierungs- und Normierungsuntersuchungen

zu den Bildungsstandards im Fach Deutsch für den Primarbereich (KMK, 2005a), für Testaufgaben zum Hörverstehen[29] eine Varianzaufklärung von 49 Prozent.

Die Ergebnisse der referierten Studien lassen folglich darauf schließen, dass auch eine systematische Analyse schwierigkeitsbestimmender Merkmale keine vollkommen akkurate Beurteilung von Aufgabenschwierigkeiten ermöglicht. Vielmehr unterscheiden sich zumindest einige der in den angeführten Forschungsarbeiten ermittelten Kennwerte der Varianzaufklärung nur in geringem Maße von den in Studien zur diagnostischen Kompetenz von Lehrkräften ermittelten Ausprägungen der Rangkomponente. So entspricht etwa die bei Leucht und Kollegen (2012) berichtete Varianzaufklärung von 31 Prozent für Englischitems zum Leseverstehen einem Korrelationskoeffizienten von $r = .56$.

Legt man die skizzierte Lesart des Linsenmodells und die in Studien zur Vorhersage von Aufgabenschwierigkeiten durch schwierigkeitsbestimmende Aufgabenmerkmale ermittelten Befunde zugrunde, ließen sich Werte für die Rangkomponente der Genauigkeit von Schwierigkeitsurteilen, wie sie in Teilstudie 1 gefunden wurden, als durchaus akzeptabel einstufen. Hiergegen könnte allerdings eingewendet werden, dass Lehrkräfte gegenüber Studien zu schwierigkeitsbestimmenden Merkmalen möglicherweise eigentlich sogar leicht im Vorteil sein müssten, da ihnen zusätzlich Informationen zu den Leistungen bzw. Leistungsunterschieden ihrer Schülerinnen und Schüler vorliegen.

Allerdings ist darüber hinaus zu bedenken, dass auch deswegen keine perfekten Korrelationen zwischen Lehrerurteilen einerseits und den Testleistungen von Schülerinnen und Schülern oder den empirischen Schwierigkeiten von Testaufgaben andererseits zu erwarten sind, weil auch Leistungstests keine perfekte Reliabilität aufweisen, so dass die mit ihnen ermittelten Ergebnisse folglich stets mit einem Messfehler behaftet sind. Auch dieser Messfehler ist also zu berücksichtigen, wenn beispielsweise die Ausprägung der Rangkomponente von Lehrerurteilen inhaltlich bewertet werden soll (vgl. Helmke, 2010).

Schließlich lassen sich die hier skizzierten Überlegungen dazu, wie genau Lehrerurteile überhaupt sein können, noch um die kritische Frage ergänzen, ob es

[29] Böhme und Kollegen (2010) verwiesen zudem darauf, dass im Leseverstehen ähnliche Werte festgestellt wurden.

überhaupt angemessen ist zu erwarten, dass Lehrkräfte im Rahmen des unter 2.1.4 beschriebenen Paradigmas zur Untersuchungen der Urteilsgenauigkeit eine hohe Akkuratheit erzielen sollten. Zwar stehen Lehrerinnen und Lehrer in ihrem Beruf immerfort vor der Herausforderung, Beurteilungen vornehmen zu müssen (vgl. Kapitel 1), jedoch dürften sie nur eher selten die Aufgabe haben, die Leistung ihrer Schülerinnen und Schüler in einem standardisieren Test einzuschätzen. Darüber hinaus sind die in empirischen Studien untersuchten Urteile möglicherweise auch insofern nur bedingt repräsentativ für die schulische Praxis, als dass viele der im Schulalltag von Lehrkräften vorzunehmenden Leistungsbewertungen womöglich nicht nur die bloße Leistung von Schülerinnen und Schülern abbilden. Dies zeigt sich zum Beispiel anhand der Vergabe von Schulnoten, die, wie unter 2.1.3 beschrieben, verschiedenen Funktionen dienen kann und daher häufig nicht allein auf der Bewertung von Leistungen, sondern auch von weiteren Aspekten, basiert. In der Forschungsliteratur wird diese „funktionale Überfrachtung“ (Sacher, 1994, S. 21) von Schulnoten jedoch kritisch gesehen; insbesondere wird bezweifelt, dass die zum Teil auch inhaltlich divergierenden Funktionen im gleichen Maße erfüllt werden können (vgl. auch Lintorf, 2012). Auch unter 2.1.3 war das Abweichen vom reinen Leistungsprinzip bei der Notenvergabe problematisiert worden (vgl. auch Dünnebier et al., 2009). Ungeachtet dessen lässt sich aus den skizzierten Vermutungen zu Gegenstand und Funktionen von Leistungsbewertungen in der Schule ein weiterer Erklärungsansatz dazu ableiten, weshalb viele empirische Untersuchungen zu dem Schluss kommen, dass Leistungsurteile von Lehrerinnen und Lehrern im Durchschnitt wenig akkurat sind (vgl. 2.1.5 sowie Feinberg & Shapiro). Vielleicht sind Lehrkräfte schlichtweg ungeübt oder sogar überfordert, wenn sie als Teilnehmerinnen und Teilnehmer einer empirischen Studie ausschließlich die Leistung ihrer Schülerinnen und Schüler beurteilen sollen?

7.2.2 Diskussion der Ergebnisse zur Sprachdiagnostik in der Grundschule (Teilstudie 2 und Teilstudie 3)

Unabhängig davon, ob man der unter 7.2.1 dargestellten Argumentation folgt und Ausprägungen der Rangkomponente von circa r = .50 als Belege dafür interpretiert, dass Lehrkräfte im Mittel vergleichsweise gut in der Lage sind, Leistungsrangfolgen von Schülerinnen und Schülern sowie Schwierigkeitsrangfolgen von Aufgaben zu beurteilen, zeigen sowohl die Ergebnisse von Teilstudie 1 als auch das Befundmuster vieler weiterer Studien, dass Lehrerurteile zu absoluten Merkmalsausprägungen (Niveaukomponente)

oftmals ungenau sind (vgl. 2.1.5 sowie Feinberg & Shapiro, 2009). Dementsprechend naheliegend ist es, bei wichtigen diagnostischen Entscheidungen nicht allein den Einschätzungen von Lehrkräften zu vertrauen, sondern noch weitere diagnostische Informationsquellen zu berücksichtigen.

Vor diesem Hintergrund wurde in Teilstudie 2 am Beispiel von Selektionsentscheidungen zur Teilnahme an additiven Sprachfördermaßnahmen untersucht, inwiefern die hierbei erzielte Klassifikationsgüte mit den verschiedenen diagnostischen Informationsquellen kovariiert, die von den untersuchten Grundschulen zur Identifikation von Kindern mit Sprachförderbedarf genutzt werden. Ein positiver Effekt auf die Klassifikationsgüte wurde einzig für die Verwendung von sprachdiagnostischen Verfahren ermittelt. Dieser Befund ist nicht zuletzt deswegen bedeutsam, weil dem schulischen Einsatz von Tests im Allgemeinen und sprachdiagnostischen Verfahren im Speziellen, wie etwa in Teilkapitel 2.2 und in Kapitel 3 dargestellt, in der Praxis noch immer mit einer verbreiteten Skepsis begegnet wird. Demgegenüber verdeutlichen die Ergebnisse von Teilstudie 2, dass der Einsatz sprachdiagnostischer Verfahren in Grundschulen sinnvoll und angemessen ist.

Dass in Teilstudie 2 auch für die Gruppe der Grundschulen, die sprachdiagnostische Verfahren nutzen, nur eine verhältnismäßig geringe Sensitivität gefunden werden konnte, wurde in Teilkapitel 7.1, unter Bezugnahme auf die Ergebnisse von Teilstudie 3, mit der starken Heterogenität der eingesetzten Instrumente und deren zum Teil geringen psychometrischen Güte sowie mit Verweis auf die oftmals geringe inhaltliche Passung zwischen der Sprachdiagnostik und der Sprachförderung an den Grundschulen erklärt. Dabei erscheint es angesichts der dort dargelegten Einschränkungen sogar bemerkenswert, dass in Teilstudie 2 überhaupt ein positiver Zusammenhang zwischen der Nutzung von Informationen aus sprachdiagnostischen Verfahren und der Klassifikationsgüte bei Sprachförderentscheidungen festgestellt wurde.

Hierzu ist zunächst festzuhalten, dass die in den Zeitschriftenbeiträgen zu den Teilstudien 2 und 3 sowie in Teilkapitel 7.1 formulierten Zweifel an der psychometrischen Güte nicht alle, sondern nur einen Teil der an den Grundschulen genutzten sprachdiagnostischen Verfahren betreffen. Mit Blick auf das skizzierte Passungsproblem zwischen Diagnostik und Förderung ist ferner anzuführen, dass zwar vielfach nur ein Ausschnitt der sprachlichen Fähigkeiten von Schülerinnen und Schülern (d. h. oftmals nur schriftsprachliche Kompetenzen) getestet wird, einzelne sprachliche

Kompetenzen jedoch hoch miteinander korreliert sind. So konnten etwa Behrens, Böhme und Krelle (2009) anhand von Daten aus einer Pilotierungsuntersuchung zu den Bildungsstandards im Fach Deutsch für den Primarbereich zwischen den Kompetenzbereichen Lesen und Zuhören eine latente Korrelation von $r = .74$ feststellen. Hieraus lässt sich schlussfolgern, dass die Ergebnisse von Lesetests (wenn auch nur eingeschränkt prognostisch valide) Hinweise auf die Ausprägung der Zuhörkompetenz von Schülerinnen und Schülern geben können.

Ein weiterer Erklärungsansatz für den in Teilstudie 2 gefundenen Effekt beinhaltet schließlich, dass der Einsatz sprachdiagnostischer Verfahren – im Gegensatz zur Nutzung der weiteren, ebenfalls untersuchten diagnostischen Informationsquellen – mit einer statistischen bzw. mechanischen Urteilsbildung einhergehen dürfte. Im Unterschied zu den Expertenurteilen von Lehrkräften oder anderen Pädagoginnen und Pädagogen basiert diese Form der Urteilsbildung auf expliziten, standardisierten Regeln. Wie unter 2.1.6.2 dargestellt, ist in empirischen Forschungsarbeiten vielfach gezeigt wurden, dass Urteile, denen eine solche statistische Urteilsbildung zugrunde liegt, im Mittel durch eine höhere Validität gekennzeichnet sind als „klinische“ Einschätzungen, die ohne einer auf expliziten Regeln basierenden Grundlage von Experten vorgenommen wurden (z. B. Meehl, 1954).

Wie bereits erläutert, wurde der in Teilstudie 2 für den Einsatz sprachdiagnostischer Verfahren berichtete Effekt auf die Klassifikationsgüte nur in Bezug auf den Gütekennwert der Sensitivität gefunden. Für den Kennwert der Spezifität konnten hingegen keine statistisch signifikanten Zusammenhänge mit der Nutzung verschiedener diagnostischer Informationsquellen ermittelt werden. Vielmehr war insgesamt (d. h. unabhängig von den verwendeten Informationsquellen) eine hohe Spezifitätsrate zu verzeichnen. Diese hohen Spezifitätswerte dürften auf die Basisrate zurückzuführen sein, also auf den Anteil an Kindern mit einem tatsächlichen sprachlichen Förderbedarf in den untersuchten Grundschulen. Fällt die Basisrate, wie in Teilstudie 2, bei der nur rund 13 Prozent der Schülerinnen und Schüler (nach der auf den Kompetenzstufen basierenden Definition) einen sprachlichen Förderbedarf hatten, eher gering aus, sind hohe Spezifitätsraten nicht schwer zu erreichen: Wenn gleichzeitig nur wenige Schülerinnen und Schüler positiv (Diagnose: Sprachförderbedarf) beurteilt werden, die sogenannte Selektionsquote also ebenfalls gering ausfällt, muss der Anteil der Kinder an allen negativ Beurteilten (Diagnose: kein Sprachförderbedarf), die tatsächlich keinen Förderbedarf

haben (d. h. der Anteil der Wahr-negativen an allen Kindern ohne Förderbedarf), „automatisch“ hoch ausfallen. Allerdings sind hohe Spezifitätswerte durchaus trügerisch, wenn gleichzeitig die Basisrate gering ist. Würde man etwa den sprachlichen Förderbedarf einer größeren Anzahl von Schülerinnen und Schülern auf Basis der in Teilstudie 2 festgestellten Gütekennwerte (d. h. Sensitivität von ca. 30 Prozent, Spezifität von ca. 93 Prozent) und Randbedingungen (d. h. Basisrate von ca. 13 Prozent) diagnostizieren, dann wäre zu erwarten, dass rund 60 Prozent der positiv beurteilten Kinder in Wirklichkeit keinen Förderbedarf haben. Aufgrund der geringen Basisrate wäre dieser Anteil auch dann noch recht hoch, wenn die Sensitivitätsrate deutlich höher ausfiele. So wäre selbst bei einer Sensitivität von 80 Prozent noch immer ein Anteil von 37 Prozent falsch-positiver Fälle an allen positiv diagnostizierten Kindern zu erwarten.

Des Weiteren ist zu vermuten, dass die hohe Spezifitätsrate zum Teil auch durch die Art der Definition des Sprachförderbedarfs in Teilstudie 2 bedingt ist. So wurde das Nichtvorliegen von Förderbedarf nur für diejenigen der in der IQB-Ländervergleichsstudie 2011 untersuchten Schülerinnen und Schüler angenommen, die sowohl im Lesen als auch im Zuhören mindestens den Regelstandard erreichen konnten. Das Vorhandensein sprachlichen Förderbedarfs wurde dann erwartet, wenn der Mindeststandard in mindestens einem der beiden Bereiche verfehlt und im jeweils anderen Bereich nicht übertroffen wurde. In dieser Definition wurden einige Schülerinnen und Schüler ausgelassen – und zwar diejenigen, die zwar in beiden Bereichen den Mindeststandard, nicht jedoch den Regelstandard erreichten. Die betreffenden Kinder wurden deswegen nicht berücksichtigt, da sie als Grenzfälle gelten müssen. Zum einen fallen ihre sprachlichen Kompetenzen zwar deutlich hinter denen ihrer Mitschülerinnen und Mitschüler zurück, sodass eine spezifische Förderung durchaus angezeigt erscheint. Zum anderen erreichen sie jedoch jeweils den Mindeststandard und sollten somit zumindest über die basalen sprachlichen Voraussetzungen verfügen, um in der Grundschule erfolgreich lernen zu können. Eine Folge der beschriebenen Nichtberücksichtigung von Grenzfällen ist, dass hierdurch die Separation von Schülerinnen und Schülern mit und ohne Förderbedarf geschärft wurde. Die Gruppe der Kinder ohne Förderbedarf umfasste somit überwiegend Fälle, für die offenbar recht einfach zu erkennen war, dass die Teilnahme an einer additiven Sprachfördermaßnahme nicht notwendig ist. Dies zeigt sich zumindest an der hohen Rate von Wahr-negativen Fällen, aus der wiederum die insgesamt hohen Spezifitätswerte resultieren.

Demgegenüber fallen die in Teilstudie 2 jeweils gefundenen Sensitivitätsraten verhältnismäßig gering aus. Dabei ist allerdings nicht auszuschließen, dass die in Teilstudie 2 berichteten Sensitivitätsraten die wahre Ausprägung der Kennwerte etwas unterschätzen. Denkbar ist zum Beispiel, dass bei einigen der als falsch-negativ klassifizierten Fälle durchaus ein Förderbedarf identifiziert wurde, allerdings aufgrund von Kapazitätsproblemen keine Förderung erfolgte. Weiterhin erscheint möglich, dass bei einigen der falsch-negativen Fälle in dem Wissen, dass die Schülerin bzw. der Schüler außerhalb der Schule an einer Fördermaßnahme teilnimmt, eine zusätzlich zum Regelunterricht stattfindende schulische Förderung des Kindes als nicht mehr notwendig erachtet wurde. Da in beiden Szenarien jeweils auf eine additive schulische Fördermaßnahme verzichtet wird bzw. werden muss, wären die betreffenden Fälle in Teilstudie 2 als falsch-negativ codiert worden, obwohl jeweils korrekt erkannt wurde, dass ein Förderbedarf vorliegt (wahr-positiv).

Im Zusammenhang mit der sich für die beschriebenen Grenzfälle ergebenden Frage, ob die betreffenden Schülerinnen und Schüler eine zusätzliche Sprachförderung erhalten sollten oder nicht, sei darauf hingewiesen, dass das der Teilstudie 2 zugrundeliegende Prinzip einer Selektionsdiagnostik in der Forschungsliteratur aus einer pädagogischen bzw. förderdiagnostischen Perspektive heraus zum Teil kritisch betrachtet wird (z. B. Kracht, 2003). So wird etwa befürchtet, dass eine Selektionsdiagnostik den Prinzipien der Inklusionspädagogik entgegenwirkt, da eine Gruppe von Schülerinnen und Schülern ausgeschlossen bzw. gesondert behandelt wird. Auch wird argumentiert, dass eine Selektionsdiagnostik den Schülerinnen und Schülern gegenüber unfair sei, für die aufgrund ihrer Kompetenzausprägungen kein Förderbedarf festgestellt wurde, da auch diese ein Recht darauf hätten, eine angemessene Sprachförderung zu erhalten.

Angesichts der skizzierten Bedenken soll an dieser Stelle betont werden, dass die inhaltliche Fokussierung von Teilstudie 2 auf die Güte der Selektionsdiagnostik zum Sprachförderbedarf nicht als Parteinahme für einen Ausschluss sprachauffälliger Kinder aus dem Regelunterricht bzw. gegen deren Inklusion missverstanden werden darf. Vielmehr wird die Position vertreten, dass eine zusätzliche Förderung dabei helfen soll, die sprachliche Kompetenzen der betreffenden Schülerinnen und Schüler so zu stärken, dass sie im Regelunterricht erfolgreich lernen können. Selektionsdiagnostik sollte also dem Leitspruch „exkludieren, um zu inkludieren“ dienen. Sprachförderung sollte darüber hinaus nicht ausschließlich additiv erfolgen, sondern primär, beispielsweise im Sinne des

Konzepts der durchgängigen Sprachbildung (z. B. FörMig-Transfer Berlin, 2009; Gogolin et al., 2011), im Regelunterricht implementiert sein.

Wie bereits weiter oben dargestellt, leisten die in Teilstudie 3 ermittelten Befunde einen wichtigen Beitrag zum Verständnis der Ergebnisse von Teilstudie 2, da sie abbilden, welche konkreten sprachdiagnostischen Verfahren an den zuvor betrachteten Grundschulen zum Einsatz kommen. Damit stellt die Teilstudie 3 die bislang einzige Forschungsarbeit dar, in der bundesweit untersucht wurde, welche diagnostischen Verfahren an Grundschulen eingesetzt werden, um festzustellen, ob Schülerinnen und Schüler einen sprachlichen Förderbedarf haben oder nicht. Bisherige Untersuchungen zur Sprachdiagnostik an Grundschulen waren entweder nur auf ein Bundesland beschränkt (Hessen) und betrachteten nicht den Primarbereich als vielmehr die vorschulische Sprachdiagnostik, die zumindest in Hessen in den Verantwortungsbereich der Grundschulen fällt (Geist, 2014), oder basierten, wie der Bericht des ZUSE-Instituts, auf Befragungen der zuständigen Ländereinrichtungen sowie auf ergänzenden Literaturrecherchen (Redder et al., 2011).

Insgesamt verdeutlichen die in Teilstudie 3 ermittelten Ergebnisse, dass die Heterogenität des Vorgehens von Grundschulen beim Einsatz von sprachdiagnostischen Verfahren zur Feststellung sprachlichen Förderbedarfs noch sehr viel größer ist, als die Angaben des ZUSE-Berichts (Redder et al., 2011) oder die Befunde von Geist (2014) vermuten lassen. Die Ergebnisse von Teilstudie 3 zeigen ferner die starke Fokussierung der in den Grundschulen eingesetzten Verfahren auf die Erfassung schriftsprachlicher Kompetenzen bzw. die Vernachlässigung mündlicher Kompetenzen bei der Sprachdiagnostik. Als Grund für diese Fokussierung ist zu vermuten, dass die Vermittlung schriftsprachlicher Kompetenzen als eine der zentralen Aufgaben des Unterrichts in der Grundschule betrachtet wird. Dementsprechend häufig werden an den Grundschulen diagnostische Verfahren eingesetzt, die der Erfassung der Lese- oder Rechtschreibkompetenz dienen. Demgegenüber scheint sich die Bedeutung, die insbesondere dem Zuhören für den schulischen Lernerfolg zukommt (vgl. Böhme, 2012), nicht in den an Grundschulen verwendeten sprachdiagnostischen Instrumente widerzuspiegeln. Allerdings mangelt es auch an Verfahren, die zur Diagnostik mündlicher Kompetenzen im Grundschulalter geeignet wären (vgl. Ehlich, 2005b).

Des Weiteren findet sich im Zeitschriftenbeitrag zu Teilstudie 3 ein umfangreicher Annex, der unter anderem Steckbriefe mit Hintergrundinformationen zur Beurteilung der

Testgüte der am häufigsten eingesetzten Verfahren sowie weitere Angaben dazu umfasst, inwiefern die genutzten Verfahren auf Schülerinnen und Schüler im Grundschulalter zielen, schriftsprachliche Kompetenzen erheben und auf Kinder mit nichtdeutscher Herkunftssprache fokussieren. Im Zusammenhang mit dem letztgenannten Aspekt ist festzuhalten, dass die in den Grundschulen verwendeten Verfahren dem mehrsprachigen Aufwachsen und seinen Besonderheiten in sehr unterschiedlicher Art und Weise Rechnung tragen (z. B. Böhme & Hoffmann, 2014; Jeuk, 2009; Schulz, 2013, vgl. auch 2.3.4). Vielfach findet Mehrsprachigkeit keine gesonderte Berücksichtigung, manchmal werden zusätzliche sprachbiografische Hintergrundinformationen, etwa zum Beginn des Zweitspracherwerbs oder zur Intensität und Qualität des Kontakts mit der Zweitsprache, erfasst. Mit einigen Verfahren kann zusätzlich der Sprachstand in der Erstsprache festgestellt werden. Vereinzelt gibt es auch Verfahren (HAVAS, Cito-Sprachtest), die für Kinder mit nichtdeutscher Herkunftssprache separate Normen (für den Sprachstand im Deutschen) bereitstellen.

Auch in der Forschungsliteratur wird an verschiedener Stelle die Anforderung formuliert, dass die im Elementar- und Primarbereich eingesetzten sprachdiagnostischen Verfahren separate Normen für Kinder mit nichtdeutscher Herkunftssprache zur Verfügung stellen sollten (Becker-Mrotzek et al., 2013; Schulz, 2013). Die Relevanz dieser Forderung ist allerdings differenziert, je nach dem Ziel der Sprachdiagnostik, zu bewerten. Wenn untersucht werden soll, ob eine Schülerin oder ein Schüler die Kriterien einer SSES erfüllt und mithin eine Sprachtherapie erhalten sollte (vgl. 2.3.1), sind separate Normen von Bedeutung. Sie ermöglichen es zu prüfen, ob der Spracherwerb im Deutschen in Relation zu einer relevanten Vergleichsgruppe verzögert ist oder nicht (Schulz, 2013).

Nicht unproblematisch kann die Nutzung separater Normen hingegen im Zusammenhang mit der Diagnostik im Rahmen einer zusätzlich zum Regelunterunterricht durchgeführten Sprachförderung sein. Diese dient der Identifikation von Schülerinnen und Schülern, die stark ausgeprägte (umgebungsbedingte) Sprachauffälligkeiten haben, aufgrund derer sie nicht in optimaler Weise von den schulischen Lerngelegenheiten profitieren können. Die Diagnostik zielt hierbei also auf eine kriteriale Bezugsnorm, die sich auf ein für das schulische Lernen notwendige Mindestniveau an sprachlichen Kompetenzen bezieht. Demgegenüber impliziert jedoch die Forderung nach separaten Normen für Kinder mit nichtdeutscher Herkunftssprache eine soziale Bezugsnorm, da

hierbei das erfasste Kompetenzniveau mit den sprachlichen Kompetenzen von anderen Schülerinnen und Schülern mit nichtdeutscher Herkunftssprache verglichen bzw. anhand dieser relativiert wird. Ob dabei die Testergebnisse ausweisen, dass ein für das schulische Lernen erforderliche Mindestniveau erreicht wird oder nicht, ist indessen nicht Gegenstand dieses Vergleichs. Separate Normen für Schülerinnen und Schüler mit nichtdeutscher Herkunftssprache tragen also wenig dazu bei, Kinder mit einem sprachlichen Förderbedarf zu identifizieren. Vielmehr erhöhen sie unter Umständen die Gefahr, dass eine eigentlich benötigte Förderung nicht erfolgt, da die Ergebnisse, die mithilfe der eingesetzten sprachdiagnostischen Verfahren erhoben werden, aufgrund des sprachlichen Hintergrunds anders bewertet werden als bei monolingualen Kindern.

Folgt man dieser Argumentation, stellt sich zudem die Frage, inwiefern es überhaupt notwendig und angemessen ist, die testtheoretische Güte von sprachdiagnostischen Verfahren, wie an verschiedener Stelle angeführt (Becker-Mrotzek et al., 2013) und auch im Zeitschriftenbeitrag zu Teilstudie 3 dargestellt, auch daran zu bewerten, ob ein Verfahren normiert wurde, sodass es also möglich ist, die jeweils ermittelten Testwerte den Ergebnissen von Normstichproben gegenüberzustellen. Tatsächlich scheinen Kriterien zur Diagnose von sprachlichem Förderbedarf bzw. von umgebungsbedingten Sprachauffälligkeiten nicht selten auf Vergleichen mit Normstichproben zu basieren. So wird, wie unter 2.3.3 skizziert, etwa von Neumann und Euler (2013) vorgeschlagen, Kinder dann als (umgebungsbedingt) sprachauffällig einzustufen, wenn ihre Leistungen eine oder anderthalb Standardabweichungen unter dem Altersdurchschnitt liegen. Angaben wie diese haben allerdings lediglich Empfehlungscharakter, da gegenwärtig keine verbindlichen resp. allgemein geteilten Kriterien dazu existieren, ab welchem Grad an Abweichung von der Altersnorm eine additive Sprachförderung erfolgen sollte. Divergierende Kriterien finden sich im Übrigen auch für die Diagnostik von SSES (AWMF, 2011; WHO, 1992); dementsprechend diskrepant sind die in Forschungsarbeiten berichteten Prävalenzraten (vgl. Law et al., 2000). Konzeptionell ist zu hinterfragen, ob es überhaupt sinnvoll ist, Sprachförderbedarf anhand eines Vergleichs mit altersüblichen Durchschnittsleistungen bestimmen zu wollen. Sollten zur Sprachdiagnostik nicht vielmehr kriteriumsorientierte Tests entwickelt und eingesetzt werden, die prüfen, ob ein inhaltlich relevantes (z. B. für schulisches Lernen erforderliches) Mindestniveau erreicht oder verfehlt wird?

Unabhängig von dieser Frage sind als zentrale Ergebnisse von Teilstudie 3 hervorzuheben, dass (1) an nicht wenigen Grundschulen sprachdiagnostische Verfahren eingesetzt werden, deren psychometrische Güte gering oder unbekannt ist, und (2) mit den genutzten Instrumenten in der Regel nur ein kleiner Ausschnitt der sprachlichen Fähigkeiten erfasst wird (z. B. ausschließlich orthografische Kompetenz). Da aber in Teilstudie 3 nicht Lehrerinnen und Lehrer, sondern Schulleiterinnen und Schulleiter befragt wurden, erlaubt dieses Befundmuster nur vorsichtige Rückschlüsse auf die Ausprägung der Assessment Literacy von Grundschullehrkräften bzw. darauf, wie gut sie in der Lage sind, geeignete sprachdiagnostische Verfahren auszuwählen (vgl. z. B. Standards zu *assessment purposes* bei DeLuca et al., 2016). Dabei deuten die Ergebnisse insgesamt darauf hin, dass in der schulischen Praxis (und vermutlich auch bei den an der Sprachdiagnostik beteiligten Lehrkräften) Unsicherheiten bezüglich der Auswahl von geeigneten Instrumenten bestehen.

7.3 Methodische Bewertung und Grenzen der Arbeit

Bereits die in den Kapiteln 4 bis 6 dargestellten Zeitschriftbeiträge beinhalten Abschnitte, in denen sowohl die Stärken als auch die zentralen methodischen Schwächen der drei in dieser Dissertationsschrift zusammengefassten Teilstudien jeweils knapp erläutert werden. An dieser Stelle soll noch auf einige weitere Aspekte hingewiesen werden, die in den Einzelbeiträgen entweder nicht oder wenig ausführlich diskutiert werden konnten.

Wie bereits in Teilkapitel 7.1 erwähnt wurde, zeigen die Ergebnisse der in Teilstudie 1 durchgeführten statistischen Analysen unter anderem, dass die Tendenz zur Unter- oder Überschätzung mit der psychometrischen Schwierigkeit der von den Lehrkräften zu beurteilenden Deutsch- und Mathematikaufgaben kovariierte. Hierbei wurde festgestellt, dass psychometrisch schwere Aufgaben eher unter- und psychometrisch leichte Aufgaben eher überschätzt werden. Dieses Ergebnis ähnelt zwar den Befunden aus den Studien von McElvany und Kollegen (2009) sowie von Lintorf und Kollegen (2011), die ebenfalls Hinweise auf einen moderierenden Effekt der psychometrischen Aufgabenschwierigkeit auf die Genauigkeit der Schwierigkeitsurteile von Lehrkräften ermitteln konnten. Dennoch sollte der gefundene Zusammenhang mit Vorsicht interpretiert werden. Wie unter 2.1.4 beschrieben, werden in empirischen Untersuchungen zur Akkuratheit von Lehrerurteilen die „tatsächlichen Merkmalsausprägungen“ mit den von Lehrkräften vorgenommenen Einschätzungen

verglichen. Dabei bestand bei Teilstudie 1 die Herausforderung, dass die „tatsächlichen Merkmalsausprägungen“ als Lösungswahrscheinlichkeiten (auf der Ebene der Schulklasse) und somit als kardinalskalierte Daten vorlagen, die Schwierigkeitsurteile der untersuchten Lehrkräfte jedoch mithilfe sechsstufiger Ratingskalen und somit auf Ordinalskalenniveau erfasst wurden. Um zu bestimmen, ob die untersuchten Lehrkräfte in der Lage waren, akkurat zu beurteilen, wie schwierig die ihnen vorgelegten Deutsch- und Mathematikaufgaben für die Schülerinnen und Schüler ihrer Klassen sind, mussten die Lösungswahrscheinlichkeiten folglich ebenfalls in sechs ordinal skalierte Kategorien überführt werden. Für den Vergleich dieser Schwierigkeitskategorien mit den Lehrerratings wurde darüber hinaus ein Toleranzbereich definiert. Dies bewirkte, dass Abweichungen in den Ausprägungen der beiden Variablen von einer Kategorie noch nicht als Fehleinschätzungen klassifiziert wurden.

Die Festlegung eines Toleranzbereichs erschien erforderlich, weil die Lehrkräfte die Aufgabenschwierigkeiten anhand von qualitativ beschriebenen Kategorien (von 1 = „sehr leicht“ bis 6 = „sehr schwer“) beurteilten, die kategorisierten Lösungswahrscheinlichkeiten hingegen quantitativ (von 1 = „100 % – 83.3 %“ bis 6 = „16.7 % – 0 %“) verankert waren. Somit war eine exakte Abbildung der Lehrerratings auf die Schwierigkeitskategorien nicht ohne weiteres möglich. Zudem erschien die Definition eines Toleranzbereichs auch deswegen sinnvoll, da, wie unter 2.1.4 angemerkt, durchaus kontrovers diskutiert werden kann, welcher Grad an Abweichung zwischen Lehrerurteilen und tatsächlichen Merkmalsausprägungen als inhaltlich relevante Fehleinschätzung zu bewerten ist bzw. bis zu welcher Diskrepanz noch von einer akkuraten Einschätzung gesprochen werden kann (vgl. Coladarci, 1986; Südkamp et al., 2012). Dieser Herausforderung wurde in Teilstudie 1 mit der Wahl eines verhältnismäßig breiten Toleranzbereichs begegnet, sodass folglich nur größere Abweichungen als Unter- bzw. Überschätzung eingestuft wurden.

Bedingt durch die Nutzung von ordinal geordneten Kategorien und durch die Wahl eines breiten Toleranzbereichs bestand allerdings die Möglichkeit, dass der Vergleich von Lehrerratings und Lösungswahrscheinlichkeiten von Boden- und Deckeneffekten überlagert wird: So konnte sowohl eine Unterschätzung besonders leichter Aufgaben (d. h. hohe Lösungswahrscheinlichkeit) als auch eine Überschätzung besonders schwerer Aufgaben (d. h. geringe Lösungswahrscheinlichkeit) nicht mehr abgebildet werden. Es ist dementsprechend zu vermuten, dass der in Teilstudie 1 gefundene und oben skizzierte

Zusammenhang zwischen der psychometrischen Schwierigkeit einer Aufgabe einerseits und der Tendenz zur Unter- oder Überschätzung andererseits zumindest teilweise auf einen Messfehler bzw. auf die Art und Weise zurückzuführen ist, wie bestimmt wurde, ob eine Lehrkraft die Schwierigkeit einer Aufgabe akkurat beurteilt oder aber unter- bzw. überschätzt hat.

In Teilstudie 1 wurde darüber hinaus festgestellt, dass die ermittelte Urteilsgenauigkeit nur in geringem Maße mit der Berufserfahrung der untersuchten Lehrkräfte kovariierte. Wie unter 7.1 dargestellt, korrespondiert auch dieses Ergebnis mit den Befundmustern anderer Studien (z. B. Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003; Schrader, 1989). Bei einem Blick auf die Resultate aus Teilstudie 1 fällt allerdings auf, dass die von den befragten Lehrkräften angegebene Lehrerfahrung sowohl im Mittel recht hoch ausfällt, als auch durch eine verhältnismäßig große Standardabweichung gekennzeichnet ist. Folglich könnte vermutet werden, dass die geringen Zusammenhänge auch durch die dergestalt schiefe Verteilung der Prädiktorvariable „Berufsdauer in Jahren“ bedingt sein könnten. Um diese Alternativerklärung zu prüfen, wurde diese Variable in einer zusätzlichen Analyse der Daten, deren Ergebnisse nicht im Zeitschriftenbeitrag zu Teilstudie 1 (Kapitel 4), sondern im Anhang A dieser Arbeit zu finden sind, dummycodiert. Hierbei wurde allen Fällen, in denen eine Berufserfahrung von weniger als 10 Jahren berichtet wurde, der Wert „0“ zugewiesen. Alle übrigen Fälle wurden mit „1“ codiert. Theoretisch abgeleitet wurde diese Klassifikation aus Befunden der psychologischen Expertiseforschung, nach denen es mindestens zehn Jahre dauert, bis eine Person in einer bestimmten Domäne einen Expertenstatus erreicht hat (z. B. Ericsson, Charness, Feltovich & Hoffman, 2006). Allerdings zeigen die im Anhang A zu dieser Dissertationsschrift dargestellten Ergebnistabellen, dass auch bei einer statistischen Auswertung der Daten aus Teilstudie 1 mit dichotom codierter Berufserfahrung nur geringe, statistisch nicht signifikante Zusammenhänge mit der Unter- oder Überschätzung der Aufgabenschwierigkeit festzustellen sind.

Schließlich ist darauf hinzuweisen, dass es im Hinblick auf die Generalisierbarkeit der Ergebnisse von Teilstudie 1 wünschenswert gewesen wäre, Schwierigkeitsurteile zu einer deutlich höheren Anzahl an verschiedenen Aufgaben zur Verfügung zu haben. Dies hätte zudem ermöglicht, die Akkuratheit von Schwierigkeitseinschätzungen differenziert nach bestimmten Aufgabenmerkmalen zu betrachten. Beispielsweise hätte untersucht werden können, ob Lehrkräften die Beurteilung der Schwierigkeit bei Aufgaben aus

bestimmten Kompetenzbereichen besonderes leicht oder schwer fällt oder inwiefern die Urteilsgenauigkeit je nach Aufgabenformat (z. B. geschlossen, halb-offen, offen) variiert.

In Teilstudie 2 wurde die bei der Sprachdiagnostik an Grundschulen erzielte Klassifikationsgüte jeweils bezogen auf die Schulebene und nicht auf der Individualebene, d. h. für einzelne Schülerinnen und Schüler, ermittelt. Hauptmotiv hierfür war, dass der in der IQB-Ländervergleichsstudie 2011 eingesetzte Leistungstest, aus dessen Ergebnissen der sprachliche Förderbedarf abgeleitet wurde, nicht für die Individualdiagnostik, sondern vielmehr zur Beschreibung von Verteilungen auf höherer Aggregationsebene (d. h. insbesondere auf Ebene der Bundesländer) entwickelt wurde. Als Maßzahlen für die Klassifikationsgüte wurden die Sensitivität und Spezifität der Klassifikation berechnet. Diese Kennwerte wurden jeweils aus den für eine Grundschule ermittelten Kompetenzstufenverteilungen (a) für die Schülerinnen und Schüler, die eine additive Sprach- und Leseförderung erhalten, und (b) für die Kinder, die nicht additiv gefördert werden, abgeleitet. Da eine Auswertung auf der Individualebene nicht angemessen erschien, wurde hingegen nicht bestimmt, ob einzelne Schülerinnen und Schüler richtig (wahr-positiv, wahr-negativ) oder falsch (falsch-positiv, falsch-negativ) zugeordnet wurden. Dadurch bedingt war es auch nicht möglich, alternative Kenngrößen für die Klassifikationsgüte zu ermitteln, die auf diesen Informationen basieren und Kennwertevergleiche unabhängig von Selektionsquoten *und* Basisraten erlauben. Zu diesen Alternativen zählt beispielsweise auch der so genannte RATZ-Index (Marx, 1992) bzw. RIOC-Index (in der englischsprachigen Fachliteratur, z. B. Copas und Loeber (1990); Farrington und Loeber (1989); Loeber und Dishion (1983)), bei dem die Gesamttrefferquote (also die Summe der „wahr-positiven“ und „wahr-negativen“ Urteile) an der Zufallstrefferquote relativiert wird (vgl. auch Tröster, 2009).

Bereits unter 7.1 und 7.2 wurde auf die Relevanz von Passungsproblemen zwischen Diagnostik und Förderung hingewiesen, die immer dann bestehen, wenn diese auf jeweils andere sprachliche Kompetenzen zielen. Eine geringe Passung zwischen Diagnostik und Förderung ist, wie bereits erwähnt, auch im Hinblick auf das Design von Teilstudie 2 festzustellen. Auf der einen Seite wurde dort der sprachliche Förderbedarf auf der Grundlage der beiden *rezeptiven* sprachlichen Kompetenzen Lesen und Zuhören bestimmt. Auf der anderen Seite wurde die Partizipation an additiven Fördermaßnahmen in der Schülerteilnahmeliste im Hinblick auf Sprach- und Leseförderung *insgesamt* erfasst und mithin nicht nur bezogen auf rezeptive sprachliche Kompetenzen, sondern

mindestens auch für produktive mündliche Kompetenzen erhoben. Bedingt durch die dergestalt undifferenzierte Abfrage der Teilnahme an additiven Fördermaßnahmen bestand für die Bestimmung der Klassifikationsgüte darüber hinaus die Notwendigkeit, die in den Leistungstests der IQB-Ländervergleichsstudie im Lesen und Zuhören jeweils erreichten Kompetenzstufen miteinander zu kombinieren.

Um bei der Untersuchung der in Teilstudie 2 fokussierten Fragestellungen nicht auf diesen Notbehelf einer Kompetenzstufenkombination zurückgreifen zu müssen und das skizzierte Passungsproblem zu vermeiden, wären weitere Hintergrundinformationen zu den additiven Sprachfördermaßnahmen erforderlich gewesen, die an den Grundschulen aus der IQB-Ländervergleichsstudie 2011 realisiert werden. Hilfreich wären insbesondere Daten dazu gewesen, welche sprachlichen Kompetenzen in den jeweils von den Schülerinnen und Schülern besuchten Maßnahmen schwerpunktmäßig gefördert werden. Mithilfe dieser Angaben hätte die Möglichkeit bestanden, die Klassifikationsgüte separat bzw. spezifisch für die einzelnen Kompetenzbereiche zu bestimmen, also beispielsweise zu prüfen, ob diejenigen Kinder, die im *Lesen* den Mindeststandard nicht erreichen, auch tatsächlich eine zusätzliche *Lese*förderung erhalten.

Schließlich wurden unter 7.2 die Ergebnisse von Teilstudie 2 auch vor dem Hintergrund der Befunde aus Teilstudie 3 diskutiert. Hierbei wurde deutlich, dass die Variable „sprachdiagnostische Verfahren“ in Teilstudie 2 eine große Anzahl von Instrumenten umfasste, die sich unter anderem hinsichtlich der erfassten Kompetenzen in Bezug auf den fokussierten Altersbereich oder im Hinblick auf die psychometrische Güte zum Teil deutlich unterscheiden. Tatsächlich wäre es aufgrund dieser Heterogenität wünschenswert gewesen, in den in Teilstudie 2 vorgenommenen statistischen Analysen stärker zu differenzieren und beispielsweise zu untersuchen, welche Klassifikationsgüte für die Nutzung einzelner, ganz bestimmter sprachdiagnostischer Verfahren festzustellen ist, oder zu prüfen, ob sich der Einsatz von Instrumenten mit hoher psychometrischer Güte in einer besseren Sensitivitäts- oder Spezifitätsrate widerspiegelt. Jedoch war eine solche Differenzierung nicht zielführend möglich: Ein sehr hoher Verbreitungsgrad wurde in Teilstudie 3 insbesondere für die „Hamburger Schreib-Probe“ (HSP, z. B. May, 2002) ermittelt. Dieses Instrument dient jedoch ausschließlich der Erfassung orthografischer Kompetenz. Für diese Kompetenz lagen im Datensatz der IQB-Ländervergleichsstudie 2011 allerdings keine Informationen dazu vor, ob einzelne Schülerinnen und Schüler eine additive Förderung erhalten oder nicht (vgl. auch das oben skizzierte Passungsproblem

zwischen Diagnostik und Förderung). Von einer separaten statistischen Auswertung an den Schulen, in denen die HSP zum Einsatz kommt, wurde folglich abgesehen.

Angesichts der Anzahl der Schulen, die den „Stolperwörter-Lesetest“ (Metze, 2005) einsetzen, wäre auch für dieses Instrument eine separate Berechnung der Klassifikationsgüte denkbar gewesen. Wie jedoch bereits im Zeitschriftenbeitrag zu Teilstudie 3 dargestellt, erscheint die psychometrische Güte des Stolperwörter-Lesetests zweifelhaft, sodass auch hier auf eine gesonderte Auswertung verzichtet wurde. Verhältnismäßig oft berichteten die befragten Schulleiterinnen und Schulleiter außerdem den Einsatz des Beobachtungsbogens „Sismik“, der allerdings nicht auf Schülerinnen und Schüler im Grundschulalter, sondern auf jüngere Kinder im Alter von 3.5 Jahren bis zur Einschulung fokussiert. Dementsprechend wurden auch für den Sismik keine separaten statistischen Analysen durchgeführt. Alle übrigen Instrumente wurden jeweils an einer geringen Anzahl von Grundschulen eingesetzt. Eine Einzelbetrachtung erschien daher auch in diesen Fällen nicht angemessen zu sein.

Neben den in den Zeitschriftenbeiträgen und in diesem Teilkapitel erläuterten Limitationen weisen die drei in dieser Dissertationsschrift zusammengefassten Teilstudien auch einige methodische Stärken auf, die an dieser Stelle nochmals unterstrichen werden sollen. So ist mit Blick auf Teilstudie 1 zuvorderst die Größe der Lehrkräftestichprobe hervorzuheben: Während viele (und insbesondere die älteren) der unter 2.1 referierten genauigkeitsorientierten Forschungsarbeiten zur diagnostischen Kompetenz auf den Urteilen von verhältnismäßig wenigen Lehrkräften (d. h. $N < 30$) basieren (z. B. Bates & Nettelbeck, 2001; Begeny et al., 2008; Coladarci, 1986), wurde in Teilstudie 1 eine große, bundesweite Stichprobe von Grundschullehrkräften untersucht. Als eine weitere Stärke von Teilstudie 1 ist außerdem zu betonen, dass alle drei Komponenten der Genauigkeit von Schwierigkeitsurteilen untersucht wurden, während sich andere Studien zumeist nur auf die Bestimmung der Rang- oder (deutlich seltener) Niveaukomponente beschränken (z. B. Anders et al., 2010; McElvany et al., 2009). Eine Besonderheit ist zudem, dass die Untersuchung von Zusammenhängen zwischen der Urteilsgenauigkeit und weiteren (z. B. lehrerseitigen) Faktoren bzw. möglichen Kovariaten nicht wie in anderen Untersuchungen (z. B. Karing, 2009) bezogen auf die Rangkomponente, sondern für die Ausprägung der Niveaukomponente vorgenommen wurde. In den betreffenden statistischen Analysen war die Urteilsgenauigkeit mithin nicht als Höhe der Korrelation zwischen Lehrerurteilen und Aufgabenschwierigkeiten

operationalisiert, sondern basierte vielmehr auf der schulpraktisch relevanteren Information, ob die Schwierigkeit einer Aufgabe akkurat beurteilt oder aber unter- oder überschätzt wurde. Dabei wurde der dadurch bedingten Dreistufigkeit der Kriteriumsvariablen Rechnung getragen, indem die Zusammenhänge zwischen der Akkuratheit der Schwierigkeitsurteile einerseits und den betrachteten Kovariaten andererseits als multinomial-logistische Mehrebenenanalysen modelliert wurden.

Auch bei Teilstudie 3 ist die Größe der Schulstichprobe als eine wesentliche Stärke anzuführen. Die Ergebnisse der Untersuchung können somit einen umfassenden Überblick über die bundesweit an Grundschulen eingesetzten sprachdiagnostischen Verfahren geben und sind nicht etwa nur auf ein einziges Bundesland beschränkt (wie z. B. bei Geist, 2014). Darüber hinaus ist hervorzuheben, dass hierbei, im Unterschied zu anderen Überblicksberichten zum Einsatz von sprachdiagnostischen Instrumenten in der Elementar- oder Primarstufe (z. B. Lisker, 2013; Redder et al., 2011), keine (gegebenenfalls um weitere Rechercheergebnisse ergänzte) Befragungen der zuständigen Landeseinrichtungen durchgeführt wurden, sondern direkt an Schulen erhoben wurde.

Im Hinblick auf die Stärken von Teilstudie 2 sei an dieser Stelle insbesondere betont, dass in dieser Forschungsarbeit nicht allein singulär geprüft wurde, inwiefern der Einsatz sprachdiagnostischer Verfahren mit der Klassifikationsgüte von Sprachförderentscheidungen kovariiert. Vielmehr zeichnet sich Teilstudie 2 dadurch aus, dass für ganz verschiedene, den Grundschulen potenziell zur Verfügung stehende, diagnostische Informationsquellen simultan betrachtet wurde, welchen Beitrag sie für die Erhöhung oder Reduzierung von Sensitivitäts- und Spezifitätsraten leisten können. Die Ergebnisse der Studie lassen somit den Schluss zu, dass eine Erhöhung der Klassifikationsgüte von Sprachförderentscheidungen offenbar einzig durch den zusätzlichen Einsatz von sprachdiagnostischen Verfahren erzielt werden kann. Dieser Befund ist, insbesondere vor dem Hintergrund der bereits zuvor skizzierten Skepsis, die im Hinblick auf die Nutzung dieser Instrumente in der Praxis noch immer verbreitet ist, von hoher bildungspolitischer und schulpraktischer Bedeutung.

7.4 Implikationen für die zukünftige Forschung und die schulische Praxis

Implikationen für zukünftige Forschungsarbeiten lassen sich zunächst aus den soeben erläuterten Limitationen der drei Teilstudien dieser Dissertationsschrift ableiten: Zum Beispiel sollte in Studien zur Genauigkeit von Schwierigkeitseinschätzungen (vgl. Teil-

studie 1) darauf geachtet werden, dass die Urteile der Untersuchungsteilnehmerinnen und -teilnehmer kardinalskaliert erfasst werden. So könnten Lehrkräfte etwa aufgefordert werden, konkret (z. B. als Prozentwert) anzugeben, wie hoch die Lösungshäufigkeit in ihrer Klasse für die von ihnen einzuschätzenden Aufgaben jeweils ausfallen wird. Dies würde Vergleiche zwischen Lehrerurteilen einerseits und empirisch gefundenen Aufgabenschwierigkeiten andererseits deutlich erleichtern.

In zukünftigen Studien zur Akkuratheit von Sprachförderentscheidungen (vgl. Teilstudie 2) ist insbesondere zu berücksichtigen, dass die zur Ermittlung der Klassifikationsgüte gegenübergestellten Informationen zum diagnostizierten Förderbedarf und zur Partizipation an sprachlichen Fördermaßnahmen jeweils auf die gleiche Kompetenz zielen sollten. Darüber hinaus sollte in statistischen Analysen zu Kovariaten der Klassifikationsgüte der Einsatz sprachdiagnostischer Verfahren nach Möglichkeit nicht nur durch eine einzelne Variable abgebildet werden. Notwendig ist vielmehr eine differenziertere Modellierung, die der starken Heterogenität der an den Grundschulen genutzten Instrumente Rechnung trägt. Als ein Desiderat könnte hierbei systematisch untersucht werden, welche konkreten Verfahren (und ggf. welche Kombinationen von Verfahren im Sinne einer diagnostischen Strategie) einen empirisch nachweisbaren Beitrag zur Erhöhung der Klassifikationsgüte leisten können.

Schließlich sollten Befragungen zu den an Grundschulen eingesetzten sprachdiagnostischen Verfahren (vgl. Teilstudie 3) nicht auf Schulleiterinnen und Schulleiter zielen, sondern mit denjenigen Lehrkräften durchgeführt werden, die jeweils an der Sprachdiagnostik beteiligt sind und mithin zuverlässigere Angaben zu den verwendeten Instrumenten und insbesondere auch zu selbstentwickelten, hausinternen Methoden machen können.

Weitere Implikationen für die Forschung lassen sich basierend auf den in dieser Dissertationsschrift zusammengefassten Teilstudien und ihren Ergebnissen sowie vor dem Hintergrund des in Kapitel 2 aufgespannten theoretischen Rahmens benennen. In diesem Zusammenhang sei zunächst auf die unter 7.2 diskutierte Frage verwiesen, wie die in Forschungsarbeiten zur Genauigkeit von Lehrerurteilen gefundenen Genauigkeitskennwerte zu bewerten sind. Hierbei wurde unter anderem erläutert, dass die Höhe der in diesen Studien ermittelten Genauigkeitskennwerte in der Regel nicht daran relativiert wird, a) welche Hinweisreize bzw. diagnostischen Informationen den untersuchten Lehrkräften bei ihren Einschätzungen zur Verfügung standen und b) welchen

Vorhersagewert diese Informationen für die jeweils zu beurteilenden Merkmalsausprägungen hatten. Während sich viele genauigkeitsorientierte Forschungsarbeiten zur diagnostischen Kompetenz vor allem darauf beschränken, die Akkuratheit von Lehrerurteilen deskriptiv in statistischen Kennwerten abzubilden und gegebenenfalls zu prüfen, inwiefern diese Kennwerte mit weiteren Faktoren kovariieren (z. B. Lehrermerkmale, Schülermerkmale, Urteilsmerkmale oder Merkmale des eingesetzten diagnostischen Verfahrens, vgl. Südkamp et al., 2012), böte eine Berücksichtigung dieser Aspekte, etwa im Rahmen quasi-experimenteller Studien (wie bei Cooksey et al., 1986), zusätzliche Betrachtungsmöglichkeiten: Als Vergleichsmaßstab zur Bewertung der festgestellten Genauigkeitskennwerte könnte ermittelt werden, welcher Grad an Genauigkeit anhand der zur Verfügung stehenden diagnostischen Informationen überhaupt maximal erreichbar wäre. Darüber hinaus könnte untersucht werden, wodurch sich die Lehrkräfte, die eine hohe Urteilsgenauigkeit erzielen, von denjenigen unterscheiden, die als Diagnostikerinnen und Diagnostiker offenbar weniger erfolgreich sind. Nutzen sie die verfügbaren diagnostischen Informationen in größerem Umfang oder stützen sie sich vielmehr auf einige wenige Hinweisreize mit hohem Vorhersagewert? Sind sie in der Gewichtung der von ihnen genutzten Informationen über eine Serie von Einschätzungen hinweg konsistenter? Fallen die Leistungsurteile von einigen Lehrerinnen und Lehrern auch deswegen verhältnismäßig ungenau aus, weil sie nicht allein die Leistung beurteilen, sondern, ähnlich wie bei der Vergabe von Schulnoten (siehe 2.1.3 und z. B. Kuhl & Hannover, 2012), bei ihren Einschätzungen aufgrund von pädagogischen Motiven noch weitere Informationen (wie etwa die wahrgenommene Anstrengungsbereitschaft) berücksichtigen?

Insbesondere die letztgenannten Forschungsfragen gehen deutlich über eine Fokussierung auf die Genauigkeit von Lehrerurteilen hinaus. Sie implizieren eine prozessorientierte Perspektive auf die diagnostische Kompetenz von Lehrkräften, da sie auf eine Betrachtung derjenigen Informationsverarbeitungsprozesse zielen, die dem eigentlichen Urteil zeitlich vorangehen. Bereits unter 2.2.1 war angeregt worden, in empirischen Studien zur diagnostischen Kompetenz von Lehrkräften nicht ausschließlich die Genauigkeit der von Lehrkräften vorgenommenen Einschätzungen in den Blick zu nehmen, sondern auch andere Aspekte des Urteilsprozesses zu untersuchen (vgl. z. B. auch Schrader, 2013). Anders als in genauigkeitsorientierten Studien, die eine eingeschränkte Sichtweise auf die diagnostische Kompetenz von Lehrkräften haben und vor

allem deskriptive Befunde (d. h. Genauigkeitskennwerte) berichten, könnte somit auch verstärkt den Ursachen und Gründen ungenauer Urteile nachgegangen werden.

Gegenwärtig ist allerdings festzuhalten, dass Untersuchungen, die den prozessorientierten Ansätzen zuzuordnen sind, im Forschungsfeld zur diagnostischen Kompetenz von Lehrkräften gegenüber den genauigkeitsorientierten Studien deutlich unterrepräsentiert sind, sodass an verschiedener Stelle ein entsprechendes Desiderat formuliert wurde (z. B. Krolak-Schwerdt et al., 2009). Die wenigen prozessorientierten Studien, die sich in der Forschungsliteratur finden (Dünnebier et al., 2009; Gräsel et al., 2010; Kishor, 1994; Krolak-Schwerdt et al., 2009, 2012; Krolak-Schwerdt & Rummer, 2005; van Ophuysen, 2006), sind zumeist sozialpsychologisch geprägt: Sie beziehen sich beispielsweise auf das heuristisch-systematische Modell von Chen und Chaiken (1999), auf das Kontinuum-Modell der Eindrucksbildung von Fiske und Neuberg (1990) (s. auch Fiske, 1993), auf das Kovariationsmodell aus der Attributionstheorie von Kelley (1967) oder auf Theorien zu Urteilsheuristiken und -verzerrungen (z. B. Tversky & Kahneman, 1974).[30]

Zumindest in einigen Fällen lassen die Befunde aus bisherigen prozessorientierten Studien die Ergebnisse aus genauigkeitsorientierten Untersuchungen in einem differenzierteren Licht erscheinen. Dies soll im Folgenden *exemplarisch* illustriert werden: So konnte beispielsweise an verschiedener Stelle gezeigt werden, dass die berufliche Erfahrung von Lehrkräften, entgegen der etwa unter 2.1.6.2 beschriebenen Befundlage (z. B. bei Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003; Schrader, 1989) und der Ergebnisse aus Teilstudie 1, für die diagnostische Kompetenz von Lehrkräften offenbar durchaus von Bedeutung sein kann: Wie bereits unter 2.2.1 skizziert, scheinen erfahrene Lehrkräfte gegenüber Berufsanfängern bzw. bzw. Lehramtsstudierenden den Vorteil zu haben, dass sie in der Lage sind, ihre Informationsverarbeitungsstrategie an die von ihnen wahrgenommenen Erfordernisse des situativen Kontexts zu adaptieren, in dem sie ihre Einschätzungen vornehmen. Wenn ihre Urteile nur einen ersten Eindruck über den Leistungsstand von Schülerinnen und Schülern abbilden sollen, nutzen sie in der Regel eine kategoriengeleitete, auf Schemata und Stereotypen basierende („*top-down*“)

[30] Ein Vortrag einer Forschergruppe aus Koblenz-Landau auf der 5. Tagung der Gesellschaft für Empirische Bildungsforschung in Heidelberg zeigt darüber hinaus, dass es mittlerweile auch wieder Bestrebungen gibt, den Urteilsprozess von Lehrkräften bzw. die von ihnen bei der Beurteilung genutzten diagnostischen Informationen, ähnlich wie schon bei Cooksey und Kollegen (1986), auf der Grundlage des Linsenmodells (Brunswik, 1955) zu untersuchen (Wahle et al., 2017).

Informationsverarbeitungsstrategie. Diese hat einerseits den Vorzug, die kognitiven Ressourcen nur in geringem Maße zu beanspruchen. Andererseits führt ist sie zu wenig differenzierten Urteilen und ist anfällig für Verzerrungen. Wenn erfahrene Lehrkräfte jedoch vor der Aufgabe stehen, Urteile vornehmen zu müssen, die für die betreffenden Schülerinnen und Schüler (z. B. Schullaufbahnentscheidungen) oder auch für sie selbst und ihre berufliche Zukunft (z. B. im Kontext von *accountability*) mit bedeutsamen Konsequenzen verbunden sind, verwenden sie hingegen eine auf Genauigkeit und Präzision ausgerichtete, „*bottom-up*“ verlaufende („systematische“ bzw. „merkmalsgeleitete“) Verarbeitungsstrategie. Diese ist zwar kognitiv aufwendig und vor allem aufmerksamkeitsintensiv, dafür aber weniger anfällig für Urteilsverzerrungen. Demgegenüber scheinen Lehrernovizen diagnostische Informationen unabhängig vom Kontext vorrangig kategoriengeleitet zu verarbeiten (vgl. Dünnebier et al., 2009; Gräsel et al., 2010; Krolak-Schwerdt et al., 2009, 2012; Krolak-Schwerdt & Rummer, 2005).

Aus den skizzierten Unterschieden in der Flexibilität der Informationsverarbeitungsstrategie von erfahrenen Lehrkräften und Lehrernovizen lassen sich zudem alternative Ansätze zur Deutung der Befundlage aus genauigkeitsorientierten Studien ableiten: So könnte vermutet werden, dass Untersuchungen zur Genauigkeit von Lehrerurteilen eine kategoriengeleitete, für Verzerrungen anfällige Verarbeitungsstrategie triggern, sofern den Probanden nicht (z. B. durch entsprechende Szenarien) suggeriert wird, dass die von ihnen erbeteten Einschätzungen mit bedeutsamen Konsequenzen verbunden sind. Einhergehend mit dieser Hypothese ließe sich fragen, ob erfahrene Lehrkräfte möglicherweise dazu in der Lage wären, die in vielen Studien berichteten Genauigkeitskennwerte deutlich zu übertreffen, wenn der situative Urteilskontext eine fehlerrobuste merkmalsgeleitete Verarbeitungsstrategie nahelegen würde.

Das angeführte Beispiel veranschaulicht, dass es sinnvoll und wissenschaftlich fruchtbar sein kann, in zukünftigen Studien deutlich stärker als bislang nicht allein die Urteilsgenauigkeit, sondern den gesamten Urteilsprozess zu fokussieren (vgl. auch Klug et al., 2013). Allerdings würden auch dergestalt inhaltlich erweiterte Forschungsarbeiten zur diagnostischen Kompetenz nur einen kleinen Teil des diagnostischen Tätigkeitsfeldes von Lehrkräften repräsentieren. Dieses ist nicht nur auf die Beurteilung von Leistungen und Aufgabenschwierigkeiten beschränkt, sondern umfasst, wie etwa in der Einleitung dieser Dissertationsschrift oder unter 2.2.1 skizziert wurde, noch viele weitere diagnostische Anforderungen. Im Rahmen empirischer Studien zur diagnostischen Kompetenz

könnte dieser Heterogenität Rechnung getragen werden, indem nicht nur die Urteilsgenauigkeit untersucht wird, sondern auch, ob Lehrkräfte in der Lage sind, weitere diagnostische Anforderungen, wie sie etwa in Konzepten zur Assessment Literacy (z. B. DeLuca et al., 2016) beschrieben werden, zu bewältigen. Hierbei könnte beispielsweise betrachtet werden, welche Kenntnisse sie über Urteilstendenzen, Urteilsfehler und Testgütekriterien haben und wie gut es ihnen gelingt, die Güte von Testverfahren zu beurteilen. Auch könnte untersucht werden, ob Lehrkräfte im Stande sind, die je nach diagnostischer Zielstellung geeigneten Instrumente auszuwählen, oder inwiefern sie es vermögen, die in individualdiagnostischen Testverfahren oder bei Lernstandserhebungen wie VERA ermittelten Kennwerte korrekt zu interpretieren und auf dieser Basis angemessene Schlussfolgerungen (z. B. im Hinblick auf Fördermaßnahmen oder für die Unterrichtsentwicklung) zu ziehen (vgl. auch Helmke, Hosenfeld & Schrader, 2004; Richter, ohne Datum). Tatsächlich sind insbesondere solche Fragestellungen zur diagnostischen *Methoden*kompetenz von Lehrkräften bislang nur selten Gegenstand von empirischen Forschungsarbeiten gewesen (Jäger, 2009). Sie sind dementsprechend, etwa auch in Übereinstimmung mit Schrader (2009), als bedeutsame Forschungsdesiderata festzuhalten.

Nicht zuletzt sollte die Forschung zur diagnostischen Kompetenz praktischen Zielen dienen, also auch darauf fokussieren, wie die diagnostischen Kenntnisse und Fähigkeiten von Lehrkräften und Lehramtsstudierenden gefördert und weiterentwickelt werden können. Hierzu finden sich in der Forschungsliteratur verschiedene Ansätze, von denen jedoch nur einige wenige in Trainingsstudien erprobt und empirisch evaluiert wurden. Als ein Beispiel hierfür sei die Untersuchung von Besser, Leiss und Klieme (2015) angeführt, die exemplarisch anhand von kompetenzorientierten Aufgaben zum mathematischen Modellieren zeigen konnten, dass die Teilnahme an einer Fortbildungsmaßnahme, deren Schwerpunkt auf der Auseinandersetzung mit den Lösungsprozessen und Schwierigkeiten von Schülerinnen und Schülern bei der Bearbeitung typischer Modellierungsaufgaben lag, zu einer Stärkung der Expertise von Lehrkräften zu formativem Assessment im Mathematikunterricht führte. In einer anderen Studie konnten Klug, Gerich und Schmitz (2012) Belege für die Wirksamkeit eines am Prozessmodell diagnostischer Kompetenz von Klug und Kollegen (2013) orientierten Trainingsprogramms zur Förderung der diagnostischen Kenntnisse und Fähigkeiten von Hauptschullehrkräften finden. Zusätzlich zum Training wurde in einer Untersuchungsbedingung ein standardisiertes Tagebuch

eingesetzt, das der Unterstützung von Reflexionsprozessen dienen sollte und offenbar eine Stärkung des diagnostischen Wissens der Probanden bewirkte.

Im Hinblick auf die Förderung der Urteilsgenauigkeit von Lehrkräften wird in der deutschsprachigen Forschungsliteratur vielmals (z. B. im Zeitschriftenbeitrag zu Teilstudie 1, bei Rjosk et al., 2011, oder bei Lorenz & Artelt, 2009) ein von Helmke und Kollegen (2004) vorgeschlagenen Ansatz referiert. Kerninhalt dieses Ansatzes ist, dass Lehrkräfte angehalten werden, das Abschneiden der von ihnen unterrichteten Klassen in Lernstandserhebungen (z. B. VERA) zu prognostizieren. Ihre Prognosen sollen sie im Weiteren mit den Angaben zum Abschneiden der Schülerinnen und Schüler vergleichen, die ihnen von den für die Auswertung zuständigen Einrichtungen zur Verfügung gestellt werden. Diesem Vergleich können sie dann entnehmen, wie akkurat die von ihnen vorgenommenen Prognosen waren. Wenn sich hierbei große Diskrepanzen ergeben, sollte dies zu einer kritischen Reflexion über die Akkuratheit der eigenen Urteile und der ihnen zugrunde liegenden Prozesse motivieren. Da die Fixierung auf einen klasseninternen Referenzrahmen als eine Hauptursache für ungenaue Lehrerurteile gilt (z. B. Feinberg & Shapiro, 2009), wird als ein Kriterium zur Wirksamkeit dieses Reflexionsprozesses hervorgehoben, dass die Ergebnisrückmeldung zur Lernstandserhebung Informationen beinhalten muss, die dazu geeignet sind, den Referenzrahmen der Lehrkraft zu erweitern. Hierzu zählen etwa Angaben zu den Leistungen anderer Klassen (z. B. aus anderen Schulen) und insbesondere dazu, inwiefern die Schülerinnen und Schüler bestimmte kriteriale Normen erfüllen (vgl. Lorenz & Artelt, 2009).

Tatsächlich wird der von Helmke und Kollegen (2004) vorgeschlagene Ansatz zur Verbesserung der Diagnosegenauigkeit in Handreichungen zu VERA aufgegriffen[31], in einigen Ländern ist er sogar als freiwilliges Angebot für interessierte Lehrkräfte in den VERA-Erhebungen implementiert[32]. Eine Untersuchung zu den Effekten des Ansatzes steht jedoch aus. Auch für andere Angebote zur Stärkung der diagnostischen Kompetenzen von Lehrkräften, wie etwa für die Studienbriefe, die in dem von der KMK geförderten Projekt *UDiKom* entwickelt wurden (Ade-Thurow et al., 2014), erfolgte keine empirische Wirksamkeitsprüfung.

[31] Zum Beispiel: http://vera-server.uni-landau.de/vera2008/download/VERA_Handreichung_Diagnosegenauigkeit_2008.pdf [03.02.2017].

[32] Zum Beispiel: https://www.kompetenztest.de/download/kt2016-landesbericht.pdf [03.02.2017].

Eine weitere Möglichkeit, Lehrkräfte bei der Bewältigung der diagnostischen Herausforderungen ihres Berufs zu unterstützen, könnte zumindest punktuell darin bestehen, Informationsmaterialien bereitzustellen, die einerseits wissenschaftlich fundiert und andererseits an den Voraussetzungen und Bedürfnissen von Praktikerinnen und Praktikern orientiert sind. So lässt die in Teilstudie 3 gefundene Vielfalt an sprachdiagnostischen Verfahren darauf schließen, dass in vielen Grundschulen und mithin auch bei vielen Lehrkräften Unsicherheiten dazu bestehen, welche Instrumente geeignet sind, um Kinder mit sprachlichem Förderbedarf valide identifizieren zu können. Unterstützung hierbei könnte eine systematische Evaluation der bundesweit am häufigsten eingesetzten Verfahren bieten, deren Ergebnisse so aufbereitet werden müssten, dass sie Grundschulen und Lehrkräfte in verständlicher Weise darüber informieren, welche Instrumente den basalen Anforderungen an sprachdiagnostischen Verfahren (vgl. 2.3.4) genügen. Zur Ergänzung sind ferner flankierende Fortbildungsmaßnahmen denkbar, die sicherstellen sollen, dass die Ergebnisse einer solchen Evaluation auch in der Praxis genutzt werden (können).

Als Vorbild für eine Evaluation könnte der vom Kölner Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache herausgegebene Bericht zur Qualität der Verfahren dienen, die in den einzelnen Bundesländern bei den Sprachstandsfeststellungen im Elementarbereich eingesetzt werden (Neugebauer & Becker-Mrotzek, 2013). In dem Bericht wird die Bewertung der evaluierten Instrumente tabellarisch anhand eines leicht verständlichen Kategoriensystems veranschaulicht, so dass gut ersichtlich ist, hinsichtlich welcher Anforderungen die einzelnen Verfahren Stärken und Schwächen aufweisen. Allerdings ist der Evaluationsbericht des Mercator-Instituts, wie unter 2.3.4 skizziert, in der wissenschaftlichen Fachöffentlichkeit zum Teil kritisch rezipiert worden (z. B. Hoffmann & Böhme, 2014a; Maihack, 2014; Maihack et al., 2014).

Bereits heute wird im Rahmen des Forschungs- und Entwicklungsprogramms „Bildung durch Sprache und Schrift“ (BiSS) ein online verfügbares Informationsportal zu empfehlenswerten sprachdiagnostischen Verfahren zur Verfügung gestellt.[33] Dabei wird auch eine rudimentäre, mittels Ampelfarben visualisierte Evaluation vorgenommen, bei der jeweils geprüft wird, inwiefern die Instrumente bestimmten Minimalstandards (Objektivität, Reliabilität, Validität, Normierung/Interpretationseindeutigkeit) genügen.

[33] Verfügbar unter: http://www.biss-sprachbildung.de/biss.html?seite=145 [13.04.2017].

Allerdings ist festzuhalten, dass eine Evaluation der an Grundschulen häufig eingesetzten sprachdiagnostischen Verfahren zwar möglicherweise eine bessere Orientierung bei der Auswahl geeigneter Instrumente bieten könnte, jedoch im Zusammenhang mit der Nutzung solcher Verfahren noch viele weitere Aufgaben bestehen, für deren Bewältigung diagnostische Kompetenz zwingend erforderlich ist. Exemplarisch seien hierbei die Durchführung und Auswertung von Tests, die Interpretation von Testergebnissen oder die Ableitung von Informationen zum Förderbedarf genannt. Tatsächlich erscheint ein Einsatz von sprachdiagnostischen Verfahren in der Grundschule nur dann sinnvoll, wenn die daran beteiligten Lehrkräfte sowohl im Zuge ihrer Ausbildung als auch durch Fortbildungsmaßnahmen dazu befähigt werden, die damit verbundenen diagnostischen Anforderungen kompetent zu meistern (vgl. auch Bredel, 2005; Geist, 2014; Lengyel, 2012). In diesem Zusammenhang könnte auch darüber nachgedacht werden, ob es nicht sinnvoll wäre, die Komplexität dieser diagnostischen Anforderungen zu reduzieren, indem den Grundschulen eine geringe Anzahl an geeigneten sprachdiagnostischen Verfahren vorgegeben wird. Lehrkräfte könnten dann spezifisch für die Arbeit mit diesen Verfahren aus- und fortgebildet werden.

Implikationen für die Aus- und Fortbildung von Lehrkräften lassen sich allerdings nicht nur für den Bereich der Sprachdiagnostik, sondern vielmehr im Hinblick auf die gesamte Breite der diagnostischen Tätigkeiten und Aufgaben von Lehrerinnen und Lehrern formulieren. Wie bereits in der Einleitung dieser Dissertationsschrift erläutert wurde, sind Lehrkräfte nicht allein nur Pädagoginnen und Pädagogen, sondern auch Diagnostikerinnen und Diagnostiker. Damit sie diese Rolle kompetent ausfüllen können, müssen sie auf die damit verbundenen Tätigkeiten und Aufgaben im Rahmen ihrer Ausbildung und durch geeignete Fortbildungsmaßnahmen angemessen vorbereitet werden. Damit einhergehend besteht für die zukünftige Forschung die Herausforderung, in einem sehr viel stärkeren Maße und Umfang als bislang, Förderkonzepte, Trainingsprogramme oder sogar ganze Ausbildungsmodule zum Aufbau und zur Stärkung von diagnostischen Kenntnissen und Fähigkeiten zu entwickeln und in empirischen Studien zu evaluieren.

Literaturverzeichnis

Ade-Thurow, M., Bos, W., Helmke, A., Helmke, T., Hovenga, N., Lebens, M., Lenske, G., Leutner, D., Pham, G. H., Praetorius, A.-K., Schrader, F.-W., Spoden, C. & Wirth, J. (Hrsg.). (2014). *Aus- und Fortbildung der Lehrkräfte in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität und individuelle Förderung*. Münster: Waxmann.

AEA (2012) = The Association of Educational Assessment – Europe. (2012). *European Framework of Standards for Educational Assessment 1.0*. Verfügbar unter: http://www.aea-europe.net/images/downloads/SW_Framework_of_European_Standards.pdf [25.11.2016].

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

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*(3), 175-193.

Artelt, C. & Gräsel, C. (2009). Gasteditorial: Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*(3–4), 157–160.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz. Testkonzeption und Ergebnisse. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiss (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). Opladen: Leske + Budrich.

Autorengruppe Bildungsberichterstattung. (2012). *Bildung in Deutschland 2012. Ein indikatorengestützter Bericht mit einer Analyse zur kulturellen Bildung im Lebenslauf*. Bielefeld: Bertelsmann.

AWMF (2011) = Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. (2011). *Diagnostik von Sprachentwicklungsstörungen (SES) unter Berücksichtigung umschriebener Sprachentwicklungsstörungen (USES)*. Verfügbar unter: http://www.awmf.org/uploads/tx_szleitlinien/049-006l_S2k_Sprachentwicklungsstoerungen_Diagnostik_2013-06-abgelaufen_01.pdf [07.04.2017].

Bach, A., Wurster, S., Thillmann, K., Pant, H. A. & Thiel, F. (2014). Vergleichsarbeiten und schulische Personalentwicklung - Ausmaß und Voraussetzungen der Datennutzung. *Zeitschrift für Erziehungswissenschaft, 17*(1), 61–84.

Bates, C. & Nettelbeck, T. (2001). Primary School Teachers' Judgements of Reading Achievement. *Educational Psychology, 21*(2), 177–187.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M. & Schiefele, U. (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

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

Baumert, J., Trautwein, U. & Artelt, C. (2003). Schulumwelten — institutionelle Bedingungen des Lehrens und Lernens. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000 — Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261–331). Wiesbaden: VS Verlag für Sozialwissenschaften.

Becker-Mrotzek, M., Ehlich, K., Füssenich, I., Günther, H., Hasselhorn, M., Hopf, M., Jeuk, S., Lengyel, D., Neugebauer, U., Panagiotopoulou, A., Stanat, P. & Wilbert, J. (2013). Qualitätsmerkmale für Sprachstandsverfahren im Elementarbereich. Ein Bewertungsrahmen für fundierte Sprachdiagnostik in der Kita. In Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache (Hrsg.). Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache.

Begeny, J. C. & Buchanan, H. (2010). Teachers' Judgments of Students' Early Literacy Skills Measured by the Early Literacy Skills Assessment: Comparisons of Teachers with and without Assessment Administration Experience. *Psychology in the Schools, 47*(8), 859–868.

Begeny, J. C., Eckert, T. L., Montarello, S. A. & Storie, M. S. (2008). Teachers' Perceptions of Students' Reading Abilities: An Examination of the Relationship between Teachers' Judgments and Students' Performance across a Continuum of Rating Methods. *School Psychology Quarterly, 23*(1), 43–55.

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.

Behrens, U., Böhme, K. & Krelle, M. (2009). Zuhören – Operationalisierung und fachdidaktische Implikation. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 357–375). Weinheim: Beltz.

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*(4), 283–293.

Belgrad, J., Eriksson, B., Pabst-Weinschenk, M. & Vogt, R. (2008). Die Evaluation von Mündlichkeit. Kompetenzen in den Bereichen Sprechen, Zuhören und szenisch Spielen. *Didaktik Deutsch (Sonderheft zum 16. Symposium Deutschdidaktik Kompetenzen im Deutschunterricht), 14*, 20–45.

Bennett, R. E., Gottesman, R. L., Rock, D. A. & Cerullo, F. (1993). Influence of Behavior Perceptions and Gender on Teachers' Judgments of Students' Academic Skill. *Journal of Educational Psychology, 85*(2), 347–356.

Berendes, K., Dragon, N., Weinert, S., Heppt, B. & Stanat, P. (2013). Hürde Bildungssprache? Eine Annäherung an das Konzept "Bildungssprache" unter Einbezug aktueller empirischer Forschungsergebnisse. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven (1)* (S. 17–41). Münster; New York; München; Berlin: Waxmann.

Bernstein, B. (1962). Linguistic Codes, Hesitation Phenomena and Intelligence. *Language and Speech, 5*(1), 31–48.

Bernstein, B. (1964). Elaborated and Restricted Codes: Their Social Origins and Some Consequences. *American Anthropologist, 66*(6), 55–69.

Besser, M., Leiss, D. & Klieme, E. (2015). Wirkung von Lehrerfortbildungen auf Expertise von Lehrkräften zu formativem Assessment im kompetenzorientierten Mathematikunterricht. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(2), 110–122.

Beswick, J., Willms, J. D. & Sloat, E. A. (2005). A Comparative Study of Teacher Ratings of Emergent Literacy Skills and Student Performance on a Standardized Measure. *Education Journal, 136*, 116–137.

Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht, 49*, 219–224.

Böhme, K. (2012). *Methodische und didaktische Überlegungen sowie empirische Befunde zur Erfassung sprachlicher Kompetenzen im Deutschen. Analysen zu den Bildungsstandards im Fach Deutsch für den Primarbereich.* Humboldt Universität zu Berlin, Philosophische Fakultät IV, Berlin.

Böhme, K. & Hoffmann, L. (2014). Sprachstandsdiagnostik bei mehrsprachigen Grundschulkindern-Empirische Befunde zum Einsatz diagnostischer Verfahren in Deutschland. *Zeitschrift für Interkulturellen Fremdsprachenunterricht, 19*(2), 20–39.

Böhme, K., Robitzsch, A. & Busè, A.-K. (2010). Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Textaufgaben. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule* (S. 81–104). Münster: Waxmann.

Bos, W., Klieme, E. & Köller, O. (2010). Vorwort. In W. Bos, E. Klieme & O. Köller (Hrsg.), *Schulische Lerngelegenheiten und Kompetenzentwicklung. Festschrift für Jürgen Baumert* (S. 7–10). Münster: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R. & Walther, G. (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster: Waxmann.

Bourdieu, P. (1977). L'économie des échanges linguistiques. *Langue française, 34*(1), 17–34.

Brattesani, K. A., Weinstein, R. S. & Marshall, H. H. (1984). Student Perceptions of Differential Teacher Treatment as Moderators of Teacher Expectation Effects. *Journal of Educational Psychology, 76*(2), 236–247.

Bredel, U. (2005). Sprachstandsmessung – eine verlassene Landschaft. In K. Ehlich (Hrsg.), *Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Förderung von Kindern mit und ohne Migrationshintergrund* (S. 77–119). Berlin: BMBF.

Brookhart, S. M. (1993). Teachers' Grading Practices: Meaning and Values. *Journal of Educational Measurement, 30*(2), 123–142.

Brookhart, S. M. (1997). A Theoretical Framework for the Role of Classroom Assessment in Motivating Student Effort and Achievement. *Applied Measurement in Education, 10*(2), 161–180.

Brookhart, S. M. (2011). Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice, 30*(1), 3–12.

Brophy, J. E. (1983). Research on The Self-Fulfilling Prophecy and Teacher Expectations. *Journal of Educational Psychology, 75*(5), 631–661.

Brophy, J. E. & Good, T. L. (1986). Teacher Behavior and Student Achievement. In M. Witrock (Hrsg.), *Handbook of Research on Teaching* (S. 328–375). New York: Macmillan.

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

Brunswik, E. (1955). Representative Design and Probabilistic Theory in a Functional Psychology. *Psychological Review, 62*(3), 193–217.

Buck, G. & Tatsuoka, K. (1998). Application of the Rule-Space Procedure to Language Testing: Examining Attributes of a Free Response Listening Test. *Language testing, 15*(2), 119–157.

Bund-Länderinitiative zur Sprachförderung [Schneider, W., Baumert, J., Becker-Mrotzek, M., Hasselhorn, M., Kammermeyer, G., Rauschenbach, T., Roßbach, H.-G., Roth, H.-J. & Rothweiler, M.]. (2012). Expertise "Bildung durch Sprache und Schrift (BISS)".

Campbell, D. T. & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56*(2), 81–105.

Chall, J. S., Jacobs, V. A. & Baldwin, L. E. (1990). *The Reading Crisis: Why Poor Children Fall Behind*. Cambridge, Massachusetts: Harvard University Press.

Chen, S. & Chaiken, S. (1999). The Heuristic-Systematic Model in its Broader Context. In S. C. Y. Trope (Hrsg.), *Dual-Process Theories in Social Psychology* (S. 73–96). New York, NY: Guilford Press.

Cicmanec, K. M., Johanson, G. & Howley, A. (2001). High School Mathematics Teachers: Grading Practice and Pupil Control Ideology. Verfügbar unter: http://files.eric.ed.gov/fulltext/ED453290.pdf [19.07.2016].

Citogroep. (2004). *CITO. Test Zweisprachigkeit*. Arnheim: National Institute for Educational Measurement.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Aufl.). Hillsdale: L. Erlbaum Associates.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*(2), 141–146.

Cooksey, R. W., Freebody, P. & Davidson, G. R. (1986). Teachers' Predictions of Children's Early Reading Achievement: An Application of Social Judgment Theory. *American Educational Research Journal, 23*(1), 41–64.

Copas, J. B. & Loeber, R. (1990). Relative Improvement Over Chance (RIOC) for 2×2 Tables. *British Journal of Mathematical and Statistical Psychology, 43*(2), 293–307.

Corno, L. & Snow, R. E. (1986). Adapting Teaching to Individual Differences Among Learners. In *Handbook of research on teaching.* (S. 605–629). New York, NY: Macmillan.

Cronbach, L. (1955). Processes Affecting Scores on "Understanding of Others" and "Assumed Similarity.". *Psychological Bulletin, 52*(3), 177–193.

Cummins, J. (1984). *Bilingualism and Special Education: Issues in Assessment and Pedagogy*. Clevedon: Multilingual Matters.

Darling-Hammond, L. (1995). Equity Issues In Performance-Based Assessment. In M. T. Nettles & A. L. Nettles (Hrsg.), *Equity and Excellence in Educational Testing and Assessment* (S. 89–114). Dordrecht: Springer Netherlands.

Davis, J. A. (1966). The Campus as a Frog Pond: An Application of the Theory of Relative Deprivation to Career Decisions of College Men. *American Journal of Sociology, 72*(1), 17–31.

de Boer, H., Bosker, R. J. & van der Werf, M. P. C. (2010). Sustainability of Teacher Expectation Bias Effects on Long-Term Student Performance. *Journal of Educational Psychology, 102*(1), 168–179.

Dehn, M. (2011). Elementare Schriftkultur und Bildungssprache. In S. Fürstenau & M. Gomolla (Hrsg.), *Migration und schulischer Wandel: Mehrsprachigkeit* (S. 129–151). Wiesbaden: VS Verlag für Sozialwissenschaften.

DeLuca, C., LaPointe-McEwan, D. & Luhanga, U. (2016). Teacher Assessment Literacy: A Review of International Standards and Measures. *Educational Assessment Evaluation and Accountability, 28*(3), 251–272.

Demaray, M. K. & Elliot, S. N. (1998). Teachers' Judgments of Students' Academic Functioning: A Comparison of Actual and Predicted Performances. *School Psychology Quarterly, 13*(1), 8–24.

Deno, S. L. (1985). Curriculum-Based Measurement: The Emerging Alternative. *Exceptional Children, 52*(3), 219–232.

Doherty, J. & Conolly, M. (1985). How Accurately can Primary School Teachers Predict the Scores of their Pupils in Standardised Tests of Attainment? A Study of some non-Cognitive Factors that Influence Specific Judgements. *Educational Studies, 11*(1), 41–60.

Doherty, M. E. & Kurz, E. M. (1996). Social Judgement Theory. *Thinking & Reasoning, 2*(2–3), 109–140.

Dompnier, B., Pansu, P. & Bressoux, P. (2006). An Integrative Model of Scholastic Judgments: Pupils' Characteristics, Class Context, Halo Effect and Internal Attributions. *European Journal of Psychology of Education, 21*(2), 119–133.

Donato, R. (1994). Collective Scaffolding in Second Language Learning. In J. P. Lantolf & G. Appel (Hrsg.), *Vygotskian Approaches to Second Language Research* (S. 33–56). Westport, CT: Ablex Publishing.

Döring, N. & Bortz, J. (Hrsg.). (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin: Springer.

Dünnebier, K., Gräsel, C. & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung. Eine experimentelle Studie zu Ankereffekten. *Zeitschrift für pädagogische Psychologie, 23*(3-4), 187-195.

Eaves, R. C., Williams, P., Winchester, K. & Darch, C. (1994). Using Teacher Judgment and IQ to Estimate Reading and Mathematics Achievement in a Remedial-Reading Program. *Psychology in the Schools, 31*(4), 261–272.

Eckert, T. L. & Arbolino, A. L. (2007). The Role of Teacher Perspectives in Diagnostic and Program Evaluation Decision-Making. In R. Brown-Chidsey (Hrsg.), *Beyond Labels: Noncategorical Individualized Assessment Methods* (S. 65–81). New York, NY: Guilford.

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*(3), 247–265.

Ehlich, K. (1995). Die Lehre der deutschen Wissenschaftssprache: sprachliche Strukturen, didaktische Desiderate. In H. L. Kretzenbacher & H. Weinrich (Hrsg.), *Linguistik der Wissenschaftssprache* (S. 325–351). Berlin: de Gruyter.

Ehlich, K. (2005a). Eine Expertise zu "Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Sprachförderung von Kindern mit und ohne Migrationshintergrund". In I. Gogolin, U. Neumann & H.-J. Roth (Hrsg.), *Sprachdiagnostik bei Kindern und Jugendlichen mit Migrationshintergrund. Dokumentation einer Fachtagung am 14. Juli 2004 in Hamburg.* (S. 33–50). Münster: Waxmann.

Ehlich, K. (2005b). Sprachaneignung und deren Feststellung bei Kindern mit und ohne Migrationshintergrund: Was man weiß, was man braucht, was man erwarten kann. In K. Ehlich (Hrsg.), *Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Förderung von Kindern mit und ohne Migrationshintergrund* (S. 3–75). Berlin: BMBF.

Ericsson, K. A., Charness, N., Feltovich, P. J. & Hoffman, R. R. (Hrsg.). (2006). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press.

Euler, H. A., Holler-Zittlau, I., Minnen, S., Sick, U., Dux, W., Zaretsky, Y. & Neumann, K. (2010). Psychometrische Gütekriterien eines Kurztests zur Erfassung des Sprachstands 4-jähriger Kinder. *HNO, 58*(11), 1116–1123.

Farrington, D. P. & Loeber, R. (1989). Relative Improvement Over Chance (RIOC) and Phi as Measures of Predictive Efficiency and Strength of Association in 2×2 Tables. *Journal of Quantitative Criminology, 5*(3), 201–213.

Feilke, H. (2012). Bildungssprachliche Kompetenzen – fördern und entwickeln. *Praxis Deutsch, 233*, 4–13.

Feinberg, A. B. & Shapiro, E. S. (2003). Accuracy of Teacher Judgments in Predicting Oral Reading Fluency. *School Psychology Quarterly, 18*(1), 52–65.

Feinberg, A. B. & Shapiro, E. S. (2009). Teacher Accuracy: An Examination of Teacher-Based Judgments of Students' Reading With Differing Achievement Levels. *The Journal of Educational Research, 102*(6), 453–462.

Fisher, C. W., Filby, N. N., Marliave, R. S., Cahen, L. S., Dishaw, M. M., Moore, J. E. & Berliner, D. C. (1978). *Teaching Behaviors, Academic Learning Time, and Student Achievement: Final Report of Phase III-B*. San Francisco, CA: Far West Laboratory for Educational Research and Development.

Fiske, S. T. (1993). Social Cognition and Social Perception. *Annual Review of Psychology, 44*, 155–194.

Fiske, S. T. & Neuberg, S. L. (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. *Advances in Experimental Social Psychology, 23*, 1–74.

Flynn, J. M. & Rahbar, M. H. (1998). Improving Teacher Prediction of Children at Risk for Reading Failure. *Psychology in the Schools, 35*(2), 163–172.

FörMig-Transfer Berlin. (2009). *Wege zur durchgängigen Sprachbildung – Ein Orientierungsrahmen für Schulen*. Berlin: Senatsverwaltung für Bildung, Wissenschaft und Forschung, Verfügbar unter http://www.foermig-berlin.de/materialien/Wege_zur_durchgaengigen_Sprachbildung___.pdf [19.08.2016].

Freedle, R. & Kostin, I. (1993). The Prediction of TOEFL Reading Item Difficulty: Implications for Construct Validity. *Language Testing, 10*(2), 133–170.

Freeman, J. G. (1993). Two Factors Contributing to Elementary School Teachers' Predictions of Students' Scores on the Gates-MacGinitie Reading Test, Level D. *Perceptual and Motor Skills, 76*(2), 536–538.

Furnham, A. (1990). Language and Personality. In H. Giles & W. P. Robinson (Hrsg.), *Handbook of Language and Social Psychology* (S. 73–95). Oxford: John Wiley & Sons.

Garb, H. N. (1989). Clinical Judgment, Clinical Training, and Professional Experience. *Psychological Bulletin, 105*(3), 387–396.

Gardner, J., Harlen, W., Hayward, L. & Stobart, G. (2008). *Changing Assessment Practice – Process, Principles and Standards (Assessment Reform Group)*. Verfügbar unter: https://www.aaia.org.uk/content/uploads/2010/06/ARIA-Changing-Assessment-Practice-Pamphlet-Final.pdf [25.11.2016].

Gasteiger-Klicpera, B., Klicpera, C. & Schabmann, A. (2001). Wahrnehmung der Schwierigkeiten lese- und rechtschreibschwacher Kinder durch die Eltern: Pygmalion im Wohnzimmer? *Praxis der Kinderpsychologie und Kinderpsychiatrie, 50*(8), 622–639.

Geist, B. (2014). *Sprachdiagnostische Kompetenz von Sprachförderkräften*. Berlin: De Gruyter.

Geist, B. & Voet Cornelli, B. (2015). Sprachdiagnostik mehrsprachiger Kinder im Elementarbereich. In M. Urban, M. Schulz, K. Meser & S. Thoms (Hrsg.), *Inklusion und Übergang. Perspektiven der Vernetzung von Kindertageseinrichtungen und Grundschulen* (S. 248–270). Bad Heilbrunn: Klinkhardt.

Genesee, F., Paradis, J. & Crago, M. B. (2004). *Dual Language Development & Disorders: A Handbook on Bilingualism & Second Language Learning* (11. Aufl.). Baltimore, MD: Paul H. Brookes Publishing.

Gibbons, P. (2002). *Scaffolding Language, Scaffolding Learning: Teaching Second Language Learners in the Mainstream Classroom*. Portsmouth, NH: Heinemann.

Givvin, K. B., Stipek, D. J., Salmon, J. M. & MacGyvers, V. L. (2001). In the Eyes of the Beholder: Students' and Teachers' Judgments of Students' Motivation. *Teaching and Teacher Education, 17*(3), 321–331.

Gläser-Zikuda, M. & Hascher, T. (2007). *Lernprozesse dokumentieren, reflektieren und beurteilen: Lerntagebuch und Portfolio in Bildungsforschung und Bildungspraxis*. Bad Heilbrunn: Klinkhardt.

Gogolin, I. (2009). Zweisprachigkeit und die Entwicklung bildungssprachlicher Fahigkeiten. In I. Gogolin & U. Neumann (Hrsg.), *Streitfall Zweisprachigkeit. 1. Aufl.* (S. 263–280). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gogolin, I. & Lange, I. (2011). Bildungssprache und Durchgängige Sprachbildung. In S. Fürstenau & M. Gomolla (Hrsg.), *Migration und schulischer Wandel: Mehrsprachigkeit* (S. 107–128). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gogolin, I., Lange, I., Hawighorst, B., Bainski, C., Heintze, A., Rutten, S. & Saalmann, W. (2011). *Durchgängige Sprachbildung. Qualitätsmerkmale für den Unterricht*. Münster: Waxmann.

Gölitz, D., Roick, T. & Hasselhorn, M. (2006). *Deutscher Mathematiktest für vierte Klassen: DEMAT 4*. Göttingen: Hogrefe.

Graney, S. B. (2008). General Education Teacher Judgments of Their Low-Performing Students' Short-Term Reading Progress. *Psychology in the Schools, 45*(6), 537–549.

Gräsel, C., Krolak-Schwerdt, S., Nölle, I. & Hörstermann, T. (2010). Diagnostische Kompetenz von Grundschullehrkräften bei der Erstellung der Übergangsempfehlung. Eine Analyse aus der Perspektive der sozialen Urteilsbildung. Projekt Diagnostische Kompetenz. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes.* (S. 286–295). Weinheim; Basel: Beltz.

Gresham, F. M., MacMillan, D. L. & Bocian, K. M. (1997). Teachers as" Tests": Differential validity of teacher judgments in identifying students at-risk for learning difficulties. *School Psychology Review, 26*, 47–60.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). Clinical versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment, 12*(1), 19–30.

Gulliksen, H. (1986). Perspective on Educational Measurement. *Applied Psychological Measurement, 10*(2), 109–132.

Hachfeld, A., Anders, Y., Schroeder, S., Stanat, P. & Kunter, M. (2010). Does Immigration Background Matter? How Teachers' Predictions of Students' Performance Relate to Student Background. *International journal of educational research, 49*(2), 78–91.

Hadley, S. T. (1954). A School Mark-Fact or Fancy. *Educational administration and supervision, 40*, 305–312.

Hadley, S. T. (1995). Feststellungen und Vorurteile in der Zensierung. In K. Ingenkamp (Hrsg.), *Die Fragwürdigkeit der Zensurengebung* (S. 159–166). Weinheim: Beltz.

Halbheer, U. & Reusser, K. (2008). Outputsteuerung, Accountability, Educational Governance – Einführung in die Geschichte, Begrifflichkeiten und Funktionen von Bildungsstandards. *Beiträge zur Lehrerbildung, 26*(3), 253–266.

Hamilton, C. & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32*(2), 228–240.

Hammel, L. (2011). *Selbstkonzepte fachfremd unterrichtender Musiklehrerinnen und Musiklehrer an Grundschulen. Eine Grounded-Theory-Studie*. Berlin: LIT-Verlag.

Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau, 63*(1), 43–49.

Hascher, T. (2008). Diagnostische Kompetenzen im Lehrberuf. In C. Kraler & M. Schratz (Hrsg.), *Wissen erwerben, Kompetenzen entwickeln. Modelle zur kompetenzorientierten Lehrerbildung.* (S. 71–86). Münster: Waxmann.

Hecht, S. A. & Greenfield, D. B. (2002). Explaining the Predictive Accuracy of Teacher Judgments of Their Students' Reading Achievement: The Role of Gender, Classroom Behavior, and Emergent Literacy Skills in a Longitudinal Sample of Children Exposed to Poverty. *Reading and Writing, 15*(7), 789–809.

Helmke, A. (Hrsg.). (2010). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (3. ed.). Seelze-Velber: 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.

Helmke, A. & Schrader, F.-W. (1987). Interactional Effects of Instructional Quality and Teacher Judgment Accuracy on Achievement. *Teaching and Teacher Education, 3*, 91–98.

Heppt, B. (2016). *Verständnis von Bildungssprache bei Kindern mit deutscher und nicht-deutscher Familiensprache.* Humboldt-Universität zu Berlin, Lebenswissenschaftlichen Fakultät, Berlin.

Heppt, B., Stanat, P., Dragon, N., Berendes, K. & Weinert, S. (2014). Bildungssprachliche Anforderungen und Hörverstehen bei Kindern mit deutscher und nicht-deutscher Familiensprache. *Zeitschrift für Pädagogische Psychologie, 28*(3), 139–149.

Hesse, I. & Latzko, B. (2009). *Diagnostik für Lehrkräfte*. Opladen: Budrich.

Hoffmann, L. & Böhme, K. (2014a). Rezension zu: Neugebauer, Uwe & Becker-Mrotzek, Michael (2013), Die Qualität von Sprachstandserverfahren im

Elementarbereich - Eine Analyse und Bewertung. *Zeitschrift für Interkulturellen Fremdsprachenunterricht – Didaktik und Methodik im Bereich Deutsch als Fremdsprache, 19*(2), 202–205.

Hoffmann, L. & Böhme, K. (2014b). Wie gut können Grundschullehrkräfte die Schwierigkeit von Deutsch-und Mathematikaufgaben beurteilen? Eine Untersuchung zur Genauigkeit aufgabenbezogener Lehrerurteile auf Klassenebene. *Psychologie in Erziehung und Unterricht, 61*(1), 42–55.

Hoffmann, L. & Böhme, K. (2017). Wird sprachlicher Förderbedarf in der Grundschule sicher erkannt? Zur Klassifikationsgüte von diagnostischen Entscheidungen. *Zeitschrift für Pädagogische Psychologie, 31*(2), 137–147.

Hoffmann, L., Böhme, K. & Stanat, P. (2017). Mit welchen diagnostischen Verfahren wird in Grundschulen Sprachförderbedarf festgestellt? Eine bundesweite Bestandsaufnahme. *Frühe Bildung, 6*(3), 116–123.

Hoge, R. D. (1983). Psychometric Properties of Teacher-Judgment Measures of Pupil Aptitudes, Classroom Behaviors, and Achievement Levels. *Journal of Special Education, 17*, 401–429.

Hoge, R. D. & Butcher, R. (1984). Analysis of Teacher Judgments of Pupil Achievement Levels. *Journal of Educational Psychology, 76*(5), 777–781.

Hoge, R. D. & Coladarci, T. (1989). Teacher-Based Judgments of Academic Achievement: A Review of Literature. *Review of Educational Research, 59*(3), 297–313.

Holler, D. (2007). Bedeutung sprachlicher Fähigkeiten für Bildungserfolge. In K. Jampert, P. Best, A. Guadatiello, D. Holler & A. Zehnbauer (Hrsg.), *Schlüsselkompetenz Sprache: Sprachliche Bildung und Förderung im Kindergarten. Konzepte, Projekte und Maßnahmen* (S. 24–28). Weimar: Verlag das Netz.

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*(3), 177–182.

Hopp, H., Thoma, D. & Tracy, R. (2010). Sprachförderkompetenz pädagogischer Fachkräfte. Ein sprachwissenschaftliches Modell. *Zeitschrift für Erziehungswissenschaft, 13*(4), 609–629.

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

Hurwitz, J. T., Elliott, S. N. & Braden, J. P. (2007). The Influence of Test Familiarity and Student Disability Status upon Teachers' Judgments of Students' Test Performance. *School Psychology Quarterly, 22*(2), 115–144.

IfBQ (2016) = Institut für Bildungsmonitoring und Qualitätsentwicklung. (2016). *KERMIT – Kompetenzen ermitteln. Hinweise und Anregungen zur Nutzung von KERMIT für die Unterrichts- und Schulentwicklung*. Verfügbar unter: http://www.hamburg.de/contentblob/4027104/20009765453b638863be82edb1624075/data/pdf-hinweise-zur-nutzung-von-kermit-fuer-die-schul-und-unterrichtsentwicklung.pdf [23.11.2016].

Ingenkamp, K. (1989). *Diagnostik in der Schule*. Weinheim: Beltz.

Ingenkamp, K. (Hrsg.). (1995). *Die Fragwürdigkeit der Zensurengebung*. Weinheim: Beltz.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6., neu ausgestattete Aufl.). Weinheim: Beltz.

IQB (2014) = Institut zur Qualitätsentwicklung im Bildungswesen. (2014). *Tätigkeitsbericht 2007–2013*. Verfügbar unter: https://www.iqb.hu-berlin.de/data/n/n063/IQBTtigkeitsberi.pdf [30.01.2017].

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

Jäger, R. S. (2009). Diagnostische Kompetenz und Urteilsbildung als Element von Lehrprofessionalität. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 105–116). Weinheim: Beltz.

Jäger, R. S. & Petermann, F. (1999). *Psychologische Diagnostik* (4. Aufl.). Weinheim: PVU.

Jeuk, S. (2009). Aktuelle Verfahren zur Einschätzung des Stands der Sprachaneignung bei mehrsprachigen Kindern im Grundschulalter. In S. Jeuk & I. Schmid- Barkow (Hrsg.), *Differenzen diagnostizieren und Kompetenzen fördern im Deutschunterricht* (S. 61–82). Freiburg: Fillibach.

Junk-Deppenmeier, A. (2009). Sprachstandserhebungen bei Schülerinnen und Schülern mit Deutsch als Zweitsprache in der Sekundarstufe. In S. Jeuk & I. Schmid-Barkow (Hrsg.), *Differenzen diagnostizieren und Kompetenzen fördern im Deutschunterricht* (S. 83–91). Freiburg: Fillibach.

Jussim, L. & Eccles, J. S. (1992). Teacher Expectations: II. Construction and Reflection of Student Achievement. *Journal of Personality and Social Psychology, 63*(6), 947–961.

Jussim, L. & Harber, K. D. (2005). Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies. *Personality and Social Psychology Review, 9*(2), 131–155.

Kaiser, J., Möller, J., Helm, F. & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift für Erziehungswissenschaft, 18*(2), 279–302.

Kaiser, J., Retelsdorf, J., Südkamp, A. & Möller, J. (2013). Achievement and Engagement: How Student Characteristics Influence Teacher Judgments. *Learning and Instruction, 28*, 73–84.

Kannengieser, S. (2012). *Sprachentwicklungsstörungen. Grundlagen, Diagnostik und Therapie*. München: Elsevier, Urban & Fischer.

Kany, W. & Schöler, H. (2014). Theorien zum Spracherwerb. In L. Ahnert (Hrsg.), *Theorien in der Entwicklungspsychologie* (S. 468–485). Berlin: Springer.

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für pädagogische Psychologie, 23*(3–4), 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*(3), 159–172.

Kelley, H. H. (1967). Attribution in Social Psychology. *Nebraska Symposium on Motivation, 15*, 192–238.

Kenny, D. T. & Chekaluk, E. (1993). Early Reading Performance: A Comparison of Teacher-Based and Test-Based Assessments. *Journal of Learning Disabilities, 26*(4), 227–236.

Kishor, N. (1994). Teachers' Judgements of Students' Performance: Use of Consensus, Consistency and Distinctiveness Information. *Educational Psychology, 14*(2), 233–247.

Klauer, K. J. (1985). Framework for a Theory of Teaching. *Teaching and Teacher Education, 1*(1), 5–17.

Klieme, E., Pauli, C. & Reusser, K. (2009). The Pythagoras Study: Investigating Effects of Teaching and Learning in Swiss and German Mathematics Classrooms. In T. Janík & T. Seidel (Hrsg.), *The Power of Video Studies in Investigating Teaching and Learning in the Classroom* (S. 137–160). Münster: Waxmann.

Klinger, D., McDivitt, P. J., Howard, B. B., Munoz, M. A., Roger, W. T. & Wylie, E. C. (2015). Classroom Assessment Standards for PreK-12 Teachers: Joint Committee on Standards for Educational Evaluation. Amazon: Kindle E-Book.

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., Keller, S. & Schmitz, B. (2012). Hängen Diagnostische Kompetenz und Beratungskompetenz von Lehrkräften zusammen? *Psychologische Rundschau, 63*(1), 3–10.

Klug, J., Gerich, M. & Schmitz, B. (2012). Ein Tagebuch für Hauptschullehrkräfte zur Unterstützung der Reflektionsprozesse beim Diagnostizieren. *Empirische Pädagogik, 26*(2), 292–311.

KMK (2002) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2002). *PISA 2000 – Zentrale Handlungsfelder. Zusammenfassende Darstellung der laufenden und geplanten Maßnahmen in den Ländern. Beschluss der 299. Kultusministerkonferenz vom 17./10.2002*. Berlin: KMK.

KMK (2004) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004*. Berlin: KMK.

KMK (2005) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005a). *Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004*. München: Luchterhand.

KMK (2005) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005b). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004*. München: Luchterhand.

KMK (2012) = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2012). *Vereinbarung zur Weiterentwicklung von VERA (Beschluss der Kultusministerkonferenz vom 08.03.2012)*. Verfügbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_03_08_Weiterentwicklung-VERA.pdf [23.11.2016].

Knapp, W. (1999). Verdeckte Sprachschwierigkeiten. *Die Grundschule, 5*(99), 30–33.

Koch, K. (2012). *Zweitspracherwerb am Übergang vom Elementar- in den Primarbereich*. München: Utz.

Koch, P. & Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz – Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch, 36*, 15–43.

Köster, J. (2008). Lern- und Leistungsaufgaben im Deutschunterricht. *Deutschunterricht, 61*(5), 4–10.

Kracht, A. (2003). Sprachliche Normen und Zielsetzungen von Sprachstandserhebungen – einige kritische Anmerkungen. In Beauftragte der Bundesregierung für Migration. Flüchtlinge und Integration (Hrsg.), *Förderung von Migranten und*

Migrantinnen im Elementar- und Primarbereich. Fachtagung am 7. März 2003 in Berlin. Dokumentation (S. 37–44). Berlin.

Krampen, G. (1984). Welche Funktionen haben Zensuren in der Schule? *Zeitschrift für erziehungs- und sozialwissenschaftliche Forschung, 18*(2), 89–102.

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*(3-4), 175–186.

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*(3), 111–122.

Krolak-Schwerdt, S. & Rummer, R. (2005). Der Einfluss von Expertise auf den Prozess der schulischen Leistungsbeurteilung. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 37*(4), 205–213.

Kronig, W. (2015). Geplante Fehler. In M. Gartmeier, H. Gruber, T. Hascher & H. Heid (Hrsg.), *Fehler: Ihre Funktionen im Kontext individueller und gesellschaftlicher Entwicklung* (S. 203–210). Münster: Waxmann.

Kuhl, P. & Hannover, B. (2012). Differenzielle Benotungen von Mädchen und Jungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44*(3), 153–162.

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

Kunz, G. C. & Schott, F. (1987). *Intelligente tutorielle Systeme. Neue Ansaetze der computerunterstuetzten Steuerung von Lehr-Lern- Prozessen*. Goettingen: Hogrefe.

Langfeldt, H.-P. & Trolldenier, H.-P. (1993a). Die Bedeutung pädagogisch-psychologischer Diagnostik in der Schule: Eine Einführung in den Themenkreis. In *Pädagogisch-psychologische Diagnostik: Aktuelle Entwicklungen und Ergebnisse* (S. 11–26). Heidelberg: Asanger.

Langfeldt, H.-P. & Trolldenier, H.-P. (1993b). *Pädagogisch-psychologische Diagnostik: Aktuelle Entwicklungen und Ergebnisse*. Heidelberg: Asanger.

Law, J., Boyle, J., Harris, F., Harkness, A. & Nye, C. (2000). Prevalence and Natural History of Primary Speech and Language Delay: Findings from a Systematic Review of the Literature. *International Journal of Language and Communication Disorders, 35*, 165–188.

Lehmann, R. H. & Hoffmann, E. (Hrsg.). (2009). *BELLA. Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf "Lernen"*. Münster: Waxmann.

Lehmann, R. H., Peek, R., Gänsfuß, R., Lutkat, S., Mücke, S. & Barth, I. (2000). Qualitätsuntersuchungen an Schulen zum Unterricht in Mathematik. In J. u. S. Ministerium für Bildung (Hrsg.), *Schulforschung in Brandenburg, Heft 1*. Teltow: Druckerei Grabow.

Leinhardt, G. (1983). Novice and Expert Knowledge of Individual Student's Achievement. *Educational Psychologist, 18*(3), 165–179.

Lengyel, D. (2012). *Sprachstandsfeststellung bei mehrsprachigen Kindern im Elementarbereich. Eine Expertise der Weiterbildungsinitiative Frühpädagogischer Fachkräfte (WiFF)*. München: DJI.

Leucht, M., Harsch, C., Pant, H. A. & Köller, O. (2012). Steuerung zukünftiger Aufgabenentwicklung durch Vorhersage der Schwierigkeiten eines Tests für die erste Fremdsprache Englisch durchDutch Grid Merkmale. *Diagnostica, 58*(1), 31–44.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: PVU.

Lintorf, K. (2012). *Wie vorhersagbar sind Grundschulnoten?: Prädiktionskraft individueller und kontextspezifischer Merkmale*. Wiesbaden: Springer.

Lintorf, K., McElvany, N., Rjosk, C., Schroeder, S., Baumert, J., Schnotz, W., Horz, H. & Ullrich, M. (2011). Zuverlässigkeit von diagnostischen Lehrerurteilen - Reliabilität verschiedener Urteilsmaße bei der Einschätzung von Aufgabenschwierigkeiten. *Unterrichtswissenschaft, 39*(2), 102–120.

Lisker, A. (2013). *Sprachstandserhebung und Sprachförderung vor der Einschulung. Eine Bestandsaufnahme in den Bundesländern. Expertise im Auftrag des Deutschen Jugendinstituts. Aktualisierung der Expertise von 2010*. München: DJI.

Loeber, R. & Dishion, T. (1983). Early Predictors of Male Delinquency: A Review. *Psychological Bulletin, 94*(1), 68–99.

Lorenz, C. (2011). *Diagnostische Kompetenz von Grundschullehrkräften. Strukturelle Aspekte und Bedingungen*. Bamberg: Univ. 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*(3–4), 211–222.

LSA (2015) – Landesschulamt und Lehrkräfteakademie. (2015). *Zentrale Lernstandserhebungen in Hessen Hinweise zur Zielsetzung und zur Durchführung – mit erläuternden Arbeitsmaterialien*. Verfügbar unter: www.kreiselternbeirat-lm.de/app/.../16_Broschüre_Zentrale_Lernstandserhebungen.pdf [23.11.2016].

Lüdtke, U. M. & Kallmeyer, K. (2007). Kritische Analyse ausgewählter Sprachstandserhebungsverfahren für Kinder vor Schuleintritt aus Sicht der Linguistik, Diagnostik und Mehrsprachigkeitsforschung. *Die Sprachheilarbeit*(6), 261–278.

Lukesch, H. (1998). *Einführung in die pädagogisch-psychologische Diagnostik* (2., vollst. neu bearb. Aufl). Regensburg: Roderer.

Madelaine, A. & Wheldall, K. (2005). Identifying Low-Progress Readers: Comparing Teacher Judgment with a Curriculum-Based Measurement Procedure. *International Journal of Disability, Development and Education, 52*(1), 33–42.

Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A. & Palumbo, P. (1998). The Accuracy and Power of Sex, Social Class, and Ethnic Stereotypes: A Naturalistic Study in Person Perception. *Personality and Social Psychology Bulletin, 24*(12), 1304–1318.

Mager, R. F. (1978). *Lernziele und Unterricht*. Weinheim.

Maihack, V. (2014). Gut gedacht – schlecht gemacht? Mercator-Institut bewertet Sprachstandsverfahren – und erntet Kritik an der Kritik. *Logos, 22*(1), 66–67.

Maihack, V., Grimm, H., Schöler, H., Becker-Mrotzek, M. & Neugebauer, U. (2014). Sprachstandsverfahren. Der Weg ist nicht das Ziel! Diskussion und Replik (Interviewpartner: Hannelore Grimm und Hermann Schöler). *Logos, 22*(2), 112–118.

Martin, S. D. & Shapiro, E. S. (2011). Examining the Accuracy of Teachers' Judgments of DIBELS Performance. *Psychology in the Schools, 48*(4), 343–356.

Marx, H. (1992). Methodische und inhaltliche Argumente für und wider eine frühe Identifikation und Prädiktion von Lese-Rechtschreibschwierigkeiten. *Diagnostica, 38*(3), 249–268.

May, P. (2002). *HSP 1+ : Hamburger Schreib-Probe für die Klassenstufen 1/2 (Mitte Klasse 1, Ende Klasse 1, Mitte Klasse 2) ; Hinweise zur Durchführung und Auswertung* (Neustandardisierung 2001). Hamburg: vpm Verl. für Pädag. Medien.

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., Horz, H. & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedienmit instruktionalen Bildern. *Zeitschrift für Pädagogische Psychologie, 23*(3–4), 223–235.

McNamara, T. (2005). Beurteilungsverfahren für die sprachliche Entwicklung von Kindern zwischen dem vierten und vierzehnten Lebensjahr. In K. Ehlich (Hrsg.), *Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Förderung von Kindern mit und ohne Migrationshintergrund* (S. 171–192). Berlin: BMBF.

Meehl, P. E. (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, MN: University of Minnesota Press.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y. & Atkins-Burnett, S. (2001). Trusting Teachers' Judgments: A Validity Study of a Curriculum-Embedded Performance Assessment in Kindergarten to Grade 3. *American Educational Research Journal, 38*(1), 73–95.

Metze, W. (2005). *Stolperwörter-Lesetest*: Verfügbar unter: http://wilfriedmetze.de/Handanweisung_2009.pdf [14.05.2014].

Meyer, H. (2004). *Was ist guter Unterricht?* Berlin: Cornelsen.

MK NI (2003) = Niedersächsisches Kultusministerium. (2003). *Fit in Deutsch. Feststellung des Sprachstandes 10 Monate vor Einschulung*. Hannover: MK NI.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Berlin, Heidelberg: Springer Berlin Heidelberg.

Murphy, D., Campbell, C. & Garavan, T. N. (1999). The Pygmalion Effect Reconsidered: Its Implications for Education, Training and Workplace Learning. *Journal of European Industrial Training, 23*(4–5), 238–251.

Nachtigal, C. (Hrsg.). (2016). *Landesbericht Thüringer Kompetenztests 2016*. Verfügbar unter: https://www.kompetenztest.de/download/kt2016-landesbericht.pdf [23.11.2016].

Neber, H. & Heller, K. A. (2004). Einführung in den Themenschwerpunkt "Hochbegabtenförderung auf dem Prüfstand". *Psychologie in Erziehung und Unterricht, 51*(1), 1–7.

Neugebauer, U. & Becker-Mrotzek, M. (2013). *Die Qualität von Sprachstandsverfahren im Elementarbereich. Eine Analyse und Bewertung*. Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache.

Neugebauer, U., Becker-Mrotzek, M. & Stanat, P. (2014). Ermittlung von Sprachförderbedarf bei Kindern im Elementarbereich aus pädagogisch-psychologischer Sicht. *Recht der Jugend und des Bildungswesens, 62*(1), 100–110.

Neumann, K. & Euler, H. A. (2013). Kann ein Sprachstandsscreening zwischen dem Bedarf für Sprachförderung und für Sprachtherapie trennen? In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven. 1.* (S. 174–198). Münster: Waxmann.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Oelkers, J. & Reusser, K. (Hrsg.). (2008). *Expertise: Qualität entwickeln, Standards sichern, mit Differenz umgehen*. Bonn: BMBF.

Oerke, B., McElvany, N., Ohle, A., Ullrich, M. & Horz, H. (2015). Verbessert sich die diagnostische Urteilsgenauigkeit von Lehrkräften bei längerem Kontakt mit der Klasse? *Psychologie in Erziehung und Unterricht, 63*(1), 34–47.

Osnes, J. (1995). Der Einfluß äußerer Faktoren bei der Aufsatzbeurteilung. In K. Ingenkamp (Hrsg.), *Die Fragwürdigkeit der Zensurengebung* (S. 131–147). Weinheim: Belz.

Owens, R. E. (2015). *Language Development: An Introduction*. Boston: Pearson.

Paetsch, J., Wolf, K. M., Stanat, P. & Darsow, A. (2014). Sprachförderung von Kindern und Jugendlichen aus Zuwandererfamilien. *Zeitschrift für Erziehungswissenschaft, Sonderband Herkunft und Bildungserfolg von der Vorschule bis zur Universität: Forschungsstand und Interventionsmöglichkeiten aus interdisziplinärer Perspektive, 17*(2), 315–347.

Paradis, J. (2005). Grammatical Morphology in Children Learning English as a Second LanguageImplications of Similarities With Specific Language Impairment. *Language, Speech, and Hearing Services in Schools, 36*(3), 172–187.

Pekrun, R., Hall, N. C., Goetz, T. & Perry, R. P. (2014). Boredom and Academic Achievement: Testing a Model of Reciprocal Causation. *Journal of Educational Psychology, 106*(3), 696–710.

Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher, 38*(2), 109–119.

Popham, W. J. (2011). Assessment Literacy Overlooked: A Teacher Educator's Confession. *The Teacher Educator, 46*(4), 265–273.

Praetorius, A.-K., Greb, K., Lipowsky, F. & Gollwitzer, M. (2010). Lehrkräfte als Diagnostiker. Welche Rolle spielt die Schülerleistung bei der Einschätzung von mathematischen Selbstkonzepten? *Journal for Educational Research Online, 2*(1), 121–144.

Praetorius, A.-K., Lipowsky, F. & Karst, K. (2012). Diagnostische Kompetenz von Lehrkräften. Aktueller Forschungsstand, unterrichtspraktische Umsetzbarkeit und Bedeutung für den Unterricht. In R. Lazarides & A. Ittel (Hrsg.), *Differenzierung im mathematisch-naturwissenschaftlichen Unterricht. Implikationen für Theorie und Praxis.* (S. 115–146). Bad Heilbrunn: Klinkhardt.

QUA-LIS NRW (2016) = Qualitäts- und Landesinstitut für Schule UnterstützungsAgentur. (2016). *Lernstandserhebungen in Klasse 8. Allgemeine Informationen und Ergebnisse des Durchgangs 2016 in Nordrhein-Westfalen*. Verfügbar unter: http://www.schulentwicklung.nrw.de/lernstand8/upload/download/mat_2016/Bericht_Lernstand8-2016_.pdf [23.11.2016].

Racherbäumer, K. & Kühn, M. S. (2013). Standardized examinations and individualized learning. *Zeitschrift für Bildungsforschung, 3*(1), 27–45.

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*(2), 335–360.

Redder, A. (2013). Sprachliches Kompetenzgitter. Linguistisches Konzept und evidenzbasierte Ausführung. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven.* (S. 108–134). Münster: Waxmann.

Redder, A., Schwippert, K., Hasselhorn, M., Forschner, S., Fickermann, D. & Ehrlich, K. (2011). *Bilanz und Konzeptualisierung von strukturierter Forschung zu "Sprachdiagnostik und Sprachförderung". ZUSE-Berichte Band 2*. Hamburg: Hamburger Zentrum zur Unterstützung der wissenschaftlichen Begleitung und Erforschung schulischer Entwicklungsprozesse (ZUSE).

Redder, A. & Weinert, S. (2013). Sprachliche Handlungsfähigkeiten im Fokus von FiSS. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven.* (S. 7–16). Münster: Waxmann.

Reich, H. H. (2003). Tests und Sprachstandsmessungen bei Schülern und Schülerinnen, die Deutsch nicht als Muttersprache haben. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache* (Bd. 2, S. 914–923). Paderborn: Schöningh.

Reich, H. H. & Roth, H.-J. (2003). *Hamburger Verfahren zur Analyse des Sprachstands Fünfjähriger – HAVAS 5*. Hamburg: Landesinstitut für Lehrerbildung und Schulentwicklung.

Reusser, K. (2009). Unterricht. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 881–896). Weinheim: Beltz.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59–71). Weinheim: Beltz.

Richter, D. (ohne Datum). *Entwicklung und Validierung eines Instruments zur Erfassung von Assessment Literacy bei Mathematiklehrkräften. Projektbeschreibung*. Verfügbar unter: http://gepris.dfg.de/gepris/projekt/273342462 [23.11.2016].

Rjosk, C., McElvany, N., Anders, Y. & Becker, M. (2011). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung der basalen Lesefähigkeit ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht, 58*(2), 92–105.

Rodriguez, M. C. (2004). The Role of Classroom Assessment in Student Performance on TIMSS. *Applied Measurement in Education, 17*(1), 1–24.

Rogalla, M. & Vogt, F. (2008). Förderung adaptiver Lehrkompetenz: eine Interventionsstudie. *Unterrichtswissenschaft, 36*(1), 17–36.

Rosenthal, R. & Rubin, D. B. (1978). Issues in summarizing the first 345 studies of interpersonal expectancy effects. *Behavioral and Brain Sciences, 1*(3), 377–415.

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

Rothweiler, M. (2007). Bilingualer Spracherwerb und Zweitspracherwerb. In M. Steinbach, R. Albert, H. Girnth, A. Hohenberger, B. Kümmerling-Meibauer, J. Meibauer, M. Rothweiler & M. Schwarz-Friesel (Hrsg.), *Schnittstellen der germanistischen Linguistik* (S. 103–135). Stuttgart: Metzler.

Rothweiler, M. (2013). Spezifische Sprachentwicklungsstörungen bei mehrsprachigen Kindern. *Sprache· Stimme· Gehör, 37*(4), 186–190.

Rothweiler, M. & Ruberg, T. (2011). *Der Erwerb des Deutschen bei Kindern mit nichtdeutscher Erstsprache. Sprachliche und außersprachliche Einflussfaktoren. Eine Expertise der Weiterbildungsinitiative Frühpädagogische Fachkräfte (WiFF). Stand: April 2011*. München: DJI.

Rupp, S. (2013). Semantisch-lexikalische Entwicklungsstörungen. In S. Rupp (Hrsg.), *Semantisch-lexikalische Störungen bei Kindern* (S. 73–106). Berlin: Springer.

Sacher, W. (1994). *Prüfen – Beurteilen – Benoten. Theoretische Grundlagen und praktische Hilfestellungen für den Primar- und Sekundarbereich*. Bad Heilbrunn: Klinkhardt.

Sachse, S., Anke, B. & von Suchodoletz, W. (2007). Früherkennung von Sprachentwicklungsstörungen - ein Methodenvergleich. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 35*(5), 323–331.

Sächsisches Bildungsinstitut. (2012). *Kompetenztests an sächsischen Schulen*. Verfügbar unter: http://www.schule.sachsen.de/download/download_sbi/ktest_broschuere.pdf [23.11.2016].

Salvesen, K. Å. & Undheim, J. O. (1994). Screening for Learning Disabilities with Teacher Rating Scales. *Journal of Learning Disabilities, 27*(1), 60–66.

Schafer, W. D. & Lissitz, R. W. (1987). Measurement Training for School Personnel Recommendations and Reality. *Journal of Teacher Education, 38*(3), 57–63.

Schleppegrell, M. J. (2001). Linguistic Features of the Language of Schooling. *Linguistics and Education, 12*(4), 431–459.

Schleppegrell, M. J. (2004). *The Language of Schooling: A Functional Linguistics Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schleppegrell, M. J. (2012). Academic Language in Teaching and Learning: Introduction to the Special Issue. *The Elementary School Journal, 112*(3), 409–418.

Schmidt-Atzert, L. & Amelang, M. (2012). Zuordnungs- und Klassifikationsstrategien. In L. Schmidt-Atzert & M. Amelang (Hrsg.), *Psychologische Diagnostik* (5. Aufl., S. 409–428). Berlin: Springer.

Schott, F., Neeb, K.-E. & Wieberg, H.-J. W. (1981). *Lehrstoffanalyse und Unterrichtsplanung: eine praktische Anleitungzur Analyse von Lehrstoffen, Präzisierung von Lehrzielen, Konstruktion von Lehrmaterialien, Überprüfung des Lehrerfolges*. Braunschweig: Westermann.

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

Schrader, F.-W. (2006). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 95–100). Weinheim: Beltz/PVU.

Schrader, F.-W. (2009). Anmerkungen zum Themenschwerpunkt Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*(3–4), 237–245.

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerinnen- und Lehrerbildung, 31*(2), 154–165.

Schrader, F.-W. & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern. Komponenten u. Wirkungen. *Empirische Pädagogik, 1*(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*(4), 312–324.

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

Schreiner, C., Breit, S. & Haider, G. (2008). Zur Validität der Mathematiknoten. Ein Vergleich von Lehrerbeurteilung und Leistungsmessung bei PISA. In F. Hofmann, C. Schreiner & J. Thonhauser (Hrsg.), *Qualitative und quantitative Aspekte. Zu ihrer Komplementarität in der erziehungswissenschaftlichen Forschung* (S. 211–223). Münster: Waxmann.

Schröder, H. (2000). *Lernen – Lehren – Unterricht. Lernpsychologische und didaktische Grundlagen* (1. Aufl.). München: Oldenbourg.

Schulz, P. (2013). Sprachdiagnostik bei mehrsprachigen Kindern. *Sprache· Stimme· Gehör, 37*(4), 191–195.

Schulz, P. & Tracy, R. (2011). *Linguistische Sprachstandserhebung-Deutsch als Zweitsprache: LiSe-DaZ (Manual)*. Göttingen: Hogrefe.

Schulz, P., Tracy, R. & Wenzel, R. (2008). Linguistische Sprachstandserhebung–Deutsch als Zweitsprache (LiSe-DaZ): Theoretische Grundlagen und erste Ergebnisse. In

B. Ahrenholz (Hrsg.), *Zweitspracherwerb. Diagnosen Verläufe Voraussetzungen. Beiträge aus dem 2. Workshop "Kinder mit Migrationshintergrund"* (S. 17–41). Freiburg: Fillibach.

Schuth, E., Heppt, B., Köhne, J., Weinert, S. & Stanat, P. (2015). Die Erfassung schulisch relevanter Sprachkompetenzen bei Grundschulkindern. Entwicklung eines Testinstruments. In A. Redder, J. Naumann & R. Tracy (Hrsg.), *Forschungsinitiative Sprachdiagnostik und Sprachförderung – Ergebnisse* (S. 93–112). Münster: Waxmann.

Seeber, S. (2009). Urteilsgenauigkeit von Lehrerinnen und Lehrern in der sonderpädagogischen Förderung. In R. Lehmann & E. Hoffmann (Hrsg.), *BELLA. Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf "Lernen".* (S. 197–208). Münster u.a.: Waxmann.

Settinieri, J. (2010). Zur Messgüte sprachstandsdiagnostischer Verfahren im Vorschulalter. Erste Ergebnisse der Korrelationsstudie Sismik/Seldak – Delfin 4. In B. Ahrenholz & W. Knapp (Hrsg.), *Sprachstand erheben–Spracherwerb erforschen. Beitrage aus dem 6. Workshop "Kinder mit Migrationshintergrund"* (S. 35–51). Stuttgart: Klett.

Shinn, M. R., Tindal, G. A. & Spira, D. A. (1987). Special Education Referrals as an Index of Teacher Tolerance: Are Teachers Imperfect Tests? *Exceptional Children, 54*(1), 32–40.

Snow, C. E. (2010). Academic Language and the Challenge of Reading for Learning About Science. *Science, 328*(5977), 450–452.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für pädagogische Psychologie, 19*(1–2), 85–95.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

Stanat, P., Weirich, S. & Radmann, S. (2012). Sprach- und Leseförderung. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 251–276). Münster: Waxmann.

Stiggins, R. J. (1991). Assessment Literacy. *Phi Delta Kappan, 72*(7), 534–539.

Stubbe, T. C. & Bos, W. (2008). Schullaufbahnempfehlungen von Lehrkräften und Schullaufbahnentscheidungen von Eltern am Ende der vierten Jahrgangsstufe. *Empirische Pädagogik, 22*(1), 49–63.

Sturzbecher, D., Mörl, S. & Rüdel, M. (2013). Die Praktische Fahrerlaubnisprüfung als systematische Fahrverhaltensbeobachtung - bewährte Fundamente und neue

Ergebnisse der Verfahrensentwicklung. *Zeitschrift für Verkehrssicherheit, 2*, 76–83.

Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology, 104*(3), 743–762.

Südkamp, A. & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum. Direkte und indirekte Einschätzungen von Schülerleistungen. *Zeitschrift für pädagogische Psychologie, 23*(3-4), 161–174.

Südkamp, A., Möller, J. & Pohlmann, B. (2008). Der Simulierte Klassenraum. Ein Instrument zur Untersuchung von diagnostischer Kompetenz. In E.-M. Lankes (Hrsg.), *Pädagogische Professionalität als Gegenstand empirischer Forschung.* (S. 87–97). Münster: Waxmann.

Teisl, J. T., Mazzocco, M. M. M. & Myers, G. F. (2001). The Utility of Kindergarten Teacher Ratings for Predicting Low Academic Achievement in First Grade. *Journal of Learning Disabilities, 34*(3), 286-293.

ter Laak, J. J. F., De Goede, M. P. M. & Brugman, G. M. (2001). Teacher's Judgements of Pupils: Agreement and Accuracy. *Social Behavior and Personality, 29*, 257–270.

Thorndike, E. L. (1920). A Constant Error in Psychological Ratings. *Journal of Applied Psychology, 4*(1), 25–29.

Tiedemann, J. (2002). Teachers' Gender Stereotypes as Determinants of Teacher Perceptions in Elementary School Mathematics. *Educational Studies in Mathematics, 50*(1), 49–62.

Trautwein, U. & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind. Referenzgruppeneffekte bei Übertrittsentscheidungen. *Zeitschrift für pädagogische Psychologie, 21*(2), 119–133.

Trautwein, U., Lüdtke, O., Köller, O., Marsh, H. W. & Baumert, J. (2006). Tracking, Grading, and Student Motivation: Using Group Composition and Status to Predict Self-Concept and Interest in Ninth-Grade Mathematics. *The journal of educational psychology, 98*(4), 788–806.

Tröster, H. (Hrsg.). (2009). *Früherkennung im Kindes- und Jugendalter. Strategien bei Entwicklungs-, Lern- und Verhaltensstörungen*. Göttingen: Hogrefe.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124–1131.

Tyler, W. (1971). *Basic Principles of Curriculum and Instruction*. Chicago: The University of Chicago Press.

Uesseler, S., Runge, A. & Redder, A. (2013). "Bildungssprache" diagnostizieren. Entwicklung eines Instruments zur Erfassung von bildungssprachlichen Fähigkeiten bei Viert- und Fünftklässlern. In A. Redder & S. Weinert (Hrsg.),

Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven. (S. 42–67). Münster: Waxmann.

Ulich, M. & Mayr, T. (2003). *Sismik. Sprachverhalten und Interesse an Sprache bei Migrantenkindern in Kindertageseinrichtungen (Beobachtungsbogen und Begleitheft)*. Freiburg: Herder.

Ulich, M. & Mayr, T. (2006). *Seldak. Sprachentwicklung und Literacy bei deutschsprachig aufwachsenden Kindern (Beobachtungsbogen und Begleitheft)*. Freiburg: Herder.

Urhahne, D., Timm, O., Zhu, M. & Tang, M. (2013). Sind unterschätzte Schüler weniger leistungsmotiviert als überschätzte Schüler? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 45*(1), 34–43.

Urhahne, D., Zhou, J., Stobbe, M., Chao, S.-H., Zhu, M. & Shi, J. (2010). Motivationale und affektive Merkmale unterschätzter Schüler. Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 24*(3–4), 275–288.

van Ophuysen, S. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 38*(4), 154–161.

van Ophuysen, S. (2010). Professionelle pädagogisch-diagnostische Kompetenz. Eine theoretische und empirische Annäherung. In N. Berkemeyer, W. Bos, H. G. Holtappels, N. McElvany & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (S. 203–234). Weinheim: Juventa.

Voet Cornelli, B., Geist, B., Grimm, A. & Schulz, P. (2012). Wie wird der Sprachstand mehrsprachiger Kinder in pädiatrischen Vorsorgeuntersuchungen erhoben? Erste Ergebnisse aus dem Projekt *cammino*. In S. Jeuk & J. Schäfer (Hrsg.), *Deutsch als Zweitsprache in Kindertageseinrichtungen und Schulen. Aneignung, Förderung, Unterricht. Beiträge aus dem 7. Workshop Kinder mit Migrationshintergrund* (S. 43–73). Freiburg: Fillibach.

Vogt, F. & Rogalla, M. (2009). Developing Adaptive Teaching Competency through coaching. *Teaching and Teacher Education, 25*(8), 1051–1060.

von Aufschnaiter, C., Cappell, J., Dübbelde, G., Ennemoser, M., Mayer, J., Stiensmeier-Pelster, J., Sträßer, R. & Wolgast, A. (2015). Diagnostische Kompetenz: Theoretische Überlegungen zu einem zentralen Konstrukt der Lehrerbildung. *Zeitschrift für Pädagogik, 61*(5), 738–758.

Wahle, C. V., Back, M. D., Nestler, S., Pretsch, J., Schrader, F.-W. & Praetorius, A.-K. (2017). Welche Hinweisreize nutzen Lehrkräfte für ihr Urteil und welche sollten sie nutzen? Eine Linsenmodellanalyse, *5. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF)*. Heidelberg.

Weinert, F. E. (2000). *Lehren und Lernen für die Zukunft – Ansprüche an das Lernen in der Schule*. Verfügbar unter: http://www2.ibw.uni-heidelberg.de/~gerstner/WeinertLehren&Lernen.pdf [26.11. 2015].

Weinert, F. E., Schrader, F.-W. & Helmke, A. (1990). Educational Expertise: Closing the Gap between Educational Research and Classroom Practice. *School Psychology International, 11*(3), 163–180.

Weinert, S. & Grimm, H. (2008). Sprachentwicklung. In R. Oerter & L. Montada (Hrsg.), *Entwicklungspsychologie* (6 Aufl., S. 502–534). Weinheim: Beltz.

Westphal, A., Gronostaj, A., Vock, M., Emmrich, R. & Harych, P. (2016). Differenzierung im gymnasialen Mathematik- und Deutschunterricht - vor allem bei guten Diagnostiker/innen und in heterogenen Klassen? *Zeitschrift für Pädagogik, 62*, 131–148.

WHO (1992) = World Health Organization. (1992). *International Classification of Diseases and Related Health Problems (10. Revision ICD-10)*. Genua: WHO.

Wild, K.-P. & Rost, D. H. (1995). Klassengröße und Genauigkeit von Schülerbeurteilungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 27*(1), 78–90.

Wildemann, A. (2010). Sprachdiagnostisches Wissen angehender Deutschlehrkräfte – Annäherungen zwischen Utopie und Wirklichkeit. In J. König & B. Hofmann (Hrsg.), *Lehrerprofessionalität. Was sollen Lehrkräfte im Lese- und Schreibunterricht wissen und können?* (S. 178–194). DGLS: Berlin.

Wilhelm, O. & Kunina, O. (2009). Pädagogisch-psychologische Diagnostik. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 307-331): Springer Berlin Heidelberg.

Wittchen, H.-U. (2011). Diagnostische Klassifikation psychischer Störungen. In H.-U. Wittchen & J. Hoyer (Hrsg.), *Klinische Psychologie & Psychotherapie* (S. 27–55). Berlin: Springer.

Wittchen, H.-U. & Hoyer, J. (2011). Diagnostische Prozesse in der Klinischen Psychologie und Psychotherapie. In H.-U. Wittchen & J. Hoyer (Hrsg.), *Klinische Psychologie & Psychotherapie* (S. 383–418). Berlin: Springer.

Wood, D., Bruner, J. S. & Ross, G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100.

Wurster, S. (2016). *Evaluationsgestützte Schul- und Unterrichtsentwicklung. Eine multiperspektivische Analyse von Vergleichsarbeiten, zentralen Abschlussprüfungen, Schulinspektion und interner Evaluation.* Humboldt-Universität zu Berlin, Berlin.

Wurster, S., Richter, D., Schliesing, A. & Pant, H. A. (2013). Nutzung unterschiedlicher Evaluationsdaten an Berliner und Brandenburger Schulen. Rezeption und Nutzung

von Ergebnissen aus Schulinspektion, Vergleichsarbeiten und interner Evaluation im Vergleich. *Die Deutsche Schule, 12*(Beiheft), 19–50.

Xu, Y. & Brown, G. T. L. (2016). Teacher Assessment Literacy in Practice: A Reconceptualization. *Teaching and Teacher Education, 58*, 149–162.

Ysseldyke, J. E., Vanderwood, M. L. & Shriner, J. (1997). Changes Over the Past Decade in Special Education Referral to Placement Probability: An Incredibly Reliable Practice. *Assessment for Effective Intervention, 23*(1), 193–201.

Ziegenspeck, J. W. (1999). *Handbuch Zensur und Zeugnis in der Schule*. Bad Heilbrunn: Klinkhardt.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang A: Ergänzung zu Teilstudie 1

Tabelle: Multinomial-logistische Mehrebenenanalyse für die Lehrerurteile zur Schwierigkeit der Mathematikaufgaben für die Schülerinnen und Schüler ihrer Klassen für das Fach Deutsch (Anhang A: Ergänzungen zu Teilstudie 1)

	Unterschätzung		*Überschätzung*	
	B	*SE (B)*	*B*	*SE (B)*
Ebene 1 (Urteile)				
Höhe des Itemparameters σ_i	1.30**	0.11	–1.45**	0.16
Zeitpunkt Thematisierung im Unterricht	–0.15	0.08	0.15*	0.06
Ebene 2 (Lehrkräfte)				
Kontaktdauer	–0.16	0.12	0.10	0.12
Dauer Lehrertätigkeit (Jahre)[1]	–0.45	0.30	0.04	0.29

Anmerkungen: $^{*}p \leq .05$, $^{**}p < .01$; [1]dichotome Kodierung: 0 (Lehrerfahrung in Deutsch < 10 Jahre), 1 (Lehrerfahrung in Deutsch ≥ 10 Jahre), Anteil von Lehrkräften mit weniger als 10 Jahre Lehrerfahrung: 32.5 Prozent

Tabelle: Multinomial-logistische Mehrebenenanalyse für die Lehrerurteile zur Schwierigkeit der Deutschaufgaben für die Schülerinnen und Schüler ihrer Klassen für das Fach Mathematik (Anhang A: Ergänzungen zu Teilstudie)

	Unterschätzung		*Überschätzung*	
	B	*SE (B)*	*B*	*SE (B)*
Ebene 1 (Urteile)				
Höhe des Itemparameters σ_i	1.37**	0.34	–0.74**	0.07
Zeitpunkt Thematisierung im Unterricht	–0.96**	0.29	0.86**	0.26
Ebene 2 (Lehrkräfte)				
Kontaktdauer	–0.04	0.20	–0.16	0.13
Dauer Lehrertätigkeit (Jahre)[1]	1.12	0.67	–0.42	0.36

Anmerkungen: $^{*}p \leq .05$, $^{**}p < .01$; [1]dichotome Kodierung: 0 (Lehrerfahrung in Mathematik < 10 Jahre), 1 (Lehrerfahrung in Mathematik ≥ 10 Jahre), Anteil von Lehrkräften mit weniger als 10 Jahre Lehrerfahrung: 29.6 Prozent