9 Schriften aus der Fakultät Humanwissenschaften
der Otto-Friedrich-Universität Bamberg

Diagnostische Kompetenz von Grundschullehrkräften

Strukturelle Aspekte und Bedingungen

von Christian Lorenz

UNIVERSITY OF
BAMBERG
PRESS

Schriften aus der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg 9

Schriften aus der Fakultät Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

Band 9

University of Bamberg Press 2011

Diagnostische Kompetenz von Grundschullehrkräften

Strukturelle Aspekte und Bedingungen

von Christian Lorenz

University of Bamberg Press 2011

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.ddb.de/ abrufbar

Diese Arbeit hat der Fakultät Humanwissenschaften der Otto-Friedrich-Universität als Dissertation vorgelegen
1. Gutachter: Prof. Dr. Cordula Artelt
2. Gutachter: Prof. Dr. Gabriele Faust
Tag der mündlichen Prüfung: 16. November 2011

Dieses Werk ist als freie Onlineversion über den Hochschulschriften-Server (OPUS; http://www.opus-bayern.de/uni-bamberg/) der Universitätsbibliothek Bamberg erreichbar. Kopien und Ausdrucke dürfen nur zum privaten und sonstigen eigenen Gebrauch angefertigt werden.

Herstellung und Druck: docupoint, Barleben
Umschlaggestaltung: Dezernat Kommunikation und Alumni der Otto-Friedrich-Universität Bamberg

© University of Bamberg Press Bamberg 2011
http://www.uni-bamberg.de/ubp/

ISSN: 1866-8674
ISBN: 978-3-86309-056-2 (Druckausgabe)
eISBN: 978-3-86309-057-9 (Online-Ausgabe)
URN: urn:nbn:de:bvb:473-opus-3956

Meiner Familie

Danksagung

Für die uneingeschränkte Unterstützung bei dieser Arbeit, für das beständige Antreiben auch in zähen Phasen, für den immer kritischen Blick, für gute Ratschläge und die stets prompte Ansprechbarkeit bei jeglichen Fragen habe ich in erster Linie Frau Prof. Dr. Cordula Artelt zu danken. Mit Verstand und Verständnis hat sie mich in meinem Vorhaben begleitet. Mein Dank gilt ebenso Frau Prof. Dr. Gabriele Faust, die meine Arbeit immer wieder durch guten Zuspruch und aufmunternde Worte vorangetrieben hat, und meinen Kollegen und Hilfskräften am Lehrstuhl. Vor allem aber danke ich meiner Familie, die meine Entscheidung zur Promotion von Anfang an voll unterstützt hat und seitdem in allen Situationen der wichtigste Rückhalt und Motivator für mich war.

Inhaltsverzeichnis

1 Zusammenfassung

In der vorliegenden Dissertation wird sich mit der diagnostischen Kompetenz von Grundschullehrern beschäftigt, wobei deren Struktur und ihre Bedingungen im Zentrum der Betrachtungen stehen. Unter diagnostischer Kompetenz wird bei Lehrern deren Fähigkeit verstanden, Schülerleistungen und -merkmale sowie die Schwierigkeit von Aufgaben korrekt einzuschätzen. Diese Fähigkeit gilt als Schlüsselkompetenz in Lehr- und Lernkontexten, da ihr eine hohe Bedeutung für adäquate Unterrichtsgestaltung sowie faire und objektive Beurteilungen beigemessen wird.

Eine Vielzahl an Forschungsbefunden belegt, dass Lehrkräfte zwar im Mittel gute Diagnostiker sind, dass jedoch große interindividuelle Unterschiede bestehen. Dabei waren die bisherigen Untersuchungen überwiegend querschnittlich angelegt und auf einzelne oder wenige Leistungsbereiche bezogen. Aussagen dazu, wie bereichsspezifisch und stabil die Güte von Lehrerurteilen ist, waren somit bislang kaum möglich. Ebenso erfolgte die Suche nach den Ursachen für die Unterschiedlichkeit zwischen Lehrern in aller Regel nur anhand weniger Lehrer- oder Klassenmerkmale, ohne dass jedoch erklärende Variablen gefunden wurden.

An diese Desiderata wird in dieser Arbeit angeknüpft, indem quer- und längsschnittlich und unter Einbezug einer Vielzahl potentiell erklärender Merkmale die Urteilsgüte in mehreren kognitiven und emotional-motivationalen Bereichen erhoben wird. Zentrale Fragestellungen beziehen sich dabei auf strukturelle Aspekte wie jenen der Bereichshomogenität und Stabilität der Urteilsgüte sowie der Reliabilität der Urteilskomponenten. Bedingungen der Urteilsgenauigkeit werden auf Ebene der Lehrer, der Klassen und der individuellen Schüler vermutet und untersucht. Darüber hinaus werden auch Zeugnisnoten als eine besonders bedeutungsvolle Form der Lehrerurteile betrachtet.

Um die Fragestellungen beantworten zu können, standen zwei Stichproben bayerischer und hessischer Grundschüler sowie ihrer Lehrer aus dem Bamberger BiKS-Projekt zur Verfügung, von denen eine (N = 2395 Schüler aus 155 Klassen) längsschnittlich über drei Erhebungszeitpunkte in der dritten und vierten Klassenstufe und die andere (N = 822 Schüler aus 146 Klassen) querschnittlich in der ersten Klassenstufe untersucht wurde. Aus der Gegenüberstellung von durch die Schüler bearbeiteten Leistungstests und Fragebögen und von den Lehrern ausgefüllten schülerbezogenen Einschätzbö-

gen und Fragebögen konnten verschiedene Komponenten der Urteilsgenauigkeit errechnet werden.

Die Ergebnisse der Arbeit bestätigen hinsichtlich der Urteilsgüte frühere Befunde, nach denen die mittlere Rangurteilsgüte der Lehrer im mittelhohen Bereich liegt und die Schüler tendenziell überschätzt werden. Die Urteilsgenauigkeit zu Leistungsmaßen liegt dabei deutlich höher als zu emotional-motivationalen Maßen wie der Leistungsängstlichkeit und dem Fachinteresse. Neuwert bringt vor allem die Erkenntnis, dass die Urteilsgüte fachspezifisch ausgeprägt ist und zwischen inhaltlich ähnlichen Bereichen höhere Zusammenhänge aufweist als zwischen unähnlichen Bereichen. Ebenfalls neu ist der Beleg, dass die Urteilsgüte sich als relativ zeitstabil erweist. Als Bedingungen für die diagnostische Kompetenz konnten vor allem individuelle Schülermerkmale wie das Geschlecht und der Sozialstatus herausgestellt werden, während sich trotz großer Auswahl an theoretisch plausiblen Einflussfaktoren weder Lehrer- noch Klassenmerkmale als bedeutsam für die Urteilsgüte erwiesen. Lediglich die Heterogenität der Klasse korrelierte mit der Ausprägung der Rangurteile. Weiterhin zeigten sich sehr enge Zusammenhänge zwischen den per Fragebogen erfassten Lehrereinschätzungen und den von ihnen erteilten Zeugnisnoten für die Schüler, wobei beide Formen des Lehrerurteils mit einem großen Streubereich der entsprechenden Leistungen auf Schülerseite einhergehen, objektiv gleiche Leistungen also von verschiedenen Lehrern teils sehr unterschiedlich bewertet werden. Nicht zuletzt scheint ein weiterer Befund der Arbeit aus methodischer Sicht aufschlussreich: so variiert die in der Forschung - im Vergleich zur Niveau- und zur Streuungskomponente - am häufigsten verwendete und als Kernstück bezeichnete Rangkomponente diagnostischer Kompetenz stark in Abhängigkeit von der Zusammensetzung der Klasse. Das Fehlen einzelner Schüler kann zu einem deutlich veränderten Indikator der Rangurteilsgüte führen, was die Aussagekraft von auf der Rangkomponente basierenden Studien einschränkt und die Bedeutung von Schülermerkmalen im diagnostischen Prozess betont.

2 Allgemeiner Teil

2.1 Einführung

Als Reaktion auf die aus deutscher Sicht enttäuschenden Ergebnisse der PISA 2000-Studie und die offenbar werdenden schlechten Leistungen deutscher Schüler legte die Kultusministerkonferenz (KMK) auf ihrer 296. Plenarsitzung am 5. und 6. Dezember 2001 als erste Konsequenz sieben Handlungsfelder vor, in denen die Länder und die Kultusministerkonferenz selbst vorrangig tätig werden sollten. Neben Maßnahmen zur Verbesserung der Sprachkompetenz oder zur Förderung bildungsbenachteiligter Kinder wurde als sechstes Handlungsfeld gefordert, „Maßnahmen zur Verbesserung der Professionalität der Lehrertätigkeit, insbesondere im Hinblick auf diagnostische und methodische Kompetenz als Bestandteil systematischer Schulentwicklung“ zu ergreifen (Kultusministerkonferenz, 2001). Auslöser für diesen plötzlichen bildungspolitischen Bedeutungsgewinn der Diagnosekompetenz von Lehrern[1] war unter anderem die aus der PISA 2000-Studie gewonnene Erkenntnis, dass Lehrer insbesondere bei der Identifizierung von schwachen Schülern im Bereich Lesen erhebliche Defizite aufwiesen. Danach gefragt, welche ihrer Schüler besonders schwache Leser seien, erkannten Hauptschullehrer weniger als 15 Prozent derjenigen Kinder, die sich im PISA-Test tatsächlich als Risikoschüler erwiesen (Artelt et al., 2005). Dabei beinhaltete die PISA-Studie - wie auch sonst keine der großen internationalen Schulleistungsstudien - gar keine umfassende und explizite Erhebung der diagnostischen Kompetenz der Lehrer (Artelt, Stanat, Schneider & Schiefele, 2001), war durch die Beschränkung auf Hauptschullehrer nicht repräsentativ, und die abgefragten Einschätzungen waren in mehrfacher Hinsicht für die Lehrer nicht einfach (u.a. fehlende Vertrautheit mit den PISA-Kompetenzstufen oder die Diskrepanz zwischen der einzuschätzenden Kompetenz und der im Test gemessenen Performanz). Dies alles wirkte sich einschränkend auf die Aussagekraft der Ergebnisse aus. Dennoch war dieser Befund ausreichend, um öffentliches Interesse an dem Thema zu wecken und sogar die Bildungspolitik zu erreichen.

[1] In der vorliegenden Arbeit findet das generische Maskulinum Verwendung, das sich nach der gegenwärtigen Konvention gleichermaßen auf Frauen und Männer bzw. Mädchen und Jungen bezieht. Falls im Zusammenhang nur männliche oder weibliche Personen gemeint sind, geht dies aus der Formulierung hervor.

Obwohl die pädagogisch-psychologische Diagnostik spätestens seit 1969, als Ingenkamp (vgl. 2005) diesen Begriff mit dem dazugehörigen Konzept populär machte, eine wichtige Rolle in Schule und Schulforschung spielt und lange vor PISA eine Vielzahl empirischer Untersuchungen die ‚Mängel' an der diagnostischen Kompetenz von Lehrkräften belegte, erlebte die - auch politische - Forderung nach verstärkter Anwendung pädagogisch-psychologischer Diagnostik im schulischen Kontext erst im Nachgang der PISA 2000-Studie eine „Renaissance" (Hesse & Latzko, 2009). Die Einschätzung zur Notwendigkeit diagnostischer Kompetenz war in den vergangenen Jahrzehnten immer wieder diskontinuierlichen Schwankungen unterworfen, und das sowohl in der Lehrerausbildung als auch in der praktischen Arbeit in den Schulen. So gab es beispielsweise schon 1970 eine ähnliche Einschätzung der diagnostischen Fähigkeiten von Lehrern durch den Deutschen Bildungsrat wie über dreißig Jahre später durch die KMK: „Ein ungerechtfertigter subjektiver Glaube an die eigene Fähigkeit, Schulleistungen objektiv richtig bewerten zu können, und das Fehlen einer ausreichenden Schulung zur Erhöhung der Objektivität und Rationalität von Leistungsbewertungen in der Lehrerbildung gehören zu den spezifischen Mängeln im deutschen Bildungswesen." (Deutscher Bildungsrat, 1970). Wie Ingenkamp (1991; Ingenkamp & Lissmann, 2008) feststellt, erlebte das Bemühen um eine Verbesserung der diagnostischen Kompetenz von Lehrern zu Anfang der 70er Jahre einen spürbaren Auftrieb, der jedoch schon ab etwa 1975 wieder deutlich abflachte. Die einsetzende „Anti-Test-Bewegung" in den 60er und 70er Jahren (Zeuch, 1973) mit ihrer strikten Ablehnung der Objektivierung von Schülerleistungen führte nicht zuletzt auch zu einem Rückgang der Bemühungen, der Diagnostik einen hohen Stellenwert in der Lehrerausbildung einzuräumen, und die Anwendung von einheitlichen Testverfahren in den Schulen, die erst kurz zuvor an Bedeutung gewonnen hatte, ging trotz zunehmender Anzahl an zur Verfügung stehender Test- und Fragebogenverfahren deutlich zurück. Erst seit etwa dem Jahr 2002 beschäftigt sich zumindest die Bildungsforschung in Deutschland spürbar intensiver als zuvor mit dem Thema. Viele wissenschaftliche Arbeitsgruppen widmeten sich dem Thema, Schwerpunkthefte bedeutender deutschsprachiger wissenschaftlicher Zeitschriften erschienen (Pädagogik, Heft 4, 2003; Zeitschrift für Pädagogische Psychologie, Heft 3-4, 2009), wenngleich auch in neuester Zeit der Einsatz von Tests in der pädagogischen Diagnostik nicht ohne Kritik bleibt und auch die alten Argumente („Vertestung" von Schule etc.) reaktiviert werden. Möglichkeiten der Verbesserung der Diagnosekompetenz der Lehrer wurden gesucht, Vergleichsarbeiten als Angebot zur Selbstevaluation gefunden. Doch am Ausgangspunkt mangelnder Lehrerkompetenzen, näm-

lich der Lehrerausbildung, hat sich in den zurückliegenden Jahren wenig geändert. Noch immer ist die Diagnostik an den wenigsten deutschen Universitäten als Pflichtveranstaltung in Lehramtsstudiengängen vorgesehen, lediglich im Lehramtsstudium für Sonderschulen taucht sie als solche auf (vgl. z.B. Helmke, 2003). Der Stellenwert der Diagnostik erfährt erst seit kurzer Zeit im Zuge der Umstellung der Lehramtsstudiengänge auf das Bachelor- und Mastersystem eine Aufwertung. Durch die bisherige Struktur der Lehramtsausbildungsgänge war und ist jedoch nicht gewährleistet, „dass Leistungsbeurteilungen im Sinne eines berufstypischen Handelns von Lehrerinnen und Lehrern als pädagogische Handlungs- oder Beurteilungskompetenz aus der praktischen und wissenschaftlichen Ausbildung und Erfahrung der Lehrenden hervorgeht“ (Beutel, 2007), so dass die Herausbildung dieser Kompetenzen vor allem eine Aufgabe der berufsbegleitenden Selbstaufklärung und Professionalisierung ist. Darüber hinaus fehlt es oft an kompetenten Fachleuten in der Lehrerfortbildung, so dass die Mängel der universitären Ausbildung auch bei Lehrkräften, die schon lange im Dienst sind, nur unzureichend ausgeglichen werden können. Insofern verwundert es nicht, dass der Ruf nach besserer diagnostischer Kompetenz der Lehrer immer wieder ertönt. Der Berufsverband Deutscher Psychologinnen und Psychologen (BDP) unterstützte erst Mitte 2008 wieder in einer Presseerklärung die Forderung der Bundesvereinigung der Deutschen Arbeitgeberverbände (BDA) nach einer Stärkung der psychologischen, diagnostischen und pädagogischen Anteile in der Aus- und Weiterbildung von Lehrern (Berufsverband Deutscher Psychologinnen und Psychologen (BDP), 2008). Während etwa im PISA-Vorzeigeland Finnland zwei Drittel der Grundschullehrerausbildung für pädagogische, psychologische und didaktische Inhalte verwendet wird, ist die Lehrerausbildung in Deutschland eher eine Ausbildung von Fachwissenschaftlern, in der fachdidaktische Elemente nur eine untergeordnete Rolle spielen. Gerade diese sind aber wichtig dafür, dass Lehrer neben der reinen Wissensvermittlung auch in der Lage sind, die Unterrichtsgestaltung an den Leistungsstand der Schüler anzupassen, Probleme zu erkennen und adäquate Einschätzungen von Leistungen und Eigenschaften vorzunehmen.

Die vorliegende Arbeit dient daher in gewisser Weise auch als Bestandsaufnahme, indem sie für zwei deutsche Bundesländer, Bayern und Hessen, die Ausprägung der diagnostischen Kompetenz der Lehrer ein knappes Jahrzehnt nach der ersten PISA-Studie und inmitten einer Phase der Rückbesinnung auf die Bedeutung guter Diagnostik nachzeichnet. Dabei wird im zweiten Kapitel dieser Arbeit nach einer Begriffsklärung und Definition di-

agnostischer Kompetenz zunächst auf deren Bedeutung eingegangen. Im dritten Abschnitt werden vordergründig aus theoretischer Sicht strukturelle Aspekte der Diagnosekompetenz betrachtet, so zum Beispiel ihre Gütekriterien, Komponenten und Maßstäbe, aber auch Urteilsfehler, die Lehrkräften im Diagnoseprozess unterlaufen können. Teil vier der Arbeit widmet sich den Bedingungsfaktoren diagnostischer Kompetenz und verknüpft den theoretischen Überblick ausführlich mit empirischen Belegen, so dass in diesem Abschnitt der Stand der Forschung expliziert wird. Anschließend werden daraus in Kapitel 5 die Fragestellungen dieser Arbeit abgeleitet, die sich grob in Fragen zur Struktur, zu den Bedingungen und zum Notenbezug diagnostischer Kompetenz unterteilen lassen. Schließlich wird im sechsten Teil relativ ausführlich auf das methodische Vorgehen eingegangen, wobei neben der Beschreibung der Instrumente und der Stichproben unter anderem auch auf die Imputation fehlender Werte eingegangen wird. Es folgen analog zu den Fragestellungen im Kapitel 7 die Ergebnisse der statistischen Analysen sowie in Kapitel 8 die abschließende Diskussion.

2.2 Definition diagnostischer Kompetenz

Der Begriff der „Diagnostischen Kompetenz"

Mittlerweile sind Begriffe wie ‚diagnostische Kompetenz', ‚Diagnosekompetenz', ‚diagnostische Expertise' oder ‚diagnostisches Wissen' geläufige Wendungen, doch wie so oft in der Wissenschaft gibt es auch für sie keine alleinige allgemeingültige Definition. Man findet ein Spektrum an Erläuterungen, die sich überwiegend sehr ähnlich sind, die sich im Detail jedoch auch unterscheiden. Im Handwörterbuch Pädagogische Psychologie, in dem der diagnostischen Kompetenz ein eigenes Kapitel gewidmet ist, wird sie als „die Fähigkeit eines Urteilers, Personen zutreffend zu beurteilen", beschrieben, die somit Grundlage für die Genauigkeit diagnostischer Urteile ist (Schrader, 2006). Sie wird nicht von vornherein auf Lehrer beschränkt, denn derartige Urteile können in vielen weiteren, sehr verschiedenen Kontexten gefällt werden: Eltern beurteilen ihre Kinder, Ärzte ihre Patienten und Vorgesetzte ihre Mitarbeiter. Da insbesondere die (Beurteilung von der) Leistungsfähigkeit einer Person auch davon abhängt, wie groß die jeweilige Anforderung ist, wird die genannte Definition mittlerweile von vielen Forschern auch um die Fähigkeit, die Schwierigkeit von Aufgaben korrekt zu erkennen, erweitert (vgl. z.B. Helmke, Hosenfeld & Schrader, 2004).

Im Englischen verwendet man üblicherweise entweder den eins zu eins übersetzten Begriff der ‚diagnostic competence', oder man spricht von ‚accuracy of judgement'. Letzterer Begriff hat auch im Deutschen eine Entsprechung, die „Urteilsgenauigkeit". Gerade in der US-amerikanischen Forschung widmet sich ein Großteil der Lehrerforschung auch den Lehrererwartungen, insbesondere hinsichtlich verschiedener Gruppenvergleiche (nach Geschlecht, ethnischer Herkunft oder Sozialstatus), weshalb dann eher von ‚teacher expectancy' die Rede ist. Sowohl methodisch als auch inhaltlich gibt es jedoch eine große Überschneidung zur Urteilsgenauigkeitsforschung, denn diese beinhaltet beinahe genauso oft differentielle Gruppenbetrachtungen wie die Erwartungsforschung den Vergleich zu aktuellen Schulleistungen heranzieht.

Die oftmals synonyme Bedeutung von „diagnostischer Kompetenz" und „Urteilsgenauigkeit" im Deutschen deutet auch auf eine Ungenauigkeit hin: genau genommen ist nämlich gute diagnostische Kompetenz eine Vorläuferbedingung von hoher Urteilsgenauigkeit. Nur wer über diese Kompetenz verfügt, ist auch in der Lage, zutreffende Einschätzungen vorzunehmen. Mit dieser Annahme ist man schnell bei der grundlegenden Begriffsklärung: Zerlegt man den zusammengesetzten Begriff in seine Bestandteile, kann man ableiten, dass es sich um eine Kompetenz handelt, die einen befähigt, Diagnosen zu stellen. Doch was bedeutet überhaupt ‚diagnostisch' oder ‚Diagnose', und was versteht man unter einer ‚Kompetenz'?

Der Begriff ‚Diagnose' leitet sich aus dem griechischen ‚diágnosi' ab, eine Zusammensetzung aus ‚dia' („durch") und ‚gnósi' bzw. ‚gignosko' („Urteil", „Erkenntnis" bzw. „erkennen"), die wörtlich also die ‚Durchforschung' im Sinne einer ‚unterscheidenden Beurteilung' bedeutet (Kluge, 1975; Wissenschaftlicher Rat der Dudenredaktion, 1997). Der Kompetenzbegriff geht hingegen auf lateinischen Ursprung zurück: ‚competere' bedeutet so viel wie ‚zu etwas fähig sein'. Es handelt sich also um die Fähigkeit, (korrekte) unterscheidende Beurteilungen abzugeben. Der Begriff der Kompetenz ist dabei seit über fünf Jahrzehnten geradezu ein Modebegriff, der in vielen Wissenschaftsdisziplinen gern und häufig, aber auch nicht immer einheitlich verwendet wird. Seine Popularität begann u.a. mit Chomskys Theorie der Sprachkompetenz, seit Anfang der 1970er Jahre wird er auch in der Erziehungswissenschaft genutzt (Roth, 1971). In der Bildungsforschung wird der Kompetenzbegriff gern verwendet, um messbare Kriterien für das, was Schülern, Studenten und anderen Teilnehmern am Bildungsprozess vermittelt werden soll, zu definieren. So werden beispielsweise in der PISA-Studie

‚Lesekompetenz' und ‚Mathematische Kompetenz' der Schüler gemessen (z.B. Artelt et al., 2001; Baumert, Klieme, et al., 2001), in der DESI-Untersuchung werden ‚sprachliche Kompetenzen' erfasst (Jude & Klieme, 2007) und auch die Kultusministerkonferenz fordert in ihren Bildungsstandards zu vermittelnde Kompetenzen (s. u.a. Kultusministerkonferenz, 2004a). Wie Weinert (1999a) in einem einflussreichen Gutachten zur Definition und Auswahl von Kompetenzen für internationale Schulleistungsstudien über die schier unüberschaubare Menge an unterschiedlichsten Bedeutungen des Kompetenzbegriffs treffend resümiert, gibt es nicht eine allgemein akzeptierte Definition, sondern die jeweils zugrunde gelegten Definitionen widersprechen sich sogar teilweise. So werden sie einmal als auf spezifische Kontexte bezogene kognitive Leistungsdispositionen, ein anderes Mal aber auch als notwendige motivationale Orientierungen für die Bewältigung anspruchsvoller Aufgaben definiert (vgl. Hartig, 2006; Klieme, 2004). Gerade dann, wenn der Kompetenzbegriff in wissenschaftlichen Studien operationalisiert werden muss oder für die Überprüfung der Einhaltung von Bildungsstandards herangezogen wird, ist jedoch ein einheitliches Verständnis davon unabdingbar. In dieser Arbeit wird daher ein erweiterter Kompetenzbegriff nach Weinert zugrunde gelegt, der in jüngster Vergangenheit auf breiten Konsens gestoßen und quasi zu einem Referenzzitat geworden ist (Klieme, 2004). „Dabei versteht man unter Kompetenzen die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können" (Weinert, 2001, 2002). Zentral an der Definition ist, dass es sich bei Kompetenzen um (bei entsprechender Absicht und Willen) prinzipiell erlernbare Kenntnisse, Fähigkeiten und Fertigkeiten handelt, und die Formulierung „um bestimmte Probleme zu lösen" deutet an, dass diese mehr oder weniger bereichsspezifisch ausgeprägt sind. Dieser Kompetenzbegriff stellt die Grundlage u.a. für viele Schulleistungsstudien dar, die ihn entsprechend ihrer Konzeption mal enger oder weiter auslegen (vgl. bspw. für die PISA-Studie Baumert, Stanat & Demmrich, 2001), doch natürlich bietet er sich gleichermaßen für die Beschreibung von bestimmten Lehrerattributen an.

Erweiterung des Begriffs der diagnostischen Kompetenz

Einigen Forschern ist der Begriff der „Diagnostischen Kompetenz" als umfassende Fähigkeitsbezeichnung noch nicht treffend genug. Für Helmke

(2009) ist diagnostische Kompetenz lediglich die Urteilsgenauigkeit („accuracy"), so verstanden als die Übereinstimmung von einer Einschätzung mit einer bestimmten Merkmalsausprägung bei einer oder mehreren Personen. Helmke ist dieser Begriff allerdings zu eng gefasst, weshalb er stattdessen von „diagnostischer Expertise" spricht, wenn es ihm um das methodische, prozedurale und konzeptuelle Wissen einer (Lehr-)Person, das für die Einschätzung von Leistungen und Aufgaben nötig ist, geht (vgl. zur Systematisierung von Lehrerwissen auch Kapitel 4.6.1.2 ab S. 79). Nach seiner Definition ist also ein Lehrer dann ein diagnostischer Experte, wenn er Methoden zur Einschätzung von Schülerleistungen sowie zur Selbstdiagnose kennt und anwenden kann, sich typischer Urteilstendenzen und -fehler bewusst ist und darüber hinaus „ein hohes Niveau an zutreffender Orientiertheit" (Helmke, 2009) besitzt.

Nicht sofort einleuchtend erscheint, warum Helmke die Urteilsgenauigkeit einer Person mit dem Begriff der „diagnostischen Kompetenz" gleichsetzt und als umfassenderes Konzept noch einen weiteren Begriff („Expertise") strapaziert. Der Begriff der Kompetenz umfasst ja bereits viel mehr als die bloße Übereinstimmung eines Urteils mit einer Ausprägung, (s. oben), die sich genau genommen auch nur auf Personen und nicht auf bloße Übereinstimmungen zweier Ausprägungen beziehen kann. Genau dies bezeichnet Helmke als Expertise. In der vorliegenden Arbeit wird sich der gebräuchlicheren Definition von diagnostischer Kompetenz angeschlossen, die die für die Urteile nötigen Wissensanteile bereits beinhaltet, wohingegen der Begriff der Expertise vorwiegend im Sinne einer herausragenden Leistung einer Person verwendet wird.

Helmke, Hosenfeld und Schrader (2004) versuchten, die diagnostische Kompetenz hinsichtlich ihrer inhaltlichen Bestandteile zu untergliedern und betonen entsprechend, dass es zu kurz greifen würde, Diagnosekompetenz lediglich auf die Genauigkeit von Urteilen zu beschränken. Stattdessen sollten auch die dafür notwendigen Grundlagen einbezogen werden, für die die Autoren mangels einer allgemein anerkannten Konzeption folgende Strukturierung vorschlagen:

1. relativ stabile Merkmale wie Intelligenz und kognitive Komplexität als Grundlage des Diagnostizierens;
2. erfahrungsabhängige bereichsspezifische Fähigkeiten und Wissensstrukturen, genauer zum einen a) methodisches Wissen wie das Wissen über Urteilsfehler oder diagnostische Methoden und b) gegen-

standsspezifisches Wissen[2], z.B. Schwierigkeitsmerkmale von Aufgaben, typische Vorgehensweisen von Schülern oder Anforderungen in verschiedenen Lerngebieten. Das gegenstandsspezifische Wissen lässt sich in Anlehnung an die etablierte Klassifikation metakognitiven Wissens in Wissen über Aufgaben, Personen, Strategien sowie deren Interaktion einteilen (Schrader, 2006);

3. spezifische Kenntnisse (Wissen über einzelne Schüler und Klassen, deren Stärken und Schwächen, Schwierigkeit und Beliebtheit von Unterrichtsstoffen etc.).

Im Sinne des Kompetenzbegriffs beinhaltet die diagnostische Kompetenz darüber hinaus aber nicht nur Wissens-, sondern auch Handlungskomponenten, so dass neben dem deklarativen auch prozedurales Wissen erworben werden muss (Hascher, 2008). Zum methodischen Wissen bzw. für die diagnostisch-methodische Kompetenz von Lehrkräften sind in Deutschland - im Gegensatz zum Beispiel zu den USA - professionelle Standards bislang kaum entwickelt und Forschungsergebnisse rar (Arnold, 1999).

Pädagogische Diagnostik

Diagnostische Kompetenz kann als Bestandteil pädagogischer Diagnostik angesehen werden. Diese unterscheidet sich von psychologischer Diagnostik nicht notwendigerweise durch eigene Theorien oder Methoden, sondern in erster Linie durch ihren engen Bezug zu pädagogischen Entscheidungen und durch ihren Fokus auf Einzelfälle statt auf allgemeine Gesetzmäßigkeiten (vgl. Hesse & Latzko, 2009). Die gebräuchlichste Definition von Pädagogischer Diagnostik stammt von Ingenkamp (2005):

„Pädagogische Diagnostik umfasst alle diagnostischen Tätigkeiten, durch die bei einzelnen Lernenden und den in einer Gruppe Lernenden Voraussetzungen und Bedingungen planmäßiger Lehr- und Lernprozesse ermittelt, Lernprozesse analysiert und Lernergebnisse festgestellt werden, um individuelles Lernen zu optimieren. Zur Pädagogischen Diagnostik gehören ferner die diagnostischen Tätigkeiten, die die Zuweisung zu Lerngruppen oder zu individuellen Förderungsprogrammen ermöglichen sowie die mehr ge-

[2] Hierzu ist bislang noch wenig bekannt, Forschungen zu subjektiven oder impliziten Theorien von Lehrkräften, zu Lehrerkognitionen und zur Lehrerexpertise können aber erste Anhaltspunkte liefern (Bromme, 1997).

sellschaftlich verankerten Aufgaben der Steuerung des Bildungsnachwuchses oder der Erteilung von Qualifikationen zum Ziel haben."

2.3 Bedeutung diagnostischer Kompetenz

Lehrprozess

Lehrern kommt im gesamten Schulsystem eine zentrale und sehr verantwortungsvolle Rolle zu. Sie stellen die Schnittstelle zwischen dem Schulsystem und allen administrativen Bemühungen für eine gute Bildung auf der einen Seite und den Schülern auf der anderen Seite dar. Die Art und Weise, wie es Lehrern gelingt, alle Vorgaben zu erfüllen, wie effektiv sie Bildungsstandards umsetzen, ist mit entscheidend dafür, ob die individuellen Schüler und das Bildungssystem insgesamt erfolgreich sind (vgl. Medley, 1979). Die Genauigkeit von Lehrereinschätzungen zu den Leistungen ihrer Schüler ist besonders vor dem Hintergrund der Vielzahl von Entscheidungen, die tagtäglich getroffen werden müssen, von maßgeblicher Bedeutung. Dabei gibt es aus verschiedensten Richtungen teils sehr hohe Anforderungen an die Lehrkräfte, denen gerecht zu werden ihnen ein hohes Maß an notwendiger Expertise (vgl. Kapitel 4.6.1.1 zur Expertise von Lehrkräften ab S. 74) abverlangt. Dies betrifft zum einen die ‚Kernaufgaben' von Lehrern, zu denen es gehört, „Unterricht zu erteilen und verständnisvolles Lernen von Schülerinnen und Schülern systematisch anzubahnen und zu unterstützen" (Baumert & Kunter, 2006, S. 470), Unterricht im Klassenverband zu gestalten (Bromme, 1992) und Lernprozesse zu organisieren (Terhart, 2000). Die diagnostischen Fähigkeiten der Lehrer werden dabei meist als Basis für adaptives und remediales Unterrichten angesehen (Schrader, 1989). Von Seiten der Schule, der Eltern, anderer Lehrer und der Schüler selbst wird - darin inbegriffen - verlangt, dass Lehrer die schulischen Fähigkeiten der Schüler erfassen. Ihr Urteil ist darüber hinaus die grundlegende Informationsquelle bezüglich der schulischen Leistung von Kindern und Jugendlichen (Eckert & Arbolino, 2005; Salvia & Ysseldyke, 2004; Shapiro & Kratochwill, 2000). Zutreffende Einschätzungen sind weiterhin Grundlage für die Beratung von Schülern und Eltern (u.a. Hertel, Bruder & Schmitz, 2009), die - hier am Beispiel Nordrhein-Westfalen - in der Allgemeinen Dienstordnung für Lehrerinnen und Lehrer (ADO, §8) sowie in der Allgemeinen Schulordnung (ASchO, §39) als wichtiger Tätigkeitsbereich von Lehrern und als Schlüsselkompetenz in ihrem professionellen Handeln festgeschrieben steht. Lehrer müssen ebenso Entscheidungen bezogen auf die Umsetzung des Curricu-

lums, die Art des Unterrichtens, die Auswahl von Unterrichtsmaterialien oder die leistungsmäßige Gruppierung der Schüler treffen, sie müssen gelegentlich festlegen, ob Schüler Förderunterricht erhalten, und nicht zuletzt obliegt ihnen natürlich die Verantwortung der Noten- und Zeugnisnotenvergabe, die Entscheidung über Bildungswege und damit in weiten Teilen auch über Entwicklungschancen im Lebenslauf (Clark & Peterson, 1986; Corno & Snow, 1986; McNair, 1978; Rogalla & Vogt, 2008; Sharpley & Edgar, 1986).

In Deutschland hat die Kultusministerkonferenz Anforderungen für die Lehrerbildung definiert, die sich auf die in den einzelnen Bundesländern formulierten Bildungs- und Erziehungsziele beziehen. Den dort beschriebenen Zielen von Schule liegt ein Berufsbild für Lehrkräfte zugrunde, dessen wichtigste Aspekte im Jahr 2000 in einer gemeinsamen Erklärung des Präsidenten der Kultusministerkonferenz und der Vorsitzenden der Lehrerverbände in fünf Punkten zusammengefasst wurden. Neben der Kernkompetenz, Fachleute für das Lehren und Lernen zu sein, sowie der Erziehungsaufgabe ist im dritten Punkt die Beurteilungsaufgabe beschrieben: „Lehrerinnen und Lehrer üben ihre Beurteilungs- und Beratungsaufgabe im Unterricht und bei der Vergabe von Berechtigungen für Ausbildungs- und Berufswege kompetent, gerecht und verantwortungsbewusst aus. Dafür sind hohe pädagogisch-psychologische und didaktische Kompetenzen von Lehrkräften erforderlich.“ (Kultusministerkonferenz, 2004b). Es folgen die Weiterentwicklung von Kompetenzen sowie die Beteiligung an der Schulentwicklung als Punkte vier und fünf. Die hier noch recht allgemein gehaltenen Anforderungen werden in den Standards noch in Form von verschiedenen Kompetenzbereichen konkretisiert und nach Standards für theoretische und praktische Ausbildungsschritte differenziert. Damit werden die Anforderungen an Lehrkräfte sehr detailliert aufgeschlüsselt und es ist Sache der Hochschulen, den Erwerb derselben im Rahmen des Lehramtsstudiums sicherzustellen.

Die so formulierten Standards der Lehrerbildung machen deutlich, dass die Beurteilung von Schülern neben der reinen Stoffvermittlung und dem Erziehungsauftrag eine der wichtigsten Aufgaben ist, die Lehrer zu erfüllen haben, da sie die Grundlage für eine Vielzahl von Entscheidungen bilden. Nahezu in jedem Arbeitsschritt eines Lehrers geht es um Schülereinschätzungen. Mal erfolgen diese Einschätzungen unbewusst und laufen automatisiert ab, mal sind sie explizite und gründlich durchdachte Urteile. Abbildung 1 zeigt eine stark vereinfachte Darstellung des Lehrprozesses. Vor der

Vermittlung von Unterrichtsinhalten steht die Erfassung des Status Quo an (1): Was können die Schüler schon und wo kann demzufolge neues Wissen angeknüpft werden? Danach kommt die Wissensvermittlung an sich (2), wo u.a. das Tempo oder konkrete Beispiele und Aufgaben gewählt werden müssen und der Lehrer nicht zuletzt zu entscheiden hat, wann ein Thema lange genug behandelt wurde. Abschließend muss der Lehrer eine geeignete Form finden, in der er den Wissensfortschritt der Schüler ermittelt (3) und diesen bewertet.

Abbildung 1: Stark vereinfachte Darstellung des Lehrprozesses in der Schule

Unter anderem dadurch wird das Unterrichten zu einer hochkomplexen Tätigkeit, bei der eine Vielzahl von Informationen auf die Lehrer einströmt. Dabei passieren viele Dinge gleichzeitig, oft ist ein sofortiges Eingreifen durch den Lehrer nötig, und selten ist vorhersehbar, was als nächstes passiert. Bei alledem ist es wichtig, dass der Lehrer nicht den Überblick verliert, was nur durch eine effiziente Organisation und Strukturierung von Unterrichtsabläufen bewältigt werden kann (Shuell, 2004) und hohe Anforderungen an die Informationsaufnahme und -verarbeitung stellt.

Modell zu Leistungserwartungen, Unterricht und diagnostischem Urteil

Schrader und Helmke (2001, S. 47) haben in einem Modell zu Leistungserwartungen, Unterricht und dem diagnostischen Urteil die Vielschichtigkeit der möglichen Einflussfaktoren auf den diagnostischen Urteilsprozess dargestellt (vgl. Abbildung 2). Es verdeutlicht insbesondere die Beeinflussbarkeit und Steuerbarkeit des Lehrerhandelns bezüglich Erwartungen über Schülerleistungen oder Schülerverhalten, auch wenn es natürlich eine vereinfachte Darstellung von in der Realität deutlich komplexeren Abläufen ist. Im Modell wirken zwei Faktoren direkt auf die Lehrererwartungen. Dies ist zum einen der Klassenhintergrund hinsichtlich der Zusammensetzung von Klassen- und Schülermerkmalen, wodurch Erwartungen vom generellen Niveau her eingeschränkt werden. Zum anderen und insbesondere hängen Erwartungen aber auch von der Diagnosekompetenz der Lehrer ab, die auf diagnostischem Wissen (z.B. über die Leistungsfähigkeit von Schülern und

die Schwierigkeit von Aufgaben) und diagnostischen Fertigkeiten (u.a. der Beobachtungsgabe der Lehrer) beruht. Darüber hinaus sind allgemeine Ziele, Orientierungen und das Rollenverständnis der Lehrer von Bedeutung, was zum Beispiel die Bezugsnormorientierung der Lehrer (Rheinberg, 2001) einschließt (vgl. Kapitel 3.3).

Entsprechend dem Modell ergeben sich aus den Erwartungen einerseits die Art, wie leicht oder schwierig Lehrer ihren Unterricht gestalten und wie sie einzelne Schüler behandeln, andererseits auch das Arrangement von Leistungssituationen. Aus den Schülerleistungen folgt über die (subjektive) Beobachtung die Interpretation der Leistungen, die ihrerseits Rückwirkung auf die Erwartungen nimmt, aber auch Basis für die mannigfaltigen diagnostischen Urteile sind, seien es Prüfungsnoten oder Übergangsentscheidungen. Ein Rückbezug vom Urteil zu den Erwartungen wird im Modell nicht vorgenommen, wäre aber denkbar.

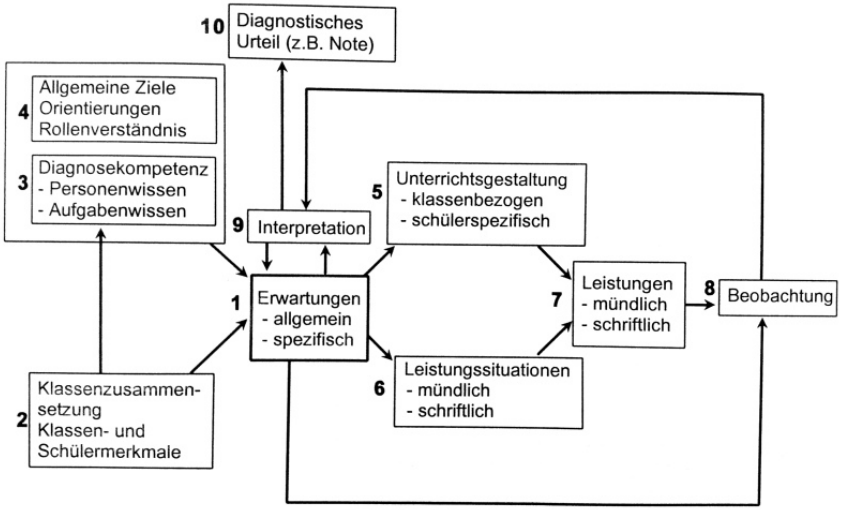

Abbildung 2: Leistungserwartungen, Unterricht und diagnostisches Urteil (Schrader & Helmke, 2001)

Jeder einzelne der angesprochenen und im Modell dargestellten Aspekte unterliegt einer gewissen Anfälligkeit für Verzerrungen und kann überdies zwischen verschiedenen Lehrern sehr unterschiedlich ausgeprägt sein. So besitzt sicher jeder Lehrer ein ihm eigenes Ensemble an Zielen und Orientierungen, Lehrkräften obliegt im Detail die Gestaltungsfreiheit für den Unterricht, und die Beobachtung der Unterrichtsprozesse unterliegt wie jede

Wahrnehmung der eigenen Fokussierung. Die Vielzahl der Ansatzpunkte für individuelle Ausprägungen des Urteilsprozesses macht es nicht nur für Lehrkräfte schwer, ein möglichst objektives Urteil zu fällen, sondern auch für die Forschung, mögliche Ursachen für Unterschiede zwischen Lehrern in Hinblick auf die Genauigkeit ihrer diagnostischen Urteile festzustellen.

Ein weiterer relevanter Einfluss auf diagnostische Urteile wird durch ein Modell aus der US-amerikanischen Forschung gegen Ende der 1970er Jahre postuliert, demzufolge Lehrer die Art ihrer Instruktionen sowie ihre didaktische Vorgehensweise im Unterricht an ihrem Eindruck vom Lernstand und den aktuellen Leistungen der Schüler orientieren (u.a. Borko, Cone, Russo & Shavelson, 1979; Shavelson & Stern, 1981). Die Einschätzung der Lehrer von den kognitiven Fähigkeiten ihrer Schüler wird darin als Einflussfaktor für a) ihre Annahmen über das Unterrichten selbst, b) die wahrgenommenen Eigenschaften von Aufgabenstellungen im Unterricht und c) die verfügbaren Informationen über die Schüler angesehen. Es ist davon auszugehen, dass es einen Unterschied für die Unterrichtsgestaltung macht, ob die zugrunde liegenden Lehrerurteile korrekt oder inkorrekt sind (vgl. Clark & Peterson, 1986).

Darüber hinaus gibt es Belege dafür, dass der interaktive Prozess der Entscheidungsfindung über Instruktionen im Unterricht auch eine Beurteilungskomponente enthält (u.a. Brophy, 1984). So gaben Lehrer in Interviews an, viele ihrer Entscheidungen auf Annahmen über individuelle Schüler zu stützen (McNair, 1978), und in einer Interviewstudie von Colker (1984) bezogen sich 41 Prozent der von mehreren Lehrern berichteten Gedanken über Lehrprozesse auf Schülerwissen.

Leistung und Leistungsentwicklung der Schüler

Theoretisch wird häufig auch eine Auswirkung der diagnostischen Kompetenz auf die Leistungsentwicklung der Schüler angenommen. Die wenigen empirischen Ergebnisse fielen bislang jedoch sehr unterschiedlich und teils widersprüchlich aus. So wurde einerseits beispielsweise in einer schon länger zurückliegenden Untersuchung aus dem amerikanischen Raum ein positiver Zusammenhang zwischen der Fähigkeit, die Aufgabenschwierigkeit in den Bereichen Deutsch und Mathematik korrekt einzuschätzen, und sowohl der Leistung als auch des Engagements der Schüler gefunden (Fisher et al., 1978). Eine neuere Studie für das Land Brandenburg kam mit ähnlicher Fragestellung allerdings zu insofern inkonsistenten Ergebnissen, als dass sich dieser Zusammenhang nur für einzelne Klassen und Schulformen

zeigte (Lehmann et al., 2000). Zu berücksichtigen sind hierbei natürlich die völlig unterschiedlichen Bildungssysteme und Zeitpunkte.

Häufig zitiert wird in diesem Zusammenhang auch ein Forschungsergebnis aus der Münchner Studie „Unterrichtsqualität und Leistungszuwachs“, in der ein von verschiedenen Bedingungen abhängiger Einfluss der Diagnosekompetenz auf die Leistungsentwicklung von Schülern im Fach Mathematik nachgezeichnet werden konnte (Schrader & Helmke, 1987). Demnach ist es für den Lernerfolg optimal, wenn hohe diagnostische Kompetenz gemeinsam mit dem gezielten Einsatz häufiger Strukturierungshilfen auftritt. Eine schlechte Diagnosekompetenz in Verbindung mit wenig Strukturierung war hingegen schon mit deutlich geringeren Lernerfolgen verbunden. Kurioserweise wirkte es sich aber besonders negativ auf die Lernfortschritte aus, wenn eine niedrige diagnostische Kompetenz mit viel Strukturierung und noch mehr wenn hohe diagnostische Fähigkeiten mit wenig Strukturierung verbunden waren. Diese Befunde sind nicht leicht zu interpretieren, und die Autoren erklären sie sich so, dass von diagnostischer Kompetenz nicht linear auf Lernerfolg geschlossen werden kann, sondern der Diagnosekompetenz eher die Rolle eines Katalysators zukommt (Helmke, 2003).

In einer aktuellen Untersuchung von Anders und Kollegen konnte anhand einer Stichprobe von 16-jährigen Schülern aus dem PISA-Längsschnitt und ihrer Lehrer gezeigt werden, dass Lehrkräfte, die die Leistungen ihrer Schüler in einzelnen Aufgaben korrekt einzuschätzen vermögen, in Klassenarbeiten Aufgaben stellen, die nach Ansicht von Experten ein höheres Potential zur kognitiven Aktivierung besitzen (Anders, Kunter, Brunner, Krauss & Baumert, 2010). Die Autoren betrachten dies als Nachweis des Zusammenhangs zwischen der Urteilsgüte und der Unterrichtsqualität. Der Gültigkeitsbereich dieser Interpretation beschränkt sich aber auch hier auf das Fach Mathematik und die Niveaukomponente diagnostischer Kompetenz; für die Rangkomponente konnten keine derartigen Zusammenhänge gefunden werden.

3 Strukturelle Aspekte diagnostischer Kompetenz

Nachdem im ersten Kapitel ein kurzer Überblick zur Bedeutung der diagnostischen Kompetenz im schulischen Kontext sowie zu ihrer Einbettung in verschiedene unterrichtliche Modelle gegeben wurde, soll nun aus theoretischer und Forschungssicht auf die Struktur, die Konstruktmerkmale sowie auf generell mit (Lehrer-)Urteilen im Zusammenhang stehende Faktoren eingegangen werden. Zunächst wird dabei darauf eingegangen, inwiefern die klassischen Gütekriterien auch auf Lehrerurteile übertragbar sind.

3.1 Gütekriterien diagnostischer Urteile

Die Literatur zu Gütekriterien von Testverfahren ist sehr umfangreich (Amelang & Schmidt-Atzert, 2006; Bortz & Döring, 2006; Bühner, 2006), doch auch für Lehrerurteile als weitere Form diagnostischer Verfahren gelten Kriterien, an denen sich die Qualität derselben ablesen lässt. Daher werden die wichtigsten Gütekriterien, nämlich Objektivität, Zuverlässigkeit (Reliabilität) und Gültigkeit (Validität), im Folgenden auf den Sachverhalt der pädagogischen Diagnostik in der Schule, also das Einschätzen und Beurteilen durch Lehrer, bezogen dargestellt.

Objektivität

Von Objektivität wird dann gesprochen, wenn Urteile möglichst unabhängig vom Urteilenden sind. Im schulischen Kontext bedeutet dies, inwieweit verschiedene Lehrer in ihren Urteilen über den gleichen Sachverhalt übereinstimmen. Lehrerurteile sind entsprechend dann objektiv, „wenn intersubjektive Einflüsse der Untersucher möglichst ausgeschaltet werden können“ (Ingenkamp, 2005). Um dies zu erreichen, müssten möglichst viele übereinstimmende Arbeitsschritte im Beurteilungsvorgang festgelegt werden, was gemeinhin mit der Unterscheidung in Durchführungs-, Auswertungs- und Interpretationsobjektivität versucht wird.

- Durchführungsobjektivität: Objektivität in der Durchführung versucht man dadurch zu erreichen, dass für alle Lernenden die gleichen Anforderungen unter gleichen Bedingungen gelten. Dazu trägt eine möglichst große Vereinheitlichung von allen durch die Lehrer beeinflussbaren Faktoren wie die Aufgabenstellung, die Bearbeitungszeit, Erläuterungen etc. bei. Völlige Gleichheit der Bedingungen ist hingegen nicht zu erwarten, da eine Reihe von Faktoren im Schüler selbst liegt, beispielsweise sein

Wohlbefinden, seine Motivation oder seine Leistungsangst. Wird bei standardisierten Testverfahren in der Regel der Ablauf detailliert von der Instruktion bis zur Zeitvorgabe festgelegt, ist dies bei alltäglichen Leistungsbeurteilungen nicht der Fall, da es keine allgemein verbindlichen Vorgaben gibt und jeder Lehrer selbst darüber entscheidet.

- Auswertungsobjektivität: Wie wiederholt gezeigt werden konnte, werden identische Schülerleistungen von verschiedenen Lehrern durchaus unterschiedlich beurteilt (vgl. Ingenkamp, 1995b). Das offensichtliche Fehlen objektiver Kriterien, nach denen Leistungen bewertet werden sollen, führt zu mangelnder Auswertungsobjektivität. Deutlich verbessern ließe sie sich zum Beispiel durch Aufgabenstellungen, zu denen eine Falschlösung zweifelsfrei von einer Richtiglösung unterschieden werden kann, wie es zum Beispiel bei Multiple-Choice-Fragen der Fall ist.
- Interpretationsobjektivität: Je zahlreicher und je unterschiedlicher die zur Verfügung stehenden Informationen sind, die bei der Beurteilung einer Leistung zur Verfügung stehen, desto schwerer fällt eine objektive Interpretation unter Ausschaltung aller intersubjektiven Einflüsse. Leiten verschiedene Lehrer aus einer Leistung die gleichen Schlussfolgerungen ab, so kann Interpretationsobjektivität angenommen werden (Tent & Stelzl, 1993). Im schulischen Kontext könnte dies beispielsweise die Zuordnung von verschiedenen Punktwerten in einer Leistungsüberprüfung zu Notenstufen sein.

Die Objektivität von Beurteilungen ist die entscheidende Voraussetzung für die anderen Gütekriterien. Ohne Objektivität können Messungen oder Einschätzungen auch nicht zuverlässig und gültig sein.

Reliabilität

Die Reliabilität (oder Zuverlässigkeit) von Messungen bezeichnet den Grad der Sicherheit oder Genauigkeit, mit dem ein Merkmal gemessen werden kann (vgl. Ingenkamp, 2005). Zuverlässig bedeutet insbesondere, dass dieselbe Leistung nach einiger Zeit immer noch genauso beurteilt wird wie beim ersten Mal. Legt man Lehrern mit zeitlichem Abstand denselben Aufsatz zweimal zur Bewertung vor, kann die Reliabilität der Beurteilung sehr genau gemessen werden. Im Idealfall würde er beide Male zur selben Einschätzung gelangen. Bei anderen Merkmalen wie beispielsweise der stark tagesformabhängigen Motivation der Schüler kann nicht erwartet werden, dass die Zuverlässigkeit von mehrfachen Einschätzungen sehr hoch ausfällt, denn die Reliabilität kann nicht höher sein als die Stabilität des einzuschätzenden Merkmals. Der Grad der Reliabilität und somit das Ausmaß, in dem

eine Messung reproduzierbar ist, kann durch einen Reliabilitätskoeffizienten angegeben werden (s. z.B. Lienert & Raatz, 1998, S. 9). Bei Aussagen zur Zuverlässigkeit wird immer davon ausgegangen, dass jedes Messergebnis einen wahren und einen verfälschenden Anteil enthält. Um das Verhältnis dieser Anteile zu schätzen, sind verschiedene Methoden verfügbar, deren gebräuchlichste die Wiederholungs- (Retest), die Halbierungs- (Split-Half) und die Paralleltestmethode sind.

- Wiederholungsmethode: Hierbei bearbeitet dieselbe Person dieselbe Aufgabe zu verschiedenen Zeitpunkten. Lehrer könnten sich selbst überprüfen, indem sie dieselben Schülerarbeiten mit zeitlichem Abstand doppelt beurteilen. Nicht außer Acht zu lassen sind hierbei natürlich Lern- oder Übungseffekte, die umso stärker zu Tage treten, je kürzer der Abstand zwischen den Zeitpunkten ist.
- Halbierungsmethode: Bei dieser Methode wird die Anzahl der zu beurteilenden Aufgaben in zwei Hälften geteilt, zum Beispiel durch die Auswahl jeder zweiten Aufgabe. Indem jede der Hälften getrennt beurteilt oder ausgewertet wird, kann durch anschließenden Vergleich der zwei Hälften die Halbierungszuverlässigkeit bestimmt werden. Bei zufälliger oder unsystematischer Zuweisung der Aufgaben zu den Hälften sollten zwischen ihnen keine großen Unterschiede bestehen. Auch dieses Verfahren könnten sich Lehrer zu Nutze machen, indem sie zum Beispiel erst eine Hälfte einer Klassenarbeit bei allen Schülern korrigieren und anschließend die andere Hälfte. Auch andere Vorgehensweisen sind denkbar.
- Paralleltestmethode: Um im Sinne eines Paralleltests zu bewerten, müssten zwei nahezu identische Leistungstests eingesetzt und bewertet werden. Die Urteile sollten dann nicht voneinander abweichen.

Alle Methoden der Reliabilitätsbestimmung zielen darauf ab, dass vom Grad der Übereinstimmung auf die Zuverlässigkeit der Messung oder Beurteilung geschlossen werden kann (Sacher, 2009).

Validität

Die Validität (oder Gültigkeit) eines Verfahrens, die Auskunft darüber gibt, ob tatsächlich das gemessen wurde, was zu messen beabsichtigt war, gilt als das wichtigste methodische Kriterium für Untersuchungsverfahren (Ingenkamp, 2005) und setzt ihrerseits hohe Objektivität und Reliabilität voraus (Jäger, 2000; Lienert & Raatz, 1998). Es gibt im Schulumfeld mannigfaltige Situationen, in denen die Frage nach der Validität relevant wird, bei-

spielsweise dann, wenn sich zum Urteil über die inhaltliche Qualität eines Aufsatzes auch die Anzahl der Rechtschreibfehler gesellt. Um zu entscheiden, ob man tatsächlich das gemessen hat, was man wollte, bedarf es eines Kriteriums, von dem es wiederum verschiedene gibt. Nachfolgend werden die bedeutsamsten vorgestellt.

- Inhaltsvalidität: Um Inhaltsvalidität zu gewährleisten, dürfen in Prüfungen nur solche Kompetenzen gemessen werden, die zu erwerben die Schüler im Vorfeld auch tatsächlich ausreichend Gelegenheit hatten (Jürgens, 2005). Das in der Schule immer wieder anzutreffende Abfragen von Inhalten, die im Unterricht nur am Rande behandelt wurden, verstößt beispielsweise dagegen. Die Form der Prüfung muss demzufolge der Form der Stoffvermittlung entsprechen. Eine Sonderform der Inhaltsvalidität ist die curriculare Validität, die dann gegeben ist, wenn Prüfungs- und Unterrichtsinhalte sich gleichermaßen an den Vorgaben durch den Lehrplan orientieren.
- Übereinstimmungs- oder Kriteriumsvalidität: Wenn gleichzeitig vorliegende Resultate zur selben inhaltlichen Facette, die jedoch mit unterschiedlichen Instrumenten gewonnen wurden, übereinstimmen, ist von Übereinstimmungsvalidität zu sprechen (Sacher, 2009). So sollte man annehmen können, dass mündliche und schriftliche Leistungsüberprüfungen zu einem bestimmten Stoffkomplex nicht zu deutlich unterschiedlichen Noten führen. Wird die Überprüfung anhand eines Außenkriteriums vorgenommen, ist der Begriff der ‚Kriteriumsvalidität' gebräuchlicher. Somit kann u.a. die Gültigkeit der von Lehrern vergebenen Noten anhand eines entsprechenden Leistungstests festgestellt werden. Die Urteilsforschung verwendet hierfür den noch exakteren Begriff der ‚Veridikalität', der konkret meint, dass ein „Prädiktor (das Lehrerurteil) mit einer möglichst guten (zumindest aber besseren) Messung des vorherzusagenden oder zu beurteilenden Merkmals, dem Kriterium, verglichen wird" (Helmke et al., 2004), wobei sich das Urteil exakt auf das Merkmal bezieht. Notwendige Voraussetzung dafür ist also, dass der Lehrer das einzuschätzende Merkmal genau kennt. Somit fallen allgemeine oder globale Einschätzungen (wie sie z.B. in der vorliegenden Arbeit von Lehrern erbeten wurden) nicht unter den Begriff der Veridikalität.
- Vorhersagevalidität: Vorhersage- oder Prognosevalidität ist dann gegeben, wenn aus Leistungsmessungen korrekte Schlüsse auf zukünftige Leistungen gezogen werden. Dies ist zum Beispiel bei Übertrittsempfehlungen am Ende der Grundschulzeit der Fall, wo die zukünftige Leistungsfähigkeit abgeschätzt werden muss. Die Überprüfung derselben

stellt Lehrer allerdings vor große Herausforderungen, da - wie im genannten Beispiel - viele andere Faktoren, u.a. die veränderte schulische Umgebung, Einfluss der Mitschüler, persönliche Ereignisse - im Prinzip gar nicht oder nur sehr schlecht vorhergesehen werden können.

- Konstruktvalidität: Während sich die beiden zuletzt genannten Formen der Validität empirisch überprüfen lassen, trifft dies auf die Konstruktvalidität nicht zu. Bei ihr liegt der Schwerpunkt zunächst darauf zu klären, ob die gemessenen Eigenschaften mit dem zugrunde liegenden theoretischen Modell übereinstimmen. Die zu messenden ‚Konstrukte' (z.B. Intelligenz oder Prüfungsangst) sind hierbei nicht unmittelbar beobachtbar, sondern müssen als latente, komplexe Merkmale abgeleitet werden. Ob sie valide gemessen wurden, kann nur abgeschätzt werden, indem geprüft wird, ob sich theoretisch erwartete Beziehungen zwischen beteiligten messbaren Eigenschaften nachweisen lassen. Bei der Schülerbeurteilung ist die Konstruktvalidität dann von besonderer Bedeutung, wenn nicht direkt beobachtbare Schülereigenschaften (z.B. ihr Fachinteresse) aus anderen - beobachtbaren - Merkmalen (z.B. ihrer Aufmerksamkeit oder ihrer Mitarbeit) abgeleitet werden müssen.

So plausibel die dargestellten Gütekriterien auch im Umfeld schulischer Leistungsbeurteilungen erscheinen, so sehr kann die alleinige Ausrichtung daran auch zu einer unangemessenen ‚Psychologisierung' führen. Ingenkamp (2005, erste Auflage 1985) warnte schon vor einem Vierteljahrhundert vor einer zu starken Dominanz von Modellen, Standards und Methoden der psychologischen Diagnostik in pädagogischen Kontexten, die sich bis heute erhalten hat, indem beispielsweise die Genauigkeit von Lehrerurteilen in Bezug auf Schülerleistungen am Maßstab standardisierter Testverfahren gemessen wird. Laut Weinert und Schrader (1986) hat es sich als unmöglich erwiesen, aufgrund psychometrisch gewonnener Kennwerte verschiedener Schüler (Intelligenz, Vorkenntnisse, Anstrengungsbereitschaft, Aufmerksamkeit etc.) die Leistungen in einzelnen Schulfächern hinreichend genau vorherzusagen, um daraus handlungsleitende Erwartungen, das passende Lehrerverhalten oder eine adäquate Unterrichtsgestaltung ableiten zu können. Sie schlagen daher alternative Gütekriterien vor.

- Lehrerdiagnosen während des Unterrichts müssten keineswegs besonders genau sein, wenn sich die Lehrer der Vorläufigkeit und Revisionsbedürftigkeit bewusst sind. Eine ungefähre Diagnose, die dafür aber im Verlauf des Unterrichts permanent überprüft wird, sei wichtiger.
- Desweiteren sei hohe Sensitivität für Verhaltens-, Wissens- und Motivationsänderungen der Schüler und gegenüber darauf einwirkender unter-

richtlicher Maßnahmen bedeutsam, wobei der Schwerpunkt auf Verlaufs- und nicht auf Zustandsdiagnostik liegen sollte.

- Wichtig sei ferner die Berücksichtigung verschiedener Beurteilungsmaßstäbe, neben sozial- und kriteriumsorientiertem vor allem das individuumszentrierte Bezugssystem, für das Rheinberg (Rheinberg, 1980, 2006) einen besonders großen unterrichtspraktischen Nutzen konstatiert (vgl. auch Kapitel 3.3 ab S. 39).
- Nicht zuletzt weisen Weinert und Schrader (1986) darauf hin, dass Lehrerdiagnosen sich nicht durch (praktisch ohnehin kaum zu erreichende) neutrale Objektivität, sondern eher durch eine pädagogisch günstige Voreingenommenheit auszeichnen sollten. Hiermit spielen sie auf den Aspekt an, dass sich eine mäßige Unterschätzung von Leistungsunterschieden zwischen Schülern einer Klasse und eine leichte Überschätzung des Leistungsniveau des Einzelnen sogar günstig und motivierend auswirken kann, wohingegen besonders exakte Urteile unterrichtspraktisch und psychologisch als unrealistisch zu bezeichnen sind.

Diese alternativen Gütekriterien sind insbesondere für den unterrichtlichen Alltag bedeutsam, während sich die Forschung zur diagnostischen Kompetenz eher an den klassischen Gütekriterien orientiert. Dennoch kann man insbesondere dem letzten der genannten Vorschläge auch kritisch gegenüberstehen, wie im Folgenden erläutert wird.

Genauigkeit von Lehrerurteilen

Gemeinhin geht man davon aus, dass ein Lehrer dann ein guter Diagnostiker ist, wenn seine Einschätzungen der Schülerleistungen besonders nahe an der Realität liegen. Im Idealfall wünschte man sich bspw. für die Rangkomponente diagnostischer Kompetenz (vgl. nachfolgendes Kapitel) eine Korrelation von $r = 1$; der Lehrer hätte dann ein perfektes Abbild der Rangverteilung der Schülerleistungen geschätzt. Gleiches trifft auf die Schätzung der Streuung sowie des Leistungsniveaus zu (vgl. Komponenten der Urteilsgenauigkeit nach Schrader und Helmke (1987)). Helmke (2009) stellt in Anlehnung an die eben erwähnten alternativen Gütekriterien nach Weinert und Schrader (1986) aber die Vermutung an, es sei günstiger, der Lehrer würde die Leistungsfähigkeit der einzelnen Schüler leicht über- sowie die Streuung der Leistungen zwischen den Schülern mäßig unterschätzen. Er meint, dass dieses leicht euphemistische Denken über die Klasse zu mehr Motivation und besserer Förderung der Klasse durch den Lehrer führe. Nichts sei wahrscheinlich so motivierend für den Lehrer wie eine leicht optimistische Erfolgserwartung.

Dieser Argumentation kann man aus mehreren Gründen skeptisch gegenüber stehen. Zum einen lässt Helmke offen, bis zu welchem Grad eine Erfolgserwartung noch „leicht optimistisch" ist. Ist es umso motivierender, je mehr ein Lehrer die Leistungen der Schüler überschätzt? Sicher nicht. Bis zu wie viel Prozent Überschätzung ist der Effekt aber noch förderlich, ab wann wirkt die Überforderung demotivierend? An dieser Stelle bleiben Unklarheiten. Zum anderen könnte man auch genau entgegengesetzt argumentieren: je schlechter ein Lehrer seine Klasse sieht, umso stärker wird er daran interessiert sein, das Leistungsniveau zumindest auf einen „normalen" Stand anzuheben. Wenn ein Lehrer es schafft, aus einer schlechten Klasse eine gute zu machen, führt dies zu Erfolgserlebnissen, zu einer Bestätigung der eigenen Lehrleistung, die umso motivierender für die eigene Arbeit und das weitere Lernen der Schüler ist. Was macht außerdem ein Lehrer, der tatsächlich die Schülerleistungen nahezu perfekt vorhersagen kann? Muss er sich aufgrund der angesprochenen Vermutung Helmkes Sorgen machen, für seine Schüler nicht motivierend genug zu sein? Soll er sich daraufhin selbst täuschen und wider besseres Wissen einen höheren Leistungsstand seiner Schützlinge annehmen?

Ein weiteres Argument gegen Helmke: Nach Vygotskijs Theorie der Stufe der nächsten Entwicklung (1978) ist der Lernerfolg dann besonders groß, wenn man kontinuierlich leicht überfordert wird. Ein Lehrer, der exakt einschätzen kann, ist daher in der Lage, bewusst das Anforderungsniveau anzuheben und damit genau jene Lernanreize zu erhöhen. Ein Lehrer, der nach Helmkes Dafürhalten ‚idealerweise' unbewusst ein höheres Leistungsniveau annimmt und außerdem nach Vygotskij seine Erwartungen noch bewusst erhöht, wird seine Schüler tatsächlich in einer Weise überfordern, die für das Lernen alles andere als förderlich ist.

Empirische Belege dafür, dass eine Überschätzung des Leistungsniveaus durch die Lehrer zu größerem Leistungszuwachs führen, sind bislang nicht gefunden worden. Aus Sicht des Autors sprechen die genannten Gründe deshalb klar dafür, Lehrern eine realistische Einschätzung der Schülerleistungen abzuverlangen. Es scheint wenig hilfreich zu sein, eine stetige leichte Überschätzung als günstig zu propagieren, solange nicht klar ist, in welchem Ausmaß und mit welchen Implikationen auf sonstige Fördermaßnahmen dies geschehen soll.

3.2 Komponenten der Diagnosegenauigkeit

Hinsichtlich der Art und Weise, wie Leistungseinschätzungen vorgenommen werden und worauf sie sich genau beziehen, gibt es unterschiedliche Möglichkeiten. Bereits vor über einem halben Jahrhundert wies Cronbach (1955) darauf hin, dass die Ermittlung der Übereinstimmung von Urteilen und tatsächlichen Merkmalsausprägungen ein nicht-triviales Problem darstellt, weil bei einer einfachen Differenzwertbildung verschiedene Komponenten der Urteilsgenauigkeit darin konfundiert wären. Auf einem von Cronbach entwickelten Komponentenmodell aufbauend unterteilten Schrader und Helmke (1987) die Diagnosegenauigkeit in drei - besonders im deutschen Sprachraum - bis heute oft zitierte und gebrauchte Komponenten, die Niveau-, die Streuungs- und die Rangordnungskomponente, auf die im Folgenden näher eingegangen wird.

Niveaukomponente

Die Niveaukomponente diagnostischer Kompetenz gibt an, ob Lehrkräfte Ausprägungen von Merkmalen zu hoch, zu niedrig oder gerade richtig einschätzen. Lehrer schätzen beispielsweise ein, wie viele Aufgaben in einem Test ein Schüler korrekt beantworten wird. Aus dem Vergleich von tatsächlichem Testwert mit dem Schätzwert mittels Differenzwertbildung resultieren entweder eine Überschätzung (z.B. von Leistungen), eine Unterschätzung oder idealerweise eine exakte Einschätzung der Merkmale. Die Genauigkeit der Niveaukomponente hängt u.a. davon ab, wie genau operationalisiert wird. Eine einfache Variante besteht darin, dass Lehrer einschätzen, wie viele Aufgaben eines Tests ein Schüler fehlerfrei lösen wird. Dabei wird jedoch die Ebene der einzelnen Aufgaben ignoriert. Es ist denkbar, dass der Lehrer zwar die Anzahl der Richtiglösungen richtig einschätzt (z.B. 10 von 20 Aufgaben korrekt), dass er jedoch dabei genau jene Aufgaben im Kopf hat, die der Schüler gerade nicht richtig löst. Ein einfaches Differenzmaß auf Basis der Anzahl korrekt gelöster Aufgaben greift daher möglicherweise zu kurz. Ein strengeres Kriterium ist ein Abgleich auf Aufgabenebene, bei dem Lehrer für einzelne Aufgaben einschätzen, ob die Schüler sie vermutlich richtig lösen werden oder nicht (aufgabenspezifische Niveaueinschätzung). Bei diesem Vorgehen ergeben sich Hinweise auf Über- oder Unterschätzung der Leistungen daraus, wie viele Aufgaben vom Schüler tatsächlich korrekt gelöst wurden, obwohl der Lehrer eingeschätzt hat, dass der Schüler diese Aufgabe nicht lösen wird (Unterschätzung) oder umgekehrt (Überschätzung). Aufgrund der Itembezogenheit dieser Einschätzung nannte Schrader

(1989) diesen Vergleich „Aufgaben-Personen-Wechselwirkungskomponente“, was z.B. in der Berliner COACTIV-Studie (Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz) als Begriff übernommen wurde (vgl. Baumert et al., 2008).

Einen Nachteil der ursprünglichen (einfachen) Niveaukomponentenberechnung sehen einige Forscher darin, dass sich bei der Mittelung über mehrere Schüler oder Lehrer Unter- und Überschätzungen gegenseitig aufheben können und insgesamt der Eindruck entsteht, dass Lehrer das Leistungsniveau optimal einschätzen können. Deshalb wurde vorgeschlagen, zusätzlich ein „globales Abweichungsmaß“ zu berechnen, das mit dem Betrag der Differenz zwischen Lehrereinschätzung und Schülerleistung rechnet und demzufolge nur ausdrückt, in welchem Ausmaß Verschätzung stattfindet, aber nicht berücksichtigt, ob es sich dabei um Über- oder Unterschätzung handelt (vgl. Südkamp, Möller & Pohlmann, 2008).

Streuungskomponente

Die Streuungskomponente diagnostischer Kompetenz gibt an, ob die Streuung der gemessenen Schülermerkmale der Streuung der Lehrerurteile entspricht. Bei deutlich geringerer Streuung der Lehrerurteile spricht man von der „Tendenz zur Mitte“, bei der man davon ausgeht, dass die Lehrer nicht in der Lage sind, die volle Bandbreite unterschiedlicher Leistungen in der Klasse zu erkennen, sondern stattdessen eine größere Leistungshomogenität vermuten. Aber auch das Gegenteil („Überdifferenzierung“) ist denkbar; hierbei werden geringe tatsächliche Leistungsunterschiede als deutlich größer vermutet.

Rangordnungskomponente

Die Rangordnungskomponente, kurz Rangkomponente oder auch ‚diagnostische Sensitivität‘ (Anders et al., 2010) genannt, wird oft als das Kernstück der diagnostischen Kompetenz bezeichnet und bringt zum Ausdruck, ob Lehrer in der Lage sind, ihre Schüler entsprechend ihrer Leistungen in eine Rangfolge zu bringen. Hier sind Korrelationskoeffizienten zwischen den Schüler- bzw. Aufgabenmerkmalen und den korrespondierenden Lehrerurteilen das zentrale Maß; im Falle einer perfekten Übereinstimmung ergäbe sich der Wert $r = 1$, eine komplett der Realität entgegengesetzte geschätzte Rangfolge ergäbe $r = -1$, und eine Zufallsangabe ohne jeglichen Zusammenhang zur tatsächlichen Rangfolge hätte den Wert $r = 0$. Die Rangkomponen-

te diagnostischer Urteile ist die am häufigsten erforschte und auch im Unterrichtsalltag gebräuchlichste Komponente der Diagnosekompetenz.

Auch die Rangordnungskomponente kann auf verschiedenen Wegen erhoben werden. Am korrektesten, aber auch am aufwendigsten ist es, dass alle einzuschätzenden Schüler in eine Rangreihe gebracht werden, indem sie entsprechend ihrer vermuteten Leistungen sortiert und nummeriert werden. Vereinfachte Vorgehensweisen können darin bestehen, beispielsweise nur die besten und schlechtesten fünf Schüler (mit oder ohne Sortierung) anzugeben oder jeden Schüler auf einer Skala einzuschätzen. Die Verwendung einer Skala fällt Lehrern in aller Regel leichter als exaktere Sortierverfahren, weil sie Differenzierungen auf einer groberen Metrik vornehmen und somit auch Schüler auf dieselbe Leistungsstufe stellen können. Je mehr Kinder einzuschätzen sind, desto deutlicher wird der daraus gewonnene Vorteil. Zudem kann bei großen Klassen mit vielen Kindern möglicherweise auch das eingesetzte Testverfahren als Kriterium gar nicht so genau zwischen den Schülerleistungen differenzieren, wie es bei einer Sortiervorgehensweise von den Lehrern erwartet würde. Daher hat sich in der Praxis die Operationalisierung mittels einer Skala am besten bewährt.

Varianten und Spezifität der Urteilskomponenten

Grundsätzlich ist die Berechnung der Urteilskomponenten nicht an bestimmte Operationalisierungen gebunden. So werden durchaus nicht selten aus personenbezogenen Aufgabenurteilen sowohl die Niveau- und Streuungskomponente als auch die Rangkomponente berechnet. Nicht zu vernachlässigen sind dabei u.a. Fragen nach dem jeweils zugrunde liegenden Maßstab und danach, ob sich Einschätzungen auf konkrete Aufgaben oder globale Fähigkeiten beziehen (vgl. nächster Absatz und folgende Kapitel 3.3 und 3.4).

Die Art und Weise, wie die vorgestellten Komponenten diagnostischer Kompetenz in der Praxis erfasst werden, ist entsprechend keinesfalls einheitlich (vgl. Begeny, Eckert, Montarello & Storie, 2008). Je nach Operationalisierung können Urteile u.a. mal im Vergleich zu Mitschülern und mal unabhängig von ihnen erfolgen, es können direkte Maße beurteilt werden, die im Klassenkontext erworbene Fähigkeiten zum Gegenstand haben (z.B. mathematische Fähigkeiten), oder Lehrerurteile werden mit Urteilen anderer Personen (z.B. Eltern) über die Schüler verglichen. Am häufigsten wird in der Forschung das Lehrerurteil standardisierten und normbezogenen Maßen gegenübergestellt, wobei die Einschätzung spezifisch oder global (gele-

gentlich - nicht ganz korrekt - auch direkt und indirekt genannt) vorgenommen werden kann. Der spezifische Weg meint, dass sich die Leistungseinschätzung der Lehrer auf bestimmte, ihnen bekannte Aufgaben bezieht, die die Schüler gelöst haben. Für beide Aspekte, Schülerleistungen und Lehrereinschätzungen, liegt also die gleiche Bezugsgröße vor, so dass ein unmittelbarer Vergleich gebildet werden kann. Das Ausmaß der Übereinstimmung solcher spezifischer Einschätzungen mit den Schülerleistungen wird auch als Veridikalität bezeichnet und stellt damit einen konkreten Sonderfall der kriterienbezogenen Validität dar (Helmke et al., 2004). Im Unterschied dazu beziehen sich globale Leistungseinschätzungen der Lehrer nicht auf die Schülerleistungen in bestimmten Aufgaben, sondern auf einen mehr oder weniger konkret benannten Leistungsbereich. Die Bitte an einen Lehrer, die Schülerleistungen im Bereich Arithmetik ganz allgemein auf einer fünfstufigen Likert-Skala (z.B. von 'sehr schlecht' bis 'sehr gut') einzuschätzen, entspräche demnach einer globalen Beurteilung. Auch hier wird aber meist diese aufgabenunabhängige globale Leistungseinschätzung den Leistungen der Schüler in standardisierten Testverfahren gegenübergestellt. Für beide Vorgehensweisen, die spezifische und globale Einschätzung, gibt es sowohl Vor- als auch Nachteile. Mittels spezifischer, aufgabenbezogener Leistungseinschätzung kann sichergestellt werden, dass Lehrer einen definierten Bezugsrahmen für ihre Urteile haben; bei globalen Urteilen riskiert man hingegen, dass Lehrer sich bei ihren Einschätzungen an Schülereigenschaften orientieren, die wenig oder nichts mit dem tatsächlich durch den Test gemessenen Konstrukt gemeinsam haben. Entsprechend zeigt sich in der Literatur auch, dass spezifische Urteile meist etwas genauer ausfallen als globale (vgl. Kapitel 4.1 ab S. 52). Dieser Vorteil kann aber auch als Nachteil angesehen werden, denn aus der Güte der Einschätzung von Schülerleistungen in einzelnen Aufgaben kann nicht induktiv auf die generelle Urteilsgenauigkeit in diesem Bereich geschlossen werden. Andererseits unterliegt auch die globale Einschätzung diesem Nachteil; hier liegt zwar ein allgemeines Urteil vor, dieses wird jedoch mit der Schülerleistung in einem speziellen Testverfahren in Beziehung gesetzt, was ebenso induktiv ist, wenn man daraus auf die generelle diagnostische Kompetenz schließen möchte. Unbestritten ist hingegen der Vorteil von globalen Leistungseinschätzungen, dass sie sich von den Lehrern deutlich schneller bearbeiten lassen, weil kein gesondertes Eindenken in konkrete Aufgaben erforderlich ist.

In ihrer Metaanalyse zur Genauigkeit von Lehrerurteilen haben Hoge und Coladarci (1989) die oben genannten Varianten der Urteilsspezifität entspre-

chend ihrer Genauigkeit folgendermaßen von unspezifisch bis sehr spezifisch kategorisiert:

1. Beurteilung von Leistungen auf einer Ratingskala, z.B. von ‚sehr schlecht' bis ‚sehr gut'
2. Schüler werden ihren Leistungen entsprechend in eine Rangfolge gebracht
3. Schülerleistungen werden Notenstufen zugeordnet
4. Schätzung der Anzahl richtig gelöster Aufgaben in einem Leistungstest
5. Einschätzung der Richtig-/Falsch-Lösung für jedes einzelne Item eines Leistungstests.

Die erste, unspezifischste Urteilsvariante, die in der Forschung am häufigsten Verwendung findet, erwies sich in der Metaanalyse gleichzeitig als diejenige, die zu den niedrigsten Übereinstimmungen von Schülerleistungen und Lehrerurteilen (der Median der Korrelationen lag bei $r = .61$). Der Median der anderen vier Kategorien lag durchweg höher (2: $r = .76$, 3: $r = .70$, 4: $r = .67$, 5: $r = .70$).

Relative Unabhängigkeit der Komponenten voneinander

Nicht zuletzt sind bei der Erfassung verschiedener Komponenten der diagnostischen Kompetenz jeweils methodische Besonderheiten zu berücksichtigen. Für die Rangkomponente werden die Lehrer beispielsweise gebeten, ihre Schüler nach Leistung (in einem bestimmten Bereich oder hinsichtlich ihrer Ergebnisse in einem konkreten Test) zu sortieren. Die Korrelation zwischen geschätzter und tatsächlicher Rangreihe ist dann der entsprechende Indikator. Analog werden für die Bestimmung der Niveaukomponente die Lehrkräfte gebeten einzuschätzen, wie viele (bzw. welche) Aufgaben eines Tests einzelne Schüler korrekt zu lösen in der Lage sind. Und für die Berechnung der Streuungskomponente ist zu fragen, in welchem Ausmaß sich die Leistungen einzelner Schüler voneinander unterscheiden. In der Forschungspraxis ist dieses Vorgehen nicht oft zu finden, denn in gewisser Weise gehen die verschiedenen Komponenten aus einander hervor. So ergibt sich aus der Frage zur Niveaueinschätzung quasi automatisch auch gleichzeitig das Maß für die Streuungseinschätzung, so dass separate Fragen dazu überflüssig erscheinen. Auch die Rangkomponente kann aus der Niveaueinschätzung abgeleitet werden. Allerdings scheint diese Ableitung nur

von der Niveaukomponente auf die Rang- und Streuungskomponente zu funktionieren, nicht umgekehrt, denn aus einer Angabe zur Rangreihe von Leistungen in einer Klasse kann allenfalls auf Umwegen auf das Leistungsniveau oder die Leistungsabstände zwischen den Schülern geschlossen werden.

Bereits vor über 20 Jahren konnte gezeigt werden, dass die verschiedenen Komponenten diagnostischer Kompetenz durchaus unterschiedlich ausfallen, ganz unterschiedliche Sachverhalte abbilden und nicht oder nur bedingt miteinander zusammenhängen (s. Schrader & Helmke, 1987). Auch in nachfolgenden Untersuchungen bestätigte sich immer wieder die Annahme geringer Korrelationen zwischen den Urteilskomponenten, besonders zwischen der Rang- und den anderen beiden Komponenten (Anders et al., 2010; Schrader, 1989; Spinath, 2005; Weinert & Schrader, 1986).

3.3 Maßstäbe diagnostischer Urteile (Bezugsnormen)

Insbesondere für Schulleistungsbeurteilungen existieren über die bereits beschriebenen Differenzierungen hinaus drei grundlegende Vergleichsmaßstäbe (vgl. Rheinberg, 2001; Sacher, 2009), die Lehrkräfte bei Leistungsbeurteilungen als Referenz bzw. Bezugsnorm mit unterschiedlichen Akzentuierungen verwenden.

Soziale Bezugsnorm

Wird die Leistung von einzelnen Schülern oder ganzen Klassen mit anderen Schülern oder Klassen verglichen, spricht man von einem sozialen oder normorientierten Vergleich. Hierbei hängt die Qualität einer Leistung davon ab, wo sie sich im Vergleich zu anderen Leistungen (z.B. denen der Mitschüler) befinden. Zu dieser Art Einschätzung sind Lehrer im Allgemeinen recht gut in der Lage. Es ist für derartige Einschätzungen allerdings immer ein Vergleichsmaßstab nötig; im Falle der Beurteilung individueller Leistungen ist die Referenz für gewöhnlich die gesamte Klasse, die der Lehrer kennt. Einschränkungen verteilungsorientierter Vergleiche treten dann auf, wenn Lehrern diese Referenz fehlt. Den mittleren Leistungsstand der eigenen Klasse festzustellen gelingt nur dann, wenn man eine Vorstellung von den Leistungen anderer Klassen hat (was meist nicht der Fall ist). Zudem entstehen Fairness- und Vergleichbarkeitsprobleme, wenn Lehrer Schülerleistungen ausschließlich am klasseninternen Bezugssystem bewerten. Je nach Leistungsstand der Mitschüler können somit gleiche Leistungen von Schü-

lern aus unterschiedlichen Klassen mal gut und mal schlecht beurteilt werden, was die Aussagekraft von Schulnoten extrem einschränkt (vgl. Ingenkamp, 1995b). Auch Jahrzehnte nach Ingenkamps Kritik daran hat sich kaum etwas an diesem Problem geändert, wie u.a. Daten aus TIMSS (Baumert et al., 2000) zeigen. Rheinberg (1980, 2001) skizziert darüber hinaus zwei weitere Einschränkungen der sozialen Bezugsnorm: Zum einen verdeckt sie bei gleichbleibender Leistungsrangordnung den gemeinsamen Leistungszuwachs einer gesamten Klasse, so dass Schüler den Eindruck gewinnen könnten, sich überhaupt nicht zu entwickeln. Zum anderen bleiben selbst deutliche Leistungssteigerungen unerkannt, wenn sie in leistungsheterogenen Klassen nicht zu einem „Überholen" anderer Schüler führen (Rheinberg, 1982). Beide letztgenannten Faktoren wirken sich ungünstig auf die Lern- und Leistungsmotivation der Schüler aus.

Kriteriale Bezugsnorm

Deutlich objektiver als verteilungsorientierte Vergleiche erscheinen daher als zweiter Maßstab die kriterialen Vergleiche. Sie orientieren sich an einem absoluten Maßstab und unterliegen daher - auch klassenübergreifend - einem festen Kriterium, an dem die Leistung gemessen wird. Beispielhaft ist hier das Erreichen einer Mindestpunktzahl zum Bestehen des Abiturs oder die Zuordnung zu einer bestimmten Kompetenzstufe in einem großen Test wie PISA zu nennen. Dabei spielt es keine Rolle, ob man besser ist als andere (soziale Bezugsnorm), sondern lediglich, ob die eigene Leistung einen bestimmten Standard erfüllt. Die sogenannten Bildungsstandards kommen der kriterialen Bezugsnorm nahe, indem sie bestimmte Fähigkeiten und Fertigkeiten (bzw. Kompetenzen) definieren, die alle Schüler erreichen sollten. Versuche, Zensuren über diese Kriterien zu bestimmen, sind jedoch kaum realistisch, da für jeden Inhaltsbereich in jedem Fach, für jede Schulform und jede Klassenstufe definiert werden müsste, welche Leistung einem „sehr gut" oder einem „ausreichend" entspricht. Dies wäre nicht nur ein nicht zu bewältigender administrativer Aufwand, dessen Tücken im Detail steckten, sondern würde Lehrer auch enorm in ihrer didaktischen Flexibilität einschränken.

Individuelle Bezugsnorm

Unter der individuellen oder auch ipsativen Bezugsnorm wird der Vergleich einer aktuellen Leistung mit einer früheren Ausgangsleistung verstanden. Die aktuelle Leistung wird dabei daran bemessen, was der Schüler im betref-

fenden Leistungsbereich zuvor erreicht hat. Relativ unabhängig vom tatsächlichen Leistungsniveau drückt sich dabei eine Leistungssteigerung in einer guten, eine Leistungsverschlechterung in einer schlechten Note aus. Schüler haben damit die Möglichkeit, den Ertrag ihrer Lernbemühungen direkt zu erkennen, allerdings fehlen ihnen Informationen zu ihrem Leistungsniveau im Vergleich zu Gleichaltrigen. Die individuelle Bezugsnorm erwies sich vor allem für leistungsschwächere Schüler als motivierend, sie tritt in der Realität jedoch überwiegend nicht allein, sondern in Kombination mit anderen Bezugsnormen auf (Rheinberg, 2006), da sie allein genommen allenfalls eine Trendaussage darstellt. Bei Leistungsbeurteilungen, die beispielsweise Berechtigungen für einen bestimmten Ausbildungsweg darstellen, tritt die Bedeutung der individuellen Entwicklung weit hinter jener des absoluten Leistungsstands zurück.

Wie Rheinberg (2001) betont, ist die Beschränkung von Lehrern auf die Verwendung einer einzigen Bezugsnorm nicht sinnvoll. Vielmehr sollten je nach Beurteilungssituation die geeigneten (Kombinationen aus) Bezugsnormen angewandt werden, da sie jeweils verschiedene Vor- und Nachteile mit sich bringen. So scheint die soziale Bezugsnorm beispielsweise zwar objektiv und konsistent zu sein, weil ein direkter Vergleich zu den Leistungen der Mitschüler möglich ist. Gleichzeitig macht sie den Schülern aber auch klar, dass sie nur dann gute Leistungen erzielen können, wenn sie besser sind als andere, was sich als besonders ungünstig für leistungsschwache Schüler herausgestellt hat. Im Gegenzug kann die Einbeziehung von individuellen Bezugsnormen bei der Leistungsbewertung deutliche Motivationseffekte mit sich bringen. In mehreren Studien konnte belegt werden, dass die Bezugsnormorientierung der Lehrer von großer Bedeutung für die Leistung und die Lernfreude von Schülern sowie für das Wohlbefinden von Schülern und Lehrern ist (vgl. Heckhausen, 1989; Jerusalem & Mittag, 1999; Martinek, 2007). In der Praxis zeigt sich tendenziell eine Bezugsnormvielfalt in dem Sinne, dass Lehrer in verschiedenen Situationen jeweils unterschiedlich große Anteile der Bezugsnormen in ihre Bewertungen einfließen lassen. Diese Vielfalt trägt ihren Teil dazu bei, dass Schulnoten von Schülern verschiedener Klassen nur bedingt miteinander vergleichbar sind.

3.4 Gegenstände und Analyseebenen

Für die Unterrichtspraxis und -forschung lassen sich weiterhin die nachfolgend genannten Dimensionen diagnostischer Urteile unterscheiden, die jeweils unterschiedliche Fokusse mit verschiedener Aussagekraft haben. Sie

sind eher von einem theoretischen Standpunkt aus interessant und liefern eine Möglichkeit der Kategorisierung verschiedener Urteilsformen, die den Lehrern im Alltag vermutlich nicht unmittelbar bewusst ist.

Explizite vs. implizite Urteile

Schrader und Helmke (2001) unterscheiden zwei Arten der Leistungsbewertung, nämlich explizite und implizite Urteile. Unter expliziten Urteilen verstehen sie Beurteilungen von ‚Daten', die speziell zum Zweck der Beurteilung ‚erhoben' wurden (z.B. Klassenarbeiten). Das diagnostische Urteil kommt dabei dadurch zustande, dass die gewonnenen Informationen mit einer Norm (s. z.B. Ingenkamp, 2005, und vorheriges Kapitel) verglichen werden. Die Beurteilung erfolgt dabei in Situationen, in denen die Lehrkraft ihre Aufmerksamkeit gezielt (wie z.B. in mündlichen Prüfungen) und im Idealfall ungeteilt (wie bei der häuslichen Korrektur schriftlicher Arbeiten) auf die Diagnose richten kann. Dies ermöglicht eine gründliche Reflexion des Urteilsvorgangs und die bewusste Weiterverwendung der gewonnenen Informationen.

Im Gegensatz zu diesen expliziten Urteilen laufen implizite Urteile meist stark verkürzt ab. Schülerleistungen werden dabei nur insoweit registriert und intuitiv eingeschätzt, wie es für Entscheidungen zu Unterrichtsprozessen (Themenwahl, Abschluss von Unterrichtseinheiten, Geben von Hilfestellungen, Auswahl von Materialien etc.) nötig ist. Derartige „Mikrodiagnosen" (Schrader & Helmke, 2001) kommen dadurch zustande, dass die Erwartungen des Lehrers an die Klasse oder an einzelne Schüler mit aktuellen Beobachtungen abgeglichen und verknüpft werden. Da diese Einschätzungen fortwährend im Unterricht getroffen werden müssen, laufen sie sehr schnell, meist unreflektiert oder sogar unbewusst ab und können nur selten überhaupt verbalisiert werden.

Leistungsdiagnostik vs. Lernprozessdiagnostik

Zentral für den Schulalltag ist die Unterscheidung danach, ob der Lernerfolg nach Abschluss einer Lernphase oder der Lernprozess selbst diagnostiziert werden soll (Scholz, 1993). Während die erste Variante als Statuserfassung die Frage danach, was wie gut gelernt wurde, in den Fokus stellt, interessiert bei der zweiten Variante, wie etwas gelernt wird. In der Bildungsforschung wird hauptsächlich die Leistungsdiagnostik erforscht und behandelt (Schrader & Helmke, 2005), und sie ist auch in der Schule maßgeblich, wenn es beispielsweise um die Notengebung oder Übergangsentscheidun-

gen geht (Scholz, 1993). Statusorientierte Diagnostik wird jedoch von vielen als für Förderanliegen unbrauchbar angesehen (Winter, 2006). Insbesondere für den Unterricht ist aber auch die Lernprozessdiagnostik (oder einfach Prozessdiagnostik) von hoher Relevanz, die Einblicke in verschiedene Lösungswege sucht, um die Korrektur falscher Lernschritte bemüht ist und somit eine Basis für gezielte Unterstützung bei Defiziten, die Planung von Lernschritten oder Aussagen zum weiteren Lernverlauf bietet.

Punktuell vs. kumulativ

Hierbei wird unterschieden zwischen Beurteilungen, die sich auf punktuelle Leistungen, z.B. in einem Test, beziehen, und zwischen Beurteilungen, die die über einen gewissen Zeitraum hinweg erbrachten Leistungen in einer Note zusammenfassen.

Global vs. spezifisch

Globale Beurteilungen beziehen sich auf einen weiten Inhaltsbereich, der nicht weiter differenziert wird. Verbalurteile über die zurückliegenden allgemeinen sprachlichen Leistungen können ein Beispiel hierfür sein. Spezifisch könnte dieses Urteil jedoch aufgegliedert werden u.a. in die Bereiche Rechtschreiben, Lesen, Grammatik, wobei auch hierfür noch feinere Differenzierungen möglich sind. In der Forschung zur diagnostischen Kompetenz liegen spezifische Urteile beispielsweise dann vor, wenn sie sich Lehrerurteile auf das Lösungsverhalten bei bestimmten, den Lehrern bekannten Aufgaben beziehen, wohingegen globale Urteile die generellen Fähigkeiten betreffen. Für spezifische Urteile ist somit einerseits zwar ebenso die (korrekte) Einschätzung der Schwierigkeit der zugrunde liegenden Aufgaben notwendig, andererseits zeigen diverse Forschungsergebnisse, dass spezifische Urteile in aller Regel etwas genauer ausfallen als globale (vgl. Kapitel 4 ab S. 52).

Kognitive vs. nicht-kognitive Merkmale

Neben der zentralen Beurteilung schulischer Leistungen sowie weiterer kognitiver, leistungsnaher Merkmale wie Intelligenz oder Begabung (Wild, 1991) ist gelegentlich auch die Einschätzkompetenz der Lehrer bezüglich affektiver, emotionaler und motivationaler Schülermerkmale im Blickpunkt der Forschung. Letztere kommt beispielsweise bei der Vergabe von Kopfnoten zum Einsatz, ist aber auch für die Unterrichtsgestaltung oder das gezielte Eingehen auf die Bedürfnisse einzelner Schüler eine wichtige Kompetenz.

Formelle vs. informelle Diagnostik

Weiterhin ist die Unterscheidung nach dem Formalisierungsgrad, genauer zwischen formeller und informeller Diagnostik, von großer Bedeutung. Die formelle Diagnostik beruht darauf, dass Lehrkräfte mit Hilfe wissenschaftlich erprobter Methoden gezielt spezifische Fragen klären können (vgl. z.B. überblicksartig Lukesch, 1998). Dazu bedarf es von Seiten der Lehrer fundierter Kenntnisse über die Methoden, die oft nicht leicht anzuwenden sind, und von Seiten der Schulen wären Rahmenbedingungen nötig, die den Einsatz dieser Methoden systematisch ermöglichen. Beides ist mit dem schulischen Alltag nicht leicht vereinbar. Dahingegen ist die informelle Diagnostik, die eher auf intuitiven Einschätzungen während des Unterrichts beruht und den Lehrpersonen mitunter wenig oder gar nicht bewusst ist, deutlich leichter in die Praxis zu integrieren (Wahl, Weinert & Huber, 1997). Ihr großer Nachteil besteht darin, dass die informelle Diagnostik eher auf Grundlage von Routinen und daher oft unsystematisch und unreflektiert erfolgt, wodurch ihre Anfälligkeit für Urteilsfehler (vgl. Kapitel 3.5) enorm verstärkt wird.

Hascher (2005, 2008) führt als weiteren, dazwischen liegenden Formalisierungsgrad die semiformelle Diagnostik an, da ihr die zweistufige Differenzierung für den Schulalltag zu kurz greift. Sie meint damit Urteile, die zwar nicht den Kriterien der formellen Diagnostik genügen, aber dennoch auf Grundlage gezielter Beobachtungen und dadurch absichtsvoll und motiviert getroffen werden. Somit wäre es auch Lehrern, die keine Experten im Einsatz wissenschaftlicher Diagnosemethoden sind, möglich, bewusste diagnostische Entscheidungen zu treffen.

Analyseeinheit

Die zu beurteilende Einheit ist im Unterrichtsalltag der einzelne Schüler, denn Leistungen werden individuell erbracht und bewertet. Vor dem Hintergrund der Einschätzung über das Erreichen von Lernzielen oder die Entscheidung für das Beginnen neuen Unterrichtsstoffes ist nicht selten aber auch die Einschätzung der Klassenleistung gefragt. Urteile über die mittlere Leistungsfähigkeit einer gesamten Schule sind ebenso denkbar. In Fragebogeninstrumenten zur diagnostischen Kompetenz begegnen einem daher auch Fragen zu Klassenleistungen im Vergleich zur Leistung einer durchschnittlichen Klasse. Lehrer müssen hier die Anforderungen meistern, zum einen aus den individuellen Leistungen ihrer Schüler einen mittleren Leistungseindruck abzuleiten, zum anderen diesen dann ins Verhältnis zum nur

schwer zu erahnenden Durchschnitt zu setzen. Aus Erfahrung und Wissen muss der Lehrer hierbei eine Ahnung vom „Durchschnitt“ haben, der sich auf empirisch ermittelte Daten stützt. Neben der Einschätzung von Leistungen einzelner Schüler oder Gruppen von Schülern ist als zweiter zentraler Punkt die Einschätzung von Aufgaben und Aufgabenschwierigkeiten von Belang. Dies ist nicht nur wichtig für die dem Leistungsvermögen der Schüler angemessene Auswahl von zu bearbeitenden Aufgaben, sondern nicht zuletzt hängt auch jede Leistungseinschätzung davon ab, dass das Niveau der zugrunde liegenden Aufgaben korrekt eingeschätzt wurde. Die Schülerleistungen hängen immer auch vom Schwierigkeitsgrad der bearbeiteten Aufgaben ab, daher muss beides gleichermaßen bei Leistungseinschätzungen berücksichtigt werden.

Natürlich ist der Gebrauch der oben genannten Dimensionen keine Entweder-Oder-Frage. In der Unterrichtspraxis treten sie in vielfältigen Kombinationen auf. Für Lehrer ist es jedoch wichtig, sich dieser grundlegenden Unterscheidungen bewusst zu sein, um jede Leistung mit dem richtigen Maßstab zu beurteilen.

Bedeutsam ist außerdem die Frage, ob Lehrerurteile rückblickend, zeitgleich oder vorhersagend gemeint sind. Alle Formen sind möglich. Im Schulalltag beziehen sich Zeugnisnoten beispielsweise auf die Leistungen innerhalb des vergangenen Schulhalbjahres, die Note für eine mündliche Leistungskontrolle auf die aktuelle Leistung und die Planung von Unterricht auf die nächsten Wochen. Auch in Studien zur diagnostischen Kompetenz gibt es diese Unterscheidungen. Voraussagen von Lehrern wurden oft im Umfeld von Lehrererwartungen und somit auch im Zusammenhang mit Forschung zu selbsterfüllenden Prophezeiungen untersucht. Dabei wurde davon ausgegangen, dass Lehrererwartungen an die Leistung der Schüler auch tatsächlich zu einer entsprechenden Leistungsveränderung führen. Ältere Schätzungen, nach denen zwischen 5 und 10 Prozent der Schüler von Erwartungseffekten betroffen sind (Brophy, 1983; Rosenthal, 1984), ließen sich auch in neueren Arbeiten bestätigen. Es lässt sich inhaltlich aber nicht immer exakt trennen, ob es sich tatsächlich um Erwartungseffekte handelt, wenn prognostizierte Leistungen tatsächlich eintreten, denn schließlich kann es auch sein, dass der Lehrer mit seiner Einschätzung über die Leistungsentwicklung einfach richtig lag, ohne dass sich sein Urteil auf die Leistung selbst auswirkt. Jussim und Harber schreiben dazu: „Prediction without causation is exactly how we define accuracy.” (Jussim & Harber, 2005).

3.5 Urteilsfehler bei der Leistungsmessung und Leistungsbewertung

Die Wahrnehmung von Schülerleistungen wird von Erwartungen, Voreinstellungen und Hypothesen der Lehrer beeinflusst, was auch unter dem Begriff der Hypothesentheorie der Wahrnehmung bekannt ist (Schrader & Helmke, 2001, vgl. auch Kapitel 1.3). Während Laien häufig der Meinung sind, die Wahrnehmung liefere eine getreue Abbildung der Realität, können Menschen mit ihrer begrenzten Aufnahme- und Verarbeitungskapazität tatsächlich immer nur einen kleinen Ausschnitt der Realität wahrnehmen. Dabei wird der Fokus der Wahrnehmung zusätzlich durch die jeweils individuellen Bedürfnisse, Interessen, Einstellungen, Werthaltungen oder Motive beeinflusst und gelenkt (Hofer, 1986), was im Idealfall zu einer gezielteren Beobachtung führt, in vielen Fällen aber auch zu oberflächlicher oder verzerrter Wahrnehmung. Welche dieser Folgen eintritt, hängt nicht zuletzt davon ab, wie zutreffend die Erwartungen sind und wie reflektiert der Wahrnehmende mit ihnen umgeht. Aus der Sozial- und Wahrnehmungspsychologie sind eine ganze Reihe von Effekten bekannt, die die Wahrnehmung beeinflussen: neben Selektionen (eigene Bedürfnisse und Einstellungen bestimmen die Sensibilität, mit der auf bestimmte Reize reagiert wird), Akzentuierungen (die Tendenz, mit eigenen Bedürfnissen im Zusammenhang stehende Dinge als wichtiger als anderes zu empfinden), Fixierungen (wiederholte bestimmte Interpretationsweise bei ähnlichen Sachverhalten) und der bedürfnisabhängigen Organisation (mehrdeutige Reize werden entsprechend der eigenen Bedürfnisse wahrgenommen) (vgl. Kebeck, 1994) sind in der Literatur zu Lehrerurteilen (u.a. Ingenkamp, 2005; Jürgens, 2005) vor allem jene Urteilsfehler beschrieben, über die im Folgenden ein kurzer Überblick gegeben werden soll, da sie Erklärungspotential für Ursachen von ungenauen Urteilen aufweisen.

Erwartungseffekte

Wie der Name bereits vermuten lässt, entstehen Erwartungseffekte dadurch, dass Lehrer an ihre Urteile mit bestimmten Erwartungen herangehen. Diese Erwartungen basieren auf Vorerfahrungen und sind entsprechend der Hypothesentheorie der sozialen Wahrnehmung (Bruner & Postman, 1951) abhängig von den eigenen Werten und Motiven mehr oder weniger stark ausgeprägt. Der Grundgedanke dieser Theorie ist, dass die Wahrnehmung schon beginnt, bevor konkrete Informationen vorhanden sind. Liegen die Informationen dann vor, können daraus diejenigen selegiert werden, die am

ehesten zu den Vorannahmen passen. Dies kann zum Beispiel dazu führen, dass bei schwachen Schülern, bei denen eine höhere Fehlerhäufigkeit vermutet wird, in Leistungskontrollen durch eine kritischere Einstellung auch tatsächlich Fehler gefunden werden, die bei guten Schülern möglicherweise übersehen worden wären, oder dass in uneindeutigen Fällen einem schwachen Schüler eher ein Falsch, einem guten Schüler in derselben Situation eher ein Richtig attestiert wird. Dieses Verhalten führt zweifelsohne zu einer Benachteiligung von leistungsschwachen Schülern. Gleichermaßen können auch Vorinformationen die eigenen Erwartungen an die Schülerleistungen beeinflussen. Während unter anderem für Vorinformationen über bisherige Leistungen kein Einfluss auf die Leistungsbewertung gefunden werden konnte, beeinflussen Informationen zum Beispiel über die häuslichen Verhältnisse des zu beurteilenden Schülers die Noten maßgeblich (Baurmann, 1995; Ingenkamp, 1989).

Die langfristigen Auswirkungen der Erwartungen von Lehrkräften auf die Leistungen ihrer Schüler sind intensiv untersucht worden und unter dem Begriff des Pygmalion-Effekts (erstmals bei Rosenthal & Jacobson, 1971, beschrieben) bekannt. Darunter wird verstanden, „dass die Erwartungen eines Lehrers sein Unterrichtsverhalten derart beeinflussen, dass Schülerleistungen nach einiger Zeit so ausfallen, wie er es erwartet - auch dann, wenn die Lehrererwartung i.d.S. unangemessen war, als der Schüler ohne diese Lehrererwartung andere Leistungen gezeigt hätte. Damit könnten Erwartungen des Lehrers Vorhersagen sein, die die Kraft haben, sich selbst zu erfüllen“ (Rheinberg, 1980). Diese sich selbst erfüllenden Prophezeiungen (‚self-fulfilling prophecies‘) treten jedoch anscheinend nicht generell, sondern nach Heckhausen (1974) nur unter folgenden Bedingungen auf: 1) der Schüler leistet weniger, als es nach seinen Fähigkeiten möglich wäre (underachievement), 2) der Lehrer unterschätzte bislang die Fähigkeiten des Schülers und machte ihm dies auch deutlich, und 3) der Schüler hat diese Einschätzungen des Lehrers auch internalisiert (vgl. auch Rheinberg, Bromme, Minsel, Winteler & Weidenmann, 2001, S. 311). In einer kritischen Betrachtung von über drei Jahrzehnten an Forschung zu sich selbst erfüllenden Prophezeiungen fassen Jussim und Harber (2005) die bisherigen Erkenntnisse so zusammen, dass Pygmalion-Effekte im Unterricht zwar existieren, jedoch nur in sehr geringem Umfang und bevorzugt bei Kindern aus benachteiligten Sozialgruppen und dass sie sich darüber hinaus über die Zeit oder in Bezug auf den Wahrnehmenden eher verringern als vergrößern. Ob sie sich überhaupt negativ z.B. auf die Leistung auswirken, wird von den Autoren angezweifelt. Sie ziehen eher in Betracht, dass Lehrerurteile die Schü-

lerleistungen deshalb gut vorhersagen, weil sie korrekt sind, und nicht, weil sie zu einer Selbsterfüllung führen.

Projektionsfehler

Wenn Lehrer dazu neigen, eigene Eigenschaften, Merkmale oder Wünsche auf die Schüler zu übertragen oder in ihnen wiederzufinden glauben, können Projektionsfehler auftreten. Dies kann in zweierlei Richtungen erfolgen (Kleber, 1992): Entweder ist der Lehrer bestrebt, sich selbst als deutlich besser als den Schüler zu empfinden (Kontrastfehler), oder er meint ihm Schüler ähnliche Eigenschaften zu sehen wie er selber besitzt (Ähnlichkeitsfehler). Derart verfälschte Wahrnehmungen können sich leicht auf Leistungsbeurteilungen auswirken.

Halo-Effekt

Unter der Bezeichnung Halo-Effekt (oder auch Hofeffekt) wird die Verzerrung von Urteilen aufgrund eines globalen Gesamteindrucks verstanden. Dieses Phänomen wurde bereits vor neunzig Jahren von dem Psychologen Edward Thorndike beschrieben, der feststellte, dass in mehreren Untersuchungen zur Personenbeurteilung einzelne Merkmale, die objektiv nicht miteinander in Zusammenhang standen, hoch korrelierten (Thorndike, 1920). Dabei scheint ein einzelnes Merkmal für Beobachter so bedeutsam zu sein, dass es wie ein Heiligenschein (halo) andere Personenmerkmale überstrahlt. Beispielsweise könnten im schulischen Kontext leistungsunabhängige Merkmale wie die Kleidung der Schüler, ihre Art des Auftretens, ihre Disziplin oder ihr Sprachverhalten die Leistungsbewertung beeinflussen. Insbesondere dann, wenn Leistungen zu bewerten sind, die nicht direkt zu beobachten oder nicht eindeutig definiert sind, kann somit z.B. ein unordentliches Auftreten des Schülers auch die Bewertung seiner Leistung negativ beeinflussen (Sacher, 1996). Wenn affektive Komponenten beteiligt sind (z.B. Sympathie oder Antipathie zwischen Lehrer und Schüler), kann dies die Auftretenswahrscheinlichkeit des Halo-Effekts vergrößern.

Logischer Fehler (theoretischer Fehler)

Nicht immer eindeutig vom Halo-Effekt abzugrenzen ist der logische Fehler in der Beurteilung. Dieser bezeichnet Situationen, in denen aus dem Auftreten eines Merkmals (z.B. großer Wortschatz eines Schülers) auf das Vorliegen eines anderen Merkmals (z.B. hohe Rechtschreibkompetenz) geschlossen wird. In Wirklichkeit müssen diese Eigenschaften keineswegs zusam-

menhängen, doch oftmals bilden Lehrer ein implizites Personenbild von ihren Schülern, das als Erwartungshintergrund in den Urteilsprozess eingreifen kann. Gerade dann, wenn viele Schüler hinsichtlich mehrerer Merkmale oder nicht genau beobachtbarer Eigenschaften eingeschätzt werden sollen, kann es zu logischen Fehlern kommen (Jürgens, 2005).

Attributionsfehler

Bei der (Kausal-)Attribution, einem auf Heider (1958) zurückgehenden Forschungsfeld, das sich mit dem psychologischen Phänomen der Ursachenzuschreibung beschäftigt, spielt neben der objektiven Leistung auch die Meinung des Lehrers über deren Zustandekommen eine große Rolle. Diese einfache Ursachenzuschreibung kann ebenfalls verzerrt sein. Wie Rieder (1990) ausführt, ist es möglich, dass Lehrer gleiche Leistungen besser oder schlechter bewerten, je nachdem, ob sie ihr Zustandekommen mehr auf Anstrengung oder größere Fähigkeiten zurückführen, wobei die Vermutung von großer Anstrengung zu einer vergleichsweise besseren Beurteilung führt. Dieser Effekt kann wiederum zusätzlich durch bestimmte Einstellungen oder Werthaltungen der Lehrer beeinflusst sein.

Perseverationstendenzen

Fällt ein Lehrer ein Urteil über einen Schüler, so ist die Wahrscheinlichkeit hoch, dass seine nächsten Urteile zum selben Schüler nicht stark vom ersten abweichen. Diese Wirkung des ersten Urteils auf die folgenden, meist auf die Leistungen innerhalb eines Schulfaches bezogen, wird als Perseverationstendenz bezeichnet (Rieder, 1990).

Reihenfolgeeffekte

Insbesondere bei der Beurteilung schriftlicher Arbeiten ist der Effekt bekannt, dass Lehrer die ersten Arbeiten strenger beurteilen als die letzten (Baurmann, 1995). Hinzu kommt eine Wechselwirkung mit der Höhe der Leistung, indem z.B. eine schlechte Leistung als noch schlechter bewertet wird, als sie eigentlich ist, wenn sie auf eine sehr gute Leistung folgt.

Kontexteffekte

Möller und Köller (1997) bezeichnen die Einwirkung eines Umgebungsreizes auf die Wahrnehmung des Leistungsverhaltens der Schüler als Kontexteffekt und unterscheiden dabei zwei Tendenzen. Auf der einen Seite stehen

Assimilationseffekte, wenn beim selben Schüler beispielsweise die Einschätzung von Deutschleistungen der Einschätzung von Mathematikleistungen (in diesem Fall der Kontext) angeglichen wird. Auf der anderen Seite handelt es sich um Kontrasteffekte, wenn - um beim Beispiel zu bleiben - die Deutschleistung möglichst unterschiedlich zur Mathematikleistung beurteilt wird. Kontrasteffekte können aber auch, ähnlich wie Reihenfolgeeffekte, in Bezug auf ein zuvor abgegebenes Urteil auftreten, das das aktuelle Urteil beeinflusst.

Neben den bis hierhin aufgelisteten Urteilsfehlern, die bei der Messung von Leistungen auftreten können, gibt es auch Fehler, die bei der Zuordnung von Leistungen zu einem Maßstab entstehen können (Jürgens, 2005).

Strenge- und Mildefehler

Liegen Strengefehler vor, neigen Lehrer dazu, kleine Mängel in Leistungen relativ stark zu gewichten, während sie gute Leistungen als weniger gut einstufen. Im Gegensatz dazu entstehen Mildefehler dadurch, dass Lehrer dazu tendieren, schlechte Leistungen als weniger schlecht und gute als besonders gut zu gewichten. Mögliche Erklärungsansätze für diese Tendenzen liefern u.a. Sacher (1996), der entsprechende Lehrertypen (Strengbeurteiler vs. Mildbeurteiler) vermutet, oder Kleber (1976), der die Ursachen in interaktionsbedingten Urteilsreaktionen sieht, die u.a. zu einer Besserbewertung von den Lehrern gut bekannten Schülern führt. Denkbar ist darüber hinaus ebenso, dass der Strengefehler besonders dann eintritt, wenn Experten Leistungen aus ihrem Fachgebiet beurteilen sollen.

Tendenz zur Mitte

Vermeiden Lehrer extreme Urteile und wählen - besonders bei auf einer (Noten-)Skala einzuschätzenden Leistungen - eher Urteile aus dem mittleren (neutralen) Bereich, dann spricht man von der Tendenz zur Mitte bzw. einer zentralen Tendenz. Dies mag zum Beispiel an einer Entscheidungsunlust der Lehrer liegen oder daran, dass solcherlei Urteile weniger Gefahr laufen anzuecken (vgl. Sacher, 1996). Auch wählen Lehrer möglicherweise dann mittlere Urteile, wenn sie sich nicht vorschnell festlegen wollen.

Tendenz zu Extremurteilen

Die gegenteilige Tendenz ist die zu Neigung zu Extremurteilen. Ist bei Schülerleistungen eine gewisse Qualitätsschwelle überschritten, wird sie gleich

als besonders gut wahrgenommen, während beim Auftreten einer bestimmten Anzahl von Fehlern gute Aspekte der Leistung leichter ausgeblendet und somit die gesamte Arbeit als schlechter wahrgenommen wird als sie tatsächlich ist. Gerade Lehrer, die leicht zu begeistern oder zu enttäuschen sind, neigen möglicherweise zu dieser Art Urteilsverzerrung.

Diese knappe Darstellung von möglichen Urteilsfehlern dient hauptsächlich dazu, die Mannigfaltigkeit verschiedener negativer Einflüsse auf die Urteilsgenauigkeit zu illustrieren. Wie im Verlauf dieser Arbeit noch gezeigt werden wird, fällen Lehrkräfte vereinzelt sehr unzutreffende Urteile. Dabei besteht der Anspruch für die Arbeit nicht darin, diese stark von den Testergebnissen abweichenden Lehrerurteile auf verschiedene Urteilsfehler zurückzuführen. Vielmehr dient der vorangegangene Überblick dazu, sich mögliche Ursachen für Fehlurteile bewusst zu machen.

Während einige der aufgeführten Urteilsfehler an die Beurteilungssituation gebunden sind, werden andere maßgeblich von der Lehrer-Schüler-Interaktion beeinflusst. Zudem können einige der genannten Urteilsfehler durchaus in Abhängigkeit vom zu beurteilenden Schüler mal stärker und mal schwächer (oder gar nicht) ausgeprägt sein, andere hingegen treten eher schülerunabhängig auf und sind auf Eigenschaften des urteilenden Lehrers zurückzuführen. Das Erkennen der Anfälligkeit der eigenen Wahrnehmungen für Verzerrungen ist nicht einfach und die Trennung von Persönlichkeit und Fachleistung von Schülern keine Selbstverständlichkeit (vgl. Rieder, 1990). Umso wichtiger erscheint es für Lehrkräfte, dass sie ihre Motive, Einstellungen und Werthaltungen kontinuierlich hinterfragen und ihre Urteile regelmäßig und systematisch selbst überprüfen.

4 Stand der Forschung zur diagnostischen Kompetenz

In diesem Kapitel werden bedeutsame bisherige Forschungsergebnisse zu verschiedenen Aspekten der diagnostischen Kompetenz zusammengefasst. Bei einem Blick über die bisherige Forschung stellt man schnell fest, dass die bereits vorgestellten Facetten und Dimensionen der diagnostischen Kompetenz sehr ungleichmäßig stark untersucht wurden. Der Großteil der Studien widmete sich der auf kognitive Schülermerkmale (also Leistungen) bezogenen Diagnosegenauigkeit und ließ nicht-kognitive Maße unbeachtet. Die nächsten Kapitel behandeln deshalb diese beiden Facetten getrennt voneinander.

4.1 Urteilsgüte in kognitiven Bereichen

Die Güte von Urteilen ist kein Maß, für das es eine feststehende Berechnungsgrundlage gibt, sondern sie kann je nach Operationalisierung ganz verschiedene Aspekte der Urteilsgenauigkeit ausdrücken. Wie in Kapitel 3.2 (ab S. 34) dargestellt, wird die diagnostische Kompetenz entsprechend je nach Art des Bezugsmaßes verschiedenen Komponenten zugeordnet. Daher werden nachfolgend bisherige Forschungsergebnisse getrennt für die Rang- und die Niveaukomponente vorgestellt.

Rangurteile zu Schülerleistungen

Am häufigsten widmete sich die Forschung zur diagnostischen Kompetenz bis heute der auf Schülerleistungen bezogenen Rangkomponente. Bereits Ende der 1980er Jahre waren die Befunde dazu so vielfältig, dass Hoge und Coladarci (1989) sie in einer Metaanalyse betrachteten. Die Autoren fassen darin 16 vorwiegend US-amerikanische Untersuchungen zur diagnostischen Kompetenz von Lehrern zusammen, die den Zeitraum zwischen 1962 und 1988 abdecken. Obwohl in den berichteten Studien zum Teil unterschiedliche Designs eingesetzt wurden, zeigte sich davon unabhängig die generelle Tendenz, dass Lehrerurteile mit den korrespondierenden Schülerleistungen moderat bis stark korrelieren. Im Mittel beträgt der Zusammenhang zwischen Schülerleistungen und Lehrereinschätzungen $r = .66$, wobei zu berücksichtigen ist, dass in der überwiegenden Zahl der einbezogenen Untersuchungen die Urteilsgüte gepoolt über die Gesamtstichprobe und nicht klassenweise berechnet wurde. Um sowohl den jeweils zugrunde liegenden Leistungsbereichen als auch der Art der abgegebenen Urteile Rechnung zu tragen, sind in Tabelle 1 Range, Median und Mittelwert (nach Fisher-Z-

Transformation) der Korrelationen getrennt nach sprachlichem und mathematischem sowie nach spezifischer (‚direct') und globaler (‚indirect') Einschätzung wiedergegeben. Wenngleich die Unterschiede zwischen diesen Gruppen nicht statistisch signifikant sind, zeigt sich doch die Tendenz, dass eine spezifische Leistungseinschätzung, bei der Lehrer die Leistung in konkreten, ihnen bekannten Aufgaben einschätzen, sowohl für den sprachlichen als auch für den mathematischen Leistungsbereich genauer ausfallen als bei einer globalen Einschätzung, wo nur allgemeine Urteile zur Leistungsfähigkeit abgegeben werden sollten. Ebenso ist ersichtlich, dass die Güte der Einschätzungen im sprachlichen Bereich[3] - statistisch ebenfalls nicht signifikant - etwas höher ist als im mathematischen Bereich[4]. Dies ist in der Hauptsache auf Unterschiede bei den globalen Einschätzungen zurückzuführen, denn bei der spezifischen Einschätzung zeigen sich keine Unterschiede in Abhängigkeit vom fachlichen Bereich. Auffällig ist nicht zuletzt die insgesamt große Spannbreite von Korrelationskoeffizienten von insgesamt r = .28 (schwache Korrelation) bis r = .92 (sehr starke Korrelation), die zum einen sicher mit der eingesetzten Bandbreite unterschiedlicher Instrumentarien zu tun hat, andererseits aber auch einen Hinweis auf Unterschiede zwischen Lehrern gibt.

[3] Unter dem „sprachlichen Bereich" wurden für diese Aufstellung folgende Bereiche subsumiert: reading, reading recognition, reading comprehension, language arts, reading vocabulary und reading word analysis.

[4] Der „mathematische Bereich" bezieht sich auf math, math problem solving, math concepts und math comprehension.

Tabelle 1: Kennwerte der Übereinstimmung zwischen Lehrerurteilen und Schülerleistungen aus der Metaanalyse von Hoge und Coladarci (1989), getrennt nach Inhaltsbereich und nach der Art der Einschätzung sowie insgesamt

	sprachlicher Bereich	*mathematischer Bereich*	*insgesamt*
spezifische Einschätzung (‚direct')	N: 10	N: 5	N: 15
	Range: r = .48 bis .92	Range: r = .67 bis .77	Range: r = .48 bis .92
	Median: r = .69	Median: r = .70	Median: r = .70
	Mittelwert: r = .72	Mittelwert: r = .71	Mittelwert: r = .72
globale Einschätzung (‚indirect')	N: 24	N: 11	N: 35
	Range: r = .41 bis .86	Range: r = .28 bis .72	Range: r = .28 bis .86
	Median: r = .66	Median: r = .45	Median: r = .63
	Mittelwert: r = .67	Mittelwert: r = .51	Mittelwert: r = .63
insgesamt	N: 34	N: 16	N: 50
	Range: r = .41 bis .92	Range: r = .28 bis .77	Range: r = .28 bis .92
	Median: r = .67	Median: r = .63	Median: r = .67
	Mittelwert: r = .69	Mittelwert: r = .58	Mittelwert: r = .66

Anm.: Es wurden nur sprachliche und mathematische Bereiche aus der Metaanalyse berücksichtigt, vier Kennwerte zu „social studies" und „science" sowie ein berichteter Beta-Koeffizient aus dem Bereich Lesen wurden für die Gegenüberstellungen in dieser Tabelle nicht einbezogen.

Verschiedene Studien widmeten sich unmittelbar der Frage, ob spezifische oder globale Leistungsbeurteilungen zu akkurateren Ergebnissen führen. Demaray und Elliott (1998) ließen Lehrer Schülerleistungen sowohl global für alle Schüler ihrer Klasse (unterdurchschnittlich, durchschnittlich oder überdurchschnittlich) als auch bezogen auf konkrete Testitems für jeweils 4 ausgewählte Schüler der Klasse einschätzen und fanden heraus, dass beide Methoden zu einer hohen Übereinstimmung führten, wenngleich der Zusammenhang bei spezifischen Urteilen mit r = .84 noch etwas enger war als bei den globalen Urteilen (r = .70). Man kann hierbei jedoch nicht ausschließen, dass die zuerst abgegebene globale Beurteilung einen Übungseffekt für die späteren spezifischen Urteile hatte. Außerdem muss erwähnt werden, dass die prozentuale Übereinstimmung auf Itemebene auch zufällige Korrekteinschätzungen produzieren kann. Korrigiert man die Werte mittels Kappa-Koeffizient, werden die Übereinstimmungen deutlich geringer und bewegen sich im Bereich zwischen .60 und .70 (ebd.).

Zu ähnlichen Ergebnissen kamen auch Feinberg und Shapiro (2003), als sie Lehrer die Fähigkeiten beim fließenden Lesen von Schülern einschätzen

ließen. Grundlage waren die Schülerleistungen in einem curriculumsbasierten Test (Curriculum-based measurement [CBM]; Deno, 1985). Waren den Lehrern die von den Schülern zu bewältigenden Aufgaben bekannt, korrelierte ihr Urteil zu r = .70 mit den Leistungen, bei einer allgemeinen Einschätzung der Leseleistung lag der Zusammenhang bei r = .62. Beide Lehrerurteile korrelierten miteinander signifikant zu r = .66, also etwa in gleicher Höhe wie jede einzelne Einschätzung mit den Schülerwerten. Trotz geringfügig genauerer Einschätzung bei spezifischen Einschätzungen scheint die Art der Lehrerbefragung keine substanziell unterschiedlichen Ergebnisse zu bringen, die Unterschiede sind in dieser Studie nicht signifikant. Die Autoren empfehlen daher den einfacheren und flexibleren Einsatz globaler, auf Ratingskalen basierender Einschätzungen.

Ein etwas anderes Vorgehen wählten Madelaine und Wheldall (2005); dort sollten Lehrer 12 zufällig ausgewählte Schüler ihrer Klasse entsprechend ihrer Leistungen im Lesen (Passage Reading Test, PRT) in eine Reihenfolge bringen. Nur 55 Prozent der Lehrer identifizierten die laut Testergebnis schlechtesten Leser, und die drei schlechtesten Schüler zu benennen (unabhängig von der Reihenfolge) schafften sogar nur 15 Prozent der Lehrer.

Die geschilderten Erkenntnisse werden auch durch die neueste Metaanalyse zur diagnostischen Kompetenz (Südkamp, Kaiser & Möller, eingereicht) gestützt. Wie Südkamp und Kollegen berichten, ist die signifikant höhere Urteisgüte bei spezifischer Erfassung gegenüber der globalen Variante einer der auffälligsten und konsistent über alle Untersuchungen, in denen ein entsprechender Vergleich stattfand, gefundenen Befunde. Nichtsdestotrotz scheinen globale Urteile die am häufigsten gewählte Operationalisierung zu sein, mutmaßlich deshalb, weil es Lehrern zwar offensichtlich leichter fällt, spezifische Einschätzungen abzugeben, der Erhebungsaufwand in den Studien und der Zeitaufwand für die Lehrer aber bei globalen Urteilen um einiges geringer sind.

Niveauurteile zu Schülerleistungen

Für eine korrekte Niveaubeurteilung ist es erforderlich, die Leistungsfähigkeit der Schüler in Beziehung zur Aufgabenschwierigkeit zu setzen. Dazu ist nicht nur eine gute Kenntnis der Schüler selbst erforderlich, sondern ebenso, dass schwierigkeitsgenerierende Merkmale realistisch eingeschätzt werden. Wie sich in der PISA-2000-Studie zeigte, fällt schon dieser aufgabenbezogene Aspekt nicht leicht. Dort sollten 60 Lehrplanexperten die Lösungswahrscheinlichkeiten für verschiedene Leseaufgaben schätzen. Wie sich

zeigte, lagen diese Schätzungen deutlich über den tatsächlich gemessenen Lösungshäufigkeiten. So erwarteten sie beispielsweise von ca. 60 Prozent der Schüler aus der neunten Klassenstufe, dass sie Aufgaben der höchsten Kompetenzstufe V lösen könnten, bis Schuljahresende trauten sie dies gar 86 Prozent zu; tatsächlich lag der Anteil insgesamt aber lediglich bei nicht einmal 20 Prozent (Artelt et al., 2001).

In aller Regel zeigt sich, dass das allgemeine Niveau von Leistungsurteilen deutlich von dem der tatsächlichen Schülerleistungen abweicht (Schrader & Helmke, 2001). Feinberg und Shapiro (2003) fanden in ihrer bereits erwähnten Studie zur Einschätzung der mündlichen Lesefähigkeit neben den relativ hohen Produkt-Moment-Korrelationen eher schwache Niveaueinschätzungen der Lehrer. Wie in Tabelle 2 ersichtlich, überschätzten die Lehrer die Lesefähigkeit der Drittklässler um über eine Standardabweichung und die der Viertklässler immerhin noch um eine halbe Standardabweichung. Lediglich in der fünften Klassenstufe korrespondieren Schülerleistungen und Lehrereinschätzungen eng.

Tabelle 2: Schülerleistungen im lauten Lesen je nach Klassenstufe im Vergleich zu den korrespondierenden Lehrereinschätzungen bei Kenntnis der zugrunde liegenden Aufgaben

Klassenstufe	*Schülerleistungen: mittlere Punktzahl (SD)*	*Lehrereinschätzungen: mittlere Punktzahl (SD)*
3 (N=13)	61,5 (30,1)	95,5 (35,5)
4 (N=10)	107,3 (41,7)	126,7 (38,0)
5 (N=6)	119,8 (38,0)	122,7 (59,2)

Anm.: Daten aus Feinberg und Shapiro (2003)

Zu einem ähnlichen Befund kam man auch in einer weiteren Studie, in der genau wie bei Feinberg und Shapiro (2003) Lehrkräfte die Schülerleistungen in curriculumsbasierten Maßen bewerten sollten. Hamilton und Shinn (2003) prüften dabei die Genauigkeit der Einschätzungen zu Schülerleistungen in verschiedenen Lesemaßen und fanden überall signifikante Überschätzungen der Leistungen. Noch schlechter schnitten Lehrer in einer Untersuchung von Eckert und Kollegen die Lese- und Mathematikleistungen ihrer Schüler einschätzen sollten: In fast allen getesteten Domänen (eingesetzt wurden verschiedene Mathematik- und Leseaufgaben) und über verschiedene Fähigkeitsstufen hinweg erwiesen sie sich als inakkurat bei der Leistungsniveaueinschätzung, indem ihre Urteile deutlich über den Schülerleistungen lagen (Eckert, Dunn, Codding, Begeny & Kleinmann, 2006).

Zusammenhänge zwischen den Komponenten

Neben der Vielzahl der aufgeführten Befunde zu einzelnen Komponenten der Urteilsgüte sind in neueren Studien auch die Zusammenhänge zwischen diesen Komponenten untersucht worden. Damit werden die einzeln betrachteten Aspekte der Urteilsgenauigkeit als sich gegenseitig ergänzende Facetten verstanden, so wie es auch Cronbach (1955) im Sinn hatte, als er zu differenzierterer Betrachtung der Diagnoseleistungen riet.

Wurde in Studien direkt untersucht, inwiefern die Einschätzleistungen nach verschiedenen Komponenten diagnostischer Kompetenz miteinander zusammenhängen, gab es stets ernüchternde Befunde. So fanden Begeny und Kollegen - wenn auch nur auf der Basis von 10 Lehrern derselben Schule - zwar gute Leistungen in der Rangkomponente, aber keine gute Akkuratheit, wenn es um das Erkennen des Leistungsniveaus ging (Begeny et al., 2008). Ein ähnliches Bild ergab sich in der Studie von Madelaine und Wheldall bei der Einschätzung der Lesegeschwindigkeit: Einer mittleren Korrelation von r = .73 standen sehr schlechte Erkennungsraten der jeweils besten und schlechtesten drei Schüler gegenüber (Madelaine & Wheldall, 2005).

4.2 Güte von Ziffernbenotungen

Sind die bisher differenzierten Formen von Lehrerurteilen hauptsächlich für die Forschung von Interesse, so ist eine andere Art des Urteils von höchster Bedeutung für schulische Prozesse: die Noten (vgl. Kapitel 1). Sie stehen seit mehreren Jahrzehnten im Fokus der (vorwiegend pädagogisch-psychologischen) Forschung, die teils ernüchternde Erkenntnisse zu Tage gebracht hat.

Noten als Lehrerurteile

Bereits vor annähernd einem halben Jahrhundert konstatierte Weiß (1965), dass die Zuverlässigkeit von Ziffernbenotungen in der Schule „betrüblich gering“ sei. Bei gleichen Schulleistungen können Schüler deutlich unterschiedliche Bewertungen erhalten, was der eigentlichen Intention von Schulnoten deutlich zuwider läuft. Noten sollen verschiedene Funktionen erfüllen, die sich jeweils aus dem Verwertungszusammenhang ergeben. An erster Stelle steht dabei die pädagogische Funktion. Danach sollen Noten Schülern Arbeitsanreize bieten, ihnen als Kontrolle der eigenen Leistungen dienen, sie mit Leistungsvergleichen und Normen vertraut machen, gleichermaßen aber auch als Informations- oder Rückmeldemedium für Schule

und Elternhaus fungieren. Weiterhin erfüllen Noten auch gesellschaftliche Funktionen, indem sie gegenüber Dritten dokumentieren und legitimieren (Berechtigungsfunktion), nach Leistung klassifizieren und - beispielsweise bei der Wahl des Ausbildungswegs - selegieren sowie im Sinne einer Kontrolle die Einhaltung der Schulpflicht oder die Effekte schulpolitischer, organisatorischer und pädagogischer Maßnahmen transparent machen (Tent, 2006). Mehrere Autoren haben in den zurückliegenden Jahrzehnten jeweils etwas unterschiedliche Kategorisierungsversuche für die Funktionen von Noten und Zeugnissen vorgenommen, die verschieden stark ausdifferenziert waren. Im Kern ist den aufgelisteten Funktionen jedoch nichts Bedeutsames hinzugefügt worden. Entscheidender ist jedoch die Kluft zwischen theoretischer Bestimmung und praktischer Bedeutung der Noten. Allein die Tatsache, dass einige Funktionen nur bedingt miteinander vereinbar sind (z.B. die Rückmeldung mit der Selektion) oder es durch mangelnde Abgrenzbarkeit zu Überschneidungen zwischen ihnen kommt, deutet auf ein substantielles Dilemma hin. Ingenkamp (1995a) bezeichnet es denn auch als schwer verständlich, wie man glauben konnte, dass Zensuren so unterschiedliche Aufgaben gleichzeitig erfüllen könnten. Seiner Ansicht nach kann die Zensur keiner der ihr zugeschriebenen Funktionen wirklich gerecht werden, und ihr Fortbestand sei nur dadurch zu erklären, dass sie die am einfachsten und am vielseitigsten handhabbare Beurteilungsform sei.

Es ist daher wenig verwunderlich, dass Ingenkamp sein vielbeachtetes Buch über ungerechtfertigte Ansprüche, Erwartungen und Hoffnungen an das Benotungssystem „Die Fragwürdigkeit der Zensurengebung“ (Ingenkamp, 1995b) nannte. Jürgens (2005, S. 66) formuliert es noch drastischer, wenn er konstatiert: „Die globale, wenig aussagekräftige Zensur, deren Zustandekommen als ein Prozess von Mängeln bezeichnet werden muss, versagt gerade in dem Bereich völlig, der für das schulische Lernen bzw. Weiterlernen entscheidend ist.“ Dem Sinn von Verfahren zur Leistungsmessung, den Lernenden eine aufklärende Rückmeldung zu geben, Schwächen zu finden und zu überwinden oder Lernzielstufen zu korrigieren, können Zensuren mit ihrer groben Rasterung überhaupt nicht erfüllen. Dazu bedürfte es weitaus differenzierterer Verfahren. So scheint am Ende die einzige Funktion, die Noten wirklich zuverlässig ausüben können, die der Informationsverdichtung zu sein.

Aus einer Reihe von Studien zur Validität der Notengebung gingen in den letzten Jahrzehnten verschiedene Mängel der Notengebung hervor, deren bedeutsamste im Folgenden kurz dargestellt werden:

- Schülerinnen erhalten - zumindest in der Primarstufe, für die Sekundarstufe sind die Befunde uneinheitlich (z.B. Roeder, Baumert, Sang & Schmitz, 1986) - durchschnittlich bessere Noten, als aufgrund von in Tests erfassten Leistungen anzunehmen wäre. Einen Hinweis auf mögliche Ursachen gibt eine Studie von Tiedemann (1995), aus der hervorgeht, dass deutsche Lehrer gute Leistungen bei Mädchen stärker auf besondere Anstrengung und weniger auf besonders ausgeprägte Fähigkeiten zurückführen als bei Jungen. Auch Trautwein und Baeriswyl (2007) kommen nach mehrebenenanalytischen Prüfungen von Referenzgruppeneffekten bei Übertrittsentscheidungen zu dem Fazit, dass es unzweifelhaft scheine, „dass geschlechterspezifische Stereotypien die Urteile der Lehrkräfte beeinflussten". Insbesondere fanden sie, dass bei gleicher Testleistung Jungen eine höhere kognitive Leistungsfähigkeit (für die Bereiche Deutsch und Mathematik) attestiert wurde als Mädchen, wohingegen bei Mädchen die schulische Motivation als vergleichsweise höher eingeschätzt wurde (vgl. auch Kapitel 4.6.2).
- Längere Aufsätze werden im Durchschnitt besser (wenn auch uneinheitlicher) bewertet als kürzere zum gleichen Thema.
- Grammatik- und Orthografiefehler beeinflussen die Beurteilung von Aufsätzen, auch wenn es ausdrücklich nur um inhaltliche Bewertung gehen soll.
- Eine langsamere Sprechweise führt - bei exakt gleichem Text - zu einer schlechteren Bewertung.
- Schüler, die Lehrern sympathisch sind, erhalten bessere Beurteilungen als andere Schüler und als ihre Leistungen rechtfertigen würden.
- In verschiedenen Unterrichtsfächern wird unterschiedlich streng geurteilt. Besondere Strenge ist in Fächern festzustellen, in denen die Leistungen mithilfe schriftlicher Tests überprüft werden und in denen die Fehlerzahl besonders einfach quantifizierbar ist, z.B. in Mathematik.
- Besonders auffallend ist auch der immer wieder festgestellte Zusammenhang zwischen sozialer Herkunft der Schüler und der Leistungsbeurteilung in der Hinsicht, dass Schüler aus sozial benachteiligten Familien bei gleicher Leistung im Durchschnitt schlechter bewertet werden als Schüler aus der gehobenen Sozialschicht (u.a. Ditton, 2010, vgl. ebenso Kapitel 4.6.2).
- Das größte Problem scheint jedoch zu sein, dass Lehrer Noten verteilen, „ohne hinreichende Informationen über den Leistungsstand ihrer Klasse im Vergleich zu dem anderer Klassen zu besitzen" (Ingenkamp, 1995a; Roeder et al., 1986). Dabei kann es passieren, dass die objektiv gleiche

Leistung in einer Klasse sehr gut, in einer anderen Klasse hingegen sehr schlecht beurteilt wird, was als sogenannter Referenzgruppeneffekt bekannt ist (vgl. hierzu z.B. Trautwein & Baeriswyl, 2007). Aber auch über verschiedene Länder, Schultypen oder Fächer sind Leistungen nur bedingt miteinander vergleichbar. In verschiedensten Untersuchungen wurde konsistent bestätigt, dass es eine große Bandbreite von Schülern mit gleichen oder vergleichbaren Testergebnissen gibt, die jedoch von ihren Lehrern völlig unterschiedliche Noten erhalten (Bos et al., 2004; Lehmann, Peek & Gänsfuß, 1997; Lehmann et al., 2004). Die dafür ursächliche Orientierung der Lehrer am klasseninternen Bezugssystem kommt einem normorientierten Notensystem sehr nahe, bei dem Noten nach dem Platz vergeben werden, den ein Schüler im Vergleich zu anderen Schülern in der Klasse hat. Diese Form der Benotung („grading on the curve") ist jedoch - zumindest in Deutschland - nicht zulässig (Woolfolk, 2008), sondern Noten müssten kriteriumsorientiert, also entsprechend der objektiv erreichten Leistung, vergeben werden, um nicht der Beziehung der Schüler untereinander und ihrer Motivation zu schaden (Guskey, 2000; Krumboltz & Yeh, 1996). Dafür fehlt den meisten Lehrern jedoch der entsprechende Maßstab.

Zumindest zwei der dargestellten Befunde wurden zum Teil durch entgegengesetzte Befunde relativiert. Tent (2006) listet einige ältere Untersuchungen auf, die allenfalls einen geringen Einfluss externer Faktoren auf die Schulnoten feststellten. In der Grundschule bleiben nach einer Studie von Tent, Fingerhut und Langfeldt (1976) beispielsweise für die Merkmale Intelligenz, Alter, Geschlecht und soziale Herkunft je nach Fach nur zwischen vier und sieben Prozent Varianzaufklärung, während die objektiv gemessenen Schülerleistungen zwischen 68 und 75 Prozent der Varianz der Fachnoten zwischen Schülern erklärten. Dieser Befund bestätigte sich u.a. auch für die sechste Klassenstufe in einer Untersuchung von Schrader und Helmke (1990).

4.3 Urteilsgüte in nicht-kognitiven Bereichen

Kognitive Schülermerkmale (Leistungen) stehen im Schulalltag zweifelsohne im Zentrum der Beurteilungs- und Einschätztätigkeiten von Lehrern. Für Leistungen werden Noten vergeben, für sonstige Eigenschaften der Schüler normalerweise nicht. Eine Ausnahme bilden die sogenannten Kopfnoten, die Auskunft über überfachliche Kompetenzen der Schüler zu Bereichen wie Betragen, Fleiß, Mitarbeit oder Selbstständigkeit auf den (meist Grundschul-)Zeugnissen geben sollen. Die Regelungen hierfür unterscheiden sich stark

von Bundesland zu Bundesland, sie werden mal als Ziffernnoten, mal als Verbalurteile vergeben. Darüber hinaus werden sie nicht in allen Bundesländern eingesetzt. Während Kopfnoten zum Beispiel in Sachsen direkt aus der DDR-Tradition übernommen und bis heute beibehalten wurden, führte beispielsweise Bayern sie erst ein, um sie nach heftigen Protesten von Lehrern und Eltern wenig später wieder abzuschaffen. In jedem Fall war oder ist es bei der Vergabe von Kopfnoten nötig, dass Lehrer auch motivationale oder emotionale Eigenschaften der Schüler einschätzen können, was mangels entsprechender Maßstäbe oder einheitlicher Kriterien ganz andere Anforderungen an die Lehrer stellt als die an Fehlern oder Richtiglösungen orientierten Leistungsbewertungen.

Die Fähigkeit, nicht-kognitive Merkmale von Schülern korrekt zu erkennen, scheint jedoch nicht nur für die Vergabe von Kopfnoten wichtig zu sein. Die Berücksichtigung ‚allgemeiner Fähigkeiten' - in Ergänzung zu den Leistungen - nennt beispielsweise die KMK schon seit Jahrzehnten in ihren Regelungen zum Übergang von der Grund- auf die weiterführenden Schulen (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2010), wenngleich diese Forderung nicht konkret mit Inhalt gefüllt wird. Schulen sind nicht nur reine ‚Expertenschmieden', in denen es allein auf das Erreichen möglichst hoher Leistungen ankommt, sondern sie sind durch ihren Erziehungsauftrag auch Institutionen, in denen u.a. die Ausbildung von sozialen Kompetenzen ebenfalls einen hohen Stellenwert einnehmen soll. Darüber hinaus werden auch Einflüsse nicht-kognitiver Eigenschaften auf das Lernen oder auf kognitive Merkmale bzw. Schülerleistungen angenommen. Zu den Aufgaben eines Lehrers gehört es deshalb ebenso, auch diese vorauslaufenden Bedingungen schulischer Leistungen nicht nur zu kennen, sondern auch korrekt zu diagnostizieren, um auf dieser Grundlage pädagogische Entscheidungen zu treffen (vgl. Spinath, 2005). Wenn die Leistungen von Schülern hinter den Erwartungen zurückbleiben, sollten Lehrer auf der Basis korrekter Einschätzungen über die zugrunde liegenden Ursachen adäquat reagieren können, indem sie z.B. geeignete Fördermaßnahmen ergreifen.

Dies erfordert ein ausreichendes Maß an Sensibilität der Lehrer. Aus diesem Grund werden in der vorliegenden Arbeit auch einige ausgewählte nicht-kognitive Variablen der Schüler, die sich in der pädagogisch-psychologischen Forschung als bedeutsam für den Lernerfolg erwiesen haben, in Beziehung zur diagnostischen Kompetenz gesetzt. Da Erforschung der Urteilsgenauigkeit von Lehrern in nicht-kognitiven Bereichen bislang

nur eine sehr untergeordnete Rolle spielte, mangelt es diesbezüglich oft an aussagekräftigen Forschungsergebnissen. Der nachfolgende Überblick stellt insofern neben den wenigen Untersuchungen die generelle Bedeutung der betrachteten Merkmale dar.

Lernfreude

Lernfreude bedeutet, dass Schülerinnen und Schüler eine Lernhandlung freiwillig und um der Sache selbst willen ausführen (Baumert et al., 2010). Sie umfasst die Freude an der alltäglichen schulischen Arbeit und die frohe Erwartungshaltung, mit der ein Kind an die Schularbeit geht (vgl. Rauer & Schuck, 2003). In der pädagogischen Psychologie wird Lernfreude hauptsächlich aus zwei verschiedenen Perspektiven betrachtet. Dies ist zum einen die normative Perspektive, die unabhängig von Effizienzüberlegungen auf die positive oder negative Besetzung des Lernens und der Schule gerichtet ist. Derart betrachtet handelt es sich um eine per se bedeutsame Zielvariable des Unterrichts, die ihren Niederschlag nicht zuletzt in Gesetzen und Regelungen des Bildungs- und Erziehungswesens gefunden hat (Helmke, 1993). Die andere Perspektive ist eine funktionale, in der die affektive Einstellung gegenüber schulischen Leistungsanforderungen gleichermaßen als Bedingung und Folge individuellen Lernens, schulischer Erfolge oder Misserfolge angesehen wird (ebd.). Gerade Misserfolge können dazu führen, dass Schüler negative affektive Einstellungen gegenüber schulischen Anforderungen entwickeln, was wiederum leistungsverringernde Vermeidungshaltungen in zukünftigen Situationen nach sich ziehen kann. Folglich kommt auch den Lehrern eine wichtige Rolle insofern zu, als dass sie die Lernfreude der Schüler versuchen sollten zu steigern, was eine Vermeidung von die Lernfreude reduzierenden Verhaltensweisen einschließt.

Über die Genauigkeit von Einschätzungen der Lernfreude durch Lehrer im Sinne der diagnostischen Kompetenz ist bislang nur wenig bekannt, allerdings weiß man, dass zumindest Eltern die Schulfreude ihrer Kinder überwiegend positiver einschätzen als die Kinder selbst (Rolff, Holtappels, Klemm, Pfeiffer & Schulz-Zander, 2002; Schneider, 2005). Aus der IGLU-Studie ist bekannt, dass sich ein positives Verhältnis der Schüler zu ihrer Lehrkraft stark auf ihre Lernfreude auswirkt. Darüber hinaus erweist sich die (wahrgenommene) diagnostische Kompetenz der Lehrer für Mädchen gar nicht und für Jungen nur minimal als bedeutend für ihre Lernfreude, anders als bspw. für die Lernmotivation, die in signifikant positivem Zu-

sammenhang zur wahrgenommenen diagnostischen Kompetenz steht (Janke, 2006).

Wird die schulische Lernmotivation, wie in der vorliegenden Studie, im Sinne von intrinsischer Motivation, Freude an schulischen Inhalten und dem Verfolgen von Lernzielen operationalisiert, so ergeben sich nur geringe positive Zusammenhänge ($r \leq .30$) zwischen Schülerselbstberichten und Lehrereinschätzungen (Helmke & Fend, 1981; Swanson, 1985). In einer Studie von Givvin, Stipek, Salmon und MacGyvers (2001) wurde die Güte von Lehrerurteilen zur Motivation der Schüler im Fach Mathematik untersucht. Siebzehn Lehrkräfte sollten dabei viermal im Schuljahr Aspekte der Motivation (wahrgenommene Fähigkeit, Lernorientierung sowie positive und negative Orientierung) von jeweils 6 Schülern einschätzen, die ihrerseits Angaben dazu machten. Wie sich zeigte, korrespondierten die Lehrerurteile nur moderat mit den Selbsteinschätzungen der Schüler, die Korrelationen lagen je nach Messzeitpunkt und Motivationsaspekt zwischen $r = .14$ und $r = .51$, im Mittel bei $r = .28$. Die Schüler differenzierten dabei deutlicher zwischen den verschiedenen Aspekten, als dies die Lehrer taten. Weiterhin fanden die Autoren, dass die Selbsteinschätzungen der Schüler über die Messzeitpunkte hinweg deutlicher schwankten und weniger stabil waren als die Lehrereinschätzungen. Nicht zuletzt zeigen die Befunde dieser Studie, dass auch bei der Einschätzung nicht-kognitiver Merkmale eine erhebliche Varianz zwischen Lehrern hinsichtlich der Genauigkeit ihrer Urteile besteht. Der Range der Korrelationen liegt zwischen $r = .10$ und $r = .80$.

Anhand von Daten aus den miteinander verschachtelten längsschnittlichen Projekten LOGIK (Longitudinalstudie zur Genese individueller Kompetenzen) und SCHOLASTIK (Schulorganisierte Lernangebote und Sozialisation von Talenten, Interessen und Kompetenzen) konnten einige wichtige Erkenntnisse zur Entwicklung der Lernfreude in der Kindergarten- und Grundschulzeit gesammelt werden (s. Helmke, 1993). Aus ihnen geht u.a. die große Bedeutung der Lehrer für die Lernfreude hervor. So erwies sich die Entwicklung der Lernfreude als beeinflussbar durch die Art des Unterrichts. Ist der Unterricht verständlich und gut organisiert, kann die Lernfreude der Schüler positiv beeinflusst werden (Helmke, 1997). Weiterhin zeigte sich in der Untersuchung, dass die Lernfreude überdurchschnittlich hoch anstieg, wenn die Lehrer auch ihren eigenen Unterricht und die eigenen didaktischen Bemühungen für Lernschwierigkeiten auf Seiten der Schüler verantwortlich machten. Und nicht zuletzt erwiesen sich ein adaptiver, schwierigkeitsangemessener Unterricht sowie die Qualität der Motivati-

on durch den Unterricht als bedeutsam für die Lernfreude(entwicklung) (ebd.). Da es der Anspruch von Schule und Lehrkräften ist, neben dem Lernerfolg der Schüler u.a. auch deren Lernfreude zu stärken, ist die fortlaufende Wahrnehmung der Lernfreude in der Klasse und bei individuellen Schülern eine wichtige Komponente, um den Erfolg des eigenen Unterrichts überprüfen zu können.

Unterschiede zwischen Schulklassen klären über die Grundschulzeit hinweg zwischen acht und zwölf Prozent der Varianz der Lernfreude auf, wobei als Prädiktoren vor allem das Klassenmanagement, die Klarheit und Verständlichkeit des Unterrichts und das soziale Klima in der Klasse eine wichtige Rolle spielen. Dies unterstreicht die bedeutsame Rolle des Lehrers für die Lernfreude. Korrelationsanalysen deuten darüber hinaus darauf hin, dass die Lernfreude im Zusammenhang mit unterstützendem Lehrerverhalten und der Schülerpartizipation (Epstein, 1981) sowie einer positiven Schüler-Lehrer-Beziehung und dem Interesse an Lerninhalten (Czerwenka et al., 1990) steht.

Leistungsängstlichkeit

Ein weiteres bedeutsames nicht-kognitives Schülermerkmal mit Bezug zur Leistung ist die Leistungs- oder Prüfungsängstlichkeit. Sie gilt neben dem Fähigkeitsselbstbild als die am meisten untersuchte individuelle Determinante der Schulleistung (Helmke, 1983a, 1983b; Helmke & Weinert, 1997; Pekrun & Helmke, 1991). Der Begriff bezeichnet die überdauernde Anfälligkeit einer Person, in als leistungsbezogen wahrgenommenen Situationen mit einem charakteristischen Muster motorischer, physiologischer o.ä. Prozesse zu reagieren, wodurch die intellektuelle Leistungsfähigkeit beeinträchtigt wird (Helmke & Weinert, 1997). Dabei sind seit Liebert und Morris (1967) zwei Hauptkomponenten zu unterscheiden. ‚Worry' bezeichnet Selbstzweifel, Sorgen oder andere aufgabenirrelevante Kognitionen, und ‚Emotionality' die Wahrnehmung stressbedingter Veränderungen des eigenen Körpers, also u.a. Nervosität und Aufgeregtheit. Besonders erstere kann sich leistungsmindernd auswirken, indem sie beim Lernen von Informationen (encoding) die Verarbeitungstiefe reduziert oder beim Abruf von Gelerntem in einer Leistungssituation (retrieval) zu Aufmerksamkeitsstörungen führt (Heckhausen, 1982).

Leidet ein Schüler unter Leistungsängstlichkeit, so kann sich dies hauptsächlich auf drei Ebenen zeigen (vgl. Rost & Schermer, 2007). Physiologische Indikatoren wie Herzklopfen oder erhöhter Puls und Blutdruck auf der ersten

sowie emotional-subjektive Indikatoren wie Unwohlsein auf der zweiten Ebene sind sehr häufig Begleiterscheinungen, die von Außenstehenden jedoch ohne Hilfsmittel nicht wahrgenommen werden können. Auf der dritten Ebene stehen beobachtbare Verhaltensweisen wie Zittern, Artikulationsstörungen oder Aggression, die zwar gut wahrgenommen werden können, für die jedoch Angst als Ursache nicht immer sofort zweifelsfrei erkennbar ist und die in der Regel nur bei sehr extremer Angstausprägung zu Tage treten. Insofern stellt die Diagnose von Leistungsängstlichkeit Lehrer vor eine große Herausforderung. Für die zuverlässige Diagnose von Ängstlichkeit haben sich standardisierte Fragebögen durchgesetzt, die Lehrern im Unterrichtsalltag jedoch nicht zur Verfügung stehen.

Gleichwohl spielen Lehrer als Bedingungsfaktor der Leistungsängstlichkeit von Schülern direkt oder indirekt eine wichtige Rolle. Rost und Schermer (2007) nennen in ihrem Überblickskapitel zur Leistungsängstlichkeit sieben wichtige Bedingungsbündel, von denen die ersten fünf auf Lehrkräfte zurückzuführen sind: 1) Lehrerverhalten. Autorität, Zuwendungsentzug, Tadel, Demütigungen oder Strafen sind exemplarische Verhaltensweisen von Lehrern, die direkt Angst bei Schülern auslösen können. 2) Inhalt und Vermittlung des Lehrstoffs. Drücken sich Lehrer unverständlich aus, ist ihr Unterricht verwirrend oder unpräzise oder wird selten Feedback gegeben, kann ebenfalls Angst entstehen. 3) Schulbezogene Fähigkeiten und Fertigkeiten. Hier wirkt sich beispielsweise die Überforderung von Schülern negativ aus. 4) Schulleistungsbewertung. Strenge Zensuren, scharfe Auslese, mangelnde Transparenz und Inkonsequenz der Bewertungskriterien können angstauslösende Faktoren sein. 5) Gestaltung von Prüfungssituationen. Einen besonders großen Einfluss auf die Leistungsangst haben auch Prüfungsanforderungen, mögliche Ungewissheit über Lernziele, hoher Zeitdruck, die Verwendung unfairer Aufgaben oder die Ankündigung schwieriger Aufgaben. Darüber hinaus nennen die Autoren noch 6) das Schüler-Schüler-Verhältnis und 7) das Verhalten und die Einstellungen der Eltern als bedeutsame Bedingungsfaktoren der Leistungsangst. Führt man sich die enorme Verantwortung der Lehrkräfte in Hinblick auf die Angstausprägung vor Augen, ist der Schluss naheliegend, dass das Ziehen von Rückschlüssen aus Schülerreaktionen auf die eigenen Verhaltensweisen ein wichtiger Ansatz für Lehrer sein sollte, um leistungsangststeigernde Verhaltensweisen abzustellen. Die Fähigkeit, Ängstlichkeit bei den Schülern wahrzunehmen, ist diesbezüglich ebenso eine wichtige Voraussetzung wie dafür, betroffenen Schülern Bewältigungsstrategien vermitteln oder in extremen Fällen auch weitere Unter-

stützungsmaßnahmen wie schulpsychologische Dienste oder Beratungen einleiten zu können (vgl. Metzig & Schuster, 2001).

Die Befunde zur Lehrerurteilsgüte in Bezug auf die Leistungsangst von Schülern sind uneinheitlich. Während einige Untersuchungen schwache positive Zusammenhänge in der Größenordnung bis maximal r = .30 fanden (Böhnke, Silbereisen, Reynolds & Richmond, 1986; Helmke, 1980), erwies sich in einer Studie von Helmke und Fend (1981) der Zusammenhang zwischen Lehrereinschätzungen und Angstausprägungen mit r = -.12 als zwar nicht signifikant, aber dennoch tendenziell negativ. In der Dresdner Kinder-Angst-Studie fand man, dass die Selbsteinschätzungen achtjähriger Kinder niedrig, aber dennoch signifikant (r = .08) mit den Lehrereinschätzungen korrespondierten, noch höher jedoch mit den Einschätzungen durch die eigenen Eltern (r = .15) (Federer, Stüber, Margraf, Schneider & Herrle, 2001). Auch aktuelle Studien konnten wiederholt belegen, dass Lehrkräfte keine hinreichend genauen Einschätzungen der Leistungsängstlichkeit ihrer Schüler vornehmen können (Faber, 2001, 2006; Spinath, 2005). Darüber hinaus gibt es Hinweise darauf, dass sich Lehrer bei der Beurteilung dieses Merkmals stark von der Kenntnis der Schülerleistung leiten lassen (z. B. Faber, 1994, 2001). In der aktuellen Forschung ist die Betrachtung der Leistungsangst im diagnostischen Kontext weniger populär, so dass hier neuere Erkenntnisse wünschenswert wären.

Fachinteresse

Interessen, meist verstanden als kognitive Gerichtetheit im Sinne einer subjektiv als bedeutsam erlebten Beziehung zwischen einer Person und einem Gegenstandsbereich (Krapp, 1989), sind als zentrales Element selbstbestimmten Handelns zu verstehen und werden als Voraussetzung und zugleich als Ziele des Lernens und der Entwicklung angesehen (Krapp, 2006). Zentrale Bezugsgrößen sind dabei Faktoren wie Inhaltsspezifität, Gegenstandsbezug und intrinsische Motivation (Helmke & Weinert, 1997), was sie u.a. von der Leistungsmotivation abgrenzt. Dabei wird die Wirkungsweise des Interesses als Bedingungsfaktor für Lernen und Leistungserreichung mit komplexen motivationalen und kognitiven Orientierungen, Lernstrategien, Aufmerksamkeitssteuerung oder emotionalen Begleitzuständen erklärt, ist jedoch insgesamt bislang nur unzureichend erforscht (Prenzel, Lankes & Minsel, 2000). Nichtsdestotrotz wurde und wird sich in der Forschung sehr intensiv mit den (Fach-) Interessen und deren Auswirkungen auf das Lernen und Leistungen

beschäftigt (Baumert, Schnabel & Lehrke, 1998; Deci, 1992; Köller, Baumert & Schnabel, 2000; Krapp, Hidi & Renninger, 1992; Schiefele, 1996; Schiefele, Krapp & Winteler, 1992).

Inwiefern das (Fach-)Interesse der Schüler von Lehrern eingeschätzt werden kann, ist bislang nur sehr selten untersucht worden. In der Unterrichtsstudie SALVE wurde für Lehrer von Fünftklässlern herausgefunden, dass diese das mittlere Interesse ihrer Klasse für den Unterricht durchschnittlich unterschätzten (Hosenfeld, Helmke & Schrader, 2002). Da hier jedoch nur der Anteil an Schülern, die den Unterricht interessant fand, erhoben wurde, ist der Indikator recht grob. Spezifischer sind die Analysen von Karing (2009), die in der BiKS-Studie u.a. das fachspezifische Interesse für Deutsch und Mathematik von Gymnasial- und Grundschullehrern zu den individuellen Schülern in ihren Klassen erhob. Ihre Ergebnisse zeigen, dass zwar Grundschullehrer die Fachinteressen ihrer Schüler - gemessen an der Rangkomponente diagnostischer Kompetenz - signifikant besser einschätzen können, dass bei beiden Lehrergruppen die Urteilsgüte im Bereich der Interessen deutlich unter derjenigen in verschiedenen Leistungsbereichen liegt ($r = .21$ bis .37).

Wie Hannover (1998) darlegt, gibt es Belege für eine gegenseitige Beeinflussung von Interessen- und Selbstkonzeptentwicklung bei Schülern. Interessen spielen für das Erreichen selbstbezogener Ziele und für die Regulation selbstbezogener Gefühle eine wichtige Rolle und helfen dabei, sich selbst zu definieren und dies auch anderen Menschen mitzuteilen. Lehrer sollten dies bei der Instruktion ihrer Schüler berücksichtigen, um ihnen Gelegenheit zu geben, eine größere und unvoreingenommenere Bandbreite an Interessen auszubilden und somit vorhandene Fähigkeiten weiterzuentwickeln. Aus der Bedeutung des Interesses der Schüler für ihr Lernen ergibt sich ferner, dass eine gezielte Unterstützung der Schüler zur Entwicklung ihrer Interessen oder die an vorhandenen Interessen orientierte Unterrichtsgestaltung nützliche Ansätze für Lehrkräfte darstellen, um Schüler optimal zu fördern. Werden beispielsweise im Leseunterricht Texte zu Themen behandelt, für die sich die Schüler auch interessieren, dann sind sie nicht nur motivierter und aufmerksamer, sondern verstehen die Texte auch schneller und erzielen im Endeffekt bessere Ergebnisse (Krapp, 2002). Um das Interesse der Schüler auch als Lernpotential für den Unterricht nutzen zu können, muss dieses jedoch von Lehrern erkannt und korrekt eingeschätzt werden. Die wenigen Studien, in denen diese Fähigkeit untersucht wurde, deuten jedoch nicht darauf hin, dass dies tatsächlich in ausreichendem Maß der Fall ist.

Fazit

Studien zur Güte von Lehrerurteilen in nicht-kognitiven Bereichen sind deutlich seltener als zur Güte in kognitiven Bereichen. Die wenigen Befunde machen jedoch deutlich, dass auch Erkenntnisse in diesem Bereich eine große Bedeutung für ein besseres Verständnis von Entscheidungsprozessen auf Seiten der Lehrer haben. Die vorliegenden Ergebnisse müssen dahingehend interpretiert werden, dass es Lehrern deutlich schwerer fällt, nicht-kognitive Schülermerkmale korrekt einzuschätzen als kognitive. Wie Schrader (2006) angibt, ist die im Vergleich zum Leistungsbereich berichtete niedrigere Urteilsgenauigkeit in emotional-motivationalen Bereichen häufig zum einen auf für die Lehrer schwerer zu erkennende Indikatoren, zum anderen aber auch auf niedrigere Reliabilitäten der jeweils eingesetzten Skalen zurückzuführen. Lernfreude, Ängstlichkeit oder Interesse sind nicht direkt beobachtbar und auch nicht Gegenstand von regelmäßig durchgeführten Kontrollen oder Tests, wie dies bei Leistungsmerkmalen der Fall ist. Hinzu kommt, dass Schüler ihre Emotionen, Gedanken etc. unterschiedlich deutlich offenbaren. Und wenn Schüler Emotionen zeigen, ist zum einen nicht sicher, ob es ihre wahren Gefühle sind, und zum anderen ist der Interpretationsspielraum für die Lehrer viel größer als bei einfach zu quantifizierenden Leistungsindikatoren. Sicher kann argumentiert werden, dass, auch wenn Schüler nicht ihre wahren Gefühle zeigen, Lehrer dennoch eben diese ‚hinter der Fassade' erkennen müssen, um angemessen darauf zu reagieren. Allerdings wird hier von Lehrern eine Wahrnehmungsgabe und ein Deutungsgeschick erwartet, das wohl kaum zu erbringen ist.

4.4 Homogenität diagnostischer Urteile

Wenn von *der* diagnostischen Kompetenz die Rede ist, verleitet es oft unbemerkt zu der Annahme, dass damit eine allgemeine Personenfähigkeit gemeint sei. Entsprechend Weinerts Kompetenzdefinition (vgl. Kapitel 2.2) sind Kompetenzen jedoch mehr oder weniger bereichsspezifisch ausgeprägt, so dass nicht davon ausgegangen werden kann, dass Lehrer über mehrere Urteilsbereiche hinweg eine ähnliche Urteilsgüte aufweisen.

Forschungsergebnisse zum Einfluss des beurteilten Leistungsbereichs auf die Güte von Urteilen sind uneinheitlich. In einer Reihe von Studien zeigte sich eine signifikant niedrigere Einschätzgüte in den naturwissenschaftlichen Fächern oder Sachkunde gegenüber den Bereichen Sprachen, Lesen und Mathematik (u.a. Hopkins, George & Williams, 1985). Im Vergleich von

Mathematik, Lesen und Rechtschreiben fanden Demaray und Elliott (1998) keine bedeutsamen Unterschiede, wohingegen sich in anderen Studien höhere Korrelationen für den Bereich Lesen als für die Einschätzungen zu mathematischen Leistungen ergaben (Eckert et al., 2006; Hinnant, O'Brien & Ghazarian, 2009). Auch innerhalb derselben Domäne wurden differente Urteilsgüten gefunden. So war bei Coladarci (1986) die Einschätzung von Fähigkeiten in mathematischen Berechnungen signifikant akkurater als die von mathematischen Konzepten. Insgesamt sind die Untersuchungsresultate zum Einfluss des Fachs auf die Urteilsgenauigkeit somit als inkonsistent zu bezeichnen. Über eine große Anzahl von Studien hinweg finden Südkamp und Kollegen (eingereicht) in ihrer Metaanalyse insgesamt keine Belege für Unterschiede in der Urteilsgenauigkeit zwischen sprachlichen und mathematischen Leistungsbereichen.

Dieser allgemeine Befund sagt allerdings noch nichts über die Zusammenhänge der Urteilsgüte in verschiedenen Bereichen auf der Ebene individueller Lehrer aus. Dass generell bspw. Mathe- und Deutschleistungen annähernd gleich gut eingeschätzt werden können, bedeutet nicht, dass dies auch auf jeden einzelnen Lehrer zutrifft. Gerade dies ist allerdings ein Aspekt, der von großer Bedeutung ist, um das Konstrukt der diagnostischen Kompetenz besser zu verstehen. Ist sie eine allgemeine Fähigkeit, die sich bereichsunabhängig zeigt, oder muss man sie als bereichsspezifisch ansehen, wie es der Begriff der Kompetenz auch nahelegt? Erstaunlicherweise liegen bislang nur sehr wenige Untersuchungen vor, die sich der Homogenität diagnostischer Urteile über verschiedene Bereiche hinweg gewidmet haben. Die wenigen Ergebnisse deuten darauf hin, dass die diagnostische Kompetenz nicht als homogenes Konstrukt angesehen werden kann. Wie Spinath (2005) zeigte, sind Lehrer, die ihre Schüler in einem Bereich recht gut einschätzen (z.B. in der Intelligenz), nicht zwangsläufig auch in anderen Bereichen (z.B. Lernmotivation) nahe an der Realität. Lorenz und Artelt (2009) konnten zeigen, dass diese Differenzen nicht nur zwischen kognitiven und nichtkognitiven Bereichen auftreten, sondern ebenso innerhalb eines dieser Felder. So gibt es keinerlei Zusammenhänge zwischen der Güte der Urteile im Fach Deutsch und im Fach Mathematik, wohl aber zwischen inhaltlich ähnlichen Bereichen (hier die dem sprachbezogenen Bereich zuzuordnenden Leistungsgebiete Wortschatz und Textverstehen).

Beurteilen Lehrer ihre Schüler, so müssen sie sich ein Bild über den zu beurteilenden Bereich sowie über den Schüler selbst machen. Bei der Beurteilung mehrerer Leistungsaspekte, wie es gerade in der Grundschule die Regel

ist, sind entsprechend bereichsspezifische Abwägungen zu treffen, wobei es durchaus plausibel erscheint, dass Leistungen in ähnlichen Bereichen auch ähnlich eingeschätzt werden, da wahrscheinlich auch in der Realität der Schüler derartige Zusammenhänge bestehen. Doch Leistungen fallen individuell unterschiedlich aus, und nur weil beispielsweise sowohl Lesefähigkeit als auch Rechtschreibleistung statistisch miteinander einhergehen, muss das nicht auch in jedem Einzelfall so sein. Von einer Leistung automatisch auf die andere zu schließen griffe daher womöglich für den Lehrer zu kurz. Vielmehr muss seinem Urteil bei jedem einzelnen Schüler ein ganz individueller Blick auf die verschiedenen zu beurteilenden Leistungs- und Merkmalsfacetten zugrunde liegen. Dies ist für Lehrer die große Herausforderung, wenn er in allen Bereichen gleichermaßen akkurate Urteile fällen soll.

4.5 Stabilität diagnostischer Urteile

Soll diagnostische Kompetenz als Lehrerfähigkeit angesehen werden, so müsste sie sich in Untersuchungen (zumindest über einen überschaubaren Zeitraum) als stabil erweisen. Voraussetzung dafür ist, dass Lehrkräfte ihre Urteile fortwährend an den (sich möglicherweise verändernden) Leistungsstand der Schüler anpassen können. Es bedarf eines besonders feinen Gespürs und der Bereitschaft zur ständigen Selbstüberprüfung, sich nicht darauf zu verlassen, dass eine einmalig aufgestellte Meinung über die Leistungsfähigkeit eines Schülers nicht einfach beibehalten, sondern permanent an aktuellen Leistungen überprüft wird. Dies fällt möglicherweise umso schwerer, wenn Lehrern die relativ große zeitliche Stabilität von Schulleistungen in unterschiedlichsten Leistungsbereichen bewusst ist, wie sie konsistent in verschiedenen Studien nachgewiesen wurde (z.B. Boland, 1993; Grube, 2004; Schneider & Stefanek, 2007). Helmke (2009) empfiehlt Lehrern deshalb, sich von Zeit zu Zeit selbst zu überprüfen, indem sie die gleichen Arbeiten ihrer Schüler mit einem Abstand von z.B. einem halben Jahr ein zweites Mal korrigieren und bewerten. Dadurch könnten sie feststellen, wie reliabel ihre Urteile über die Leistungen sind.

Anpassung von Einschätzungen an den Leistungsstand betreffende Umstände

In dem Moment, in dem man von einer diagnostischen Kompetenz spricht, unterstellt man implizit bereits, dass es sich um eine Fähigkeit der diagnostizierenden Person handelt, die - im Sinne von Weinerts Kompetenzbegriff (Weinert, 1999b) - stabil ist. Dies bedeutet auch, dass der Diagnostiker (hier:

die Lehrkraft) in der Lage ist, sein Urteil an sich möglicherweise verändernde Umstände anzupassen. In einer schon etwas älteren Laborstudie zur Genauigkeit von Lehrerurteilen über Schüler und den Einfluss dieser Urteile auf Entscheidungen im Instruktionsprozess wurde dieser Aspekt auch überprüft (Borko et al., 1979). Eine der Fragestellungen der Studie lautete, ob Lehrer ihre Leistungseinschätzungen über einen fiktiven Schüler ändern, wenn sie zusätzliche Informationen über diesen Schüler erhielten. Es zeigte sich, dass Lehrer ihre Einschätzungen tatsächlich anpassten, allerdings nur dann, wenn die zusätzliche Information als verlässlich wahrgenommen wurde. Stuften die Lehrkräfte die zusätzlichen Informationen über diesen Schüler als unreliabel ein, z.B., weil es sich lediglich um die Meinung eines Mitschülers handelte, neigten die Lehrer dazu, diese zu ignorieren.

In einer weiteren Studie wurde untersucht, ob sich berufserfahrene Lehrer hinsichtlich der so genannten Bestätigungstendenz (confirmation bias) von Lehramtsstudenten unterscheiden (van Ophuysen, 2006). Die Bestätigungstendenz meint die Tendenz, ein einmal gefälltes Urteil auch nach dem Erhalt widersprüchlicher oder uneindeutiger Informationen nicht zu ändern. Sozialpsychologisch lässt sich dieses Phänomen durch verzerrte oder selektive Wahrnehmung, Vermeidungsstrategien und Dissonanzreduktion erklären. Diese Mechanismen führen dazu, dass einmal aufgestellte Hypothesen oder Urteile relativ lange aufrechterhalten werden (Nisbett & Ross, 1980). Es besteht die generelle Annahme, dass Novizen eher für die Bestätigungstendenz anfällig sind als Experten. Van Ophuysen setzte mit ihrer Untersuchung an diesem Punkt an und verglich 83 Lehramtsstudierende mit 57 Grundschullehrern. Die Probanden sollten für einen fiktiven Schüler, der durch verschiedene Aussagen charakterisiert wurde, eine Schullaufbahnempfehlung abgeben. Während eine Hälfte der Stichprobe alle 12 Statements gleichermaßen berücksichtigen sollte, wurde die andere Hälfte gebeten, zunächst nur auf Grundlage der ersten drei (positiven und auf eine Gymnasialempfehlung hindeutenden) Statements zu urteilen, und anschließend auch die weiteren neun Items - von denen 6 nicht auf eine Gymnasialempfehlung hindeuteten - zu berücksichtigen. Für diese zweite Hälfte der Stichprobe zeigte sich, dass nur eine von 17 Lehrerinnen (6%) bei ihrer zuerst getroffenen Empfehlung blieb, alle anderen aber ihre Einschätzungen revidierten. Im Gegensatz dazu blieben trotz der eher negativen Informationen über den Schüler 13 von 37 Novizen (35%) bei ihrem anfangs abgegebenen Urteil. Dieses Ergebnis deutet darauf hin, dass erfahrene Lehrer gut in der Lage sind, ihre Urteile an neue Bedingungen oder Informationen anzupassen und damit eine wichtige Voraussetzung für das Fällen korrekter

Urteile auch unter sich verändernden Umständen erfüllen. Sie unterscheiden sich darin deutlich von unerfahrenen Lehramtsstudenten.

Zur Stabilität von Lehrerurteilen in kognitiven Schulleistungsbereichen

Erstaunlicherweise ist während der jahrzehntelangen Erforschung der diagnostischen Kompetenz - abgesehen von sehr wenigen Ausnahmen - stets querschnittlich vorgegangen worden. Entsprechend gab es zu Fragen der Dimensionalität und zeitlichen Stabilität diagnostischer Kompetenz ein großes Forschungsdefizit (vgl. Baumert & Kunter, 2006). Der Annahme, dass es sich bei der Fähigkeit zur korrekten Schülereinschätzung um eine Personenfähigkeit handelt und dass der verwendete ‚Kompetenz'-Begriff entsprechend gerechtfertigt ist, fehlte somit ein wichtiger empirischer Beleg. Erst seit jüngster Zeit liegen dazu Erkenntnisse vor. Einerseits konnte Spinath (2005) zeigen, dass Lehrereinschätzungen zu den Schülermerkmalen Intelligenz, schulische Fähigkeitsselbstwahrnehmung, schulische Lernmotivation und schulbezogene Leistungsängstlichkeit ein halbes Jahr nach der ersten Messung immer noch Test-Retest-Korrelationen in Höhe von .50 bis .72 aufwiesen. Dieser in diesem Beitrag eher beiläufig erwähnte Befund stellt die erste empirische Überprüfung der Stabilität von Lehrerurteilen dar, die dem Autor bekannt ist. Andererseits, und darin liegt ein besonderer Neuigkeitswert auch der vorliegenden Arbeit, konnte auf Grundlage der hier verwendeten, längsschnittlich begleiteten BiKS-Stichprobe gezeigt werden, dass die Annahme einer (fachspezifischen) Stabilität der diagnostischen Kompetenz gerechtfertigt zu sein scheint, da sich auch für die schulbezogenen Leistungsbereiche Arithmetik, Wortschatz und Textverstehen Stabilitäten in Höhe von r_{tt} = .38 bis .58 zeigten (vgl. Artelt, 2009; Lorenz & Artelt, 2009).

Zur Stabilität von Lehrerurteilen in nicht-kognitiven Bereichen

Auch bei der Beurteilung von nicht-kognitiven Schülermerkmalen spielen, ähnlich wie bei der Einschätzung von Leistungen, die Erwartungen der Lehrer eine große Rolle. Forschungsergebnisse legen nahe, dass die anfänglichen Erwartungen oder Einschätzungen von Lehrern, auch wenn sie auf falschen oder lückenhaften Informationen beruhen, nur schwer geändert werden können. Auch wenn sich die Merkmale der Schüler ändern, sie z.B. motivierter, interessierter oder fleißiger werden, tendieren Lehrer - falls sie es überhaupt wahrnehmen - eher dazu, dies als Ausnahme anzusehen, als ihre generelle Meinung zu ändern. Selbst wenn sie sich bewusst bemühen, ihre Urteile anzupassen, verändern sich ihre Erwartungen nicht als direkte Kon-

sequenz von Änderungen im Schülerverhalten über die Zeit oder verschiedene Situationen hinweg (Brophy, 1983; Givvin et al., 2001; Wigfield & Harold, 1992). Die wenigen Forschungsbefunde auf diesem Gebiet deuten somit eher darauf hin, dass die Stabilität der Urteilsgüte in nicht-kognitiven Bereichen niedriger ausfällt als in kognitiven.

4.6 Bedingungsfaktoren diagnostischer Kompetenz

Individuelle Lehrer unterscheiden sich mitunter extrem in ihrer Urteilsgenauigkeit. In nahezu jeder Studie zur diagnostischen Kompetenz von Lehrern wird die große Streubreite der Urteilsgüte festgestellt (siehe überblicksartig die Metaanalyse von Hoge und Coladarci, 1989). Die Korrelationen einzelner Lehrer decken dabei einen Bereich ab, der in manchen Studien im negativen Bereich beginnt und wiederum auch nicht selten annähernd bis zu perfekten Übereinstimmungen reicht. Daher ist es nur folgerichtig, dass nach den Gründen für diese Unterschiedlichkeit gesucht wird.

Es sind viele Ursachen für diese interindividuellen Unterschiede denkbar. In der vorliegenden Arbeit sollen mögliche Bedingungsfaktoren untersucht werden, die sich grob den folgenden drei Gruppen zuweisen lassen: 1) Merkmale der Lehrperson selbst sind besonders naheliegend, aber auch 2) Merkmale einzelner Schüler oder 3) der gesamten Klasse lassen sich als plausible Einflussfaktoren auf die Güte von Einschätzungen theoretisch begründen. Darüber hinaus sind weitere Bedingungsfaktoren denkbar, die über diese Kategorisierung hinausgehen, beispielsweise Eigenschaften der zu beurteilenden Leistungsbereiche und der entsprechenden Anforderungen. In den folgenden Abschnitten wird anhand der hier betrachteten Einflussfaktoren gleichsam der weitere Stand der Forschung zur diagnostischen Kompetenz dargestellt.

4.6.1 Lehrermerkmale

In den folgenden Abschnitten soll unter Bezugnahme auf verschiedene theoretische Ansätze die Bedeutung von Lehrermerkmalen für die diagnostische Kompetenz hergeleitet werden. Die diagnostische Kompetenz als eine der Kernkompetenzen von Lehrkräften steht in engem Zusammenhang zur Suche nach dem „guten“ oder „optimalen“ Lehrer. Wie Weinert (1996) beschreibt, beschäftigt sich die pädagogisch-psychologische Forschung seit über hundert Jahren mit der „Persönlichkeit von Lehrern, ihrem pädagogi-

schen Handeln, mit der Bedeutung didaktischer Expertise und den Wirkungen des Unterrichts ...“, stets auf der Suche nach jenen Eigenschaften, die erfolgreiche Lehrer von weniger erfolgreichen unterscheiden. Diese Suche verlief über die Jahrzehnte keineswegs einheitlich, sondern war verschiedenen Strömungen unterworfen, die heute als Paradigmen der Lehrerforschung bezeichnet werden.

4.6.1.1 Expertisestatus

Unter dem Einfluss der „kognitiven Wende“ entstand ab den 1980er Jahren die Expertiseforschung, die auf der Grundannahme beruht, dass erfolgreiches Lehrerhandeln zu großen Teilen vom persönlichen Wissen und Können des Lehrers abhängig sei (Besser & Krauss, 2009). Im sogenannten Expertenparadigma werden Lehrer als kompetente Fachleute für die „Kunst des Unterrichtens“ betrachtet (u.a. Berliner, 1986; Bromme, 1992; Weinert, Helmke und Schrader, 1992). Hierbei geht es jedoch nicht so sehr um Persönlichkeitseigenschaften des guten Lehrers, sondern um ein Ensemble von Fertigkeiten und Wissen, das für die Bewältigung der beruflichen Anforderungen nötig ist. Im Unterschied zum Persönlichkeitsparadigma wird angenommen, dass dieses Wissen und die Fertigkeiten durch Lernprozesse (Studium, Weiterbildung etc.) erlernbar und entwicklungsfähig sind.

Die Anhaltspunkte, nach denen Expertenlehrer in der Forschung ausgewählt werden, sind uneinheitlich. Gemein ist lediglich eine grobe Definition von Experten, die eine besondere Leistungsstärke in relevanten Gegenstandsbereichen beinhaltet (Gruber, 2006). Je nach Tradition, die mit unterschiedlichen Operationalisierungen und Gruppenaufteilungen einhergehen, findet man in empirischen Arbeiten jedoch sehr unterschiedliche Auswahlkriterien für Experten:

- Entweder werden Expertenlehrer von Vorgesetzten ausgewählt (z. B. Strahan, 1989),
- oder es ist die Ausbildung oder die Berufserfahrung entscheidend (z. B. Berliner, 1986),
- oder es gelten solche Lehrer als Experten, die sich in längsschnittlichen Erhebungen als „Optimallehrer“ erwiesen haben (z. B. Leinhardt, 1987)
- oder deren Klassen besonders hohe Leistungszuwächse in bestimmten Fächern aufweisen (z. B. Leinhardt & Greeno, 1986),

- oder es wird eine Verbindung von Vorgesetztennominierungen und Unterrichtsbeobachtung vorgenommen (z. B. Berliner, 1986).

Die jeweils zugrunde gelegten Ankerpunkte, z.B. was genau einen Optimallehrer ausmacht, in welchem konkreten Bereich er besonders gut oder wie hoch der Leistungszuwachs der Schüler sein muss, damit deren Lehrer als Experte gelten kann, sind darin freilich noch nicht definiert und können je nach Studie und theoretischer Perspektive variieren. Darüber hinaus gibt es noch weitere Abwandlungen der Expertenauswahl, die hier jedoch nicht weiter thematisiert werden sollen. Ein kurzer Überblick zur reichhaltigen Forschungslage in Bezug auf Lehrerexpertise und mit besonderem Bezug zur diagnostischen Kompetenz und deren Bedingungsfaktoren soll im Folgenden gegeben werden.

Wie eben gezeigt, bestehen in verschiedenen theoretischen Ansätzen sehr unterschiedliche Definitionen von Expertise. Beispielsweise werden in der allgemeinen kognitiven Expertiseforschung mal „umfangreiches Wissen über eine kleine Klasse von Fragen und Problemen, [das] durch langandauernde Erfahrung erworben" wurde (Zimbardo & Gerrig, 1999, S. 786), und mal spezielle ‚Fähigkeiten' (Gobet, 2001) als Expertise bezeichnet, oft scheint sich die Definition aber einfach der Operationalisierung in der jeweiligen Arbeit anzupassen (Gruber, 2006). Kann als Schnittmenge der verschiedenen Begriffsverwendungen von Expertise noch angesehen werden, dass es sich bei einem Experten um jemanden handelt, der auf einem spezifischen Gebiet möglichst dauerhaft herausragende Leistungen erzielt (Gruber, 2006), so fallen bei der Übertragung auf das Gebiet der Lehrerexpertise besonders dann Antworten schwer, wenn man danach fragt, wann eine Leistung eines Lehrers als „herausragend" gilt oder wie stark eingegrenzt man das entsprechende „Gebiet" verstehen soll (Besser & Krauss, 2009).

Besonders in der englischsprachigen Literatur wird der Expertenbegriff („expert teacher") sehr häufig für Lehrer mit besonders hohem Wissen oder Können gebraucht, die sich von unerfahreneren Lehrern unterscheiden (sog. Experten-Novizen-Vergleich). Im Gegensatz dazu versteht Bromme, der sich im deutschsprachigen Raum in besonderer Weise um Versuche zur Vereinheitlichung der Begriffe verdient gemacht hat, unter Expertenlehrern generelle Fachleute (für das Lehren und Lernen in der Schule) im Vergleich zu nicht der Berufsgruppe der Lehrer angehörigen Laien (Bromme, 2008). Für ihn ist somit die Profession an sich die Expertise, und nicht, wie im englischsprachigen Bereich (z.B. Berliner, 1994; Leinhardt & Greeno, 1986; Livingston & Borko, 1989; Sternberg & Horvath, 1995), eine dauerhafte Spit-

zenleistung auf einem eingegrenzten Gebiet. Trotz unterschiedlicher Begriffsdefinitionen sind beide Forschungsansätze auf der Suche nach Merkmalen des ‚guten Lehrers'.

Im Zentrum des pädagogischen (Novizen-)Experten-Paradigmas, das die Analyse der für den größeren Lehrerfolg von Lehrerexperten gegenüber Novizen verantwortlichen kognitiven Kompetenzen zum Gegenstand hat, steht das professionelle Expertenwissen der Lehrer, für das angenommen wird, dass ihm eine große Bedeutung für die Klassenführung und den Lernerfolg der Schüler zukommt (vgl. Helmke & Weinert, 1997). Aus der Fülle der dazu vorliegenden Forschungen lässt sich als wichtigster Faktor die Erkenntnis herauslesen, dass Experten im Vergleich zu Novizen auch in schwierigen Unterrichtssituationen auf ein größeres und effektiveres Wissen zurückgreifen können, auf dessen Grundlage sie durchdachtere Entscheidungen treffen und Handlungsroutinen flexibler nutzen können (Berliner, 1994; Bromme, 1992).

Im Expertenansatz steckt der Aspekt der Erlernbarkeit. Dies ist nicht nur bereits im Namen dieses Paradigmas erkennbar, sondern schlägt sich auch in Fragestellungen und Forschungsmethoden nieder (Chi, Glaser & Farr, 1988). Im Vordergrund steht darin die Frage, wie sich lange Erfahrung und Übung auf die Bewältigung komplexer Aufgaben auswirken. In der Lehrerforschung wurden darauf aufbauend verschiedene Studien zum Vergleich von Experten und Novizen hinsichtlich verschiedener Facetten durchgeführt. Berliner (1992) fand in einer Untersuchung zur kognitiven Verarbeitung von Unterrichtsfotos, -videos und schriftlichen Informationen deutliche Unterschiede in der kategorialen Wahrnehmung der Unterrichtsereignisse zwischen (von Schuldirektoren benannten, besonders guten) „Expertenlehrern", Anfängern und Anwärtern. Die Experten zeichneten sich dadurch aus, dass sie die Situationen ganzheitlich und nicht nur auf einzelne Schüler bezogen betrachteten und eher interpretierten, was sie sahen. Ihre weniger erfahrenen Kollegen beschrieben die Situationen hingegen eher mit dem Fokus auf einzelnen Details. Auch hinsichtlich der Wahrscheinlichkeit systematischer Urteilsverzerrungen bei Übergangsempfehlungen konnten Unterschiede zwischen Experten und Novizen festgestellt werden. In einer quasiexperimentellen Untersuchung von van Ophuysen (2006) zeigte sich ebenfalls eine differenziertere Wahrnehmung und Bewertung von Informationen durch Experten. Darüber hinaus unterlagen sie deutlich seltener systematischen Urteilsverzerrungen im Sinne einer Tendenz, einmal gefällte Urteile auch bei Vorliegen widersprüchlicher Informationen aufrecht zu erhalten

(ähnliche Ergebnisse lassen sich u.a. auch bei Krems, 1996, finden). Bei den Novizen zeigte sich hingegen eine gewisse Voreingenommenheit in der Bewertung konsistenter bzw. inkonsistenter Informationen in dem Sinne, dass zum eigenen Urteil in Widerspruch stehende Informationen in ihrer Bedeutung abgewertet wurden. Weiterhin zeigen Forschungsergebnisse, dass Experten Informationen schneller und angemessener verarbeiten und speichern können als Novizen (Chi, Feltovich & Glaser, 1981; Gobet, 1996). Es wird außerdem davon ausgegangen, dass Experten seltener die sogenannte Bestätigungstendenz (confirmation bias) aufweisen. Darunter wird in der kognitiven Sozialpsychologie der Effekt verstanden, dass ein einmal gefälltes Urteil nicht verändert wird, selbst wenn neue Informationen uneindeutig oder widersprüchlich in Bezug auf dieses Urteil sind. Erklärt wird dies dadurch, dass uneindeutige Informationen verzerrt als zustimmend interpretiert und widersprüchliche Informationen mehr oder weniger ignoriert werden, so dass bestehende Hypothesen lange Zeit unverändert aufrecht erhalten bleiben (Fiedler, 1983). Ein weiteres Merkmal besonders berufserfahrener Lehrer fand Bromme (1987), indem Mathematiklehrer direkt nach dem Unterricht nach ihrer Erinnerung an Probleme oder Lernfortschritte einzelner Schüler gefragt wurden. Der zunächst enttäuschende Befund, dass diese Lehrer kaum Erinnerungen an einzelne Schüler hatten, entpuppte sich bald als Ausdruck ihrer ganzheitlichen Wahrnehmung von Unterrichtsepisoden.

Chi, Glaser & Farr (1988) zählen im Vorwort ihres Buches „The nature of expertise“ mehrere von der Forschung bestätigte Charakteristika von Experten auf, darunter die folgenden (Übersetzungen nach Lingelbach, 1995):

- Experten zeigen nur in ihrer eigenen Domäne hervorragende Leistungen. Eine Übertragung der Fähigkeiten auf andere Gebiete scheint nicht stattzufinden.
- Experten sind schnell: Sie sind schneller als Novizen in der Ausführung von Fertigkeiten ihrer Domäne, und sie lösen Probleme ihrer Domäne schneller und mit weniger Fehlern.
- Experten sehen und repräsentieren ein Problem ihrer Domäne auf einer tieferen (mehr prinzipiellen) Ebene als Novizen; diese tendieren dazu, Probleme auf der Oberflächenebene zu repräsentieren.
- Experten brauchen anfänglich mehr Zeit als Novizen, um ein Problem qualitativ zu analysieren.

Als Experten angesehene Lehrer sind jedoch nicht automatisch auch erfolgreiche Lehrer. Bereits in früheren Studien zeigte beispielsweise das fachbezogene curriculare Wissen von Expertenlehrern zum Teil erhebliche Lücken (Leinhardt & Smith, 1985). Wie Bromme (1992) meint, zeigt sich die Qualität des Wissens in der Flüssigkeit des Unterrichts, wobei dies wiederum eine gute Anpassung an die Schülerkenntnisse und ihre Verhaltensweisen erfordert und u.a. persönliches Können - und nicht nur das Wissen darüber - voraussetzt. Aus Interviewstudien ist beispielsweise bekannt, dass viele Lehrer zwar wissen, wie sie im Unterricht vorgehen sollten, aber nicht über die persönlichen Routinen verfügen, um dieses Wissen auch umzusetzen. Daher verwundern auch kaum die größtenteils als enttäuschend angesehenen Ergebnisse von Studien, die den Zusammenhang zwischen dem curricularen Fachwissen und dem Unterrichtserfolg (z.B. Leistungszuwachs der Schüler) untersuchten (Gage & Berliner, 1977). In Folge dieser Befunde wurde der Expertenansatz, der ja gerade das Wissen als Bedingung der beruflichen Leistung untersucht, infrage gestellt, doch es erwies sich bald als nicht vernünftig, den korrelativen Zusammenhang beider Faktoren kausal zu interpretieren, weil dabei eine Vielzahl von stattfindenden Vermittlungsprozessen, z.B. die Instruktionsqualität, außer Acht gelassen wird.

Ähnliches trifft möglicherweise auf die ebenfalls recht ernüchternden Befunde hinsichtlich der Güte diagnostischer Urteile zu. Trotz der oben beschriebenen Erkenntnisse, dass Lehrerexperten bei ihrer Urteilsfindung differenzierter vorgehen als Novizen, brachten Untersuchungen, die die Lehrerexpertise über die Berufserfahrung operationalisierten und sie mit der Urteilsgüte in Verbindung brachten, nicht die erwarteten Ergebnisse ans Licht. Bereits vor über zwanzig Jahren vermutete Coladarci (1986) durchaus plausibel, dass die Berufs- bzw. Lehrerfahrung von Lehrkräften einen Einfluss auf die Güte ihrer Schülerbeurteilungen habe, da sich im Laufe der Zeit und durch den sich stets wiederholenden Prozess von Hypothesenbildung und darauf folgender Überprüfung entsprechende kognitive Strukturen ausbilden und zu einer Verbesserung der Urteilsgenauigkeit könnten. Dabei fand er ebenso wenig eine Erklärung für Unterschiede zwischen Lehrern wie Demaray und Elliott (1998) in ihrer Studie: die Prüfung, ob die Anzahl der Arbeitsjahre als Lehrer und der Bildungsgrad (Master oder Bachelor) einen systematischen Einfluss auf ihre Urteilsgenauigkeit hat, verlief auch hier ohne entsprechenden Befund. Eine Vielzahl anderer Untersuchungen ist ebenso dieser These nachgegangen, ohne jedoch auch nur den geringsten Hinweis auf einen derartigen Zusammenhang zu finden (Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003; Impara & Plake, 1998;

Leinhardt, 1983; Mulholland & Berliner, 1992; Wild & Rost, 1995). In einer aktuellen Metaanalyse, die Studien zur diagnostischen Kompetenz aus den letzten zwanzig Jahren zusammenfasst, wurde neben der Berufserfahrung auch das Alter der Lehrer, das eng an ihre Berufserfahrung gekoppelt ist, überprüft, jedoch ebenfalls ohne Effekte zu finden (Südkamp et al., eingereicht).

Zu beachten ist bei all diesen Studien jedoch, dass Expertise und Berufserfahrung nicht deckungsgleich sind. Versteht man unter dem Grad der Expertise das jeweilige „Ausmaß an thematisch relevanter Erfahrung“ (van Ophuysen, 2006), so wäre davon auszugehen, dass Lehrer mit längerer Berufserfahrung auch ein höheres Expertiseniveau erreicht haben als Kollegen mit weniger Berufsjahren. Im Gegensatz zu einer Grundannahme des Novizen-Experten-Paradigmas erwies sich die Zunahme von Lehrerexpertise jedoch nicht als lineare Funktion der Dauer unterrichtlicher Erfahrung. Hinweise stützen hingegen die Annahme, dass sich mit der Dauer der professionellen Tätigkeit verschiedene kognitive (z.B. Expertisezuwachs) und motivationale (z.B. Selbstwirksamkeitsüberzeugungen) Entwicklungen überlagern (Helmke & Weinert, 1997). Sicher spielt dabei auch eine große Rolle, in welchem Maße Lehrer ihre gesammelten Erfahrungen reflektieren, um aus ihnen überhaupt einen Expertisezuwachs generieren zu können.

4.6.1.2 Wissenskomponenten

Das Lehrerwissen, das in der Expertiseforschung als Indikator für die Expertise herangezogen wird, ist selbst auch Gegenstand umfangreicher Forschungen, da es als basale Voraussetzung für erfolgreiches Lehrerhandeln angesehen werden kann. Implizit wird in der wissenschaftlichen Beschäftigung mit Lehrern davon ausgegangen, dass verschiedene Wissenskomponenten für erfolgreichen Unterricht notwendig sind. Die Notwendigkeit, diese Komponenten zu untersuchen, um ein besseres Verständnis insbesondere der Voraussetzungen für erfolgreiches Lehren zu gewinnen, wird schon lange gefordert: „To understand teachers‘ thinking, then, it is necessary to describe the tasks teachers face in classrooms and explicate the knowledge structures that underlie the interpretation and accomplishment of these tasks.” (Carter & Doyle, 1987). Das Unterrichten von Schulklassen stellt vielseitige Anforderungen an die Lehrkräfte und stellt sie vor Aufgaben, die Sabers, Cushing und Berliner (1991) als „simultaneity, multidimensionality and immediacy“ bezeichnen. Einzelne Elemente von gutem und erfolgreichem Unterricht, wie sie innerhalb des Prozess-Produkt-Paradigmas ange-

nommen wurden, können von Lehrern gelernt und geübt werden, doch die Schwierigkeit besteht in der effektiven „Orchestrierung“ dieser Elemente. Die speziell an die Bedürfnisse der Schüler angepasste Komposition ist laut Gage (1978) gerade das, was die „Kunst des Unterrichtens“ ausmacht.

Modell nach Shulman

Ein besonders populäres Modell zur Strukturierung des professionellen Lehrerwissens, in das auch die diagnostische Kompetenz als Facette eingebettet und erklärt werden kann, stammt von Shulman (Shulman, 1986, 1987). Darin wird aufgegriffen, dass die oben beschriebene Suche nach dem ‚guten Lehrer‘, die den Expertiseansatz antreibt, sich für das typische Wissen und Können von Lehrern interessierte (Besser & Krauss, 2009). In seinem Modell versuchte Shulman zu systematisieren, welche konkreten Inhalte und Bereiche das professionelle Wissen von Lehrkräften umfasst bzw. umfassen sollte. Shulmans Konzeptualisierung ist bis heute die verbreitetste und gebräuchlichste Taxonomie des professionellen Lehrerwissens und lässt sich auch auf das im Deutschen übliche Professionsverständnis übertragen. Sie soll im Folgenden vor dem Hintergrund einer der dafür relevanten Facetten - der diagnostischen Kompetenz - beschrieben werden. Trotz vielschichtiger Befunde lassen die Erkenntnisse aus der Expertiseforschung insgesamt unter anderem darauf schließen, dass professionelles Wissen domänenspezifisch und ausbildungs- bzw. trainingsabhängig sowie sehr gut vernetzt und hierarchisch angeordnet ist und dass Basisprozeduren zwar automatisiert, aber dennoch flexibel an die spezifischen Bedingungen des Einzelfalls oder Kontextes anpassbar sind (s. die ausführliche Zusammenfassung verschiedenster Forschungsergebnisse bei Baumert & Kunter, 2006). Zentral an Shulmans Modell (vgl. Abbildung 3) ist die Differenzierung von (in der Abbildung hervorgehobenen) Kompetenzbereichen, nämlich dem Fachwissen (subject-matter content knowledge), dem allgemeinen pädagogischem Wissen (general pedagogical knowledge) und dem fachdidaktischen Wissen (pedagogical content knowledge). Sie hat sich in der Praxis durchgesetzt und wurde in nahezu allen Übersichtsartikeln übernommen (Borko & Putnam, 2004; Helmke, 2003; Lipowsky, 2006). Dabei beinhaltet das Fachwissen das lehrereigene Verständnis des zu unterrichtenden Lernstoffes, das pädagogische Wissen schließt u.a. Prinzipien und Strategien des Klassenmanagements ein, die über Fachinhalte hinausgehen, und das fachdidaktische Wissen der Lehrer bildet gleichsam eine Schnittstelle der eben genannten Bereiche als eine spezielle und je Lehrer einzigartige Form des professionellen Verständnisses ab. Shulmans Modell ist u.a. von Bromme (1997) um ver-

schiedene Facetten erweitert worden, die an dieser Stelle aber nicht detaillierter ausgeführt werden sollen. Im Folgenden werden die einzelnen Kompetenzbereiche des unten abgebildeten Modells ausführlicher beschrieben.

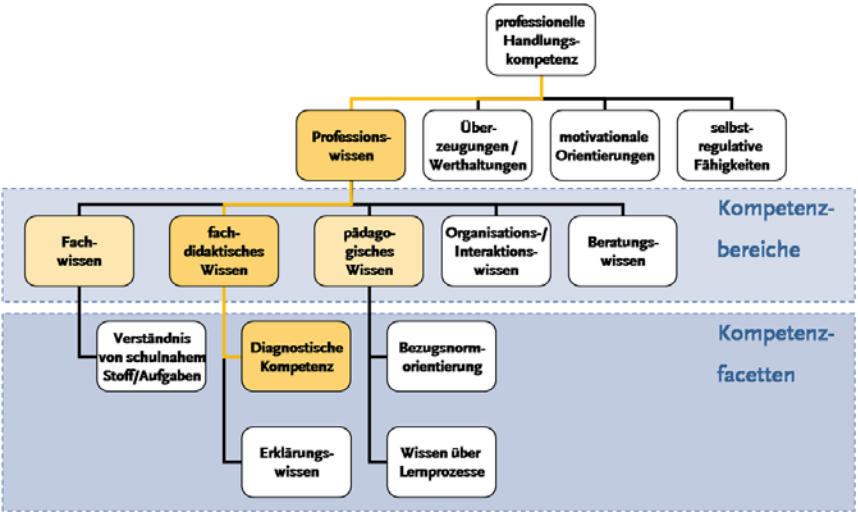

Abbildung 3: Topologie der professionellen Wissensdomänen von Lehrkräften, Abbildung nach Krauss et al. (2004) (vgl. Shulman, 1986)

- pädagogisches Wissen

Das (generelle) pädagogische Wissen (pedagogical knowledge) umfasst nach dem Modell Facetten wie das Wissen über allgemeines didaktisches Vorgehen, über Lernprozesse, Unterrichtsmanagement oder über Bewertungsstandards und ihre Wirkungen. Das pädagogische Wissen, das sowohl prozedurale als auch deklarative Aspekte des Professionswissens umfasst und durch seinen generellen Charakter fachunspezifisch ist, ist für die Gestaltung und Optimierung der Lehr-Lern-Situationen und somit für einen reibungslosen und effektiven Ablauf des Unterrichts von großer Bedeutung. Darüber hinaus dient es auch der Aufrechterhaltung eines förderlichen sozialen Klimas in der Klasse (Krauss et al., 2004).

- Fachwissen und fachdidaktisches Wissen

Auch die diagnostische Kompetenz von Lehrern lässt sich in das Shulman'sche Modell einordnen, nämlich als Kompetenzfacette des fachdidaktischen Wissens. Das fachdidaktische Wissen und Können der Lehrer ist u.a. besonders dann gefragt, wenn es darum geht, Aufgaben auszuwählen oder

Arbeitsaufträge zu formulieren, mit denen Lehrer besonders gut die Stärken und Schwächen der Schüler abschätzen können, und das nicht erst in Form von Leistungskontrollen, sondern bereits während des Lernprozesses. Obwohl allgemeiner Konsens darüber besteht, dass beide Wissensformen zum Kern professioneller Kompetenz von Lehrern gehört, sind sowohl in der deutschen als auch in der englischsprachigen Literatur empirische Arbeiten zur Rolle des fachdidaktischen Wissens und - mehr noch - des Fachwissens verhältnismäßig selten (s. Baumert & Kunter, 2006; Minnameier, 2005). Somit liegen statt gesicherter Kenntnisse eher ausschnittartige Befunde vor. Hinzu kommt, dass die theoretisch recht ausdifferenzierte Unterscheidung von Fachwissen und fachdidaktischem Wissen in der Praxis kaum in unterschiedlichen Indikatoren Ausdruck findet, so dass ihr Verhältnis zueinander empirisch ungeklärt ist. Auch konzentrieren sich die vorhandenen Untersuchungen in der Regel auf querschnittliche Betrachtungen im Fach Mathematik oder in den Naturwissenschaften.

Ein großer Mangel besteht weiterhin darin, dass beide Wissenskomponenten überwiegend über distale Maße wie staatliche Zertifizierung, Abschlüsse oder die Zahl besuchter Kurse oder Weiterbildungen operationalisiert werden (s.u.). Da diese Indikatoren jedoch kaum in der Lage sind, Auskunft über Inhalt, Struktur und Qualität des Lehrerwissens zu geben, können Unterrichtsprozesse oder die Leistungsentwicklung von Schülern damit nur sehr unzureichend erklärt werden.

Grundsätzlich wird das Fachwissen (‚content knowledge') aus theoretischer Sicht als Grundvoraussetzung zum Erteilen von Fachunterricht angesehen (vgl. Terhart, 2002). Dabei handelt es sich im Gegensatz zu Alltagswissen, über das in vielen schulrelevanten Bereichen grundsätzlich alle Erwachsenen verfügen sollten, um vertieftes Hintergrundwissen über Inhalte des jeweiligen Curriculums. Ungeklärt ist dabei, ob das Fachwissen als im Lehramtsstudium erworbenes oder als strukturell davon zu unterscheidender Wissensbestand angesehen werden muss. Ergebnisse aus der Berliner COACTIV-Studie deuten jedoch - zumindest für die untersuchten Mathematiklehrkräfte - darauf hin, dass Lehrer, die im Studium bessere Noten erzielten, sowohl ein höheres Fach- als auch ein höheres fachdidaktisches Wissen aufwiesen (Brunner, M. et al., 2006). Die Autoren vermuten aufgrund ihrer Ergebnisse, dass ein umfangreiches (mathematisches) Fachwissen den Erwerb von (ebenfalls mathematikbezogenem) fachdidaktischem Wissen begünstigt. Während das Vorhandensein von Fachwissen jedoch nicht automatisch zu einer für die Schüler effektiven und lehrreichen Vermittlung des-

selben führt, ist genau dies der Ansatzpunkt des fachdidaktischen Wissens (‚pedagogical content knowledge'). Erst durch gutes fachdidaktisches Wissen kann ein Lehrer die Unterrichtsinhalte für die Schüler verständlich vermitteln (vgl. Shulman, 1986).

Für den Erwerb von ‚pedagogical content knowledge' ist eine Umwandlung von Fachwissen in handlungsrelevantes Wissen nötig, was als zentral an der Weiterentwicklung von Lehramtsstudenten angesehen wird (Wilson, Shulman & Richert, 1987). Der Prozess dieser Transformation setzt sich aus vier Subprozessen zusammen, nämlich der kritischen Interpretation (d.h., einen eigenen Standpunkt dem Unterrichtsmaterial gegenüber zu entwickeln), der Repräsentation und der Adaptation (das eigene Repertoire zu erweitern und an das Verständnis der Schüler anzupassen), sowie dem ‚Tailoring' (den Unterrichtsstoff an die Voraussetzungen individueller Schüler anpassen, i.d.S. ‚maßzuschneidern'). In diesem Prozess wird das reine Sachwissen durch Wissen über Schüler, das Curriculum und spezifische Lehrkontexte angereichert. Weitere Bestandteile des ‚pedagogical content knowledge' sind Strategien, Techniken und Prinzipien des Vermittelns bestimmter Inhalte. Durch seine Bindung an spezifische Fachinhalte hebt sich das im Deutschen oft ‚fachdidaktisches Wissen' oder ‚didaktisches Fachwissen' genannte Wissen von dem generellen pädagogischen Wissen ab.

Inwieweit sich das Ausmaß des fachlichen und des fachdidaktischen Wissens von Lehrern direkt auf ihre diagnostische Kompetenz auswirkt, ist bisher nicht untersucht worden. Es gibt jedoch eine Vielzahl von - zum Teil widersprüchlichen - Befunden zu den Effekten auf das Leistungsniveau der Schüler. Ergebnisse aus quantitativen Studien sind aufgrund der oft mangelhaften Operationalisierung des Wissens nicht sehr belastbar. Aus Reanalysen der US-amerikanischen National Educational Longitudinal Study (NELS: 88) ist beispielsweise bekannt, dass die fachbezogene staatliche Zertifizierung besonders im Fach Mathematik tendenziell positiv mit dem Leistungsstand der Schüler zusammenhängt (Goldhaber & Brewer, 1997, 2000). Die Auswirkung von Abschlüssen der Lehrer und die besuchten Kurse sind hingegen uneindeutiger. Hier finden sich vor allem Zusammenhänge in der Sekundarstufe und wiederum vor allem im Fach Mathematik, wohingegen für die Grundschulen mal positive (Druva & Anderson, 1983, für die Naturwissenschaften) und mal negative Effekte (Rowan, Correnti & Miller, 2002, für Mathematik) auftraten. Auch Studien, die nach Zusammenhängen zwischen den in den USA verbreiteten Berufseingangstests (z.B. National Teacher Examination [NTE] oder der Praxis II-Test) und den Fachleistungen

suchten, konnten keine konsistenten Belege für eine prädiktive Validität der Testbatterien finden (Wayne & Youngs, 2006; Wilson & Youngs, 2005). Studien, die sich auf qualitativem Weg dem Fachwissen und fachdidaktischen Wissen von Lehrkräften genähert haben, liefern insgesamt starke Argumente dafür, Untersuchungen zum Unterricht und zur Lehrerkompetenz domänenspezifisch anzulegen (Shulman & Sherin, 2004). Insgesamt deuten die qualitativen Studien, die zum Beispiel über die Analyse des Lehrerverhaltens in kritischen Unterrichtssituationen konzeptionalisiert waren, darauf hin, dass das fachliche Verständnis der unterrichteten Sachverhalte eine notwendige, aber nicht hinreichende Bedingung für einen verständnisorientierten Unterricht darstellt (Baumert & Kunter, 2006; Leinhardt, 2001). Wie Bromme (1995) darlegte, besteht noch viel Unklarheit darüber, was genau unter den Begriffen Fachwissen und fachdidaktischem Wissen zu verstehen ist, um welchen Wissenstyp es sich dabei handelt und inwiefern das vorauszusetzende Wissensniveau von der Schulform, an der ein Lehrer unterrichtet, abhängig gemacht werden sollte.

Baumert und Kunter (2006, S. 489) weisen darauf hin, dass empirisch weder zur Entwicklung noch zur Stabilität diagnostischer Kompetenz belastbare Befunde existieren und ebenso unklar ist, inwiefern die Urteilsgenauigkeit für die Adaptivität des Unterrichts eine notwendige Bedingung darstellt. Ihrer Meinung nach sprechen jedoch Plausibilitätsargumente dafür, „dass Leistungsdiagnostik, die i.d.R. im Handlungsrahmen von Fächern erfolgt, vom fachdidaktischen Wissen der Lehrkräfte abhängig sein könnte“. Da Lehrer nicht erst in Klassenarbeiten oder Tests das Verständnis ihrer Schüler abfragen sollten, sondern in ihrem Unterricht permanent adaptiv auf den Lernstand, das Lerntempo der Schüler eingehen müssen, erscheint es insbesondere für ihr fachdidaktisches Können als eine große Herausforderung, Aufgaben auszuwählen und Arbeitsaufträge zu formulieren, die selbst Ergebnis eines andauernden Diagnoseprozesses sind.

Modell nach Weinert, Schrader und Helmke

Eine alternative Unterteilung der Wissenskomponenten, die sich in vielen Punkten dennoch mit dem Shulman'schen Modell deckt, ist nur wenige Jahre später vorgeschlagen worden. Nach Weinert, Schrader und Helmke (1990a) sowie Weinert, Helmke und Schrader (1992) lassen sich die Erkenntnisse aus der Lehrerexpertiseforschung am besten strukturieren, wenn man eine Unterteilung der notwendigen Wissensanforderungen in Sachwissen, unterrichtsmethodisches Wissen, Klassenführungswissen und diagnos-

tisches Wissen vornimmt. Im Rahmen erfolgreichen unterrichtlichen Handelns, so die Autoren, müssten diese Komponenten miteinander verknüpft werden, so dass die Unterteilung zwar theoretisch berechtigt, praktisch aber eher künstlich erscheint.

- Sachwissen

Das Sachwissen beinhaltet das bereichsspezifische Wissen über den Aufbau und die Prinzipien einer fachlichen Disziplin und ist somit wichtig dafür, den Schülern Informationen verständlich und spezifisch zu präsentieren. Es umfasst ferner sowohl die Kenntnis von Konzepten und Fakten eines Bereichs als auch notwendige Algorithmen und Heuristiken (Leinhardt & Smith, 1985). Bereits die universitäre Ausbildung deutet darauf hin, dass Lehrer ein breites deklaratives Wissen über Fakten, organisierende Prinzipien und zentrale Konzepte der zu unterrichtenden Inhalte haben sollen. Untersuchungen, die nach dem Einfluss des Sachwissens der Lehrer auf die Leistungen ihrer Schüler fragten, führten jedoch nicht zu den erwarteten Ergebnissen, weshalb davon ausgegangen wurden, dass lediglich ein Mindestmaß an Wissen nötig ist, um guten Unterricht zu geben, dass darüber hinausgehendes Mehrwissen jedoch keinen weiteren Einfluss auf die Schülerleistungen hat (Begle, 1972).

- Klassenführungswissen

Das klassenführungsbezogene Wissen stellt die Rahmenbedingungen für eine effektive Stoffvermittlung und die Herstellung einer störungsfreien Unterrichtsatmosphäre dar, bei dem insbesondere prozedurale Wissensbestände zum Tragen kommen, um schnell und zielgerichtet auf verschiedenste Unterrichtseinflüsse reagieren zu können.

- Unterrichtsmethodisches Wissen

Das unterrichtsmethodische Wissen beschreibt die enge wechselseitige Beziehung zwischen Inhalt und Methodik des Unterrichts. Es wird benötigt, um fachspezifische Unterrichtsinhalte mit einer geeigneten Auswahl an Methoden optimal zu vermitteln. Darunter fällt ebenfalls das implizite und explizite Wissen des Lehrers, wie der Unterricht aufgebaut und gestaltet werden sollte, um die angestrebten Unterrichtsziele zu erreichen.

- Diagnostisches Wissen

Das diagnostische Wissen bildet gleichsam die Grundlage für die diagnostische Kompetenz der Lehrer, und während es im Modell von Shulman nur

einen Unteraspekt darstellt, wird es hier prominenter und gleichberechtigt zu den anderen drei Wissenskomponenten angenommen. Nach Weinert, Schrader und Helmke (1990b) umfasst es sowohl das allgemeine Wissen über Schüler bestimmter Alters- und Schulstufen, über deren typische Leistungsfähigkeiten, zu erwartende Stärken oder Schwächen als auch das Wissen über Besonderheiten der eigenen Klasse bzw. eigener Schüler.

In Analysen stellen die Autoren substantielle Einflüsse jeder dieser vier genannten Faktoren auf die Schülerleistungen im (dort untersuchten) Fach Mathematik fest (vgl. Weinert et al., 1990a), wobei die Leistungsheterogenität in der Schulklasse sowie das Vorkenntnisniveau ihrerseits einen deutlichen Einfluss auf die Lehrerkompetenzen haben (vgl. zu den Auswirkungen der diagnostischen Fähigkeiten von Lehrern auf Schülerleistungen auch Kapitel 2.3, S. 25).

Eine Erweiterung des diagnostischen Wissens nehmen Helmke, Hosenfeld und Schrader (2004) in Bezug auf die diagnostische Kompetenz von Lehrern vor, indem sie dem Wissen über Personen (Schüler) jenes über Aufgaben hinzufügen. Begründen lässt sich dies damit, dass gerade die Genauigkeit der Einschätzung der Leistungsfähigkeit stark davon abhängt, ob die Schwierigkeit bzw. die Anforderungen der zugrunde liegenden Aufgaben korrekt erkannt werden. Beides ist - gerade im schulischen Kontext - nicht unabhängig voneinander.

4.6.1.3 Weitere Lehrermerkmale

Neben den beschriebenen professionellen Kompetenzen Expertise und Wissenskomponenten sind eine Reihe weiterer Lehrermerkmale denkbar, die mit der diagnostischen Kompetenz in Zusammenhang stehen könnten, bislang aber nie oder nur sehr selten untersucht wurden. Sie werden im Folgenden eingeführt und lassen sich grob unterteilen in demografische (Geschlecht), berufsbezogene (Lehrdauer in der Klasse) und persönliche Merkmale (Perspektivenübernahmefähigkeit, Perfektionsstreben, Aus- und Weiterbildung, Einstellung zur diagnostischen Kompetenz, Selbstwahrnehmung der eigenen diagnostischen Kompetenz, Schwierigkeiten beim Beurteilen). Letztgenannte persönliche Merkmale lassen sich theoretisch in einem Modell zur Entwicklung professioneller Kompetenz von Lehrkräften verorten, wie es dem Forschungsprogramm COACTIV zugrunde liegt (Kunter, Kleickmann, Klusmann & Richter, 2011). Darin werden persönliche Voraussetzungen wie kognitive Fähigkeiten, Motivation und die Persönlichkeit als

Einflussfaktoren für die Kompetenzentwicklung der Lehrer angesehen. So, wie der Kompetenzaufbau bei Lehrern genauso wie bei Schülern nicht passiv und automatisch verlaufe, sondern auch von der individuellen Nutzung abhänge, wird dies auch für die diagnostische Kompetenz als einer Facette der professionellen Lehrerkompetenz (s. Kapitel 4.6.1.2) angenommen und anhand der genannten Variablen untersucht.

Geschlecht

Während das Geschlecht der Schüler als Moderatorvariable für die Erklärung der Urteilsgüte schon oft untersucht wurde (vgl. Kapitel 4.6.2), spielte das Lehrergeschlecht bislang nur eine untergeordnete Rolle in der Forschung. Dies liegt möglicherweise auch daran, dass in Schulen - und insbesondere in Grundschulen - in der Regel deutlich mehr Frauen als Männer unterrichten. So konnten beispielsweise Südkamp, Kaiser und Möller (eingereicht) in ihrer Metaanalyse das Lehrergeschlecht nicht berücksichtigen, weil in den einzelnen Untersuchungen oftmals entweder gar keine Informationen dazu oder nur das Verhältnis von Männern zu Frauen berichtet wurden oder die Lehrerstichprobe ausschließlich aus Frauen bestand. Empirische Belege dazu, inwiefern sich Frauen und Männer im Lehrerberuf voneinander unterscheiden, liegen entsprechend ebensowenig vor.

Lehrdauer in der Klasse

Personen lassen sich umso besser einschätzen, je länger man sie kennt. Diese Erkenntnis scheint zunächst trivial zu sein, im Kontext Schule ist sie jedoch von großer Bedeutung. Vor dem Hintergrund der häufig stattfindenden Lehrerwechsel, in der Grundschule zumeist nach der zweiten Klasse, macht es möglicherweise einen Unterschied für die Urteilsgenauigkeit, ob ein Lehrer seine Klasse erst vor kurzem übernommen hat und die Schüler noch kennenlernen muss oder ob er schon längere Zeit in der Klasse unterrichtet. Als Bedingungsfaktor für die diagnostische Kompetenz ist dieses Merkmal bislang nicht untersucht worden. Zumindest deutet die Nennung der durchschnittlichen Lehrdauer in der Klasse in einigen wenigen Untersuchungen zur Diagnosegenauigkeit (z.B. Bates & Nettelbeck, 2001; Kenny & Chekaluk, 1993) darauf hin, dass ihr eine Bedeutung in diesem Kontext zugeschrieben wird.

Persönliche Voraussetzungen und Eigenschaften

Die nachfolgend genannten Lehrermerkmale werden im Sinne des Modells der Determinanten der professionellen Kompetenz von Lehrkräften (Kunter et al., 2011, S. 59) als persönliche Voraussetzungen für das Abgeben akkurater Schülerbeurteilungen betrachtet. Es handelt sich dabei keinesfalls um eine vollständige Liste von Merkmalen, sondern lediglich um eine Auswahl, die im Kontext der vorliegenden Arbeit von Bedeutung ist. Die Zusammenstellung ist insofern nicht willkürlich, auch wenn sie durch das Fehlen empirisch belegter Theorien in Teilen einen explorativen Charakter trägt. Dies ist jedoch allein der Tatsache geschuldet, dass die betrachteten Eigenschaften in der Forschungsliteratur zur diagnostischen Kompetenz bislang nicht behandelt wurden und empirische Erkenntnisse dazu entsprechend nicht vorliegen.

Die Fähigkeit der Lehrer, sich in die Sichtweise ihrer Schüler versetzen zu können, kann als Bedingungsfaktor für hohe diagnostische Kompetenz angesehen werden. Unterricht, der sich idealerweise an den Leistungsvoraussetzungen der Schüler orientiert, setzt bei Lehrenden die Fähigkeit voraus, den Unterricht nicht allein von den Inhalten her zu denken, sondern ebenso die Lernvoraussetzungen der Schüler zu berücksichtigen. In diesem Sinne ist die Fähigkeit zur Einbeziehung der Schülerperspektive eine wesentliche Facette fachdidaktischen Wissens (vgl. vorhergehendes Kapitel sowie Borko & Putnam, 2004; Shulman, 1986). In der Schulpraxis sollte eine hohe Fähigkeit zur Perspektivenübernahme daher auch mit hoher diagnostischer Kompetenz einhergehen. Allerdings deutet auch einiges darauf hin, dass es insbesondere Experten mit hohem Fachwissen schwerfällt, sich in die Perspektive von Schülern hineinzuversetzen und mögliche Lernprobleme vorherzusehen, was für verschiedene Unterrichtsfächer nachgewiesen werden konnte (u.a. Herppich, Wittwer, Nückles & Renkl, 2010; Hinds, 1999). Auch für die Eigenschaft, die Arbeit möglichst perfekt zu verrichten, kann angenommen werden, dass sie die Urteilsgenauigkeit positiv beeinflusst. Dabei wird nicht von einem übersteigerten Perfektionsstreben ausgegangen, das oftmals sogar als Risikofaktor für Stress und Burnout angenommen wird (Flett, Hewitt & Hallett, 1995; Schaarschmidt, Kieschke & Fischer, 1999), sondern von dem Bemühen, die Lehrtätigkeit und damit auch die Schülerbeurteilung möglichst ohne Fehler durchzuführen. Ebenfalls positive Effekte lassen sich vom Besuch von Aus- und Weiterbildungsveranstaltungen, die die diagnostische Kompetenz thematisierten, erwarten, da angenommen wird, dass sich Kompetenz durch die aktive Nutzung von Lerngelegenheiten entwickelt und festigt (Kunter et al., 2011).

Einige weitere Merkmale werden zwar nicht direkt als Bedingungsfaktoren diagnostischer Kompetenz angesehen, aber durchaus als Eigenschaften, die mit ihr in Zusammenhang stehen können. Dazu zählen die Bedeutung, die Lehrer diagnostischen Fähigkeiten im allgemeinen zuschreiben (Einstellung zur diagnostischen Kompetenz), die Selbstwahrnehmung der eigenen diagnostischen Fähigkeiten und auch die Schwierigkeiten, die Lehrer für ihre eigenen Urteilsprozesse angeben. Auch wenn dies nicht ursächlich für die Urteilsgüte ist, haben diese Selbstangaben der Lehrer womöglich Erklärungspotential für die diesbezügliche Unterschiedlichkeit von Lehrern. Da auch diesem Aspekt in der Forschung bislang nicht nachgegangen wurde, mangelt es an entsprechenden empirischen Belegen für die aufgestellten Hypothesen.

4.6.2 Schülermerkmale

Nach dem Überblick zu zentralen Lehrermerkmalen, für die ein Zusammenhang zur Güte diagnostischer Urteile theoretisch begründbar erscheint, folgt in diesem Abschnitt die Betrachtung lehrerunabhängiger Einflussmerkmale, für die ebenso eine Beziehung zur Urteilsgenauigkeit angenommen werden kann.

Aus der Forschung ist mittlerweile hinreichend bekannt, dass die Güte von Urteilen nicht nur in erheblichem Ausmaß zwischen Lehrern, sondern auch beim selben Lehrer in Bezug auf verschiedene Schüler deutlich variieren kann (u.a. Coladarci, 1986; Dünnebier, Gräsel & Krolak-Schwerdt, 2009). Es stellt sich somit die Frage, ob oder inwiefern diagnostische Urteile allen Schülern gegenüber gleichermaßen fair ausfallen und ob sie möglicherweise von bestimmten Merkmalen der Schüler oder entsprechenden Stereotypen auf Lehrerseite beeinflusst sind. Gerade Stereotype sind in vielen Studien Gegenstand der Untersuchung gewesen, da die Befürchtung bestand, dass sie durch sich selbst erfüllende Prophezeiungen zu sozialen Problemen führen könnten, wenn sich herausstellte, dass die Personenwahrnehmung und mithin nicht zuletzt die Leistungsbewertung in der Schule durch sie in Bezug auf Geschlecht, Sozialstatus, ethnische Herkunft oder ähnliches systematischen verzerrt würde. Die Mehrheit der Befunde deutet jedoch darauf hin, dass derartige Stereotype zum einen nicht sehr ausgeprägt sind (Judd & Park, 1993; Jussim, McCauley & Lee, 1995; McCauley, 1995) und zum anderen die Wahrnehmung weit weniger beeinflussen als die tatsächlichen Personenmerkmale (Jussim & Eccles, 1995; Madon et al., 1998). Beurteilten Lehrer Schüler aufgrund eines dieser Merkmale schlechter als andere, dann

erwies es sich in der Regel in den genannten amerikanischen Studien auch als realistisch. Allerdings gibt es auch gegensätzliche Befunde. So erwies sich beispielsweise, dass Schüler, deren Benehmen den Lehrern missfiel, unabhängig von ihren tatsächlichen Leistungen von Lehrern schlechtere Bewertungen erhielten als Schüler mit positiv wahrgenommenem Verhalten (Bennett, Gottesmann, Rock & Cerullo, 1993). In einer anderen Untersuchung wurden Leistungen von Schülern mit Behinderung ungenauer eingeschätzt als die anderer Schüler (Hurwitz, Elliott & Braden, 2007). Und die Beeinflussbarkeit von Lehrern zeigte sich in einer Studie von Ritts, Patterson und Tubbs (1992) sogar insoweit, als dass äußerlich attraktivere Schüler bei gleicher Leistung vorteilhaftere Bewertungen erhielten.

Auch in deutschen Untersuchungen wurden häufig Ergebnisse gefunden, die auf eine Beeinflussung von Lehrerurteilen durch bestimmte Schülereigenschaften hindeuten. Besonderes Gewicht erhalten hier diagnostische Urteile vor dem sozioökonomischen Hintergrund und dem Migrationsstatus der Schüler. Forschungsergebnisse zeigten wiederholt, dass die objektiven Schulleistungen von verschiedenen Faktoren wie den individuellen kognitiven Voraussetzungen (Helmke & Weinert, 1997) oder der sozialen Herkunft (Pietsch, 2007; Schwippert, 2007; Schwippert, Bos & Lankes, 2004) beeinflusst werden. Wenn sich zu diesen ohnehin schon „von Hause aus“ benachteiligenden Effekten zusätzlich auch noch eine durch Vorurteile oder Stereotypen begründete Benachteiligung bei der Leistungseinschätzung gesellen sollte, wäre dies eine doppelte Schlechterstellung, beinahe ein umgekehrter Matthäuseffekt. Tatsächlich gibt es Erklärungsansätze, die von einer negativen Beeinflussung der Leistungsentwicklung durch systematische - zum Nachteil gereichende - Schlechterbewertung ausgehen (Hinnant et al., 2009). Im Folgenden soll der Stand der Forschung zum Einfluss der Schülermerkmale Geschlecht, Leistung und soziale Herkunft auf die Genauigkeit von Lehrerurteilen überblicksartig skizziert werden.

Geschlecht

Das Geschlecht von Schülern spielt in der Schule eine zentrale Rolle, und dies nicht nur in Bezug auf den Umgang miteinander, das Einnehmen von Rollen, die Disziplin und ähnliches. Vor allem auch in Hinblick auf die Schulleistungen unterscheiden sich Mädchen von Jungen ganz erheblich. Generell belegen statistische Daten, dass Mädchen in den meisten Bereichen bessere Schulleistungen als Jungen erbringen, dass sie an Gymnasien überrepräsentiert sind und deutlich seltener die Schule ohne Abschluss ver-

lassen (Statistisches Bundesamt, 2009; Stürzer, 2003). Im Gegenzug stellen Jungen an Haupt- und Sonderschulen den größten Anteil.

In jüngerer Zeit haben nationale und internationale Schulleistungsstudien auch immer wieder Geschlechtsunterschiede belegt, die überwiegend in dieselbe Richtung weisen: Mädchen sind im sprachlichen Bereich den Jungen deutlich überlegen, wohingegen Jungen in den mathematischen und naturwissenschaftlichen Bereichen tendenziell bessere Ergebnisse erreichen (Rohrmann, 2007). Konkret erwies sich beispielsweise in der PISA-2000-Studie, dass Jungen eine im Unterschied zu Mädchen besonders schwache Lesemotivation und Schwierigkeiten haben, „Texte und ihre Merkmale kritisch zu reflektieren und zu bewerten" (Stanat & Kunter, 2001, S. 257). Einen noch größeren Vorsprung der Mädchen in sprachbezogenen Bereichen wurde in der DESI-Studie bei Neuntklässlern gefunden, und das besonders bei schriftlichen produktiven Aufgaben sowohl im Deutschen als auch im Englischen. Allerdings schnitten hier die Jungen bei Aussprache und Sprechflüssigkeit im Englischen besser ab (DESI-Konsortium, 2006). Auch IGLU zeigte für den Grundschulbereich die Überlegenheit der Mädchen im sprachlichen Bereich, dies aber deutlicher bei deutschen Schülern als im internationalen Vergleich (Bos et al., 2005; Bos et al., 2003). Schließlich konnte die Hamburger LAU-Studie in ihren längsschnittlichen Untersuchungen von der 5. bis zur 9. Klasse aufzeigen, dass der Vorsprung der Mädchen im Laufe der Schulzeit sogar noch wächst (Stürzer, 2003).

Die beschriebenen Leistungsunterschiede zwischen den Geschlechtern entsprechen mit hoher Sicherheit auch den Erfahrungen der meisten Lehrkräfte. Gerade dann, wenn derartiges Wissen immer wieder rezipiert wird, sei es im eigenen Unterricht oder in der Literatur, besteht die Gefahr, dass es schnell zum Gemeinplatz wird, der verallgemeinert und unreflektiert auf den Einzelfall übertragen wird. Nicht auszuschließen ist darüber hinaus, dass durch dieses Wissen auch Erwartungseffekte (vgl. Kapitel 3.5, S. 46) auf Seiten der Lehrer entstehen, die die geschlechtsspezifischen Leistungsdifferenzen zusätzlich vergrößern. Somit interessierte sich auch die Forschung für den Einfluss des Schülergeschlechts auf die Genauigkeit von Lehrerurteilen.

So wurde beispielsweise in der Metaanalyse von Hoge und Coladarci (1989) auch verglichen, ob sich Unterschiede in der Urteilsgüte in Abhängigkeit vom Geschlecht der Schüler finden ließen. Entsprechende Analysen boten ebenfalls u.a. Helwig, Anderson und Tindal (2001), Jussim und Eccles (1995), Doherty und Conolly (1985), Hoge und Butcher (1984), Demaray und

Elliott (1998) sowie Sharpley & Edgar (1986), doch in keiner dieser Studien zeigten sich Geschlechtseffekte auf die Urteilsgenauigkeit. Auch eine Metaanalyse zu Erwartungseffekten von Lehrern konnte keinen Einfluss des Geschlechts aufdecken (Dusek & Joseph, 1983). In einer weiteren Studie fand man jedoch, dass Lehrer einem indirekten Einfluss des Geschlechts bei der Einschätzung schulischer Kompetenzen unterliegen, indem ihre Einschätzung vom Schülerverhalten auch die Leistungseinschätzung beeinflusst (Bennett et al., 1993). Da Jungen öfter problematisches Verhalten zeigen als Mädchen, wurde demzufolge auch ihre Leistungsfähigkeit als geringer eingeschätzt. Tiedemann (2000) berichtet, dass Lehrer die mathematischen Fähigkeiten von Jungen in der vierten und fünften Klassenstufe etwas höher bewerteten als die der Mädchen, obwohl sich die Leistungen nicht signifikant voneinander unterschieden.

Die IGLU-Studie förderte zutage, dass sich Geschlechtsunterschiede deutlicher bei den Zeugnisnoten als bei den gemessenen Testleistungen zeigten. Bei gleicher Leistung erhielten Mädchen im Durchschnitt die besseren Noten in den Fächern Deutsch und Sachunterricht (Bos et al., 2005). Diesen Befund stützen auch neuere Ergebnisse von Hinnant, O'Brien und Ghazarian (2009), wo ebenfalls die Lesefähigkeiten von Mädchen über-, die der Jungen hingegen von den Lehrern unterschätzt wurden.

Auch in einer aktuelleren Arbeit von Trautwein und Baeriswyl (2007) zeigten sich nicht erwartete Geschlechtereffekte in den Einschätzungen von Lehrkräften. Dabei wurden Jungen bei gleicher Testleistung als kognitiv leistungsfähiger eingeschätzt, Mädchen hingegen erhielten im Vergleich zu Jungen eine günstigere Beurteilung der schulischen Motivation. Die Autoren erklären dies mit geschlechterspezifischen Stereotypen der Lehrer, was auch durch eine Studie von Tiedemann gestützt wird (Tiedemann, 1995), in der Lehrkräfte gute Leistungen bei Mädchen stärker auf Anstrengungen und weniger stark auf Fähigkeiten attribuierten als bei ihren männlichen Mitschülern.

Mehrere Studien deuten darüber hinaus darauf hin, dass der Leistungsrückstand von Mädchen im Fach Mathematik zu einem Großteil vom geringeren Selbstvertrauen der Mädchen erklärt wird. Selbst bei gleich guten Leistungen trauen sich Schülerinnen deutlich weniger zu als ihre männlichen Mitschüler (Moser & Rhyn, 2000). Eine Erklärung hierfür könnte das Lehrerverhalten sein, indem die (unwahre) Vorstellung der Lehrer, Mädchen seien im mathematischen Bereich weniger leistungsfähig als Jungen, unbewusst auf ihr Verhalten den Schülern gegenüber abfärbt und Mädchen somit im Sinne

einer sich selbst erfüllenden Prophezeiung tatsächlich schlechtere Leistungen abliefern.

Die bisherigen Ergebnisse zum Einfluss des Schülergeschlechts auf die Urteilsgüte von Lehrern sind nicht durchgängig konsistent, zumindest in den jüngeren Arbeiten zeigen sich jedoch relativ einheitlich Effekte, die möglicherweise auf eine Stereotypisierung durch die Lehrer hindeuten und somit ohnehin bestehende Leistungsunterschiede noch verstärken.

Schülerleistung

Als ein weiterer Moderator für die Güte diagnostischer Urteile wurde die Schülerleistung selbst vermutet. Coladarci (1986) berichtet von substanziellen Zusammenhängen (je nach Fach zwischen r = .78 und .89) zwischen dem eingesetzten „prozentualen Zustimmungsindex" der Lehrer und den Schülerleistungen. In jedem der vier eingesetzten Subtests lag die Übereinstimmung zwischen Lehrerurteilen und Schülerleistungen für Schüler im untersten Leistungsquartil bei ca. 60 Prozent, im obersten Leistungsquartil hingegen bei 88 Prozent. Allerdings kann dieses Ergebnis auch durch die Art der Übereinstimmungsmessung erklärt werden, denn die Lehrer sollten für jedes zu Grunde liegende Item angeben, ob die jeweiligen Schüler sie im Test korrekt gelöst hatten oder nicht. Gewertet wurde der Anteil an Übereinstimmungen. Bei sehr guten Schülern schätzten Lehrer demzufolge bei vielen Items ein, dass sie korrekt gelöst wurden, und die Schwierigkeit bestand darin, die wenigen nicht korrekt beantworteten Items herauszufinden. Schlechtere Schüler geben hingegen - naturgemäß - mehr falsche Antworten, und hier muss also für mehr Aufgaben herausgefunden werden, welche die falsch beantworteten sind. Dass die Übereinstimmungen von Lehrererwartung und Schülerantwort bei guten Schülern höher sind als bei schlechteren, könnte möglicherweise ein rein stochastisches Phänomen sein.

Auch bei Demaray und Elliott (1998) waren die Lehrerurteile über durchschnittliche bis leistungsstärkere Kinder etwas akkurater als Urteile über leistungsschwächere Schüler, wenn auch bei weitem nicht in dem deutlichen Ausmaß wie bei Coladarci (1986). Dennoch waren auch hier die Unterschiede zwischen leistungsschwachen (r = .56 bzw. 77% Übereinstimmung) und leistungsstarken Schülern (r = .75 bzw. 80% Übereinstimmung) sowohl als Korrelation als auch in prozentualer Übereinstimmung auf Itemebene - bezogen auf die direkten Einschätzungen in der Studie - signifikant.

Hoge und Butcher (1984) konnten nachweisen, dass über das Leistungsniveau hinaus auch die Intelligenz der Schüler einen signifikanten Beitrag (β = 0.18) zur Erklärung der Urteilsgüte beitrug, auch wenn das Leistungsniveau selbst den deutlich größeren Anteil (β = 0.71) hatte. Das Ausmaß des IQ-Einflusses variierte jedoch stark zwischen den untersuchten Lehrern; vier von zwölf Lehrern tendierten jedoch zu einer Überschätzung von intelligenteren Schülern.

Ähnliche Ergebnisse berichten auch Schrader und Helmke (1990), in deren Untersuchung sich zeigte, dass in ihren Mathematikleistungen überschätzte Schüler sich tendenziell durch höhere Intelligenz (sowie ein günstigeres Selbstkonzept) auszeichneten, wohingegen Schüler mit niedrigeren Ausprägungen dieser beiden Merkmale in ihren Leistungen eher unterschätzt wurden.

Sozioökonomischer Hintergrund der Schüler

Immer wieder bestätigte sich - auch international, wenn auch je nach Land unterschiedlich stark ausgeprägt - nicht nur die enge Kopplung von Schulerfolg und sozialer Herkunft (vgl. Bos, Schwippert & Stubbe, 2007), sondern - und dies besonders in Deutschland und dort in der Sekundarstufe - ebenso die Tatsache, dass Kinder aus Familien mit niedrigem Sozialstatus selbst bei gleicher Leistung eine geringere Wahrscheinlichkeit für den Übertritt auf das Gymnasium haben oder generell schlechter eingeschätzt werden. Baumert und Schümer (2001) etwa quantifizieren, dass Kinder der oberen Dienstklasse im Vergleich zu Arbeiterkindern selbst bei Kontrolle von Unterschieden in kognitiven Grundfähigkeiten und in der Lesekompetenz eine ca. dreimal so hohe Chance haben, auf das Gymnasium zu wechseln. Vor dem Hintergrund diagnostischer Kompetenz interessiert deshalb vor allem, ob diagnostische Urteile allen Kindern gegenüber gleichermaßen fair sind, oder ob die Urteilsgüte trotz objektiv gleicher Leistungen von Merkmalen der sozialen Herkunft oder dem Migrationsstatus beeinflusst ist.

International kommen Untersuchungen, die sich dem Verhältnis von Lehrerurteilen und der sozialen Herkunft von Schülern widmeten, zu inkonsistenten Ergebnissen. Während in einigen Studien kein Einfluss des Sozialstatus gefunden wurde (Jussim & Eccles, 1995; Wigfield, Galper, Denton & Seefeldt, 1999), wird er in anderen Artikeln als entscheidender Faktor für Lehrererwartungen herausgestellt (Alexander, Entwisle & Thompson, 1987; Jussim, Eccles & Madon, 1996). So wurden in einer Untersuchung von Alvidrez und Weinstein (1999) in Hinblick auf das Abschneiden in einem In-

telligenztest Schüler mit hohem sozioökonomischen Status über-, Schüler mit niedrigem sozioökonomischen Status hingegen unterschätzt. In eine ähnliche Richtung weisen Ergebnisse von Kennedy (1995), in dessen Studie der Anteil sozial schlechter gestellter Schüler signifikant negativ mit der Fähigkeitswahrnehmung durch die Lehrer zusammenhing.

Die Literaturlage zum Einfluss der ethnischen und sozialen Herkunft der Schüler auf Urteile wie Schulnoten oder Übergangsempfehlungen in Deutschland ist relativ umfangreich, die Erkenntnisse allerdings nur zum Teil homogen. Bereits in den 60er und 70er Jahren des vergangenen Jahrhunderts konnte gezeigt werden, dass Leistungsbewertungen durch Lehrer durch das Wissen um den sozialen Hintergrund der Schüler beeinflusst werden (z.B. Oevermann, Kieper, Rothe-Bosse, Schmidt & Wienskowiski, 1976; Weiß, 1965) und für viele Lehrer bei Übergangsentscheidungen auch leistungsferne Kriterien wie Charaktereigenschaften eine wichtige Rolle spielen (Steinkamp, 1967). An diesen Befunden hat sich auch in neuerer Zeit nicht viel geändert, so dass in einer ganzen Reihe von Studien (Ditton, 1992, 2004, 2010; Lehmann et al., 1997; Schneider, 2009) Verzerrungen von Lehrerurteilen durch sozialspezifische Stereotype sowie deren Einfluss auf Übergangsempfehlungen bestätigt werden konnten. Wiederholt zeigte sich, dass Kinder, die den unteren sozialen Schichten angehören, bezogen auf ihre tatsächlichen Leistungen zu schlecht benotet werden, wohingegen Schüler aus Familien mit hohem Sozialstatus besser benotet werden, als es ihre Leistungen rechtfertigen würden (Ditton, 2010). Um die gleiche Bewertung zu erhalten, müssen sozial benachteiligte Schüler also mehr leisten als ohnehin durch ihre Herkunft privilegierte Schüler.

Stahl (2007) fasst nach gründlicher Analyse von KOALA-S-Daten für den Grundschulbereich zusammen, dass Übergangsempfehlungen zwar in erster Linie durch tatsächliche Leistungsunterschiede der Schüler erklärt werden können, sie weist jedoch auch darauf hin, dass Kinder aus Familien mit niedrigem sozialen Status wegen ihrer schlechteren Leistungen in stärkerem Maße von Erwartungseffekten betroffen sind als andere Kinder. Darüber hinaus lassen sich auch auf Grundlage der KOALA-S-Daten negative Urteilsverzerrungen gegenüber Kindern aus Familien mit niedrigem sozioökonomischen Status nachweisen, wenn auch - gegen Ende der hier untersuchten vierten Klassenstufe - in vergleichsweise geringem Umfang. Die Autorin hält es für möglich, dass dies auch zur Vergrößerung der Leistungsdifferenzen von Kindern unterschiedlicher sozialer Herkunft über die Zeit beiträgt (Stichwort: Schereneffekt).

Zur naheliegenden Vermutung, dass die beschriebenen Effekte in Abhängigkeit von der sozialen Herkunft in ähnlicher Weise auch auf Schüler mit Migrationshintergrund zutreffen, liegen widersprüchliche Befunde vor. Während auf der einen Seite - auf Grundlage der IGLU-Daten - Migrantenkindern eine signifikant niedrigere Wahrscheinlichkeit der Gymnasialempfehlung attestiert wird (Arnold, Bos, Richert & Stubbe, 2007), wurde beispielsweise anhand der LAU-Daten das Gegenteil konstatiert (Lehmann et al., 1997). Die meisten Studien sehen diesbezüglich jedoch gar keine Unterschiede zwischen Deutschen und Migranten (Ditton, Krüsken & Schauenberg, 2005; Kristen, 2002, 2006; Schneider, 2009; Tiedemann & Billmann-Mahecha, 2007).

Sympathie gegenüber den Schülern

Nicht zuletzt gibt es auch Hinweise darauf, dass die Sympathie oder Antipathie den Schülern gegenüber einen Einfluss auf die Genauigkeit und Fairness ihrer Beurteilungen hat. So wurde in einige Studien ein Effekt der Einstellung der Lehrer zu den Schülern vermutet. Einen Hinweis darauf fanden Itskowitz, Navon und Strauss (1988), in deren Untersuchung Lehrkräfte die Selbstwahrnehmung von Schülern ihrer Klasse einschätzen sollten. Wie sich zeigte, variierte die Ungenauigkeit ihrer Einschätzungen als eine Funktion der selbst wahrgenommenen Nähe zu den Schülern. Verglichen mit den Selbsteinschätzungen der Schüler tendierten die Lehrer meist zu einer positiveren Darstellung (Überschätzung) jener Schüler, denen sie sich nahe fühlten, und zu einer Unterschätzung jener Schüler, denen sie neutral oder ablehnend gegenüberstanden. Hierbei spielen möglicherweise Halo-Effekte eine Rolle, bei denen das positiv wahrgenommene Bild der Schüler selbst auf die Bewertung ihrer Leistungen abfärbt.

4.6.3 Klassenmerkmale

Neben Merkmalen der einzelnen Schüler sind ebenso Merkmale gesamter Klassen als Einflussfaktoren auf die Urteilsgüte theoretisch begründbar, von denen nachfolgend einige zentrale beschrieben werden.

Klassengröße

Nur selten ist bislang die Auswirkung der Klassengröße auf die diagnostische Kompetenz von Lehrern untersucht worden. Auf Grundlage eines Teildatensatzes der Marburger Hochbegabtenstudie kamen Wild und Rost

(1995) zu dem Schluss, dass in kleineren Klassen weder milder noch strenger geurteilt wird als in großen und dass sich auch die Genauigkeit von Lehrerurteilen nicht als von der Klassengröße abhängig erweist. Es erscheint logisch, dass Lehrkräfte mit zunehmender Klassengröße mehr Zeit für die Klassenorganisation aufwenden müssen und weniger Zeit für die Zuwendung, Unterstützung und gezielte Diagnose zu einzelnen Schüler bleibt. Dies trifft sich auch mit der Sicht der Lehrer selbst (Bennett, 1996; Pate-Bain, Achilles, Boyd-Zaharias & McKenna, 1992). Empirische Befunde zur Auswirkung der Klassengröße, die größtenteils aus dem angloamerikanischen Raum stammen, sind jedoch uneinheitlich. Während einige Forscher keine statistisch signifikanten Unterschiede im Lehrerverhalten und ihrer Zuwendungszeit zur gesamten Klasse bzw. zu individuellen Schülern je nach Klassengröße fanden (Bressoux, Kramarz & Prost, 2008; Ehrenberg, Brewer, Gamoran & Willms, 2001; Shapson, Wright, Eason & Fitzgerald, 1980) oder konstatierten, dass kleinere Klassengrößen sich eher auf das Engagement der Schüler als auf das Lehrverhalten der Lehrer auswirke (Finn, Pannozzo & Achilles, 2003), wird in einer Reihe anderer Studien auch Gegenteiliges dargestellt. So zeigte beispielsweise Anderson (2000) in einem umfassenden Modell Verbindungen zwischen Klassengröße und Schülerleistung auf, die auch auf das Lehrerverhalten zurückzuführen sind, nämlich größeres Wissen und mehr Engagement der Schüler, tiefergehende Behandlung des Unterrichtsstoffes und auch mehr individuelle Instruktionszeit durch den Lehrer. Letzteres wird auch gestützt durch Ergebnisse von Betts und Shkolnik (1999), die weniger klassenbezogene und mehr individuelle Instruktionen in kleinen Klassen fanden. Die bessere Individualisierung des Unterrichts, das gezieltere Eingehen auf den einzelnen Schüler, scheint somit in kleineren Schulklassen eher möglich zu sein (s. auch Blatchford, Russell, Bassett, Brown & Martin, 2007; Molnar et al., 1999). Eine Schwachstelle der zitierten Untersuchungen besteht darin, dass ihnen keine einheitliche Definition von ‚großen' und ‚kleinen' Klassen zugrunde liegt. So können Klassen mit beispielsweise 15 oder 20 Schülern im Vergleich zu Klassen mit 50 Schülern als klein, im Vergleich zu Klassen mit weniger als 10 Schülern hingegen als groß gelten. Dies erschwert die Interpretation der Ergebnisse erheblich.

Auch wenn die oben geschilderten Befunde zu den Auswirkungen der Klassengröße auf das Lehrerverhalten inkonsistent sind, erscheint es dennoch plausibel anzunehmen, dass gerade durch den größeren Spielraum für die Unterstützung des Einzelnen in kleinen Klassen ebenso die Genauigkeit individueller Lehrerurteile (und in der Summe die Genauigkeit von Einschät-

zungen in Bezug auf die gesamte Klasse) zunimmt. Mangels weiterer Forschungsbefunde in diesem Bereich bleibt die Annahme jedoch spekulativ.

Klassenklima und Unterrichtsstörungen

Bislang unerforscht ist der Einfluss, den das Klassenklima oder das Ausmaß an Unterrichtsstörungen auf die Genauigkeit von Lehrerurteilen haben kann. Der Begriff des Klimas bezieht sich auf die Wahrnehmungen und Beurteilungen von Aspekten des Unterrichts oder des Schüler-Lehrer- bzw. Schüler-Schüler-Verhältnisses (Helmke, 2003). Wie beispielsweise die PISA-Studie zeigte, scheint ein positives Schulklima förderlicher für die Lernleistungen der Schüler zu sein als beispielsweise personelle und materielle Ressourcen oder die jeweilige Schulpolitik (OECD, 2005). Es ist anzunehmen, dass mit besserem Schul- und Klassenklima und niedrigerer Belastung durch Unterrichtsstörungen Lehrer im Unterricht umso motivierter sind und sich intensiver auf die Schüler einlassen können, um Stärken und Schwächen zu erkennen und optimal individuell zu fördern.

5 Fragestellungen

Trotz der bis hierhin beschriebenen Erkenntnisse über das Konstrukt der diagnostischen Kompetenz von Lehrkräften sind nach wie vor viele Fragen ungeklärt. Dies liegt vor allem daran, dass die überwiegende Mehrzahl der Studien querschnittlich angelegt war und somit keinerlei Aussagen über Entwicklungen und Verläufe innerhalb derselben Stichprobe gemacht werden konnten. Und auch die untersuchten Bereiche, auf die sich die Urteile bezogen, waren meist der Mathematik zugeordnet oder konzentrierten sich auf die Leseleistungen von Schülern. Ein Vergleich von Urteilen zu verschiedenen (kognitiven und nicht-kognitiven) Bereichen fand bislang kaum statt, so dass sich vorhandene Erkenntnisse in aller Regel auf den Vergleich von Ergebnissen aus unterschiedlichen Studien stützen. Dabei schränken die mitunter sehr unterschiedlichen methodischen Herangehensweisen sowie die Betrachtung verschiedener Klassenstufen, Schulsysteme etc. diese Vergleiche enorm ein.

In der vorliegenden Studie können derartige Limitierungen vermieden werden, indem durch längsschnittliche Erfassung diagnostischer Lehrerurteile Aussagen zu ein- und derselben Grundschulstichprobe in einer Vielzahl unterschiedlichster Urteilsbereiche ermöglicht werden. Eine weitere Stichprobe aus dem Grundschulbereich mit ansonsten gleichem Erhebungskontext erlaubt eine Überprüfung der Hauptergebnisse in einer anderen Klassenstufe. Im Fokus stehen dabei die nachfolgend genannten Fragestellungen, die sich überwiegend auf die Struktur diagnostischer Urteile sowie die Bedingungen der Urteilsgüte konzentrieren.

5.1 Struktur der diagnostischen Kompetenz von Grundschullehrkräften

Der erste Komplex an Fragestellungen widmet sich der Struktur diagnostischer Lehrerurteile und deren Güte bzw. Genauigkeit. Während zur Urteilsgüte an sich schon zahlreiche Ergebnisse vorliegen, ist zu strukturellen Aspekten der Urteile, insbesondere über verschiedene Urteilsbereiche hinweg, kaum geforscht worden.

Struktur

An den Anfang soll die Frage nach der Struktur von Lehrerurteilen über verschiedene Leistungsbereiche hinweg im Vergleich zur Struktur der entsprechenden Schülermerkmale gestellt werden. Davon werden Erkenntnisse erwartet, die Aufschluss über das Zustandekommen der Urteilsakkuratheit bzw. von Urteilsverzerrungen geben können. Voraussetzung dafür, dass Lehrerurteile auch über verschiedene Leistungsbereiche hinweg akkurat ausfallen, ist eine hohe Übereinstimmung von Lehrerurteilen und Schülermerkmalen unter Berücksichtigung der jeweiligen Bereichsspezifika. Wer gute Deutschleistungen zeigt, muss eben noch lange nicht auch in Mathematik leistungsstark sein, da unterschiedliche Bereiche jeweils spezifische Anforderungen an die Schüler stellen, die nicht immer gleichermaßen gut bewältigt werden können.

Hypothetisch sollte die Struktur der Lehrerurteile und die Stärke ihrer Zusammenhänge nicht deutlich von jener der Schülerleistungen und -merkmale abweichen. Vielmehr wäre idealerweise zu erwarten, dass Lehrkräfte in ihren Urteilen zwischen Bereichen differenzieren, die auch in der Realität der Schüler unterschiedliche Leistungsaspekte darstellen. Bei den Analysen hierzu wird noch nicht die im Zentrum dieser Arbeit stehende Urteilsgenauigkeit betrachtet, sondern es werden zunächst nur die Zusammenhänge von Urteilen zu verschiedenen Leistungen und Eigenschaften im Vergleich zu den Zusammenhängen der entsprechenden Schülerausprägungen. Damit einhergehend ist von Interesse, wie stabil die Lehrerurteile im Vergleich zu den Schülerleistungen und -eigenschaften über die drei Messzeitpunkte hinweg sind, ob Lehrer also die (möglicherweise bereichsspezifische) Veränderlichkeit von Schülerleistungen und -eigenschaften bei ihren Beurteilungen berücksichtigen. Akkurate Einschätzungen vorausgesetzt, sollten die Stabilität der Schülerwerte und jene der Lehrerurteile in etwa die gleiche Höhe aufweisen.

Güte

Mittlerweile gibt es zur Güte diagnostischer Urteile umfangreiche Forschungserkenntnisse (vgl. besonders Kapitel 4.6). In verschiedensten Studien mit ganz unterschiedlichen Designs zeigte sich immer wieder, dass es bezüglich der Rangkomponente diagnostischer Kompetenz mittelhohe Übereinstimmungen zwischen Leistungen von Schülern und den Leistungseinschätzungen ihrer Lehrkräfte gibt und dass das Leistungsniveau von den Lehrkräften in aller Regel überschätzt wird. Urteile in Bezug auf

nicht-kognitive (emotional-motivationale) Merkmale der Schüler fallen in der Regel deutlich ungenauer aus als hinsichtlich kognitiver Fähigkeiten. Im Sinne einer Prüfung der Passung der vorliegenden Daten auf den Forschungsstand soll auch in dieser Arbeit - zunächst deskriptiv - geprüft werden, welche Güte diagnostische Urteile von Grundschullehrkräften besitzen. Dabei soll jedoch ein direkter Vergleich der Urteilsgenauigkeiten in Abhängigkeit vom zu beurteilenden Bereich vorgenommen werden. Was bisherige Untersuchungen in aller Regel nicht bieten, ist eine umfassende Betrachtung der Urteilsgüte über verschiedene Leistungsbereiche hinweg. In der vorliegenden Studie werden zudem Leistungsmerkmale (u.a. aus den Bereichen Arithmetik, Wortschatz und Textverstehen) um Merkmale aus dem emotional-motivationalen Bereich (u.a. Leistungsängstlichkeit und fachspezifisches Interesse der Schülerinnen und Schüler) ergänzt und hinsichtlich der Urteilsgenauigkeit betrachtet. Hierbei wird entsprechend der vorhandenen bisherigen Forschungsergebnisse vermutet, dass Rangurteile der Lehrer auch in dieser Untersuchung im Mittel in allen Bereichen mittelhohe Zusammenhänge mit den entsprechenden Ausprägungen auf Schülerseite aufweisen. Unterschiede zwischen den Fächern werden insofern vermutet, als dass in besonders gut beobachtbaren Bereichen eine höhere Urteilsgenauigkeit zu finden ist als in nur indirekt schlussfolgerbaren Bereichen. So sollte im arithmetischen Bereich, wo sich die Rechenfähigkeit direkt im Lösungsverhalten bei Rechentests niederschlägt, genauere Urteile möglich sein als beispielsweise im Wortschatz, der zwar in jeder Äußerung der Schüler zum Ausdruck kommt, aber nie direkt abgeprüft wird. Auch hinsichtlich der Niveaueinschätzung wird ein zur Forschungslage passendes Bild der tendenziellen Leistungsüberschätzung erwartet. Dies lässt sich aufgrund der methodischen Restriktionen in dieser Arbeit allerdings nur in der Zusatzstichprobe prüfen. Da Einschätzungen im nicht-kognitiven Bereich aber weniger eng an der täglichen Unterrichtspraxis liegen und einen Randaspekt von Beurteilungen darstellen, wird angenommen, dass die Rangurteilsgüte in den kognitiven Bereichen höher ausfällt als in den emotional-motivationalen Bereichen.

Homogenität

Gleichsam als Verbindung des Vergleichs der Urteilskompetenz in verschiedenen Bereichen mit der ersten Fragestellung nach der Struktur stellt sich die Frage, ob die Güte der Urteile von Lehrkräften in diesen unterschiedlichen einzuschätzenden Bereichen miteinander zusammenhängt, also homogen ist, oder ob die Urteilsgenauigkeit heterogen und somit bereichsspe-

zifisch ist. Dieser Aspekt ist in der langen Tradition der Erforschung der diagnostischen Kompetenz erst relativ spät aufgekommen. Die wenigen diesbezüglichen Befunde (vgl. Spinath, 2005) legen die Vermutung nahe, dass es sich bei der diagnostischen Kompetenz um eine bereichsspezifisch ausgeprägte Fähigkeit der Lehrkräfte handelt und somit hinsichtlich der Übereinstimmung von Lehrereinschätzung und Merkmalsausprägung zwischen verschiedenen Bereichen keine signifikanten Zusammenhänge bestehen. Inhaltlich ähnliche Bereiche wie die dem sprachlichen Bereich zuzuordnenden Leistungen im Wortschatz und im Textverstehen, in denen vermutlich auch die Schülerleistungen hoch miteinander korrelieren, sollten vom selben Lehrer hingegen auch ähnlich gut eingeschätzt werden. Dem liegt die Annahme zugrunde, dass Lehrer um die Ähnlichkeit der Anforderungen in inhaltlich verwandten Bereichen wissen, ihnen demzufolge auch die anzunehmende hohe Kovarianz der Schülerleistungen in diesen Bereichen bewusst ist und sie somit für diese Bereiche zu vergleichbaren Urteilen kommen. Bei inhaltlich getrennten Bereichen (z.B. Leistungsfacetten aus der Mathematik und dem Fach Deutsch) ist diese enge Übereinstimmung der Urteile nicht zu erwarten, wenn man diagnostische Kompetenz nicht als eine Universalfähigkeit annimmt. Eine hohe Urteilshomogenität geht allerdings nicht zwangsläufig mit einer hohen Urteilsgüte einher, sondern kann sich ebenso aus zwei gleichermaßen unzutreffenden Urteilen ergeben.

Stabilität

Obwohl die Frage, ob es sich bei der diagnostischen Kompetenz um ein stabiles Persönlichkeitsmerkmal handelt, zentral für das Verständnis von diesem Konstrukt ist, hat sich die Forschung bislang so gut wie gar nicht diesem Punkt gewidmet. Dafür notwendige längsschnittliche Studiendesigns gab und gibt es nur sehr selten. Offenbar schien es selbstverständlich zu sein, dass diagnostische Kompetenz etwas ist, das Lehrkräfte in einer gewissen Qualität besitzen und das nach dem allgemein üblichen Verständnis vom Kompetenzbegriff kaum Veränderungen über die Zeit unterliegt. Diese Annahme gilt es allerdings erst noch zu belegen, und deshalb ist sie ein weiterer Gegenstand dieser Arbeit. Indem Weinerts Definition von „Kompetenz“ (Weinert, 1999b) zugrunde gelegt wird, ist die Hypothese, dass sich die Güte der Rangeinschätzungen auf Ebene der einzelnen Lehrer in den verschiedenen Bereichen innerhalb eines halben bzw. ganzen Jahres nicht deutlich verändert. Ein akkurat einschätzender Lehrer sollte auch nach einem halben oder ganzen Jahr noch zum Abgeben treffender Urteile in der Lage sein, wohingegen von einem ‚schlechten‘ Diagnostiker nach dieser Zeit

zwar zutreffendere Urteile wünschenswert, aber auch nicht unbedingt zu erwarten sind.

Reliabilität

In der Literatur sind seit vielen Jahren Zweifel daran nachzulesen, ob oder inwiefern es sinnvoll erscheint, die diagnostische Kompetenz von Lehrern an psychologischen Gütekriterien zu messen. Damit wurde aber stets entweder gemeint, dass Lehrerurteile nicht zwangsläufig objektiv sein müssen (u.a. Weinert & Schrader, 1986), oder es wurde vor einer „Psychologisierung" des pädagogischen Handlungsfelds Schule durch den häufigen Einsatz standardisierter Testverfahren gewarnt (Ingenkamp & Lissmann, 2008). Kann darüber hinaus aber einfach davon ausgegangen werden, dass es sich bei der Rang- und der Niveaukomponente diagnostischer Kompetenz auch um ein reliables und damit valides Urteilsmaß handelt? In keiner der unzähligen Untersuchungen zur diagnostischen Kompetenz wurde - soweit dem Autor bekannt - überprüft, ob sich die Gütemaße der Rang- und Niveaukomponente als reliables, schülerunabhängiges Maß zeigen. In jüngster Zeit wird gehäuft auf diese Forschungslücke hingewiesen (Lorenz & Artelt, 2009; McElvany et al., 2009; Schrader, 2009). Soll angenommen werden, dass es sich bei der diagnostischen Kompetenz um ein von Lehrerfähigkeiten abhängiges Maß handelt, wäre eine hohe Reliabilität die Bedingung dafür. Daher wird in der vorliegenden Arbeit mittels eines Split-half-Reliabilitätstests geprüft, inwiefern die Urteilsgenauigkeit der Lehrer für eine Klassenhälfte mit der Urteilsgenauigkeit für die andere Klassenhälfte vergleichbar ist. Wenn die diagnostische Kompetenz als eine Personenfähigkeit verstanden werden soll, müsste sie unabhängig davon sein, welche Schüler der eigenen Klasse beurteilt werden. Eine leistungsbezogene Rangfolge der Schüler (Rangkomponente) und die individuelle Leistungsniveaueinschätzung (Niveaukomponente) sollte sich hypothetisch gleichermaßen für beide - zufällig festgelegten - Klassenhälften zeigen.

5.2 Bedingungen der diagnostischen Kompetenz von Grundschullehrkräften

Seit empirisch belegt ist, dass nicht alle Lehrer gleichermaßen treffende Leistungsdiagnosen zu stellen in der Lage sind, beschäftigt sich die Forschung mit der Suche nach den Ursachen dafür. In dieser Arbeit soll die Liste der potentiellen Einflussfaktoren im Vergleich zu früheren Untersuchungen deutlich erweitert und darüber hinaus nicht nur nach Merkmalen von

Lehrern selbst, sondern auch der Klasse und einzelner Schüler differenziert werden. Während bei Lehrermerkmalen die Annahme ist, dass Eigenschaften der beurteilenden Personen selbst auf die Urteilsgenauigkeit Einfluss haben, wird bei den lehrerexternen Variablen vermutet, dass bestimmte Klassenzusammensetzungen bzw. individuelle Schülereigenschaften die Genauigkeit von Beurteilungen beeinflussen. Es wird angenommen, dass sich auf jeder dieser drei Ebenen Merkmale zeigen, die einen Einfluss auf die Güte der Lehrerurteile haben.

Lehrermerkmale

Aus der Literatur ist bereits bekannt, dass es zwischen Lehrern große interindividuelle Unterschiede in Bezug auf die Güte ihrer diagnostischen Einschätzungen gibt. Dabei wurde auch immer wieder versucht, den Ursachen für diese Unterschiede auf den Grund zu gehen bzw. beeinflussende Faktoren in der Lehrperson selbst zu finden. So wurde beispielsweise vermutet, dass die Qualität der Urteile mit längerer Berufserfahrung zunimmt (u.a. Wild & Rost, 1995), doch diese unikausalen Annahmen zu wenigen einzelnen Merkmalen führten bislang nicht zu dem gewünschten Erkenntnisfortschritt. In der vorliegenden Untersuchung sind neben den oft betrachteten Lehrervariablen wie der Berufserfahrung und dem Geschlecht eine Reihe weiterer, bislang nicht beleuchteter Merkmale erfasst worden, die sich unter anderem auf ihre Aus- und Weiterbildung und schul- und unterrichtsbezogene Facetten erstrecken und von denen angenommen werden kann, dass sie im Zusammenhang mit besonderer Lehrerexpertise und somit hoher diagnostischer Kompetenz stehen. Auf dieser Grundlage soll geprüft werden, ob sich einzelne dieser Lehrermerkmale als Bedingungsmerkmale guter Diagnostiker herausstellen lassen. Nicht als Bedingung, aber als damit verwandter Aspekt wird schließlich zusätzlich geprüft, ob die Selbsteinschätzung der Lehrer in Hinblick auf ihre diagnostischen Fähigkeiten mit der gemessenen Urteilsakkuratheit korrespondiert.

Klassenmerkmale

Neben Einflussfaktoren auf die Güte diagnostischer Urteile auf Lehrerebene kann vor dem Hintergrund anderer Forschungsergebnisse ebenfalls davon ausgegangen werden, dass Klassenmerkmale die Urteilsgüte von Lehrern beeinflussen. Auch wenn der oft vermutete Einfluss der Klassengröße auf die Urteilsgenauigkeit bislang nicht bestätigt werden konnte, so wird sie - neben der Anzahl einzuschätzender Schüler - dennoch auch in dieser Un-

tersuchung eine der betrachteten Variablen sein. Eine damit zusammenhängende Vermutung ist, dass die Urteilsgüte der Lehrer nicht davon abhängt, wie viele Kinder sie normalerweise in der Klasse zu unterrichten haben, sondern für wie viele Schüler sie in unserer Befragung Einschätzungen vornehmen sollten. Es wird angenommen, dass Lehrer, die viele individuelle Einschätzbögen auszufüllen hatten, insgesamt weniger genau urteilten, vielleicht ermüdeten oder unmotivierter waren. Die Anzahl der Kinder als unabhängige Variable wird ergänzt durch das Leistungsniveau und durch die Leistungsstreuung in der Klasse. Die Anforderung, die Schüler entsprechend ihrer kognitiven und nicht-kognitiven Attribute in eine Rangfolge zu bringen, sollte umso besser gelingen, je deutlicher sich die Schüler hinsichtlich dieser Attribute voneinander unterscheiden. Befinden sich in einer Klasse leistungs- und merkmalsmäßig viele ähnliche Schüler, so geschieht es leicht, dass sich beispielsweise Lehrer verschätzen, indem sie stärker zwischen den Schülern differenzieren wollen, als es der Realität entspricht, oder die im Test gemessenen Leistungen zweier ansonsten gleich guter Schüler unterscheiden sich aufgrund von Tagesformunterschieden, was der Lehrer in seinem auf längerer Erfahrung und Kenntnis der Schüler beruhenden Urteilen nicht einbeziehen kann und somit zu einem niedrigeren Wert diagnostischer Kompetenz kommt. In heterogenen Klassen hingegen sollten sich die Schüler in der Rangreihe mit größerer Sicherheit platzieren lassen, und explizite Differenzierungsbemühungen der Lehrer treffen auf eine tatsächliche größere Bandbreite an Kompetenzen und nicht-kognitiven Eigenschaften.

Als ein weiteres Klassenmerkmal wird auch der Anteil von Schülern mit Migrationshintergrund (gemessen an der elterlichen Herkunft) betrachtet. Aus dem Migrantenanteil lässt sich nicht automatisch ein Einfluss auf Leistungsniveau, Unterrichtsqualität oder Urteilsgenauigkeit der Lehrer ableiten, aber es gibt Zusammenhänge, die dahingehende Vermutungen stützen. So ist beispielsweise anzunehmen, dass sich mit zunehmendem Ausländeranteil in der Klasse auch der Anteil derjenigen Schüler erhöht, die erhebliche Sprachverständnisprobleme im Deutschen haben und somit nicht nur ihr eigenes Lernen erschweren, sondern auch die Unterrichtsgestaltung durch den Lehrer für die gesamte Klasse negativ beeinflussen (vgl. Tiedemann & Billmann-Mahecha, 2004). Dies könnte in der Folge auch zu ungenaueren Lehrerurteilen führen, da durch den Einsatz kompensatorischer Maßnahmen weniger Zeit für den „normalen“ Unterricht mit allen Schülern bleibt. Denkbar ist jedoch auch, dass die Lehrerurteilsgüte durch einen hohen Mig-

rantenanteil in der Klasse steigt, da er die Differenzierbarkeit der Leistungen in der Klasse erhöht.

Darüber hinaus wird vermutet, dass sich ein angenehmes Klassenklima sowie ein geringes Ausmaß an Unterrichtsstörungen und Zeitverschwendung positiv auf die mittlere Urteilsgüte der Lehrkräfte auswirken, weshalb auch diese Merkmale in die Untersuchung einfließen. Während man bei Klassen, für die die Lehrer ein gutes Klassenklima attestiert haben, davon ausgehen kann, dass sie sich somit enger mit ihrer Klasse verbunden fühlen und dies einen positiven Einfluss auf ihre Urteilsgenauigkeit hat, verhält es sich beim Ausmaß der Unterrichtsstörungen und der Zeitverschwendung theoretisch umgekehrt: je mehr Zeit im Unterricht verschwendet wird, desto weniger Gelegenheit haben die Lehrer, die Leistungen und Eigenschaften ihrer Schüler in reinen Unterrichtssituationen wahrzunehmen, und durch den somit entstehenden Nachteil gegenüber Kollegen mit weniger Zeitverschwendung kann angenommen werden, dass die Güte ihrer Einschätzungen weniger genau ist.

Schülermerkmale

Auf individueller Ebene liegen Merkmale der Schüler selbst, für die eine Beeinflussung der ebenfalls individuell erteilten Lehrerurteile vermutet werden kann. Dabei ist insbesondere das Geschlecht der Schüler von Interesse, für das entsprechend früherer Forschungsbefunde eine je nach Leistungsbereich unterschiedliche Auswirkung auf die Güte der Lehrerurteile - für sprachbezogene Leistungsbereiche eher zu Gunsten der Mädchen, für den mathematischen Bereich eher zu Gunsten der Jungen - angenommen wird. In den emotional-motivationalen Bereichen ist hingegen kein Unterschied in der Urteilsgenauigkeit je nach Geschlecht zu vermuten. Auch soll geprüft werden, ob Wechselwirkungen zwischen dem Lehrer- und dem Schülergeschlecht bestehen, die bspw. in einer vorteilhafteren Bewertung von Jungen durch männliche Lehrkräfte Ausdruck finden könnten. Darüber hinaus soll untersucht werden, inwiefern sich die soziale Herkunft der Schüler auf die Güte der Einschätzungen durch ihre Lehrer auswirkt. Wie bereits mehrfach gezeigt wurde, spielt der Sozialstatus der Kinder eine wichtige Rolle bei verschiedenen Formen von diagnostischen Urteilen. Am häufigsten konnte ein Einfluss der Sozialschicht auf die Übergangsempfehlung in die Sekundarstufe derart nachgewiesen werden, dass Kinder aus unteren sozialen Schichten für die gleiche Empfehlung bessere Leistungen erbringen müssen als Schüler oberer sozialer Schichten (vgl. Baumert & Schümer, 2001; Ditton &

Krüsken, 2006). Doch nicht nur bei den wegweisenden Übertrittsentscheidungen, sondern auch bei der tagtäglichen Notenvergabe im Unterricht zeigte sich, dass Schüler aus sozial schwachen Elternhäusern, gemessen an ihren tatsächlichen Leistungen, deutlich zu schlecht, Kinder aus der mittleren und vor allem der oberen Sozialgruppe hingegen zu gut benotet werden (Ditton, 2010). Die Vermutung, dass derartige Verzerrungen auf Stereotypisierungen durch die Lehrer infolge einer sozialschichtspezifischen Attribuierung von Begabung zurückzuführen sind, lassen es plausibel erscheinen, dass auch notenunabhängige diagnostische Urteile insofern einer Beeinflussung durch den Sozialstatus der Kinder unterliegen, als dass Kinder mit niedrigem Sozialstatus im Vergleich zu sozial besser gestellten Mitschülern eher in ihren Leistungen unterschätzt bzw. weniger genau eingeschätzt werden. Weiterhin soll als individuelles Schülermerkmal das Leistungsniveau in die Analysen einbezogen werden, da in Anlehnung an frühere Ergebnisse von Hoge und Butcher (1984) angenommen wird, dass bessere Schüler über ihren tatsächlichen Leistungsvorsprung hinaus noch besser von den Lehrern eingeschätzt werden. Und um schließlich der Frage nachzugehen, ob auch persönliche Sympathien eine Rolle im Beurteilungsprozess spielen, soll geprüft werden, ob sich Schüler, die sich von ihren Lehrern besonders angenommen fühlen, auch in den Beurteilungen einen Vorteil haben. Hypothetisch wird davon ausgegangen, dass höhere Sympathien des Lehrers für einzelne Schüler im Sinne eines Halo-Effekts (vgl. Kapitel 3.5) positive Auswirkungen auf die Leistungsbeurteilung haben, Urteile also genauer werden oder sogar zu einer Bevorteilung gegenüber weniger sympathisch gefundenen Schülern führen.

Für die Prüfung des Einflusses von Lehrer-, Klassen- und Schülermerkmalen auf die Güte und Genauigkeit diagnostischer Lehrerurteile findet eine Eingrenzung der untersuchten Leistungs- und Merkmalsbereiche auf unterrichtsrelevante Leistungsbereiche statt. Im Unterschied zu den Analysen zu Struktur, Güte, Homogenität und Stabilität diagnostischer Urteile wird nun die Auswahl interessierender Bereiche deshalb um das logisch-abstrakte Denken sowie die emotional-motivationalen Schülermerkmale Lernfreude, Schuleinstellung und Leistungsängstlichkeit reduziert. Zudem sind die Analysen zur Niveaukomponente diagnostischer Kompetenz ohnehin nur für die Leistungsbereiche möglich, wie in Kapitel 6.3 beschrieben wird.

5.3 Vergleichbarkeit von Leistungseinschätzungen und Zeugnisnoten

Noten sind im schulischen Kontext die für die Schüler entscheidendsten Manifestationen von Lehrerurteilen. Über die „Fragwürdigkeit der Zensurengebung“ (Ingenkamp, 1995b) ist schon viel geschrieben worden, mit dem immer gleichen Tenor, dass gleiche Leistungen von verschiedenen Lehrern teils sehr unterschiedlich bewertet werden. Gleiches ist von Einschätzungen, die zu Forschungszwecken erhoben wurden, bekannt (u.a. Hoge & Coladarci, 1989). Nun umfassen Zensuren in der Regel mehr als Einschätzungen in einem wissenschaftlichen Fragebogeninstrument (vgl. Kapitel 4.2, S. 57) und sind in ihrer praktischen Bedeutung für den Schüler ungleich wichtiger; dennoch liegen beiden Formen des Lehrerurteils Ansichten über die Leistungsfähigkeit der Schüler zugrunde, die möglichst objektiv und fair sein sollen und somit auch der Gegenüberstellung mit objektiven Leistungstests standhalten müssen. Einschätzungen für die interessierten Forscher sollten daher genauso zutreffend sein wie Zensuren, weshalb von einem hohen Zusammenhang zwischen beidem ausgegangen wird.

In der vorliegenden Arbeit wird der Zusammenhang von Lehrerurteilen (sowohl in Form von Fragebogeneinschätzungen als auch in Form von Zeugnisnoten) mit den gemessenen Schülerleistungen in verschiedenen Leistungsbereichen überprüft. Im Zentrum steht dabei die Frage, ob Zeugnisnoten in der Relation zu den gemessenen Testleistungen die gleiche Güte aufweisen wie die - für die Schüler bedeutungslosen - Fragebogenurteile und in welchem Ausmaß beide Urteilsformen miteinander zusammenhängen. Weiterhin ist von Interesse, inwiefern gleiche Testleistungen über verschiedene Klassen hinweg von den Lehrern auch gleich bewertet werden. Die hierzu vorliegenden Forschungsergebnisse lassen vermuten, dass auch in der hier verwendeten Stichprobe gleiche Leistungen je nach Kontext unterschiedlich beurteilt werden und somit ein objektiver klassenübergreifender Vergleich von Noten und Einschätzungen stark eingeschränkt wird.

5.4 Ergänzende Fragestellungen anhand von Daten zur Klassenstufe 1

Die bis hierher formulierten Fragestellungen werden mit einer im nächsten Kapitel beschriebenen, längsschnittlich verfolgten Stichprobe von Dritt- und Viertklässern sowie ihren Klassenlehrern zu beantworten versucht. Zum einen ist man damit - wie in jeder Untersuchung - in der Reichweite der Aus-

sagekraft selbst bei hoher Repräsentativität auf diesen Klassenstufenbereich eingeschränkt, zum anderen treten einige methodische Limitationen auf, die in Kapitel 6.3 näher beschrieben sind. Diese beiden Effekte können möglicherweise etwas gemindert bzw. die Aussagekraft der Ergebnisse erhöht werden, indem eine Überprüfung ausgewählter Ergebnisse anhand einer zweiten Stichprobe vorgenommen wird. Zu diesem Zweck steht eine Gruppe von Erstklässlern samt ihrer Lehrer zur Verfügung, die in weiten Teilen ähnliche Instrumentarien bearbeitet hat wie die Dritt- und Viertklässler und deren Lehrer. Mit ihr kann geprüft werden, inwiefern die verschiedenen Urteilsgüten zwischen Klasse 1 und Klasse 3 und 4 miteinander vergleichbar sind und welche Akkuratheit andere Urteilskomponenten, die in der Hauptstichprobe nicht gebildet werden konnten, aufweisen. Darüber hinaus wird es mit dieser Stichprobe möglich sein zu prüfen, inwiefern sich Operationalisierungsvarianten zur Niveaukomponente diagnostischer Kompetenz aus der Hauptstichprobe bestätigen lassen, wenn man sie mit korrekt erfassten Niveauurteilen vergleicht.

Es wird erwartet, dass die Analysen mit der Zusatzstichprobe nicht zu wesentlich anderen Ergebnissen führen als die der Hauptstichprobe, da es sich - wenngleich in einer niedrigeren Klassenstufe und mit anderen Lehrkräften - dennoch um Schüler und Lehrer aus der Grundschule handelt. Die Prüfung der Operationalisierung der Niveaukomponente in der Hauptstichprobe sollte deren Korrektheit bestätigen.

6 Methodisches Vorgehen

Die Beschreibung der Untersuchungsmethoden erfolgt in diesem Kapitel entsprechend der üblichen Unterteilung nach Stichprobe, Untersuchungsdesign und Instrumenten. Im Anschluss daran wird genauer auf die Bildung geeigneter Indikatoren zur Prüfung der Fragestellungen eingegangen und das Problem fehlender Werte sowie der gewählte Umgang damit geschildert. Da im Rahmen dieser Arbeit zwei verschiedenen Schülergruppen mit ihren jeweiligen Lehrern Gegenstand der Analysen sind, wird auch bei der Darstellung des methodischen Vorgehens dahingehend differenziert. Zunächst wird immer auf die Hauptstichprobe eingegangen, die längsschnittlich vom zweiten Halbjahr der dritten bis zum zweiten Halbjahr der vierten Klassenstufe untersucht wurden. Daran schließt sich jeweils die Beschreibung der Zusatzstichprobe an, bei der es sich um querschnittlich untersuchte Erstklässler mit ihren Lehrern handelt. Inhaltlich liegt der Schwerpunkt auf der Hauptstichprobe, so dass sie auch in den Methodenbeschreibungen ausführlicher behandelt wird.

6.1 Stichproben

Beide Stichproben wurden im Rahmen des DFG-geförderten Bamberger BiKS-Projektes[5] („Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor- und Grundschulalter“) rekrutiert. Die Hauptstichprobe entspricht dabei dem sogenannten Längsschnitt BiKS-8-12, in dem Kinder vom Alter von acht Jahren an (3. Klasse) bis in die Sekundarstufenzeit hinein untersucht werden, die Zusatzstichprobe ist hingegen der Längsschnitt BiKS-3-8, an dessen Erhebungen Kindergartenkinder ab 3 Jahren teilnehmen und der sich bis in die Grundschulzeit erstreckt. Dabei wurde auf eine möglichst hohe Repräsentativität geachtet. Weiterhin wurde unter anderem eine regionale Fokussierung vorgenommen, indem zum einen Erhebungen in zwei Bundesländern - Bayern und Hessen - durchgeführt werden. Dies hat den Vorteil, dass verschiedenen institutionellen und politischen Rahmenbedingungen, zum Beispiel bezüglich der Einschulungs- und Übergangsregelungen oder den Schulsystemen, Rechnung getragen werden kann. Zum anderen wurden aber auch innerhalb beider Bundesländer verschiedene Regionen einbezogen, so dass eine hohe Varia-

[5] Für eine ausführliche Stichprobenbeschreibung siehe von Maurice et al. (2007).

bilität soziostruktureller Kontextfaktoren bestand. Die Regionen unterscheiden sich untereinander zum Teil stark hinsichtlich Bevölkerungsdichte, Wirtschaftsstruktur, Arbeitsmarktlage und dem Anteil von Bürgern mit Migrationshintergrund, sind jedoch jeweils zwischen den Bundesländern relativ vergleichbar. Tabelle 3 gibt einen Überblick über die einbezogenen städtischen und ländlichen Erhebungsregionen.

Tabelle 3: Erhebungsregionen für beide Stichproben

Region	*Bayern*	*Hessen*
Großstädtische Region	Nürnberg	Frankfurt am Main
Städtische Region	Bamberg Stadt	Darmstadt
Ländliche Region 1	Landkreis Bamberg	Landkreis Bergstraße
Ländliche Region 2	Landkreis Forchheim	Landkreis Odenwald

Repräsentativität

Die Auswahl der jeweiligen Schulen in den Regionen ergab sich in Form einer verbundenen Stichprobe auf Grundlage der Auswahl der Kindergärten aus dem Längsschnitt BiKS-3-8. Es wurden solche Schulen ausgewählt, auf die die Kindergartenkinder aus dem Längsschnitt BiKS-3-8 in aller Regel wechseln. Die Rekrutierung der Kindergärten wiederum basierte auf einer mehrfach geschichteten Zufallsstichprobe nach folgenden fünf Kriterien:

- Bundesland: 60% der Kindergärten sollten in Bayern, 40% in Hessen liegen.
- Region: je ein Drittel der bayerischen und hessischen Kindergärten sollte aus Großstädten stammen.
- Migrationshintergrund: in den Großstädten sollte jeweils ein Drittel der Kindergärten einen niedrigen (unter 10%), einen mittleren (zwischen 10% und 50%) und einen hohen (über 50%) Anteil an Kindern mit Migrationshintergrund aufweisen.
- Gruppenzahl: die Zahl der Gruppen in den jeweiligen Kindergärten sollte proportional berücksichtigt werden.
- Zahl der Einmündungsschulen: bei 90% der Kindergärten sollten die Kinder in der Regel nur auf eine Grundschule überwechseln, bei den übrigen 10% sollte sich der Wechsel in der Regel auf drei oder mehr Grundschulen verteilen.

Die jeweils betroffenen Grundschulen bildeten somit die Grundlage für den Längsschnitt BiKS-8-12. Da die Kindergärten und Schulen jedoch der Beteiligung an der BiKS-Studie zustimmen mussten und die Ausschöpfungsquote hierfür insgesamt bei den Kindergärten bei 76,0 Prozent und bei den Schulen bei nur 67,8 Prozent lag, ist mit einer leichten Verzerrung der angestrebten Auswahl zu rechnen.

6.1.1 Hauptstichprobe

Im Folgenden wird zunächst auf die Eigenschaften der Hauptstichprobe eingegangen.

Erhebungszeitpunkte

Die vorliegende Arbeit bezieht sich überwiegend auf Daten aus der Grundschulzeit, die zu drei Messzeitpunkten erhoben wurden. Der Zeitpunkt der Erhebungen wurde jeweils derart gewählt, dass sie in beiden Bundesländern nicht mit den Ferienterminen kollidierten und alle Testungen eines Erhebungszeitpunkts möglichst eng beieinander lagen. Dass dies nicht immer gleichermaßen gut gelang, wird aus Tabelle 4 ersichtlich, in der dargestellt ist, wie viele Tage vor Ende des jeweiligen Schulhalbjahres die Erhebung im Mittel (sowie frühestens und spätestens) stattfand. Zu beachten ist, dass die Schulhalbjahre in Bayern und Hessen immer zu unterschiedlichen Zeitpunkten endeten; deshalb sind die Werte nicht nur insgesamt, sondern auch getrennt nach Bundesland dargestellt. Während zum ersten Messzeitpunkt die Erhebung im Mittel noch deutlich über hundert Tage vor Ende des Schuljahres stattfand, rückte sie in den folgenden beiden Wellen näher an den letzten Schultag im Halbjahr heran. In Hessen fanden die Erhebungen zudem stets näher am Halbjahresende statt als in Bayern, was bei der Interpretation der Leistungsdaten berücksichtigt werden muss. Ein weiteres Problem stellt die hohe Streuung dar, die sich nicht nur zwischen den Bundesländern, sondern auch innerhalb zeigt. Insgesamt lagen zwischen der frühesten und der spätesten Erhebung je Welle zwischen 71 und 90 Tage, aber auch innerhalb eines Bundeslandes lagen maximal 75 Tage zwischen erstem und letztem Erhebungstermin.

Tabelle 4: Verteilung der Abstände zwischen den Erhebungszeitpunkten in den Klassen und dem letzten Schultag im jeweiligen Halbjahr

		früheste Erhebung	*späteste Erhebung*	*Mittelwert*	*Standardabweichung*
Bayern	t1	139	115	126	8
	t2	103	67	91	10
	t3	88	45	72	13
Hessen	t1	124	49	105	16
	t2	81	47	67	10
	t3	61	17	42	12
Gesamt	t1	139	49	118	15
	t2	103	47	82	15
	t3	88	17	60	19

Anmerkung: Alle Werte geben den Abstand zwischen Testung und dem letzten Schultag des jeweiligen Schulhalbjahres in Tagen an.

Die hohe Varianz zwischen Klassen hinsichtlich des Zeitpunktes im Schuljahr, zu dem die Testung stattfand, führt zu der Frage, inwiefern dadurch die gemessenen Leistungsdaten beeinflusst wurden. Zu vermuten wäre, dass eine später getestete Klasse gegenüber einer früh getesteten Klasse durch mehr Unterricht, der in der Zwischenzeit stattgefunden hat, einen Leistungsvorsprung erzielen konnte. Dies wäre ein Problem für die Interpretation der Leistungsvariablen, da sie über Klassen hinweg nicht mehr vergleichbar wären. In Tabelle 5 ist dargestellt, inwieweit es einen Zusammenhang zwischen dem Testzeitpunkt (operationalisiert über die Tage, die zwischen Testtag und letztem Schultag im Halbjahr liegen) und den gemessenen Kompetenzen gibt. Die Daten sind getrennt nach Bundesland dargestellt, da sich die Bundesländer systematisch sowohl hinsichtlich des Testzeitpunkts im Schulhalbjahr als auch hinsichtlich der mittleren Schülerkompetenzen unterscheiden. Obwohl einige Korrelationen signifikant ausfallen, ist die Höhe der Zusammenhänge als unbedenklich anzusehen. Zu erwarten wären außerdem negative Korrelationen gewesen, die anzeigen, dass die Leistungen umso niedriger ausfallen, je früher sie erhoben wurden. Stattdessen sind die meisten der signifikanten Korrelationen positiv.

Tabelle 5: Zusammenhang zwischen dem Abstand des Testzeitpunktes vom Schulhalbjahresende und den Schülerleistungen

	Bayern			*Hessen*		
	t1	t2	t3	t1	t2	t3
Arithmetik	.06*	.09**	-.04	.09*	-.01	-.01
Wortschatz	.03	.07*	.00	.03	.02	-.02
Textverstehen	.05*	.08**	.06*	-.06	-.02	-.09*
Rechtschreibung	-.01	-	.00	-.04	-	.01
log.-abstr. Denken	-.00	-	-.01	.02	-	.04

**p < .01, *p < .05

Geschlechterverteilung der Schülerinnen und Schüler in der Studie

Nach Vorlage aller Schul- und Elterngenehmigungen ergab sich eine Größe der Ausgangsstichprobe von insgesamt 2395 Schülerinnen und Schülern. Der Anteil der Jungen betrug dabei insgesamt 52,2 Prozent, in Bayern 51,9 Prozent und in Hessen 52,8 Prozent. Auch nach der Verkleinerung der Stichprobe durch diverse Ausfälle in den tatsächlichen Erhebungen bleibt der Anteil der Jungen insgesamt und in den Bundesländern bei ca. 52 Prozent stabil. In Tabelle 6 sind die absoluten Häufigkeiten der an den Erhebungen teilnehmenden Jungen und Mädchen dargestellt.

Tabelle 6: Entwicklung der Schülerstichprobengröße über die drei Messzeitpunkte hinweg

	Größe der Ausgangsstichprobe		*N zu t1*		*N zu t2*		*N zu t3*	
	m	w	m	w	m	w	m	w
Bayern	807	749	771	718	740	698	705	643
Hessen	443	396	412	375	381	363	352	332
Gesamt	1250	1145	1183	1093	1121	1061	1057	975
Gesamt (m + w)	2395		2276		2182		2032	
Prozent der Ausgangsstichprobe			94,6%	95,5%	89,7%	92,7%	84,6%	85,2%

m = männlich, w = weiblich

Stichprobenausfall

Ebenfalls in Tabelle 6 ist zu erkennen, dass über die drei Messzeitpunkte ein kontinuierlicher Stichprobenausfall zu verzeichnen ist. Bereits zum Zeit-

punkt der ersten Erhebung haben rund 5 Prozent weniger Kinder tatsächlich teilgenommen als entsprechend der Ausgangsstichprobe zu vermuten war. Auch zum zweiten und dritten Messzeitpunkt reduzierte sich die Stichprobengröße um jeweils ca. 5 Prozent, wobei der Ausfall bei den Jungen etwas größer war als bei den Mädchen. Hauptsächlich ist der Ausfall auf Krankheit einzelner Schüler am Testtag zurückzuführen, zweimal nahmen ganze Klassen nicht mehr an der Studie teil (vgl. Tabelle 8).

Alter der Schüler

Die Schülerinnen und Schüler der Hauptstichprobe befanden sich zum ersten Messzeitpunkt im zweiten Halbjahr der dritten Klassenstufe. Die zum ersten Messzeitpunkt teilnehmenden Schüler waren damals im Mittel 111 Monate (9;3 Jahre) alt mit einer Standardabweichung von 5,6 Monaten. Das jüngste Kind war zum ersten Erhebungszeitraum gerade einmal 88 Monate alt, das älteste hingegen 140 Monate. Dies deutet darauf hin, dass sich sowohl extrem früh eingeschulte Kinder in der Stichprobe befinden als auch Kinder, die mindestens einmal eine Klasse wiederholen mussten. Die hessischen Kinder waren im Schnitt etwas älter als die bayerischen bei einer ebenfalls höheren Streuung, und in beiden Bundesländern waren Jungen etwas älter als die Mädchen (vgl. Tabelle 7).

Tabelle 7: Alter der Schüler zu t1, getrennt nach Geschlecht und Bundesland sowie insgesamt

	Alter der Jungen zu t1 in Monaten		*Alter der Mädchen zu t1 in Monaten*		*Alter der Jungen und Mädchen zu t1 in Monaten*	
	M	SD	M	SD	M	SD
Bayern	111,1	5,7	110,1	5,1	110,6	5,4
Hessen	112,7	6,3	110,8	5,4	111,8	6,0
Gesamt	111,7	5,9	110,3	5,2	111,0	5,6

M = Mittelwert, SD = Standardabweichung

Klassen

Die Schülerinnen und Schüler wurden in ihren jeweiligen Klassenkontexten getestet und befragt. Da jedoch nicht alle Eltern einer Teilnahme ihres Kindes zustimmten, konnte je Klasse nur ein Teil der Schüler an der Erhebung teilnehmen, während die restlichen Schüler währenddessen in einem anderen Raum anderweitig beschäftigt wurden. Tabelle 8 gibt Auskunft über die Anzahl teilnehmender Schulen und Klassen zu den drei Messzeitpunkten, zur durchschnittlichen Klassengröße und der durchschnittlichen Anzahl tat-

sächlich an der Untersuchung teilnehmender Kinder. Für diese Gruppe wird außerdem der Anteil an Migranten (mindestens ein Elternteil ist nicht in Deutschland geboren) berichtet. Die Angaben zur Klassengröße stammen von den Klassenlehrern, die zum Migrationshintergrund von den Eltern. Die Daten zur Anzahl der Schulen und Klassen beziehen sich auf die tatsächliche Testung und sind unabhängig davon, ob jede Klasse später in die Analysen einfließt. Bedingt durch fehlende Lehrerangaben oder eine zu geringe Anzahl eingeschätzter Kinder werden einige Klassen nicht in allen Analysen berücksichtigt.

Zu t2 und t3 hat jeweils eine gesamte Schulklasse die weitere Teilnahme an der Studie verweigert, die Anzahl der beteiligten Schulen ist jedoch gleich geblieben. Während die mittlere Klassengröße über die Wellen relativ stabil geblieben ist, reduzierte sich sowohl die mittlere Teilnahequote in den Klassen als auch der Anteil teilnehmender Migranten. Konnten zu Beginn der Studie noch durchschnittlich 14,7 Kinder je Klasse (65% Beteiligung) getestet und befragt werden, reduzierte sich diese Zahl bis zur dritten Welle auf durchschnittlich 13,3 (59% Beteiligung), wie in Tabelle 8 abzulesen ist. Dies entspricht einem Rückgang der Beteiligungsquote um ca. 6 Prozent. Gleichzeitig verkleinerte sich der Anteil teilnehmender Migranten von t1 zu t3 um ca. 15 Prozent.

Tabelle 8: Schülerzahl, Migrantenanteil und Teilnahmequote der teilnehmenden Klassen zu den drei Messzeitpunkten, getrennt nach Bundesland sowie insgesamt

		Anzahl Schulen	*Anzahl Klassen*	*durchschnittliche Anzahl Kinder pro Klasse insgesamt (SD)**	*durchschnittliche Anzahl an BiKS teilnehmender Kinder pro Klasse (SD)*	*durchschnittliche Anzahl Kinder mit Migrationshintergrund in der Gruppe der teilnehmenden Kinder (SD / %)*
t1	Bayern	51	97	23,6 (3,7)	15,4 (4,2)	3,3 (3,2 / 23,2%)
	Hessen	31	58	21,2 (3,6)	13,6 (4,4)	4,9 (3,1 / 38,5%)
	Gesamt	82	155	22,7 (3,8)	14,7 (4,4)	3,9 (3,2 / 28,9%)
t2	Bayern	51	96	23,5 (3,4)	15,0 (4,2)	3,0 (2,9 / 22,8%)
	Hessen	31	58	21,3 (3,4)	12,8 (4,4)	4,5 (3,0 / 38,9%)
	Gesamt	82	154	22,7 (3,6)	14,2 (4,4)	3,6 (3,0 / 28,8%)
t3	Bayern	51	96	23,5 (3,6)	14,0 (4,1)	2,8 (2,6 / 22,6%)
	Hessen	31	57	21,0 (3,5)	12,0 (4,3)	4,2 (3,0 / 37,9%)
	Gesamt	82	153	22,6 (3,8)	13,3 (4,3)	3,3 (2,8 / 28,3%)

*Diese Angaben stammen von den Lehrern und beziehen sich auf die gesamten Klassen.

Durch die relativ geringe Teilnahmequote je Klasse stellt sich die Frage, inwiefern der Ausfall zufällig war bzw. inwieweit die teilnehmenden Kinder noch repräsentativ für die gesamte Klasse waren. Anhand der Zeugnisnoten, die durch die Angabe eines Notenspiegels durch die Lehrer jeweils für die gesamte Klasse bekannt sind, soll dies beleuchtet werden. Dazu werden die Zeugnisnoten der teilnehmenden Schüler in den Fächern Deutsch und Mathematik in einem t-Test für unabhängige Stichproben mit den nicht zuzuordnenden Noten aus dem Notenspiegel verglichen. Die durch Lehrerangaben bekannten individuellen Noten der teilnehmenden Schüler wurden von den vorliegenden Klassenspiegeln abgezogen, die übrig bleibende Notenverteilung aus den Notenspiegeln der Gesamtklassen bildeten die Werte für die nicht teilnehmenden Schüler.

Dabei zeigt sich sowohl für jeden der drei Messzeitpunkte als auch für beide Unterrichtsfächer, dass der Notendurchschnitt der teilnehmenden Kinder signifikant besser ist als der Durchschnitt jener Kinder, die von ihren Eltern keine Genehmigung zur Teilnahme erhielten (Tabelle 9). Der mittlere Unterschied beträgt dabei in etwa eine halbe Notenstufe.

Zu berücksichtigen ist jedoch, dass die angegebenen Werte nur eine Annäherung sind, denn einige Lehrer gaben sowohl für individuelle teilnehmende Kinder als auch für gesamte Klassen keine Zeugnisnoten bzw. keinen Notenspiegel an. Der Anteil der Missings für die Noteninformationen lag bei den teilnehmenden Schülern je nach Zeitpunkt und Fach zwischen 5,7 und 17,9 Prozent, in Bezug auf die Notenspiegel für die Gesamtklassen zwischen 10,9 und 25,6 Prozent. Für die vorliegende Berechnung wurden diese Missings ignoriert und nur mit Klassen gerechnet, bei denen sowohl individuelle als auch klassenbezogene Notenangaben vorlagen.

Tabelle 9: Unterschiede in den Zeugnisnoten für Deutsch und Mathematik zwischen an der Untersuchung teilnehmenden Schülern und ihren nicht teilnehmenden Klassenkameraden

	Deutsch: Notenmittelwert BiKS-Kinder (SD; N)	*Deutsch: Notenmittelwert andere Kinder (SD; N)*	*t*	*Mathematik: Notenmittelwert BiKS-Kinder (SD; N)*	*Mathematik: Notenmittelwert andere Kinder (SD; N)*	*t*
t1	2,63 (0,92; 1542)	2,88 (0,96; 686)	-5,8	2,39 (0,93; 1568)	2,66 (1,03; 640)	-5,8
t2	2,59 (0,93; 1482)	3,04 (1,01; 881)	-10,8	2,48 (0,91; 1511)	2,83 (0,99; 921)	-8,7
t3	2,65 (0,91; 1645)	3,02 (0,99; 1109)	-10,0	2,58 (0,99; 1648)	2,95 (1,05; 1115)	-9,2

Anm.: Es wurden nur jene Klassen in die Analyse einbezogen, in denen sowohl Angaben zu den Noten der teilnehmenden Schüler als auch der Notenspiegel der Klasse vorlagen. Alle angegebenen Mittelwertsunterschiede erweisen sich beim t-Test für unabhängige Stichproben als signifikant ($p < .001$). In Klammern sind jeweils die zugrunde liegenden Fallzahlen angegeben.

Somit scheinen die an der Untersuchung teilnehmenden Schüler nicht repräsentativ für die Gesamtklassen zu sein. Offenbar gehören leistungsschwächere Schüler mit höherer Wahrscheinlichkeit zur Gruppe der Teilnahmeverweigerer, was bei der Interpretation der Ergebnisse zur diagnostischen Kompetenz berücksichtigt werden muss, denn die Lehrereinschätzungen beziehen sich dadurch überdurchschnittlich häufig auf die leistungsstärkeren Schüler.

Klassenlehrerwechsel

Eine längsschnittliche Untersuchung in Schulen ist oftmals mit dem Problem des Lehrerwechsels konfrontiert. Die diesbezüglichen Bestimmungen und Regelungen unterscheiden sich von Bundesland zu Bundesland. Vor allem in Bayern findet in der Regel zwischen der zweiten und dritten Klassenstufe ein Klassenlehrerwechsel statt, über den der Schulleiter unter Abwägung der schulischen Gegebenheiten entscheidet. Wie sich anhand der vorliegenden Stichprobe zeigt, finden Klassenlehrerwechsel jedoch zu einem bedeutsamen Anteil auch zum Ende des dritten Schuljahres statt (Tabelle 10). In Bayern ist dies in knapp dreißig Prozent der untersuchten Klassen der Fall, in Hessen mit fünf Prozent der Fälle eher die Ausnahme. In beiden Bundesländern findet zu einem geringen Anteil außerdem ein Klassenlehrerwechsel innerhalb des vierten Schuljahres statt, in durchschnittlich drei Vierteln aller Klassen unterrichtet jedoch in den untersuchten drei Halbjahren derselbe Klassenlehrer.

Tabelle 10: Übersicht über Klassenlehrerwechsel nach Messzeitpunkt und Bundesland

*Teilnahme**			*Bayern*		*Hessen*		*insgesamt*	
t1	t2	t3	Häufigkeit	Prozent	Häufigkeit	Prozent	Häufigkeit	Prozent
A	A	A	65	67,0	51	87,9	116	74,8
A	B	B	28	28,9	3	5,2	31	20,0
A	A	B	3	3,1	4	6,9	7	4,5
A	B	C	1	1,0	-	-	1	0,6

*Die Buchstaben A, B und C stehen für jeweils unterschiedliche Lehrer zu den drei Messzeitpunkten (t1, t2 und t3).

Die Klassenlehrerwechsel haben Auswirkungen auf die Analysen. Bei querschnittlichen Berechnungen werden alle zum Messzeitpunkt teilnehmenden Lehrer einbezogen, bei längsschnittlichen Analysen erscheint es sinnvoller, sich auf jene Lehrer zu beschränken, die zu jedem relevanten Messzeitpunkt auch tatsächlich in der Klasse unterrichtet haben.

In Tabelle 11 wird die nach Bundesland getrennte und gemeinsam berechnete Zusammensetzung der Lehrerschaft für jeden der drei Messzeitpunkte hinsichtlich einiger demografischer und berufsbezogener Merkmale beschrieben. Es fällt auf, dass in Hessen die Quote an Klassenlehrerinnen deutlich höher ist als in Bayern, insgesamt liegt sie bei etwas über 80 Prozent. Dieses Bild steht im Widerspruch zu aktuellen - vom Statistischen Bundesamt stammenden - Daten aus dem Jahresgutachten 2009 des Aktionsrats Bildung (Blossfeld et al., 2009), in dem das Bundesland Hessen im Bundesvergleich mit über 22 Prozent den höchsten Anteil an männlichen Grundschullehrkräften aufweist, wohingegen in Bayern nur ca. 14 Prozent aller Grundschullehrkräfte Männer sind, was in etwa dem gesamtdeutschen Durchschnitt entspricht. Männliche Lehrer scheinen in der vorliegenden Stichprobe demzufolge unterrepräsentiert zu sein. Ein weiterer bundeslandspezifischer Aspekt scheint der Anteil der auch Mathematik unterrichtenden Klassenlehrkräfte zu sein. Er liegt in Bayern bei ca. 95 Prozent, in Hessen je nach Messzeitpunkt ungefähr zehn Prozentpunkte niedriger. Für das Durchschnittsalter der Lehrer hingegen lassen sich genau wie hinsichtlich der Berufserfahrung, der Quote der Deutsch unterrichtenden Klassenlehrer oder in Bezug auf das Unterrichtsdeputat keine Unterschiede zwischen Bayern und Hessen feststellen. Ebenso verändern sich die Lehrermerkmale im Mittel trotz Lehrerwechseln zwischen den Messzeitpunkten kaum.

Tabelle 11: Merkmale der Klassenlehrer zu den drei Messzeitpunkten, getrennt nach Bundesland

		Anteil Klassenlehrerinnen	*durchschnittliches Geburtsjahr*	*durchschnittliche Berufsjahre an einer Grundschule*	*Anteil Klassenlehrer, die Deutsch in dieser Klasse unterrichten (%)*	*Anteil Klassenlehrer, die Mathematik in dieser Klasse unterrichten (%)*	*mittleres Unterrichtsdeputat pro Woche in Stunden*
t1	Bayern	77,4	1960	15,4	97,8	95,7	22,7
	Hessen	92,3	1960	15,5	98,1	84,6	23,8
	Gesamt	82,8	1960	15,4	97,9	91,7	23,1
t2	Bayern	75,8	1958	17,8	100,0	95,7	22,6*
	Hessen	94,2	1959	17,1	100,0	88,0	18,3*
	Gesamt	82,3	1958	17,5	100,0	93,1	22,2*
t3	Bayern	75,8	1960	17,9	100,0	94,6	17,5*
	Hessen	94,2	1960	16,2	100,0	81,1	24,3*
	Gesamt	82,3	1960	17,3	100,0	89,7	20,9*

*das Unterrichtsdeputat zu t2 und t3 wurde nur bei den neuen Lehrern erfragt und wird somit auch nur für sie berichtet

6.1.2 Zusatzstichprobe

Für ausgewählte Analysen fand in dieser Arbeit eine weitere Gruppe von Schülern Berücksichtigung, nämlich ein Teil der Stichprobe aus dem BiKS-3-8-Längsschnitt. Dies sind jene Kinder, die in der BiKS-Studie seit ihrem dritten Lebensjahr begleitet werden und im Jahr 2009 die erste Grundschulklasse besuchten. Da auch die Lehrkräfte dieser Kinder um diagnostische Einschätzungen gebeten und gleichzeitig Testergebnisse erfasst wurden, ist der Vergleich zur Hauptstichprobe in der dritten und vierten Jahrgangsstufe möglich und aufschlussreich. Dennoch kommt dieser Gruppe von Schülern in der vorliegenden Arbeit nur der Status einer Zusatzstichprobe zu.

Die Zusatzstichprobe ist deutlich kleiner als die Hauptstichprobe und besteht aus insgesamt 822 Kindern, wobei die Teilnahmequote in Bayern mehr als doppelt so hoch war wie in Hessen und in beiden Bundesländern mehr Mädchen als Jungen partizipierten (s. Tabelle 12). Ursprünglich war die Ausgangsstichprobe dieses Längsschnitts deutlich größer, doch nach dem Schuleintritt konnte nur noch ein Teil dieser Kinder weiterverfolgt werden.

Tabelle 12: Schülerstichprobengröße (Zusatzstichprobe aus Klassenstufe 1)

	männlich	weiblich	Gesamt
Bayern	270	296	566
Hessen	110	146	256
Gesamt	380	442	822

Wie in Tabelle 13 abzulesen, waren die Kinder zum Erhebungszeitpunkt in der ersten Klasse durchschnittlich 89,6 Monate (ca. 7,5 Jahre) alt. Wie in der Hauptstichprobe sind die hessischen Kinder etwas älter als die bayerischen und Jungen etwas älter als Mädchen.

Tabelle 13: Alter der Schüler (Zusatzstichprobe)

	Alter der Jungen in Monaten		Alter der Mädchen in Monaten		Alter der Jungen und Mädchen in Monaten	
	M	SD	M	SD	M	SD
Bayern	88,7	4,6	88,5	4,8	88,6	4,7
Hessen	91,6	4,4	90,0	4,1	90,6	4,3
Gesamt	89,5	4,7	88,9	4,6	89,2	4,3

Die Ausschöpfung in den Schulklassen ist deutlich schlechter als in der Hauptstichprobe. In Bayern nahmen im Mittel nur 4,4 Schüler pro Klasse teil, also 20 Prozent, in Hessen waren es durchschnittlich 8,6 Schüler pro Klasse, also etwa 40 Prozent (s. Tabelle 14). Dies ist auf die Schwierigkeiten zurückzuführen, die es sowohl beim Nachverfolgen der früheren Kindergartenkinder nach dem Schuleintritt als auch bei der Akquise der Klassenkameraden, die bis dahin nicht an der Untersuchung teilgenommen hatten, gab. Es ist anzunehmen, dass die verbliebenen Schüler nicht vollkommen repräsentativ für die Klassen sind. Eine dahingehende Überprüfung, wie sie für die Hauptstichprobe anhand der Verteilung der Zeugnisnoten in den gesamten Klassen im Vergleich zu denen der teilnehmenden Kinder stattfand (vgl. Tabelle 9), ist hier mangels Notenangaben durch die Lehrer nicht möglich. Zumindest hinsichtlich des Anteils von Kindern mit Migrationshintergrund in den Gesamtklassen scheint es große Ähnlichkeit zur Hauptstichprobe zu geben, denn die Werte von durchschnittlich 4,4 Schülern mit Migrationshintergrund in Bayern und 8,6 in Hessen entsprechen recht genau denen der Dritt- und Viertklässlerstichprobe (vgl. Tabelle 8).

Tabelle 14: Schülerzahl, Migrantenanteil und Teilnahmequote (Zusatzstichprobe)

	Anzahl Schulen	Anzahl Klassen	durchschnittliche Anzahl Kinder pro Klasse (SD)	durchschnittliche Anzahl an BiKS teilnehmender Kinder pro Klasse (SD)	durchschnittliche Anzahl Kinder mit Migrationshintergrund in der Gruppe der teilnehmenden Kinder (SD / %)
Bayern	59	90	22,1 (3,5)	6,3 (4,8)	4,4 (5,4)
Hessen	31	56	21,2 (2,6)	4,6 (4,1)	8,6 (6,3)
Gesamt	90	146	21,8 (3,2)	5,6 (4,6)	5,9 (6,0)

Die Lehrkräfte der Erstklässler sind zum Erhebungszeitpunkt zirka fünf Jahre jünger als ihre Kollegen aus der Hauptstichprobe und haben entsprechend etwas weniger Berufsjahre aufzuweisen (s. Tabelle 15). Das mittlere Unterrichtsdeputat dieser Lehrer liegt in beiden Bundesländern etwas über dem der Hauptstichprobe.

Tabelle 15: Merkmale der Lehrer (Zusatzstichprobe)

	Anteil Lehrerinnen	durchschnittliches Geburtsjahr	durchschnittliche Berufsjahre an einer Grundschule	Anteil Lehrer, die Deutsch in dieser Klasse unterrichten (%)	Anteil Lehrer, die Mathematik in dieser Klasse unterrichten (%)	mittleres Unterrichtsdeputat pro Woche in Stunden
Bayern	95,1	1963	15,6	96,8	96,8	24,6
Hessen	84,8	1966	11,1	100,0	85,3	24,6
Gesamt	91,5	1965	13,4	97,9	92,8	24,6

Die bis hierher angegebenen deskriptiven Werte der Schüler, Klassen und Lehrkräfte beziehen sich auf all jene Schüler, die zu diesem Messzeitpunkt an der Erhebung teilgenommen haben. Insbesondere aufgrund der sehr niedrigen Beteiligung in manchen Klassen ist für die späteren Analysen u.a. zur Rangkomponente diagnostischer Kompetenz die Stichprobe nochmals reduziert worden, nämlich auf jene Klassen, in denen mindestens fünf Schüler an den Leistungstests teilgenommen haben und zu denen gleichzeitig Lehrereinschätzungen vorlagen.

6.2 Eingesetzte Erhebungsinstrumente

Im Folgenden werden die eingesetzten Instrumente hinsichtlich jener Tests, Skalen und Items, die für die vorliegende Arbeit relevant sind, genauer beschrieben. Da diese Arbeit im Rahmen des BiKS-Projekts mit seinen vielfältigen Fragestellungen aus verschiedenen Wissenschaftsdisziplinen entstan-

den ist, gehören zum vollständigen Instrumentarium eine Vielzahl weiterer Skalen, auf die hier jedoch nicht weiter eingegangen wird.

6.2.1 Leistungstests für die Schüler

6.2.1.1 Leistungstests für die Schüler der Hauptstichprobe

Dieser Arbeit liegen für die Hauptstichprobe drei Messzeitpunkte zugrunde. Die erste Erhebung fand im Frühjahr 2006 statt, als sich die Schülerinnen und Schüler am Ende der dritten Klassenstufe befanden. Mit ungefähr halbjährlichem Abstand folgten dann Messzeitpunkt 2 (Mitte bis Ende 1. Halbjahr Klassenstufe 4) und 3 (Mitte bis Ende 2. Halbjahr Klassenstufe 4). Der halbjährliche Abstand wurde bewusst gewählt, um die in diesem Altersbereich massiven Entwicklungen der Kinder möglichst genau abbilden zu können. Zugleich fanden die Erhebungen immer gegen Ende der Schulhalbjahre statt, um den empfohlenen Einsatzzeitraum der eingesetzten Testverfahren zu treffen.

Um in der festgelegten Testzeit von drei Unterrichtsstunden je Erhebungszeitpunkt möglichst viele Kompetenzfacetten der Schüler erfassen zu können, wurde die Stichprobe von vornherein in zwei annähernd gleich große Substichproben geteilt, wobei jeder Hälfte ein anderes Kompetenztestheft zugewiesen wurde. Es bekamen immer ganze Klassen dieselbe Testheftversion vorgelegt, und dies zu jedem der drei Messzeitpunkte. Die enthaltenen Tests bildeten zwischen den Heftversionen eine große Schnittmenge, einige Testverfahren wurden jedoch nur von jeweils einer Hälfte der Stichprobe bearbeitet.

Tabelle 16: Administrationsdesign der eingesetzten Leistungstests zu den drei Messzeitpunkten

	t1	*t2*	*t3*
	2. Halbjahr 3. Klasse	*1. Halbjahr 4. Klasse*	*2. Halbjahr 4. Klasse*
Testheftversion 1	Wortschatz (CFT 20)	Wortschatz (CFT 20)	Wortschatz (CFT 20)
	Arithmetik (DEMAT 3+)	Arithmetik (DEMAT 4)	Arithmetik (DEMAT 4)
	Rechtschreiben (DRT 3)*	Textverstehen (ELFE 1-6)	Textverstehen (ELFE 1-6)
			logisch-abstraktes Denken 2 (CFT 20-R)
	- Pause -	- Pause -	- Pause -
	Hörverstehen (KNUSPEL-L)	Grammatik (TROG-D)	Rechtschreiben (DRT 4)
	Textverstehen (ELFE 1-6)*	Metagedächtnis (Würzburger Testbatterie)	logisch-abstraktes Denken 1 (CFT 20-R)
	logisch-abstraktes Denken 1 (CFT 20-R)*		
	Basale Lesefertigkeit (SLS 1-4)		
Testheftversion 2	Wortschatz (CFT 20)	Wortschatz (CFT 20)	Wortschatz (CFT 20)
	Arithmetik (DEMAT 3+)	Arithmetik (DEMAT 4)	Arithmetik (DEMAT 4)
	Rechtschreiben (DRT 3)*	Textverstehen (ELFE 1-6)	Textverstehen (ELFE 1-6)
			logisch-abstraktes Denken 1 (CFT 20-R)
	- Pause -	- Pause -	- Pause -
	Metagedächtnis (Würzburger Testbatterie)	Lesekompetenz (IGLU)	Rechtschreiben (DRT 4)
	logisch-abstraktes Denken 1 (CFT 20-R)*		Metagedächtnis (Würzburger Testbatterie)
	Textverstehen (ELFE 1-6)*		
	logisch-abstraktes Denken 2 (CFT 20-R)		

Anm.: Die für die vorliegende Arbeit relevanten Leistungstests sind grau hinterlegt.

* Die mit Sternchen markierten Kompetenzbereiche werden in diesem Kapitel zwar deskriptiv dargestellt, insbesondere um die Entwicklung über die Zeit besser beschreiben zu können, sie sind jedoch nicht Gegenstand der später folgenden Analysen zur diagnostischen Kompetenz, da für sie auf Seiten der Lehrkräfte keine Einschätzung erhoben wurde.

Bei der Auswahl der Leistungstests wurde überwiegend auf normierte, erprobte und bewährte Verfahren zurückgegriffen. Da insgesamt nur eine be-

grenzte Testzeit von drei Unterrichtsstunden je Erhebung für die Kompetenztestungen und die Schülerbefragungen zur Verfügung stand, wurden einige Testverfahren nicht in ihrem Originalformat, sondern in gekürzter Form administriert. In den folgenden Abschnitten wird auf die für die Berechnungen zur diagnostischen Kompetenz relevanten Leistungstests genauer eingegangen, die in Tabelle 16 grau hinterlegt sind.

Mathematische Kompetenz

Die mathematische Kompetenz der Schüler wurden zum ersten Messzeitpunkt im zweiten Halbjahr der dritten Klassenstufe mit dem Subtest Arithmetik des DEMAT 3+ (Roick, Gölitz & Hasselhorn, 2004) erfasst, der vier Zahlenstrahlaufgaben sowie jeweils 4 Aufgaben zur Addition, Subtraktion und Multiplikation enthält. Dafür standen je Rechenart und für die Zahlenstrahlaufgaben 3 Minuten Zeit zur Verfügung, für die Multiplikation 4 Minuten. Zu den beiden folgenden Messzeitpunkten in der vierten Klassenstufe kam der DEMAT 4 (Gölitz, Roick & Hasselhorn, 2006) zum Einsatz, der zusätzlich auch vier Divisionsaufgaben enthält und die Aufgaben zu den anderen Rechenarten auf ein dem Leistungsniveau der vierten Klasse angemessenes Niveau anhebt. Hier waren für die Bearbeitung der Zahlenstrahlaufgaben 1:30 Minuten, für die Strichrechnungen 3 Minuten und für die Punktrechnungen 3:30 Minuten Zeit vorgegeben. Die Testverfahren der DEMAT-Reihe nehmen für sich in Anspruch, mit ihren Aufgaben eine Schnittmenge der Lehrpläne aller deutschen Bundesländer als Grundlage zu haben. Aus Testleiterprotokollen von den Erhebungen lässt sich jedoch ablesen, dass insbesondere zum zweiten Messzeitpunkt über die Hälfte der Schülerinnen und Schüler die Divisionsaufgaben unbearbeitet ließ, weil diese Rechenart noch nicht im Unterricht behandelt wurde. Hier stoßen zwei nachteilige Aspekte aufeinander. Zum einen war es aus organisatorischen Gründen nicht möglich, die Schüler genau im vom Test vorgegebenen Zeitfenster zu testen. Statt den DEMAT 3+ im Zeitraum der letzten sechs Schuljahreswochen und den DEMAT 4 drei Wochen vor und nach dem Halbjahreswechsel bzw. sechs Wochen vor Ende des vierten Schuljahres einzusetzen, wie es die Handbücher nahelegen, fanden die BiKS-Erhebungen zu t1 bis zu 19 Wochen vor dem Schuljahresende statt, zu t2 bis zu 14 Wochen vor dem Halbjahr und zu t3 immer noch bis zu 12 Wochen vor dem Schuljahresende statt (vgl. Tabelle 4, S. 113). Diese deutlich zu frühe Administration könnte ein Grund dafür sein, dass Rechenarten den Schülern einiger Klassen noch nicht bekannt waren und dass die ermittelten Testleistungen im Mittel schlechter ausfielen als in der Normierungsstichprobe. Ein weiterer

nachteiliger Aspekt ist jedoch, dass die bayerischen und hessischen Lehrpläne nicht konkret festlegen, wann im Schuljahr bestimmte Inhalte gelehrt werden sollen. Stattdessen wird den Lehrern viel Spielraum gelassen, indem lediglich festgeschrieben ist, welche Lernziele am Ende eines zweijährigen Blocks (3. und 4. Klassenstufe) erreicht worden sein sollen. Wann Lehrer jedoch welche Inhalte vermitteln, steht ihnen weitestgehend frei. In der Elementarstufe orientieren sich Lehrer aber in aller Regel an einem Spiralcurriculum, in dem zunächst alle wesentlichen Inhalte auf basaler Ebene eingeführt werden, um sie anschließend in ähnlicher Reihenfolge wiederholt zu behandeln. Dieses Prinzip soll die Wissensfestigung verstärken, führt allerdings auch dazu, dass Inhalte, die zum Zeitpunkt unserer Testung schon längere Zeit nicht mehr besprochen und geübt wurden, von den Schülern auch schlechter gelöst werden. Somit lässt sich u.a. das Phänomen erklären, dass die Schüler am Ende der dritten Klasse Multiplikationsaufgaben besser lösen konnten als ein halbes Jahr später. Die vom Testverfahren in Anspruch genommene curriculare Validität ist demnach eher eine per Augenschein vorgenommene Annahme darüber, dass die abgefragten Aufgaben theoretisch zum empfohlenen Testzeitpunkt in den Schulen behandelt worden sein müssten; prüfen und bestätigen lässt sich dies im statistischen Sinne jedoch nicht, u.a. auch deshalb, weil die den Bildungsplänen zugrunde liegenden Absichten von Lehrern in unterschiedlichem Maße erfüllt werden können (vgl. Langfeldt, 1984, S. 80). Die genannten Einschränkungen haben Implikationen für die Analysen zur diagnostischen Kompetenz. Schneiden die Schüler im Test nämlich schlechter ab als erwartet, und sind sich die Lehrer - wie bei globalen Leistungseinschätzungen - der Schwierigkeit der eingesetzten Aufgaben bzw. ihrer geringeren Übereinstimmung zum Lehrplan nicht bewusst, kann es zu einer scheinbaren Überschätzung der Leistungen kommen, die jedoch nicht auf mangelnde Fähigkeiten seitens der Lehrer sondern auf methodische Schwächen der Erhebung zurückzuführen sind.

Ein weiteres Problem der Arithmetikaufgaben aus dem DEMAT 3+/4 stellte sich bei der Skalierung ein- vs. mehrdimensionaler Rasch-Modelle heraus. Es zeigte sich, dass die Zahlenstrahlaufgaben zusammen mit den Rechenaufgaben kein eindimensionales Konstrukt bildeten. Für die Rechenaufgaben allein ist dies nach Analysen mit ConQuest (Wu, Adams, Wilson & Haldane, 2007) deutlich besser der Fall, so dass für alle Berechnungen auch nur aus den Rechenaufgaben ein Summenscore gebildet und die drei Zahlenstrahlitems je Messzeitpunkt ignoriert wurden.

In Tabelle 17 sind die psychometrischen Kennwerte der Tests zur mathematischen Kompetenz der Schülerinnen und Schüler dargestellt. Wie auch in den nachfolgenden Tabellen zu den anderen eingesetzten Kompetenztests werden Angaben sowohl auf Grundlage der Originaldaten als auch - jeweils darunter stehend - auf Grundlage der imputierten Daten dargestellt (eine nähere Beschreibung des gewählten Imputationsvorgehens folgt in Kapitel 6.4 ab S. 161). Imputiert wurden für die jeweiligen Schülermerkmale nur die Skalensummenwerte, aber nicht auf der u.a. für die Berechnung von Reliabilitätsmaßen relevanten Einzelitemebene, so dass Cronbachs Alpha nur für die originalen Daten berichtet werden kann. Für die übrigen Maße können die Tabellen nicht nur als deskriptive Statistiken, sondern durch den Vergleich von originalen und imputierten Daten gleichzeitig als Prüfung dahingehend angesehen werden, inwiefern durch die Imputation die Originaldaten über die Gesamtstichprobe verändert wurden.

Die Schülerleistungen im mathematischen Bereich sind durchweg gekennzeichnet von einem hohen Anteil nicht bearbeiteter Aufgaben, so dass die Schüler in diesem Test zu allen drei Messzeitpunkten die schlechtesten Leistungen zeigten. Dies ist sicherlich einerseits auf die oben erwähnte ungünstige Passung zu den Lehrplänen und den etwas zu frühen Einsatzzeitpunkt zurückzuführen, was die Testschwierigkeit beeinflusst. Andererseits aber gewiss auch darauf, dass Schüler gegenüber dem mathematischen Bereich die größten Vorbehalte und die geringste Leistungsmotivation aufweisen, wie u.a. Ergebnisse dahingehender Untersuchungen vermuten lassen (z.B. Sparfeldt, Buch, Schwarz, Jachmann & Rost, 2009). Abgesehen von den niedrigen erreichten Punktzahlen sind die Mathematiktests jedoch durch eine zufriedenstellend hohe interne Konsistenz von $\alpha = .74$ bis .78 gekennzeichnet. Die Intraklassenkorrelation (ICC) fällt bei diesem Verfahren besonders hoch aus. Zwischen 12,3 und 28 Prozent der gesamten Varianz ist zu den drei Messzeitpunkten auf Unterschiede zwischen Klassen zurückzuführen (mit imputierten Daten jeweils leicht niedriger). Dies verwundert bei der beschriebenen Lehrplanabhängigkeit der Leistungen nicht, denn es sind große Leistungsunterschiede zwischen Klassen denkbar, je nachdem, ob die im Test vorkommenden Inhalte schon bzw. kürzlich im Unterricht behandelt wurden oder nicht (s.o.).

Tabelle 17: Psychometrische Kennwerte der Tests zur mathematischen Kompetenz (DEMAT 3+, DEMAT 4)

	Anzahl Bearbeitungen	*Anzahl missings*	*erreichbare Punkte*	*Mittelwert*	*SD*	*min.*	*max.*	*Schiefe*	*Cronbachs α*	*ICC*
t1	2274	121	12	5,0	2,8	0	12	0,3	.76	28,0%
	imputiert:			5,0	2,8	0	12	0,3		27,0%
t2	2180	215	16	4,5	2,6	0	13	0,3	.74	12,3%
	imputiert:			4,5	2,6	0	13	0,3		11,9%
t3	2016	379	16	9,6	3,7	0	16	-0,4	.78	17,8%
	imputiert:			9,4	3,7	0	18	-0,3		15,5%

Wortschatz

Der Wortschatz (s. Tabelle 18) der Schüler wurde - als einziger Kompetenztest - zu jedem der drei Messzeitpunkte mit demselben Instrument erfasst, nämlich mit dem Ergänzungstest Wortschatz aus dem CFT 20 (Weiß, 1998). Dieser aus dreißig Items bestehende Test stellt einen schulnahen Ergänzungstest zum eigentlichen Grundintelligenztest dar. Die Schüler sollen je Item zu einem vorgegebenen Begriff aus einer Auswahl von fünf Wörtern dasjenige auswählen, das am ehesten dazu passt. Dafür standen in der dritten Klasse 10 Minuten und in der vierten Klasse 8 Minuten Zeit zur Verfügung. Cronbachs Alpha liegt durchweg um .80 und damit im akzeptablen Bereich. Die Unterschiede zwischen Klassen gehören zu den niedrigsten von allen eingesetzten Verfahren.

Tabelle 18: Psychometrische Kennwerte der Tests zum Wortschatz (CFT 20)

	Anzahl Bearbeitungen	*Anzahl missings*	*erreichbare Punkte*	*Mittelwert*	*SD*	*min.*	*max.*	*Schiefe*	*Cronbachs α*	*ICC*
t1	2272	123	30	14,4	5,0	0	29	-0,1	.79	5,8%
	imputiert:			14,4	4,9	0	29	-0,1		6,1%
t2	2180	215	30	17,0	4,8	3	28	-0,3	.79	5,9%
	imputiert:			16,9	4,8	2	28	-0,3		6,2%
t3	2015	380	30	18,9	4,8	2	30	-0,6	.80	6,5%
	imputiert:			18,7	4,7	2	30	-0,5		6,4%

Textverstehen

Für die Erfassung der Textverstehensleistung kam über alle drei Messzeitpunkte der Subtest „Textverständnis“ aus dem Lesetest ELFE 1-6 (Lenhard & Schneider, 2006) zum Einsatz. Das Textverständnis der Kinder wird dabei auf einer sehr basalen Ebene erfasst, indem zunächst ein aus zwei bis drei Sätzen bestehender kurzer Text gelesen und danach dazu aus einer Auswahl von vier Aussagen die zum Text passende angekreuzt werden soll. Im Original besteht der Subtest aus 13 solcher Texte, zu denen jeweils zwischen einer und drei Fragen gestellt werden, so dass innerhalb von sieben Minuten Testzeit insgesamt 20 Punkte zu erreichen sind. Während dieser Test zu t1 noch sehr gut funktioniert hat, wies er schon zu t2 deutliche Deckeneffekte auf. Die Maximalpunktzahl von 20 war gleichzeitig der am häufigsten erreichte Wert, über 12 Prozent der Schüler machten keinen Fehler. Deshalb wurde der Test in der dritten Welle um weitere 2 Texte mit jeweils drei Aufgaben aus dem Repertoire der Testautoren erweitert. Dadurch differenzierte der Test im oberen Leistungsbereich wieder deutlich besser. Die interne Konsistenz kann für alle drei Messzeitpunkte als sehr gut bezeichnet werden, zu einem geringen Anteil lassen sich auch Unterschiede zwischen den einzelnen Klassen abbilden (vgl. Tabelle 19).

Tabelle 19: Psychometrische Kennwerte der Tests zum Textverstehen (ELFE)

	Anzahl Bearbeitungen	*Anzahl missings*	*erreichbare Punkte*	*Mittelwert*	*SD*	*min.*	*max.*	*Schiefe*	*Cronbachs α*	*ICC*
t1	2270	125	20	12,0	4,4	0	20	-0,1	.90	8,5%
	imputiert:			11,9	4,4	0	20	-0,1		8,6%
t2	2177	218	20	14,9	4,1	0	20	-0,7	.86	6,2%
	imputiert:			14,7	4,1	0	20	-0,6		6,1%
t3	2013	382	26	17,6	4,7	1	26	-0,5	.88	6,6%
	imputiert:			17,4	4,7	1	26	-0,4		6,3%

Rechtschreibleistung

Die Rechtschreibfähigkeiten der Kinder wurden zweimal, jeweils zum Ende des dritten und vierten Schuljahres, erfasst. Zum Einsatz kamen der DRT 3 (Müller, 2003) und der DRT 4 (Grund, Haug & Naumann, 2003), wobei von beiden Verfahren nur jeweils die Hälfte der Items (jedes zweite) administriert wurde, um Testzeit zu sparen. Obwohl sie derselben Testreihe angehören und nur jeweils auf den Altersbereich angepasst sind, gibt es zwi-

schen beiden Verfahren keine Schnittmenge. Das Vorgehen, bei dem von den Testleitern Sätze diktiert werden, die den Kindern mit einer Lücke vorliegenen und die Aufgabe darin besteht, das fehlende Wort in die Lücke zu schreiben, unterscheidet sich hingegen nicht zwischen den beiden Tests. Im Gegensatz zu den Originalvorgaben wurden in dieser Studie nicht die Gesamtzahl der Fehler, sondern die Anzahl korrekt geschriebener Wörter gezählt. Es ging bei der Auswertung weniger darum, konkrete Fehleranalysen vorzunehmen, sondern einen Eindruck von der generellen Kompetenz im Bereich Rechtschreiben zu erhalten. Wie aus Tabelle 20 zu ersehen ist, schneiden die Kinder in der vierten Klasse etwas besser ab als in der dritten. Große Unterschiede gibt es auch hier zwischen Klassen, und Cronbachs Alpha liegt im sehr guten Bereich.

Tabelle 20: Psychometrische Kennwerte der Tests zur Rechtschreibung (DRT 3, DRT 4)

	Anzahl Bearbeitungen	*Anzahl missings*	*erreichbare Punkte*	*Mittelwert*	*SD*	*min.*	*max.*	*Schiefe*	*Cronbachs α*	*ICC*
t1	2274	121	22	12,1	4,8	0	22	-0,1	.84	17,6%
	imputiert:			12,1	4,7	0	22	-0,1		18,0%
t3	1979	416	21	15,9	4,1	2	21	-0,9	.87	14,9%
	imputiert:			15,7	4,0	2	21	-0,8		14,1%

Logisch-abstraktes Denken

Zur Testbatterie gehörte ebenfalls der CFT 20-R (Weiß, 2006), ein Test für die Grundintelligenz im Sinne der „General Fluid Ability“ nach Cattell (1963). Aus diesem Verfahren kam in BiKS lediglich der Matrizentest zum Einsatz, daraus jedoch sowohl der 12 Items umfassende Teil 2 aus dem Originalinstrument sowie der 15 Items lange Teil 1. Wie aus Tabelle 16 hervorgeht, wurden diese beiden Subtests in den Wellen 1 und 3 und abhängig von der Testheftversion in unterschiedlichen Kombinationen administriert. Da die Analysen zeigen, dass jene Schüler, die beide Testteile zu einem Zeitpunkt bearbeitet haben, im zweiten Testteil offensichtlich durch einen Übungseffekt besser abschneiden als jene Schüler, die zuvor keinen anderen Matrizentest vorgelegt bekamen, werden die Analysen in dieser Arbeit auf den 15 Items umfassenden Matrizentest (Teil 1) beschränkt. Die Kinder sollten dabei innerhalb einer Matrix aus Symbolen ein fehlendes Kästchen ergänzen, indem sie das fehlende Symbol aus einer Auswahl aus fünf ähnlichen Symbolen ankreuzten. Dafür standen zu beiden Messzeitpunkten jeweils drei Minuten Zeit zur Verfügung. Nach einem halben Jahr konnte eine

leichte Verbesserung in den Leistungen festgestellt werden (s. Tabelle 21), auch hier ist die interne Konsistenz zufriedenstellend, die Unterschiede zwischen den Klassen aber besonders zu t1 nur gering.

Tabelle 21: Psychometrische Kennwerte der Tests zum logisch-abstrakten Denken (CFT 20-R)

	Anzahl Bearbeitungen	*Anzahl missings*	*Anzahl Items*	*Mittelwert*	*SD*	*min.*	*max.*	*Schiefe*	*Cronbachs α*	*ICC*
t1	2268	127	15	8,1	2,5	0	14	-0,3	.77	2,8%
	imputiert:			8,1	2,4	0	14	-0,3		3,2%
t3	1980	415	15	9,8	2,4	0	15	-0,7	.80	6,6%
	imputiert:			9,8	2,3	0	15	-0,7		6,2%

Alles in allem stand für die Erfassung der verschiedenen Leistungsmaße der Schülerinnen und Schüler eine umfangreiche und vielschichtige Testbatterie zur Verfügung, deren psychometrische Kennwerte den Ansprüchen in hohem Maße genügen. Um einen Eindruck von den Zusammenhängen der Maße untereinander zu bekommen, werden in Tabelle 22 alle Interkorrelationen dargestellt. Besonders hohe Korrelationen finden sich für gleiche Konstrukte zu verschiedenen Messzeitpunkten (in der Tabelle grau hinterlegt), wenngleich dies für den sprachlichen Bereich deutlicher der Fall ist als in den Bereichen Arithmetik und logisch-abstraktes Denken. Auch innerhalb der Messzeitpunkte zeigen sich die höchsten Zusammenhänge zwischen sprachlichen Leistungsmaßen. So korrelieren beispielsweise Wortschatz und Textverstehen durchweg zu ungefähr r = .6. Erwartungsgemäß schwach, wenngleich auch begünstigt durch die hohen Fallzahlen immer noch signifikant, hängen die Kompetenzen in der Arithmetik und dem logisch-abstrakten Denken mit den Kompetenzen in sprachlichen Bereichen zusammen.

Tabelle 22: Interkorrelationen zwischen den Leistungstests zu allen drei Messzeitpunkten

		t1					*t2*			*t3*			
		Ari.	Wor.	Text.	Re.	log.	Ari.	Wor.	Text.	Ari.	Wor.	Text.	Re.
t1	Wortschatz	.30											
	Textverstehen	.34	.60										
	Rechtschreiben	.35	.47	.56									
	log.-abstr. Denken	.32	.28	.30	.29								
t2	Arithmetik	.47	.32	.38	.42	.34							
	Wortschatz	.33	.76	.62	.46	.27	.33						
	Textverstehen	.30	.57	.74	.54	.27	.37	.65					
t3	Arithmetik	.46	.30	.39	.44	.31	.58	.35	.39				
	Wortschatz	.31	.71	.58	.45	.27	.33	.79	.62	.37			
	Textverstehen	.29	.56	.74	.53	.26	.36	.62	.78	.42	.62		
	Rechtschreiben	.31	.43	.53	.70	.25	.42	.46	.55	.51	.48	.57	
	log.-abstr. Denken	.34	.28	.28	.27	.45	.35	.31	.29	.39	.34	.31	.34

Anm.: Alle Korrelationen sind auf dem 1%-Niveau signifikant. Grau hinterlegt sind zur besseren Übersicht die Korrelationen desselben Konstrukts zwischen den Messzeitpunkten (Stabilitäten).

6.2.1.2 Leistungstests für die Schüler der Zusatzstichprobe

Von der umfangreichen Testbatterie, die die Schüler in der ersten Klasse zu bearbeiten hatten, sind für diese Untersuchung nur die drei Leistungsbereiche Arithmetik, Wortschatz und Textverstehen relevant, da sie durch ihre große Ähnlichkeit zu den Verfahren aus der Hauptstichprobe die Möglichkeit eröffnen, als Grundlage von diagnostischen Urteilen der Lehrer Vergleiche zwischen beiden Altersstufen herzustellen. In Tabelle 23 sind die psychometrischen Kennwerte der drei Tests für den einen Messzeitpunkt gegen Ende des ersten Schuljahres dargestellt. Sowohl für die Deskription der Schülerleistungen als auch für die weiterführenden Rechnungen damit werden fehlende Werte - im Gegensatz zur Hauptstichprobe und obwohl es bei Missingquoten zwischen 7,6 und 11,5 Prozent durchaus nützlich gewesen wäre - in der Zusatzstichprobe nicht imputiert, da hier aufgrund der einmaligen Messung und vieler fehlender Elternangaben deutlich weniger Hintergrundinformationen vorliegen.

Tabelle 23: Psychometrische Kennwerte der Leistungstests aus der Zusatzstichprobe

	Anzahl Bearbeitungen	*Anzahl missings*	*Anzahl Items*	*Mittelwert*	*SD*	*min.*	*max.*	*Schiefe*	*Cronbachs α*	*ICC*
AR	820	62	141	43,5	13,0	4	100	0,4	.96	24,2%
WS	820	62	15	7,1	2,5	1	14	0,1	.60	10,0%
TV	791	91	20	4,8	3,3	0	19	1,4	.85	12,0%

Anm.: AR = Arithmetik, WS = Wortschatz, TV = Textverstehen

Arithmetik

Im Unterschied zur Hauptstichprobe wurden die mathematischen Fähigkeiten der Erstklässler mit dem Heidelberger Rechentest (HRT 1-4, Haffner, Baro, Parzer & Resch, 2005), einem Speedtest, erfasst. Dieser bestand aus den folgenden Subtests: Addition (40 Items, 2 Min. Zeit), Subtraktion (40 Items, 2 Min. Zeit), Zahlen ergänzen (40 Items, 2 Min. Zeit) sowie Mengen zählen (21 Items, 1 Min. Zeit), wobei dem Speed-Charakter folgend die einzelnen Items recht leicht lösbar waren und es somit auf die Bearbeitungsgeschwindigkeit ankam. Die Arithmetikleistung wurde durch die Summenwerte aller Subtests ermittelt. Neben der ausgezeichneten internen Konsistenz ($\alpha = .96$) ist eine vergleichsweise hohe Intraklassenkorrelation (24,2%) auffällig, die - ähnlich wie bereits zu t1 in der Hauptstichprobe - auf große Leistungsunterschiede zwischen Klassen hindeutet.

Wortschatz

Für die Bestimmung des Wortschatzes der Kinder wurde der 15 Aufgaben umfassende Subtest Sprachverständnis (Wortschatz) aus dem Kognitiven Fähigkeitstest für 1. bis 3. Klassen (KFT 1-3, Heller & Geisler, 1983) verwendet. Dabei sollten die Schüler ohne Zeitvorgabe dasjenige von fünf Bildern von Objekten ankreuzen, das dem vom Testleiter vorgelesenen Wort entspricht. Dieses Vorgehen soll den Wortschatztests insbesondere für Schulanfänger, die schwierige Wörter noch nicht ohne weiteres lesen können, vereinfachen. Die verwendeten Bilder sind allerdings aufgrund des Alters dieses Testverfahrens nicht mehr zeitgemäß und - bedingt durch ihre schwarz-weiß-Darstellung und die ‚Strichmalweise' ohne klare Konturen - teilweise nicht eindeutig zu erkennen. Möglicherweise dadurch ist die niedrige Reliabilität von $\alpha = .60$ zu erklären. Dies ist eine Einschränkung bei diesem Verfahren, welches mangels Alternativen im frühen Primarschulbereich trotzdem eingesetzt wurde.

Textverstehen

Das Textverstehen wurde analog zur Hauptstichprobe mit dem 20 Items umfassenden Leseverständnistest für Erst- bis Sechstklässler (ELFE 1-6, Lenhard & Schneider, 2006) gemessen. Dafür standen wie in der 3. und 4. Klassenstufe 7 Minuten Zeit zur Verfügung. Im Gegensatz zur 3. und 4. Klasse liegt die Intraklassenkorrelation hier mit 12 Prozent etwas höher, die gute interne Konsistenz (α = .85) ist hingegen bei deutlich niedrigeren erreichten Punktzahlen vergleichbar.

6.2.2 Fragebogen für die Schüler

Zeitgleich zur Erhebung der Kompetenzen der Schüler wurde ihnen in der Hauptstichprobe auch zu jedem Messzeitpunkt ein Fragebogen vorgelegt. Dieser enthält jeweils eine Vielzahl von Items und Skalen, die aus soziologischer, psychologischer und pädagogischer Sicht im Rahmen des BiKS-Projekts bedeutsam sind, so zum Beispiel bezogen auf familiäre Unterstützung, Bildungsaspiration oder das Erledigen von Hausaufgaben. In Welle 3 wurde auch für den Fragebogen die Stichprobe analog zu den Kompetenzheften gesplittet, um die große Menge an interessierenden Items besser zu verteilen und die Schüler nicht zu überlasten. Sämtliche soziodemografischen Variablen der Schüler waren bereits im Zuge der Teilnahmegenehmigungen von den Eltern erfragt worden, so dass im Schülerfragebogen tatsächlich nur die Meinungen der Kinder erfasst wurden. Die für die vorliegende Arbeit relevanten Items und Skalen, zu denen es auch Entsprechungen in den individuellen Einschätzbögen der Lehrer gab, werden im Folgenden beschrieben und in Tabelle 24 deskriptiv zusammengefasst.

Schul-/Leistungsängstlichkeit

Die Leistungsängstlichkeit der Schülerinnen und Schüler ist nur zum dritten Messzeitpunkt mittels vier Items auf einer vierstufigen Skala (stimmt nicht - stimmt eher nicht - stimmt eher - stimmt) erfasst worden. Die Items stammen aus der IGLU-Studie und lauten: a) „Ich habe Angst, mich im Unterricht zu melden.“, b) „Ich fürchte mich davor, aufgerufen zu werden.“, c) „Ich habe Angst, etwas falsch zu machen.“ und d) „Ich traue mich nicht, etwas nachzufragen.“. Die interne Konsistenz ist mit α = .75 zufriedenstellend hoch.

Schuleinstellung

Die Ausprägung der Schuleinstellung von Schülern bezieht sich darauf, wie sehr sich das Kind in der Schule wohl fühlt. Ein hoher Wert bedeutet, dass sich das Kind in der Schule wohl fühlt und gerne dorthin geht. Um die Schuleinstellung der Schüler zu messen, wurden drei Items aus der entsprechenden Skala aus dem ‚Fragebogen zur Erfassung emotionaler und sozialer Schulerfahrungen von Grundschulkindern dritter und vierter Klassen' (FEESS 3-4; Rauer & Schuck, 2003) verwendet. Sie lauteten a) „Ohne Schule wäre alles viel schöner.", b) „Ich gehe gerne in die Schule." und c) „Schule ist ganz schön nervig." und wurden auf derselben vierstufigen Skala gemessen wie die Leistungsängstlichkeit. Die interne Konsistenz der Skala steigerte sich im Verlauf des Längsschnitts von $\alpha = .81$ bis .86 und ist somit sehr gut. Im Originalinstrument umfasst die Skala ‚Schuleinstellung' 14 Items. Aus testökonomischen Gründen wurde für die vorliegende Untersuchung eine sehr reduzierte Auswahl daraus verwendet, die auf Grundlage der Trennschärfe sowie der inhaltlichen Eignung anhand der Formulierung gebildet wurde.

Lernfreude

Ebenfalls aus dem FEESS 3-4 stammen folgende drei Items aus der im Original 13 Aussagen umfassenden Skala, mit denen die Lernfreude der Schüler auf derselben vierstufigen Skala erfasst wurde: a) „Ich arbeite im Unterricht gerne mit.", b) „Ich spiele lieber, als etwas zu lernen." und c) „Ich habe keine Lust, etwas zu lernen.". Cronbachs Alpha liegt hier bei $\alpha = .55$ (t1), .62 (t2) und .67 (t3) und damit sehr niedrig, zumindest für t2 und t3 aber noch im akzeptablen Bereich. Die Lernfreude wird im FEESS 3-4 gemeinsam mit der Skala zur Schuleinstellung sowie mit einer Skala zur Anstrengungsbereitschaft zur Messung der emotionalen und motivationalen Auswirkungen von Selbstwirksamkeit und Selbstbestimmung der Schüler erfasst.

Gefühl des Angenommenseins

Eine weitere Skala aus dem FEESS 3-4 wurde zu allen drei Messzeitpunkten eingesetzt, um mittels fünf Items das Gefühl des Angenommenseins der Schüler zu erheben. Die Formulierungen hierzu waren a) „Meine Lehrer sind gerecht zu mir.", b) „Meine Lehrer mögen mich.", c) „Meine Lehrer kümmern sich um mich.", d) „Meine Lehrer schimpfen zu viel mit mir." und e) „Meine Lehrer helfen mir, wenn ich Hilfe brauche." Auch bei dieser Skala liegt die interne Konsistenz mit $\alpha = .78$ bis .84 zufriedenstellend hoch.

Fachinteresse Deutsch

Aus der BIJU-Studie stammen die vier Items, mit denen das Fachinteresse der Schüler für Deutsch erhoben wurde. Zu den Fragen a) „Wie sehr freust du dich auf eine Stunde im Fach Deutsch?“, b) „Wie viel liegt dir daran, den Stoff des Faches Deutsch zu behalten?“, c) „Wie viel liegt dir daran, im Fach Deutsch viel zu wissen?“ und d) „Wie gerne würdest du im Fach Deutsch noch mehr Stunden haben als bisher?“ antworteten die Schüler auf einer fünfstufigen Skala (gar nicht - wenig - mittel - ziemlich - sehr). Die interne Konsistenz dieser Fachinteressensskala liegt bei $\alpha = .83/.85$. Da das Fachinteresse für Deutsch und Mathematik zu t1 nur bei der Hälfte der Stichprobe erhoben wurde, zu t2 und t3 hingegen bei der gesamten Stichprobe, wird in den Analysen auf die Verwendung des ersten Messzeitpunktes verzichtet.

Fachinteresse Mathematik

Die vier Items zur Erfassung des Fachinteresses für Mathematik stammen ebenfalls aus BIJU und sind analog zu den Fachinteresseitems für Deutsch formuliert. Cronbachs Alpha liegt mit $\alpha = .84/.86$ in vergleichbarer Höhe.

Tabelle 24: Zusammenfassung der Kennwerte für ausgewählte Skalen und Items aus dem Schülerfragebogen

Bereich	*MZP*	*N*	*M*	*SD*	*min.*	*max.*	*Schiefe*	*Cronbachs α*	*ICC*
Schul-/Leistungsängstlichkeit	3	2027	1,5	0,7	1	4	1,5	.75	
	imp.	2395	1,6	0,7	1	4	1,1		3,0%
Schuleinstellung	1	2198	2,9	1,0	1	4	-0,6	.81	
	imp.	2395	3,0	1,1	1	4	-0,6		7,1%
	2	2156	2,9	1,0	1	4	-0,6	.85	
	imp.	2395	2,9	1,0	1	4	-0,6		8,6%
	3	2026	2,8	1,0	1	4	-0,4	.86	
	imp.	2395	2,8	1,0	1	4	-0,4		8,3%
Lernfreude	1	2200	2,9	0,7	1	4	-0,4	.55	
	imp.	2395	3,0	0,8	1	4	-0,4		7,5%
	2	2157	2,9	0,7	1	4	-0,5	.62	
	imp.	2395	3,0	0,8	1	4	-0,4		7,1%
	3	2029	2,9	0,7	1	4	-0,4	.67	
	imp.	2395	2,9	0,8	1	4	-0,3		5,5%
Gefühl des Angenommenseins	1	1047	3,4	0,6	1	4	-1,4	.78	
	imp.	2395	3,4	0,6	1	4	-0,8		5,6%
	2	2079	3,3	0,7	1	4	-1,2	.79	
	imp.	2395	3,3	0,7	1	4	-0,9		8,4%
	3	1946	3,3	0,7	1	4	-1,2	.84	
	imp.	2395	3,3	0,8	1	4	-0,9		10,4%
Fachinteresse Deutsch	2	2082	3,6	1,1	1	5	-0,5	.83	
	imp.	2395	3,7	1,1	1	5	-0,5		7,9%
	3	2026	3,5	1,1	1	5	-0,4	.85	
	imp.	2395	3,6	1,1	1	5	-0,5		5,0%
Fachinteresse Mathematik	2	2084	3,9	1,0	1	5	-0,9	.84	
	imp.	2395	4,0	1,0	1	5	-0,8		3,0%
	3	2026	3,8	1,1	1	5	-0,8	.86	
	imp.	2395	3,9	1,1	1	5	-0,8		3,4%

6.2.3 Einschätzbogen für die Lehrkräfte

Zu jedem Messzeitpunkt wurden die Lehrer gebeten, für jeden ihrer Schüler einen individuellen Einschätzbogen auszufüllen. Somit wurden in der Hauptstichprobe u.a. die jeweiligen letzten Zeugnisnoten für die Fächer Deutsch und Mathematik, die erreichte Punktzahl bei den zurückliegenden Orientierungsarbeiten und - für diese Arbeit von besonderer Bedeutung - Einschätzungen für die bereits aufgelisteten Leistungsbereiche und emotional-motivationalen Eigenschaften abgefragt. Die Formulierung der einzelnen Items wurde jeweils so gewählt, dass sie möglichst nahe an der Formulierung der Fragen im Schülerfragebogen bzw. am erfassten Leistungsmerkmal lag. Die folgende Übersicht listet alle Formulierungen auf. Mit einem Sternchen (*) markierte Einschätzungen wurden so auch in der Zusatzstichprobe erhoben.

Einschätzung der Arithmetikleistung

- Er/sie ... ist mathematisch sehr begabt. (nur zu t1 als Einzelitem)
- Er/sie ... beherrscht die Grundrechenarten.*
- Er/sie ... kann Rechenaufgaben gut lösen.*
- Er/sie ... hat ein gutes Verständnis für Zahlen.*

Einschätzung der Wortschatzleistung

- Er/sie ... ist sprachlich sehr begabt. (nur zu t1 als Einzelitem)
- Er/sie ... verfügt über einen umfangreichen Wortschatz.*

Einschätzung der Textverstehensleistung

- Er/sie ... kann Texte gut verstehen.*

Einschätzung der Leistung im logisch-abstrakten Denken

- Er/sie ... kann logisch-abstrakt denken.

Einschätzung der Rechtschreibleistung

- Er/sie ... macht kaum Fehler beim Schreiben.

Einschätzung der Schul-/Leistungsängstlichkeit

- Er/sie ... hat Angst davor, sich im Unterricht zu melden.
- Er/sie ... hat Angst, etwas falsch zu machen.

- Er/sie ... fürchtet sich davor, aufgerufen zu werden.
- Er/sie ... traut sich nicht, etwas nachzufragen.

Einschätzung der Schuleinstellung

- Er/sie ... geht gerne in die Schule.

Einschätzung der Lernfreude

- Er/sie ... hat viel Freude am Lernen in der Schule.

Einschätzung des Fachinteresses Deutsch

- Er/sie ... hat Interesse am Deutschunterricht.

Einschätzung des Fachinteresses Mathematik

- Er/sie ... hat Interesse am Mathematikunterricht.

Die Kompetenzeinschätzungen sowie die Urteile zur Lernfreude und zur Schuleinstellung sollten dabei auf einer fünfstufigen Skala (trifft voll und ganz zu - trifft eher zu - teils/teils - trifft eher nicht zu - trifft überhaupt nicht zu) bewertet werden, wohingegen die Einschätzungen der fachbezogenen Interessen sowie der Leistungsängstlichkeit aus Gründen der besseren Vergleichbarkeit in Analogie zum Schülerfragebogen auf vier Stufen (stimmt genau - stimmt fast - stimmt ein wenig - stimmt gar nicht) bewertet werden sollten. In der Zusatzstichprobe fehlte im Vergleich zur Hauptstichprobe die Mittelkategorie, und die Skalenendpunkte waren weniger absolut formuliert (trifft zu - trifft eher zu - trifft eher nicht zu - trifft nicht zu).

Zu beachten ist, dass zu t1 die Einschätzitems für die Bereiche Arithmetik und Wortschatz von den Formulierungen zu t2 und t3 abweichen und weniger passend sind. Um für diese wichtigen Bereiche nicht auf Einschätzungen und die Berechnungen zur diagnostischen Kompetenz verzichten zu müssen, wird diese Einschränkung in Kauf genommen, sie muss allerdings bei der Interpretation der Ergebnisse deutlich berücksichtigt werden.

In Tabelle 25 sind die Kennwerte für alle relevanten Lehrereinschätzungen wiedergegeben. Zunächst werden die Werte für die Hauptstichprobe beschrieben. Für die Bereiche Arithmetik und Wortschatz fällt auf, dass sich die Mittelwerte der Einschätzungen zu t1 von den Mittelwerten zu t2 und t3 unterscheiden, was wahrscheinlich auf die unterschiedlichen Formulierungen der Fragen zurückzuführen ist. Die Intraklassenkorrelation liegt im Leistungsbereich zwischen 3,7 (Rechtschreiben) und 9,9 Prozent (Wort-

schatz zu t3), für die nicht-kognitiven Maße gibt es jedoch mit Werten zwischen 11,1 (Fachinteresse Mathematik zu t2) und 23,7 Prozent (Schuleinstellung zu t2) weit größere Unterschiede zwischen Klassen bei gleichzeitig niedrigerer Streuung der Mittelwerte. Bei den wenigen Skalen, die aus mehr als nur einem Item bestehen, ist die Reliabilität mit Cronbachs α = .87 bis .96 sehr zufriedenstellend. Bis auf die Schul-/ Leistungsängstlichkeit weisen alle Skalen eine leicht rechtssteile Verteilung auf, so dass die Annahme einer Normalverteilung eingeschränkt werden muss.

Bei den angegebenen Daten der Zusatzstichprobe (MZP = Z) muss berücksichtigt werden, dass im Unterschied zur Hauptstichprobe die Lehrereinschätzungen auch für die Leistungsbereiche nur auf vierstufigen Skalen (1-4) erfasst wurden. Die dargestellten Mittelwerte sind dementsprechend nicht direkt mit jenen der Hauptstichprobe vergleichbar. Ansonsten sind keine direkten Unterschiede zwischen den Daten der ersten und der dritten und vierten Klasse auffällig, einzig die Intraklassenkorrelation für den Bereich Wortschatz fällt in der ersten Klasse deutlich höher aus (16,9%) als am Ende der Grundschulzeit (8,8 bis 9,9%).

Tabelle 25: Zusammenfassung der Kennwerte für ausgewählte Skalen und Items aus den Einschätzbögen über alle Schüler und alle Messzeitpunkte

Bereich	*MZP*	*N*	*M*	*SD*	*Schiefe*	*Cronbachs α*	*ICC*
Arithmetik	1	2238	3,30	1,09	-0,17		9,6%
	2	2054	3,76	1,00	-0,56	.95	7,0%
	3	1952	3,73	1,01	-0,54	.96	6,5%
	Z*	609	3,28	0,73	-0,82	.95	5,9%
Wortschatz	1	2264	3,22	1,15	-0,07		8,8%
	2	2071	3,55	1,18	-0,39		9,2%
	3	1986	3,59	1,14	-0,45		9,9%
	Z*	595	2,95	0,88	-0,38		16,9%
Textverstehen	2	2067	3,69	1,09	-0,49		7,2%
	3	1987	3,76	1,04	-0,55		8,8%
	Z*	606	3,30	0,75	-0,74		8,0%
logisch-abstraktes Denken	3	1955	3,47	1,10	-0,24		5,6%
Rechtschreiben	3	1980	3,17	1,32	-0,17		3,7%
Schul-/ Leistungsängstlichkeit	3*	1986	1,61	0,66	1,23	.87	16,6%
Schuleinstellung	1	2263	3,99	0,82	-0,55		23,3%
	2	2072	3,92	0,89	-0,52		23,7%
	3	1983	3,80	0,92	-0,49		21,7%
Lernfreude	1	2263	3,76	0,93	-0,36		16,8%
	2	2069	3,65	1,01	-0,35		14,0%
	3	1986	3,59	0,98	-0,25		14,3%
Fachinteresse Deutsch	2*	2065	2,93	0,87	-0,28		14,0%
	3*	1981	2,83	0,86	-0,17		14,3%
Fachinteresse Mathematik	2*	2037	3,09	0,83	-0,48		11,1%
	3*	1948	2,95	0,86	-0,34		15,0%

*) Mit einem Sternchen gekennzeichnete Konstrukte wurden mittels vierstufiger Skalen erhoben, alle übrigen mittels fünfstufiger Skalen.

Anm.: Es werden nur jene Skalen, Items und Messzeitpunkte dargestellt, zu denen es auch eine Entsprechung in der Einschätzung durch die Lehrer gibt. Ist Cronbachs α nicht angegeben, so handelt es sich nur um ein Einschätzitem.

6.2.4 Fragebogen für die Lehrkräfte

In separaten Fragebögen wurden außerdem zu jedem Messzeitpunkt Selbstauskünfte der Lehrkräfte erfragt. Diese bezogen sich neben Fragen zum Beruf und zur Person u.a. auch auf persönlichen Eigenschaften wie die Einstellung zur diagnostischen Kompetenz. Viele dieser Angaben werden nicht nur zur Beschreibung der Stichprobe, sondern auch als unabhängige Variablen bei der Analyse von Bedingungen der diagnostischen Kompetenz verwendet. Nachfolgend werden die für diese Arbeit wichtigsten Variablen deskriptiv dargestellt, wobei jeweils nur jene Lehrkräfte berücksichtigt werden können, von denen auch Angaben vorliegen. Am Anfang jedes Abschnitts sind die jeweiligen Itemformulierungen (und ggf. die Skalenausprägungen) vermerkt.

Berufserfahrung

> Seit wie vielen Jahren unterrichten Sie an einer Grundschule? Bitte geben Sie die Anzahl der Jahre ohne Referendariat an.

Hinsichtlich der Anzahl der Berufsjahre an Grundschulen (hier verkürzt auch als Berufserfahrung bezeichnet) gibt es in der Stichprobe eine große Spanne, die von absoluten Berufsanfängern bis zu Lehrern mit fast vier Jahrzehnten Unterrichtstätigkeit reicht (vgl. Tabelle 26). Die durchschnittliche Berufserfahrung variiert leicht zwischen den Messzeitpunkten.

Tabelle 26: Deskriptive Angaben zur Berufserfahrung der Lehrer

MZP	*N*	*M*	*SD*	*min*	*max*	*Schiefe*
t1	144	15,4	11,1	0	37	0,2
t2	146	17,5	10,5	0,5	37,5	0,1
t3	147	17,3	10,8	0	38	0,2

Lehrdauer in der jetzigen Klasse

> Seit wann unterrichten Sie in der jetzigen Klasse?
>
> seit Beginn der Grundschulzeit / seit Beginn der 2. Klasse / seit Beginn der 3. Klasse / seit Beginn der 4. Klasse / seit einem anderen Termin

Was die Lehrdauer in der Klasse betrifft, unterrichtet ca. ein Fünftel der Lehrer bereits seit Beginn der Grundschulzeit in der Klasse und sieben Prozent seit Beginn der 2. Klasse. Gute zwei Drittel der Stichprobe unterrichtete zum ersten Messzeitpunkt seit Beginn der dritten Klassenstufe in der Klasse, doch durch in der vierten Klasse neu hinzugekommene Lehrer reduzierte sich ihr Anteil ab t2 auf etwa die Hälfte. Somit ergibt sich zum dritten Mess-

zeitpunkt eine relativ große Bandbreite zwischen Lehrern, was die genaue Kenntnis der Schüler aufgrund von direkten Unterrichtserfahrungen mit ihnen betrifft (vgl. Tabelle 27).

Tabelle 27: Deskriptive Angaben zur Lehrdauer der Lehrer in der jeweiligen Klasse

MZP	*N*	*seit 1. Klasse*	*seit 2. Klasse*	*seit 3. Klasse*	*seit 4. Klasse*	*sonstiges*
t1	145	32 (22%)	11 (8%)	99 (68%)	-	3 (2%)
t2	147	32 (22%)	11 (8%)	76 (52%)	27 (18%)	1 (1%)
t3	148	29 (20%)	11 (7%)	73 (49%)	27 (18%)	8 (5%)

Geschlecht

Frauen dominieren unter den Grundschullehrkräften stark, denn ihr Anteil liegt in jeder Erhebungswelle bei über achtzig Prozent (s. Tabelle 28).

Tabelle 28: Deskriptive Angaben zur Geschlechterverteilung der Lehrer

MZP	*N*	*männlich*	*weiblich*
t1	145	25 (17%)	120 (83%)
t2	147	26 (18%)	121 (82%)
t3	147	26 (18%)	121 (82%)

Aus- oder Weiterbildung zur Diagnostik

> Haben Sie während Ihres Lehramtsstudiums eine Veranstaltung zur Diagnostik besucht? (ja/nein)
>
> Haben Sie nach Ihrem Lehramtsstudium an einer Weiterbildung zum Thema „Diagnostik" teilgenommen? (ja/nein)

Eine weitere Frage an die Lehrkräfte betraf ihre Teilnahme an Lehrveranstaltungen während des Studiums und Weiterbildungen nach dem Studium, die sich mit Diagnostik befassten. In Tabelle 29 ist kreuztabellarisch dargestellt, wie viele Lehrer die eine oder andere Form solcher Veranstaltungen (oder beides) besucht haben. Jeweils nur ca. ein Drittel aller Lehrer in der Stichprobe haben entweder während des Studium oder danach eine Diagnostikveranstaltung besucht, und nur 19 Prozent nahmen sowohl während als auch nach dem Studium daran teil, während 53 Prozent der Lehrer noch nie Besucher einer expliziten Diagnostikveranstaltung waren.

Tabelle 29: Deskriptive Angaben zur Teilnahme der Lehrkräfte an Lehrveranstaltungen oder Weiterbildungen zur diagnostischen Kompetenz

		Lehrveranstaltung besucht		
		Ja	Nein	Gesamt
Weiterbildung besucht	Ja	N = 23	N = 16	N = 39
		18,9%	13,1%	32,0%
	nein	N = 18	N = 65	N = 83
		14,8%	53,3%	68,0%
	Gesamt	N = 41	N = 81	
		33,6%	66,4%	

Einstellung zur diagnostischen Kompetenz

Inwieweit stimmen Sie folgenden Aussagen zu?

[1=stimme nicht zu, 2=stimme eher nicht zu, 3=stimme eher zu, 4=stimme zu]

Für Lehrende ist es wichtig, die Persönlichkeit ihrer Schülerinnen und Schüler richtig einschätzen zu können. / Als Lehrer kann man auch ohne Diagnostik-Ausbildung die Leistungen und Eigenschaften seiner Schüler richtig einschätzen.

Um die Einstellung der Lehrer zur diagnostischen Kompetenz zu erfassen, wurden ihnen zu t2 zwei entsprechende Fragen gestellt. Nahezu alle Lehrer waren der Ansicht, dass es wichtig für Lehrer sei, die Persönlichkeit der Schüler richtig einschätzen zu können. Nur ein Lehrer stimmte dieser Aussage nicht zu, alle anderen stimmten zu oder eher zu. Dahingegen war nur knapp ein Drittel der Lehrkräfte der Ansicht, dass es für die richtige Schülereinschätzung auch einer Diagnostikausbildung bedarf. Der überwiegende Teil der Lehrer stimmte der Aussage, dass man Leistungen und Eigenschaften der Schüler auch ohnedies richtig einschätzen könne (vgl. Tabelle 30).

Tabelle 30: Deskriptive Angaben über die Einstellung der Lehrer zur diagnostischen Kompetenz

	N	*M*	*SD*	*min*	*max*	*Schiefe*
Wichtig, die Persönlichkeit der Schüler richtig einschätzen zu können	141	3,9	0,4	1	4	-3,5
Einschätzung der Leistung auch ohne Diagnostikausbildung möglich	139	2,8	0,8	1	4	-0,4

Selbstwahrnehmung der eigenen diagnostischen Kompetenz im Leistungsbereich

Wie sicher sind Sie sich in Ihrer Schülerdiagnose im Leistungsbereich?

[1=trifft nicht zu, 2=trifft eher nicht zu, 3=trifft eher zu, 4=trifft zu]

Ich kenne die Stärken und Schwächen der einzelnen Kinder. / Es fällt mir leicht festzustellen, ob ein Kind eine Aufgabe verstanden hat. / Ich weiß, bei welchen Aufgaben die einzelnen Kinder Schwierigkeiten haben. / Ich merke sehr schnell, wenn jemand etwas nicht verstanden hat. / Ich merke sofort, wenn ein Kind im Unterricht nicht mitkommt.

Fünf Fragen im Lehrerfragebogen bezogen sich auf die Selbstwahrnehmung der eigenen diagnostischen Kompetenz im Leistungsbereich. Sie gehören einer aus COAKTIV übernommenen Skala an, die mit Werten zwischen α = .73 und .78 eine akzeptable interne Konsistenz aufweist und zu jedem Messzeitpunkt sehr vergleichbare Ergebnisse hervorbringt (Tabelle 31). Auf der vierstufigen Skala liegt der Mittelwert bei 3,2 bzw. 3,3.

Tabelle 31: Deskriptive Angaben zur Selbstwahrnehmung der eigenen diagnostischen Kompetenz der Lehrer

MZP	*N*	*M*	*SD*	*min*	*max*	*Schiefe*	*Cronbachs α*
t1	145	3,2	0,4	2,2	4,0	0,1	.78
t2	144	3,2	0,4	2,2	4,0	0,1	.73
t3	138	3,3	0,4	2,2	4,0	0,2	.77

Schwierigkeiten beim Beurteilen

Wie lange haben Sie für die Beantwortung der Fragen A01 und A02 [individuelle Einschätzung von Leistungen und nicht-kognitiven Schülermaßen] gebraucht?

[Minuten]

Wie sicher waren Sie sich bei der Einschätzung der Leistung dieses Schülers/dieser Schülerin im Fach Deutsch (Wortschatz, Rechtschreibung, Lesegeschwindigkeit, Textverstehen)? / ... im Fach Mathematik (Grundrechenarten, Rechenaufgaben, Zahlenverständnis)? / ... hinsichtlich nicht-leistungsbezogener Eigenschaften (Lernfreude, Anstrengungsbereitschaft, Interesse, Ängstlichkeit)?

[1=sehr unsicher, 2=eher unsicher, 3=eher sicher, 4=sehr sicher]

Um einschätzen zu können, wie schwer den Lehrern die Schülerbeurteilungen gefallen sind und für spätere Zusammenhangsanalysen zur diagnostischen Kompetenz erfragten wir eine Schätzung des Zeitaufwandes sowie der Sicherheit bei der Beurteilung der Schüler. Die Formulierungen hierzu waren selbst entwickelt und an die eingesetzten Fragen angepasst. In Tabelle 32 sind die Durchschnittswerte je Lehrer angegeben, die sich aus den klassenweise aggregierten individuellen Angaben je Schüler ergeben. Hinsichtlich der Zeit, die die Lehrer für die Beurteilung der Schüler benötigten, gibt es große interindividuelle Unterschiede. Während der schnellste Lehrer im Durchschnitt für jeden seiner Schüler nur ca. 36 Sekunden benötigte, waren es beim langsamsten ganze zehn Minuten pro Schüler. Da es sich hierbei um von den Lehrern vorgenommene Schätzungen handelt, ist eine gewisse

Unsicherheit einzurechnen. Dahingegen gab es weniger Streuung, was die Sicherheit bei der Einschätzung angeht. Sie ist im Mittel sehr hoch ausgeprägt.

Tabelle 32: Deskriptive Angaben zu Schwierigkeiten und Zeitbedarf der Lehrer bei der Beurteilungen

	N	*M*	*SD*	*min*	*max*	*Schiefe*
durchschnittliche Zeit für Einschätzung je Schüler (in Min.)	141	2,7	1,5	0,6	10,0	1,7
durchschn. Einschätzsicherheit Deutschleistungen	144	3,5	0,3	2,7	4,0	-0,1
durchschn. Einschätzsicherheit Mathematikleistung	142	3,4	0,4	2,1	4,0	-0,5
durchschn. Einschätzsicherheit nicht-kognitive Eigenschaften	144	3,3	0,4	1,9	4,0	-0,3

Fähigkeit zur Perspektivenübernahme

Bitte schätzen Sie ein, wie sehr folgende Aussagen auf Sie zutreffen!

[1=trifft nicht zu, 2=trifft eher nicht zu, 3=trifft eher zu, 4=trifft zu]

Bei Meinungsverschiedenheiten versuche ich, mir die Sache von allen Seiten aus anzuschauen, bevor ich eine Entscheidung treffe. / Ich versuche manchmal, meine Freunde besser zu verstehen, indem ich mir vorstelle, wie die Dinge aus ihrer Sicht aussehen. / Ich glaube, dass jedes Problem zwei Seiten hat, und ich versuche, mir beide Seiten anzusehen. / Wenn ich mich über jemanden aufrege, versuche ich normalerweise erst einmal, in seine Haut zu schlüpfen. / Bevor ich Leute kritisiere, versuche ich mir vorzustellen, wie ich mich fühlen würde, wenn ich an ihrer Stelle wäre.

Zu t2 wurden 5 aus BIJU entnommene Items eingesetzt, die die Fähigkeit der Lehrer zur Perspektivenübernahme erfassen sollten. Die interne Konsistenz lag dabei mit Cronbachs α = .68 zufriedenstellend hoch. Insgesamt schätzten die Lehrer diese Fähigkeit bei sich sehr hoch ein (s. Tabelle 33).

Tabelle 33: Deskriptive Angaben zur Fähigkeit der Lehrer zur Perspektivenübernahme

	N	*M*	*SD*	*min*	*max*	*Schiefe*	*Cronbachs α*
t2	144	3,3	0,4	2,4	4,0	-0,1	.68

Perfektionsstreben

Bitte bewerten Sie, inwiefern folgende Einstellungen auf Sie persönlich zutreffen.

[1=trifft überhaupt nicht zu, 2=trifft eher nicht zu, 3=teils, teils, 4=trifft eher zu, 5=trifft voll und ganz zu]

Meine Arbeit soll stets ohne Fehl und Tadel sein. / Ich kontrolliere lieber noch dreimal nach, als dass ich fehlerbehaftete Arbeitsergebnisse abliefere. / Bei meiner Arbeit habe ich den Ehrgeiz, keinerlei Fehler zu machen. / Was immer ich tue, es muss perfekt sein. / Für mich ist die Arbeit erst dann getan, wenn ich rundum mit dem Ergebnis zufrieden bin. / Es widerstrebt mir, wenn ich eine Arbeit abschließen muss, obwohl sie noch verbessert werden könnte.

Auch das Perfektionsstreben der Lehrer wurde zum dritten Messzeitpunkt mit sechs Fragen erhoben. Auf der fünfstufigen Skala liegt der Mittelwert etwas oberhalb der Mittelkategorie. Nur ein gutes Viertel aller Lehrer erreichte hier Werte unterhalb der Skalenmitte, insgesamt ist das volle Spektrum vertreten. Wie die meisten anderen Skalen ist auch diese von sehr hoher interner Konsistenz gekennzeichnet (s. Tabelle 34).

Tabelle 34: Deskriptive Angaben zum Perfektionsstreben der Lehrer

	N	*M*	*SD*	*min*	*max*	*Schiefe*	*Cronbachs α*
t3	145	3,4	0,8	1,0	5,0	-0,5	.89

Nach der Auflistung von im Lehrerfragebogen erhobenen Merkmalen, die sich auf die Lehrer selbst beziehen, werden abschließend noch die ebenfalls mit diesem Instrument erfassten Skalen zu den Klassenmerkmalen ‚Klassenklima', ‚Unterrichtsstörung' und ‚Zeitverschwendung' dargestellt.

Klassenklima

Bitte sagen Sie mir zu jeder der folgenden Aussagen, in wie weit diese auf Ihre Klasse zutrifft. In dieser Klasse ...

[1=trifft nicht zu, 2=trifft eher nicht zu, 3=trifft eher zu, 4=trifft zu]

... arbeiten die Kinder gut zusammen / ... verstehen sich die meisten Kinder gut miteinander / ...werden Konflikte rasch gelöst / ... ist es selbstverständlich, dass die besseren Kinder den schlechteren helfen / ... arbeiten die meisten Kinder nur sehr zögerlich mit / ... ist jeder nur auf seinen eigenen Vorteil bedacht

Anhand von sechs Items sollten die Lehrer zu zwei Messzeitpunkten das Klassenklima beschreiben. Die Items wurden aus der Studie ‚Schule & Co' (Holtappels & Leffelsend, 2003) adaptiert. Der Skalenmittelwert liegt dabei unterhalb einer mittleren Ausprägung, die interne Konsistenz ist gut.

Tabelle 35: Deskriptive Angaben zum Klassenklima

	N	*M*	*SD*	*min*	*max*	*Schiefe*	*Cronbachs α*
t1	145	3,1	0,5	1,5	4,0	-0,5	.78
t2	144	3,1	0,5	1,2	4,0	-0,7	.85

Unterrichtsstörung

> Einzelne Klassen können sich sehr unterscheiden. Sie sind unterschiedlich schwer zu führen. In welcher Weise versuchen Sie, Ihre Klasse zu führen und wie gut gelingt Ihnen dies in dieser Klasse?
>
> [1=trifft nicht zu, 2=trifft eher nicht zu, 3=trifft eher zu, 4=trifft zu]
>
> Ich muss in dieser Klasse viel ermahnen, um für Ruhe zu sorgen. / In dieser Klasse wird viel Blödsinn gemacht. / In dieser Klasse wird der Unterricht oft sehr gestört. / In dieser Klasse wird viel geschwätzt.

Adaptiert aus COACTIV-Items kam weiterhin eine Skala zum Einsatz, die das Ausmaß der Unterrichtsstörungen in der Klasse erfassen soll. Wie Tabelle 36 zeigt, erweist sich die Skala zu beiden Erhebungszeitpunkten als sehr reliabel.

Tabelle 36: Deskriptive Angaben zur Unterrichtsstörung

	N	*M*	*SD*	*min*	*max*	*Schiefe*	*Cronbachs α*
t1	145	2,5	0,7	1,0	4,0	-0,1	.88
t3	143	2,4	0,7	1,0	4,0	0,1	.89

Zeitverschwendung

> Einzelne Klassen können sich sehr unterscheiden. Sie sind unterschiedlich schwer zu führen. In welcher Weise versuchen Sie, Ihre Klasse zu führen und wie gut gelingt Ihnen dies in dieser Klasse?
>
> [1=trifft nicht zu, 2=trifft eher nicht zu, 3=trifft eher zu, 4=trifft zu]
>
> In dieser Klasse ist es schwer, den Unterricht pünktlich zu beginnen. / Es dauert zu Beginn der Stunde in dieser Klasse sehr lange, bis die Kinder ruhig werden und zu arbeiten beginnen. / Ich habe oft den Eindruck, dass im Unterricht in dieser Klasse viel Zeit vertrödelt wird. / Es fehlt meistens bei irgendjemandem etwas, wenn ich anfangen will zu arbeiten.

Ebenfalls aus der COACTIV-Studie stammen weitere vier Items, die gemeinsam das Ausmaß der Zeitverschwendung im Unterricht erfassen (s. Tabelle 37). Die Skalenmittelwerte fallen hier etwas niedriger aus als bei der Skala zur Unterrichtsstörung, die interne Konsistenz ist auch hier erfreulich hoch.

Tabelle 37: Deskriptive Angaben zur Zeitverschwendung

	N	*M*	*SD*	*min*	*max*	*Schiefe*	*Cronbachs α*
t1	145	2,2	0,7	1,0	4,0	0,4	.80
t3	143	2,1	0,7	1,0	4,0	0,5	.80

6.3 Indikatoren für die diagnostische Kompetenz

Für die vorliegende Arbeit ist die Operationalisierung von diagnostischer Kompetenz von besonderer Bedeutung. Diagnostische Kompetenz ist als latente Variable nicht direkt beobachtbar und messbar, sondern muss über geeignete Indikatoren erschlossen werden. Ganz allgemein wird diagnostische Kompetenz in dieser Arbeit als die Übereinstimmung von Schülermerkmalen mit Einschätzungen derselben durch Lehrer verstanden. Die aufgrund der eingesetzten Instrumente zur Verfügung stehenden Möglichkeiten sind zum Teil mit methodischen Problemen behaftet und bedürfen daher besonderer Erläuterung.

Personenbezogene Diagnosekennwerte

Die Lehrer wurden gebeten, jeden einzelnen (an der Studie teilnehmenden) Schüler ihrer Klasse hinsichtlich verschiedener Eigenschaften einzuschätzen. Dazu zählen zunächst die Leistungen in den nachfolgend genannten Bereichen. Das Prinzip war für jeden Bereich dasselbe; die Lehrer sollten für alle Bereiche Auskunft zur Kompetenz des einzelnen Schülers auf einer vier- oder fünfstufigen Skala einschätzen. Während die Lehrer in anderen Studien, z.B. „Unterrichtsqualität und Leistungszuwachs bei Formen direkter Instruktion im Mathematikunterricht fünfter Hauptschulklassen“ (Schrader, 1989), explizit dazu aufgefordert wurden, unter Berücksichtigung der Situation und der konkreten Aufgaben die Performanz der Schüler einzuschätzen, wurde in der vorliegenden Untersuchung bewusst nach einer globalen Einschätzung gefragt. Dies hat vordergründig forschungspragmatische Gründe; im Rahmen der Gesamtuntersuchung wurde das Globalurteil der Lehrer erfasst, weil es sich leichter und schneller erfassen lässt als aufgabenbezogene spezifische Urteile, zu denen sich die Lehrer zusätzlich noch in die verschiedenen Aufgabenformate hätten eindenken müssen. Die Ökonomie der Forschung, nach der der Aufwand für die Lehrer beim Ausfüllen der Fragebögen so gering wie möglich gehalten werden muss, um sie im Längsschnitt nicht zu verlieren, gestattete ferner nur eine Art der Frageformulierung an alle Lehrer, so dass eine parallele Erfassung spezifischer Urteile ausschied.

Rangordnungskomponente

Das wichtigste Maß für die diagnostische Kompetenz der Lehrer ist für uns die Fähigkeit, die Schüler hinsichtlich ihrer Leistung oder der Höhe ihrer Merkmalsausprägung (z.B. beim Fachinteresse) in eine Rangfolge zu brin-

gen. Sie zeigt sich als Produkt-Moment-Korrelation zwischen geschätzten und tatsächlichen Werten. Dazu baten wir die Lehrer, jeden einzelnen an der BiKS-Studie teilnehmenden Schüler ihrer Klasse in einem individuellen Einschätzbogen zu beurteilen (vgl. hierzu die genaue Beschreibung der Einschätzitems in Kapitel 6.2, S. 138). Zunächst sollten die Lehrer Fähigkeiten und Voraussetzungen der Schüler auf einer fünfstufigen Likert-Skala (trifft voll und ganz zu - trifft eher zu - teils/teils - trifft eher nicht zu - trifft überhaupt nicht zu) einschätzen. Die einzuschätzenden Merkmale wurden jeweils in positiver Ausprägung genannt, z.B. „Er/sie ... macht kaum Fehler beim Schreiben“ (Einschätzung der Rechtschreibleistung des Schülers).

Eine Ausnahme stellen die Einschätzfragen zum Deutsch- und Mathematikinteresse sowie zur Leistungsängstlichkeit dar. Diese sind analog zur Fragestellung in den Schülerfragestellung auf einer vierstufigen Skala (stimmt genau - stimmt fast - stimmt ein wenig - stimmt gar nicht) eingeschätzt worden. Die analogen Fragen im Schülerfragebogen haben eine ebenfalls vierstufige, aber etwas abweichende Skalenausprägung (stimmt nicht - stimmt eher nicht - stimmt eher - stimmt), was seine Ursache in den Bestrebungen hat, Skalen und Formulierungen innerhalb der Instrumente für die jeweilige Zielgruppe angemessen zu gestalten und konsistent zu halten. Die dadurch etwas abweichenden Formulierungen sind zwar nicht ideal, sie sollten aber auch keinen großen Einfluss auf das Antwortverhalten haben.

Die Korrelationen zwischen Schülermerkmalen und Einschätzungen durch die Lehrer basierten zum Teil auf Einzelitems, zum Teil aber auch auf kurzen Skalen. Während beispielsweise im Bereich Wortschatz einem 30 Items langem Schülertest nur ein einzelnes Item des Lehrers gegenübersteht, ist im Bereich Arithmetik auch auf Lehrerseite eine drei Items umfassende Skala zum Einsatz gekommen. Dabei zeigte sich über die Messzeitpunkte konsistent, dass der Zusammenhang des Skalenmittelwertes mit den Arithmetikleistungen der Schüler etwas höher war als zwischen jedem einzelnen Item auf Lehrerseite und der Schülerleistung (s. Tabelle 38).

Tabelle 38: Over-all-Korrelationen zwischen der Arithmetikleistung zu t2 und t3 und einerseits den Einzelitems im Einschätzbogen und andererseits dem daraus gebildeten Skalenmittelwert

	Er/sie beherrscht die Grundrechenarten gut.	*Er/sie kann Rechenaufgaben gut lösen.*	*Er/sie hat ein gutes Verständnis für Zahlen.*	*Skalenmittelwert*
Arithmetikleistung zu t2	.46	.47	.47	.49
Arithmetikleistung zu t3	.54	.54	.54	.56

Auch wenn die leicht höhere Korrelation mit der Skala bei weitem nicht signifikant unterschiedlich gegenüber den Einzelitems ausfällt, kommt dieses Vorgehen dennoch in gewisser Weise den Lehrern entgegen. Über die Einzelitems haben Lehrer die Möglichkeit, ein differenzierteres Urteil zu den Kindern abzugeben, was sich positiv auf die Güte des Arithmetikurteils auswirkt. Einiges spräche somit dafür, auch für die Einschätzung der anderen Leistungsbereiche kleine Skalen statt Einzelitems einzusetzen, was jedoch aus testökonomischen Gründen nicht realisiert werden konnte. Zumindest für den Bereich Arithmetik wird jedoch die Skala für die Berechnungen zugrunde gelegt, auch wenn sich dadurch ein minimaler Unterschied zu den per Einzelitems eingeschätzten Bereichen ergibt.

Auch im nicht-kognitiven Bereich gibt es uneinheitliche Passungen zwischen Schülermerkmalen und Lehrereinschätzungen. Hier steht in der Regel ein Einschätzitem einer aus drei oder vier Items bestehenden Skala gegenüber. Die Skalen sind dabei aus bestehenden Instrumenten adaptiert, wobei sich die Formulierung des Lehrereinschätzitems eng an jeweils einem der Schüleritems innerhalb der Skala orientiert.

Es ist zu beachten, dass nur einige Items und Fragen zu allen drei Messzeitpunkten administriert wurden. Bei vielen Fragen wäre eine halbjährliche Administration nicht sinnvoll gewesen, weil nicht damit zu rechnen ist, dass sich jene abgefragten Eigenschaften in so kurzen Intervallen ändern. In anderen Fällen war es wiederum aus testökonomischen Gründen und vor dem Hintergrund der Vielzahl von Fragen, die in einem derart großen Längsschnittprojekt beantwortet werden sollen, nicht möglich, Fragebatterien längsschnittlich vorzugeben. Insbesondere für die Schüler sollten die Instrumente überschaubar kurz gehalten werden, und auch die Lehrer, die die Instrumente oft in ihrer Freizeit ausfüllten, durften nicht über die Maßen beansprucht werden.

Niveaukomponente

In vielen Studien wird auch geprüft, ob Lehrer das Leistungs- oder Merkmalsniveau ihrer Schüler korrekt bestimmen können. Dazu kann unter anderem den Lehrern die Frage gestellt werden, wie viele Aufgaben eines den Lehrern bekannten Tests die Schüler korrekt beantworten. Bedingung für einen einfachen Vergleich ist, dass Schüler- und Lehrerausprägung eine gemeinsame Metrik haben. Zu den Erhebungszeitpunkten der BiKS-Studie, die dieser Arbeit zugrunde liegen, wurden derartige Einschätzfragen in der Hauptstichprobe nicht gestellt. Für Fragestellungen, die sich auf die Niveaueinschätzung von Lehrern beziehen, muss daher ein Umweg beschritten werden. Um die Schülerausprägungen mit jeweils unterschiedlichen Wertebereichen in den verschiedenen Leistungs- und Eigenschaftsdomänen der immer fünfstufigen Ausprägung der Lehrerurteile anzupassen, sind jeweils Transformationen der Schülersummenwerte in ebenfalls fünf Gruppen notwendig, für die es allerdings verschiedene Vorgehensweisen gibt, von denen einige im nächsten Absatz beschrieben werden. Die Höhe der Über- oder Unterschätzung drückt sich dabei jeweils in der Differenz zwischen Schülerleistung und Lehrerurteil aus, nachdem der fünfstufige Leistungswert jedes Schülers von dem ebenfalls fünfstufigen Einschätzwert der Lehrer subtrahiert wurde. Dabei können sich theoretisch Werte zwischen -4 (maximale Leistungsunterschätzung) und +4 (maximale Leistungsüberschätzung) ergeben, ein Wert von Null bedeutet - unabhängig vom Leistungsniveau der Schüler - eine gemessen am Testwert des Schülers korrekte Einschätzung durch den Lehrer.

- Quintilbildung anhand der Schülerleistungen

Die einfachste Vorgehensweise für die Kategorisierung der Schülerleistungen in fünf Gruppen wäre eine bereichsspezifische Quintilbildung über das Leistungsspektrum der Schüler. Dabei würden entsprechend der Leistungsverteilung fünf annähernd gleich große Gruppen gebildet. Dieses Vorgehen ist für die beabsichtigten Analysen, bei denen eine Differenz zwischen Lehrerurteil und Schülerleistung auf individueller Ebene gebildet werden soll, völlig ungeeignet, weil es zu einer starken Verzerrung führen würde. In der Realität entspricht die Leistungsverteilung - ein valides Testverfahren vorausgesetzt - annähernd einer Normalverteilung, so dass die Mittelkategorie 3 am stärksten besetzt sein sollte, wohingegen die Randkategorien 1 und 5 die wenigsten Schüler zugewiesen bekommen sollten. Die Quintilbildung berücksichtigt diesen Aspekt gar nicht, so dass sich bei der Gegenüberstellung mit den Lehrerurteilen, die sich sehr wahrscheinlich ebenfalls grob an

einer Normalverteilung orientieren werden, automatisch viele Über- bzw. Unterschätzungen ergeben würden. Die Aussagekraft der Ergebnisse bei dieser Herangehensweise wäre sehr eingeschränkt, so dass sie von vornherein nicht in Frage kommt.

- Anpassung der Schülerleistungen an die Verteilung der Lehrereinschätzungen

Wesentlich geeigneter erscheint eine Herangehensweise, bei der der Verteilung der Lehrereinschätzungen Rechnung getragen würde. Die Idee hierbei ist zu prüfen, wie viele Schüler die Lehrer auf den Stufen 1 bis 5 eingeschätzt haben. Anschließend wird ein ebenso großer Anteil der Schüler, angefangen am unteren Ende der Leistungsverteilung, der Stufe 1 zugewiesen, wie es Lehrereinschätzungen für Stufe 1 gibt. Als nächstes werden von den verbliebenen Schülern wiederum so viele der Stufe 2 zugewiesen, wie es Lehrerurteile für Stufe 2 gibt. Diese Prozedur wird bis Stufe 5 fortgeführt. Die Verteilung der Schülerleistungen wird somit an die Verteilung der Lehrereinschätzungen angepasst. Werden nun Differenzen zwischen den fünfstufigen Schülerleistungen und den ebenfalls fünfstufigen Lehrereinschätzungen gebildet, so ergeben sich für jene Schüler von Null abweichende Werte, die deutlich besser oder schlechter im Test abgeschnitten haben als vermutet. In differenziellen Analysen könnte daraufhin geprüft werden, ob sich die (besonders stark) falsch eingeschätzten Schüler hinsichtlich bestimmter Merkmale von korrekt oder anders eingeschätzten Schülern unterscheiden. Der Vorteil dieser Strategie liegt darin, dass die Verteilungen aufeinander abgestimmt sind und sich Differenzen nicht allein aufgrund von künstlich erzeugten Gruppen ergeben können. Allerdings kann dieser Vorteil auch ein Nachteil sein, denn es wird unterstellt, dass die Lehrerurteile ein wahres und geeignetes Kriterium sind. Das muss aber nicht der Fall sein, und so ist es in gewisser Weise als problematisch anzusehen, dass die eigentlich zu erklärende Variable „Lehrerurteil“ hier zum Kriterium avanciert. Gewichtiger ist jedoch noch ein weiterer Nachteil dieses Verfahrens. Da Lehrer sich vermutlich bei ihren Urteilen meist an der eigenen Klasse als Referenz orientieren (vgl. soziale Bezugsnorm, Kapitel 3.3 ab S. 39) und die Klassen in der Stichprobe unterschiedliche Leistungsniveaus aufweisen, ist eine Anpassung der Schülerleistungen an die Lehrereinschätzungen über die Gesamtstichprobe nicht angemessen. Vielmehr müsste diese Anpassung klassenweise erfolgen. Dafür sind die Daten jedoch nicht geeignet. Bei im Durchschnitt nur fünfzehn und in einigen Fällen sogar nur fünf einzuschätzenden Schülern pro Klasse, die eine jeweils unterschiedliche Leistungsver-

teilung aufweisen, ist es nahezu unmöglich, die fünf zu bildenden Leistungsgruppen der Schüler zahlenmäßig an die fünf Lehrerurteilskategorien anzupassen. Alle Schüler mit gleicher Punktzahl müssten selbstverständlich derselben Leistungsgruppe zugeteilt werden, und wenn beispielsweise der Lehrer nur sechs Prozent der Schüler in der höchsten Leistungsgruppe sieht, aber allein fünfzehn Prozent der Schüler in der Klasse die höchste Punktzahl in einem Test erreicht haben, ergeben sich Diskrepanzen, die eine klassenweise Gruppenbildung, so sehr sie aus theoretischen Aspekten plausibel erscheint, unmöglich machen. Also bleibt nur die Gruppenbildung über die Gesamtstichprobe, da durch die sehr hohe Fallzahl die Abstimmung der Gruppengrößen aufeinander leichter - wenn auch immer noch nicht perfekt - möglich ist. Problematisch ist dann jedoch, dass besonders in den Außenkategorien 1 und 5 jeder Lehrer eventuell eine - je nach Leistungsniveau der Gesamtklasse - andere Fähigkeitsausprägung zu beschreiben meint. Die schlechtesten Schüler einiger Klassen würden von ihren Lehrern auf Stufe 1 beurteilt, obwohl ihr Fähigkeitslevel in anderen Klassen mit noch schlechteren Schülern vermutlich einer höheren Stufe entspräche. Indem diese unterschiedlichen Bezugssysteme bei der Transformation über alle Klassen vereinheitlicht werden, ist eine Interpretation der Übereinstimmung auf Klassenebene kaum mehr möglich, da den individuellen Lehrerurteilen nicht zwangsläufig die eigenen Schüler gegenübergestellt werden, sondern die aus der Gesamtverteilung passenden.

- Unterteilung der Bandbreite der Schülerleistung in fünf gleich große Abschnitte

Eine weitere Herangehensweise ist es, die Bandbreite der tatsächlich erreichten Schülerleistungen in fünf gleich große Abschnitte zu unterteilen. Die Abschnitte richten sich hier nur an der Punktzahl aus, nicht an der Anzahl der Schüler in den Gruppen. Das tatsächliche Spektrum der Gesamtgruppe wird deshalb als Kriterium angesetzt, weil somit der testspezifischen Schwierigkeit in einem gewissen Rahmen Rechnung getragen werden kann. Bei sehr schweren Tests (wo hohe Punktwerte unbesetzt blieben) läuft man genauso wie bei sehr leichten Tests (wo niedrige Punktwerte unbesetzt blieben) nicht Gefahr, dass die Leistungsgruppen 5 oder 1 gar nicht oder nur gering besetzt werden. Die Verteilung der Leistungen bleibt im Vergleich zu den originalen Rohwerten bei der Einteilung in fünf Gruppen erhalten. Dennoch kann bei diesem Vorgehen nicht ausgeschlossen werden, dass Leistungsbereiche (v.a. der niedrigste und der höchste) sehr dünn besetzt werden, weil nur sehr wenige Schüler in den Tests sehr schlecht bzw. sehr

gut abgeschnitten haben. Die dadurch entstehenden Differenzen zu den Lehrerurteilen würden dann möglicherweise keine tatsächliche Leistungsüber- oder -unterschätzung ausdrücken, sondern wären auf Effekte der Testschwierigkeit zurückzuführen, die die Lehrer mangels Kenntnis der eingesetzten Testverfahren natürlich nicht berücksichtigen konnten. Dies könnte ein Nachteil dieser Methode gegenüber der zuvor erläuterten sein. Möglich ist aber auch, dass dies gerade ein Vorteil ist, nämlich dann, wenn die Schülerleistungen einer ähnlichen Verteilung unterliegen wie die Lehrerurteile. In der vorliegenden Stichprobe ist dies für einige Testverfahren (zu einigen Messzeitpunkten) gegeben, für andere wiederum nicht (vgl. Kapitel 6.2.1 und 6.2.3). Auch dieses Verfahren müsste genaugenommen klassenweise angewendet werden, indem für jede Klasse individuell das erreichte Leistungsspektrum in fünf gleich große Abschnitte geteilt wird. Somit würde dem von den Lehrern wahrscheinlich angewendeten klasseninternen Bezugsrahmen (vgl. Kapitel 3.3) Rechnung getragen werden. Dies ist eine weitere Variante für die Transformation der Schülerleistungen auf fünf Stufen. Allerdings ist vorstellbar, dass besonders Lehrer in sehr leistungsstarken oder leistungsschwachen Klassen gar nicht die komplette Bandbreite der Urteile benutzen, wohingegen eine Unterteilung der Schülerleistungen anhand des klasseneninternen Leistungsspektrums immer dazu führen würde, dass die volle Bandbreite von 1 bis 5 ausgeschöpft würde. Auf diese Weise würde man sich künstlich Differenzen zwischen Lehrerurteilen und Schülerleistungen schaffen, die es de facto gar nicht gibt.

Vergleich verschiedener Möglichkeiten einer Komponententransformation

Die folgenden Übersichten dienen dem Vergleich der vorgestellten Transformationsvarianten, wobei die Quintilbildung wegen offensichtlicher Schwächen nicht weiter berücksichtigt wird. Stattdessen wird zwischen der Unterteilung des Leistungsspektrums in fünf gleich große Abschnitte anhand der Gesamtleistungsverteilung und auf Grundlage der einzelnen Klassen sowie der Anpassung an die Lehrerurteile differenziert.

Inwiefern sich die Zuweisungen zu den fünf gebildeten Leistungsgruppen je nach Transformationsvariante unterscheiden, wird aus den folgenden Abbildungen deutlich, die sich auf den Bereich Wortschatz zu t3 beziehen, aber stellvertretend auch für die anderen Leistungsbereiche gelten, da es dort ein ähnliches Bild gibt. Darin wird gezeigt, welche Bandbreite an tatsächlichen Leistungen den fünf Gruppen jeweils zugewiesen wird. Die fünf abgebildeten Linien stellen jeweils die Leistungsgruppen 1 bis 5 dar, wobei auf der X-

Achse abzulesen ist, welche tatsächlichen Punktwerte Schüler der jeweiligen Leistungsgruppe im Wortschatztest erreicht haben, während auf der Y-Achse abzulesen ist, wie viele Schüler aus der Stichprobe den Gruppen (und Punktwerten) zuzuordnen sind.

Abbildung 4 zeigt die entstehenden Leistungsgruppen, wenn die Schüler entsprechend ihrer Leistungen je nach Häufigkeit der durch die Lehrer insgesamt vergebenen Leistungszuordnungen gruppiert werden. Dabei wird deutlich, dass das Leistungsspektrum ungleichmäßig aufgeteilt wurde, weil die Lehrer die Urteilskategorien in unterschiedlichen Häufigkeiten gewählt haben. Die Mittelkategorie (3) deckt nur die Punktwerte 15 bis 17 ab, während die besten Schüler Leistungswerte zwischen 22 bis zum Maximum von 30 haben.

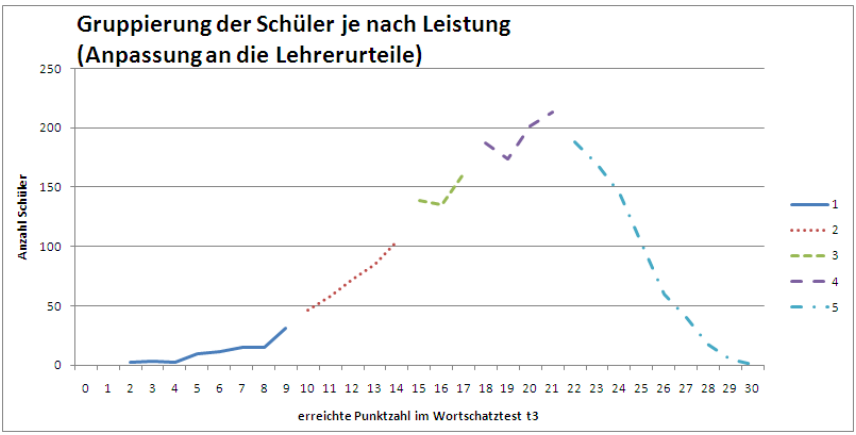

Abbildung 4: Transformation der Schülerleistungen in fünf Gruppen entsprechend der Verteilung der Lehrerurteile

Im Gegensatz dazu werden bei einer über alle Schüler gleichmäßigen Unterteilung des tatsächlichen Leistungsspektrums gleichgroße Leistungsbereiche gebildet (s. Abbildung 5), bei denen es wie bei der ersten Transformationsvariante nicht zu Überschneidungen kommen kann, denn jeder Punktwert aus dem Test entspricht eindeutig einer Leistungsgruppe.

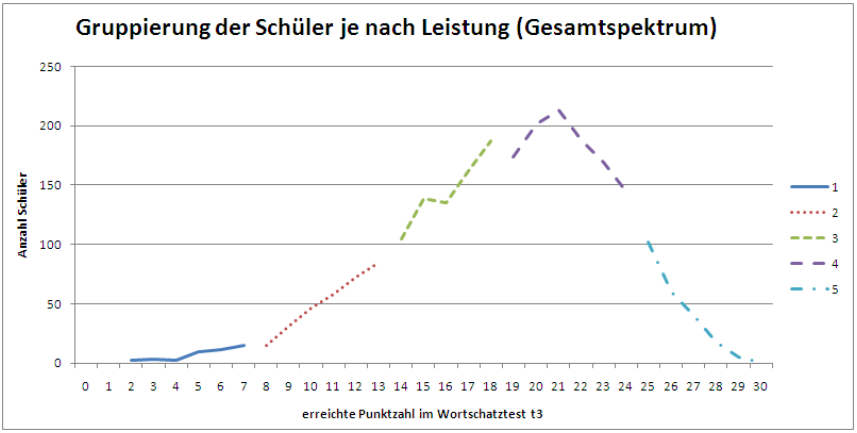

Abbildung 5: Transformation der Schülerleistungen in fünf Gruppen, wobei das insgesamt erreichte Leistungsspektrum in fünf gleich große Abschnitte geteilt wurde

Ein gänzlich anderes Bild ergibt sich, wenn man die Schülerleistungen je nach dem in der jeweiligen Klasse erreichtem Leistungsspektrum in die fünf Gruppen unterteilt (Abbildung 6). Dadurch, dass die Zuteilung zu den Gruppen von den Leistungen der Mitschüler abhängen und das Leistungsniveau von Klasse zu Klasse unterschiedlich sein kann, ergeben sich bei der Gruppenzuweisung sich überschneidende Leistungsbereiche. So ist es in den Extremfällen möglich, dass der Punktwert 17 im Wortschatz zu t3 allen fünf Leistungsgruppen zugewiesen werden kann. In insgesamt sehr leistungsschwachen kann der Punktwert 17 bedeuten, dass man im obersten Fünftel der Leistungsverteilung rangiert, während in insgesamt sehr leistungsstarken Klassen ein Schüler mit 17 Punkten zum schlechtesten Fünftel gehört.

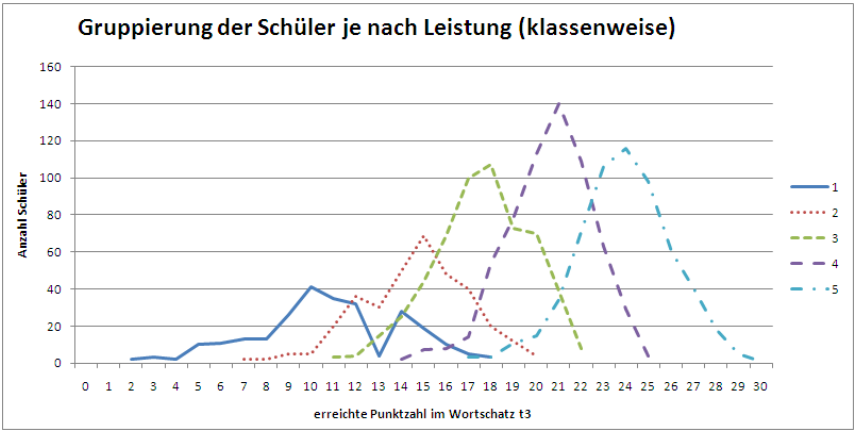

Abbildung 6: Transformation der Schülerleistungen in fünf Gruppen, wobei separat für jede Klasse das erreichte Leistungsspektrum in fünf gleich große Abschnitte geteilt wurde

Ob für die Analysen eine einheitliche Transformation über die Gesamtstichprobe oder auf Basis der einzelnen Klassen vorgenommen wird, hängt entscheidend davon ab, ob sich die Lehrer bei ihren Urteilen eher an einem generellen oder einem klasseninternen Bezugsrahmen orientiert haben. Da die Fragestellung im Einschätzbogen dahingehend unpräzise formuliert war, bleibt nur der Blick in die Daten, um daraus Erkenntnisse zu gewinnen. Wenn die Lehrer anhand eines klasseninternen Maßstabes geurteilt haben, so sollte es zwischen dem mittleren Leistungsniveau der Klasse und den mittleren Urteilen durch die Lehrer nur niedrige Korrelationen geben, wohingegen eine Orientierung an einem absoluten Maßstab eher zu hohen Korrelationen führen sollte. Tabelle 39 zeigt die sich ergebenden Zusammenhänge für t2 und t3 in allen Bereichen auf. In den Leistungsbereichen ergeben sich dabei durchweg Korrelationen von über r = .30 bis maximal r = .42, die eher als mittelhoch zu bezeichnen, aber allesamt signifikant sind. Für die emotional-motivationalen Bereiche trifft dies nicht zu, lediglich für das Fachinteresse in Mathematik ergeben sich mit r = .20 signifikante Korrelationen. Für die Leistungsbereiche ist jedoch festzuhalten, dass, je höher die durchschnittliche Klassenleistung ist, desto höher auch die Leistungsbeurteilungen durch die Lehrer ausfallen, so dass davon ausgegangen werden kann, dass Lehrkräfte sich dabei eher an einem externen Rahmen orientieren. Auch wenn eine eindeutige Entscheidung für oder wider die klassenweise Transformation bei der doch recht niedrigen Höhe der Korrelationen schwerfällt, sprechen die Daten eher für eine Orientierung an der Gesamtverteilung.

Tabelle 39: Korrelationen zwischen mittlerem Klassenniveau je Bereich und den mittleren entsprechenden Lehrerurteilen

	t2	*t3*
Arithmetik	.35**	.30**
Wortschatz	.32**	.33**
Textverstehen	.42**	.32**
logisch-abstraktes Denken	-	.32**
Rechtschreiben	-	.38**
Lernfreude	-.03	.08
Schuleinstellung	.06	.13
Leistungsängstlichkeit	-	.15
Fachinteresse Deutsch	.16	.13
Fachinteresse Mathematik	.20*	.20*

**p < .01, *p < .05

Im Folgenden wurden beispielhaft dieselben Analysen mit jeder der drei Transformationen gerechnet und die Ergebnisse (hier nicht ausführlich dargestellt) miteinander verglichen. Zugrunde gelegt wurden dafür Rechnungen, die in Abschnitt 7.2.3 ab Seite 215 beschrieben sind und die Frage beantworten sollen, ob es hinsichtlich der individuellen Schülereinschätzung zu differentiellen Effekten z.B. je nach Geschlecht der Schüler oder ihrer sozialen Herkunft kommt. Dabei wichen die Ergebnisse je nach Rechenvariante nur minimal voneinander ab. Einzig in einem Fall (Arithmetik zu t3 hinsichtlich sozialer Herkunft) führte nur die erste Transformationsvariante (Anpassung der Schülerleistungen an die Lehrerurteile) zu signifikanten Gruppenunterschieden.

Letztendlich kann trotz aller oben dargestellten Analysen keine zweifelsfreie Entscheidung für oder gegen eine bestimmte Transformation gefällt werden. Jede könnte sicher gerechtfertigt werden, aber jede hat auch ihre Schwächen und Nachteile, die sich auch deshalb nicht beseitigen lassen, weil die Lehrer wahrscheinlich verschiedenartige Antwortverhaltensweisen gezeigt haben. Eine direkte Operationalisierung der Niveaukomponente diagnostischer Kompetenz hätte viel Unsicherheit vermieden. Immerhin konnte nun aber gezeigt werden, dass die Annäherung über Transformationen im Endergebnis nicht wesentlich verschieden voneinander sind und sie allesamt eine gute Annäherung an die Niveaukomponente sein können, solange nicht das absolute, sondern nur das im Gruppenvergleich zu Tage tretende relative

Niveau der Urteile berücksichtigt wird. Für den Autor überwiegen die Argumente, die für die Verwendung der Unterteilung des Leistungsspektrums, gemessen an der Leistungsverteilung der Gesamtstichprobe, sprechen. Daher wird ihr in dieser Arbeit der Vorzug gegeben.

Es resultieren Werte, die mit der Niveaueinschätzung vergleichbar sind, aber dennoch nur ein grobes Hilfsmittel darstellen. Das sich ergebende absolute Niveau der errechneten Über- oder Unterschätzung wird nicht interpretierbar sein. Dies liegt daran, dass die Art der Frageformulierung (nur die Einschätzung verschiedener Fähigkeiten auf einer fünfstufigen Skala von „trifft voll und ganz zu“ bis „trifft überhaupt nicht zu“) und das Fehlen eines absoluten Bezugspunktes keine Aussagen über das Ausmaß von Über- oder Unterschätzung zulässt. Bei dem gewählten Vorgehen wird so getan, als hätte der Lehrer nicht eine Rang-, sondern eine Niveaueinschätzung vorgenommen. Vor dem Hintergrund, dass die Genauigkeit spezifischer Urteile im Vergleich zu globalen Urteilen in den meisten Untersuchungen höher ausfiel (vgl. Kapitel 4.1 ab S. 52), müssen hinsichtlich der Vergleichbarkeit von (spezifischer) originaler und (globaler) abgeleiteter Niveaukomponente weitere Einschränkungen in Kauf genommen werden, die sich wahrscheinlich jedoch nur in bescheidenem Umfang manifestieren. Um es nochmals zu betonen: Die so gewonnenen Ergebnisse stellen für die Richtung und Höhe der Niveauurteile nicht viel mehr als einen groben Richtwert dar und können keinesfalls im Sinne einer echten Niveaueinschätzung interpretiert werden. Ihre Eignung für differentielle Analysen zwischen Gruppen ist davon jedoch unberührt, und darin liegen die Stärke und der Vorteil der Transformation.

Die beschrieben Transformationen lassen sich alle nur bei den Leistungsvariablen sinnvoll anwenden, weil die Bandbreite der in Gruppen zu unterteilenden Punktwerte groß genug ist. Bei den nicht-kognitiven Variablen versagt dieses Vorgehen. Unglücklicherweise passen die Stufen der Skalen im Schülerfragebogen nicht zu den Stufen in den Einschätzbögen, die die Lehrer ausfüllten. Das Bestreben, die Fragebögen für die Ausfüllenden möglichst homogen zu konstruieren, führte leider dazu, dass für die Bereiche Lernfreude und Schuleinstellungen den vierstufigen Schülerskalen fünfstufige Lehrerskalen gegenüberstehen. Für die Fachinteresseitems ist dies genau andersherum. Allein die Leistungsängstlichkeit ist in beiden Instrumenten mittels vierstufiger Skala erhoben worden, dies aber auch nur zu t3. Vier- und fünfstufige Skalen lassen sich nicht aneinander anpassen, um anschließend eine die Über- oder Unterschätzung ausdrückende Differenz zu

errechnen. Die Analysen zur Niveaukomponente diagnostischer Kompetenz werden daher von vornherein auf den Leistungsbereich beschränkt.

6.4 Imputation fehlender Werte

Die dieser Arbeit zugrunde liegenden Daten beruhen auf einer Ausgangsstichprobe von knapp 2400 Schülern aus 145 Klassen. Wie in Längsschnittstudien oft nicht zu vermeiden, tritt jedoch auch hier über die Erhebungszeitpunkte ein zunehmender Stichprobenausfall auf (siehe Tabelle 6, S. 114). Neben diesem Ausfall, der auf das Fehlen einzelner Schüler oder ganzer Klassen zurückzuführen ist, kam es darüber hinaus auch vor, dass nur bestimmte Tests oder Fragen von den Schülern nicht beantwortet wurden, z.B. weil die Testzeit im Einzelfall zu kurz war und die Testung vor Fertigstellung aller Aufgaben abgebrochen wurde. Aber nicht nur auf Schülerseite entstanden fehlende Werte (Missings), sondern es kam auch vor, dass Lehrer für einzelne Items, Schüler oder ganze Klassen keine Einschätzungen vornahmen. Für die Analyse der Daten stellte sich für diese Arbeit daher grundsätzlich die Frage, wie mit fehlenden Werten umzugehen ist. Dabei stehen zwei generelle Möglichkeiten zur Verfügung: a) Fehlende Werte werden hingenommen und Berechnungen nur mit Daten durchgeführt, die vollständig und ohne Lücken vorhanden sind. Somit nutzt man den paarweisen oder fallweisen Ausschluss, der in gängiger Software für statistische Analysen meist standardmäßig vorgesehen ist. b) Fehlende Werte werden mittels geeigneter Imputationsverfahren neu geschätzt.

Nachteil von Vorgehensweise a) ist, dass die Fallzahl mitunter sehr reduziert ist und Analysen somit weniger zuverlässig sind, gerade dann, wenn man den Datensatz klassenweise gruppiert. Allerdings kann man so sicher sein, dass alle verwendeten Werte auch der Realität entsprechen. Verfechter elaborierter Imputationsverfahren halten jedoch dagegen, dass der paar- oder fallweise Ausschluss zum einen zu verzerrten Parameterschätzungen führen kann, wenn die Missings nicht zufällig sondern systematisch auftreten, und dass der fallweise Ausschluss zum anderen nur eingesetzt werden sollte, wenn weniger als 5 Prozent der Fälle ausgeschlossen werden müssen (Graham, Cumsille & Elek-Fisk, 2003).

Allein die wichtigen Kompetenzdaten der Schüler (Summenwerte) weisen je nach Leistungsbereich und mit jedem Messzeitpunkt steigende Missing-Quoten zwischen 5,1 und 17,4 Prozent auf, die größtenteils darauf zurückzuführen sind, dass Schüler zum Erhebungstermin nicht anwesend waren.

Dies führt zu einem deutlichen Rückgang der Fallzahlen und damit zu einem Verlust an Power für die Analysen. Da in der Forschung zunehmend der fall- bzw. listenweise Ausschluss von Personen mit fehlenden Werten zugunsten der Schätzung jener Missings mit leistungsstarken Algorithmen aufgegeben wird, wurde auch in der vorliegenden Arbeit das Verfahren der multiplen Imputation mit der Software R (R Development Core Team, 2009) genutzt, um auf Grundlage aller drei Grundschul-Messzeitpunkte fehlende Summen- bzw. Mittelwerte neu zu schätzen. Die dafür notwendige Bedingung, dass die fehlenden Werte zufällig und nicht systematisch zustande kommen (vgl. Lüdtke, Robitzsch, Trautwein & Köller, 2007), konnte in diesem Fall angenommen werden, da der Ausfall in der überwiegenden Mehrheit durch zufällige Abwesenheit am Testtag begründet war und nicht etwa durch einen systematischen Stichprobenschwund. In die Imputation wurden nicht nur Informationen aus Kompetenztests, sondern auch aus Schüler- und Elternbefragungen einbezogen, so dass möglichst umfassende Hintergrundvariablen für die Schätzung zugrunde lagen. Die Variablen im erstellten Imputationsdatensatz lagen alle als Summen- bzw. Mittelwerte der jeweiligen Skalen vor, so dass nicht auf Itemebene imputiert wurde, sondern auf Skalenebene und nur dann, wenn komplette Tests oder Skalen unbearbeitet geblieben waren. Der Anteil der Lehrkräfte, in deren Klassen Schülerwerte imputiert wurden, betrug zu t2 ca. 33 Prozent, zu t3 ca. 47 Prozent.

Als Ergebnis der multiplen Imputation mit R lagen fünf Datensätze mit neu geschätzten fehlenden Werten (‚plausible values') vor. Im eigentlichen Sinne dieses Verfahrens hätten nun alle Analysen mit jedem einzelnen dieser Datensätze durchgeführt und die daraus resultierenden fünf Ergebnisse mit Hilfe der entsprechenden Rubin'schen Formeln gemittelt werden müssen, um eine korrekte Schätzung des Standardfehlers zu erhalten (s. Rubin, 1987). Die Besonderheit an den dieser Arbeit zugrunde liegenden Fragestellungen ist jedoch, dass die große Mehrheit der Analysen auf klassenweise gebildeten Korrelationskoeffizienten zwischen Schülermerkmalen und Lehrereinschätzungen basiert und somit meist nicht der gesamte Datensatz, sondern klassenweise aggregierte Werte benötigt werden. Für diese besondere Anforderung ist der Rechenaufwand nach der mathematisch korrekten Methode sehr hoch. Um abschätzen zu können, inwiefern sich die Ergebnisse je nach gewählter Rechenmethode unterscheiden und ob eine einfachere Methode zu den gleichen Ergebnissen führt, wurde exemplarisch die Güte diagnostischer Kompetenz je Lehrkraft bezogen auf drei Leistungsbereiche (Arithmetik, Wortschatz, Textverstehen) auf verschiedene Weisen berechnet.

Es wurde verglichen, inwiefern sich die Ergebnisse unterscheiden, wenn man

a) mit dem lückenhaften Originaldatensatz,

b) mit dem Mittelwert von fünf imputierten Datensätzen oder

c) (im Rubin'schen Sinne korrekt) mit 5 imputierten Datensätzen separat und mit anschließender Mittelung der Ergebnisse

die Korrelationen zwischen Lehrereinschätzungen und Schüleritems berechnet. Wie sich zeigte, wichen die Ergebnisse bei der Rechnung mit dem unvollständigen Originaldatensatz (a) deutlich von den Rechnungen mit imputierten Daten ab, was als Hinweis darauf gewertet wird, dass beide Imputationsvorgehen zu einer besseren Schätzung der diagnostischen Kompetenz führen. Zwischen den Ergebnissen der vollkommen korrekten, aber besonders aufwendigen Rechenweise mit fünf Datensätzen (c) und jenen der Rechnungen mit dem aus den imputierten Daten gemittelten Werten (b) ergaben sich nur minimale Unterschiede. Die Differenzen zwischen den Korrelationen sind sowohl je nach Messzeitpunkt als auch in Abhängigkeit vom eingeschätzten Leistungsbereich als marginal zu bezeichnen und liegen bei maximal 0,01.

In Abbildung 7 sind die Unterschiede in den Korrelationen, die sich durch den unterschiedlichen Umgang mit den imputierten Werten ergeben, exemplarisch für den Bereich Arithmetik zu t3 grafisch veranschaulicht. Dabei fällt vor allem auf, dass die Korrelationen auf Basis der beiden Imputationsvarianten sich geringfügig von den Korrelationen auf Basis der Originaldaten unterscheiden, jedoch so gut wie gar nicht untereinander. Abweichungen sind vor allem bei solchen Lehrkräften zu finden, in deren Klassen die Daten mehrerer Schüler imputiert wurden. Dass es hier zu etwas unterschiedlichen Werten kommt, entspricht auch den Erwartungen, schließlich liegen den Urteilen dieser Lehrkräfte mehr Schülereinschätzungen zugrunde. Lehrerurteile zu Schülern, zu denen zu einem Messzeitpunkt keine Leistungsdaten vorliegen, werden bei Benutzung der Originaldaten ignoriert; werden die Schülerleistungsdaten imputiert, können auch die entsprechenden Lehrerurteile mit in die Berechnungen einfließen und deren Güte verändern.

Somit wird davon ausgegangen, dass die wesentlich einfachere und pragmatischere Variante c gut gerechtfertigt werden kann. Dem Vorteil der drastischen Vereinfachungen beim Rechnen steht lediglich der Nachteil gegen-

über, dass nun die Standardfehler unterschätzt werden, was sich bei Analysen auf die Signifikanzprüfung auswirken kann.

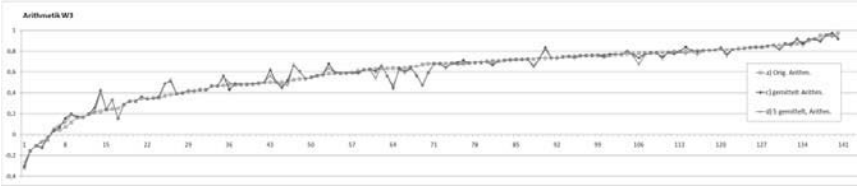

Abbildung 7: Differenzen zwischen den Indikatoren diagnostischer Kompetenz für jeden einzelnen Lehrer, exemplarisch für den Bereich Arithmetik zu t3

Die gemittelten ‚plausible values' von Vorgehensweise c) sind in aller Regel Dezimalzahlen, obwohl viele der beobachteten Werte nur als ganze Zahlen vorliegen können, z.B. Zeugnisnoten oder Punktzahlen in einem Kompetenztest. Für die Berechnung von Gruppenstatistiken ist dies im Grunde genauso unproblematisch wie für die Bildung von Korrelationskoeffizienten. Mitunter fällt jedoch die Interpretation solcher Werte schwer. Deshalb wurden die gemittelten imputierten Daten für diese Arbeit einerseits auf ganze, plausible Werte gerundet. Andererseits kam es in Einzelfällen bei einigen Variablen vor, dass die gemittelten imputierten Daten außerhalb des theoretisch möglichen Wertebereichs lagen (z.B. Schulnoten kleiner 1 oder größer 6 oder Testwerte unter dem Minimal- bzw. über dem Maximalwert). Sie werden trunkiert, also auf den möglichen Wertebereich beschränkt, indem über dem Maximalwert oder unterhalb des Minimalwertes liegende Werte in den jeweiligen Maximal- oder Minimalwert umgewandelt werden.

Den Analysen in dieser Arbeit wird also ein vollständig imputierter Datensatz zugrunde gelegt. Dies betrifft insbesondere die Leistungs- und Fragebogendaten der Schüler, aber auch die lückenhaften Elterninformationen z.B. zum Bildungshintergrund in der Familie. Für die oftmals als unabhängige Variablen benutzten Angaben z.B. zur sozialen Herkunft der Schüler werden keine Schätzungen, sondern nur die tatsächlich erfassten Werte benutzt. Ebenso wäre es nicht sinnvoll, fehlende Lehrereinschätzungen zu imputieren, da dies subjektive Urteile sind, die vermutlich nicht ohne weiteres durch andere Variablen erklärt und deshalb auch nicht einfach geschätzt werden können. Somit wird der Umfang der Stichprobe allein durch nicht imputierte Missings in Lehrervariablen reduziert, nicht aber durch fehlende Schülerantworten.

7 Ergebnisse

7.1 Struktur der diagnostischen Kompetenz von Grundschullehrkräften

Im ersten Teil der Ergebnisdarstellung wird sich der Struktur diagnostischer Kompetenz gewidmet. Bevor jedoch das Augenmerk auf die auf Lehrerebene gebildeten Korrelationen zwischen Schülermerkmalen und den getroffenen Einschätzungen gelegt wird, sollen die Einschätzungen selber genauer beschrieben werden. Dazu zählt, die Ausprägungen auf den Einschätzskalen zu beleuchten, genauso wie die Prüfung der faktoriellen Struktur der Urteile an sich und im Vergleich zu den Schülermerkmalen.

7.1.1 Struktur diagnostischer Urteile und Schülermerkmale

Interkorrelationen der Lehrereinschätzungen und Schülermerkmale

Erkennen Lehrer, welche emotional-motivationalen und leistungsbezogenen Eigenschaften der Schüler miteinander zusammenhängen oder einander ähnlich sind bzw. welche unterschiedlich ausgeprägt sind? Spiegelt sich die Unterschiedlichkeit der Schülerleistungen und -eigenschaften demzufolge auch in den entsprechenden Lehrerurteilen wider? Um dies genauer zu beleuchten, werden im Folgenden die Interkorrelationen von Schülermerkmalen einerseits und Lehrerurteilen andererseits für die drei Messzeitpunkte gegenübergestellt. Dabei finden auch die jeweils zurückliegenden Zeugnisnoten aus den Fächern Deutsch und Mathematik Berücksichtigung, da sie ebenso als Lehrerurteile angesehen werden können wie die Antworten der Lehrkräfte in den Fragebögen.

Zu t1 zeigt sich, dass alle Urteile einschließlich der Zeugnisnoten untereinander überwiegend im mittelhohen Bereich und signifikant miteinander zusammenhängen (s. Tabelle 40). Für die Korrelationen auf Lehrerseite liegen zwischen N = 1664 und 1699 individuelle Lehrerurteile zugrunde, die Basis für die Schülerwerte ist der vollständig imputierte Datensatz mit 2395 Schülern. Die niedrigsten Korrelationen gibt es zwischen der Schuleinstellungseinschätzung und den anderen Bereichen, insbesondere zur Arithmetikleistung und der Zeugnisnote in Mathematik (r = .36). Die größten Korrelationen gibt es hingegen zwischen den Zeugnisnoten und den inhaltlich

ähnlichen Leistungsbereichen. Die Mathematiknote hängt zu r = .72 mit der Mathematikleistungseinschätzung zusammen, die Deutschnote zu r = .69 mit der Wortschatzeinschätzung. Die Zusammenhänge auf Schülerseite erweisen sich hingegen als durchweg niedriger, wenngleich auch sie insbesondere aufgrund der großen Stichprobe signifikant ausfallen. Es gibt allerdings einen wesentlichen Unterschied. Lernfreude und Schuleinstellung korrelieren nach Selbsteinschätzung der Schüler vergleichsweise hoch miteinander (r = .58), während beide Eigenschaften gar keine Zusammenhänge zu den Leistungen aufweisen. Lediglich zu den Zeugnisnoten gibt es geringe Zusammenhänge zwischen r = -.08 und -.17. Anders als die Lehrer annehmen, haben Lernfreude und Schuleinstellung der Schüler also anscheinend überhaupt nichts mit den Leistungen in Arithmetik und im Wortschatz zu tun.

Tabelle 40: Interkorrelationen sowohl der Schülerleistungen und -merkmale als auch der Lehrereinschätzungen zu t1

t1 (Lehrereinschätzungen / Schülermerkmale)	*Arithmetik*	*Wortschatz*	*Lernfreude*	*Schuleinstellung*	*Zeugnisnote Ma*
Wortschatz	.54** /				
	.31**				
Lernfreude	.47** /	.54** /			
	-.04	-.01			
Schuleinstellung	.36** /	.44** /	.80** /		
	.00	.01	.58**		
Zeugnisnote Ma	-.72** /	-.50** /	-.47** /	-.36** /	
	-.43**	-.44**	-.04	-.08**	
Zeugnisnote Deu	-.49** /	-.69** /	-.51** /	-.40** /	.63**
	-.29**	-.48**	-.12**	-.17**	

**p < .01

Anm.: In jeder Zelle steht der obere Wert für die Korrelation der Lehrerurteile, der untere Wert für die Korrelation der Schülermerkmale.

Zu t2 zeigt sich auch bei erweiterter Variablenauswahl grundsätzlich ein sehr ähnliches Bild (Tabelle 41). Wiederum korrelieren die Lehrerurteile untereinander in ausnahmslos allen Bereichen höher als die Schülermerkmale. Die hohe Übereinstimmung von Zeugnisnote und Einschätzungen im korrespondierendem Leistungsbereich wird für die Lehrerurteile nun auch durch den Bereich Textverstehen bestätigt (r = .70). Während die Lehrerur-

teile wieder überwiegend stark miteinander zusammenhängen und sich Differenzierungen allenfalls für die nicht-kognitiven Variablen an niedrigeren Korrelationen ablesen lassen, ist für die Schüler eine deutlich markantere Trennung zwischen nicht-kognitiven Eigenschaften und den Leistungen zu erkennen. Auch im ersten Halbjahr der vierten Klassenstufe sind die Korrelationen zwischen Lernfreude und Schuleinstellung der Schüler zu ihren Leistungen sehr niedrig bzw. gar nicht vorhanden, eine Ausnahme stellen auch jetzt wieder die Beziehungen zu den Zeugnisnoten dar, die tendenziell - für das Fach Deutsch noch mehr als für das Fach Mathematik - signifikant ausfallen. Einen noch höheren Zusammenhang weisen die beiden motivationalen Merkmale jedoch zu den zu diesem Messzeitpunkt neu hinzugekommenen Maßen des Fachinteresses auf, und dort besonders zum Fachinteresse Deutsch ($r = .41/.42$).

Tabelle 41: Interkorrelationen sowohl der Schülerleistungen und -merkmale als auch der Lehrereinschätzungen zu t2

t2 (Lehrereinschätzungen / Schülermerkmale)	*Arithmetik*	*Wortschatz*	*Textverstehen*	*Lernfreude*	*Schuleinstellung*	*Fachinteresse Deutsch*	*Fachinteresse Mathematik*	*Zeugnisnote Mathematik*
Wortschatz	.53** / .35**							
Textverstehen	.55** / .39**	.78** / .66**						
Lernfreude	.48** / .01	.49** / .02	.50** / .06**					
Schuleinstellung	.38** / .05*	.37** / .04	.39** / .09**	.78** / .61**				
Fachinteresse Deu.	.42** / -.01	.58** / .04*	.59** / .08**	.67** / .41**	.60** / .42**			
Fachinteresse Mat.	.74** / .25**	.40** / .03	.39** / .02	.50** / .25**	.41** / .25**	.49** / .14**		
Zeugnisnote Mat.	-.71** / -.52**	-.50** / -.47**	-.51** / -.46**	-.45** / -.04	-.35** / -.08**	-.40** / .00	-.60** / -.28**	
Zeugnisnote Deu.	-.54** / -.40**	-.66** / -.52**	-.70** / -.55**	-.52** / -.14**	-.42** / -.19**	-.58** / -.21**	-.42** / -.05*	.64**

**p < .01, *p < .05

Zum dritten Messzeitpunkt ändert sich die Korrelationstabelle nicht wesentlich, auch hier vermuten die Lehrkräfte deutlich höhere Zusammenhänge zwischen den einzelnen Schülerleistungen und -eigenschaften, als sich dies durch die Daten der Schüler selbst bestätigen lässt. Im Vergleich zu den beiden vorherigen Messzeitpunkten gibt es aber auch einige Besonderheiten. Interessant erscheint zunächst, dass Lernfreude und Schuleinstellung der Schüler nun enger mit den verschiedenen Leistungen und auch mit den Fachinteressen zusammenhängen, besonders in den Bereichen Arithmetik und Textverstehen. Darüber hinaus erstaunt eine besonders große Diskrepanz den neu hinzugekommenen Bereich des logisch-abstrakten Denkens betreffend. Während die Lehrkräfte ihn in engem Zusammenhang zu den arithmetischen Fähigkeiten der Schüler wähnen, ist die tatsächliche Korrela-

tion mit r = .42 sogar noch geringer als zu den Leistungsbereichen Textverstehen (r = .46) und Rechtschreiben (r = .54). Die ebenfalls zu t3 neu erfasste Leistungsängstlichkeit der Schüler sehen die Lehrer in besonders großem Zusammenhang zur Leistung im logisch-abstrakten Denken (r = .37), während dieser sich auf Schülerseite als der am geringsten mit der Ängstlichkeit korrelierende Bereich erweist (r = .11). Beim Fachinteresse zeigt sich wie bereits zu t2 nur im Bereich Mathematik ein Zusammenhang zur Leistung (r = .28), während das Fachinteresse Deutsch wider Erwarten nur eine geringe Beziehung zu den deutschunterrichtsbezogenen Leistungen im Wortschatz und Textverstehen zu haben scheint (r = .07 bis .15) und zur Rechtschreibleistung eine vergleichbare Korrelation von r = .25 besteht.

Tabelle 42: Interkorrelationen sowohl der Schülerleistungen und -merkmale als auch der Lehrereinschätzungen zu t3

t3 (Lehrereinsch. / Schülermerkm.)	*Arithmetik*	*Wortschatz*	*Textverstehen*	*log.-abstr. Denken*	*Rechtschreiben*	*Lernfreude*	*Schuleinstellung*	*Leistungsängstl.*	*Fachinteresse Deu.*	*Fachinteresse Ma.*	*Zeugnisnote Ma*
Wortschatz	.49** / .40**										
Textverstehen	.56** / .46**	.83** / .65**									
log.-abstr. Denken	.81** / .42**	.56** / .36**	.63** / .34**								
Rechtschreib.	.49** / .54**	.62** / .50**	.64** / .60**	.46** / .36**							
Lernfreude	.57** / .12**	.53** / .07**	.55** / .12**	.52** / .02	.54** / .19**						
Schuleinst.	.45** / .18**	.39** / .09**	.43** / .14**	.38** / .06**	.44** / .21**	.81** / .65**					
Leistungsängstl.keit	-.34** / -.19**	-.31** / -.15**	-.34** / -.17**	-.37** / -.11**	-.22** / -.21**	-.32** / -.17**	-.26** / -.17**				
Fachinteresse Deu.	. 46** / .07**	.60** / .07**	.61** / .15**	.44** / -.03	.60** / .24**	.71** / .43**	.63** / .43**	-.30** / -.14**			
Fachinteresse Ma.	.74** / .28**	.38** / .06**	.41** / .05**	.66** / .15**	.37** / .07**	.60** / .29**	.50** / .29**	-.33** / -.20**	.57** / .16**		
Zeugnisnote Ma.	-.78** / -.57**	-.52** / -.47**	-.57** / -.47**	-.72** / -.43**	-.53** / -.49**	-.52** / -.12**	-.39** / -.13**	.30** / .28**	-.47** / -.01	-.63** / -.35**	
Zeugnisnote Deu.	-.56** / -.50**	-.68** / -.55**	-.71** / -.59**	-.55** / -.34**	-.73** / -.63**	-.56** / -.18**	-.43** / -.21**	.27** / .27**	-.60** / -.22**	-.43** / -.12**	.68**

**p < .01

Stabilitäten der Lehrereinschätzungen und Schülermerkmale

Neben den querschnittlichen Analysen soll in Bezug auf den Strukturvergleich zwischen Schülermerkmalen und Lehrereinschätzungen ein abschließender Blick auf die Stabilität derselben geworfen werden. Dem liegt

die Frage zugrunde, ob und inwiefern Lehrer die Veränderlichkeit der Schülerleistungen und der nicht-kognitiven Merkmale in ihren Urteilen berücksichtigen. Gerade in der Grundschulzeit unterliegen nicht nur Leistungen, sondern gerade auch Interessen und Motivationen mitunter größeren Veränderungen. Zu berücksichtigen ist bei den nachfolgenden Analysen, dass zwischen den Messzeitpunkten teils unterschiedliche Messinstrumente auf Schülerseite eingesetzt wurden, was die Stabilität der Schülerleistungen und -eigenschaften möglicherweise mindert, während die Lehrer zu allen drei Messzeitpunkten dieselben globalen Einschätzfragen (ohne Kenntnis der Testaufgaben) beantworteten.

Wie in Tabelle 43 dargestellt, sind besonders die Schülerleistungen in den Bereichen Wortschatz und Textverstehen sehr stabil über den Zeitraum eines halben Jahres (r = .79 - .80). Liegt ein ganzes Jahr zwischen den Messzeitpunkten (t1 - t3), verringern sich die Stabilitäten minimal (nachweisbar für die Wortschatzleistung, r = .74). Im arithmetischen Bereich scheint es hingegen stärkere Veränderungen innerhalb von sechs Monaten zu geben, zwischen Ende dritter und Mitte vierter Klasse (Ende des ersten Halbjahres) noch mehr (r = .48) als innerhalb der vierten Klasse (r = .61). In ähnlichem Ausmaß sind auch Veränderungen für Lernfreude und Schuleinstellung sowie das Fachinteresse für Deutsch und Mathematik feststellbar.

Auf Seiten der Lehrerurteile spiegeln sich diese Werte nur bedingt wider. Zunächst einmal ist augenscheinlich, dass die Stabilitäten zwischen den Messzeitpunkten in allen einzuschätzenden Bereichen nahezu dasselbe Niveau aufweisen, Differenzierungen scheinen in den Urteilen kaum vorzukommen. Während dadurch das Ausmaß der Veränderung für die dem Fach Deutsch zuzurechnenden Bereiche Wortschatz und Textverstehen sehr genau durch die Lehrerurteile abgebildet wird, liegt die Stabilitätsannahme beispielsweise für Arithmetik viel höher als aus den Testleistungen ersichtlich. Ähnliches gilt auch für die Einschätzungen von Lernfreude und Schuleinstellung, die deutlich stabiler sind als die Konstrukte auf Schülerseite. Die von den Lehrern als zu hoch angesehene Stabilität der Leistungen kommt ebenso in den vergebenen Zeugnisnoten zum Ausdruck.

Tabelle 43: Stabilitäten sowohl der Schülerleistungen und -merkmale als auch der Lehrereinschätzungen über die drei Messzeitpunkte

Bereich (Lehrereinschätzungen / Schülermerkmale)	*t1 - t2*	*t1 - t3*	*t2 - t3*
Arithmetik	.73** /	.70** /	.80** /
	.48**	.49**	.61**
Wortschatz	.72** /	.72** /	.83** /
	.79**	.74**	.81**
Textverstehen	-	-	.81** /
			.80**
Lernfreude	.73** /	.66** /	.74** /
	.43**	.40**	.52**
Schuleinstellung	.66** /	.59** /	.69** /
	.53**	.48**	.64**
Fachinteresse Deu	-	-	.68** /
			.58**
Fachinteresse Ma	-	-	.67** /
			.60**
Zeugnisnote Deu	.84**	.78**	.82**
Zeugnisnote Ma	.83**	.74**	.77**

**p < .01

Mit den bis hierher vorgenommenen Analysen zur Struktur der Lehrerurteile im Vergleich zu den korrespondierenden Schülereigenschaften ist ein erster wichtiger Überblick zum Urteilsverhalten der Lehrer geschaffen worden. Es konnte gezeigt werden, dass den Einschätzungen der Lehrer offensichtlich eine Annahme über Zusammenhänge von Schülereigenschaften zugrunde liegt, die sich nur teilweise mit den gemessenen Schülerdaten deckt, und dass Lehrkräfte diese Eigenschaften als weniger variabel über die Zeit empfinden, als dies in der Realität der Fall zu sein scheint. Wie dicht die Lehrerurteile an den tatsächlichen Schülermerkmalen liegen, inwiefern sich verschiedene Lehrer dahingehend voneinander unterscheiden und welche Leistungen und Eigenschaften besonders gut oder schlecht beurteilt werden können, konnte damit noch nicht gezeigt werden. Diesem Aspekt widmet sich der folgende Abschnitt, indem darin die spezifischen Übereinstim-

mungen von Einschätzungen und Ausprägungen anhand der Rangkomponente diagnostischer Kompetenz betrachtet werden.

7.1.2 Güte diagnostischer Urteile

Güte der Rangkomponente diagnostischer Urteile

In diesem Abschnitt wird auf Grundlage der Rangkomponente diagnostischer Kompetenz der Frage nachgegangen, wie valide diagnostische Urteile zu kognitiven und nicht-kognitiven Schülermerkmalen sind. Aufgrund der bereits umfangreichen Forschungsbefunde dazu sind für die Leistungsbereiche Korrelationen im mittelhohen Bereich zwischen r = .5 und .6 zu erwarten, für Einschätzungen zu nicht-kognitiven Maßen sollten die Zusammenhänge niedriger ausfallen. Die Rangkomponente diagnostischer Kompetenz wird berechnet als die klassenweise Produkt-Moment-Korrelation (Pearson) von pro Kind abgegebenen und auf der Konstruktebene angesiedelten Lehrerurteilen und den entsprechenden Schülermerkmalen. Somit entstehen für jede Lehrkraft und jeden einzuschätzenden Bereich separate Indikatoren der diagnostischen Kompetenz. Soll eine mittlere diagnostische Kompetenz über mehrere Lehrer angegeben werden, werden zunächst alle Korrelationen der Lehrkräfte in Fisher-Z-Werte transformiert, um eine annähernde Normalverteilung der Werte zu erhalten. Die geschieht über die Formel

$$Z=\frac{1}{2}\cdot\ln\left(\frac{1+r}{1-r}\right).$$

Im Gegensatz zu normalen Korrelationswerten haben Fisher-Z-Werte Kardinalskalenniveau, so dass ein doppelt so hoher Wert auch in etwa einer doppelt so hohen Güte diagnostischer Kompetenz entspricht (Bortz, 2005, S. 219) und die Berechnung des arithmetischen Mittels zulässig ist. Der errechnete Mittelwert wird anschließend mithilfe der Eulerschen Zahl (e) über die Formel

$$r=\frac{e^{2Z}-1}{e^{2Z}+1}$$

wieder in einen Korrelationskoeffizienten umgewandelt (vgl. Charter & Larsen, 1983). Im Gegensatz zu einer - früher oft berechneten - einfachen Korrelation aller Lehrer- und Schülerwerte hat dieses Vorgehen den Vorteil, dass die jeweils von verschiedenen Faktoren abhängigen individuellen Lehrerwerte der Einschätzgüte gleichgewichtig in den Mittelwert einfließen.

Die Rolle unterrichtlicher Erfahrbarkeit

Die Stichprobe der Lehrkräfte weist dahingehend eine Besonderheit auf, dass es sich durchweg um die Klassenlehrer der getesteten Kinder handelt, dass aber nicht alle Lehrer auch dieselben Fächer unterrichten. Insbesondere ist für diese Untersuchung wichtig zu berücksichtigen, dass - abhängig vom Messzeitpunkt - 10 bis 15 Lehrer die Schüler nicht im Fach Mathematik unterrichten. Dennoch haben diese Lehrer die Arithmetikleistungen und das mathematische Fachinteresse der Schüler eingeschätzt. Da diesen Lehrerangaben jedoch andere (subjektivere) Eindrücke zugrunde liegen als bei Lehrern, die die Schüler tatsächlich in Mathematik unterrichten, werden sie in den folgenden Analysen nicht berücksichtigt. Die Stichprobe reduziert sich dadurch zwar um 5 bis 9 Prozent je nach Messzeitpunkt, aber die Aussagekraft der Ergebnisse steigt, weil alle Lehrer zumindest gleiche Beobachtungsgelegenheiten als Grundlage für ihre Urteile hatten.

Güte der Rangurteile in den Leistungsbereichen

Dies berücksichtigend, sind in Tabelle 44 die mittleren Korrelationen zwischen Lehrereinschätzungen und den jeweiligen Schülerleistungen in Leistungsbereichen Arithmetik, Wortschatz, Textverstehen, Rechtschreiben und logisch-abstraktem Denken dargestellt. Dabei sind Zellen jener Messzeitpunkte und Bereiche unbesetzt, zu denen entweder keine Schülervariablen oder Lehrereinschätzungen oder beides vorliegen. Zu beachten ist weiterhin, dass für manche Klassen keine Lehrerurteile vorliegen. Außerdem gingen Klassen, in denen zu weniger als fünf Schülern Urteile abgegeben wurden, nicht in die Analysen ein, da Korrelationen bei so niedriger Fallzahl zu anfällig für zufällige Verzerrungen und damit kaum aussagekräftig sind. In beiden Fällen flossen diese Klassen nicht in die Analysen ein, was die unterschiedlichen Fallzahlen je nach Bereich und Messzeitpunkt erklärt. Die Grundgesamtheit von 155 Klassen ist daher je nach eingeschätztem Bereich unterschiedlich stark reduziert.

Die angegebenen Korrelationen liegen größtenteils im aus der Literatur bekannten Bereich zwischen r = .5 und .6 und entsprechen damit den Erwartungen. Während die Güte der Einschätzungen in den Bereichen Arithmetik, Wortschatz und Textverstehen in vergleichbarer Höhe mit nur geringfügigen Unterschieden liegt, ist sie im Bereich Rechtschreiben deutlich besser, für das logisch-abstrakte Denken der Schülerinnen und Schüler jedoch schlechter. Die vergleichsweise niedrige Urteilsgüte im Bereich des logisch-abstrakten Denkens unterscheidet sich signifikant von allen anderen Ur-

teilsgenauigkeiten zu allen Messzeitpunkten (Ausnahme: Arithmetik zu t1), die hohe Urteilsgüte im Bereich Rechtschreiben hebt sich ihrerseits signifikant von den meisten anderen ab (genauer: von allen Werten kleiner oder gleich r = .58), unter anderem von den Werten im Wortschatz zu allen Messzeitpunkten. In jenen Leistungsbereichen, in denen Einschätzungen zu verschiedenen Messzeitpunkten erfolgten, zeigen sich nur sehr geringe Veränderungen über die Zeit. Lediglich in der Arithmetik ist die gestiegene mittlere Güte der Einschätzungen zu t3 auffällig, aber nicht signifikant höher als zu t1 und t2.

Tabelle 44: mittlere Einschätzgüte in den Leistungsbereichen (Rangkomponente)

	Arithmetik	*Wortschatz*	*Textverstehen*	*Rechtschreiben*	*logisch-abstraktes Denken*
t1	.52 (0,35)*	.56 (0,37)*	-	-	-
	N=137	N=149			
t2	.55 (0,30)	.55 (0,37)	.58 (0,32)	-	-
	N=132	N=141	N=141		
t3	.65 (0,37)	.55 (0,36)	.61 (0,38)	.73 (0,37)	.34 (0,49)
	N=128	N=143	N=143	N=143	N=142

Anm.: In Klammern ist jeweils die Standardabweichung angegeben.

*Diese Konstrukte wurden im Einschätzbogen zu t1 anders erfragt als zu den anderen Messzeitpunkten.

Güte der Rangurteile in den nicht-kognitiven Bereichen

Im Unterschied zu den Leistungseinschätzungen ist die auf nicht-kognitive Maße bezogene Urteilsgüte - korrespondierend mit Befunden aus anderen Untersuchungen (u.a. Karing, 2009; Spinath, 2005) - überwiegend signifikant niedriger ausgeprägt und liegt zwischen r = .18 und r = .38 (vgl. Tabelle 45). Dabei gibt es kaum Unterschiede in Abhängigkeit vom Bereich, signifikante Differenzen treten nicht auf. Allerdings fällt auf, dass die Einschätzgüte für jedes der Merkmale mit jedem Messzeitpunkt kontinuierlich ansteigt.

Dies ist auch für die Einschätzungen in Arithmetik und minimal für das Textverstehen der Fall, nur im Bereich Wortschatz nicht, wo die durchschnittliche Urteilsgüte auf konstantem Niveau bleibt. Dadurch hat es insgesamt den Anschein, als würde sich die diagnostische Kompetenz der Lehrer im Verlauf der dritten und vierten Klasse (unter Vernachlässigung der stattfindenden Lehrerwechsel in dieser Zeit) verbessern. Ob dies allerdings auf zunehmende Fähigkeiten der Lehrkräfte zurückzuführen ist, kann nicht ge-

klärt werden. Denkbar wäre auch, dass Einschätzungen bei älteren Grundschulkindern generell leichter fallen als bei jüngeren, weil sich beispielsweise die Interessen in diesem Alter erst herausbilden und demzufolge vorher deren Einschätzungen entsprechend schwieriger sind bzw. die Selbstauskünfte der Schüler mangels ausgeprägtem Bewusstsein darüber kein geeignetes Kriterium darstellen.

Tabelle 45: mittlere Einschätzgüte für das Fachinteresse und in emotional-motivationalen Bereichen (Rangkomponente)

	Fachinteresse Deutsch	*Fachinteresse Mathematik*	*Schuleinstellung*	*Lernfreude*	*Leistungsängstlichkeit*
t1	-	-	.23 (0,36)	.18 (0,34)	-
			N=146	N=149	
t2	.25 (0,31)	.33 (0,35)	.29 (0,35)	.26 (0,33)	-
	N=140	N=139	N=139	N=140	
t3	.29 (0,35)	.38 (0,38)	.33 (0,42)	.30 (0,35)	.26 (0,33)
	N=142	N=140	N=142	N=142	N=142

Anm.: In Klammern ist jeweils die Standardabweichung angegeben.

Die mittlere diagnostische Kompetenz weist sowohl bezogen auf die Leistungsvariablen als auch bezogen auf das Fachinteresse und die emotional-motivationalen Variablen recht hohe Standardabweichungen von durchweg über 0,3, teilweise bis fast 0,5 (beim logisch-abstrakten Denken) auf. Dies deutet auf große interindividuelle Unterschiede zwischen den einzelnen Lehrkräften hin, wie sie ebenfalls in der Literatur beschrieben werden. Da die Standardabweichungen in allen Bereichen in etwa gleich hoch ausgeprägt sind, ist Unterschiedlichkeit zwischen Lehrern offenbar nicht bereichsspezifisch, sondern ein generelles Phänomen.

7.1.3 Homogenität der Güte diagnostischer Urteile

Homogenität der Rangurteilsgüte

Die hohen Standardabweichungen für die Güte diagnostischer Urteile legen den Schluss nahe, dass es unter den Lehrern sowohl gute als auch weniger gute oder schlechte Diagnostiker gibt. Daraus ergibt sich die Frage, ob es generell und unabhängig vom einzuschätzenden Bereich gute und schlechte Diagnostiker gibt, oder ob die Urteilsgüte für jeden Lehrer von Bereich zu Bereich unterschiedlich ausgeprägt sein kann. Um herauszufinden, inwie-

weit die bereichsspezifischen Indikatoren der diagnostischen Kompetenz miteinander zusammenhängen (synchrone Zusammenhänge), werden Pearsons Korrelationen auf Grundlage Fisher-Z-transformierter Urteilsgüten für alle Lehrer zwischen allen Bereichen gerechnet und die Ergebnisse getrennt nach Messzeitpunkt in den folgenden Tabellen dargestellt. Betrachtet werden dieselben Leistungs- und emotional-motivationalen Bereiche, die auch in den vorhergehenden Analysen Gegenstand waren. Dadurch wird das Spektrum möglichst groß gehalten, und durch das Einbeziehen von untereinander mal ähnlicheren und mal unähnlicheren Inhaltsbereichen ergibt sich die Möglichkeit, die Hypothese zu testen, dass die Güte der Urteile in sich inhaltlich ähnelnden Bereichen stärker miteinander zusammenhängt als in Bereichen, die inhaltlich kaum etwas oder nichts miteinander zu tun haben.

Zu t1 sind nur vier verschiedene Maße eingeschätzt worden. Wie in Tabelle 46 dargestellt, hängt die Güte der Einschätzung für den arithmetischen Bereich mit r = .25 signifikant mit der Güte der Einschätzungen für den Wortschatz der Kinder zusammen, jedoch nicht mit der Einschätzgüte für die Lernfreude und die Schuleinstellung. Letztere beiden Bereiche hängen untereinander jedoch zu r = .51 noch deutlich enger miteinander zusammen. Darüber hinaus erweisen sich keine Zusammenhänge zwischen den Urteilsgüten in verschiedenen Bereichen als bedeutsam. Es zeigt sich also, dass die Güte von Leistungseinschätzungen untereinander korrelieren, genau wie dies bei der Güte der Einschätzungen in den motivationalen Bereichen der Fall ist. Zwischen diesen zwei wesentlichen Inhaltsbereichen gibt es aber kaum Zusammenhänge mit Blick auf die Urteilsgüte. Die Fähigkeit, Leistungen gut einschätzten zu können, geht also nicht zwangsläufig mit der Fähigkeit, nicht-kognitive Eigenschaften korrekt beurteilen zu können, einher.

Tabelle 46: Synchrone Zusammenhänge der Urteilsgüte für alle zu t1 eingeschätzten Bereiche (Homogenität)

t1	*Arithmetik*	*Wortschatz*	*Lernfreude*
Wortschatz	.25**		
	N=132		
Lernfreude	.04	.09	
	N=132	N=148	
Schuleinstellung	.06	.13	.51**
	N=129	N=145	N=145

**p < .01

Zum zweiten Messzeitpunkt stehen doppelt so viele Variablen für die Homogenitätsprüfung zur Verfügung als zu t1. Dabei ergibt sich ein etwas anderes Bild als zum ersten Messzeitpunkt (s. Tabelle 47). Die Güte der Einschätzungen in den beiden der Domäne Deutsch zuzuordnenden Leistungsbereichen Wortschatz und Textverstehen hängt über alle Lehrer hinweg signifikant miteinander zusammen (r = .41). Wie bereits zu t1 korreliert die Güte der Matheleistungseinschätzung - wenn nun auch nur noch auf dem 5-Prozent-Niveau - signifikant mit der Güte der Wortschatzeinschätzung, zum Textverstehen jedoch nicht bedeutsam. Weitere signifikante Zusammenhänge bestehen zwischen der Urteilsgüte für die Lernfreude und der Urteilsgüte für die Schuleinstellung der Schülerinnen und Schüler (r = .45) sowie für ihr Fachinteresse in Deutsch (r = .29) und etwas geringer in Mathematik (r = .22). Lernfreude- und Schuleinstellungseinschätzung hängen damit in ähnlichem Ausmaß zusammen wie zum ersten Messzeitpunkt. Ein weiterer bedeutsamer, wenn auch recht niedriger Zusammenhang findet sich zwischen der Güte der Schuleinstellungseinschätzung und jener für das Fachinteresse in Deutsch (r = .18). Interessanterweise gibt es hingegen gar keinen Zusammenhang zwischen den beiden fachinteressenbezogenen Urteilsgüten. Zum zweiten Messzeitpunkt kristallisiert sich demnach ein Muster heraus, nach dem inhaltlich ähnliche Schülerleistungen und -eigenschaften, die untereinander auch eng zusammenhängen (vgl. Tabelle 41 auf S. 168), von den Lehrkräften auch gleichermaßen gut bzw. schlecht eingeschätzt werden.

Tabelle 47: Synchrone Zusammenhänge der Urteilsgüte für alle zu t2 eingeschätzten Bereiche (Homogenität)

t2	*Arithmetik*	*Wortschatz*	*Textverste-hen*	*Lernfreude*	*Schulein-stellung*	*Fachinte-resse Deu.*
Wortschatz	.20*					
	N = 130					
Textverste-hen	.12	.41**				
	N = 130	N = 140				
Lernfreude	-.09	-.00	.02			
	N = 129	N = 139	N = 139			
Schuleinstel-lung	-.01	.12	.08	.45**		
	N = 129	N = 138	N = 138	N = 137		
Fachinteresse Deutsch	-.02	.06	.09	.29**	.18*	
	N = 129	N = 139	N = 139	N = 139	N = 137	
Fachinteresse Mathematik	.14	.08	-.16	.22*	.16	-.04
	N = 130	N = 130	N = 130	N = 129	N = 129	N = 129

**p < .01, *p < .05

Das Befundmuster aus t2 wird auch zum dritten Messzeitpunkt tendenziell bestätigt (vgl. Tabelle 48) und stützt sich nun auf zehn Variablen. Wiederum zeigen sich signifikante Zusammenhänge innerhalb des sprachlichen Bereichs. Die Wortschatzeinschätzgüte hängt zu r = .55 mit der Textverstehenseinschätzung zusammen, und auch zwischen dem Wortschatz und der neu hinzugekommenen Einschätzung der Rechtschreibleistung besteht ein signifikanter Zusammenhang von r = .35. Die Korrelation zwischen Rechtschreib- und Textverstehenseinschätzung liegt mit r = .17 noch knapp oberhalb der Signifikanzgrenze. Unerwartet hoch ist der Zusammenhang zwischen Matheeinschätzung und Rechtschreibeinschätzung ausgefallen, der bei r = .31 liegt, und auch die Einschätzung des mathematischen Fachinteresses weist zur Rechtschreibeinschätzung einen signifikanten Zusammenhang auf (r = .23). Ein etwas weniger enger, aber dennoch signifikanter Zusammenhang besteht zu diesem Messzeitpunkt zwischen den Fähigkeiten der Lehrkräfte, die Leistungen in Mathematik und im logisch-abstrakten Denken einzuschätzen (r = .27), was aufgrund der inhaltlichen Nähe auch erwartungsgemäß ist. Wie schon zu den beiden ersten Messzeitpunkten ist ein hoher Zusammenhang zwischen den Einschätzungen von Schuleinstellung und Lernfreude der Schüler zu verzeichnen (r = .59), und auch Lern-

freude- und Fachinteresse Deutsch-Einschätzungen korrelieren mit r = .39 relativ hoch. Die Urteilsgüte für das mathematische Fachinteresse hängt zu t3 wie beim vorherigen Messzeitpunkt mit der Lernfreudeeinschätzung zusammen (r = .20), darüber hinaus gleichermaßen auch mit den Einschätzungen von logisch-abstraktem Denken und der Rechtschreibleistung (r = .18 bzw. .23).

Tabelle 48: Synchrone Zusammenhänge der Urteilsgüte für alle zu t3 eingeschätzten Bereiche (Homogenität)

t3	*Arithmetik*	*Wortschatz*	*Textverstehen*	*log.-abstr. Denken*	*Rechtschreiben*	*Lernfreude*	*Schuleinstellung*	*Leistungsängstl.*	*Fachinteresse Deu.*
Wortschatz	.02								
	N=126								
Textverstehen	.05	.55**							
	N=126	N=143							
log.-abstr. Denken	.27**	-.05	.12						
	N=126	N=142	N=142						
Rechtschreiben	.31**	.35**	.17*	-.05					
	N=126	N=143	N=143	N=142					
Lernfreude	.04	-.07	.08	.18*	.07				
	N=126	N=142	N=142	N=141	N=142				
Schuleinstellung	.11	.09	.16	.03	.16	.59**			
	N=125	N=142	N=142	N=141	N=142	N=141			
Leistungsängstl.	-.09	.09	.09	-.13	.09	-.03	.04		
	N=125	N=142	N=142	N=141	N=142	N=141	N=141		
Fachinteresse Deu.	.01	-.12	-.02	.19*	-.03	.39**	.11	.15	
	N=125	N=142	N=142	N=141	N=142	N=141	N=141	N=141	
Fachinteresse Ma.	.23*	.00	-.02	.18*	.23*	.20*	.12	-.10	.13
	N=126	N=126	N=126	N=126	N=126	N=126	N=125	N=125	N=125

**p < .01, *p < .05

Insgesamt kann man zusammenfassen, dass die Mehrheit der Güte von Einschätzungen statistisch unabhängig von anderen Einschätzungsgenauigkeiten ist. Signifikante Korrelationen sind vor allem zwischen inhaltlich ähnlichen Bereichen zu finden, sowohl bei den Leistungsvariablen innerhalb der Domäne Deutsch als auch wiederholt zwischen den Einschätzungen von

Schuleinstellung und Lernfreude. Auffällig ist der wiederholt gefundene Zusammenhang zwischen der Urteilsgüte für das Fachinteresse Deutsch und für die Lernfreude und die Schuleinstellung der Schüler. Hinsichtlich der Niveauentwicklung der signifikanten Zusammenhänge ist das Bild nicht einheitlich. In Tabelle 49 sind die wichtigsten Korrelationen aus den eben dargestellten Ergebnistabellen noch einmal im zeitlichen Verlauf abgetragen. Der erste Messzeitpunkt ist aufgrund seiner anderen Formulierungen bei den Leistungseinschätzungen vorsichtig zu interpretieren. Abgesehen davon findet sich zwischen der Arithmetik- und der Wortschatzeinschätzung ein abnehmender Zusammenhang über die Zeit, ebenso beim Verhältnis von Wortschatz- und Lernfreudeeinschätzung. Zunehmende Übereinstimmung ist im Mittel jedoch besonders für den Zusammenhang der Genauigkeit von Wortschatz- und Textverstehens- sowie der Lernfreude- und Schuleinstellungseinschätzung zu finden.

Tabelle 49: Entwicklung der Homogenität der Urteile

	t1	*t2*	*t3*
Arithmetik - Wortschatz	.25**	.20*	.02
Wortschatz - Textverstehen		.41**	.55**
Wortschatz - Lernfreude	.09	-.00	-.07
Lernfreude - Schuleinstellung	.51**	.45**	.59**
Lernfreude - Fachinteresse Deutsch		.29**	.39**
Lernfreude - Fachinteresse Mathematik		.22*	.20*
Schuleinstellung - Fachinteresse Deutsch		.18*	.11

**p < .01, *p < .05

Schülermerkmale in einem Bereich zutreffend einschätzen zu können, bedeutet offenbar nicht automatisch, auch in anderen Bereichen zu korrekten Urteilen zu kommen. In Abbildung 8 ist exemplarisch die Verteilung der Indikatoren diagnostischer Kompetenz für die Bereiche Arithmetik, Wortschatz und Textverstehen zu t3 dargestellt. Die Lehrer auf der X-Achse sind dabei nach ihrem Mittelwert dieser drei Indikatoren sortiert, die Indikatoren selbst sind pro Lehrer mit einer Spannweitenlinie verbunden. Dabei wird zunächst deutlich, dass die Verteilung der Bereiche keiner Systematik unterliegt und dass die Streuung zwischen den Bereichen auch für viele Lehrer

beachtlich ist, wie bereits die angegebenen Standardabweichungen in Tabelle 44 andeuteten. Rein optisch könnte der Eindruck entstehen, dass die Streuung zwischen den Bereichen mit zunehmender mittlerer Einschätzgüte geringer wird. Dieser Eindruck ist auch dadurch bedingt, dass die insgesamt sehr genau urteilenden Lehrer im rechten Bereich der X-Achse dem Optimum (mittlere Korrelation = 1) recht nahe kommen und allein deshalb keine große Varianz zwischen den verschiedenen Bereichen mehr auftreten kann, denn da kein größerer Wert als 1 erreicht werden kann, würde sich sonst ihr Mittelwert der drei Bereiche reduzieren und sie gehörten eben nicht mehr zu den insgesamt sehr gut einschätzenden Lehrern. Im hier abgebildeten Beispiel beträgt der Zusammenhang zwischen Streuung und Gütemittelwert jedoch nur r = -.09, so dass allenfalls von einer Tendenz gesprochen werden kann. Es ist also nicht so, dass insgesamt besser urteilende Lehrkräfte auch eine größere Homogenität der Einschätzgüte über die Bereiche aufweisen würden. Vielmehr scheint sich bei ähnlicher oder gleichbleibender Streuung das Einschätzniveau insgesamt zu verbessern.

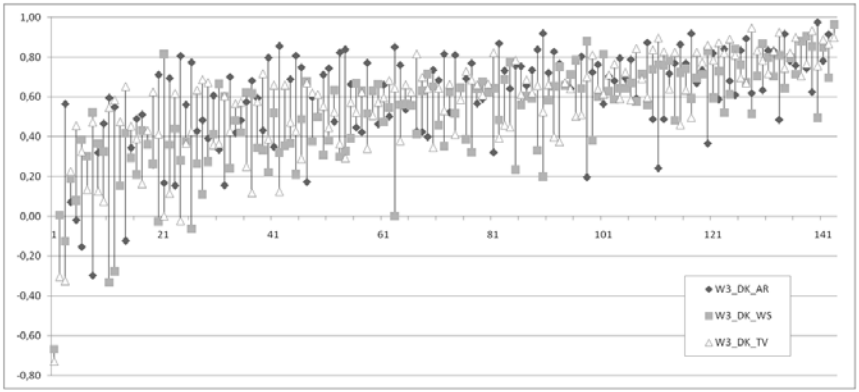

Abbildung 8: Verteilung der Güte diagnostischer Urteile für die Bereiche Arithmetik (AR), Wortschatz (WS) und Textverstehen (TV) zu t3 über alle Lehrer

7.1.4 Stabilität der Güte diagnostischer Urteile

Das Bild, das bei der Betrachtung der Güte diagnostischer Urteile über mehrere Messzeitpunkte gezeichnet werden konnte, scheint sich auch über die Halbjahre hinweg nicht wesentlich zu verändern. Bis auf einige Ausnahmen bleiben hinsichtlich der Rangkomponente hohe Korrelationen auch über die Messzeitpunkte hinweg bestehen. Dies allein ist jedoch noch kein Beleg da-

für, dass die Güte der Urteile auch für die einzelnen Lehrer ähnlich bleibt, denn theoretisch wäre denkbar, dass zu einem späteren Zeitpunkt nur die mittlere Urteilsgüte vergleichbar ausfällt, während auf individueller Lehrerebene jedoch große Verschiebungen stattfinden. Somit kann mit der mittleren Veränderung über alle Lehrer nicht festgestellt werden, wie die Variation der Urteilsgüte auf Ebene der einzelnen Lehrer über die Messzeitpunkte hinweg ausfällt. Mit der wiederholten Messung der diagnostischen Kompetenz in dieser Untersuchung kann erstmals umfassend die individuelle Stabilität dieses Konstrukts überprüft werden. Nur dann, wenn sich diese Stabilität nachweisen lässt, kann diagnostische Kompetenz auch zu Recht als Fähigkeitsdimension der Lehrer betrachtet werden.

Mangels elaborierterer statistischer Verfahren, für die die Daten geeignet wären, wird auch die Stabilität der Güte von Schülerurteilen als Korrelationsmaß berechnet. Genauer formuliert handelt es sich hierbei um die Korrelation der klassenweise erzeugten und wiederum Fisher-Z-transformierten Korrelationen zwischen Schülermerkmalen und Lehrereinschätzungen zu den drei Messzeitpunkten, wie sie schon für die beiden vorhergehenden Fragestellungen erzeugt wurden. In Tabelle 10 (S. 119) wurde bereits dargestellt, dass es zu jedem Messzeitpunkt eine gewisse Fluktuation der Klassenlehrer in der Stichprobe gab. Nur drei Viertel aller Klassen hatten vom ersten bis zum dritten Messzeitpunkt denselben Klassenlehrer. Die Berechnungen der Stabilität diagnostischer Kompetenz können daher nur sinnvoll für jene Klassen durchgeführt werden, in denen zwischen den betrachteten Zeitpunkten kein Lehrerwechsel stattgefunden hat. Darüber hinaus werden auch für diese Analysen und in Hinblick auf die Bereiche Arithmetik und mathematisches Fachinteresse nur jene Lehrer berücksichtigt, die tatsächlich Mathematik in den untersuchten Klassen unterrichten.

In der folgenden Tabelle 50 werden die Stabilitäten der Güte diagnostischer Rangurteile für jene Bereiche dargestellt, die auch zu aufeinanderfolgenden Messzeitpunkten erfasst wurden. Darüber hinaus ist auch die Stabilität von t1 zu t3 berechnet worden. In der Tabelle erkennt man, dass fast alle Stabilitäten signifikant ausfallen, nur die Stabilitäten für die Urteilsgüte in Arithmetik, bei denen der erste Messzeitpunkt beteiligt ist, erreichen nicht das Signifikanzniveau. Zu berücksichtigen ist hierbei wiederum, dass die Arithmetik- und auch die Wortschatzeinschätzungen zu t1 in ihren Formulierungen etwas von den Formulierungen zu den folgenden Erhebungszeitpunkten abwichen und daher die Ergebnisse nur als Annäherung angesehen werden sollten. Am aussagekräftigsten sind sicher die Stabilitäten zwischen

t2 und t3, da hier zum einen alle Maße identisch sind und zum anderen der Halbjahresabstand auch am geeignetsten für die Stabilitätsanalysen erscheint. Erwartungsgemäß sind diese Korrelationen in der mittleren Spalte auch durchweg am höchsten. Zwischen dem zweiten und dem dritten Messzeitpunkt sind die Stabilitäten bei den Leistungseinschätzungen deutlich höher als für t1-t2 oder t1-t3 und liegen gleichauf mit den Stabilitäten für die nicht-kognitiven Merkmale. Allein die Stabilität der Fachinteresse-Einschätzungen, besonders für Mathematik, fällt etwas hinter den anderen Stabilitäten zurück.

Tabelle 50: Stabilität der Urteilsgüte zwischen den Messzeitpunkten (Rangkomponente)

	t1 zu t2	*t2 zu t3*	*t1 zu t3*
Arithmetik	.16	.44**	.21
	N = 94	N = 118	N = 88
Wortschatz	.35**	.49**	.36**
	N = 109	N = 130	N = 106
Textverstehen	-	.47**	-
		N = 130	
Lernfreude	.38**	.42**	.29**
	N = 108	N = 129	N = 104
Schuleinstellung	.38**	.47**	.28**
	N = 107	N = 128	N = 104
Fachinteresse Deutsch	-	.40**	-
		N = 129	
Fachinteresse Mathematik	-	.28**	-
		N = 118	

**p < .01

Die durchweg signifikanten und mittelhohen Korrelationen zwischen den Ausprägungen der Urteilsgüte zu verschiedenen Zeitpunkten sind ein Indiz dafür, dass die Urteilsgenauigkeit von Lehrern in einem Bereich und zu einem Zeitpunkt keine Zufallstreffer sind. Vielmehr darf man davon ausgehen, dass ein guter Diagnostiker in einem Bereich auch ein halbes oder ein ganzes Jahr später noch ein guter Diagnostiker in diesem Bereich ist - allerdings bleibt auch ein schlechter Diagnostiker ein schlechter Diagnostiker. Diesen Schluss legen die Ergebnisse der vorangegangenen Analysen nahe, auch wenn die insgesamt nur mittelhohen Korrelationen darauf hindeutet,

dass es bezüglich der Rangplatzstabilität eine gewisse Variabilität zu geben scheint. Insgesamt bewegt sich die Höhe der Stabilitäten der Urteilsgüten etwas unterhalb der reinen Urteilsgüte.

Von Belang ist in diesem Zusammenhang auch die Frage, wie die Stabilität der Urteilsgüte im Vergleich zur Stabilität der Schülerleistungen und -merkmale sowie im Vergleich zu den Lehrerurteilen an sich einzuschätzen ist. Sehen die Lehrer über die Zeit mehr Veränderung, als dies in der Realität der Fall ist, oder verhält es sich genau umgekehrt? In Tabelle 43 (S. 172) wurden vergleichend die Stabilitäten der Schülermerkmale und der analogen Lehrereinschätzungen dargestellt. Die höchsten Stabilitäten waren auch da - wie bereits bei den Stabilitäten zur Urteilsgüte festgestellt - zwischen dem zweiten und dem dritten Messzeitpunkt zu finden, was sicher darin begründet liegt, dass hier die größten Ähnlichkeiten zwischen den eingesetzten Skalen und Items bestehen. Davon unabhängig zeigte sich für die meisten Bereiche (Ausnahme ist der Bereich Wortschatz t1-t2 und t1-t3), dass die Stabilitäten der Lehrerurteile überwiegend gleich hoch oder etwas höher ausfallen als die der Schülerleistungen und -eigenschaften. Dies lässt sich als Hinweis darauf deuten, dass sich die Leistungen und emotional-motivationalen Eigenschaften der Schüler entsprechend unserer Testverfahren als weniger stabil erweisen, als die Lehrer dies im Allgemeinen annehmen. Insgesamt scheinen auf Schülerseite die Stabilitäten der Leistungen (besonders für das Fach Deutsch) etwas höher zu sein als die der emotional-motivationalen Maße. Auch die Lehrer vermuten dort etwas höhere Konstanz.

Die vorangegangenen Analysen bezogen sich bewusst auf jene Lehrer, die zu mehreren Messzeitpunkten Klassenlehrerfunktion in den untersuchten Klassen hatten. Nur so ist die Berechnung von Stabilitäten überhaupt sinnvoll. Da Lehrer interindividuell unterschiedlich akkurat urteilen, sollten die gezeigten Stabilitäten deutlich größer ausfallen als in Klassen mit Lehrerwechsel beim Vergleich der Urteilsgüte von alten und neuen Lehrern. Entsprechende Analysen, die hier nicht gesondert dargestellt werden, bestätigten dies - die Übereinstimmungen in der Urteilsgüte zwischen verschiedenen Lehrkräften derselben Klassen schwankte bei teils sehr niedrigen Fallzahlen (N = 4 bis 29) erheblich (je nach Urteilsbereich zwischen r = -.44 und .93).

7.1.5 Reliabilität diagnostischer Urteile

In den vorherigen Abschnitten wurde auf die Struktur sowie die Güte, Homogenität und Stabilität der diagnostischen Kompetenz eingegangen. Allen jeweiligen Analysen lag - analog zur Literatur über diagnostische Kompetenz - die Annahme zugrunde, dass es sich bei dem untersuchten Konstrukt um eine Fähigkeit des Lehrers handelt. Es wurde geprüft, wie gut diese ausgeprägt ist (Güte), wie sehr sie über verschiedene Bereiche hinweg Ähnlichkeiten aufweist (Homogenität) und ob sie zeitlich stabil bleibt (Stabilität). Dabei lagen den Rechnungen jeweils Einschätzungen zugrunde, die sich immer auf Schüler derselben Klassen bezogen. Der Vorteil dieser Datengrundlage, nämlich ein längsschnittliches Design mit all seinen Vergleichsmöglichkeiten, liegt auf der Hand. Nichtsdestotrotz wäre es darüber hinaus wünschenswert, wenn Lehrerurteile nicht nur zu dieser einen, sondern zu mindestens einer weiteren Klasse vorliegen würden, um vergleichen zu können, inwiefern die Urteilsgüte möglicherweise von der Klasse selbst abhängt und - damit im Zusammenhang - ob sich eine konsistente Lehrerfähigkeit als Grundlage für die diagnostischen Urteile belegen lässt, die unabhängig von der Klasse den Lehrer charakterisiert. Dies ist insbesondere für die Rangkomponente diagnostischer Kompetenz von Bedeutung, da hierbei - im Gegensatz zu schülerspezifischen Werten bei der Niveaukomponente - immer nur ein Wert je Lehrer auf Grundlage der gesamten Klasse vorliegt und nicht ersichtlich ist, inwiefern dieser Wert Ausdruck einer lehrerspezifischen Fähigkeit oder z.B. der Zusammensetzung und Struktur der Klasse mit ihren individuellen Schülern ist. Ein derartiger Inter-Klassen-Vergleich wäre eine gute Möglichkeit, die Reliabilität der Güte der Lehrerurteile zu prüfen. Leider liegen Lehrereinschätzungen aber nur für immer dieselbe Klasse vor. Es lässt sich aber eine Annäherung an die genannte Reliabilitätsprüfung erreichen, wenn die eingeschätzten Klassen zufällig in zwei Hälften geteilt werden und anschließend die Urteilsgüte für jede Klassenhälfte bestimmt wird. Die sich somit ergebenden zwei Werte je Lehrer sollten sich, wenn eine Personenfähigkeit zugrunde liegt, sehr ähnlich sein. Anders ausgedrückt: Es sollte für die Ausprägung der Urteilsgenauigkeit keine Rolle spielen, ob in der Klasse ein paar Kinder mehr oder weniger eingeschätzt werden. Die beurteilten Schüler in dieser Untersuchung sind ohnehin niemals die komplette Klasse, da immer Schüler wegen fehlender Zustimmung der Eltern oder Abwesenheit am Testtag fehlten. Dennoch wird aus der auf Urteilen über die restlichen Kinder basierenden Urteilsgüte auf die generelle diagnostische Kompetenz geschlossen. Werden nun also die Klassen nochmals

künstlich halbiert, sollte dies auch das Ergebnis, die Urteilsgüte der Lehrer, nicht wesentlich beeinflussen.

Mittels eines Split-half-Tests soll demnach geprüft werden, wie reliabel die Rangurteilsgüte tatsächlich ist und ob man davon ausgehen kann, dass sie nicht wesentlich durch Eigenschaften der eingeschätzten Schüler bedingt ist. Dazu wurde der Datensatz derart in zwei Hälften geteilt, dass jeder zweite Schüler der anderen Gruppe zugeordnet wurde (odd-even-Methode). Vorher wurde der Datensatz zuerst nach Klassenzugehörigkeit, dann nach dem Geschlecht und zuletzt nach dem Vornamen sortiert, so wie es dem anonymisierten Schülercode entspricht. Somit kann von einer Zufallsziehung ausgegangen werden, und t-Tests bestätigen, dass sich beide Gruppen in keinem der geprüften Merkmale Geschlecht, Sozialstatus und Leistungen in allen Bereichen signifikant unterscheiden. Neben den t-Tests wurde zusätzlich die Leistungsmittelwerte und -standardabweichungen je Klassenhälfte miteinander korreliert, um auch die mittlere Stärke des Zusammenhangs auf Klassenebene prüfen zu können. Dabei erwies sich der Zusammenhang zwischen den Klassenmittelwerten in fast allen getesteten Bereichen als signifikant, während die Streuung augenscheinlich deutlich zwischen den Hälften variierte und nur in Arithmetik zu t1 signifikant korrelierte (s. Tabelle 51, Spalten 2 und 3). Weiterhin wurden alle Klassen aus dem Datensatz entfernt, für die weniger als fünf Einschätzungen je Gruppe vorlagen, so dass eine sinnvolle Berechnung von Korrelationen möglich wird. Es verblieben danach 117 verschiedene Klassen im Datensatz, wobei je nach Bereich weitere einzelne Klassen wegen fehlender Werte in den Einschätzdaten wegfielen. Für Arithmetik und das Fachinteresse Mathematik wurden wieder nur jene Klassen einbezogen, deren Lehrer auch Mathematik in den Klassen unterrichteten.

Vorgehen bei der Berechnung

Um die Split-half-Reliabilität zu bestimmen, wurden je nach Urteilskomponente unterschiedliche Vorgehensweisen gewählt. Für die Rangkomponente wurde für jede Klassenhälfte jeweils die Korrelation aus Lehrerurteilen und Schülerleistungen berechnet, wie es das grundsätzliche Vorgehen für die Berechnung der diagnostischen Kompetenz ist. Die resultierenden gruppenspezifischen Werte wurden dann Fisher-Z-transformiert, um sie auf Intervallskalenniveau zu heben. Dies ist nötig, um im Anschluss die beiden Gruppenwerte über alle Lehrer miteinander zu korrelieren (wie üblich mittels Produkt-Moment-Korrelation nach Pearson), um somit die Testhalbie-

rungsreliabilität (s. z.B. Bortz & Döring, 2006, S. 198) berechnen zu können. Während damit üblicherweise die Reliabilität eines Testverfahrens bestimmt wird, indem das Abschneiden von Probanden jeweils für eine Hälfte des Tests miteinander verglichen wird, ist die Datengrundlage im vorliegenden Anwendungsfall etwas anders. Statt der Güte eines halbierten Testverfahrens wird die Urteilsgüte der Lehrer jeweils für eine Klassenhälfte bestimmt; was sonst die Probanden des Tests sind, sind nun die Lehrkräfte selbst, und die (Beurteilungen der) Schüler stehen entsprechend für den Test. Die sich ergebende Korrelation zwischen den Rangurteilsgüten je Klassenhälfte über alle Lehrer drückt demnach die Testhalbierungsreliabilität aus.

Im Gegensatz dazu wurde für die Niveaukomponente derart vorgegangen, dass je Gruppe die mittlere Abweichung zwischen Lehrerurteilen und den fünfstufigen Schülerleistungen berechnet wurde. Auch wenn, wie bereits beschrieben, die absoluten Ausprägungen der Niveaukomponente nicht interpretierbar sind, so sollte man dennoch davon ausgehen können, dass sich zwischen zwei Klassenhälften keine bedeutsamen Unterschiede ergeben. In Erweiterung zu den bisherigen Analysen wird nun bei der Reliabilitätsbestimmung auch auf Daten aus der Zusatzstichprobe zurückgegriffen, da dort die Niveaukomponente ohne die einschränkende Transformation direkt berechnet werden kann und somit ein Vergleich zu den auf Transformation beruhenden Daten der Hauptstichprobe möglich ist. Für die Messzeitpunkte 1 bis 3 beruht die Niveaueinschätzung wiederum auf den Rangurteilen der Lehrkräfte, für die Zusatzstichprobe ist die tatsächliche Niveaukomponente zugrunde gelegt, die die Diskrepanz zwischen der Anzahl tatsächlich und geschätzt richtig gelöster Aufgaben ausdrückt. Weiterhin ist es auf Grundlage der Zusatzstichprobe möglich, auch die Rangkomponente einer weiteren Prüfung zu unterziehen, indem die Split-half-Reliabilität für die aus den Niveauurteilen gebildete Rangkomponente (auf Grundlage der sieben Aufgabeneinschätzungen statt auf Grundlage der globalen Einschätzung) berechnet wird.

Die Split-half-Reliabilität von Testverfahren wird in der Regel als r_{tt} angegeben. Dazu wird lediglich aus dem Korrelationskoeffizient mittels der Spearman-Brown-Formel

$$r_{tt} = \frac{2 \cdot r_{12}}{1 + r_{12}}$$

ein Testhalbierungskoeffizient berechnet, um den niedrigeren Werten für die Zusammenhänge, die durch die Halbierung des Stichprobenumfangs entstehen, Genüge zu leisten (s. z.B. Moosbrugger & Kelava, 2007, S. 122).

Dieses Verfahren wurde auch in der vorliegenden Untersuchung angewendet.

Ergebnisse der Reliabilitätsprüfung

Die Ergebnisse der Analysen sind in Tabelle 51 dargestellt. Bereits auf den ersten Blick stechen gravierende Unterschiede zwischen der Rang- und der Niveaukomponente ins Auge. Die Korrelationen für die Niveaukomponente (ohne Berücksichtigung des aufgabenbezogenen Treffers aus der Zusatzstichprobe) sind mit Werten zwischen r = .34 und .73 durchgängig signifikant, wobei je Messzeitpunkt die größten Zusammenhänge für den Bereich Arithmetik zu finden sind. Nichtsdestotrotz schwankt der mittlere Unterschied der Abweichung des Niveauurteils vom tatsächlichen Leistungsniveau der Schüler über alle Bereiche relativ einheitlich um 0,4 zwischen beiden Klassenhälften (g1, g2), was für die von -4 bis +4 reichende Skala aber ein moderater Wert ist. Diesem Wert liegen absolute Differenzwerte zugrunde, so dass es unerheblich ist, ob Gruppe 1 oder Gruppe 2 höher eingeschätzt wurde.

Gravierender und mit den Erwartungen nicht in Einklang stehend zeigen sich die Inter-Gruppen-Korrelationen für die Rangkomponente diagnostischer Kompetenz. Hier erweist sich über alle Messzeitpunkte und sowohl kognitive als auch nicht-kognitive Bereiche hinweg - bis auf drei Ausnahmen - kein Zusammenhang zwischen den Gruppenwerten als signifikant, die Korrelationen liegen deutlich unter jenen für die Niveaukomponentenzusammenhänge. Die Ausnahmen stellen beide Leistungsbereiche aus dem Messzeitpunkt in der dritten Klasse (Arithmetik und Wortschatz) sowie der Bereich Arithmetik aus der Zusatzstichprobe in Klassenstufe 1 dar, die - im Vergleich überraschend - sogar signifikant ausfallen.

Da auch für diese Analyse berücksichtigt werden muss, dass die angegebenen Korrelationen der Niveaukomponente auf transformierten Schülerleistungen und originären Rangurteilen der Lehrkräfte beruhen, wurde anhand der Zusatzstichprobe eine Gegenprüfung vorgenommen. In der dritt- und viertletzten Zeile von Tabelle 51 sind die Ergebnisse der Split-half-Prüfungen auf Grundlage der korrekt erfassten Niveaukomponente (auf Grundlage der sieben vorgegebenen Aufgaben) dargestellt. Diese unterscheiden sich für die beiden Bereiche Wortschatz und Textverstehen (nur für diese Bereiche war die Prüfung durch die vorliegenden Niveauurteile überhaupt möglich) trotz der deutlich geringeren Fallzahl nicht von den auf Transformation beruhenden Werten. Zum Vergleich wurde in den unters-

ten beiden Zellen derselben Tabelle ebenfalls die Split-half-Reliabilität auf Grundlage des aufgabenbezogenen Treffers berechnet, der die Anzahl von Items, bei denen das Lösungsverhalten der Schüler exakt mit den Lehrereinschätzungen übereinstimmt, angibt. Der aufgabenbezogene Treffer ist in diesem Sinne das noch akkuratere Niveaueinschätzungsmaß, das an den Lehrer aber auch deutlich höhere Anforderungen stellt. Für ihn sind die Korrelationen zwischen den Klassenhälften nur noch für das Textverstehen auf dem 5-Prozent-Niveau signifikant (r = .63), für den Bereich Wortschatz mit r = .50 jedoch nicht mehr.

Für die Rangkomponente wurde zu Prüfzwecken ein ähnliches Vorgehen wie für die Niveaukomponente angewendet. In den letzten beiden Zeilen von Tabelle 51 sind für Wortschatz und Textverstehen Rangkomponentenkorrelationen dargestellt, die auf den Niveauurteilen der Lehrer, bezogen auf die sieben einzuschätzenden Aufgaben, beruhen. Grundlage sind also die je Klassenhälfte gebildeten Korrelationen zwischen Anzahl richtig gelöster Aufgaben und der Anzahl der als richtig gelöst eingeschätzten Aufgaben durch die Lehrer. Auf diese Weise ist es möglich, auch aus Lehrerurteilen, die sich auf das Leistungsniveau der Schüler beziehen, die Rangkomponente zu bilden. Die so gewonnenen Werte unterscheiden sich nicht von den übrigen Split-Half-Reliabilitäten für die Rangkomponente diagnostischer Kompetenz. Bei deutlich geringerer Fallzahl liegen sie ebenfalls im nicht signifikanten Bereich. Somit ergeben sich sowohl für die Rang- als auch für die Niveaukomponente für beide Varianten - aus der Transformation jeweils anderer Urteile gewonnen oder direkt aus den entsprechenden Urteilen gebildet - keine unterschiedlichen Ergebnisse in Bezug auf die Testhalbierungsreliabilität.

Tabelle 51: Split-half-Reliabilitäten der Rang- und Niveaukomponente diagnostischer Kompetenz (Korrelation) sowie Korrelationen über Mittelwerte und Standardabweichungen der mittleren Leistungen und Eigenschaften je Klassenhälfte

	Korrelation über Mittelwerte (M) und Standardabweichungen (SD) der Leistungen je Klassenhälfte			*Rangkomponente*	*Niveaukomponente*
	Leistungsbereich	r_M	r_{SD}	r_{tt}	r_{tt}
t1	Arithmetik	.75**	.31**	.53** (N = 102)	.84** (N = 102)
	Wortschatz	.32**	.16	.56** (N = 114)	.66** (N = 114)
	Lernfreude	.29**	.05	.11 (N = 109)	-
	Schuleinstellung	.22*	.16	.11 (N = 103)	-
t2	Arithmetik	.50**	.11	-.33 (N = 103)	.67** (N = 104)
	Wortschatz	.32**	.12	.13 (N = 109)	.58** (N = 111)
	Textverstehen	.20*	.18	.17 (N = 111)	.56** (N = 111)
	Lernfreude	.41**	.05	.25 (N = 107)	-
	Schuleinstellung	.44**	.16	.29 (N = 103)	-
	Fachinteresse Deu.	.51**	.17	.08 (N = 112)	-
	Fachinteresse Ma.	.13	.01	-.17 (N = 103)	-
t3	Arithmetik	.65**	.15	-.08 (N = 101)	.77** (N = 105)
	Wortschatz	.28**	.09	.26 (N = 114)	.55** (N = 115)
	Textverstehen	.29**	-.01	.20 (N = 114)	.68** (N = 115)
	Rechtschreiben	.55**	.06	.13 (N = 115)	.72** (N = 115)
	logisch-abstr. Denken	.32**	.13	.04 (N = 113)	.51** (N = 113)
	Leistungsängstlichkeit	.12	.10	-.02 (N = 102)	-
	Lernfreude	.32**	.10	.17 (N = 107)	-
	Schuleinstellung	.41**	.17	.02 (N = 109)	-
	Fachinteresse Deu.	.35**	.08	-.11 (N = 114)	-
	Fachinteresse Ma.	.22*	.11	.29 (N = 100)	-
Z	Arithmetik[1]	.67**	.33	.68** (N = 27)	-
	Wortschatz[1]	.31	-.20	-.30 (N = 26)	.80** (N = 26)
	Textverstehen[1]	.30	.06	-.60 (N = 26)	.73** (N = 26)
	Wortschatz[2]	-	-	-.15 (N = 21)	.50 (N = 26)
	Textverstehen[2]	-	-	.17 (N = 21)	.63* (N = 26)

Anm.: Für die Split-Half-Reliabilitätsberechnungen für die Rangkomponente wurden - ähnlich wie bei den Stabilitätsberechnungen in Kapitel 7.1.4 - Korrelationen über die je Klassenhälfte Fisher-Z-transformierten Korrelationen zwischen Lehrerurteilen und Schülerleistungen gerechnet. Den Reliabilitätsberechnungen für die Niveaukomponente liegen die je Klassenhälfte gemittelten Niveauurteilsabweichungen zugrunde. Die fünf untersten Zeilen der Tabelle beziehen sich zum Vergleich auf die Zusatzstichprobe (Z) aus Klassenstufe 1.

[1]Die Rangkomponente wurde analog zu t1 bis t3 berechnet, die Niveaukomponente auf Grundlage der gruppenweise gemittelten Differenz zwischen Anzahl richtig gelöster und Anzahl als richtig eingeschätzter Aufgaben.

[2]Die Rangkomponente wurde auf Grundlage der Niveauurteile berechnet, die Niveaukomponente auf Grundlage des aufgabenbezogenen Treffers (Übereinstimmung von Lehrerurteilen und tatsächlichem Lösungsverhalten für sieben Aufgaben).

**p < .01, *p < .05

Es muss somit konstatiert werden, dass sich in dieser Arbeit die Rangurteile von Lehrkräften, operationalisiert als die Fähigkeit, unabhängig von konkreten Schülern und Aufgaben eine leistungsbezogene Rangfolge anzugeben, bis auf wenige Ausnahmen als nicht reliabel erweisen. Möglicherweise liegt dies weniger an tatsächlich mangelhaften Lehrerfähigkeiten, sondern eher an der Art der Operationalsierung, indem globale Lehrerurteile erfasst wurden, den Lehrern also keine konkreten Aufgaben für ihre Urteile zugrunde lagen, im Gegenzug diese globalen Urteile aber dem Lösungsverhalten bei konkreten Aufgaben gegenübergestellt wurden. Denkbar ist, dass die Rangurteile der Lehrkräfte durchaus anders ausgefallen wären, hätten sie sich an den im Test eingesetzten Aufgaben orientieren können. Allerdings deuten u.a. Ergebnisse aus dem Bereich der Sekundarstufe - im Gegensatz zum Primarbereich - darauf hin, dass globale Lehrerurteile in aller Regel sogar genauer ausfallen als spezifische (Karing, 2009).

Ganz offensichtlich kann die Korrelation zwischen Lehrerurteilen und Schülerleistungen als Maß für die diagnostische Kompetenz im Sinne der Rangkomponente stark in Abhängigkeit von der An- oder Abwesenheit einzelner Schüler, ihrer Motivation oder Konzentration variieren und ist somit als generelles oder gar alleiniges Maß für die diagnostische Kompetenz der Lehrer nur bedingt geeignet.

7.2 Bedingungen der diagnostischen Kompetenz von Grundschullehrkräften

Insbesondere die großen interindividuellen Unterschiede hinsichtlich der Güte diagnostischer Urteile zwischen Lehrkräften führen zu der Frage, wodurch diese zu erklären sind. Die infrage kommenden und in dieser Arbeit aufgegriffenen Faktoren können verschiedenen Bezugsrahmen zuge-

ordnet werden, wie es bereits in den Fragestellungen dieser Arbeit differenziert wurde (vgl. Kapitel 5.2, S. 103). Als erstes sind Ursachen dafür, ob ein Lehrer seine Schüler akkurat einzuschätzen vermag, bei Personenmerkmalen des Lehrers selbst zu suchen. Eine Reihe von Faktoren wie seine Berufserfahrung oder Perfektionsstreben werden hierbei überprüft. Weiterhin können auch Merkmale der Klasse bzw. ihre Zusammensetzung die Güte diagnostischer Urteile beeinflussen, beispielsweise durch die Anzahl der Schüler in der Klasse oder die Menge an Störungen durch die Schüler. Besonders wichtig ist aber auch der einzelne Schüler, dessen individuelle Leistungs- oder Merkmalseinschätzungen durch Lehrkräfte beispielsweise durch sein Geschlecht oder seinen sozioökonomischen Hintergrund beeinflusst sein können.

Auch für die Überprüfung der entsprechenden Zusammenhangs- oder Unterschiedshypothesen kann auf verschiedene Komponenten der diagnostischen Kompetenz zurückgegriffen werden. Nicht jede Komponente eignet sich aufgrund ihrer unterschiedlichen Datengrundlage für jede Analyse. Wie in Tabelle 52 zu sehen, bietet sich die Rangkomponente als abhängige Variable für Merkmale an, die auf der Ebene der Lehrer und/oder Klassen liegen. Soll jedoch der Einfluss individueller Schülermerkmale auf die Rangkomponente diagnostischer Kompetenz analysiert werden, so wäre dies nur möglich, wenn sie auf der Klassenebene aggregiert würden, denn nur so könnten sie mit dem ebenfalls auf Klassenebene vorliegenden Maß der Rangkomponente gegenübergestellt werden. Für diese individuellen Merkmale besser geeignet wäre ein ebenfalls auf individueller (Schüler-)Ebene vorliegendes Maß, nämlich die Niveaukomponente der Urteilsgenauigkeit. Hier wird den Merkmalen der einzelnen Schüler jeweils die Abweichung der Lehrereinschätzung vom tatsächlich gemessenen Merkmalswert gegenübergestellt. Aufgrund der transformierten Werte als Grundlage der Niveaukomponente können hier nur Unterschiedshypothesen sinnvoll geprüft werden, eine absolute Interpretation der Niveaudifferenzen ist nicht möglich. In die Analysen zur Rangkomponente werden je Messzeitpunkt grundsätzlich nur jene Lehrkräfte eingeschlossen, für die auch Daten vorliegen und bei denen die Anzahl von Einschätzungen größer oder gleich fünf ist. In Analysen zur Niveaukomponente fließen hingegen die Daten aller Lehrer ein, da das mittlere Abweichungsmaß je Klasse sich auch für weniger als fünf Schüler sinnvoll berechnen lässt.

Tabelle 52: Geeignete Kombinationen für die Überprüfung von Zusammenhangs- und Unterschiedshypothesen zwischen Lehrer-, Klassen- und Schülermerkmalen sowie den Komponenten diagnostischer Kompetenz

	Rangkomponente	*Niveaukomponente*
Lehrermerkmale	Zusammenhang zwischen Lehrermerkmalen und der Güte diagnostischer Rangurteile auf Lehrer- und Klassenebene	-
Klassenmerkmale	Zusammenhang zwischen Klassenmerkmalen und der Güte diagnostischer Rangurteile auf Lehrer- und Klassenebene	-
Schülermerkmale	-	Mittelwertsunterschiede zwischen Gruppen von Schülern bezüglich der Urteilsabweichung

7.2.1 Lehrermerkmale

Bei der Analyse der Bedingungsfaktoren diagnostischer Kompetenz sind für die Bereiche Arithmetik und Fachinteresse Mathematik jeweils nur jene Lehrkräfte einbezogen worden, die auch tatsächlich Mathematik in den untersuchten Klassen unterrichten. Weiterhin ist für Analysen, bei denen für bestimmte Gruppen, z.B. weibliche und männliche Lehrkräfte, Gruppenmittelwerte gebildet werden mussten, mit den Fisher-Z-transformierten Korrelationskoeffizienten gerechnet worden, um Unterschiede auf Signifikanz prüfen zu können. In den Ergebnistabellen sind dann wiederum die für die Gruppen zurücktransformierten „normalen“ Korrelationen angegeben worden.

Für die nachfolgenden Analysen wurden verschiedene Lehrermerkmale berücksichtigt, die teils zu allen, teils auch nur zu einigen Messzeitpunkten über den Lehrerfragebogen von den Lehrkräften erfasst wurden. Sie lassen sich grob in die Bereiche Demografie, Aus- und Weiterbildung, schul- und unterrichtsbezogene Merkmale sowie Wahrnehmung der eigenen diagnostischen Fähigkeiten unterteilen.

Berufserfahrung

Die Berufserfahrung der Lehrkräfte, operationalisiert über die Anzahl der Berufsjahre an einer Grundschule, ist in der vorliegenden Stichprobe nicht gleichmäßig verteilt. Die größte Gruppe sind zum ersten Messzeitpunkt 13 Lehrer, die weniger als ein Jahr an einer Grundschule gearbeitet haben. Ein

Viertel der Lehrer hat bis zu fünf Jahre Berufserfahrung, wohingegen das Viertel mit der meisten Berufserfahrung bereits zwischen 26 und 37 Jahre an einer Grundschule unterrichtet. Der Mittelwert liegt bei 15,4 Jahren, und die große Standardabweichung von 11,1 Jahren ist Ausdruck der großen Bandbreite zwischen den Lehrkräften. Die Werte verschieben sich für die zwei folgenden Erhebungszeitpunkte durch Lehrerwechsel und das jeweils in der Zwischenzeit verstrichene halbe Jahr geringfügig (vgl. auch Tabelle 26).

Tabelle 53: Zusammenhang zwischen der Anzahl der Berufsjahre der Lehrer an einer Grundschule und ihrer diagnostischen Kompetenz (Rangkomponente)

	t1	*t2*	*t3*
Arithmetik	.01	.06	.11
Wortschatz	.09	-.08	-.04
Textverstehen	-	-.10	.03
Rechtschreiben	-	-	.16
Fachinteresse Deutsch	-	.03	-.03
Fachinteresse Mathematik	-	-.10	-.16

*p < .05

Wie in Tabelle 53 dargestellt, lassen sich für die Rangkomponente zu keinem Messzeitpunkt Zusammenhänge zwischen der Anzahl der Grundschulberufsjahre der Lehrer und ihrer diagnostischen Kompetenz finden. Die Korrelationen schwanken geringfügig um Null. Auch alternativ gerechnete Varianzanalysen, bei denen die Lehrer entsprechend ihrer Berufserfahrung in zwei, drei oder vier Gruppen unterteilt und dann die Gruppenmittelwerte miteinander verglichen wurden, führten nicht zu signifikanten Unterschieden je nach Berufserfahrung (hier nicht separat dargestellt). Somit kann entgegen der landläufigen Vermutung davon ausgegangen werden, dass die Berufserfahrung, die häufig auch in engem Zusammenhang mit dem Grad der Expertise gesehen wird, nicht mit der Güte von Rangurteilen im Leistungsbereich und im Interessenbereich zusammenhängt.

Geschlecht

Die Geschlechterverteilung in der Stichprobe fällt deutlich zugunsten der Frauen aus. Zu t1 waren knapp 83 Prozent der untersuchten Lehrer, die Angaben über ihre Geschlechterzugehörigkeit machten, Frauen (N = 120). Ihnen standen 25 männliche Lehrer gegenüber (vgl. Tabelle 28, S. 143). Dies

ist bei der Bewertung der Analyseergebnisse entsprechend zu berücksichtigen.

Tabelle 54: Mittlere Urteilsgüte in Abhängigkeit vom Geschlecht der Lehrer

	t1		*t2*		*t3*	
	m/w Rang	m/w Niv.	m/w Rang	m/w Niv.	m/w Rang	m/w Niv.
Arithmetik	.55/.51	4,02/3,90*	.62/.54^{+}	4,20/4,10	.56/.53	4,14/4,17
Wortschatz	.60/.55	4,18/4,17	.51/.57	4,13/4,10	.48/.46	4,26/4,20
Textverstehen	-	-	.58/.58	4,20/4,25	.52/.52	4,27/4,30
Rechtschreiben	-	-	-	-	.62/.59	4,20/4,10
Fachinteresse Deutsch	-	-	.22/.26	-	.27/.29	-
Fachinteresse Mathem.	-	-	.31/.34	-	.29/.32	-

Anm.: Dargestellt sind für die Rangkomponente die mittlere Korrelation je Geschlecht, für die Niveaukomponente die mittlere Abweichung auf der Skala von 1 bis 5 (Optimum bei 5). Signifikanzen der t-Tests sind als Sternchen dargestellt. m=männlich, w=weiblich, Rang=Rangkomponente, Niv.=Niveaukomponente.
*p < .05, $^{+}$p < .10

Bezogen auf die Rangkomponente weisen männliche Lehrer zu t1 in den untersuchten Bereichen zwar eine etwas höhere Einschätzgüte auf als weibliche, die Unterschiede sind jedoch nach den verwendeten t-Tests nicht signifikant. Zu t2 zeigt sich ein signifikanter Unterschied für die Arithmetikeinschätzungen zugunsten der männlichen Lehrer, in den meisten anderen Bereichen sind nun aber die Lehrerinnen die - wenn auch nicht signifikant - besseren Urteiler. Zu t3 gibt es keine bedeutsamen Unterschiede mehr zwischen Männern und Frauen.

Für die Niveaukomponente zeichnet sich ein vergleichbares Bild ab. Auch hier urteilen männliche Lehrkräfte etwas genauer im Bereich Arithmetik, allerdings zu t1 und nicht zu t2, wohingegen es in allen anderen Bereichen und Messzeitpunkten keine signifikanten Geschlechterunterschiede in Bezug auf die Einschätzgenauigkeit des Leistungsniveaus gibt. Insgesamt lässt sich somit keine einheitliche Aussage dazu treffen, ob männliche oder weibliche Lehrkräfte (in bestimmten Bereichen) die genaueren Urteiler sind.

Eine besondere Bedeutung kommt in diesem Zusammenhang sicher auch der Relation aus Lehrer- und Schülergeschlecht zu. Analysen hierzu erfordern jedoch die Betrachtung nicht nur der Lehrer-, sondern auch der individuellen Schülerebene, weshalb sie erst in Kapitel 7.2.3 (ab S. 215) aufgegriffen werden.

Lehrdauer in der jetzigen Klasse

Die Verteilung der Lehrdauer in den Klassen ist sehr uneinheitlich (vgl. Tabelle 10, S. 119). Gut ein Fünftel der Lehrkräfte unterrichtete zum ersten Messzeitpunkt am Ende der dritten Klasse bereits seit Beginn der Grundschulzeit, knapp 8 Prozent seit Beginn der 2. Klasse und knapp 70 Prozent seit dem dritten Schuljahr. Somit sind die verschiedenen Lehrergruppen sehr ungleich besetzt, was die Aussagekraft der Analysen - ähnlich wie bei der vorherigen Betrachtung der Geschlechtsunterschiede - stark beeinflusst. Die mittlere diagnostische Kompetenz variiert nach varianzanalytischer Betrachtung zwischen den Gruppen und für die jeweils betrachteten Leistungs- und Fachinteressenbereiche für beide Komponenten kaum und bis auf den Bereich Arithmetik zu t3 nicht signifikant (Tabelle 55). Ein einheitlicher linearer Trend, demzufolge mit steigender Lehrdauer auch die Urteilsgüte höher ausfällt, ist nicht zu erkennen, auch wenn Lehrer, die die Schüler seit der ersten Klassenstufe kennen, in aller Regel akkuratere Urteile fällen als Lehrer, die erst seit der 4. Jahrgangsstufe in den Klassen unterrichten. Dem gegenüber erscheint häufig die Urteilsgüte jener Lehrer, die die Schüler seit der zweiten Klasse unterrichten, besonders hoch, allerdings nicht in den arithmetikbezogenen Bereichen und im Bereich Rechtschreiben. Trotz signifikanter Unterschiede zwischen den vier Lehrergruppen im Bereich Arithmetik zu t3 zeigt sich auch hier kein linearer Zusammenhang zwischen Urteilsgüte und Lehrdauer.

Tabelle 55: Mittlere Rangurteilsgüte in Abhängigkeit von der Lehrdauer in der jetzigen Klasse

	t1	*t2*	*t3*
	1./2./3.	1./2./3./4.	1./2./3./4.
Arithmetik	.58/.48/.50	.57/.49/.54/.54	.60/.39/.50/.55*
Wortschatz	.60/.66/.53	.59/.70/.53/.51	.49/.55/.45/.45
Textverstehen	-	.61/.70/.56/.56	.50/.57/.50/.54
Rechtschreiben	-	-	.60/.55/.61/.58
Fachinteresse Deutsch	-	.31/.34/.22/.22	.25/.39/.25/.25
Fachinteresse Mathematik	-	.37/.42/.33/.27	.37/.21/.35/.29

Anm.: 1. = Lehrer unterrichtet die Klasse seit Beginn der Grundschulzeit, 2. = ... seit Beginn der 2. Klasse, 3. = ... seit Beginn der 3. Klasse, 4. = ... seit Beginn der 4. Klasse
*p < .05

Fähigkeit zur Perspektivenübernahme

Als sehr bedeutsam für die Einschätzung von Personen oder Personenmerkmalen wird die Fähigkeit zur Perspektivenübernahme angesehen. Sie gilt als Voraussetzung dafür, die Positionen anderer Personen einzunehmen und aus deren Sicht zu denken. Damit steht sie in engem Verhältnis zu diagnostischen Prozessen, bei denen genau dies - z.B. aus Schülersicht den Lernfortschritt oder den Wissensstand abschätzen - erforderlich ist. In der vorliegenden Untersuchung ist diese Fähigkeit mittels einer vierstufigen Skala (Werte 1 bis 4) mit fünf Items zum zweiten Messzeitpunkt erfasst worden (vgl. Tabelle 33, S. 146). Die Lehrkräfte schätzten dabei ihre eigene Fähigkeit im Mittel als relativ hoch ein (M = 3,3, SD = 0,4). Auch hier erreichen nur 14,6 Prozent der Lehrer einen Skalenmittelwert unter 3,0, allerdings ist die Mittelwertsverteilung im Wertebereich zwischen 2,4 und 4,0 annähernd normalverteilt (Schiefe = -0,05).

Erstaunlicherweise zeigen sich als Ergebnis der Korrelation zwischen der selbst eingeschätzten eigenen Fähigkeit zur Perspektivenübernahme und der Güte diagnostischer Urteile für die Rangkomponente überwiegend negative Korrelationen, die jedoch für keinen der Bereiche signifikant ausfallen (vgl. Tabelle 56). Der vermutete Zusammenhang, dass mit größerer Fähigkeit zur Perspektivenübernahme auch die Genauigkeit der diagnostischen Urteile zunimmt, kann somit nicht bestätigt werden.

Tabelle 56: Zusammenhang zwischen der Fähigkeit der Lehrer zur Perspektivenübernahme und ihrer diagnostischen Kompetenz (Rangkomponente)

	t2
Arithmetik	.06
Wortschatz	.00
Textverstehen	-.14
Fachinteresse Deutsch	-.08
Fachinteresse Mathematik	-.02

Weiterbildung

Unter der Annahme, dass diagnostische Kompetenz ebenso wie andere Kompetenzen prinzipiell erlernbar ist, sollten Lehrkräfte, die während ihres Studiums eine explizite Diagnostik-Lehrveranstaltung oder nach ihrem Studium eine Weiterbildung zum Thema Diagnostik besucht haben, Vorteile gegenüber anderen Kollegen und somit eine bessere Einschätzgüte aufwei-

sen. Im Lehrerfragebogen zu t2 sind die Lehrer deshalb gefragt worden, ob sie sowohl während ihres Lehramtsstudiums oder aber in einer Weiterbildung explizit eine Veranstaltung zur diagnostischen Kompetenz besucht haben. Wie in Tabelle 29 (Seite 144) dargestellt, hat etwas mehr als die Hälfte (53,3%) aller Lehrkräfte weder während des Studiums noch danach eine Diagnostikveranstaltung besucht, wohingegen ungefähr jeder fünfte Lehrer schon einmal an beiden Arten von Diagnostikveranstaltung teilgenommen hat. Um zu überprüfen, ob dies einen Einfluss auf die Güte diagnostischer Urteile hat, werden jene 65 Lehrer, die keine solcher Veranstaltungen besucht haben, jenen Lehrern gegenübergestellt, die mindestens einmal an einer teilgenommen haben (N = 57), so dass sich annähernd eine hälftige Verteilung ergibt.

Wie Tabelle 57 verdeutlicht, lässt sich statt des vermuteten Effekts eher das Gegenteil finden. Die Unterschiede hinsichtlich der Güte diagnostischer Rangurteile in Abhängigkeit vom Besuch einer Veranstaltung zur Diagnostik sind zwar lediglich in den Bereichen Wortschatz und Textverstehen tendenziell auf dem 10-Prozent-Niveau signifikant, jedoch liegt die Urteilsgüte jener Lehrer, die weder in ihrem Studium noch danach eine Diagnostikveranstaltung besucht haben, zahlenmäßig stets über der Güte jener Lehrer, die den Besuch einer derartigen Veranstaltung angegeben haben. Ausnahmen sind die Fachinteressensbereiche, wo dieses Verhältnis sich zugunsten der Diagnostikveranstaltungsteilnehmer umkehrt. Dies entspricht nicht den Erwartungen. Lehrkräften, die in speziellen Veranstaltungen für wichtige Faktoren im Urteilsprozess, ein Bewusstsein zu den Komponenten der Urteilsgenauigkeit oder mögliche Urteilsverzerrungen sensibilisiert wurden, wäre eine höhere diagnostische Kompetenz als ihren Kollegen ohne spezielle Ausbildung zuzutrauen gewesen. Dies ist jedoch offensichtlich nicht der Fall.

Tabelle 57: Zusammenhänge zwischen der Teilnahme an mindestens einer und keiner Lehr- oder Weiterbildungsveranstaltung zur Diagnostik und der diagnostischen Kompetenz der Lehrer (Rangkomponente)

	Ja, Veranstaltung besucht	*Nein, Veranstaltung nicht besucht*
Arithmetik	.53	.57
Wortschatz	.51^{+}	.59^{+}
Textverstehen	.55^{+}	.61^{+}
Fachinteresse Deutsch	.26	.25
Fachinteresse Mathematik	.36	.32

Anm.: Die Signifikanz bezieht sich auf den Unterschied in der Urteilsgüte je nachdem, ob eine Veranstaltung zur Diagnostik besucht wurde.
$^{+}p < .10$

Perfektionsstreben

Streben Lehrer nach tadelloser Arbeit und Perfektion, so sollten sie sich auch bei der Schülereinschätzung besonders große Mühe geben, Fehler zu vermeiden. Entsprechend wird angenommen, dass Lehrer mit hohem Perfektionsstreben auch gute Diagnostiker sind. In der Stichprobe liegt der Mittelwert der eingesetzten fünfstufigen Sechs-Item-Skala bei 3,4 (SD = 0,8) bei einem Range von 1,0 bis 5,0. 73,1 Prozent der Lehrer liegen damit über dem Mittelwert der Skala.

Wie die bisher betrachteten Lehrermerkmale, für die ein Zusammenhang zur diagnostischen Kompetenz angenommen wurde, erweist sich ebenfalls das Streben nach Perfektion als unbedeutend für deren Erklärung. Es finden sich keine signifikanten Zusammenhänge, sondern stattdessen eher noch eine Häufung von minimal negativen Korrelationen, was eine zur Erwartung gegenläufige Tendenz zeigt.

Tabelle 58: Zusammenhang zwischen dem Perfektionsstreben der Lehrer und ihrer diagnostischen Kompetenz (Rangkomponente)

	t3
Arithmetik	-.03
Wortschatz	-.02
Textverstehen	-.02
Rechtschreiben	-.07
Fachinteresse Deutsch	.01
Fachinteresse Mathematik	-.02

Einstellung zur Diagnostischen Kompetenz

Auch die Einstellung der Lehrer zur diagnostischen Kompetenz spielte eine Rolle in der Befragung. Dass es wichtig sei, die Persönlichkeit der Schülerinnen und Schüler in der Klasse einschätzen zu können, sahen 99,3 Prozent der Lehrer so („stimme zu", „stimme eher zu"), nur ein Lehrer stimmte dieser Aussage nicht zu. Geteilter war das Meinungsbild bei der Frage danach, ob man auch ohne Diagnostik-Ausbildung Leistungen und Eigenschaften der Schüler richtig einschätzen kann. Ein knappes Drittel der Lehrkräfte stimmte dieser Aussage nicht oder eher nicht zu. Wie zu erwarten, weist die Einstellung zu dieser Frage laut Chi-Quadrat-Test ($\chi^2 = 22{,}53$, df = 3, p = .000) einen signifikanten Zusammenhang dazu auf, ob die Lehrer selbst im Studium oder als Weiterbildung eine Veranstaltung zur diagnostischen Kompetenz besucht haben.

Für die Frage, ob die Bedeutungsbeimessung zur Diagnostikausbildung im Zusammenhang mit der gemessenen diagnostischen Kompetenz der Lehrer steht, wurde beides miteinander korreliert. Die sich ergebenden Koeffizienten für die Rangkomponente deuten jedoch nicht darauf hin, dass die diesbezügliche Lehrermeinung mit der Urteilsgüte korreliert (vgl. Tabelle 59). Wie schon bei vorhergehenden Fragestellungen fällt auch hier das Ergebnis nicht erwartungskonform aus. Erwartet wurde, dass Lehrkräfte, die die Bedeutung einer Diagnostikausbildung für groß halten, auch akkuratere Einschätzungen der Leistungen und Fachinteressen ihrer Schüler vornehmen können.

Tabelle 59: Zusammenhang zwischen der Lehrermeinung zur Wichtigkeit einer Diagnostikausbildung für die korrekte Schülereinschätzung und ihrer diagnostischen Kompetenz (Rangkomponente)

	t2
Arithmetik	-.01
Wortschatz	.10
Textverstehen	.12
Fachinteresse Deutsch	-.09
Fachinteresse Mathematik	-.05

Selbstwahrnehmung der eigenen diagnostischen Kompetenz im Leistungsbereich

Zu allen drei Erhebungszeitpunkten wurde auch eine Skala eingesetzt, die die Selbstwahrnehmung der eigenen diagnostischen Fähigkeiten im Leistungsbereich mit fünf Items erfasst (s. Tabelle 31, S. 145). Mit ihr soll überprüft werden, inwiefern oder ob zwischen der von uns ermittelten diagnostischen Kompetenz und den Selbsteinschätzungen ein Zusammenhang besteht. Eine realistische Selbsteinschätzung ist im Unterschied zu den anderen untersuchten Variablen zwar nicht direkt als Bedingung diagnostischer Kompetenz anzunehmen, dennoch verspricht sie aufschlussreiche Einblicke in das Konzept aus Lehrersicht und die Brauchbarkeit als Vorhersagemöglichkeit für die tatsächliche Einschätzgenauigkeit. Da sich die Skala auf die Einschätzung von Leistungen bezieht, werden für die Analyse die emotionalen und motivationalen Variablen vernachlässigt. Gerechnet wurden Korrelationen zwischen der erfassten Güte diagnostischer Urteile je Leistungsbereich und dem Skalenwert je Lehrer. Hypothetisch sollten sich positive Zusammenhänge zeigen, so dass bei realistischer Selbsteinschätzung Lehrer mit hoher diagnostischer Kompetenz sich auch als besser einschätzen als Lehrer mit niedrigerer diagnostischer Kompetenz. Nachteilig wirkt sich bei dieser Analyse aus, dass die Selbsteinschätzung nicht bereichsspezifisch, sondern nur allgemein erfragt wurde. Gegeben die Erkenntnis, dass die diagnostische Kompetenz je nach Bereich unterschiedlich ausfallen kann, wäre für diese Analyse wichtig zu wissen, ob, und wenn ja, an welchem Bereich sich die Lehrer bei ihrer Selbsteinschätzung orientierten. Diese Information liegt jedoch nicht vor.

Die Ergebnisse der Analysen sind in Tabelle 60 dargestellt. Die Selbsteinschätzung der diagnostischen Fähigkeiten korreliert für die Rangkomponente in keinem der Bereiche mit der gemessenen Urteilsgüte. Die Aussagekraft

dieser Analyse wird wahrscheinlich auch dadurch verringert, dass nur recht wenige Lehrer sich selbst für weniger gute Diagnostiker gehalten haben. Nur zwischen 7,2 und 11,8 Prozent der Lehrer erreichten - je nach Messzeitpunkt - auf der vierstufigen Skala Werte unter 3,0, alle anderen stimmen den Aussagen eher oder ganz zu. Diese geringe Varianz auf Lehrerseite ist für die Analyse nicht unproblematisch.

Tabelle 60: Zusammenhang zwischen der Selbstwahrnehmung der eigenen diagnostischen Kompetenz der Lehrer im Leistungsbereich und ihrer diagnostischen Kompetenz (Rangkomponente)

	t1	*t2*	*t3*
Arithmetik	.03	-.00	.06
Wortschatz	-.05	-.02	-.04
Textverstehen	-	-.01	.04
Rechtschreiben	-	-	.11

Parallel wurden auch die Schüler zu t2 und t3 gebeten, die diagnostischen Fähigkeiten ihrer Lehrer auf derselben Skala einzuschätzen, wenngleich die Skalenausprägungen mit ‚stimmt nicht', ‚stimmt eher nicht', ‚stimmt eher' und ‚stimmt' sich minimal von der ‚trifft zu'-Formulierung im Lehrerfragebogen unterschied. Auch sie beurteilen die diagnostische Kompetenz ihrer Lehrer überwiegend positiv. Sie beurteilten ihre Lehrer zu t2 und t3 jeweils im Mittel mit 3,2 (SD = 0,7) auf der vierstufigen Skala, nur 14,6 (t2) bzw. 15,0 (t3) Prozent der Schüler schätzten die diagnostischen Fähigkeiten mit Werten unter 3 ein. Die Mittelwerte dieser Einschätzungen des Lehrers auf Klassenebene verteilen sich annähernd normal und variieren je nach Messzeitpunkt zwischen 2,4 und 3,9, wobei die Standardabweichung auf Klassenebene mit 0,3 bei gleichem Mittelwert deutlich geringer ausfällt als auf der Individualebene. Die sowohl auf Lehrer- als auch auf Schülerseite insgesamt sehr positiv wahrgenommenen diagnostischen Fähigkeiten der Lehrkräfte führen allerdings nicht zu einer erwartbaren Korrelation von Lehrer- und Schülersicht. Die mittlere Schülereinschätzung pro Klasse korreliert mit der Selbstwahrnehmung der Lehrer zu t2 nur zu r = .14, zu t3 zu r = .10. Die Wahrnehmungen sind damit alles andere als deckungsgleich, es besteht nahezu kein Zusammenhang.

Schließlich interessiert, ob die Güte der Lehrerurteile analog zu Tabelle 60 im Zusammenhang mit den mittleren Schülereinschätzungen der diagnostischen Lehrerfähigkeiten steht (s. Tabelle 61). Doch auch hier gibt es in keinem Leistungsbereich signifikante Korrelationen, wenngleich die sehr nied-

rigen Zusammenhänge zumindest zu t3 tendenziell alle in Richtung eines positiven Zusammenhangs weisen. Es zeigt sich, dass weder die Selbst- noch die Fremdeinschätzung der Fähigkeit, Leistungen korrekt einzuschätzen, ein valider Indikator für die gemessene Rangkomponente diagnostischer Kompetenz sind.

Tabelle 61: Zusammenhang zwischen der Wahrnehmung der diagnostischen Kompetenz im Leistungsbereich durch die Schüler und der gemessenen diagnostischen Kompetenz der Lehrer (Rangkomponente)

	t2	*t3*
Arithmetik	-.06	.07
Wortschatz	.02	.11
Textverstehen	.07	.10
Rechtschreiben	-	.14

Schwierigkeiten beim Beurteilen

Zum dritten Erhebungszeitpunkt wurden die Lehrer um eine Einschätzung gebeten, wie lange sie für die individuelle Einschätzung jedes Schülers hinsichtlich seiner Leistungen und emotional-motivationalen Eigenschaften benötigt haben und wie sicher sie sich dabei jedes Mal bei ihrem Urteil waren. Um aus diesen Angaben ein lehrerspezifisches Merkmal zu generieren, wurden die individuellen Angaben zunächst klassenweise aggregiert und dann mit der Urteilsgüte korreliert. Lehrer, die die Urteile schnell fällen können und sich dabei sicher sind, so die Annahme, sollten über eine höhere diagnostische Kompetenz verfügen als unsichere, langsamer urteilende Lehrer. Zu berücksichtigen ist auch bei dieser Analyse, dass es bezüglich der Sicherheit bei der Einschätzung nur wenig Varianz gibt. Die Mittelwerte liegen auf der vierstufigen Skala bei 3,5 (Deutsch), 3,4 (Mathematik) und 3,3 (emotional-motivationale Eigenschaften), die Standardabweichungen dazu zwischen 0,3 und 0,4. Nur 2,1 Prozent der Lehrer haben für den Deutschbereich einen Mittelwert unter 3,0, für Mathe sind es - ohne nicht Mathematik unterrichtende Lehrer - 3,2 Prozent, und für die emotional-motivationalen Maße 15,3 Prozent.

Tabelle 62: Zusammenhang zwischen dem mittleren Zeitbedarf für die Einschätzung der einzelnen Schüler sowie der Sicherheit bei der Einschätzung der Fähigkeiten in den Fächern und des Fachinteresses und der diagnostischen Kompetenz der Lehrer (Rangkomponente)

t3	*mittlerer Zeitbedarf für die Schülereinschätzungen*	*Sicherheit bei Einschätzung von mathematikbezogenen Leistungen*	*Sicherheit bei Einschätzung von sprachbezogenen Leistungen*	*Sicherheit bei Einschätzung emotional-motivationaler Eigenschaften*
Arithmetik	.03	.03	-	-
Wortschatz	.01	-	.03	-
Textverstehen	-.03	-	.05	-
Rechtschreiben	-.13	-	-	-
Fachinteresse Deutsch	-.03	-	-	-.13
Fachinteresse Mathematik	-.09	-	-	-.13

Aus Tabelle 62 ist abzulesen, dass sowohl der mittlere Zeitbedarf für die Einschätzungen als auch das Ausmaß der Sicherheit dabei für keinen Bereich einen Einfluss auf die Urteilsgüte je Lehrer hat. Ist die Selbsteinschätzung der Lehrer als realistisch anzusehen, so sind also auch die mittlere Schnelligkeit und Sicherheit bei der Einschätzung keine Indikatoren für hohe diagnostische Kompetenz. Darüber hinaus zeigt sich jedoch, dass die Sicherheit bei der Einschätzung relativ homogen zwischen den Bereichen ausgeprägt ist. Die Korrelation zwischen der Einschätzsicherheit für sprach- und für mathematikbezogene Bereiche liegt bei signifikanten r = .88, zwischen sprachlichen und den emotional-motivationalen Maßen beträgt sie r = .60, und zwischen mathematischen und den emotional-motivationalen Maßen bei r = .63, beides ebenfalls signifikant auf dem 1-Prozent-Niveau. Dies kann als Hinweis darauf gewertet werden, dass die Lehrer bei ihrer Einschätzung der eigenen Sicherheit nicht besonders nach Bereichen differenziert haben. Weiterhin zeigen sich schwache negative, jedoch nicht signifikante Korrelationen zwischen dem durchschnittlichen Zeitbedarf für die Einschätzungen und der Sicherheit der Einschätzungen zwischen r = -.15 und r = -.13. Tendenziell sind sich somit die Lehrer umso sicherer bei ihren Urteilen, je schneller sie sie fällen.

Neben den auf Klassenebene aggregierten Angaben zur Sicherheit und zum Zeitbedarf bei den Einschätzungen kann deren Einfluss auch auf individuel-

ler Ebene untersucht werden, indem sie nicht mit der klassenweise gebildeten Rangkomponente, sondern mit den durch Transformation gewonnenen individuellen Niveauurteilen (vgl. Erläuterung in Kapitel 6.3) in Beziehung gesetzt werden. Lehrkräfte, die sich bei den Einschätzungen eher oder sehr unsicher waren, unterschieden sich hinsichtlich ihrer Niveaueinschätzung in keinem der Leistungsbereiche von jenen Kollegen, die sich eher oder sehr sicher bei den Schülereinschätzungen waren (Tabelle 63). Zu berücksichtigen ist hier allerdings, dass entsprechend der Angaben nur 1,9 Prozent der Lehrer bei der Einschätzung von Items aus dem sprachlichen Bereich und 5,5 Prozent der tatsächlich Mathematik unterrichtenden Lehrer bei der Einschätzung (Abweichungsmaß) von Mathematikitems unsicher waren. Die Korrelation der transformierten Niveaueinschätzungen mit den Angaben der Lehrer zum Zeitbedarf bei den Einschätzungen werden - bis auf die Ausnahme Wortschatz - trotz niedriger negativer Werte - signifikant. Je schneller die Lehrer die Leistungseinschätzungen vornahmen, desto weniger wichen ihre Niveauurteile von den Schülerleistungen ab. Die niedrigen Werte deuten allerdings auch darauf hin, dass dieser Zusammenhang weit von einem linearen Trend entfernt ist. Bei gleichem Zeitbedarf gibt es eine erhebliche Streuung der entsprechenden Differenzwerte bei der Niveaueinschätzung.

Tabelle 63: Unterschiede in der Niveaukomponente diagnostischer Kompetenz in Abhängigkeit vom Ausmaß der Schwierigkeiten beim Beurteilen, getrennt nach oberer und unterer Hälfte der Verteilung (Mediansplit), sowie Korrelation des Abweichungsmaßes der Niveauurteilsgüte mit dem Zeitbedarf bei der Einschätzung

		eher oder sehr unsicher bei der Einschätzung	*eher oder sehr sicher bei der Einschätzung*	*T*	*Korrelation der Güte der Niveauurteile (Abweichungsmaß) mit dem Zeitbedarf bei den Einschätzungen*
t3	Arithmetik	0,52	0,57	-0,50	-.12**
	Wortschatz	-0,13	0,08	-1,24	-.03
	Textverstehen	0,05	0,04	0,07	-.05*
	Rechtschreiben	-0,95	-0,71	-1,05	-.09**

Anm.: Für den Bereich Arithmetik wurden die Lehrerantworten zur Frage, wie sicher sie sich bei der Einschätzung der Schüler im Fach Mathematik waren, für alle anderen Bereiche wurde die entsprechende Frage bezogen auf das Fach Deutsch als unabhängige Variable verwendet.
**p < .01, *p < .05

Die Annahme, dass einzelne der in dieser Arbeit untersuchten Lehrermerkmale Prädiktoren für die Güte diagnostischer Urteile sind, lässt sich nach den vorliegenden Ergebnissen nicht bestätigen. Wenn sich signifikante

Zusammenhänge zeigten, dann in aller Regel nur für einzelne Bereiche, was aufgrund der Annahme der Bereichsspezifität diagnostischer Kompetenz noch nicht ungewöhnlich wäre. Signifikante Zusammenhänge zeigten sich darüber hinaus aber auch nicht stabil über mehrere Messzeitpunkte und oftmals eher in eine der Erwartung entgegengesetzte Richtung. Im Folgenden wird der Fokus von Merkmalen der Lehrer auf Merkmale der Klasse verlegt.

7.2.2 Klassenmerkmale

Neben den eben untersuchten Lehrermerkmalen als potentiellen Einflussfaktoren auf die Güte diagnostischer Urteile sind auch Merkmale der Klasse plausibel. Sie sind wie die Lehrermerkmale auf der Klassenebene angesiedelt, so dass auch hier wieder Korrelationen als Analysemethode zum Einsatz kommen.

Klassengröße

Die Klassengröße wird - besonders von Lehrern selbst - immer wieder als wichtiger Faktor für Lehrbelastung und Unterrichtsqualität angesehen. Dies kann zweifellos richtig sein, doch ob oder inwieweit sich die Anzahl der Kinder in der Klasse auch auf die Urteilsgenauigkeit der Lehrer auswirkt, ist wenig untersucht. Bislang deutet sich allerdings an, dass die Urteilsgenauigkeit nicht mit der Klassengröße in Zusammenhang steht (u.a. Wild & Rost, 1995).

Zwischen der Klassengröße und der Urteilsgüte der Lehrer in Bezug auf die Rangkomponente zeigt sich nur zum dritten Messzeitpunkt für den Bereich des Fachinteresses Deutsch eine signifikante Korrelation, allerdings nicht in der erwarteten Richtung (s. Tabelle 64). Da alle anderen Zusammenhänge jedoch gänzlich unbedeutend ausfallen und keine einheitliche Tendenz zu erkennen ist, ist auch ein Zufallszusammenhang nicht auszuschließen.

Tabelle 64: Zusammenhang zwischen der Klassengröße und der diagnostischen Kompetenz der Lehrer (Korrelation)

	t1	*t2*	*t3*
Arithmetik	-.08	-.10	-.04
Wortschatz	-.09	.08	.00
Textverstehen	-	.04	.12
Rechtschreiben	-	-	-.09
Fachinteresse Deutsch	-	.04	.20*
Fachinteresse Mathematik	-	-.07	-.06

*p < .05

Die mittlere Klassengröße liegt für die ersten beiden Messzeitpunkte jeweils bei 22,7 (SD = 3,8 bzw. 3,6), für den dritten Messzeitpunkt bei 22,6 (SD = 3,8) und weist dabei einen Range von 13 bis 31 auf (t2: 14 bis 31). Es ist nicht anzunehmen, dass ein Schüler mehr oder weniger in der Klasse die Urteilsgüte der Lehrer maßgeblich beeinflusst. Denkbar sind aber Unterschiede zwischen besonders kleinen und größeren Klassen. So könnte in Klassen mit maximal 20 Schülern ein intensiverer Unterricht stattfinden, in dem Lehrer besser auf Einzelne eingehen können, wohingegen dies in Klassen mit mehr als 20 Schülern nicht mehr der Fall wäre. Jene Klassen, in denen maximal 20 Schüler sind, machen in der vorliegenden Stichprobe ca. 25 Prozent aus. Um zu überprüfen, ob möglicherweise Unterschiede zwischen diesem Viertel und den restlichen größeren Klassen bezogen auf die diagnostische Kompetenz der Lehrkräfte zeigen, werden sie im Folgenden mittels t-Test verglichen. Dieses Vorgehen unterscheidet sich vom Vorhergehenden, da es keinen linearen Zusammenhang annimmt.

In Tabelle 65 sind jeweils nach Bereich und Messzeitpunkt getrennt die mittlere Urteilsgüte für alle Lehrer mit 20 oder weniger Kindern im Vergleich zu den Lehrern mit größeren Klassen dargestellt. Die t-Tests für die Rangkomponente wurden wiederum mit Fisher-Z-transformierten Werten gerechnet und anschließend rücktransformiert. Anhand der Rangkomponentenergebnisse lassen sich keine eindeutigen Gruppenunterschiede feststellen. Die größte Differenz zwischen kleinen und großen Klassen gibt es auch hier für den Bereich des Deutsch-Fachinteresses zu t3, es wird aber knapp das fünfprozentige Signifikanzniveau verfehlt. Somit kann ein Einfluss der Klassengröße auf die Güte diagnostischer Urteile de facto ausgeschlossen werden.

Tabelle 65: Unterschiede in der diagnostischen Kompetenz der Lehrer (Rangkomponente), getrennt für Klassen mit 20 Kindern oder mehr

		t1	*t2*	*t3*
Arithmetik	bis 20	.56	.56	.66
	ab 21	.50	.55	.65
Wortschatz	bis 20	.58	.51	.56
	ab 21	.55	.57	.54
Textverstehen	bis 20	-	.59	.62
	ab 21		.58	.61
Rechtschreiben	bis 20	-	-	.76
	ab 21			.72
Fachinteresse Deutsch	bis 20	-	.29	.18
	ab 21		.24	.33
Fachinteresse Mathematik	bis 20	-	.41	.37
	ab 21		.31	.36

Anzahl einzuschätzender Schüler

In der vorliegenden Stichprobe ist die Klassengröße allerdings nicht identisch mit den tatsächlich eingeschätzten Schülern. Als nächstes wird daher untersucht, ob die Anzahl der einzuschätzenden Schülern mit der Urteilsgüte im Zusammenhang steht. Die durchschnittliche Klassengröße lag für die drei Messzeitpunkte bei 22,7 (t3: 22,6), wohingegen die Teilnahmequote (nach Imputation) immer 15,5 Schüler pro Klasse betrug. Der Zusammenhang zwischen Klassengröße und Teilnahmequote ist mit Werten zwischen r = .46 und .52 zwar immer signifikant, aber dennoch lässt sich die Quote nicht automatisch aus der Klassengröße ableiten oder umgekehrt. Deshalb wird im nächsten Schritt separat geprüft, inwiefern sich die Anzahl einzuschätzender Schüler auf die Güte der Lehrerurteile auswirkt.

Dabei zeigt sich - wie in Tabelle 66 dargestellt - zu t3 für die Rechtschreibung (Rangkomponente) mit r = -.21 ein auf dem 5-Prozent-Niveau signifikanter Zusammenhang. Dieser Befund trifft die Erwartungen, dass bei niedrigerer Anzahl von Einschätzungen durch die damit verbundene geringere Arbeitsbelastung und erhöhte Konzentration die Urteilsgüte ansteigt. Er lässt sich aber nicht durch ähnliche Korrelationen zu anderen Messzeitpunkten oder in anderen Bereichen replizieren, weshalb davon auszugehen ist,

dass es auch bezüglich der Anzahl einzuschätzender Schüler keine systematischen Effekte auf die Urteilsgenauigkeit gibt.

Tabelle 66: Zusammenhang zwischen der Anzahl einzuschätzender Schüler und der diagnostischen Kompetenz der Lehrer (Rangkomponente)

	t1	*t2*	*t3*
Arithmetik	.02	-.05	.02
Wortschatz	-.16	.12	.08
Textverstehen	-	.08	.10
Rechtschreiben	-	-	-.21*
Fachinteresse Deutsch	-	.09	.12
Fachinteresse Mathematik	-	-.17	.02

*p < .05

Anteil Migranten

In dieser Untersuchung werden Kinder dann als Migranten angesehen, wenn mindestens ein Elternteil nicht in Deutschland geboren wurde. Der Anteil von Kindern mit derart definiertem Migrationshintergrund variiert zwischen den Klassen erheblich. Nur fünfzehn Prozent der Klassen bestehen ausschließlich aus Kindern ohne Migrationshintergrund, dahingegen bestehen zehn Prozent aller Klassen zu drei Vierteln oder mehr aus Migranten. Auch eine Klasse ohne einzigen deutschstämmigen Schüler ist in der Stichprobe vertreten. Im Mittel beträgt der Ausländeranteil pro Klasse 28,4 Prozent bei einer Standardabweichung von 27,4 (N = 140, alle Angaben mit Bezug zum dritten Messzeitpunkt). Die Zahl der Migranten in der Klasse hängt zu r = .41 signifikant mit der Anzahl der Schüler mit großen sprachlichen Einschränkungen zusammen. Zudem ist der Migrantenanteil signifikant negativ mit dem mittleren Leistungsniveau der Klasse in allen Bereichen korreliert, zwischen r = -.40 im logisch-abstrakten Denken und -.58 in der Rechtschreibleistung. In der folgenden Tabelle 67 wird der Anteil der Migranten in der Klasse mit der diagnostischen Kompetenz der Lehrer korreliert.

Tabelle 67: Zusammenhang zwischen dem Anteil der Kinder mit Migrationshintergrund in der Klasse und der diagnostischen Kompetenz der Lehrer (Rangkomponente)

	t1	*t2*	*t3*
Arithmetik	-.09	-.05	.02
Wortschatz	.10	.07	.27**
Textverstehen	-	.13	.09
Rechtschreiben	-	-	.11
Fachinteresse Deutsch	-	.02	-.04
Fachinteresse Mathematik	-	.05	.00

**p < .01

Die Ergebnisse zeigen in keine einheitliche Richtung. Es ergibt sich nur eine signifikante positive Korrelation für den Bereich Wortschatz zu t3 (r = 27). Auch die übrigen Zusammenhänge für die sprachbezogenen Leistungsbereiche weisen eher in Richtung eines positiven Zusammenhangs, erreichen jedoch nicht das Signifikanzniveau. Für Arithmetik und die Fachinteressen schwanken die Korrelationen um Null, so dass hier kein Zusammenhang zwischen der Rangurteilsgüte und dem Migrantenanteil in der Klasse besteht.

Leistungs- und Fachinteressenniveau der Klasse

Der folgende Abschnitt geht der Frage nach, ob oder inwiefern das Leistungs- und Merkmalsniveau der Klasse (genauer: der eingeschätzten Schüler) im Zusammenhang mit der diagnostischen Kompetenz der Lehrer steht. Wie bereits ab Tabelle 17 (S. 128) gezeigt wurde, gibt es hinsichtlich des Leistungs- und Merkmalsniveaus in den verschiedenen erfassten Bereichen mitunter hohe Variabilität zwischen verschiedenen Klassen, wie an den dargestellten Intraklassenkorrelationen abgelesen werden konnte. In den Leistungsbereichen waren die zu erwartenden Veränderungen im durchschnittlichen Niveau zwischen den Messzeitpunkten gut zu erkennen, was in den motivationalen Bereichen nicht der Fall ist. Die Streuung des mittleren Niveaus bleibt hingegen meist auch nach einem halben Jahr gleich.

Wie in Tabelle 68 abzulesen ist, weist das Leistungsniveau der Klasse in keinem der untersuchten Bereiche signifikante Zusammenhänge zur Urteilsgüte auf. Tendenziell sind in den meisten sprachbezogenen Leistungsbereichen negative Korrelationen zu erkennen, die sich allerdings auch wegen der

Veränderungen zwischen den Messzeitpunkten zufällig ergeben haben können.

Tabelle 68: Zusammenhang zwischen dem mittleren Niveau je Bereich in der Klasse und der diagnostischen Kompetenz der Lehrer (Rangkomponente)

	t1	*t2*	*t3*
Arithmetik	.06	.16	-.05
Wortschatz	-.11	-.13	-.13
Textverstehen	-	-.09	.02
Rechtschreiben	-	-	-.12
Fachinteresse Deutsch	-	-.03	-.15
Fachinteresse Mathematik	-	-.03	-.13

Leistungs- und Fachinteressenstreuung in der Klasse

Besonders plausibel erscheint die Annahme, dass eine große Leistungs- und Merkmalsheterogenität in der Klasse sich positiv auf die Rangkomponente diagnostischer Kompetenz auswirkt. Wie vermutet, zeigen sich für die Streuungen der Leistungen fast durchweg deutliche Korrelationen zur diagnostischen Kompetenz der Lehrer, die Rangkomponente betreffend. Lediglich für das Fachinteresse in Deutsch und Mathematik fallen die Zusammenhänge meist nicht signifikant aus. Die signifikanten Zusammenhänge in den meisten Bereichen und die dahinter stehende Heterogenität der Klasse sind allerdings weniger als Ursache dafür anzusehen, ob Lehrer gute oder schlechte Diagnostiker sind. Vielmehr ist davon auszugehen, dass es die Urteilsbildung der Lehrer maßgeblich erleichtert, wenn die Schüler sich voneinander deutlich unterscheiden, bzw. dass es sie erschwert, wenn die Schülermerkmale einander sehr ähnlich sind.

Tabelle 69: Zusammenhang zwischen der Streuung je Bereich in der Klasse und der diagnostischen Kompetenz der Lehrer (Rangkomponente)

	t1	*t2*	*t3*
Arithmetik	.20*	.32**	.25**
Wortschatz	.30**	.40**	.26**
Textverstehen	-	.32**	.40**
Rechtschreiben	-	-	.28**
Fachinteresse Deutsch	-	.05	.28**
Fachinteresse Mathematik	-	.15	.17

**p < .01, *p < .05

Klassenklima

In den meisten Klassen der Stichprobe herrscht ein sehr gutes Klassenklima, in dem sich die Kinder gut miteinander verstehen und zusammenarbeiten. Dies ist an den hohen Skalenmittelwerten abzulesen, die zu beiden Messzeitpunkten, zu denen das Klassenklima erfasst wurde, bei 3,1 auf der von eins bis vier reichenden Skala liegen (SD = 0,5). Ob das Ausmaß des guten Miteinanders einen Einfluss auf die Güte ihrer Urteile hat, ist bislang nicht bekannt. Da das Klassenklima von den Lehrern selbst eingeschätzt wurde, kann davon ausgegangen werden, dass sich in diesen Urteilen auch eine Art persönliche Betroffenheit ausdrückt, die möglicherweise im Zusammenhang mit der Güte ihrer Urteile steht.

Die sich ergebenden Korrelationen, die in Tabelle 70 dargestellt sind, widersprechen mit ihren teils zweistelligen negativen Korrelationen der Vermutung, dass sich ein gutes Klassenklima förderlich auf die Güte von Lehrerurteilen auswirkt. In keinem der Bereiche wird der Zusammenhang signifikant.

Tabelle 70: Zusammenhang zwischen dem Klassenklima und der diagnostischen Kompetenz der Lehrer (Rangkomponente)

	t1	*t2*
Arithmetik	.04	.01
Wortschatz	-.05	-.10
Textverstehen	-	-.01
Fachinteresse Deutsch	-	.07
Fachinteresse Mathematik	-	-.11

Unterrichtsstörung

Als weiteres Klassenmerkmal wird das Ausmaß der Unterrichtsstörungen auf seinen Zusammenhang mit der diagnostischen Urteilsgüte hin geprüft. In der Stichprobe ist die Unterrichtsstörung zu beiden Messzeitpunkten, zu denen sie erfasst wurde (t1 und t3), sehr gut normalverteilt und weist auf der vierstufigen Skala (1-4) Mittelwerte von 2,5 bzw. 2,4 (SD = 0,7) auf. Somit ist für die Analysen das volle Spektrum abgedeckt.

Wie in Tabelle 71 abzulesen ist, gibt es für keinen Bereich signifikante Zusammenhänge zwischen dem Ausmaß der Unterrichtsstörungen und der diagnostischen Kompetenz der Lehrer, außer für die Niveaueinschätzung der Arithmetik zu t1 (r = -.17). Auch wenn die meisten Korrelationen über r = -.10 und somit sehr nahe an Null liegen, so ist doch zumindest bemerkenswert, dass bis auf eine Ausnahme alle Werte negativ sind. Dies zeigt einen minimalen Trend dahingehend an, dass eine höhere Störungsbelastung mit niedrigerer Urteilsgüte einhergeht.

Tabelle 71: Zusammenhang zwischen dem Ausmaß der Unterrichtsstörungen und der diagnostischen Kompetenz der Lehrer (Korrelation)

	t1	*t3*
Arithmetik	-.13	-.02
Wortschatz	-.03	-.02
Textverstehen	-	-.05
Rechtschreiben	-	-.05
Fachinteresse Deutsch	-	-.06
Fachinteresse Mathematik	-	.05

Zeitverschwendung

Die Zeitverschwendung, die zu t1 und t3 mittels vier Items erhoben wurde, ist im Mittel etwas niedriger ausgeprägt als die Unterrichtsstörung und liegt auf der vierstufigen Skala (1-4) zu t1 bei 2,2, zu t3 bei 2,1 (SD jeweils 0,7). Die Befunde aus den Analysen ähneln aufgrund der inhaltlichen Nähe denen der Unterrichtsstörung sehr stark. Kein Zusammenhang wird signifikant, aber die Tendenz zeigt geringfügig in die erwartete negative Richtung.

Tabelle 72: Zusammenhang zwischen dem Ausmaß der Zeitverschwendung im Unterricht und der diagnostischen Kompetenz der Lehrer (Rangkomponente)

	t1	*t3*
Arithmetik	-.06	-.04
Wortschatz	-.02	.01
Textverstehen	-	-.03
Rechtschreiben	-	-.03
Fachinteresse Deutsch	-	-.08
Fachinteresse Mathematik	-	.02

7.2.3 Schülermerkmale

Bislang konnte gezeigt werden, dass weder Merkmale der Lehrkräfte selbst als auch Merkmale der gesamten Klassen maßgeblich mit der Güte der diagnostischen Lehrerurteile im Zusammenhang stehen. Allein die Streuung der Leistungen und Fachinteressen in den Klassen korreliert signifikant mit der Rangurteilsgüte in nahezu allen untersuchten Bereichen. Im letzten Abschnitt zu den Bedingungen der diagnostischen Kompetenz soll nun geprüft

werden, inwiefern die Urteilsgenauigkeit mit Merkmalen der individuellen Schüler in Verbindung steht. Während Lehrer- und Klassenmerkmale jeweils mit den über alle Schüler aggregierten Einschätzungen betrachtet wurden und somit schon aufgrund der Mittelung eine gewisse Ungenauigkeit nicht zu vermeiden ist, bietet sich die Ebene der einzelnen Schüler an, um Einflussfaktoren mit deutlich höherer Genauigkeit abzubilden. Voraussetzung dafür ist, dass die Urteile auch für jeden einzelnen Schüler vorliegen, und dieser große Vorteil ist bei der vorliegenden Stichprobe gegeben. Nachteilig wirkt sich an dieser Stelle jedoch aus, dass von den Lehrern lediglich die grobe globale, an fünf Stufen orientierte Einschätzung hinsichtlich der verschiedenen Merkmale erfragt wurde. Dies ist für die Bildung der Rangkomponente diagnostischer Urteile, so wie sie in dieser Arbeit bislang auch eingesetzt wurde, ein adäquates und sehr gut geeignetes Vorgehen. Urteile zum Leistungs- bzw. Fachinteressenniveau wären eine sinnvolle Ergänzung zu den globalen Rangurteilen, diese liegen für die Hauptstichprobe aber nicht vor. Dies wären Fragen wie zum Beispiel danach, wie viele Aufgaben eines Tests ein Schüler wahrscheinlich richtig gelöst hat. Hierbei kann die Einschätzung direkt mit dem tatsächlichen Lösungsverhalten der Schüler verglichen werden. Die Rangkomponente ergibt sich hingegen erst aus einem sozialen Vergleich mit Klassenkameraden oder z.B. einer Jahrgangsstufe, so dass für den einzelnen Schüler kein direktes Maß der Übereinstimmung zur entsprechenden Lehrereinschätzung berechnet werden kann. In diesem Fall ist es, sollen dennoch Aussagen zur Urteilsgenauigkeit auf individueller Schülerebene getroffen werden, notwendig, die erfassten globalen Urteile nicht in die Rangkomponente, sondern als individuenbezogenen Indikator in die Niveaukomponente zu überführen. Dafür gibt es jedoch keine festgelegte Vorgehensweise, sondern es obliegt dem Forscher, für die gegebenen Daten die bestmögliche Transformationsstrategie zu wählen.

Für die vorliegenden Daten wurde nach gründlicher Abwägung ein Verfahren gewählt, das wie im Kapitel 6.3 ab Seite 149 beschrieben das Leistungsspektrum der Schüler - gemessen an der Gesamtstichprobe - in fünf gleich große Abschnitte unterteilt und jeden Schüler dem von ihm erreichten Leistungsabschnitt zuweist. Somit konnten die Schülerleistungen auch auf fünf Stufen gebracht und eine direkte Differenz zur Lehrereinschätzung berechnet werden. Die Werte in den folgenden Tabellen sind so zu interpretieren, dass ein höherer Wert im Vergleich zu einem niedrigeren eine positivere Leistungseinschätzung bedeutet, wobei ‚positiver' heißt, dass das Urteil im Vergleich zur tatsächlichen Leistung des Schülers entweder weniger stark

unterschätzt oder stärker überschätzt wird. Könnte man das absolute Niveau interpretieren (was aufgrund der Transformation der Schülerleistungen nicht möglich ist), so wäre ein Wert von Null als - im Mittel - perfekte Übereinstimmung zwischen Leistungsniveau und Urteilsniveau zu verstehen, negative Werte bedeuteten eine Leistungsunterschätzung, wohingegen positive Werte als Leistungsüberschätzung zu verstehen wären. Der Wertebereich läge zwischen minus vier und plus vier, wobei die Maxima die größtmögliche Fehleinschätzung anzeigen würden. Da die nachfolgenden Analysen darauf abzielen, Unterschiede zwischen Gruppen zu prüfen und nicht das absolute Niveau der Einschätzungen bestimmt werden soll, ist die Datenbasis dennoch geeignet.

Geschlecht

Als erstes Schülermerkmal, für das ein Einfluss auf die Beurteilung durch die Lehrer vermutet wird, wird das Geschlecht der Schüler analysiert. Dafür erfolgt zunächst eine Gegenüberstellung von geschlechtsspezifischen Testleistungen der Schüler und geschlechtsspezifischen Urteilen der Lehrer. Sowohl über die Leistungen als auch über die Urteile wurden t-Tests mit dem Schülergeschlecht als unabhängiger Gruppenvariable gerechnet, deren Ergebnisse in Tabelle 73 dargestellt sind. Dabei fallen besonders zwei Aspekte auf: Zum einen drückt sich in den Lehrerurteilen aus, dass für alle Leistungsbereiche und über alle Messzeitpunkte hinweg signifikante Leistungsunterschiede zwischen den Geschlechtern vermutet werden, was sich in den Testleistungen jedoch nur bedingt widerspiegelt. Im Bereich Wortschatz treten nur zu t1 signifikante Unterschiede zwischen Jungen und Mädchen zu Tage, zu t2 und t3 gar keine. Und auch im Bereich Arithmetik zu t2 sind beide Geschlechter leistungsmäßig gleichauf. Zum anderen sticht aber auch ins Auge, dass die Lehrer mitunter die Leistungsverteilung zwischen den Geschlechtern genau entgegen der Testergebnisse einschätzen, nämlich für den Bereich Wortschatz zu t1 und t2 sowie für Arithmetik zu t3. Während über die Messzeitpunkte hinweg in den verschiedenen Leistungsbereichen nicht immer dasselbe Geschlecht die besseren Leistungen erbringt, sondern eben zu t3 die Mädchen bessere Arithmetikleistungen zeigen als Jungen und Jungen im Gegenzug zu t1 und t2 den besseren Wortschatz besitzen, widerspricht dies offenbar dem Eindruck der Lehrer, die konsistent die Jungen für bessere Mathematiker und die Mädchen für besser im Bereich Deutsch halten.

Tabelle 73: Leistungsunterschiede in den eingesetzten Testverfahren sowie mittlere Leistungseinschätzung durch die Lehrer, jeweils getrennt nach Geschlecht der Schüler

			Schülerleistungen				*Lehrerurteile*		
			N[1]	M	SD	t	M	SD	t
t1	Arithmetik	m	1099	5,3	2,9	6,02**	3,5	1,1	7,75**
		w	1001	4,7	2,6		3,1	1,1	
	Wortschatz	m	1182	14,6	5,0	2,03*	3,1	1,1	-7,75**
		w	1082	14,2	4,9		3,4	1,1	
t2	Arithmetik	m	1016	4,5	2,7	0,67	3,9	1,0	8,40**
		w	943	4,4	2,4		3,6	1,0	
	Wortschatz	m	1069	17,0	4,8	1,58	3,5	1,2	-3,70**
		w	1002	16,7	4,8		3,7	1,1	
	Textverstehen	m	1070	14,3	4,1	-4,98**	3,6	1,1	-5,14**
		w	997	15,2	4,0		3,8	1,0	
t3	Arithmetik	m	956	9,2	3,7	-2,65**	3,9	1,0	6,55**
		w	882	9,6	3,5		3,6	1,0	
	Wortschatz	m	1027	18,6	4,9	-0,57	3,5	1,2	-5,09**
		w	959	18,7	4,6		3,7	1,1	
	Textverstehen	m	1027	16,7	4,8	-7,49**	3,7	1,1	-4,81**
		w	960	18,1	4,4		3,9	1,0	
	Rechtschreiben	m	1024	14,9	4,2	-10,15**	2,9	1,3	-8,04**
		w	956	16,6	3,6		3,4	1,3	

[1] Die angegebenen Häufigkeiten beziehen sich auf die vorliegenden Lehrerurteile. Auf Seiten der Schüler handelt es sich um einen durch die Imputation vervollständigten Datensatz mit 1250 Jungen und 1145 Mädchen.

**p < .01, *p < .05

Während die vorangegangene Analyse reine Niveauunterschiede zwischen geschlechtsbezogenen Leistungen und Urteilen zum Gegenstand hatte, interessiert im Folgenden, inwiefern die Lehrkräfte in ihren Urteilen unabhängig vom wirklichen Leistungsniveau der Schüler zu einer geschlechtsspezifischen Über- oder Unterschätzung der Leistungen neigen. Die Antwort darauf kann in gewisser Weise bereits aus den eben beschriebenen Diskrepanzen geschlussfolgert werden, denn wenn Mädchen tatsächlich besser in einem Test abschneiden als Jungen und dies zur Einschätzung durch die Leh-

rer im Widerspruch steht, ergibt sich daraus zwangsläufig eine Unterschätzung der Mädchen.

Tabelle 74 zeigt die Ergebnisse der Berechnungen, bei denen die mittleren geschlechtsspezifischen Niveaudifferenzen zwischen Lehrerurteilen und Schülerleistungen per t-Test auf Signifikanz geprüft wurden. Die absolute Höhe der angegebenen mittleren Differenzen kann dabei nicht als tatsächliche Über- bzw. Unterschätzung interpretiert werden. Bewusst wurde in den Analysen aus demselben Grund auf eine gleiche Skalierung von Schülerleistungen und Lehrerurteilen verzichtet, da eine Vergleichbarkeit durch die Transformation der Niveau- aus der Rangkomponente ohnehin nicht gegeben wäre. Interpretierbar sind jedoch die Unterschiede zwischen den Geschlechtern, die überall signifikant ausfallen, nur nicht in jenen Bereichen, in denen sich bereits die Lehrerurteile als korrekte Einschätzungen der Geschlechtsunterschiede erwiesen haben (s. Tabelle 73). Zusätzliche Information kann aus dem Vergleich der Differenzwerte in Tabelle 74 insofern gezogen werden, als dass nun auch ersichtlich ist, in welchen Leistungsbereichen die geschlechtsspezifischen Abweichungen der Niveauurteile von den Testwerten besonders groß ausfallen, nämlich - mit leichten Unterschieden je nach Messzeitpunkt - im Bereich Arithmetik sowie im Bereich Wortschatz.

In den geschlechtsspezifischen Abweichungen der Leistungseinschätzungen von den gemessenen Leistungen drückt sich aus, dass die Lehrkräfte - wahrscheinlich unbewusst - Geschlechtsunterschiede in jenen Leistungsbereichen vermuten, in denen de facto keine sind. Tatsächlich vorhandene Vorteile der Mädchen im Textverstehen sowie im Bereich Rechtschreiben drücken sich hingegen nicht in zusätzlichen Über- bzw. Unterschätzungen in den Lehrerurteilen aus, da die Leistungsunterschiede hier im Mittel korrekt von den Lehrern eingeschätzt wurden (vgl. auch ähnliche Befunde für die Zusatzstichprobe in Tabelle 82, S. 243).

Tabelle 74: Unterschiede in der Genauigkeit der Niveauurteile der Lehrer in Abhängigkeit vom Geschlecht der Schüler

		männlich	*weiblich*	*t*
t1	Arithmetik	0,75	0,66	1,69
	Wortschatz	0,06	0,47	-9,41**
t2	Arithmetik	1,66	1,35	6,61**
	Wortschatz	0,23	0,48	-5,09**
	Textverstehen	-0,41	-0,34	-1,44
t3	Arithmetik	0,74	0,37	8,08**
	Wortschatz	-0,04	0,20	-5,06**
	Textverstehen	0,06	0,03	0,65
	Rechtschreiben	-0,74	-0,68	-1,27

**p < .01, *p < .05

Im Folgenden werden die Geschlechtseffekte noch differenzierter betrachtet, indem zusätzlich das Geschlecht der Lehrer berücksichtigt wird. Dafür wurden für jeden Leistungsbereich univariate Varianzanalysen mit Schüler- und Lehrergeschlecht als festen Faktoren gerechnet. Zunächst sind in Tabelle 75 die Niveauurteilsgenauigkeiten, getrennt nach Schüler- und Lehrergeschlecht sowie nur getrennt nach Lehrergeschlecht (Spalte „Schüler gesamt"), angegeben. Dabei fällt auf, dass die in Tabelle 74 berichteten Unterschiede in der Urteilsgenauigkeit zwischen Schülerinnen und Schülern sich überwiegend auch dann zeigen, wenn man sie nach dem Lehrergeschlecht getrennt ausweist. Unter Berücksichtigung der Tatsache, dass in den Analysen aufgrund der Stichprobenverteilung wenige männliche Lehrkräfte vielen weiblichen gegenübergestellt sind, zeigt sich dennoch, dass auf Seiten der Lehrerinnen offenbar beinahe überall größere Leistungsunterschiede zwischen Jungen und Mädchen angenommen werden als auf Seiten ihrer männlichen Kollegen. Der insgesamt signifikante Geschlechtereffekt für Arithmetik zu t2 ist sogar ausschließlich auf die geschlechtsspezifischen Beurteilungsunterschiede der Lehrerinnen zurückzuführen, da die Urteile männlicher Lehrer sich nicht als signifikant geschlechtsspezifisch erweisen.

Die drei letzten Spalten der Tabelle geben nochmals als F-Wert an, ob Unterschiede in der Genauigkeit der Niveauurteile in Abhängigkeit vom Schülergeschlecht vorliegen, darüber hinaus wird aber auch die Abhängigkeit vom Lehrergeschlecht sowie die Wechselwirkung aus Schüler- und Lehrergeschlecht ausgewiesen. Für das Schülergeschlecht decken sich die Ergeb-

nisse naturgemäß mit jenen der vorangegangenen t-Tests, allerdings zeigen sich durchaus Unterschiede je nach Lehrergeschlecht. Berücksichtigt werden muss, dass die Anzahl männlicher Lehrer deutlich unter der Anzahl der Lehrerinnen liegt. Trotzdem ist zunächst auffällig, dass sich - mit Ausnahme von Wortschatz zu t1 und Arithmetik zu t2 - bei Lehrerinnen in allen Bereichen Annahmen über Leistungsunterschiede zwischen Jungen und Mädchen deutlicher in ihren Urteilen widerspiegeln als bei ihren männlichen Kollegen, was an den $F_{Lehrergeschlecht}$-Werten in Tabelle 75 abgelesen werden kann.

Was sich nicht zeigt, sind Bevorzugungen von Schülern des eigenen Geschlechts bei der Beurteilung ($F_{Wechselwirkung}$). Allerdings ist zu erkennen, dass Frauen in fast allen Leistungsbereichen grundsätzlich für Jungen und für Mädchen leicht positivere Differenzen ihrer Urteile zu den Schülerleistungen aufweisen, ganz so, als urteilten sie etwas wohlwollender oder als würden sie eher zur Leistungsüberschätzung neigen. In der mittleren Spalte von Tabelle 75 (Schüler gesamt) sind unabhängig vom Geschlecht der Schüler die mittleren Differenzwerte für Lehrerinnen und Lehrer getrennt aufgeführt. Dabei zeigen sich für die meisten Bereiche signifikante Unterschiede, die darauf hindeuten, dass Lehrerinnen unter Kontrolle der tatsächlichen Leistungen insgesamt etwas mehr überschätzen als männliche Lehrer (t-Werte s. dritte Spalte).

Tabelle 75: Unterschiede in der Genauigkeit der Niveauurteile der Lehrer in Abhängigkeit vom Geschlecht Lehrer und der Schüler (t-Tests und univariate Varianzanalyse)

		Geschlecht Lehrer	Schüler männlich	Schüler weiblich	$t_{getrennt}$	Schüler gesamt	$F_{Schülergeschlecht}$	$F_{Lehrergeschlecht}$	$F_{Wechselwirkung (Lehrergeschlecht * Schülergeschlecht)}$
t1	Arithmetik	m	0,54	0,45	1,04	0,50	1,63	12,41**	0,00
		w	0,79	0,70	1,95	0,75			
	Wortschatz	m	0,07	0,46	-3,78**	0,26	46,53**	0,00	0,08
		w	0,05	0,48	-8,62**	0,25			
t2	Arithmetik	m	1,56	1,44	1,12	1,50	15,75**	0,01	3,52
		w	1,68	1,31	7,27**	1,50			
	Wortschatz	m	0,04	0,30	-2,30*	0,17	15,98**	13,01**	0,02
		w	0,28	0,52	-4,56**	0,39			
	Textverstehen	m	-0,48	-0,49	0,07	-0,49	0,39	5,78*	0,56
		w	-0,39	-0,31	-1,61	-0,35			
t3	Arithmetik	m	0,62	0,28	3,21**	0,46	37,56**	4,21*	0,10
		w	0,79	0,39	7,80**	0,60			
	Wortschatz	m	-0,21	0,10	-2,92**	-0,06	18,70**	6,92**	0,46
		w	-0,00	0,23	-4,26**	0,11			
	Textverstehen	m	-0,14	-0,24	0,98	-0,18	0,99	23,71**	0,61
		w	0,10	0,09	0,25	0,09			
	Rechtschreiben[1]	m	-0,94	-0,94	0,02	-0,94	0,35	20,37**	0,39
		w	-0,70	-0,62	-1,41	-0,66			

Anm.: [1] Für den Bereich Rechtschreiben deutet der Levene-Test eine nicht gegebene Fehlervarianzhomogenität an.
**p < .01, *p < .05

Leistungsniveau

Als nächstes Merkmal wird das individuelle Leistungsniveau der Schüler auf seinen Zusammenhang zur Leistungseinschätzung hin betrachtet. Angenommen wird, dass die ohnehin guten Schüler für noch besser gehalten werden als die leistungsschwächeren Schüler. Wie in Tabelle 76 dargestellt, verhält es sich aber genau andersherum. In allen Bereichen werden die Schüler, die der unteren Hälfte der jeweiligen bereichsspezifischen Leis-

tungsverteilung zuzuordnen sind, unter Berücksichtigung des tatsächlichen Leistungsniveaus besser als die Schüler in der oberen Verteilungshälfte eingeschätzt. Die Ursache hierfür ist vermutlich technischer Natur. Gute Schüler, die auf der fünfstufigen Skala ohnehin schon mit hohen Werten (oftmals sicher mit dem Maximalwert fünf) eingeschätzt werden, können darüber hinaus nicht noch stärker überschätzt werden, wohingegen schlechte Schüler wenig Spielraum für eine weitergehende Unterschätzung haben. Verschätzungen gehen daher mit hoher Wahrscheinlichkeit gerade in die Richtung des entgegengesetzten Endes des Leistungsspektrums.

Tabelle 76: Unterschiede in der Genauigkeit der Niveauurteile der Lehrer in Abhängigkeit vom Leistungsniveau des Schülers, getrennt nach oberer und unterer Hälfte der Verteilung (Medianspilt)

		untere Hälfte der Leistungsverteilung	*obere Hälfte der Leistungsverteilung*	*t*
t1	Arithmetik	1,68	1,05	9,95**
	Wortschatz	0,66	0,19	7,52**
t2	Arithmetik	2,14	1,56	9,39**
	Wortschatz	0,88	0,41	6,95**
	Textverstehen	0,22	-0,32	8,26**
t3	Arithmetik	1,22	0,86	5,36**
	Wortschatz	0,46	0,17	3,92**
	Textverstehen	0,42	0,25	2,37*
	Rechtschreiben	-0,40	-0,83	5,76**

**p < .01, *p < .05

Exemplarisch ist dieser Zusammenhang in Abbildung 9 für den Bereich Arithmetik zu t2 dargestellt. Zwischen den Arithmetikleistungen der Schüler und den Niveauurteilsabweichungen durch die Lehrer besteht nahezu ein negativer linearer Zusammenhang. Je höher die Schülerleistung ist, desto größer wird auch die Wahrscheinlichkeit für eine Unterschätzung durch den Lehrer. Regressionsanalytisch zeigt sich, dass ein Fünftel der Varianz der Niveauurteilsabweichungen durch die Arithmetikleistung der Schüler erklärt werden kann ($R^2 = .202$).

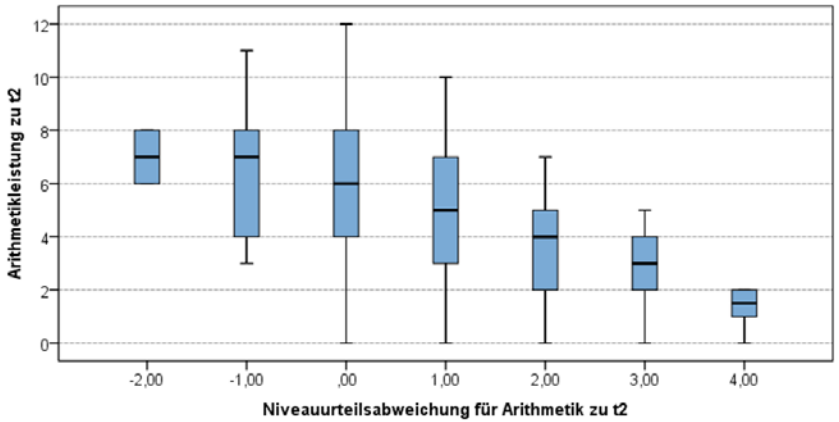

Abbildung 9: Zusammenhang zwischen Schülerleistung und Niveauurteilsabweichung, exemplarisch für den Bereich Arithmetik zu t2

Sozialstatus

Neben dem Geschlecht der Schüler, für das immer wieder ein Einfluss auf die Leistungsbeurteilung durch Lehrer beschrieben wird, steht vor allem auch die soziale Herkunft häufig im Fokus von Untersuchungen. Auch in der vorliegenden Untersuchung wurde geprüft, inwiefern sich der Sozialstatus auf die (durch Transformation gebildeten) Niveauurteile der Lehrer auswirkt. Als unabhängige Variable wurde der höchste ISEI-Wert im Haushalt (HISEI) verwendet und per Mediansplit geteilt, so dass die Urteilsgenauigkeit für Schüler der oberen Hälfte der Sozialverteilung mit der Urteilsgenauigkeit für Schüler der unteren Hälfte der Sozialverteilung verglichen werden konnte.

Wie in Tabelle 77 abzulesen ist, weist lediglich der Bereich Arithmetik zu t3 eine nicht signifikante Differenz in der Urteilsgenauigkeit auf, deutet aber dennoch in die Richtung, die sonst durchgehend zu finden ist: Schüler, die der oberen Hälfte der Sozialverteilung angehören, werden - wieder unter Berücksichtigung der tatsächlichen Leistung - in ihren Leistungen stärker überschätzt (bzw. weniger stark unterschätzt) als Schüler, die nach dem Sozialstatus der Eltern der unteren Hälfte der Verteilung zuzuordnen sind.

Tabelle 77: Unterschiede in der Genauigkeit der Niveauurteile der Lehrer in Abhängigkeit vom sozioökonomischen Status der Eltern (HISEI), getrennt nach oberer und unterer Hälfte der Verteilung (Mediansplit)

		untere Hälfte der HISEI-Verteilung	*obere Hälfte der HISEI-Verteilung*	*t*
t1	Arithmetik	0,60	0,81	-4,06**
	Wortschatz	0,07	0,44	-8,35**
t2	Arithmetik	1,40	1,60	-4,35**
	Wortschatz	0,15	0,54	-8,11**
	Textverstehen	-0,54	-0,22	-7,25**
t3	Arithmetik	0,54	0,60	-1,26
	Wortschatz	-0,11	0,26	-7,71**
	Textverstehen	-0,12	0,19	-7,22**
	Rechtschreiben	-0,85	-0,58	-6,37**

**p < .01

Kein grundsätzlich anderes Bild erhält man, wenn man die individuellen Leistungseinschätzungen durch die Lehrer mit dem höchsten sozioökonomischen Status der Eltern korreliert (vgl. Tabelle 78). Auch hierbei ist der Zusammenhang für jeden Bereich und jeden Messzeitpunkt - außer für Arithmetik zu t3 - signifikant, wenn auch auf insgesamt relativ niedrigem Niveau. Wie im vorherigen Abschnitt gezeigt, bestehen überwiegend signifikante Zusammenhänge zwischen dem Leistungsniveau der Schüler und den Lehrerurteilen. Da die Schülerleistungen bekanntermaßen mit dem Sozialstatus korrelieren, sozial schwache Schüler in der Regel also auch niedrigere Leistungen erbringen als Schüler mit hohem Sozialstatus, liegt die Frage auf der Hand, wie die Zusammenhänge ausfallen, wenn für die Leistung kontrolliert wird.

Wie in Tabelle 78 zu sehen ist, werden die Zusammenhänge tatsächlich in allen Bereichen höher, wenn man die Schülerleistung auspartialisiert. Dabei wird auch der Zusammenhang für Arithmetik zu t3 signifikant, der sich bislang sowohl bei den t-Tests als auch bei der einfachen Korrelation als sehr schwach erwiesen hatte. In der letzten Spalte wird belegt, dass die partiellen Korrelationen - bis auf die Ausnahme Arithmetik zu t2 - durchgängig signifikant höher ausfallen als die bivariaten.

Tabelle 78: Zusammenhang zwischen dem sozioökonomischen Status der Eltern (HISEI) und der Güte der individuellen Leistungseinschätzungen (Niveaukomponente)

		r	*Partialkorrelation unter Kontrolle der jeweiligen Schülerleistung*	*Signifikanzunterschied zwischen r und Partialkorrelation*
t1	Arithmetik	.08**	.22**	4,65*
	Wortschatz	.21**	.28**	2,58*
t2	Arithmetik	.19**	.24**	1,64
	Wortschatz	.20**	.32**	4,46*
	Textverstehen	.16**	.27**	3,99*
t3	Arithmetik	.03	.18**	4,60*
	Wortschatz	.20**	.30**	3,69*
	Textverstehen	.14**	.27**	4,70*
	Rechtschreiben	.11**	.18**	2,47*

**p < .01, *p < .05

Gefühl des Angenommenseins

Kommen in den Lehrerurteilen auch persönliche Sympathien oder Antipathien zum Ausdruck? Dieser Frage kann sich mit den vorliegenden Daten nur auf einem Umweg genähert werden, indem angenommen wird, dass jene Schüler, die sich von ihren Lehrern nicht gut angenommen fühlen, von ihnen auch möglicherweise anders behandelt werden als andere Schüler. Der Anteil der Schüler, die sich von ihrem Lehrer nicht oder eher nicht angenommen fühlen, liegt - gemessen am Skalenmittelwert - zwischen 5,9 Prozent (t1) und 13,0 Prozent (t3) und ist damit sehr gering, was die Aussagekraft des eingesetzten t-Tests zur Prüfung der Mittelwertunterschiede zwischen den Gruppen reduziert. Die gefunden Ergebnisse sind relativ uneinheitlich, es lässt sich keine einheitliche Tendenz ablesen (s. Tabelle 79). Während es im Bereich Wortschatz (über die tatsächlichen Leistungen hinaus) zu allen drei Messzeitpunkten zu einer Schlechterbewertung der sich nicht oder eher nicht angenommen fühlenden Schüler kommt, werden im Bereich Arithmetik die sich nicht angenommen fühlenden Schüler mal signifikant besser (t3), mal signifikant schlechter (t1) und mal genauso gut (t2) eingeschätzt wie die sich angenommen fühlenden Schüler. Auch für das Textverstehen sind die Ergebnisse uneinheitlich, während sich für die Rechtschreibung keine Effekte zeigen.

Wegen der ungleichmäßigen Verteilung der Schülergruppen ist der Zusammenhang zwischen dem Ausmaß des Gefühls des Angenommenseins und der Niveaueinschätzung durch die Lehrer zusätzlich als Korrelation berechnet worden (s. letzte Spalte in Tabelle 79). Im Vergleich zum t-Test ändert sich jedoch kaum etwas an den Befunden, lediglich das Signifikanzniveau ist in zwei der Bereiche leicht unterschiedlich. Die signifikanten Korrelationen bewegen sich insgesamt im niedrigen Bereich, so dass sich der vermutete Zusammenhang zwischen beiden Variablen nur bedingt zeigt.

Tabelle 79: Unterschiede in der Genauigkeit der Niveauurteile der Lehrer in Abhängigkeit vom Gefühl des Angenommenseins des Schülers, getrennt nach oberer und unterer Hälfte der Verteilung (Mediansplit) sowie als Korrelation zwischen Skalenausprägung des Gefühls des Angenommenseins auf Schülerseite und dem Niveauurteil der Lehrer

		Schüler fühlt sich (eher) nicht angenommen	*Schüler fühlt sich (eher) angenommen*	*t*	*Korrelation zwischen Gefühl des Angenommenseins und Lehrereinschätzung*
t1	Arithmetik	0,50	0,72	-2,06*	.07**
	Wortschatz	-0,27	0,29	-6,00**	.20**
t2	Arithmetik	1,45	1,51	-0,71	.03
	Wortschatz	0,06	0,39	-4.17**	.14**
	Textverstehen	-0,55	-0,35	-2,74**	.05*
t3	Arithmetik	0,78	0,54	3,33**	-.07**
	Wortschatz	-0,18	0,12	-4,00**	.12**
	Textverstehen	0,07	0,04	0,40	.00
	Rechtschreiben	-0,76	-0,70	-0,77	.00

**p < .01, *p < .05

7.3 Vergleichbarkeit von Leistungseinschätzungen und Zeugnisnoten

Wie bereits in Kapitel 7.1 (ab S. 165) dargestellt, hängen bereichsspezifische Lehrerurteile mit den vergebenen Zeugnisnoten in den entsprechenden Unterrichtsfächern signifikant zusammen. Obwohl sich in Zeugnisnoten in der Regel mehr widerspiegelt als die reine Leistung im jeweiligen Fach und mehrfach zu Recht darauf hingewiesen wurde, dass Zensuren und Tests nur teilweise übereinstimmende Sachverhalte messen (Roeder et al., 1986; Schrader & Helmke, 1987), sind Noten dennoch ein guter Indikator für die

Kompetenzen der Schüler. Insofern ist der Gedanke naheliegend, dass die Lehrer sich bei ihrer Einschätzung im Einschätzbogen an den von ihnen selbst erteilten Noten orientieren. Dies trifft umso mehr zu, da die Lehrer im selben Instrument sogar gebeten werden, für jeden Schüler auch die letzten Zeugnisnoten zu notieren. Es ist ihnen demzufolge bewusst, wie sie selbst die Leistungen auf dem letzten Zeugnis bewertet haben, und die Anforderung besteht nun darin, von den Fachnoten in Deutsch und Mathematik auf die im Einschätzbogen erfragten Konstrukte wie Wortschatz, Textverstehen, Arithmetik, aber zum Beispiel auch auf das logisch-abstrakte Denken zu abstrahieren. Schrader und Helmke (1990) fanden für den Bereich Mathematik, dass die Korrelation zwischen Zensur und Testleistung mit r = .70 praktisch genauso hoch ausfiel wie zwischen direkter Leistungseinschätzung und Testleistung (r = .67).

In der folgenden Abbildung 10 ist exemplarisch für den Bereich Arithmetik zu t3 der Zusammenhang zwischen der Arithmetikeinschätzung und der letzten Mathematik-Zeugnisnote als Streudiagramm inklusive Regressionsgeraden dargestellt. Deutlich zu erkennen ist der negative Zusammenhang. Je besser die Zeugnisnoten der Schüler sind, desto besser wird auch ihre Arithmetikkompetenz eingeschätzt. Die Zeugnisnote klärt dabei 59 Prozent der Varianz der Leistungseinschätzungen auf (R^2 = .591).

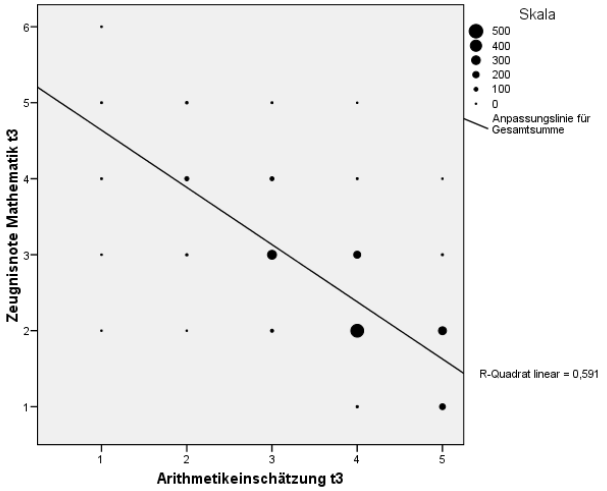

Abbildung 10: Zusammenhang zwischen der Einschätzung der Schülerleistungen im Bereich Arithmetik zu t3 und der zuletzt vergebenen Zeugnisnote in Mathematik

Was sich beispielhaft schon für den Bereich Arithmetik anhand von Abbildung 10 erkennen lässt, beschreibt Tabelle 80 für alle eingeschätzten und getesteten Leistungsbereiche zu den drei Messzeitpunkten, die auch grob einer Zeugnisnote zuzuordnen sind. Die Korrelationen zwischen den zuletzt vergebenen Zeugnisnoten und den Leistungseinschätzungen der Lehrer in den entsprechenden Leistungsbereichen liegt durchweg im hohen Bereich zwischen r = .67 und .77. Damit kann gezeigt werden, dass die beiden verschiedenen Formen des Lehrerurteils einander sehr ähnlich sind und es nur geringe bis mäßige Unterschiede zwischen den Fächern Deutsch und Mathematik und den jeweils damit verbundenen Leistungsbereichen gibt. Bezieht man die im Test gemessenen Schülerleistungen aus den entsprechenden Bereichen ein und stellt sie sowohl den vergebenen Zeugnisnoten als auch den Lehrereinschätzungen gegenüber, so fallen diese Korrelationen immer noch signifikant, aber dennoch deutlich niedriger aus. Beide Lehrerurteile, Zeugnisnoten und abgefragte Einschätzungen, sind demzufolge annähernd gleich nah (oder fern) an den Testleistungen der Schüler, und die Unterschiede je nach Bereich und Messzeitpunkt sind minimal.

Tabelle 80: Zusammenhänge zwischen Zeugnisnoten, Lehrerurteilen und Schülerleistungen

t	*Bereich*	*Korrelation zwischen Noten und Einschätzungen*	*Korrelation zwischen Noten und Leistung*	*Korrelation zwischen Einschätzung und Leistung*
t1	Arithmetik	-.73**	-.43**	.43**
	Wortschatz	-.70**	-.48**	.49**
t2	Arithmetik	-.71**	-.53**	.48**
	Wortschatz	-67**	-.52**	.51**
	Textverstehen	-.70**	-.55**	.55**
t3	Arithmetik	-.77**	-.59**	.55**
	Wortschatz	-.69**	-.55**	.50**
	Textverstehen	-.72**	-.59**	.59**
	Rechtschreibung	-.73**	-.63**	.62**

**p < .01

Anm.: Die Korrelationen im Bereich Arithmetik wurden nur für jene Schüler berechnet, deren Lehrer auch Mathematik unterrichten. Alle Korrelationen sind Overall-Korrelationen ohne Berücksichtigung der Klassenebene.

Im Folgenden wird der Zusammenhang zwischen Lehrerurteilen und Zeugnisnoten detaillierter betrachtet. Wie in Abbildung 11 wiederum exemplarisch, diesmal für den Bereich Wortschatz zu t3, dargestellt, ist die

Zuordnung der Zeugnisnoten zu den Lehrereinschätzungen nicht exakt deckungsgleich. Stattdessen ist es so, dass die Noten 2, 3 und 4 sogar über alle fünf Einschätzstufen streuen, und selbst einem Schüler, der im Zeugnis die Note 5 im Fach Deutsch hatte, wurde für den Bereich Wortschatz bescheinigt, dass es eher zuträfe, dass er/sie über einen umfangreichen Wortschatz verfüge. Die Deutschnote 6 kommt nur zweimal vor, und die betreffenden Schüler bekamen auch die niedrigste Einschätzung für den Wortschatz. Am anderen Ende der Notenskala verteilen sich die Einser-Schüler aber über die ersten drei Einschätzstufen. Absolute Deckungsgleichheit ist hierbei zwar auch gar nicht zu erwarten, weil schulfachbezogene Zeugnisnoten eben viel mehr beinhalten als bereichsspezifische Leistungsurteile, auffällig große Abweichungen sind jedoch auch eher unplausibel.

Dieses Bild sieht für die anderen Leistungsbereiche und zu anderen Messzeitpunkten sehr ähnlich aus und ist der entscheidende Hinweis darauf, dass Noten und Urteile letztendlich trotz hoher Korrelationen und im Mittel großer Ähnlichkeit dennoch verschiedene Urteilsformen sind, die auch verschiedene Dinge ausdrücken und sich deshalb nicht gegenseitig substituieren.

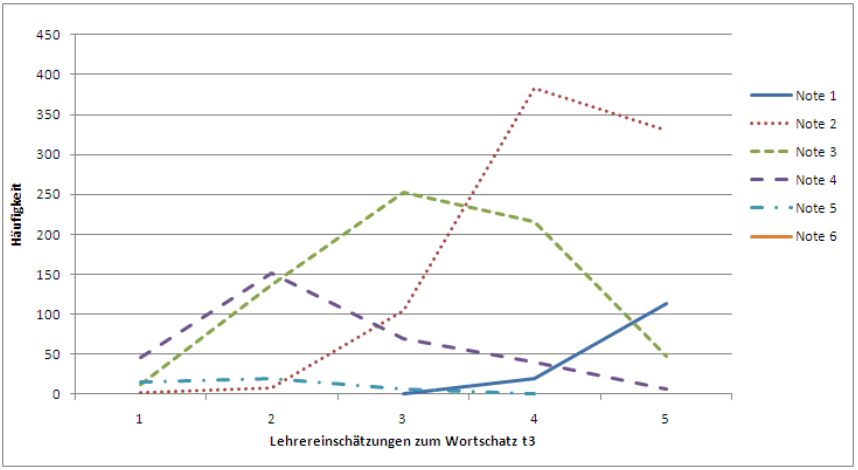

Abbildung 11: Zuordnung der Zeugnisnoten im Fach Deutsch zu den fünfstufigen Lehrereinschätzungen für den Wortschatz zu t3

Als nächstes wird der Blick auf den Zusammenhang zwischen den Wortschatzeinschätzungen und den Leistungen im Wortschatztest zu t3 gelenkt. Dabei soll betrachtet werden, in welchem Ausmaß vergleichbare Schülerleistungen unterschiedliche Leistungseinschätzungen von ihren Lehrern erhal-

ten. Fast die Hälfte aller Schüler (N = 1157) wurden bei der Frage nach der Ausprägung des Wortschatzes der höchsten und der zweithöchsten Leistungsstufe zugeordnet, was die Verteilung der Einschätzungen insgesamt rechtssteil ausfallen lässt (Schiefe = -0,45, vgl. Tabelle 25, S. 141). Abbildung 12 macht die riesigen Überschneidungen deutlich, die es zwischen den Einschätzstufen gibt. Die über alle Lehrer und Schüler und ohne Berücksichtigung der Klassenzugehörigkeit berechneten Werte zeigen, dass Schüler mit Leistungswerten zwischen 9 und 23 Punkten bei den Einschätzungen jeder der fünf Leistungsstufen zugeordnet wurden. Auch wenn im Mittel höhere Leistungen auch mit besseren Beurteilungen einhergehen, lässt sich gerade im mittleren Leistungsbereich das Lehrerurteil nur schwer aus der Leistung ableiten.

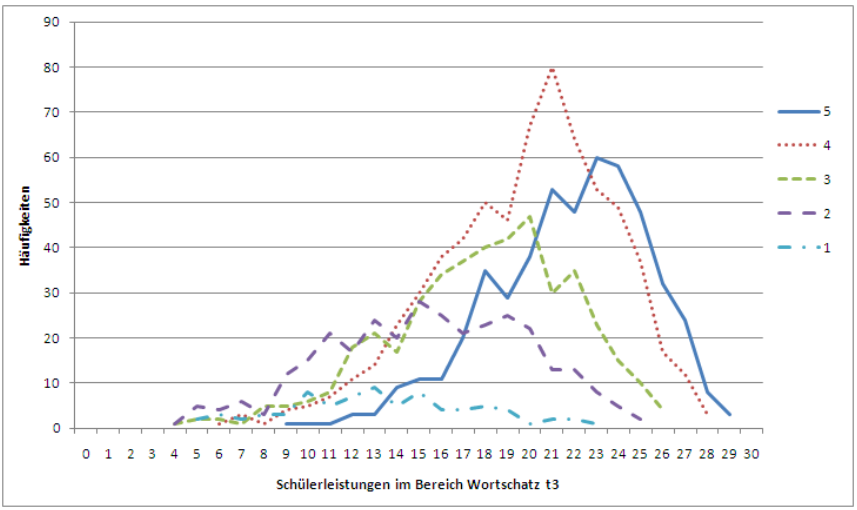

Abbildung 12: Zuordnung der Leistungseinschätzungen für den Bereich Wortschatz zu den entsprechenden Schülerleistungen zu t3

Im Vergleich zu obiger Abbildung kommt es bei der Gegenüberstellung von Wortschatzleistung und Zeugnisnoten (t3) zu nicht ganz so deutlichen Überschneidungen, aber von einer eindeutigen Zuordnung von Leistungswerten zu den Zensuren kann keine Rede sein. Viele Schüler, die auf dem Zeugnis die Noten 2 oder 3 hatten, erreichten im Wortschatztest dieselben hohen Punktwerte wie Schüler mit der Zeugnisnote 1, und unter den wenigen Schülern, die im letzten Zeugnis nur eine 5 hatten, sind ebenso einige darunter, die im Wortschatztest nicht schlechter abschneiden als andere Kinder mit den Noten 4, 3, 2 oder 1.

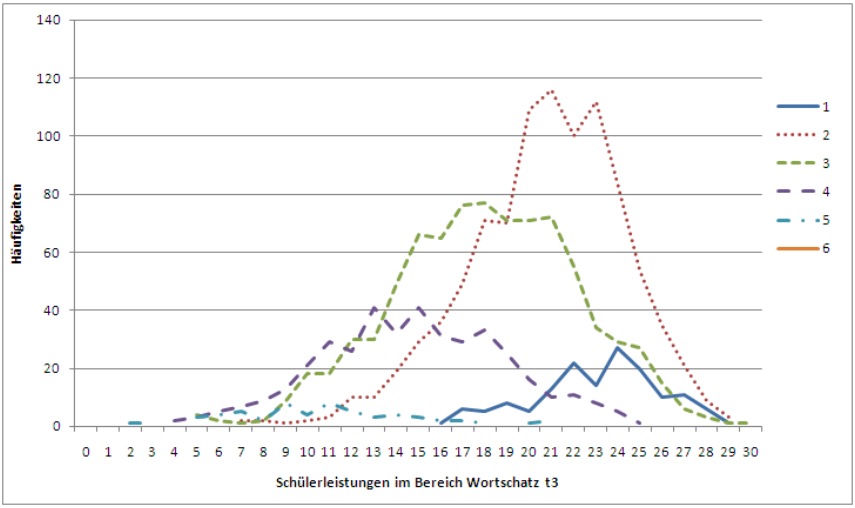

Abbildung 13: Zuordnung der Zeugnisnoten im Fach Deutsch zu Schülerleistungen im Wortschatz zu t3

Die letzten Abbildungen machen gleichermaßen die Schwächen der Orientierung am klasseninternen Bezugssystem deutlich. Durch eine Vielzahl von Studien ist belegt, dass die Vergleichbarkeit von Noten zwischen Schülern verschiedener Klassen, Schulen, Schulformen oder (Bundes-)Länder nur sehr bedingt gegeben ist (vgl. u.a. Baumert, Trautwein & Artelt, 2003; Ingenkamp, 1989; Klauer, 1989; Ziegenspeck, 1999). Objektiv gleiche Schülerleistungen können von verschiedenen Lehrern je nach Kontext und abhängig von einer Vielzahl von Faktoren sehr unterschiedlich bewertet werden. Anhand der vorliegenden Daten soll nun noch genauer geprüft werden, inwieweit Schüler mit gleichen oder ähnlichen Leistungen von den Lehrern auch entsprechend eingeschätzt werden und welche Rolle dabei der Bezugsrahmen spielt.

Hierfür wurden zunächst alle Schüler, die von ihrem jeweiligen Lehrer auf ein- und derselben Kompetenzstufe (1-5) eingeschätzt wurden, hinsichtlich ihrer tatsächlich erreichten Leistung im entsprechenden Test verglichen. Abbildung 14 stellt dies für alle fünf zu t3 erfassten Leistungsbereiche als Boxplot-Diagramme grafisch dar. Die auf der y-Achse abzulesenden Leistungswerte sind z-standardisierte Punktwerte aus den Tests, um eine Vergleichbarkeit zwischen den im Original auf unterschiedlichen Metriken erfassten Leistungsbereichen zu ermöglichen. Die farbigen Blöcke zeigen jenen Bereich an, in dem 50 Prozent der Schüler liegen, die horizontale Linie

innerhalb der Blöcke kennzeichnet die Lage des Medians. Die schwarzen Linien über- und unterhalb der Blöcke zeigen die Spannweite der gesamten Leistungsverteilung ohne Ausreißer (Kreise) und Extremwerte (Sternchen) an. Folgende Botschaften lassen sich aus der Abbildung erkennen:

- Für jeden der fünf Leistungsbereiche ist die mittlere Schülerleistung umso höher, je stärker die Lehrer die Schüler im jeweiligen Bereich eingeschätzt haben. Dieser Befund deckt sich mit der relativ hohen Ausprägung der Rangkomponente diagnostischer Kompetenz, wie sie bereits im Kapitel 7.1.2 zur generellen Güte diagnostischer Kompetenz berichtet wurde.
- Durch die z-Standardisierung wird eine sinnvolle Interpretation der Rangeinschätzung der Lehrer in Bezug auf das Leistungsniveau der Schüler möglich. Es zeigt sich, dass die Mittelkategorie („teils/teils“ hinsichtlich der Kompetenzausprägung) im Durchschnitt auch den mittleren Leistungsbereich der Schüler trifft, im Fall der Einschätzung des logisch-abstrakten Denkens sogar ziemlich genau. Entsprechend liegen die als schwächer eingeschätzten Schüler auch tatsächlich unterhalb der mittleren Leistung, die als stark eingeschätzten Schüler oberhalb.
- Die Varianz der Schülerleistungen ist bei den am schwächsten eingeschätzten Schülern in jedem Leistungsbereich am größten und wird deutlich kleiner, je besser die Schüler in den Tests abgeschnitten haben. Dies deutet darauf hin, dass es Lehrern leichter fällt, gute und sehr gute Schüler einzuschätzen. Der Effekt, dass ein für leistungsstark gehaltener Schüler (z.B., weil er einen schlechten Tag hatte, unkonzentriert war etc.) im Test eine schlechtere Leistung abgeliefert hat, als es normalerweise der Fall wäre, ist offenbar deutlich seltener der Fall als dass - im umgekehrten Fall - ein für leistungsschwach gehaltener Schüler eine ungewöhnlich hohe Testleistung erzielt hat.
- Darüber hinaus sind deutliche Überlappungseffekte erkennbar. Auch wenn die mittlere gemessene Schülerleistung je nach Lehrerurteil variiert, so ist es doch in jedem Leistungsbereich möglich und auch tatsächlich der Fall, dass Schüler mit derselben Testleistung auf jeder der Kompetenzstufen eingeschätzt wurden.

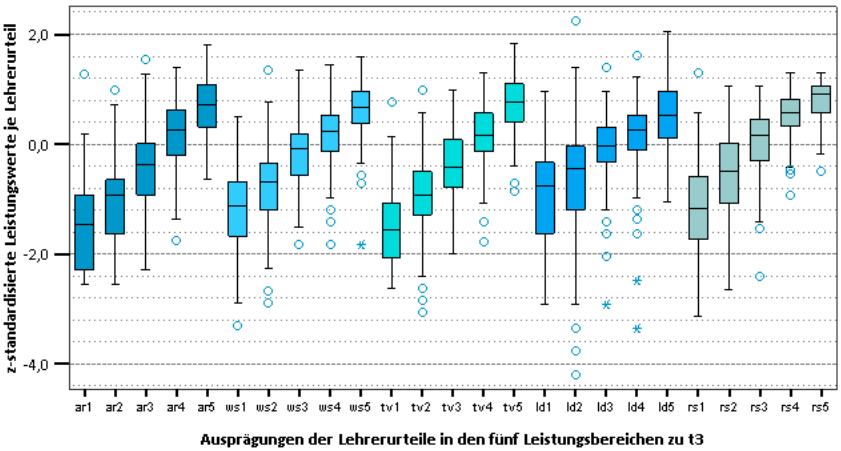

Abbildung 14: Verteilung der Schülerleistungen zu t3, gruppiert nach Lehrereinschätzungen

Anmerkung: Die Abkürzungen auf der x-Achse haben folgende Bedeutungen: ar = Arithmetik, ws = Wortschatz, tv = Textverstehen, ld = logisch-abstraktes Denken, rs = Rechtschreiben. Die Zahlen 1 bis 5 dahinter stehen jeweils für die Stufe des Lehrerurteils (1 = trifft überhaupt nicht zu ... 5 = trifft voll und ganz zu)

7.4 Ergänzende Betrachtungen anhand der Zusatzstichprobe

In einem abschließenden Schritt sollen zentrale Ergebnisse aus den bisherigen Analysen anhand der Zusatzstichprobe mit Kindern aus dem ersten Grundschuljahr ergänzt werden. Darüber hinaus soll geprüft werden, ob oder inwiefern Daten aus der Hauptstichprobe, die auf der aus Rangurteilen abgeleiteten Niveaukomponente beruhen, mit den Daten aus der korrekt operationalisierten Niveaukomponente übereinstimmen, da hierbei nicht selbstverständlich davon ausgegangen werden kann, dass durch die Transformation tatsächlich ein mit der direkt erfassten Niveaukomponente vergleichbares Maß entsteht.

Güte von Leistungsurteilen im ersten Schuljahr - Betrachtung verschiedener Komponenten diagnostischer Kompetenz

Zunächst soll geprüft werden, inwiefern die Urteilsgüte von Lehrkräften der ersten Klassen vergleichbar ist mit denen, die in der dritten und vierten Klasse gefunden wurden (vgl. hierzu Kapitel 7.1.2 ab S. 173). Die beiden Stichproben ähneln sich insofern, als dass sie beide die Grundschulzeit betreffen, und durch den Erhebungszeitpunkt zur Mitte des zweiten Halbjah-

res kannten die Lehrer der Erstklässler ihre Schüler auch etwa genauso lange wie die meisten Lehrer zu t1 in der Hauptstichprobe. Es gibt allerdings auch deutliche Unterschiede zwischen beiden Untersuchungsgruppen. Dies betrifft einerseits den Lernstand der Kinder, der in der ersten Klasse naturgemäß viel niedriger ist als am Ende der Grundschulzeit. Viele basale Kompetenzen werden im ersten Schuljahr gerade erst ausgebildet, so zum Beispiel die Lesefähigkeit. Somit stellt die Einschätzung dieser Kompetenzen andere Anforderungen an die Lehrer als in der dritten und vierten Klasse, in denen alle Schüler zumindest alle Buchstaben sicher beherrschen sollten und das Lesetempo sowie das Leseverständnis höher ausgeprägt sind. Andererseits werden in der ersten Klasse nur Verbalurteile vergeben, aber noch keine Noten. Die Verbalurteile sind zwar recht eng an den Notenstufen orientiert, aber angesichts der hohen Korrelationen, die sich zwischen Zeugnisnoten und Lehrerurteilen in der dritten und vierten Klassenstufe gezeigt haben, ist davon auszugehen, dass den Lehrern der Erstklässler mit den Noten ein bedeutsamer Anhaltspunkt für die Einschätzungen der Schülerleistungen im Einschätzbogen fehlt.

In der Zusatzstichprobe konnte eine umfangreichere Erfassung der diagnostischen Kompetenz realisiert werden, was die Urteilskomponenten betrifft. Neben der global erfassten Leistungseinschätzung im Sinne der Rangkomponente wurden außerdem aus den Bereichen Wortschatz und Textverstehen je sieben das gesamte Schwierigkeitsspektrum des Tests abdeckende Items vorgegeben, zu denen die Lehrer einschätzen sollten, ob sie von sieben zufällig ausgewählten Schülern aller Voraussicht nach korrekt gelöst werden können. In der Zusatzstichprobe ist die Bandbreite der erfassten Konstrukte im Vergleich zur Hauptstichprobe jedoch deutlich eingeschränkter, indem dort nur Einschätzungen für die Bereiche Arithmetik, Wortschatz und Textverstehen (spezifische Urteile wurden nur für die beiden letztgenannten Leistungsbereiche erfasst) erfragt wurden. Dennoch lassen sich aus der zusätzlich abgefragten Niveaueinschätzung gleich vier weitere Komponenten diagnostischer Kompetenz bilden, die im Folgenden betrachtet und - wo möglich - mit der Hauptstichprobe verglichen werden sollen.

In Tabelle 81 (S. 238) sind die Mittelwerte über alle Lehrkräfte für die verschiedenen Komponenten dargestellt. Die Rangkomponente (1) ist dabei jenes Maß, das direkt mit den Ergebnissen aus der Hauptstichprobe vergleichbar ist. Dabei ist die mittlere Güte für die Bereiche Arithmetik und Textverstehen den Werten aus der dritten und vierten Klasse sehr ähnlich (vgl. Tabelle 44, S. 175). Die mittlere Rangurteilsgüte im Bereich Wortschatz

ist in der ersten Klasse jedoch im Vergleich zur dritten und vierten Klasse deutlich geringer (r = .31 im Vergleich zu r = .56/.55./.55), wenn auch aufgrund des niedrigen Stichprobenumfangs in der Zusatzstichprobe nicht signifikant niedriger. Hierbei muss in Betracht gezogen werden, dass die Diskrepanz zwischen dem Inhalt des eingesetzten Wortschatztests und dem, was die Lehrer sich vorstellen, wenn sie nach einer Einschätzung des Wortschatzes der Erstklässler gefragt werden, auch besonders groß sein kann. Hier kann die globale Einschätzung von Kompetenzen schnell an ihre Grenzen stoßen. Darüber hinaus hat sich der in der ersten Klasse eingesetzte Wortschatztest KFT 1-3 als nicht besonders reliabel erwiesen (α = .60).

Die anderen vier Komponenten beruhen allesamt auf der Niveaueinschätzung der Lehrer, bei der für jeweils sieben vorgegebene Aufgaben aus dem Wortschatz- und dem Textverstehenstest für sieben zufällig ausgewählte Schüler eingeschätzt werden sollte, ob sie diese Aufgaben aller Wahrscheinlichkeit nach richtig gelöst haben. Für den Bereich Arithmetik fand eine derartige Einschätzung nicht statt, da sich die sehr einfachen Items des Heidelberger Rechentests - bei dem es sich um einen Speedtest handelt, bei dem es eher auf die Summe richtig gelöster Aufgaben in der vorgegebenen Zeit ankommt - nicht für die Aufgabeneinschätzung eignen. Zwar ist die Vergleichbarkeit zwischen Rang- und Niveaueinschätzung insofern eingeschränkt, als dass sich die Urteile für die Rangkomponente auf alle Schüler der Klasse beziehen, für die Niveaueinschätzungen jedoch nur auf eine Auswahl von sieben Schülern. Einerseits entsprechen jedoch in vielen Klassen die sieben (oder - je nach Teilnahmequote am Testtag - weniger) einzuschätzenden Schülern bereits der kompletten Teilnehmerschaft, da die Teilnahmequote in der Zusatzstichprobe deutlich geringer war als in der Hauptstichprobe (vgl. Tabelle 14, S. 122), andererseits kann davon ausgegangen werden, dass durch die Zufallsziehung der sieben Schüler im Mittel über alle Lehrer die Niveau- mit den Rangurteilen vergleichbar sein sollten. Auf Basis der spezifischen Urteile konnte eine weitere Rangkomponente (2) berechnet werden, die sich nun darauf bezieht, wie gut Lehrer die sieben Schüler hinsichtlich ihres Lösungsverhaltens für die sieben Items in eine Rangfolge bringen konnten. Hierbei handelt es sich nun nicht mehr um eine globale Einschätzung, sondern um eine spezifische, da den Lehrern genau bekannt war, auf welche Leistungsanforderung sich ihr Urteil bezieht. Für den Bereich Textverstehen ergibt sich daraus eine sehr ähnliche, minimal geringere mittlere Urteilsgüte von r = .56, und im Wortschatz ist die Urteilsgüte (r = .42) besser als bei der globalen Rangkomponente auf Konstruktebene, jedoch immer noch niedriger als die globale Rangkomponente in der Haupt-

stichprobe. Der Vorteil einer genaueren Rangeinschätzung durch Vorgabe der zugrundeliegenden Aufgaben (spezifische Einschätzung), wie er in anderen Studien berichtet wird (z.B. Demaray & Elliott, 1998), zeigt sich hier also nur für den Bereich Wortschatz, nicht jedoch für das Textverstehen.

Die weiteren drei Komponenten stellen die Genauigkeit der Niveaueinschätzungen im eigentlichen Sinne dar. In Spalte (3) der Tabelle 81 ist die Abweichung zwischen Lehrereinschätzung und Schülerleistung, bezogen auf die korrekt gelösten Aufgaben (Items), angegeben. Ein Wert von Null würde hier eine optimale Niveaueinschätzung indizieren, negative Werte bis zu -7 eine Unterschätzung, positive Werte bis +7 eine Überschätzung. Im Mittel findet in der Zusatzstichprobe also in beiden Bereichen eine leichte Leistungsüberschätzung statt, die im Textverstehen (+2,28) um einiges höher ausfällt als für den Wortschatz (+0,69). Dieser Befund der Überschätzung von Schülerleistungen fügt sich in die Reihe vieler gleichartiger Untersuchungsergebnisse (u.a. Feinberg & Shapiro, 2003; Hamilton & Shinn, 2003; Schrader & Helmke, 1987). Direkt aus den niveaubezogenen Urteilen ableitbar ist das Abweichungsmaß der Niveaueinschätzung (Spalte 4). Dieses wird in einen umso größeren Wert transformiert, je geringer die Niveauverschätzung bzw. die Abweichung von einer exakten Einschätzung ist. Somit kann zwar keine Aussage mehr über das Ausmaß der Über- oder Unterschätzung getroffen werden, doch während eine mittlere Niveaukomponentenausprägung nahe Null darüber hinwegtäuschen kann, dass Schüler möglicherweise deutlich über- und unterschätzt werden und sich dies insgesamt ausmittelt, kann das Abweichungsmaß als Angabe zur mittleren absoluten Verschätzung interpretiert werden. Für die vorliegenden Daten wurde die Transformation derart durchgeführt, dass der optimale Niveaukomponentenwert von Null in den Wert 7 und die maximale Verschätzung von -7 bzw. +7 in den Wert Null sowie die dazwischen liegenden Werte entsprechend ebenfalls linear umcodiert wurden. Die mittleren Abweichungsmaße von 5,37 für den Wortschatz und 4,56 für das Textverstehen bestätigen im Wesentlichen die Befunde für die Niveaukomponente (3), dass die Wortschatzleistungen etwas genauer als die Textverstehensleistungen eingeschätzt wurden.

Während die Niveaukomponenten (3) und (4) sich nur auf die Angaben stützten, wie viele Aufgaben von den Schülern gelöst wurden und ob dies mit der Schätzung durch die Lehrer übereinstimmt, ist der aufgabenbezogene Treffer (5) noch konkreter, indem dabei für jedes einzelne Item geprüft wird, ob die Lehrereinschätzung mit dem Lösungsverhalten der Kinder übereinstimmt. Der von den Lehrern erreichte Punktwert gibt an, für wie

viele Items korrekt vorhergesagt wurde, ob es vom Schüler richtig oder falsch beantwortet wurde. Der aufgabenbezogene Treffer verlangt von den Lehrern also zusätzlich eine hohe Einschätzgenauigkeit auf Aufgabenebene, nicht nur für den Test per se. Der mögliche Wertebereich variiert zwischen 0 (keine Aufgabe korrekt eingeschätzt) und 7 (alle Aufgaben korrekt eingeschätzt). Die Ergebnisse zeigen hier, dass die aufgabenbezogenen Einschätzungen im Mittel recht gut gelingen, erwartungsgemäß aber etwas niedriger ausfallen als das Abweichungsmaß. Anders als bei der reinen Niveaukomponente werden hierbei die Textverstehensitems geringfügig genauer eingeschätzt als die Wortschatzitems, was belegt, dass Niveaukomponente und aufgabenbezogener Treffer nicht zwangsläufig zu denselben bereichsspezifischen Ergebnissen führen.

Tabelle 81: Mittlere Urteilsgüte und Standardabweichungen der Lehrer aus der Zusatzstichprobe, bezogen auf verschiedene Komponenten diagnostischer Kompetenz

	1. Rangkomponente (Konstruktebene)[1]	*2. Rangkomponente (Aufgabenebene)*[2]	*3. Niveaukomponente*[3]	*4. Niveaukomponente (Abweichungsmaß)*[4]	*5. aufgabenbezogener Treffer*[5]
Arithmetik	.65 (0,49)	-	-	-	-
Wortschatz	.31 (0,38)	.42 (0,48)	0,69 (1,15)	5,37 (0,60)	4,07 (0,75)
Textverstehen	.59 (0,37)	.56 (0,45)	2,28 (1,22)	4,56 (1,05)	4,17 (0,94)

Die angegebenen Werte entsprechen …

[1] … dem Mittelwert der klassenweisen Produkt-Moment-Korrelationen zwischen individuellen Leistungseinschätzungen auf globaler Konstruktebene und den Schülerleistungen im entsprechenden Leistungsbereich (Fisher-Z-transformiert).

[2] … dem Mittelwert der klassenweisen Produkt-Moment-Korrelation zwischen der individuell eingeschätzten Summe der richtig gelösten Aufgaben (aus der Auswahl von 7 einzuschätzenden konkreten Aufgaben) und der Summe der tatsächlichen Richtiglösungen der Schüler (Fisher-Z-transformiert).

[3] … dem Mittelwert der klassenweisen mittleren Differenz zwischen Anzahl der als richtig gelöst eingeschätzten Aufgaben (aus der Auswahl von 7 Aufgaben) und der tatsächlichen Anzahl von den Schülern korrekt gelöster Aufgaben.

[4] … dem Mittelwert der in ein positives Abweichungsmaß transformierten klassenweisen mittleren Differenz zwischen Anzahl der als richtig gelöst eingeschätzten Aufgaben (aus der Auswahl von 7 Aufgaben) und der tatsächlichen Anzahl von den Schülern korrekt gelöster Aufgaben.

[5] … dem Mittelwert der klassenweisen Übereinstimmung zwischen dem vom Lehrer eingeschätzten Richtig-Falsch-Lösungsverhalten für jede der 7 einzuschätzenden Aufgaben und dem tatsächlichen Richtig-Falsch-Lösungsmuster der Schüler. Hierbei gilt die Lehrereinschätzung als korrekt, wenn entweder eine Richtiglösung oder eine Falschlösung der jeweiligen Aufgabe zutreffend vorhergesagt wurde.

Überprüfung der Transformation der Rang- in die Niveaukomponente aus der Hauptstichprobe anhand der Zusatzstichprobe

Für die differentiellen Analysen bezüglich des Einflusses von Schülermerkmalen auf die individuellen Lehrerurteile wurde mit der Differenz aus den Lehrerurteilen und auf fünf Stufen transformierten Schülerleistungen ein Abweichungsmaß gebildet. Dieses Vorgehen ist allerdings nur eine grobe Annäherung an die eigentlich benötigten Niveauurteile der Lehrer, die jedoch nicht erfasst wurden. Die gleichzeitige Erhebung von Rang- und Niveauurteilen in der Zusatzstichprobe ermöglicht es nun, die Berechnungen zur Niveaukomponente aus der Hauptstichprobe auf ihre Ähnlichkeit zu Niveauurteilen und somit ihre Zuverlässigkeit hin zu prüfen. Auch wenn dieser Vergleich zum einen durch Unterschiede in der Klassenstufe, außerdem wegen gänzlich neuer Lehrer und zum anderen durch den Vergleich von globaler und spezifischer Lehrereinschätzung etwas eingeschränkt werden muss, bietet er dennoch eine gute Möglichkeit, die Analysen zur Niveaukomponente aus der Hauptstichprobe, insbesondere in Hinblick auf die Bedingungen diagnostischer Kompetenz, zu untermauern. Für diese Überprüfung wird der Frage nachgegangen, wie sehr die Güte der individuellen und direkt erfassten Niveauurteile der Lehrkräfte mit der Güte der individuellen Niveauurteile auf Grundlage der transformierten Rangeinschätzungen, wie sie in der Hauptstichprobe berechnet wurden, zusammenhängt. Dazu wurde in nahezu identischer Vorgehensweise zur Hauptstichprobe auch hier das gesamte Leistungsspektrum für den Wortschatz- und den Textverstehenstest in gleich große Abschnitte unterteilt und die individuellen Schüler entsprechend ihrer Testleistung diesen Gruppen zugeordnet. Der einzige Unterschied zur Hauptstichprobe bestand darin, dass nicht fünf, sondern vier Leistungsgruppen gebildet wurden, weil in der ersten Klasse die Lehrerurteile nur vierstufig erfasst wurden. Anschließend wurde die Differenz zwischen vierstufigen Lehrereinschätzungen und ebenfalls vierstufiger Schülerleistung gebildet, wobei sich auch hier eine umgekehrt U-förmige Werteskala ergibt, die zwischen -3 (maximale Unterschätzung) und +3 (maximale Überschätzung) variiert und ihr Optimum bei Null hat. Dies ist dann die transformierte Niveaukomponente, die für die Überprüfung auf Klassen- bzw. Lehrerebene aggregiert (gemittelt) wurde.

Im Bereich Wortschatz ergibt sich zwischen der aus den Rangurteilen gebildeten Niveaukomponente und der direkt erhobenen Niveaukomponente eine signifikante Korrelation von r = .52 auf Lehrerebene, der Zusammenhang für den Bereich Textverstehen liegt mit r = .37 - wenn auch immer noch signifikant - etwas niedriger. Die zugrunde liegende Fallzahl beträgt jeweils 374

bzw. 369 Schüler. Dies deutet in dieselbe Richtung, wie sie auch Feinberg und Shapiro (2003) für den Zusammenhang von spezifischer und globaler Leistungseinschätzung im Bereich Lesen fanden. Beides korrelierte in ihrer Studie zu r = .66. Die genannten Korrelationen liegen zwar nur im mittelhohen Bereich und sind damit zunächst nur bedingt als Beleg für die Zulässigkeit der Transformation zu betrachten. Berücksichtigt man jedoch, dass verschiedene Komponenten diagnostischer Kompetenz in den meisten Studien nur bedingt oder gar nicht miteinander zusammenhängen, so ist die Übereinstimmung hier vergleichsweise hoch.

In ähnlicher Größe bewegen sich auch die Zusammenhänge, wenn die individuellen Differenzen zwischen Lehrerurteil und tatsächlich von den Schülern gelösten Aufgaben zunächst in ein Abweichungsmaß transformiert und danach auf Lehrerebene aggregiert werden. Die Umwandlung in ein Abweichungsmaß, bei der die optimale Niveauabweichung von Null in den größten und eine maximale Über- oder Unterschätzung in den kleinsten Wert umcodiert wird (0=5, 1/-1=4, 2/-2=3, 3/-3=2, 4/-4=1), ist sinnvoll, da sich ansonsten bei einer Mittelung auf Lehrerebene Über- und Unterschätzungen zu einem scheinbar optimalen Wert nahe Null mitteln könnten. Hierbei liegen die Zusammenhänge auf Lehrerebene für den Bereich Wortschatz bei r = .43 (signifikant auf dem 1-Prozent-Niveau), für den Bereich Textverstehen wiederum etwas geringer bei r = .30 und nur noch auf dem 5-Prozent-Niveau signifikant. Grundlage sind dabei nur noch 61 bzw. 62 Fälle, wobei für die Korrelation alle zur Verfügung stehenden Lehrkräfte in die Rechnung einbezogen wurden, auch wenn sie weniger als fünf Schüler eingeschätzt hatten.

Beispielrechnung mit den zur Niveaukomponente transformierten globalen Rangurteilen

Die eben berichteten reinen Korrelationen zwischen den auf unterschiedliche Weise gebildeten Komponenten der Urteilsgenauigkeit allein sind jedoch noch nicht allzu aussagekräftig. Zudem kann die relativ niedrige Höhe der Korrelationen bislang nicht unbedingt als Beleg dafür angesehen werden, dass die durch Transformation gewonnenen Niveauurteile tatsächlich direkt mit korrekt erfassten Niveauurteilen vergleichbar sind. Weitere Aufschlüsse bringt gegebenenfalls der Versuch, mittels beider Komponenten dieselben weiterführenden Analysen zu rechnen, wie sie in der Hauptstichprobe mit den transformierten Globalurteilen durchgeführt wurden. Wenn sich dabei zeigen sollte, dass mit beiden Komponenten vergleichbare Ergeb-

nisse erzielt werden, wäre dies ein zusätzlicher - und sichererer - Beleg für die Zulässigkeit der Komponententransformation.

Mit dem Geschlecht der Schüler und ihrem Sozialstatus stehen auch in der Zusatzstichprobe zwei Individualmerkmale zur Verfügung, die sich in der Hauptstichprobe als bedeutsame Bedingungsvariablen für die Genauigkeit der Lehrerurteile herausgestellt haben (vgl. Kapitel 7.2.3 ab S. 215). Im Folgenden wird daher verglichen, inwiefern es zu ähnlichen Ergebnissen führt, wenn mit Daten der Zusatzstichprobe dieselben Bedingungsanalysen durchgeführt werden, und zwar einmal mit der auf spezifischen Lehrerurteilen beruhenden ‚originalen' Niveaukomponente und einmal mit der aus den globalen Lehrereinschätzungen basierenden transformierten Niveaukomponente. Für letztere wurde wiederum je Leistungsbereich das insgesamt erreichte Leistungsspektrum in - diesmal vier und nicht fünf - gleich große Abschnitte eingeteilt und die Schüler entsprechend ihrer eigenen Leistung einer dieser vier Gruppen zugewiesen. Die Differenz dieses Wertes zur ebenfalls vierstufigen globalen Lehrereinschätzung sollte wie in der Hauptstichprobe eine Annäherung an die Niveaukomponente sein, wobei auch hier die absolute Abweichung vom Optimalwert Null eben aufgrund der Unkenntnis der Lehrer über die zugrunde liegenden Aufgaben nicht interpretiert werden kann. Differentielle Unterschiede je nach Geschlecht und Sozialstatus sind hingegen sehr aussagekräftig.

Die Ergebnisse der differentiellen Analysen sind in Tabelle 82 dargestellt. Für die beiden Leistungsbereiche Wortschatz (WS) und Textverstehen (TV) sind jeweils mittlere Abweichungen vom Optimalwert Null sowie die t-Werte aus den gerechneten t-Tests aufgelistet. Dabei erfolgt eine weitere Differenzierung nach der Fragegenauigkeit. Da spezifische Urteile pro Lehrer für maximal sieben Schüler vorliegen, globale Urteile aber immer für die gesamte Klasse, sind die zugrundeliegenden Stichproben unterschiedlich groß. Um zu kontrollieren, dass die Mittelwerte und Signifikanzen nicht womöglich durch die Verschiedenheit der jeweiligen Stichproben beeinflusst werden, wurden als dritte Datenbasis nur jene Schüler für die auf globalen Einschätzungen beruhende Niveauberechnungen verwendet, für die auch die spezifischen Urteile vorlagen [in der Tabelle als ‚global (gefiltert)' bezeichnet]. Minimale Abweichungen in der Stichprobengröße ergeben sich dabei dadurch, dass für einige wenige Kinder, für die spezifische Urteile vorhanden waren, die globalen Urteile fehlten.

Wie aus den Ergebnissen abzulesen ist, sind die Niveaudifferenzen je nach Fragegenauigkeit nicht deckungsgleich. Während es offenbar nur einen ge-

ringen Unterschied ausmacht, ob für die Analysen zu den globalen Einschätzungen die gesamte Klasse [global (transf.)] oder nur die Stichprobe für die spezifischen Fragestellungen [global (gefiltert)] zugrundegelegt wird (Ausnahme: Wortschatzurteile je nach Sozialstatus), unterscheiden sich die differentiellen Niveaudifferenzen je nach Art der Fragegenauigkeit und je nach Leistungsbereich. Interessanterweise fällt auf, dass sich auch in der ersten Grundschulklasse deutliche Geschlechts- und Sozialstatusunterschiede in den globalen Lehrerurteilen nur für den Bereich Wortschatz, nicht jedoch für das Textverstehen finden lassen (vgl. Tabelle 74, S. 220, wo das Schülergeschlecht ebenfalls keinen Einfluss auf die Urteilsgenauigkeit der Lehrer im Bereich Textverstehen hatte). Somit stellt das Textverstehen möglicherweise einen Leistungsbereich dar, für den differentielle Effekte mit geringerer Wahrscheinlichkeit auftreten. Für die eigentliche Fragestellung ist deshalb vor allem der Wortschatzbereich relevant, für den sowohl für das Geschlecht als auch für den Sozialstatus Gruppenunterschiede bei globalen Lehrerurteilen signifikant ausfallen, nicht jedoch für spezifische Lehrerurteile. Bei spezifischen Urteilen gibt es allenfalls Tendenzen in dieselbe Richtung, im Bereich Wortschatz in Bezug auf das Geschlecht aber noch nicht einmal diese. Somit gelangt man in Bezug auf die Fragestellung zu unterschiedlichen Ergebnissen, je nachdem, ob man sie auf Grundlage der originalen spezifischen Lehrerurteile oder auf Grundlage der aus den globalen Rangurteilen transformierten Werte berechnet. Auch wenn für die Überprüfung der Transformation der Rang- in die Niveaukomponente nur zwei Leistungsbereiche zur Verfügung standen und durch die unterschiedliche Klassenstufe, verschiedene Lehrer etc. die Vergleichbarkeit eingeschränkt ist, kann der Vergleich mit der ersten Klassenstufe nicht als erfolgreich angesehen werden. Positiv hervorzuheben ist zumindest, dass die verschiedenen Operationalisierungen nicht zu gänzlich widersprüchlichen Ergebnissen geführt haben. Vielmehr zeigen sich deutlich stärkere differentielle Effekte in Abhängigkeit von Geschlecht und Sozialstatus der Schüler - zumindest im Bereich Wortschatz -, wenn transformierte Globalurteile betrachtet werden.

Tabelle 82: Vergleich der Ergebnisse differentieller Analysen zu Unterschieden in der Genauigkeit der Niveauurteile der Lehrer in Abhängigkeit vom Geschlecht und vom Sozialstatus der Schüler, basierend sowohl auf spezifischen als auch auf globalen Urteilen der Lehrer

		Geschlecht			*Sozialstatus*		
	Fragegenauigkeit	männlich (N)	weiblich (N)	t	unterste Hälfte (N)	oberste Hälfte (N)	t
WS	spezifisch (orig.)	0,71 (179)	0,71 (206)	0,00	0,60 (159)	0,69 (196)	-0,44
	global (transf.)	0,38 (269)	0,67 (325)	-3,58**	0,42 (249)	0,64 (303)	-2,56**
	global (gefiltert)	0,42 (169)	0,76 (196)	-3,24**	0,52 (151)	0,67 (186)	-1,35
TV	spezifisch (orig.)	2,07 (176)	2,23 (200)	-0,83	2,01 (155)	2,20 (193)	-0,94
	global (transf.)	1,71 (271)	1,78 (315)	-1,12	1,72 (245)	1,80 (302)	-1,15
	global (gefiltert)	1,69 (172)	1,77 (197)	-0,96	1,73 (153)	1,78 (188)	-0,60

Anm.: Die Fragegenauigkeit ‚global (gefiltert)' reduziert die Stichprobe auf jene Schüler, die auch den spezifischen Urteilen zugrunde lagen, während bei ‚global (transf.)' alle Fälle in die Rechnungen einbezogen wurden.

**p < .01

8 Diskussion

Die vorliegende Arbeit beschäftigte sich hauptsächlich mit der Struktur sowie den Bedingungen der diagnostischen Kompetenz von Grundschullehrern am Beispiel einer bayerisch-hessischen Stichprobe (N = 155 Lehrer sowie deren 2395 Schüler). Neben der Replikation bestehender Befunde besitzt die Arbeit insbesondere dadurch Neuigkeitswert, dass sie sich auf drei Messzeitpunkte stützt und somit Aussagen zur Stabilität und zur Homogenität der diagnostischen Kompetenz ermöglicht, zwei Aspekten, die bislang noch kaum erforscht sind. Darüber hinaus wurde in dieser Untersuchung die Liste möglicher Einflussfaktoren auf die Urteilsgüte gegenüber früheren Studien deutlich erweitert.

8.1 Zusammenfassung zentraler Befunde

Zunächst wird die Vielzahl an Einzelanalysen in dieser Arbeit noch einmal zusammengefasst und um mögliche Erklärungen ergänzt. Danach folgt eine Beschreibung der spezifischen Vor- und Nachteile der Studie, bevor ein Fazit den Abschluss bildet.

8.1.1 Struktur

Im ersten Abschnitt des Ergebnisteils wurde zunächst die generelle Struktur diagnostischer Urteile und der Urteilsgüte betrachtet. Die durchgeführten Analysen sind hierbei teils Replikationen früherer Befunde, wobei die Bandbreite der untersuchten Leistungsbereiche und emotional-motivationalen Aspekte deutlich umfassender ist als in bestehenden Untersuchungen, so dass auch Vergleiche zwischen den Bereichen möglich waren. Anschließend folgten mit Analysen zur Homogenität, Stabilität und Reliabilität in diesem Umfang Nova für die Forschung zur diagnostischen Kompetenz.

8.1.1.1 Struktur der diagnostischen Kompetenz von Grundschullehrkräften

Die erste Fragestellung der vorliegenden Arbeit widmete sich dem Zusammenhang von Lehrerurteilen in verschiedenen Urteilsbereichen im Vergleich zu den jeweiligen gemessenen Eigenschaften auf Schülerseite. Anhand der drei verwendeten Messzeitpunkte und der insgesamt verwendeten fünf Leistungsbereiche, der Zeugnisnoten aus zwei Fächern sowie weiterer

fünf motivational-emotionaler und interessensbezogener Bereiche stand eine umfassende Datenbasis mit breitem inhaltlichen Spektrum für die Analysen zur Verfügung. Die Ergebnisse lassen sich dahingehend zusammenfassen, dass sich konsistent über alle drei Messzeitpunkte hinweg die Zusammenhänge der gemessenen Schülermerkmale als deutlich niedriger erwiesen als die Zusammenhänge der korrespondierenden Lehrereinschätzungen. Die Lehrerurteile korrelieren ausnahmslos über alle Bereiche signifikant, auch die emotional-motivationalen, während dies auf Schülerseite, wo insbesondere die Lernfreude und das Fachinteresse teils unkorreliert zu verschiedenen Leistungsvariablen sind, nicht immer so ist. Dies legt die Vermutung nahe, dass die untersuchten Grundschullehrer, die ihre Schüler in der Regel sowohl im Fach Deutsch als auch im Fach Mathematik unterrichten, nicht ausreichend zwischen den fachspezifischen Schülerkompetenzen einerseits und den sonstigen Eigenschaften andererseits differenzieren. Möglicherweise liegen ihren Urteilen Annahmen eines generellen Leistungsniveaus für jeden Schüler zugrunde, ein generelles Fähigkeitskonstrukt, das nicht oder nur geringfügig nach Inhalts- oder Fachaspekten aufgegliedert wird. Die Hypothese, der zufolge die Struktur und Stärke der Zusammenhänge zwischen verschiedenen Bereichen auf Lehrer- und Schülerseite nicht voneinander abweichen, bestätigt sich somit nicht bzw. nur teilweise.

Interessant wäre hier zu wissen, wonach sich der generelle Eindruck der Lehrer, der anscheinend bereichsunspezifisch ihren Urteilen zugrunde liegt, richtet. Wären es einheitlich die Leistungen in jenem Fach, in dem die Schüler höhere Kompetenzen haben, könnten Schüler durch diese ‚Generalisierung' in gewisser (ungerechtfertigter) Weise profitieren, indem sie in Leistungsbereichen, in denen sie eigentlich schlechtere Leistungen erbringen, dennoch genauso gut beurteilt werden wie in dem Fach bzw. den Fächern, in denen sie besser abschneiden. Denkbar wäre aber auch eine Orientierung an den schlechteren Leistungen oder ein ganz und gar unsystematischer Zusammenhang. Die Frage danach war allerdings nicht Gegenstand dieser Arbeit und aufgrund unterschiedlicher Skalierung von Schülereigenschaften und Lehrerurteilen sowie unterschiedlicher Konkretheit auf Schüler- und Lehrerseite auch schwierig zu beantworten.

In der Arbeit wurde sich auch der Stabilitäten von bereichsspezifischen Lehrerurteilen und Schülerleistungen sowie -merkmalen gewidmet. Während es bei den leistungsbezogenen Stabilitäten fachspezifische Unterschiede gab und Mathematikleistungen halbjährlich zwischen r_{tt} = .48 und .61 korrelier-

ten, sprachbezogene Leistungen hingegen mit Werten zwischen r_{tt} = .74 und .81 deutlich höhere Stabilitäten aufwiesen, zeigten Lehrerurteile entgegen der aufgestellten Hypothese keine bereichsabhängigen Stabilitätsunterschiede und lagen durchweg hoch zwischen r_{tt} = .70 und .83. In den emotional-motivationalen Bereichen war auf Schüler- und auf Lehrerseite das mittlere Ausmaß der Zusammenhänge geringer als in den Leistungsbereichen, bei den Urteilen (r_{tt} = .59 bis .74) aber durchweg um ca. 0,1 höher als bei den Schülern (r_{tt} = .40 bis .64). Die bereichsübergreifenden höheren Stabilitäten auf Seiten der Lehrerurteile könnten ein Anzeichen dafür sein, dass Lehrer ihre Einschätzungen (und dies betrifft nicht nur die Antworten im Fragebogen, sondern ebenso die vergebenen Zeugnisnoten) langsamer verändern, als es die Schülerleistungen und -eigenschaften tun, dass sie sich also möglicherweise nicht schnell genug an veränderte Dispositionen anpassen. Einmal gewonnene Eindrücke von den Schülerleistungen scheinen sich zu verfestigen, statt variabel anpassbar zu sein, wie es für eine gerechte Leistungsbeurteilung erforderlich wäre.

Einschränkend gilt hier wie auch generell, dass die in den Testverfahren gemessenen Leistungen der Kinder allenfalls ein grober Indikator für die Leistungen sind, auf deren Grundlage die Lehrer die Zeugniszensuren ebenso wie die erfragten Leistungseinschätzungen vergeben. Somit könnten die Schüler sich z.B. zwar im eingesetzten Testverfahren, ggf. auch unter dem Einfluss von Gewöhnungs- oder ‚Testwiseness'-Effekten, verbessern, was sich in ihren alltäglichen Unterrichtsleistungen jedoch nicht widerspiegelt und somit für die Lehrer nicht wahrnehmbar ist. Dieser mögliche Effekt sollte jedoch dadurch abgemildert werden, dass er sich primär auf das Leistungsniveau der Schüler auswirkt und weniger auf die Leistungsreihenfolge innerhalb der Klasse. Die Reihenfolge der Schülerleistungen im Test sollte insofern, auch wenn die zugrunde liegenden Aufgaben den Lehrern für ihr Urteile nicht zur Verfügung stehen, nicht wesentlich von den im Unterricht gezeigten Leistungen und deren Rangfolge abweichen. Somit dürfte die Auswirkung auf die per Korrelation bestimmte Rangkomponente diagnostischer Kompetenz und deren Stabilität allenfalls gering sein.

8.1.1.2 Güte diagnostischer Urteile

Die zentrale Fragestellung der meisten Untersuchungen zur diagnostischen Kompetenz betrifft die Güte von Urteilen. Oft wird vergleichend gefragt, welche Personen eine höhere Urteilsgüte aufweisen, bspw. Lehrer oder Eltern, Experten oder Laien, Gymnasial- oder Hauptschullehrer. Verall-

gemeinerbare Befunde dazu sind rar, nur wenige Erkenntnisse scheinen wirklich Bestand zu haben. Wie Begeny et al. (2008) nach einem Überblick über die Ergebnisse neuerer Arbeiten zur diagnostischen Kompetenz zusammenfassen, scheint die Güte von Lehrerurteilen als eine Funktion mehrerer Variablen zu variieren. Häufig fiel die Urteilsgüte dann höher aus, wenn

- Lehrerurteile normierten Leistungstests gegenübergestellt wurden (Hoge & Butcher, 1984; Hoge & Coladarci, 1989),
- korrelative Zusammenhänge anstatt prozentualer Übereinstimmungen berechnet wurden (Eckert et al., 2006; Feinberg & Shapiro, 2003),
- ältere Schüler eingeschätzt wurden (Kenny & Chekaluk, 1993) und wenn
- Lehrern die der Einschätzung zugrunde liegenden Testaufgaben für die Schüler bekannt waren (Demaray & Elliott, 1998).

In einer Reihe von Studien wurde wiederholt gezeigt, dass Lehrer ungenauere Schätzungen abgeben, wenn sie die Leistungen von Schülern nicht in normierten Leistungstests, sondern z.B. bezogen auf das sogenannte ‚Curriculum-based measurement'[6] (CBM; Deno, 1985) einschätzen sollten (Eckert et al., 2006; Feinberg & Shapiro, 2003; Hamilton & Shinn, 2003). Unbestreitbar ist ebenso der oft replizierte Befund, dass die Güte diagnostischer Lehrerurteile für die Rangkomponente im Mittel zwischen Werten von r = .50 und .60 liegt. Diese Höhe wird überwiegend als relativ gute Fähigkeitsausprägung interpretiert, wenn mögliche Messfehler auf Schüler- wie auf Lehrerseite berücksichtigt werden. Auch in der vorliegenden Untersuchung finden die genannten Werte erwartungskonform ihre Bestätigung, insbesondere in den Leistungsbereichen Arithmetik, Wortschatz und Textverstehen. Damit werden Facetten von Schülerkompetenzen tangiert, die den untersuchten Grundschullehrern aus ihrem Unterrichtsalltag sehr vertraut sein sollten. Dabei ist davon auszugehen, dass arithmetische Kompetenzen etwas besser erkennbar sein sollten als der Umfang des Wortschatzes, denn während die Rechenfähigkeit beinahe für jede Aufgabe im Mathematikunterricht gebraucht wird und Lehrer aus den Schülerleistungen leicht Rückschlüsse auf diese Fähigkeit ziehen können, ist der Wortschatz - gerade im Grundschulbereich, wo es selbst im Deutschunterricht nur selten auf kunst-

[6] Hierunter versteht man kurze, regelmäßig eingesetzte Tests mit Speedcharakter in verschiedenen - meist sprachbezogenen und mathematischen - Leistungsbereichen, anhand derer die Lehrkräfte den Lernfortschritt der Schüler besser ablesen können sollen.

volle Formulierungen ankommt - schwerer zu fassen und zu erkennen. Eine eloquente Ausdrucksweise, die Lehrer in mündlichen sowie schriftlichen Beiträgen der Schüler erkennen können, könnte zwar ein Hinweis auf einen großen Wortschatz sein, muss aber nicht unbedingt mit dem Abschneiden im verwendeten Wortschatztest in Beziehung stehen. Entscheidend ist hierbei eben auch, in welcher Art Leistungsfacetten im Unterricht direkt beobachtbar bzw. in Leistungskontrollen direkt messbar sind und inwieweit diese Facetten durch die eingesetzten Kompetenztests abgebildet werden. Plausibel erscheint in dieser Hinsicht, dass innerhalb der drei genannten Bereiche Arithmetik in Hinblick auf die mittlere Urteilsgüte vor dem Textverstehen steht und der Wortschatz der Kinder am schwersten zu beurteilen ist. Tatsächlich kann diese Vermutung auch durch die Daten bestätigt werden. Die Unterschiede zwischen den Bereichen sind zwar nicht substantiell, und die Einschätzgüte in Arithmetik ist zu den ersten beiden Messzeitpunkten entgegen der Theorie geringfügig niedriger als in den beiden anderen Bereichen. Zu t3 stimmt die angenommene Reihenfolge aber mit den Befunden überein, denn Arithmetik (r = .65) liegt vor der Güte im Bereich Textverstehen (r = .61) und für den Wortschatz (r = .55).

Mittels desselben Schemas sind ebenfalls die Urteilsgüten für die Bereiche Rechtschreiben und logisch-abstraktes Denken zu erklären. Die Rechtschreibleistung ist ähnlich wie die Rechenfertigkeit direkt mess- und beobachtbar, daher verwundert es nicht, dass auch hier die mittlere Urteilsgüte sehr hoch, sogar am höchsten, ausfällt (r = .73). Im Gegensatz dazu steht mit der niedrigsten mittleren Urteilsgüte der Bereich des logisch-abstrakten Denkens (r = .34), der kein explizit zu lehrender oder zu testender Unterrichtsgegenstand ist und für den angenommen werden kann, dass die Kluft zwischen dem, was der Test als Maß für die fluide Intelligenz tatsächlich misst, und dem, was die Lehrer sich darunter vorstellen, besonders groß ist.

Im Unterschied zur Urteilsgüte in den Leistungsbereichen erweisen sich die Einschätzungen in den untersuchten emotional-motivationalen Bereichen als deutlich niedriger, wenn auch mit zunehmender Klassenstufe die Korrelationen generell immer höher werden. Hierbei gibt es kaum Unterschiede zwischen den Bereichen, egal ob die Lehrer das Fachinteresse in Deutsch bzw. in Mathematik, die Schuleinstellung, die Lernfreude oder die Leistungsängstlichkeit der Schüler einschätzen sollten. Die Korrelationen schwanken im Bereich von r = .18 und .38. Nur sehr wenige Studien widmeten sich bislang der Urteilsgüte in nicht-kognitiven Bereichen, dennoch erwiesen sie sich konsistent als deutlich niedriger im Vergleich zu denen in

kognitiven Bereichen. Obwohl es durchaus als wichtig angesehen wird, dass Lehrer auch ein Gespür für Ängste, Vorlieben, Interessen der Schüler und ähnliches haben und entsprechend darauf reagieren können (Kultusministerkonferenz, 2004b), ist die Einschätzung derartiger Merkmale ein untergeordneter Aspekt im Schulalltag. Während Leistungen permanent beobachtet und schließlich in (begründbare) Noten übersetzt werden müssen, ist dies bei anderen Merkmalen nicht der Fall. Leistungsbeobachtung und -bewertung findet im Unterricht demnach explizit statt, Emotions- oder Motivationswahrnehmung hingegen nur als Randaspekt, der nur selten gerechtfertigt werden muss und dessen Richtigkeit der Lehrer selbst auch nur schwer überprüfen kann. Dennoch ist nicht zuletzt in den Lehrplänen explizit gefordert, die Entwicklung der Persönlichkeit zu unterstützen, was ein Erkennen von Stärken und Schwächen in verschiedenen nicht-kognitiven Bereichen voraussetzt. So heißt es beispielsweise im bayerischen Grundschullehrplan, in dem der Aspekt der Bildung und Erziehung und somit auch der Punkt ‚Persönlichkeitsentwicklung' als erstes noch vor dem Bereich Lernen und Lehren angesprochen wird: „Berücksichtigt werden mit dem Ziel der umfassenden Persönlichkeitsentwicklung nicht nur kognitive, sondern auch emotionale Aspekte und alle Bereiche des Handelns." (Bayerisches Staatsministerium für Unterricht und Kultus, 2000).

Ein weiterer Grund, weshalb die Einschätzung nicht-kognitiver Schülermerkmale so deutlich schlechter ausfällt als die Beurteilung von Leistungen, ist möglicherweise die Tatsache, dass die Übereinstimmung mit Selbsteinschätzungen der Schüler berechnet wurde. Hierbei kann nicht ausgeschlossen werden, dass die Selbsteinschätzungen der Schüler hinsichtlich der erfassten Merkmale - im Gegensatz zu den mit standardisierten und erprobten Leistungstests gemessenen kognitiven Eigenschaften - weniger verlässlich sind. Einen Hinweis darauf geben beispielsweise die Befunde von Nicholls (1978) oder Wigfield, Eccles, Harold, Freedman-Doan und Blumenfeld (1997), die zeigen, dass Kinder erst gegen Ende der Grundschulzeit eine realistische Vorstellung von der eigenen Leistungsfähigkeit zu haben scheinen. Erst ab der dritten Klasse korrelieren ihre Schätzungen des eigenen Rangplatzes in der Leistungsverteilung der Klasse signifikant mit jener ihrer Lehrer. Für nicht-kognitive Eigenschaften, über die sie selbst keine Rückmeldung erhalten, fällt Kindern eine realistische Selbsteinschätzung sicher noch schwerer als für Leistungen, zumal Konzepte wie Interesse, Ängstlichkeit u.ä. erstens für sie schwerer zu fassen sind und zweitens gerade bei jungen Menschen auch besonders situationsabhängig sein können. Deshalb ist nicht auszuschließen, dass die niedrigere Urteilsgenauigkeit weniger an

mangelnden Einschätzfähigkeiten der Lehrer als vielmehr an unzureichend ausgeprägter Selbstbeurteilungskompetenz der Grundschüler liegt.

Diese Annahme erfährt Unterstützung durch die Tatsache, dass die Urteilsgüte in allen nicht-kognitiven Bereichen von Messzeitpunkt zu Messzeitpunkt sukzessive zunimmt. In den Leistungsbereichen ist dies nur für Arithmetik der Fall. Die über die Zeit ansteigende Urteilsgenauigkeit für die Rangkomponente muss deshalb nicht zwangsläufig als ständige mittlere Verbesserung der Lehrer interpretiert, sondern könnte in dieser Hinsicht auch als Verbesserung der Selbsteinschätzungen mit zunehmendem Alter gedeutet werden (s. ebenfalls Marsh, 1990; Marsh & Craven, 1991). Ebenso ist allerdings auch eine tatsächliche Verbesserung der Lehrer bei der Einschätzung emotional-motivationaler Schülereigenschaften denkbar. Gerade deshalb, weil diese Art der Beurteilung im normalen Unterrichtsalltag nicht vorkommt, könnte es sein, dass die Lehrer sich im Laufe der drei Messzeitpunkte damit intensiver auseinandergesetzt und somit an der entsprechenden Urteilskompetenz hinzugewonnen haben.

8.1.1.3 Homogenität der Güte diagnostischer Urteile

Sowohl für die Einschätzung kognitiver als auch nicht-kognitiver Schülermerkmale fallen trotz der im Mittel sicher zufrieden stellenden Höhe der Urteilsgüte die durchweg hohen Standardabweichungen auf. Sie sind Ausdruck hoher interindividueller Variabilität in allen Bereichen. Somit lag die Frage auf der Hand, ob Lehrer, die in einem Bereich treffende Einschätzungen abgeben, dies auch in anderen Bereichen tun. Dahinter steht nicht zuletzt auch die Frage, ob es sich bei der diagnostischen Kompetenz um eine Fähigkeit handelt, die Lehrer unabhängig vom zu beurteilenden Bereich besitzen bzw. nicht (in ausreichendem Maße) besitzen. In bisherigen Forschungsarbeiten ist dieser Aspekt nur äußerst selten behandelt worden, so dass wenige Erkenntnisse darüber vorliegen und die hier vorgestellten Analysen somit großen Neuigkeitswert besitzen. Dabei ist zum einen sowohl die Vielzahl der hier untersuchten Leistungs- und emotional-motivationalen Bereiche von Vorteil, die somit mehr als nur exemplarischen Charakter haben, sondern mehrere unterrichtsrelevante Facetten gleichzeitig abdecken. Zum anderen kommt die wiederholte Messung der Aussagekraft der Analysen zugute, da damit eine größere Reliabilität der Befunde verbunden ist.

Über die drei verwendeten Messzeitpunkte hinweg zeigt sich, dass die höchsten Zusammenhänge hinsichtlich der Einschätzgüte zwischen inhalt-

lich sehr ähnlichen Bereichen auftreten. Insbesondere trifft dies auf die Urteilsgenauigkeit für die Lernfreude und die Schuleinstellungen der Schüler zu (r = .45 bis .59). Ebenso hängt die Urteilsgüte für die Lernfreude statistisch signifikant mit der Urteilsgüte für die Fachinteressen der Schüler zusammen, und dies besonders für das Fachinteresse Deutsch (r = .29 bis .39) und etwas weniger für das Fachinteresse Mathematik (r = .20 bis .22). Interessanterweise gibt es auch einen ähnlich hohen und signifikanten Zusammenhang zwischen der Urteilsgenauigkeit im logisch-abstrakten Denken und den beiden Fachinteressenbereichen (r = .18/.19). Anhand dieses zuletzt genannten Befundes lässt sich zeigen, dass sich Homogenität keineswegs nur zwischen Bereichen ergibt, die sowohl auf Lehrerurteils- als auch auf Schülerseite eng miteinander zusammenhängen. Wie in Tabelle 42 (S. 170) gezeigt, bestehen (als einige der wenigen Fälle) zwischen den Leistungen im logisch-abstrakten Denken der Schüler und sowohl ihrer Lernfreude als auch ihrem Fachinteresse im Fach Deutsch keine Zusammenhänge, während die Lehrer dort - wie überall - hohe Zusammenhänge in ihren Urteilen ausdrücken. Dass dennoch ausgerechnet bei dieser Kombination signifikante Korrelationen der Urteilsgüte zutage treten, lässt sich als Indiz dafür deuten, dass hohe synchrone Zusammenhänge nicht (nur) ein mathematisches Phänomen sind, sondern dass auch andere Gründe für diese Kovarianz bestehen. Inhaltlich kann dies nicht immer leicht erklärt werden, wie es auch beim signifikanten Zusammenhang zwischen der Urteilsgüte zum mathematischen Fachinteresse und der Rechtschreibleistung der Fall ist (r = .23).

Von besonderem Interesse sind auch hier die Leistungsbereiche. Konsistent über die Messzeitpunkte korrelieren die Lehrerurteilsgüten für Wortschatz und das Textverstehen (r = .41 bis .55), beides Bereiche, die dem Fach Deutsch bzw. dem sprachlichen Bereich zuzuordnen sind. Dies ist erwartungskonform, da beide Leistungsbereiche einander recht ähnlich sind und somit ähnliche Anforderungen an die einschätzenden Lehrkräfte stellen. Um einiges niedriger, aber immer noch signifikant, fallen die synchronen Zusammenhänge zwischen den beiden genannten Bereichen und der Urteilsgüte für die Rechtschreibung aus (r = .17 für Textverstehen, .35 für Wortschatz). Rechtschreiben ist zwar auch dem Fach Deutsch zuzuordnen, korrespondiert aber weniger mit den Leistungen in den beiden anderen eher sprachbezogenen Bereichen, und auch Lehrer sehen hier geringere Zusammenhänge (vgl. Tabelle 42, S. 170). Weiterhin plausibel erscheint, dass die Urteilsgenauigkeit für Arithmetik signifikant mit jener im logisch-abstrakten Denken zusammenhängt (r = .27), da logisches und mathematisches Denken oftmals als miteinander einhergehend angesehen werden

(Stern, 1997, 1998). Was jedoch erstaunt, sind die zu t1 (r = .25) und t2 (r = .20) ebenfalls signifikanten Beziehungen zwischen den Bereichen Arithmetik und Wortschatz, für die sich zu t3 schließlich eine Nullkorrelation ergibt. Die Schülerleistungen in beiden Bereichen hängen hingegen von Messzeitpunkt zu Messzeitpunkt kontinuierlich enger zusammen (r = .31 zu t1 bis .40 zu t3, vgl. Tabelle 40 ff.). Auch die signifikante Korrelation zwischen der Urteilsgüte für Arithmetik und Rechtschreiben zu t3 (r = .31) lässt sich nicht ohne weiteres inhaltlich begründen.

Tendenziell zeigt sich insgesamt, dass die diagnostische Kompetenz der Lehrkräfte eher bereichsspezifisch ausgeprägt ist. Vergleichbare Urteilsgenauigkeiten treten vorwiegend in inhaltlich ähnlichen Bereichen auf, und dies auch kontinuierlich über die Zeit. Somit kann die entsprechende Hypothese als bestätigt angesehen werden. Darüber hinaus gibt es auch Zusammenhänge zwischen Bereichen, die inhaltlich wenig gemeinsam haben. Woran dies genau liegt und ob dahinter eine Systematik steht, lässt sich nicht genau beantworten. Tatsache ist jedoch, dass die Annahme einer über verschiedenste Bereiche hinweg gleichermaßen ausgeprägten diagnostischen Kompetenz nicht berechtigt ist, sondern man vielmehr von einer Bereichs- bzw. Domänenspezifität dieser Fähigkeit ausgehen muss.

8.1.1.4 Stabilität der Güte diagnostischer Urteile

Ebenso wie die Betrachtung der Homogenität ist auch die Frage nach der Stabilität diagnostischer Urteilsgüte in der Forschung bislang noch nie systematisch untersucht worden. Dies erstaunt vor allem deshalb, weil der Nachweis dafür, dass sich zutreffende Lehrerurteile nicht nur zufällig, sondern auch replizierbar und wiederholt ergeben, essentiell für das Verständnis von diagnostischer Kompetenz als einer Personenfähigkeit ist. Sämtliche Untersuchungen beispielsweise, die nach den generellen Ursachen für hoch oder niedrig ausgeprägte Diagnosekompetenz fragten, wären gegenstandslos, wenn sich zeigte, dass die Genauigkeit von Beurteilungen durch Lehrer nach einem halben oder ganzen Jahr (zumindest im selben Bereich) völlig unterschiedlich ausfallen kann.

In der vorliegenden Untersuchung ist diese Forschungslücke aufgenommen worden, indem nach wiederholter Einschätzung derselben Leistungen und Eigenschaften geprüft wurde, inwiefern sich die Rangurteilsgenauigkeit der Lehrer innerhalb der Stichprobe veränderte bzw. verschob, ob also gute Diagnostiker auch nach einem halben oder ganzen Jahr noch gute Diagnosti-

ker sind. Dazu wurden Korrelationen zwischen der Urteilsgüte zu verschiedenen Zeitpunkten, aber bezogen auf dieselbe Kompetenz bzw. Eigenschaft, gerechnet. Insbesondere bei den Vergleichen zwischen dem zweiten und dem dritten Messzeitpunkt, bei denen die Konstrukte im Wesentlichen gleich geblieben sind, fallen die Stabilitäten vergleichsweise hoch aus (r_{tt} = .40 bis .49). Nur geringe Differenzen treten zwischen kognitiven und nichtkognitiven Bereichen auf, lediglich der Bereich der mathematischen Fachinteressen stellt mit r_{tt} = .28 den niedrigsten, wenngleich immer noch signifikanten Stabilitätswert. Erwartungsgemäß etwas niedriger ist die Stabilität, wenn sie zwischen dem ersten und zweiten Messzeitpunkt berechnet wird, da zu t1 die Formulierungen der Einschätzfragen etwas von den folgenden Erhebungen abwichen, sowie zwischen t1 und t3, zum einen ebenfalls wegen der unterschiedlichen Formulierungen, zum anderen aber hier besonders wegen des einjährigen statt halbjährigen Abstands zwischen den Erfassungen. Dies erklärt möglicherweise insbesondere die niedrigeren Stabilitäten in den Bereichen Lernfreude und Schuleinstellung. Einzig für die Arithmetik erreichen die Stabilitäten von r_{tt} = .16 (zwischen t1 und t2) und .21 zwischen t1 und t3 - im Unterschied zur recht hohen Stabilität zwischen t2 und t3 (r_{tt} = .44) nicht das Signifikanzniveau. Die Ausnahmestellung dieser beiden niedrigen Werte deutet demnach nicht auf einen systematischen Effekt, etwa in Bezug auf das Fach Mathematik, hin, sondern scheint andere Gründe zu haben, die möglicherweise in der abweichenden Frageformulierung für die Lehrereinschätzung zu t1 zu suchen sind.

Mittels einer weiteren Analyse auf Grundlage von Klassen, in denen es zwischen den Messzeitpunkten zu einem Lehrerwechsel gekommen war, konnte nachgewiesen werden, dass sich die oben beschriebenen systematischen Stabilitäten nur dann ergeben, wenn sie sich auch auf gleiche Lehrer beziehen. Somit kann die Annahme bestätigt werden, dass die Güte diagnostischer Urteile nicht nur bereichsspezifisch, sondern ebenso replizierbar und relativ überdauernd ausgeprägt ist. Spricht man von diagnostischer „Kompetenz“ im Sinne Weinerts (2001), so schließt der Begriff gleichzeitig auch die prinzipielle Erlernbarkeit mit ein, was insbesondere bei Lehrern, deren Urteile sich als wenig zutreffend erwiesen haben, wünschenswert wäre. Wie viel Stabilität - bei welchen Lehrern - ist demnach eigentlich gewollt? Korrekt urteilende Lehrer sollten diese Fähigkeit nach Möglichkeit beibehalten, während für ungenau urteilende Lehrer eine sich zeigende Stabilität im Grunde ein ungünstiger Befund ist. Genauere Analysen dazu wurden im Rahmen dieser Arbeit nicht durchgeführt. Die Tatsache, dass in allen emotional-motivationalen Bereichen sowie in den Bereichen Arithmetik und Textver-

stehen die mittlere Rangurteilsgüte über alle Lehrer leicht, aber kontinuierlich ansteigt (vgl. Tabelle 44 f.), ist möglicherweise ein Hinweis darauf, dass es tatsächlich zu einer minimalen Verbesserung der Lehrer über die Zeit kommt. Ob dies systematisch auf anfänglich ungenauer urteilende Lehrer zurückzuführen ist, bleibt allerdings offen.

Bei allen Ergebnissen sowohl zur Stabilität als auch zur Homogenität der diagnostischen Urteilsgüte ist zu berücksichtigen, dass sie durch Korrelation von Korrelationswerten gewonnen wurden. Dieses Vorgehen ist für die hier vorliegende Datenstruktur methodisch derzeit die einzige Möglichkeit, die entsprechenden Werte zu berechnen. Dabei vergrößern sich allerdings die Messfehler, die aufgrund der Reliabilitätseinschränkungen der beteiligten Variablen bereits in jeder der einzelnen Korrelationen stecken. Insofern sind sowohl die berichteten synchronen als auch die diachronen Zusammenhänge mit Werten um r = .40 höher zu gewichten, als es die reine Höhe zunächst erscheinen lässt. Nicht zuletzt vor dem Hintergrund, dass die mittlere Urteilsgüte insgesamt nur zwischen r = .5 und .6 liegt, sind die Homogenitäts- und Stabilitätsindikatoren als sehr bedeutsam zu bezeichnen.

8.1.1.5 Reliabilität diagnostischer Urteile

Reliabilitätsprüfung für die Rang- und Niveaukomponente diagnostischer Urteile

In der bisherigen Forschung zur diagnostischen Kompetenz ist zwar immer wieder betont worden, dass eine Unterteilung in die verschiedenen Komponenten (Rang-, Niveau- und Streuungskomponente) vorgenommen werden muss, da sie unterschiedliche Sachverhalte des Lehrerurteils abbilden. In der überwiegenden Anzahl der Untersuchungen stand - wie auch in der vorliegenden - jedoch die Rangkomponente im Fokus. Dies mag nicht zuletzt auch am Stellenwert liegen, den die im deutschsprachigen Raum seit Jahrzehnten renommiertesten Forscher auf diesem Gebiet ihr seit vielen Jahren beimessen: „Während wir in den anderen beiden Genauigkeitskomponenten [Anm. d. Verf.: gemeint sind die Niveau- und die Streuungs- bzw. Differenzierungskomponente] in erster Linie den Ausdruck bestimmter pädagogisch wichtiger Urteilsvoreingenommenheiten gesehen haben, betrachten wir die Vergleichskomponente [Anm. d. Verf.: gemeint ist hiermit die Rangkomponente] als Indikator für die diagnostische Kompetenz im eigentlichen Sinne.“ (Schrader & Helmke, 1987, S. 35) An der Popularität der Rangkomponente in der Forschung hat sich bis in die Gegenwart nicht viel geändert, weshalb Schrader auch im Jahr 2009 konstatiert: „Zur Erfassung diagnosti-

scher Kompetenzen sind klassenweise berechnete Korrelationen zwischen Urteilen und Kriterien das vorrangig eingesetzte Verfahren." (Schrader, 2009, S. 242). Gleichzeitig, und auch darauf weist Schrader (2009) hin, werden in jüngster Zeit aber auch im Rahmen der Generalisierbarkeitstheorie Methoden entwickelt, bei denen Personen anhand mehrerer Merkmale beurteilt und die resultierenden Genauigkeitskomponenten miteinander in Beziehung gesetzt werden (vgl. Greb & Lipowsky, 2009).

Wenn über die Ausprägung der diagnostischen Kompetenz von Lehrkräften geschrieben wird, dass sie im allgemeinen hinreichend ausgeprägt sei, wird dies in der Regel damit begründet, dass Lehrer die Rangreihe von Schülerleistungen einigermaßen adäquat durch Urteile oder Noten bestimmen können (z.B. Langfeldt, 2006). Somit wird implizit ausgesagt, dass die Rangkomponente ein verlässlicher Indikator für die Diagnosekompetenz sei. Dabei ist die empirische Befundlage rund um die Rangkomponente durchaus durchwachsen. Einerseits spricht allein die jahrelange Verwendung dieses Maßes, seine häufige Nennung in Publikationen dafür, dass es einer ausgiebigen wissenschaftlichen Prüfung standgehalten haben muss. Auch der Nachweis der zeitlichen Stabilität, der erst jüngst (und insbesondere durch die vorliegende Arbeit) erbracht wurde, kann als Beleg für deren Verlässlichkeit angesehen werden. Andererseits können auch entgegengesetzte Befunde Zweifel wecken. Dazu zählt als wichtiger Aspekt die bis heute erfolglose Suche nach Lehrermerkmalen, die sich als ursächlich für die diagnostische Kompetenz in bestimmten Bereichen erweisen. Wenn die diagnostische Kompetenz (im Sinne der Rangkomponente) ein Personenmerkmal, eine Fähigkeit, die den Lehrkräften in verschiedenem Ausmaß innewohnt, sein sollte, dann wäre zu erwarten, dass sich dazu irgendein messbares Korrelat in den Lehrereigenschaften findet. Dieses ist trotz jahrelanger Forschung bislang nicht geschehen. Auch dann, wenn die Rangkomponente in Beziehung zum Beispiel zum Lernerfolg der Schüler gesetzt wurde, waren einheitliche Ergebnisse selten. Oft zitiert ist die Erkenntnis, dass sich eine hohe Diagnosekompetenz der Lehrer förderlich auf den Leistungszuwachs der Schüler (im Fach Mathematik) auswirkt, allerdings nur dann, wenn er mit einem hohen Maß an Strukturierung im Unterricht einhergeht (Schrader & Helmke, 1987). Interessant und irritierend gleichzeitig ist an diesem Befund, dass eine hohe Diagnosekompetenz gepaart mit wenig Strukturierung sogar zu schlechteren Leistungszuwächsen führte als bei Lehrern, die - unabhängig vom Ausmaß ihrer Strukturierung - über eine niedrige Diagnosekompetenz verfügen. Die Interpretation der Autoren, dass Lehrer mit hoher diagnostischer Kompetenz und gleichzeitig geringer Strukturierung die

Schüler verunsichern würden und so der niedrigere Leistungszugewinn zu erklären sei, kann durchaus infrage gestellt werden.

Auch wenn nach dem Zusammenhang von Diagnosekompetenz und anderen vom Lehrer beeinflussten und leistungsbeeinflussenden Unterrichtsmerkmalen gesucht wurde, ergaben sich für die Rangkomponente keine Effekte. So fanden Anders et al. (2010) zwar für die Niveaukomponente einen Zusammenhang mit der Unterrichtsqualität (gemessen an der hilfreichen Auswahl von Aufgaben in Klassenarbeiten) im Fach Mathematik, nicht jedoch für die Rangkomponente.

Aus den beschriebenen Befunden ergab sich für diese Arbeit die Frage nach der Reliabilität der Rang- und Niveaukomponente diagnostischer Kompetenz. Wünschenswert wäre in diesem Zusammenhang gewesen, wenn Urteile derselben untersuchten Lehrer zu mindestens einer weiteren Klasse vorgelegen hätten, um zu prüfen, ob sich eine charakteristische, personenspezifische Urteilsgenauigkeit unabhängig von der einzuschätzenden Schulklasse zeigt. Da dies jedoch nicht gegeben war, wurden behelfsmäßig die einzelnen Klassen per Zufallsprinzip in zwei gleich große Hälften geteilt und für jede Klassenhälfte daraufhin die Urteilsgenauigkeiten berechnet. Die Erwartung, eine hohe Kongruenz zwischen jeweils beiden lehrerspezifischen Werten zu finden, erfüllte sich sehr deutlich für die Niveaukomponente. Obwohl das Ausmaß der Über-, Unter- oder Korrekteinschätzung zwischen den beiden Klassenhälften im Mittel zwar nicht identisch ist und es zu einer durchschnittlichen Abweichung der beiden Werte in Höhe von etwa 0,4 auf der neunstufigen Skala kommt, fallen die Zusammenhänge zwischen den Klassenhälften über alle Lehrer mit Werten zwischen r = .51 und .84 für alle Leistungsbereiche und Messzeitpunkte durchweg signifikant aus (s. Tabelle 51, S. 191). Dies wird auch durch die Ergebnisse aus der Zusatzstichprobe bestätigt, wenn dort die reine Niveaukomponente (auf Grundlage der Summe korrekt gelöster Items) betrachtet wird (r = .80 und .73); beim schwerer einzuschätzenden Treffervektor wird das Ergebnis des Reliabilitätstests nur für den Bereich Textverstehen auf dem 5-Prozent-Niveau signifikant, allerdings nicht mehr für den Bereich Wortschatz, was jedoch auch an der hier deutlich geringeren Stichprobengröße von N = 26 liegt.

Somit findet sich in den geschilderten Ergebnissen bestätigt, dass das Ausmaß der Verschätzung (Über-, Unter- oder Korrekteinschätzung, i.d.S. also die Güte der Niveauurteile) relativ unabhängig von den konkret eingeschätzten Schülern zu sein scheint und sich vielmehr in sehr ähnlicher Weise auch für unterschiedliche Schülergruppen innerhalb der eigenen Klasse zeigt.

Unberührt von diesem Befund bleibt die Möglichkeit, dass einzelne Lehrer dennoch bestimmte Schüler stärker über- bzw. unterschätzen als andere. Dadurch, dass die durch die Auswahl jedes zweiten Schülers gebildeten Klassenhälften in den relevanten Aspekten Geschlecht, Leistung und soziale Herkunft miteinander vergleichbar sind, würde sich also auch eine je nach Schülermerkmal differenzierte Über- oder Unterschätzung gleich verteilen und die in Tabelle 51 (S. 191) dargestellten hohen Korrelationen kaum beeinflussen.

Ähnlich hohe Zusammenhänge wie die bei der Niveaukomponente gefundenen sind auch für die Rangkomponente diagnostischer Kompetenz erwartet worden. Da die Korrelation zwischen Schülerleistungen bzw. -eigenschaften und Lehrerurteilen als zentraler Indikator für die diagnostischen Fähigkeiten von Lehrern angesehen wird, sollte es - theoretisch - keinen Unterschied machen, ob die ganze Klasse eingeschätzt wird oder nur ein Teil der Klasse. Schließlich sind Erhebungen, bei denen alle Kinder der Klasse teilnehmen, die große Ausnahme, da durch Krankheit am Testtag, fehlende Genehmigung o.ä. immer einige Schüler nicht anwesend sind. Die errechnete Rangkomponente sollte von derartigen Ausfällen unberührt bleiben. Die in der vorliegenden Untersuchung gefundenen Ergebnisse sprechen jedoch eine ganz andere Sprache. Von 21 untersuchten Bereichen zu drei Messzeitpunkten fanden sich lediglich bei zweien signifikante Zusammenhänge zwischen der Urteilsgüte in der einen und der anderen Klassenhälfte, konkret in Arithmetik ($r = .53$) und Wortschatz ($r = .56$) in der Klassenstufe drei. In allen anderen Bereichen bewegten sich die Korrelationen unsystematisch nahe Null und streuten von $r = -.33$ bis .17. Gleichfalls Ausdruck der beinahe riesigen mittleren Güteunterschiede zwischen den beiden Klassenhälften sind die einfachen Differenzen zwischen den gruppenbezogenen Korrelationen, die durchschnittlich 0,37 über alle 21 Bereiche betragen (nicht separat berichtet). Bedenkt man, dass die mittlere Urteilsgüte gerade für die nicht-kognitiven Bereiche in den Gesamtklassen nur zwischen $r = .18$ und .38 liegt (Tabelle 45, S. 176), dann erscheinen die gruppenbezogenen Unterschiede, die in vielen Bereichen höher sind als die Gesamtkorrelationen, geradezu enorm.

Vergleichbar fällt das Bild für die Zusatzstichprobe aus. Auch hier hängen die klassenhälftigen Rangkomponenten für Wortschatz und Textverstehen nicht miteinander zusammen, weder auf Grundlage der globalen Urteile noch auf Grundlage der aus den spezifischen Niveauurteilen abgeleiteten Rangkomponente. Auch hier gibt es allerdings eine Ausnahme, nämlich -

wie zu t1 in der Hauptstichprobe - den Bereich Arithmetik, für den die Rangurteilsgüte für die eine Klassenhälfte zu r = .68 signifikant mit jener für die andere Klassenhälfte korreliert. Die Hypothese, dass die Höhe der Urteilsgüte in beiden Klassenhälften vergleichbar ausfällt, kann somit nur für die Niveau-, nicht jedoch für die Rangkomponente bestätigt werden. Ganz offensichtlich kann die Korrelation zwischen Lehrerurteilen und Schülerleistungen als Maß für die diagnostische Kompetenz im Sinne der Rangkomponente stark in Abhängigkeit von der An- oder Abwesenheit einzelner Schüler, ihrer Motivation oder Konzentration variieren und ist somit als generelles oder gar alleiniges Maß für die diagnostische Kompetenz der Lehrer nur bedingt geeignet.

Es schließt sich die Frage an, wie die eklatanten Reliabilitätsunterschiede zwischen Niveau- und Rangkomponente zu erklären sind. Eine Antwort darauf ist nicht leicht zu finden, zumal sie durch die drei abweichenden (und signifikanten) Korrelationen für die Rangkomponente differenziert ausfallen muss. Zunächst soll dabei noch einmal auf die Vergleichbarkeit der Klassenhälften eingegangen werden, die durch die Auswahl jedes zweiten Schülers (odd-even-Methode) aus der nach Geschlecht und Namen sortierten Schülerliste gebildet wurden. Während sich in t-Tests keine Unterschiede hinsichtlich der mittleren Leistung und des mittleren Sozialstatus der Schüler in den beiden Hälften zeigte, die Geschlechterverteilung schon durch die Teilungsmethode gleichmäßig war und dies als Hinweis auf die gute Vergleichbarkeit der beiden Substichproben gab, waren die korrelativen Analysen aufschlussreicher. Die Zusammenhänge zwischen den mittleren Leistungen je Klasse korrelierten - mit wenigen Ausnahmen - in allen Bereichen und zu allen Messzeitpunkten signifikant, die mittleren Standardabweichungen je Klassenhälfte jedoch nur in einem einzigen Bereich (Arithmetik zu t1). In Anbetracht der Befunde zum Zusammenhang von Leistungsheterogenität in der Klasse und diagnostischer Kompetenz der Lehrer (vgl. Kapitel 7.2.2 ab S. 207) ist dies ein wichtiger, zu beachtender Aspekt. Da in dieser Arbeit nachgewiesen wurde, dass die Rangurteilsgüte umso besser ausfällt, je größer die Streuung der Leistungen und emotional-motivationalen Eigenschaften in der Klasse ist, sollte idealerweise auch die Streuung in den Klassenhälften vergleichbar sein, wenn die Reliabilität der Urteile untersucht werden soll. Dies ist aber offenbar nur in der dritten Klassenstufe im Bereich Arithmetik der Fall, und gerade dort ergibt sich auch eine signifikante Korrelation der klassenweisen Rangurteilsgüte zwischen den Klassenhälften. Die vergleichbare Streuung scheint hier also zu ebenfalls vergleichbarer Bewertungsakkuratheit geführt zu haben. Allerdings kann dies nicht die allei-

nige Erklärung sein, denn in den beiden anderen Leistungsbereichen mit vergleichbarer Übereinstimmung der Urteilsgenauigkeit zwischen den Klassenhälften bei der Rangkomponente diagnostischer Kompetenz (Wortschatz zu t1 und Arithmetik in der Zusatzstichprobe) liegt keine signifikante Ähnlichkeit hinsichtlich der Streuung zwischen den Klassenhälften vor.

Für eine alternative Erklärung hilft es, sich noch einmal die Berechnungswege zu vergegenwärtigen, die zu den Korrelationen geführt haben. Im Fall der Niveaukomponente handelt es sich um den Zusammenhang von je Klassenhälfte gemittelten Differenzen von Lehrerurteilen und Schülerleistungen. Da aus vielen Untersuchungen hervorgeht, dass besonders Grundschullehrkräfte zu einer generellen Überschätzung von Leistungen neigen (Bates & Nettelbeck, 2001; Begeny et al., 2008; Hamilton & Shinn, 2003; Schrader & Helmke, 1987), könnte davon ausgegangen werden, dass sich diese Tendenz gleichmäßig für beide - zufällig erzeugten und hinsichtlich zentraler Eigenschaften einander sehr ähnlichen - Klassenhälften zeigt. Lehrer, die generell zu einer höheren Leistungsüberschätzung (oder auch zur Überschätzung von nicht-kognitiven Merkmalen) neigen, zeigen diese Tendenz aller Wahrscheinlichkeit nach gleichermaßen stark für beide Klassenhälften (wenn auch nicht notwendigerweise für alle Schüler innerhalb dieser Gruppen in gleicher Weise). Weicht eine Schülereinschätzung in der vorliegenden Befragung einmal von dieser generellen Tendenz ab, führt dies zwar zu einer Reduktion des mittleren Differenzwertes, ohne jedoch etwas an der grundsätzlichen Überschätzungstendenz zu ändern. Gleiches trifft in ähnlicher Weise für grundsätzlich eher unterschätzende Lehrer zu. Einzelne ‚Fehleinschätzungen' oder Abweichungen wirken sich somit nur in geringem Maße auf die generell vorhandene Urteilstendenz aus.

Anders scheint dies bei der Rangkomponente zu sein. Zum einen kann man hier nicht von generellen Tendenzen sprechen, denn es geht nur darum, Schüler entsprechend ihrer Leistungen oder Persönlichkeitseigenschaften korrekt in eine Reihenfolge zu bringen. Etwas Analoges zur Über- oder Unterschätzung existiert hierfür nicht. Zum anderen wirkt sich - gerade in kleineren Klassen - eine geringe Abweichung der eingeschätzten von der tatsächlichen Rangfolge gleich in einem spürbar niedrigeren Korrelationskoeffizienten aus, und dies umso stärker, als dass ja nicht eine tatsächliche Rangpositionszuweisung für jeden einzelnen Schüler durch den Lehrer erfolgte, sondern die Rangfolge nur auf Grundlage der fünfstufigen Einschätzungen gebildet wurde. Dies gegenübergestellt zu den viel feinschrittiger ausgeprägten Schülerleistungen reduziert ohnehin schon die Höhe der Kor-

relationen, und zwar bei Anwendung der Pearson'schen Produkt-Moment-Korrelation noch mehr als bei nicht-parametrischer Korrelation nach Spearman. Der ohnedies schon vorhandene Einfluss von Messfehlern, die sich auf Seiten der Schüler wie der Lehrer kumulieren und die Korrelation beeinträchtigen, wird somit noch vergrößert. Die Wahrscheinlichkeit einer Reduktion der Split-half-Reliabilität ist für die Rangkomponente diagnostischer Kompetenz entsprechend deutlich höher als für die Niveaukomponente. Hinzu kommt, dass Abweichungen zwischen Schülerleistungen und Lehrerurteilen nicht automatisch als Fehleinschätzung des Lehrers ausgelegt werden dürfen, da sie natürlich auch durch eine von den sonstigen (dem Lehrerurteil zugrunde liegenden) Schülerleistungen abweichende Testleistung entstehen können. Dies trifft zwar auch für die Niveaukomponente zu, wie beschrieben wirken sich dort die Abweichungen aber weniger stark aus als bei der Rangkomponente.

Die gefundene Nicht-Reliabilität der Rangkomponente diagnostischer Kompetenz ist somit möglicherweise weniger als Unfähigkeit der Lehrer, ihre Schüler in eine Rangfolge zu bringen, anzusehen, als vielmehr ein Beleg dafür, dass die Verwendung eines Korrelationskoeffizienten als Maß für die diagnostische Kompetenz deutlich fehleranfälliger ist als beispielsweise das Abweichungsmaß für die Niveaueinschätzung. Selbstredend geben die verschiedenen Urteilskomponenten auch Auskunft zu ganz unterschiedlichen diagnostischen Fähigkeiten der Lehrer, wie insbesondere Schrader und Helmke mehrfach beschrieben haben. Dennoch erscheint es zumindest fragwürdig, ein so anfälliges Maß wie die Rangkomponente als zentralen Indikator für die diagnostische Kompetenz in der Forschung zu verwenden und in Publikationen darzustellen. Eine ähnliche Reliabilitätsprüfung, wie in dieser Arbeit angestellt, ist in der Literatur, soweit dem Autor bekannt, nicht zu finden. Es wäre jedoch überaus interessant zu sehen, inwiefern die Vielzahl publizierter Ergebnisse zur Rangkomponente noch Bestand hat, wenn sich erwiese, dass durch Hinzunahme oder Weglassen von Schülern in den untersuchten Klassen, was sich durch Fehlen von Schülern und somit Testen in nicht vollständigen Klassen in vielen Fällen fast zwangsläufig ergibt, die Urteilsgüte der individuellen Lehrer massiv variieren kann.

8.1.2 Bedingungen der diagnostischer Kompetenz von Grundschullehrkräften

Insbesondere die Erkenntnis, dass es große interindividuelle Unterschiede hinsichtlich der Urteilsgüte zwischen verschiedenen Lehrern gibt, führte zu

der Frage nach den zugrunde liegenden Ursachen. Für jeden einzelnen Lehrer, in besonderer Weise aber auch für die Lehrerausbildung, wäre es von großer praktischer Bedeutung, wenn man wüsste, welche Merkmale im Zusammenhang mit hoher Diagnosekompetenz stehen. Erkenntnisse dazu könnten u.a. genutzt werden, um in der Lehreraus- und -weiterbildung Lehrkräfte dafür zu sensibilisieren, dass bestimmte Effekte oder Faktoren ihre Urteilsgenauigkeit beeinflussen können. Nachdem die bisherigen, mitunter durchaus intensiven Forschungsbemühungen im Wesentlichen ergebnislos verliefen, wurde in der vorliegenden Arbeit in Erweiterung früherer Studien sowohl differenziert für verschiedene Leistungsbereiche und emotional-motivationale Faktoren als auch für mehrere Messzeitpunkte nach Variablen gesucht, die mit der Güte diagnostischer Urteile im Zusammenhang stehen könnten. Eine Vielzahl von unterschiedlichen, theoretisch begründbaren Faktoren wurde auf verschiedenen Ebenen - die des Lehrers selbst, die der Klasse und die der individuellen Schüler - in die Analysen einbezogen und sowohl die Rang- als auch die Niveauurteilsgenauigkeit berücksichtigt. Dabei wurde davon ausgegangen, dass die Urteilsgüte eben nicht nur auf Eigenschaften des Lehrers zurückzuführen ist, sondern immer im Zusammenspiel mit den zu beurteilenden Schülern individuell oder im Klassenkontext entsteht.

Insgesamt führte die Suche nach Bedingungen der diagnostischen Kompetenz in dieser Arbeit nur zu wenigen neuen Erkenntnissen. Die Versuche, einen unikausalen Zusammenhang zwischen einzelnen Merkmalen von Lehrern, Klassen oder Schülern und der Güte der Lehrerurteile in verschiedenen Bereichen aufzudecken, verliefen auch hier weitestgehend erfolglos. Wenn überhaupt, zeigten sich nur für einzelne Bereiche geringe Zusammenhänge, so dass kaum von einem systematischen Einfluss der untersuchten Faktoren ausgegangen werden kann.

8.1.2.1 Lehrermerkmale

Auf Seiten der Lehrermerkmale wurde eine ganze Reihe von beeinflussenden Faktoren angenommen, nämlich die Berufserfahrung, das Geschlecht, die Lehrdauer in der derzeitigen Klasse, die Fähigkeit der Lehrkräfte zur Perspektivenübernahme, die Anzahl besuchter relevanter Weiterbildungen bzw. Studienseminare, ihr Perfektionsstreben sowie ihre Einstellung gegenüber der Bedeutung diagnostischer Kompetenz. Für keine dieser Variablen zeigte sich - auch nicht für einzelne Urteilsbereiche - die jeweils erwartete Tendenz.

Da manche Zusammenhänge in der vierten Klassenstufe für die Rangkomponente sogar ein negatives Vorzeichen tragen, kann noch nicht einmal eine Tendenz angenommen werden. Vereinzelt traten auch für die sonstigen Lehrermerkmale signifikante Korrelationen auf, die beispielsweise andeuteten, dass männliche und weibliche Lehrkräfte sich zu manchen Messzeitpunkten hinsichtlich der Arithmetikeinschätzungen voneinander unterscheiden. Da sich diese Befunde jedoch nicht auch für weitere Messzeitpunkte oder Leistungsbereiche bestätigten, muss eher von Zufallsbefunden ausgegangen werden. An dieser Stelle zeigt sich zumindest ein Vorteil dessen, dass drei Erhebungen und mehrere Leistungs- und emotional-motivationale Bereiche für die Analysen zur Verfügung standen, denn somit kann eine Überinterpretation singulärer Signifikanzen relativiert bzw. ausgeschlossen werden.

Neben den genannten Personenmerkmalen wurden auch weitere Lehreraspekte mit in die Untersuchung einbezogen, die nicht direkt als Bedingungen zu verstehen sind, sondern eher Aufschluss über die Selbstwahrnehmung der Lehrer geben. Dazu zählen die Selbstwahrnehmung der eigenen diagnostischen Kompetenz im Leistungsbereich sowie das selbst eingeschätzte Ausmaß an Schwierigkeiten beim Beurteilen und der jeweilige Zeitbedarf für die Einschätzungen. Dahinter stand die Frage, ob Lehrer ein Gespür für die eigenen Fehler haben und ob genauer urteilende Lehrer möglicherweise länger für ihre Urteile brauchen, weil sie gründlicher darüber nachdenken, oder sogar schneller sind, weil ihnen Urteile ohne langes Grübeln schnell von der Hand gehen. Die Ergebnisse zeigen jedoch, dass auch keiner dieser Faktoren mit der Urteilsgüte einhergeht, und dies über alle Bereiche und Messzeitpunkte hinweg.

Die Vermutung, dass einzelne Lehrermerkmale einen Zusammenhang zur diagnostischen Kompetenz aufweisen und somit möglicherweise Erklärungspotential für die Unterschiedlichkeit zwischen Lehrkräften bieten, konnte in dieser Arbeit somit auch nach deutlicher Erweiterung der untersuchten Variablen nicht bestätigt werden. Es bleibt folglich die Erkenntnis analog zu dem Persönlichkeits- und dem Prozess-Produkt-Paradigma der Lehrerforschung, dass von einzelnen Lehrermerkmalen also nicht nur keine Effekte auf den Lernerfolg der Schüler ausgehen (vgl. Rheinberg et al., 2001), sondern ebenso wenig auf die diagnostische Kompetenz der Lehrer. Es spricht einiges dafür, auch in diesem Forschungsfeld die Expertise der Lehrer stärker zu betrachten und zum Beispiel auf die den diagnostischen Urteilen zugrunde liegenden Wissensaspekte zu fokussieren. Ausgehend von

Shulman's Taxonomie des Professionswissens von Lehrern (Shulman, 1986, vgl. auch Kapitel 3.6.1.2) besteht ein möglicher Ansatz beispielsweise darin, das fachdidaktische Wissen bzw. zusätzlich - wie in der COACTIV-Studie (Brunner, Anders, Hachfeld & Krauss, 2011) theoretisch zugrunde gelegt - das pädagogisch-psychologische Wissen der Lehrer genauer zu erfassen und in Beziehung zur Urteilsgenauigkeit zu setzen. Bislang fehlen dazu aber noch empirische Erkenntnisse.

Darüber hinaus ist auch die Rolle der in Kapitel 3.5 beschriebenen Urteilsfehler und Urteilstendenzen im Zusammenhang mit diagnostischer Kompetenz noch nicht systematisch untersucht worden. Eine genaue Prüfung, welche Art Fehler bei unzutreffenden Lehrerurteilen vorliegen und ob sich hier möglicherweise systematische Effekte in Hinblick auf Merkmale der Urteilenden oder zu Beurteilenden nachweisen lassen, könnte zu einer stärkeren Bewusstmachung dieser Einflüsse bei den Lehrern genutzt werden und durch derart ausgelöste selbstreflexive Prozesse zu einer höheren Urteilsgenauigkeit führen. Die empirische Erforschung von Urteilsfehlern kommt jedoch kaum um experimentelle Untersuchungsdesigns herum, so dass dieser Aspekt auch in der vorliegenden Arbeit nur theoretisch angeschnitten, aber nicht praktisch in zu beantwortende Fragestellungen umgesetzt werden konnte.

8.1.2.2 Klassenmerkmale

In der Annahme, dass die Güte der Lehrerurteile auch von Merkmalen und der Zusammensetzung der Schulklasse beeinflusst sein kann, wurden in der vorliegenden Untersuchung die Klassengröße, die Anzahl einzuschätzender Schüler, der Migrantenanteil, Leistungs- sowie Fachinteressenniveau und -streuung sowie Indikatoren für das Klassenklima untersucht. Die Befundlage ähnelt allerdings jener der eben beschriebenen Lehrereigenschaften. Obwohl auch hier verschiedene Leistungsbereiche sowie das Fachinteresse in Deutsch und Mathematik zu mehreren Messzeitpunkten in die Analysen einbezogen wurden, zeigten sich kaum signifikante Zusammenhänge zur Urteilsgüte in der jeweils erwarteten Richtung. Gar keine Effekte - für keinen Bereich zu keinem Messzeitpunkt - ließen sich für das Leistungs- und Fachinteressenniveau, das Klassenklima sowie das Ausmaß an Unterrichtsstörungen und Zeitverschwendung finden. Diese Merkmale hängen offensichtlich in keiner Weise mit der Urteilsgenauigkeit der Lehrer zusammen.

Für die Klassengröße, die als Indikator für quantitative Arbeitsanforderungen an die Lehrkraft gelten kann, erwies sich einzig die Güte der Fachinteresseneinschätzung für das Fach Deutsch als umso genauer, je mehr Schüler in der Klasse unterrichtet wurden. Plausibel wäre eine höhere Urteilsgüte bei kleineren Klassen gewesen. Da der signifikante Zusammenhang aber auch durch den zweiten Messzeitpunkt nicht bestätigt wird und sich darüber hinaus keine ähnlichen Korrelationen zeigten, sollte dieses Ergebnis nicht überbewertet werden.

Während bei der Betrachtung der Klassengröße vermutet wurde, dass es Lehrern mit zunehmender Schülerzahl schwerer fällt, jeden Einzelnen exakt einzuschätzen, stand hinter der Analyse zur Anzahl einzuschätzender Schüler in der Befragung die Überlegung, dass die Belastung beim Ausfüllen der Fragebögen umso größer (und damit möglicherweise ungenauer) wird, je mehr Schüler zu beurteilen sind. In den Ergebnissen zeigt sich jedoch, dass diese Hypothese lediglich für den Bereich Rechtschreibung bestätigt wird ($r = -.21$). Da die Rechtschreibleistung nur zum dritten Messzeitpunkt erfasst wurde, ist eine Replikation an einem anderen Zeitpunkt nicht möglich. Eine Replikation hätte Aufschluss darüber geben können, ob dies spezifisch für die Rechtschreibeinschätzung ist. Theoretisch lässt sich nicht leicht erklären, warum gerade die Einschätzung der Rechtschreibleistung einen negativen Zusammenhang zur Anzahl einzuschätzender Schüler haben soll, während dies für die anderen betrachteten Bereiche nicht zutrifft. Wahrscheinlicher ist deshalb auch hier ein Zufallsergebnis.

Weiterhin wurde der Anteil der Migranten in der Klasse betrachtet. Vermutet wurde, dass ein hoher Migrantenanteil in der Klasse auch die Urteilsgenauigkeit der Lehrkräfte negativ beeinflusst, indem sich Lehrer nicht so sehr auf ihre eigentlichen Kernaufgaben konzentrieren können, sondern statt dessen in hohem Maße kompensatorisch tätig werden müssen, um die ohnehin benachteiligten Kinder mit Migrationshintergrund nicht noch weiter abrutschen zu lassen. Ein Gegenargument zu dieser Hypothese war, dass gerade dadurch, dass viele Migrantenkinder vergleichsweise schlechtere Leistungen erbringen, ihre Leistungseinschätzung im Grunde sehr eindeutig ausfällt, was zu einer höheren Urteilsgenauigkeit bei ihren Lehrern führt. Tatsächlich zeigen die Ergebnisse, dass vor allem in den sprachbezogenen Leistungsbereichen (Wortschatz, Textverstehen und Rechtschreiben) positive Korrelationen zur Urteilsgüte auftreten, von denen aber nur eine (Wortschatz zu t3) das Signifikanzniveau erreicht ($r = .27$). Die Zusammenhänge im Bereich Arithmetik und den Fachinteressen pendeln hingegen deutlich

um Null. Da sich Leistungsdefizite von Migranten besonders bei sprachbezogenen Leistungen zeigen (vgl. u.a. Lehmann et al., 1997; Tiedemann & Billmann-Mahecha, 2004), scheint sich somit tendenziell die Gegenhypothese zu bestätigen, nach der ein höherer Migrantenanteil in der Klasse - zumindest eben für die sprachlichen Leistungsbereiche - zu einer akkurateren Rangurteilsgenauigkeit bei den Lehrern führt.

Ein letzter der untersuchten Aspekte auf Klassenebene scheint allerdings einen erheblichen Einfluss auf die Urteilsgenauigkeit zu haben, nämlich die Streuung der verschiedenen Leistungen. Sie erwies sich für ausnahmslos alle untersuchten Leistungsbereiche als signifikanter Zusammenhangsfaktor zur Güte der diagnostischen Urteile (r = .20 bis .40). Für die Fachinteressen traf dies nur bedingt zu, denn hier konnte lediglich zum dritten Messzeitpunkt für das Fachinteresse Deutsch ein statistisch bedeutsamer Zusammenhang gefunden werden (r = .28), während sich für die anderen Messzeitpunkte und Interessen nur schwach positive Werte zeigten. Zu vermuten ist allerdings, dass die beschriebenen hohen Effekte weniger auf eine tatsächlich höhere diagnostische Kompetenz von Lehrern in heterogenen Klassen hindeuten, als vielmehr darauf, dass in leistungsgemischten Klassen Urteile mit einer höheren Wahrscheinlichkeit zutreffender ausfallen. In Klassen, in denen das Leistungsspektrum sehr einheitlich ist, fällt es grundsätzlich schwerer, diese Leistungen mit der zur Verfügung stehenden 5-stufigen Einschätzskala treffend zu differenzieren. Hinzu kommt, dass schon kleine minimale Beeinflussungen der individuellen Tagesform der Schüler zu einer veränderten Rangfolge der Testergebnisse führen können. Dies ist in Klassen mit breit streuenden Leistungen weniger wahrscheinlich, so dass deren Lehrer bei der Beurteilung quasi einen „natürlichen“ Vorteil gegenüber ihren Kollegen in leistungshomogenen Klassen haben.

8.1.2.3 Schülermerkmale

Während in dieser Arbeit der Einfluss von Lehrer- und Klassenmerkmalen der Urteilsgenauigkeit im Sinne der Rangkomponente gegenübergestellt und somit jeweils zwei Merkmale auf Klassenebene miteinander verglichen wurden, war das Vorgehen für die Überprüfung des Einflusses von Schülermerkmalen etwas anders. Diese liegen auf individueller Ebene vor, so dass sie auch den individuellen Lehrerurteilen gegenübergestellt werden. Dafür wurden die Schülerleistungen an die Metrik der Lehrerurteile angeglichen und anschließend die Differenz zwischen beiden berechnet. Die Art der Erhebung (globale Urteile) und die Transformation erlauben keine In-

terpretation der absoluten Differenzen, differentielle Unterschiede lassen aber sehr wohl Rückschlüsse zu. Als individuelle Merkmale sind das Geschlecht, das Leistungsniveau, der Sozialstatus sowie das Gefühl des Angenommenseins untersucht worden.

Individuelle Merkmale der Schüler scheinen insgesamt einen deutlich stärkeren Zusammenhang zu individuellen Einschätzungen zu haben als aggregierte Klassenmaße zu Lehrer- oder Klasseneigenschaften. Obwohl sich die Testleistungen nicht immer als geschlechtsabhängig erweisen und Leistungsunterschiede je nach Geschlecht im Bereich Wortschatz beispielsweise deutlich geringer ausgeprägt (oder gar nicht vorhanden) sind als im Bereich Rechtschreiben, drücken sich in den Lehrerurteilen für alle Bereiche Annahmen über Leistungsunterschiede aus. Entsprechend der aufgestellten Hypothese folgen diese durchweg den oft anzutreffenden Stereotypen, dass Jungen den Mädchen im mathematischen Bereich überlegen seien, wohingegen Mädchen in den sprachbezogenen Leistungsbereichen besser abschnitten als Jungen. Die in der Untersuchung gezeigten Leistungen entsprechen dem aber nur teilweise. In Arithmetik wandelt sich ein Leistungsvorsprung der Jungen am Ende der dritten Klasse über einen Nullzusammenhang Mitte der vierten zu einem Vorteil der Mädchen am Ende der vierten Klasse. Und auch was den Wortschatz der Kinder anbelangt, haben zu zwei Messzeitpunkten die Jungen leichte Vorteile gegenüber den Mädchen.

Entsprechend zeigt sich, dass die Urteilsgenauigkeit je nach Schülergeschlecht variiert, und zwar insbesondere in jenen Leistungsbereichen, in denen die tatsächlichen Schülerleistungen sich deutlich von den Lehrereinschätzungen unterscheiden (Wortschatz und Arithmetik). Bezieht man außerdem auch das Geschlecht der Lehrer mit ein, fällt auf, dass sich weibliche Lehrkräfte in allen Bereichen (wenn auch nicht zu allen Messzeitpunkten) durch höhere Differenzwerte auf der Niveaukomponente von ihren männlichen Kollegen unterscheiden. Dabei scheint es so, als würden Lehrerinnen etwas ‚wohlwollender' urteilen bzw. Leistungen tendenziell eher überschätzen (bzw. zumindest positiver einschätzen). Eine Wechselwirkung zwischen Lehrer- und Schülergeschlecht existiert hingegen nicht, was entsprechende Annahmen, nach denen bspw. Lehrerinnen Mädchen positiver beurteilen als Jungen, für keinen der untersuchten Bereiche bestätigt.

Die Annahme, dass leistungsstärkere Schüler für noch stärker und leistungsschwächere Schüler für noch schwächer gehalten werden, als sie laut Leistungstest tatsächlich sind, konnte in dieser Arbeit nicht bestätigt werden. Stattdessen verhält es sich genau andersherum, und dies bereichs- und

messzeitpunktübergreifend. Dieser Effekt ist im Arithmetikbereich besonders ausgeprägt. Ein ähnlicher Befund trat auch in der Studie von Bates und Nettelbeck (2001) für die dort untersuchten Leistungsbereiche Lesegenauigkeit und Leseverstehen zutage. Auch in ihrer australischen Stichprobe wurden leistungsschwache Schüler überschätzt, leistungsstarke hingegen sogar leicht unterschätzt. Die Ursache für diesen negativen Zusammenhang ist sehr wahrscheinlich technischer Natur. Weil nur eine begrenzte fünfstufige Einschätzskala zur Verfügung stand, war der Spielraum für die Lehrer für eine Überschätzung sehr guter und für eine Unterschätzung sehr schwacher Schüler im Grunde nicht gegeben. Je leistungsstärker ein Schüler ist, desto weniger Möglichkeiten zur Überschätzung bieten sich dem Lehrer überhaupt, und vice versa für leistungsschwache Schüler. Dafür wäre beispielsweise das Schätzen der konkreten erreichten Punktzahl im Test geeigneter gewesen. So aber besteht annähernd ein linearer Zusammenhang zwischen der Urteilsabweichung und der Leistung: je schwächer der Schüler ist, desto höher ist die Wahrscheinlichkeit für eine Leistungsüberschätzung. Beruhigend ist zumindest die Erkenntnis, dass der beschriebene Effekt im Wesentlichen auf Schüler mit Leistungsextremen zurückgeht und das Gros der Schüler korrekt beurteilt wird.

Einen deutlichen Einfluss hat den Analysen zu Folge der Sozialstatus der Schüler auf die Urteilsgenauigkeit der Lehrer. Vor allem in den sprachbezogenen Leistungsbereichen zeigen sich signifikante Effekte derart, dass Schüler, deren Elternhaus der oberen Hälfte der Sozialverteilung zuzuordnen sind, durchweg für besser gehalten (d.h., stärker überschätzt bzw. weniger stark unterschätzt) werden als Schüler aus Familien mit einem HISEI unterhalb des Medians. Am geringsten ist dieser Effekt im Bereich Arithmetik, wo er zu t3 sogar das Signifikanzniveau verfehlt. Auch korrelative Analysen bestätigen die genannten Ergebnisse der t-Tests, und unter Kontrolle der Leistung, die selbst auch mit dem Sozialstatus korreliert, zeigten sich sogar noch höhere Zusammenhänge.

Diese Art von Fehlbeurteilungen birgt insofern eine besondere Bedeutung in sich, als dass die im Durchschnitt ohnehin leistungsschwächeren Kinder aus sozial schwachen Familien (u.a. Ditton & Krüsken, 2009) eine zusätzliche Benachteiligung durch die systematische Schlechtereinschätzung im Vergleich zu gut situierten Schülern erfahren. Statt Schüler mit Nachteilen, die sich aus ihrer familiären Herkunft ergeben, besonders zu fördern, werden Ausgangsunterschiede so eher zementiert oder sogar vergrößert.

Eine weitere Hypothese war, dass die in dieser Arbeit nicht explizit untersuchte Sympathie oder Antipathie den Schülern gegenüber einen Einfluss auf die Genauigkeit und Fairness ihrer Beurteilungen hat. So wurde in einige Studien ein Effekt der Einstellung der Lehrer zu den Schülern vermutet. Einen Hinweis darauf fanden Itskowitz, Navon und Strauss (1988), in deren Untersuchung Lehrkräfte die Selbstwahrnehmung von Schülern ihrer Klasse einschätzen sollten. Wie sich zeigte, variierte die Ungenauigkeit ihrer Einschätzungen als eine Funktion der selbst wahrgenommenen Nähe zu den Schülern. Verglichen mit den Selbsteinschätzungen der Schüler tendierten die Lehrer meist zu einer positiveren Darstellung (Überschätzung) jener Schüler, denen sie sich nahe fühlten, und zu einer Unterschätzung jener Schüler, denen sie neutral oder ablehnend gegenüberstanden.

Als relativ distaler Indikator dafür, wie hoch die Sympathie zwischen Schülern und ihren Lehrern ausgeprägt ist, wurde in der vorliegenden Arbeit das Gefühl des Angenommenseins auf Schülerseite erfasst. Vermutet wurde, dass Schüler ihren Lehrern umso sympathischer sind, je mehr sie sich auch von ihnen gerecht behandelt, gemocht und unterstützt fühlen, und dass dies Rückschlüsse darauf zulässt, wie freundlich die Lehrkraft ihnen gegenüber tatsächlich auftritt. Auch wenn nur ein kleiner Teil der Schüler angab, sich nicht gut angenommen zu fühlen, ergeben sich für die meisten Leistungsbereiche signifikante Zusammenhänge zum Lehrerurteil. Besonders deutlich zeigt sich dies für den Bereich Wortschatz, in dem Schüler, die sich (eher) angenommen fühlen, auch deutlich besser beurteilt werden als Schüler, die sich (eher) nicht angenommen fühlen. Gleiches zeigt sich zwar auch für Arithmetik zum ersten Messzeitpunkt, hier geht dieser Zusammenhang jedoch zum zweiten Messzeitpunkt zurück und kehrt sich kurz vor dem Übergang in die Sekundarstufe sogar um.

Unabhängig von den tatsächlichen Leistungen der Schüler erweist sich somit auch dieser emotionale Aspekt, der das Fühlen und Erleben der Schüler und ihr Verhältnis zum Lehrer zum Gegenstand hat, als bedeutsame Variable bei der Erklärung der Urteilsgenauigkeit. Unklar bleibt hier freilich, ob Lehrer die Gefühle der Schüler bestätigen können bzw. ob sich die Schülergefühle tatsächlich aus der Verhaltensweise der Lehrer ableiten lassen oder nicht vielleicht eher aus anderen Aspekten wie der Einstellung zu den eigenen Unterrichtsleistungen. Nicht auszuschließen ist, dass Schüler eigene (nicht zufrieden stellende) Leistungen auf ihre Lehrer statt auf die eigene Leistungsfähigkeit attribuieren.

8.1.3 Vergleichbarkeit von Leistungseinschätzungen und Zeugnisnoten

Die Indikatoren, aus denen auf die diagnostischen Fähigkeiten der Lehrer geschlossen wird, beruhen überwiegend auf Einschätzfragen aus entsprechenden Fragebogeninstrumenten. Auch wenn davon ausgegangen wird, dass alle Lehrer diese Fragen gewissenhaft beantwortet haben, so ist deren unterrichtspraktische Relevanz doch nicht mit jener von alltäglich vergebenen Zensuren zu vergleichen: Mit dem Ausfüllen des Fragebogens tun die Lehrer Forschern einen Gefallen, während sie mit der Notenvergabe verlässliche Leistungsurteile fällen sollen. Daher war eine weitere Fragestellung dieser Arbeit, inwiefern Noten und Leistungseinschätzungen miteinander zusammenhängen. Die Ergebnisse bestätigen frühere Befunde (z. B. Schrader & Helmke, 1990), nach denen die Testleistungen mit den Einschätzungen genauso hoch zusammenhängen wie mit den Noten. In der vorliegenden Arbeit wurde als Referenz die dem jeweiligen Leistungsbereich zuzuordnende letzte Zeugnisnote zugrunde gelegt. Unabhängig vom Leistungsbereich waren praktisch keine Unterschiede in der Höhe der Korrelationen zwischen Noten und Leistung bzw. Einschätzung und Leistung zu finden, alle Korrelationen fielen signifikant aus, lagen jedoch mit Werten zwischen r = .43 und .63 und einiges niedriger als die von Schrader und Helmke berichteten Werte. Bemerkenswert ist allerdings, dass die Zusammenhänge zwischen den von den Lehrern vergebenen Noten und ihren Einschätzungen in den Fragebögen um einiges höher lagen (r = -.67 bis .73). Dies lässt zum einen darauf schließen, dass Lehrer sich bei ihren Einschätzungen möglicherweise in vielen Fällen direkt an den Noten orientieren, zumal im selben Fragebogen die Noten direkt erfragt wurden und somit den Lehrern besonders präsent waren. Zum anderen bedeutet es, dass sowohl Einschätzungen als auch Zeugnisnoten eben einen vergleichsweise geringen Zusammenhang zu den im Test gemessenen Leistungen der Schüler aufweisen.

Die weiteren Analysen zeigten, dass Schüler trotz gleicher Ergebnisse im Leistungstest sehr unterschiedliche Leistungseinschätzungen und Zeugnisnoten von ihren Lehrern erteilt bekamen. Gerade im mittleren Leistungsbereich ist nahezu das gesamte Notenspektrum vertreten, so dass es - über die Gesamtstichprobe hinweg - faktisch unmöglich ist, von der Testleistung auf die Zeugnisnote im selben Bereich zu schließen. Eine mögliche Ursache hierfür ist, dass sich Lehrer bei ihren Urteilen für gewöhnlich am Maßstab der eigenen Klasse orientieren (vgl. 3.3 ab S. 39). Somit kann ein relativ leis-

tungsschwacher Schüler in einer insgesamt leistungsstarken Klasse mit hoher Wahrscheinlichkeit schlechter eingeschätzt werden als ein objektiv gleich starker Schüler, der in einer insgesamt leistungsschwachen Klasse mit dieser Leistung zu den Besten gehört und entsprechend gut bewertet wird. Für alle Leistungsbereiche zeigte sich, dass das tatsächliche Leistungsspektrum der als besonders leistungsschwach beurteilten Schüler deutlich größer war als das Leistungsspektrum der als sehr leistungsstark angesehenen Schüler. Somit fallen Lehrerurteile bei leistungsstarken Schülern genauer aus als bei schwachen Schülern, bzw. wird eine Zuordnung der Schüler zur Gruppe der Leistungsschwachen - ggf. je nach Klassenkontext - für eine größere Bandbreite an Leistungen vorgenommen als eine Zuordnung zur Gruppe der Leistungsstarken.

Dieses Phänomen der unterschiedlichen Bewertung objektiv gleicher Leistungen deckt sich mit Befunden einer Reihe anderer Untersuchungen (z.B. Baumert et al., 2003; Ingenkamp, 1989; Ziegenspeck, 1999). Es tritt vorwiegend dann auf, wenn Leistungen nicht an einem absoluten Maßstab gemessen werden, der transparent und über die eigene Klasse hinaus vergleichbar ist, sondern zum Beispiel an Leistungen anderer Schüler (sozialer Vergleich), indem etwa eine Normalverteilung der Leistungen angenommen wird. Dies ist in der Unterrichtsrealität fast immer der Fall, da Leistungsüberprüfungen nicht klassen-, schul- oder gar landesübergreifend standardisiert vorgegeben, sondern in aller Regel von den Lehrern selbst entworfen werden. Lehrpläne und Lehr- und Lernziele wären eine Möglichkeit, die erreichten Leistungen objektiv zu beurteilen. Da Lehrer bei der Umsetzung der Lehrpläne aber gewisse Freiheiten in puncto Reihenfolge und Tempo der Stoffvermittlung genießen, ist auch ein derartiger Vergleich nur sehr schwer umzusetzen. Insofern kann man den Lehrern die Nicht-Vergleichbarkeit von Bewertungen objektiv gleicher Leistungen nicht zum Vorwurf machen. Lehrer orientieren sich an ihren Erfahrungen, die sie nur in ihrem gewöhnlichen Arbeitsumfeld, also anhand der Schüler ihrer Schule, machen können. Daran machen sie ihre Entscheidungen fest, wie einzelne Leistungen zu bewerten sind. In einigen Ländern wird daher versucht, diesem Problem durch den Einsatz klassenübergreifender Leistungsüberprüfungen, die sich eng an den Lehrplänen orientieren, zu begegnen (vgl. Schrader & Helmke, 2001). Allerdings ist auch dieses Vorgehen mit Nachteilen behaftet, da dabei alle Schüler gleichzeitig dieselben Tests bearbeiten müssten und Einflüsse von unterschiedlicher Tagesform, Abwesenheit am Testtag o.ä. die angestrebte bessere Vergleichbarkeit wiederum reduzieren. Ein Kompromiss wäre deshalb eine Kombination aus herkömmlicher Leis-

tungsbewertung und standardisierten diagnostischen Verfahren. Letzteres steht bislang jedoch nicht in ausreichendem Umfang zur Verfügung.

8.1.4 Ergänzende Betrachtungen der Ergebnisse anhand der Zusatzstichprobe

Mittels einer zusätzlichen Stichprobe aus Erstklässlern, die dem sogenannten Längsschnitt 1 des BiKS-Projekts entspricht, sollten zum einen die Befunde zur Güte diagnostischer Urteile aus der Hauptstichprobe überprüft und erweitert, und zum anderen die Transformation der Rangurteile in die Niveaukomponente aus der Hauptstichprobe überprüft werden. Die Befunde aus der dritten und vierten Klasse konnten dabei im Wesentlichen bestätigt werden, allerdings fiel die Rangkomponente für den Bereich Wortschatz - im Gegensatz zu den Bereichen Arithmetik und Textverstehen - deutlich geringer aus. Aufgrund aufgabenspezifischer Einschätzungen ließen sich in der Zusatzstichprobe auch weitere Urteilskomponenten bilden, deren so gemessene Güte in Teilen mit Ergebnissen anderer Untersuchungen übereinstimmte. So zeigte sich auch hier, dass die auf individuellen Aufgabeneinschätzungen beruhende Rangkomponente vergleichbar zur Rangkomponente auf Konstruktebene ausfällt (z.B. Demaray & Elliott, 1998; Feinberg & Shapiro, 2003), allerdings nur für die Wortschatzeinschätzung, während im Textverstehen eine minimal höhere Urteilsgüte für die globale Einschätzung bestehen bleibt. Letzterer Befund deckt sich mit ebenfalls aus dem BiKS-Projekt stammenden Ergebnissen aus der fünften Klassenstufe im Bereich Lesen (Karing, Matthäi & Artelt, in Druck). Weiterhin wurde auch in der vorliegenden Studie in der ersten Klasse analog zu früheren Untersuchungen im Grundschulbereich (u.a. Hamilton & Shinn, 2003) und im frühen Sekundarstufenbereich (u.a. bei Schrader & Helmke, 1987) das Leistungsniveau der Schüler entsprechend der aufgestellten Hypothese überschätzt, und zwar für das Textverstehen deutlich stärker als für den Wortschatz. Der bislang nur selten in anderen Untersuchungen einbezogene aufgabenspezifische Treffer liegt wie bei Karing und Kollegen (in Druck) ebenfalls im mittelhohen Bereich.

Die teilweise aufgetretenen bereichsspezifischen Unterschiede lassen sich nur bedingt mit Spezifika der jeweiligen Leistungen erklären. Am Ende des ersten Schuljahres, in dem das Lesenlernen einen wesentlichen Bestandteil des Unterrichts darstellt, sollte insbesondere die Einschätzung des Textverstehens (bzw. der Lesefähigkeit) leicht fallen, da diesem im Unterricht erhöhte Aufmerksamkeit durch die Lehrer gewidmet wird. Darüber hinaus

sollte aber auch der Wortschatz der Kinder zum Beispiel gut aus ihrer Ausdrucksfähigkeit abgeleitet werden können. Eine plausible Erklärung dafür, dass die Rangeinschätzung beim Wortschatz ebenso wie die Niveaueinschätzung beim Textverstehen besonders schwer gefallen sein könnte, liegt nicht auf der Hand. Vielmehr bestätigt sich auch hierbei, dass die diagnostischen Fähigkeiten der Lehrer bereichsspezifisch ausfallen, auch wenn dies hier nur anhand zweier im Grunde vergleichbarer Leistungsbereiche exemplifiziert werden konnte, und es scheint außerdem die Referenz des Urteils (im Sinne der Urteilskomponente) von Bedeutung zu sein. Ob Leistungseinschätzungen im Vergleich zu anderen Schülern oder im Vergleich zu einem festgelegten Maßstab akkurat ausfallen, sind offenbar verschiedene Aspekte des Urteils.

Anhand der Zusatzstichprobe sollte außerdem die Zulässigkeit der durchgeführten Überführung von Rangurteilen in die Niveaukomponente überprüft werden. Transformierte und original erfasste Niveaukomponente korrelierten dabei zwar signifikant, aber nur in moderater Höhe für die Bereiche Wortschatz und Textverstehen. Dies scheint auf den ersten Blick ein gewisser Beleg für die Zulässigkeit der durchgeführten Transformation zu sein. Anhand einer Beispielrechnung mit beiden Varianten der Niveaukomponente sollte überprüft werden, inwiefern dabei vergleichbare Ergebnisse entstehen. Dabei ging es um die Frage, inwiefern die individuellen Urteile vom Geschlecht und vom Sozialstatus der Schüler beeinflusst werden. In der Hauptstichprobe hatten sich beide Merkmale als bedeutsame Moderatoren erwiesen. Dies war für die Zusatzstichprobe nur noch bedingt der Fall. Hatte das Geschlecht der Schüler in der Hauptstichprobe bei globaler Einschätzung noch einen signifikanten Einfluss auf die Urteilsgenauigkeit, so bestätigte sich dies auch in der Zusatzstichprobe für die globalen, aber nicht mehr für die spezifischen Urteile. Gleiches trifft für den Einfluss des Sozialstatus zu, wobei hier zusätzlich festzuhalten ist, dass dieser sich im Bereich Textverstehen - im Gegensatz zur Hauptstichprobe - auch nicht mehr auf die globalen Urteile auswirkt.

Ist dieser Befund ein Beleg dafür, dass die Transformation globaler Urteile zu verzerrten Ergebnissen führt? Auch wenn die teils unterstützenden, teils widersprüchlichen Ergebnisse zumindest Zweifel an der Zulässigkeit der in der Arbeit vorgenommenen Transformation aufkommen lassen, sollten Unterschiede der beiden Stichproben bei der Interpretation berücksichtigt werden. Abgesehen von deutlich verschiedenen Stichprobengrößen und Beteiligungsquoten handelt es sich bei der Zusatzstichprobe um Erstklässler, de-

ren zu bewertende Leistungsfacetten gerade erst in basaler Form erlernt wurden, während sie gegen Ende der Grundschulzeit schon gefestigter sind. Womöglich haben Lehrer in der ersten Klasse ganz andere Vorstellungen von den einzuschätzenden Bereichen als Lehrer in der dritten oder vierten Klasse, gerade wenn sie dabei an den zurückliegenden Unterricht denken. Darüber hinaus sollte der Umstand, dass die Zusatzstichprobe einen deutlich höheren Stichprobenausfall aufweist, gerade vor dem Hintergrund der Befunde zum Einfluss der Gruppenzusammensetzung auf die Urteilsgenauigkeit (vgl. Kapitel 7.1.5) nicht vernachlässigt werden. Dass globale Urteile überwiegend zu signifikanten Gruppenunterschieden führen, spezifische aber nicht, kann möglicherweise auch auf eine andere Ursache zurückzuführen sein, die die Eignung dieses Vergleichs zwar etwas einschränkt, aber nicht in Frage stellt. Wie Daten aus weiteren Erhebungen im Rahmen des BiKS-Projekts aus der Sekundarstufe vermuten lassen, macht es für die Genauigkeit der Lehrerurteile insgesamt (vgl. bspw. auch Feinberg & Shapiro, 2003) und entsprechend eben auch für differentielle Effekte einen Unterschied, ob Urteile global oder spezifisch gefällt werden (Lorenz & Artelt, 2010). Während Lehrer bei aufgabenspezifischen Einschätzungen genau abwägen müssen, ob einzelne Schüler die vorgegebenen Items korrekt lösen können und dazu eine exakte Abschätzung von Fähigkeitseinschätzungen, Wissen um den bereits im Unterricht behandelten Stoff etc. vonnöten ist, spielt bei globalen Urteilen die generelle Meinung, die anfälliger für Vorurteile und Stereotypisierungen ist, eventuell eine größere Rolle. Diese Annahme sollte auch bei der Interpretation der Ergebnisse aus der Hauptstichprobe, die ja gänzlich auf globalen Urteilen beruht, berücksichtigt werden.

Insgesamt ergibt sich für die Differenzierungen und die Vergleichbarkeit der verschiedenen Urteilskomponenten somit eine Befundlage, aus der sich aufgrund vieler Einschränkungen keine klaren Erkenntnisse ableiten lassen. Weitere Forschung wird nötig sein, um die teils widersprüchlichen Ergebnisse separat für verschiedene Klassenstufen, Schularten und -formen sowie Inhaltsbereiche zu prüfen.

8.2 Vorteile und Einschränkungen der vorliegenden Untersuchung

Vorteile

Die vorliegende Arbeit zeichnet sich im Vergleich zu anderen Untersuchungen, die sich der diagnostischen Kompetenz von Lehrkräften widmeten,

durch einige spezifische Vorteile aus, die in der Summe zu einer besonders hervorzuhebenden Datenbasis geführt haben. Dazu zählt zunächst, dass es sich nicht nur um eine querschnittlich angelegte Studie handelt, sondern drei Messzeitpunkte für die Analysen zur Verfügung stehen, die im halbjährlichen Abstand das Ende der Grundschulzeit abdecken. Dass es sich dabei um den Zeitraum direkt vor der wegweisenden Übergangsempfehlung handelt, weckt zudem berechtigte Hoffnung, dass die involvierten Lehrer ihre Urteile gewissenhaft fällten, da sie wahrscheinlich ohnehin in dieser Phase besonders gründlich die Schülerleistungen beobachten. Durch die wiederholten Messungen sind zudem in dieser Arbeit Aussagen über die Stabilität der Urteilsgüte möglich, was bisher in der Forschung noch gar nicht systematisch untersucht wurde. Die bislang genannten Vorteile gewinnen weiterhin besonders dadurch an Bedeutung, dass es sich einerseits mit knapp 2400 beteiligten Kindern aus über 150 Schulklassen einerseits um eine vergleichsweise große Stichprobe handelt, die zu untersuchen nur im Rahmen eines großen Forschungsprojekts wie BiKS möglich ist. Andererseits ist die große Bandbreite der erfassten Leistungsbereiche und emotional-motivationalen Aspekte nahezu einzigartig und eröffnet somit mannigfaltige Vergleichsmöglichkeiten. Die Frage danach, ob es sich bei diagnostischer Kompetenz um ein homogenes Fähigkeitskonstrukt handelt bzw. zwischen welchen Urteilsbereichen besonders hohe Übereinstimmungen hinsichtlich der Urteilsgüte vorliegen, ist so überhaupt erst auf belastbarer Basis möglich. Zu einigen berücksichtigten Bereichen wie der Rechtschreibleistung oder dem fachspezifischen Interesse der Schüler lagen bislang gar keine neueren Erkenntnisse in Bezug auf die Akkuratheit der jeweiligen Urteile vor, und emotional-motivationale Schülermerkmale standen ebenfalls äußerst selten im Fokus der Forschung zu Lehrerurteilen. Was die Frage nach den Bedingungen diagnostischer Kompetenz anbelangt, ist in dieser Arbeit ebenfalls eine bisher nicht dagewesene Bandbreite an möglichen Faktoren untersucht worden, die weit über die oft singulären betrachteten Variablen hinausgehen und sich auf mehrere Ebenen (Lehrer, Klasse, Schüler) erstreckten. Und nicht zuletzt liegt ein Vorteil darin, dass für die Überprüfung und Ergänzung der Befunde eine weitere Stichprobe innerhalb desselben Projekts zur Verfügung stand, die sich hinsichtlich Klassenstufe sowie der untersuchten Lehrer von der Hauptstichprobe unterschied, aber in anderen Punkten (z.B. Erhebungsregion, Untersuchungsinstrumente) vergleichbar war. Dies alles sind deutlich hervorzuhebende Vorteile dieser Untersuchung, die Antworten auf bisher vernachlässigte Fragen ermöglichten und als Besonderheit zu bezeichnen sind.

Einschränkungen

Wie jede andere Untersuchung weist aber auch diese ebenfalls Nachteile bzw. Einschränkungen auf, die sich auf die Verallgemeinerbarkeit der Ergebnisse auswirken. Ein zentrales Problem ist dabei sicher die teils geringe Teilnahmequote der Schüler an der Untersuchung, insbesondere in der Zusatzstichprobe. Da die Eltern der Beteiligung ihrer Kinder zustimmen mussten, entstand über die Messzeitpunkte hinweg ein immer größerer Ausfall, der in der Hauptstichprobe zur dritten Erhebung gut vierzig Prozent ausmachte. In der Zusatzstichprobe nahmen gar nur zwanzig Prozent der Schüler aus den untersuchten Klassen teil. Dies führte insofern zu einer Verzerrung der Stichprobe, als dass diese Kinder nicht zufällig ausfielen, sondern überdurchschnittlich häufig leistungsschwächere Schüler davon betroffen waren, die somit leicht unterrepräsentiert sind. Optimal wären Vollerhebungen in den Klassen gewesen, gerade für die Untersuchung der diagnostischen Fähigkeiten und vor dem Hintergrund, dass fehlende Schüler einen bedeutsamen Einfluss auf die Rangurteilsgüte haben können. Dies ist jedoch bei Untersuchungen mit freiwilliger Beteiligung nicht zu erreichen. Eine generelle Einschränkung besteht insofern, als dass die Selbsteinschätzungen der Kinder zu ihren emotionalen oder motivationalen Eigenschaften im Fragebogen im Unterschied zu den per Test gemessenen Leistungen nicht objektiv sein müssen. Gerade bei Grundschülern muss man davon ausgehen, dass eine realistische Selbstbeurteilung noch nicht in jedem Fall gelingt. Entsprechend sind die niedrigeren Urteilsgenauigkeiten der Lehrer in diesen Bereichen nicht unbedingt ein Hinweis auf schlechtere diagnostische Fähigkeiten, sondern möglicherweise basieren die höheren Differenzen zwischen Schüler- und Lehrerwerten hier einfach auf weniger validen Selbstauskünften. Natürlich können darüber hinaus auch die Testleistungen der Schüler durch verschiedene denkbare Einflüsse wie niedrige Motivation, schlechte Konzentration o.ä. von ihren sonstigen Leistungen abweichen und die Urteilsgenauigkeit damit verringern. Ein besseres, gleichermaßen ökonomisches Untersuchungsdesign, das viele Leistungsfacetten mit möglichst wenig Testaufwand erhebt, konnte im Rahmen der Studie jedoch nicht umgesetzt werden. Aus dem gleichen Grund konnten auch die Einschätzitems für die Lehrer nicht umfassender gestaltet werden, um den ohnehin immensen Aufwand für die Lehrer durch die vielfältigen individuellen Einschätzungen nicht noch weiter zu steigern. Dies hätte sehr wahrscheinlich zu einer Teilnahmeverweigerung der Lehrer geführt.

Was den Einsatz von standardisierten Testverfahren zur Erfassung der Schülerleistungen betrifft, gibt es einen weiteren, möglicherweise einschränken-

den Aspekt zu berücksichtigen. In der Forschung zur diagnostischen Kompetenz besteht im allgemeinen Konsens darüber, dass (meist eben mit standardisierten Tests gemessene) Schülerleistungen das Kriterium sind, an dem die Lehrerurteile gemessen werden. In der Regel geht man davon aus, dass die Testergebnisse valide sind und die Lehrereinschätzungen Fehlern unterliegen, die es zu quantifizieren gilt. Wahrscheinlich ist allerdings auch, dass die Leistungserfassung auf Schülerseite bestimmten Messfehlern unterliegt, die berücksichtigt werden sollten, und einige Autoren schlugen genau dies vor (Gerber & Semmel, 1984; Gresham, Reschly & Carey, 1987). Testverfahren erfüllen allerdings üblicherweise die Testgütekriterien, so dass an ihrer Aussagekraft weniger Zweifel bestehen als an den Lehrerurteilen. Um die Kompetenzen der Schüler zu erfassen und somit einen Bezugswert für die Lehrerurteile zur Verfügung zu haben, wurden in dieser Arbeit standardisierte Schulleistungstests eingesetzt, die zum Großteil etablierte Instrumente darstellen. Die Güte von Lehrerurteilen anhand der Übereinstimmung zu mit standardisierten Tests gemessenen Schulleistungen zu bestimmen, ist aufgrund nicht direkt gegebenen Praxisnähe (vgl. Wahl et al., 1997) nicht unproblematisch. Im Vorfeld wurde überprüft, ob die in den Tests vorkommenden Aufgaben auch mit den Lehrplänen der relevanten Bundesländer Bayern und Hessen übereinstimmen. Dazu gibt es oft bereits in den Testhandbüchern Hinweise, in denen nicht zuletzt ein konkreter Erhebungszeitpunkt empfohlen wird, um sicherzustellen, dass der Stoff auch mit hoher Wahrscheinlichkeit schon in den Klassen behandelt wurde. Allerdings gibt es natürlich auch Testverfahren wie bspw. den Matrizentest zum logisch-abstrakten Denken, bei denen keine Korrespondenz zu Lehrplänen existiert, weil damit eher generelle kognitive Fähigkeiten abgeprüft werden sollen, die kein expliziter Unterrichtsgegenstand sind. Zwei weitere Einschränkungen gibt es darüber hinaus. Zum einen erstreckte sich der Testzeitraum durch die hohe Stichprobengröße je nach Messzeitpunkt auf einen Zeitraum von ca. 10 bis 13 Wochen, wodurch zuletzt getestete Klassen grundsätzlich über einen höheren Wissensbestand bzw. höhere Kompetenzen verfügen konnten als zuerst getestete Klassen. Wie sich zeigte, war dieser Effekt jedoch eher gering (vgl. Tabelle 5, S. 114). Zum anderen konnte durch die häufige Verwendung von Spiralcurricula, bei denen keine exakte Abfolge von Schulstoff vorgegeben ist, sondern einzelne Themen im Laufe der Schuljahre mehrmals, auf jeweils höherem Niveau, wiederkehren und bei denen Lehrern überwiegend freie Hand bei der konkreten Abfolgeplanung gelassen wird, nicht garantiert werden, dass sich alle Klassen zu den Testzeitpunkten gerade an vergleichbaren Stellen im Lehrplan befanden. Beide Einschränkungen ließen sich in zukünftigen Studien zum Beispiel

dadurch reduzieren, dass der Testzeitraum verkürzt wird, also mehr Klassen in kürzerer Zeit (durch mehr Testleiter) getestet werden, und dass durch eine zusätzliche Lehrerbefragung kontrolliert wird, ob für die abgefragten Kompetenzen tatsächlich schon Lerngelegenheiten im Unterricht bestanden.

Die Arbeit weist eine Reihe weiterer methodischer Einschränkungen auf, die in zukünftigen Untersuchungen vermieden werden sollten. Dazu zählt als wichtigste die Tatsache, dass in der Hauptstichprobe ausschließlich globale Einschätzungen abgefragt wurden. Dies führt dazu, dass kein direkter Abgleich zum individuellen Leistungs- oder Merkmalsniveau der Schüler möglich ist und somit die bedeutsame Niveaukomponente diagnostischer Kompetenz nicht direkt berechnet werden kann. Der hier gewählte Umweg über eine Differenzmaßbildung auf Grundlage transformierter Schülerwerte und der Lehrerurteile erwies sich als nicht deckungsgleich mit dem standardmäßigen Vorgehen, wie es in der Zusatzstichprobe angewendet wurde, so dass es allenfalls als ‚Notlösung', aber nicht als adäquater Ersatz angesehen werden kann. Inwiefern die originale Niveaukomponentenberechnung zu anderen Ergebnissen als den hier berichteten geführt hätte, kann nur gemutmaßt, aber nicht sicher bestimmt werden. Ebenfalls unsicher ist, ob auch die Rangurteilsgüte der Lehrer in der Hauptstichprobe anders ausgefallen wäre, hätte man ihr spezifische Einschätzungen zugrunde gelegt. In der Zusatzstichprobe erwies sie sich nur für den Bereich Wortschatz als exakter. Allerdings deuten Ergebnisse aus dem Bereich der Sekundarstufe - im Gegensatz zum Primarbereich - darauf hin, dass globale Lehrerurteile sogar genauer ausfallen können als spezifische (Karing, 2009).

Weiterhin sollten zukünftige Untersuchungen, die sich der Leistungsängstlichkeit widmen, diese und deren Einschätzung fachspezifisch erheben, so wie es andere Forschungsbefunde (z.B. Sparfeldt, Schilling, Rost, Stelzl & Peipert, 2005) nahelegen. Eine Erfassung der allgemeinen, nicht nach Fachbereich differenzierten Leistungsangst, wie in dieser Arbeit geschehen, hat sich als nicht adäquat erwiesen. Als ein weiterer Nachteil hat sich herausgestellt, dass den Lehrern für ihre Einschätzungen kein klar definierter Bezugsrahmen vorgegeben wurde. Die Items nach dem Schema „Er/sie kann gut rechnen" lassen offen, ob der Schüler oder die Schülerin im Vergleich zu ihren Mitschülern, im Vergleich zu einem durchschnittlichen Schüler dieses Alters oder im Vergleich zu anderen Maßstäben ‚gut' ist. Durch diesen Interpretationsspielraum ist es möglich, dass verschiedene Lehrer auch verschiedene Bezugsrahmen zugrundegelegt haben, wodurch die Vergleichbar-

keit ihrer Urteile eingeschränkt würde. Dieses Problem wirkte sich auch auf die Wahl der Transformationsmethode zur Gewinnung von schülerspezifischen Niveauurteilen aus und könnte mit verantwortlich dafür sein, dass die so gewonnenen Daten nur mäßig mit den originalen Niveauurteilen übereinstimmen.

Weitere kleinere Einschränkungen der vorliegenden Arbeit beziehen sich beispielsweise darauf, dass insbesondere zum ersten Messzeitpunkt die Formulierungen der Einschätzitems von jenen der beiden anderen Messzeitpunkte abwichen. Auch dies schränkt die Vergleichbarkeit zwischen den Erhebungszeitpunkten ein und sollte vermieden werden. Im Zusammenhang damit erweist es sich auch - gerade vor dem Hintergrund globaler Beurteilungen - als problematisch, wenn die Testinstrumente selbst im Laufe der Erhebung verändert werden. Dies ist in vielen Fällen aber nur schwer zu verhindern, denn bei sich weiter entwickelnden Fähigkeiten der Schüler kann in einer längsschnittlich angelegten Untersuchung nicht immer derselbe Test, bspw. zur Erfassung der Rechenfähigkeit, eingesetzt werden.

8.3 Fazit

Lehrerurteile sind nicht perfekt. Vermutlich können sie es auch nie sein. Eine zuverlässige Diagnose und Prognose von schulischen Leistungen und Schülereigenschaften ist schwierig und fordert jeden einzelnen Lehrer in den vielfältigsten Unterrichtssituationen heraus. Dabei darf nicht vergessen werden, dass an Lehrer nicht dieselben Maßstäbe angelegt werden dürfen wie an standardisierte diagnostische Verfahren, wie sie beispielsweise in der Intelligenzdiagnostik oder der klinischen Verhaltensdiagnostik zum Einsatz kommen. Weder läuft der Unterricht im Allgemeinen ausreichend standardisiert ab, noch kann man erwarten, dass Lehrer Leistungen rein technisch und ohne Berücksichtigung anderer Faktoren bewerten.

Die vorliegende Arbeit hat neben den vielen zu Anfang der Diskussion zusammengefassten inhaltlichen Befunden vor allem zwei Facetten aufgezeigt, die große Bedeutung für die Unterrichtspraxis, die Lehrerausbildung und das Forschungsgebiet beinhalten. Dies ist einerseits, dass die Akkuratheit von Urteilen zu Leistungen, Interessen, Ängstlichkeit und Motivation anscheinend nicht auf einzelne Lehrermerkmale zurückführbar ist, sondern vor allem stark von Eigenschaften der einzuschätzenden Schüler selbst abzuhängen scheint. Den Lehrern sieht man also nicht an und man kann es auch nicht durch Zulassungs- oder Einstellungstests feststellen, ob sie gute

Diagnostiker sind, sondern diese Eigenschaft tritt vor allem in der individuellen Konstellation mit den Schülern zutage. Dabei ist von Bedeutung, dass sich Lehrer offensichtlich zu stark von allgemeinen Eindrücken, eventuell auch vorhandenen Stereotypen leiten lassen statt ganz objektiv ausschließlich von Leistungen. Wie sich diese Erkenntnis nutzen lässt, um Lehrer und Lehramtsanwärter dahingehend in Ausbildung und Praxis zu unterstützen, wird im Ausblick am Ende dieses Kapitels angerissen.

Darüber hinaus offenbarte diese Arbeit als zweiten wesentlichen Aspekt, dass Forschungsbefunde zur diagnostischen Kompetenz oftmals durch methodische Besonderheiten beeinflusst werden. So ergibt sich die niedrige Reliabilität der Rangkomponente vor allem daraus, dass geringe Abweichungen zwischen Urteilen und Schülermerkmalen gleich zu deutlich niedrigeren Korrelationskoeffizienten führen. Auch die Frage nach der Abhängigkeit der Niveauurteilsgüte vom Leistungsniveau der Schüler ist vor allem durch methodische Restriktionen limitiert, was zu erwartungswidrigen Ergebnissen führt. Man kann zu Recht die praktische Relevanz solcher Ergebnisse anzweifeln. Forschung, die ihren Sinn und Zweck darin sieht, Fragen des praktischen Alltags zu durchleuchten und nützliche Antworten zu präsentieren, muss sich mit ihrer Methodik möglichst nah an der Praxis orientieren.

Ausblick: Verbesserung der diagnostischen Kompetenz von Lehrern

Wie Lüders (2001) anführt, verfügen Lehrer bei der Beurteilung ihrer Schüler über teils erhebliche Freiheiten und Dispositionsspielräume, die sich an drei Punkten festmachen lassen. Erstens besteht für sie keine Verpflichtung und auch keine Rechtfertigungspflicht für wissenschaftlich exakte Diagnostik, was sich auch darin ausdrückt, dass Berufsanfänger Grundkenntnisse und Fähigkeiten dafür erst im Berufsleben und nicht bereits während der Ausbildung erlernen und erwerben. Dies liegt nicht zuletzt daran, dass nach wie vor eine verbindliche pädagogisch-diagnostische Grundausbildung in der Lehrerbildung fehlt. Zweitens gewähren die Richtlinien zur Leistungsbeurteilung ausdrücklich einen mehr oder weniger großen Beurteilungsspielraum als Bestandteil pädagogischer Freiheit (z.B. ASchO NRW § 21 Abs. 1; Margies, Gampe & Knapp, 2001), wenn auch dadurch beispielweise nicht die Chancengleichheit verletzt werden darf. Und drittens, argumentiert Lüders, gebe es nicht nur unterschiedliche Auffassungen über Sinn und Zweck schulischer Leistungsbeurteilungen unter den Lehrern (bspw. Selektion vs. Förderung), sondern darüber hinaus ebenso unterschiedlichste

Formen der Leistungserhebung und -bewertung, was summa summarum eine Vergleichbarkeit von Urteilen verschiedener Lehrer unmöglich macht.

Schule soll Heranwachsende aber nicht nur einfach für das Leben qualifizieren, sondern dabei jedem einzelnen auch gleiche Entwicklungsmöglichkeiten bieten. Den Lehrerurteilen, die mitunter richtungsweisend sind und Laufbahnen mitbestimmen, kommt dabei ein besonders wichtiger Stellenwert zu, weshalb sie hohe Erwartungen an Objektivität und Fairness erfüllen sollten. Das bedeutet jedoch nicht, dass gleiche Leistungen immer identisch bewertet werden müssen. Der große Handlungsspielraum, den Lehrer für ihre Beurteilungen haben, reicht von der Orientierung an verschiedenen Bezugsnormen (vgl. Rheinberg, 2006) bis zum Einbeziehen weiterer Kriterien. Dies kann durchaus funktional sein, wenn die zugrundeliegenden Kriterien transparent und nachvollziehbar offenliegen. Da dies oftmals leider nicht der Fall ist, ergeben sich fast zwangsläufig Ungerechtigkeiten, die u.a. in der Notengebung oder den Übergangsempfehlungen ihren Niederschlag finden. Mit dem Geschlecht und dem Sozialstatus der Schüler haben sich zwei Merkmale als besonders bedeutsam für die Güte diagnostischer Urteile erwiesen, auf die die Schüler keinen Einfluss haben. Ob darüber hinaus auch das Verhalten der Schüler die Lehrerurteile verzerrt, ist in dieser Studie nicht untersucht worden. Es wäre nicht angemessen, Lehrern gezielte Diskriminierung zu unterstellen, wie auch Ditton (2010) vor dem Hintergrund von Urteilsverzerrungen durch den Sozialstatus der Schüler zu bedenken gibt. Wie er weiterhin annimmt, scheinen sich eher implizite Persönlichkeits- und Begabungstheorien, möglicherweise in Form stereotyper Erwartungshaltungen, einen Einfluss auf die Urteilsgenauigkeit zu haben. Dieser Vermutung kann sich auch nach den Ergebnissen der vorliegenden Untersuchung angeschlossen werden. Umso wichtiger ist jedoch, dass sich Lehrer dieser Einflüsse bewusst sind. Nur wer die Mechanismen im Urteilsprozess kennt, kann ihnen beispielsweise durch besonders gründliches Reflektieren und Hinterfragen der eigenen Beurteilungen entgegenwirken, um sie zu vermeiden.

Tent (2006) resümiert in einem Überblickskapitel zu Schulnoten, dass diese Art der Leistungsbeurteilung weder so schlecht sei, wie sie hingestellt würde, noch so gut, wie sie es ihrem Anspruch nach sein müsste, gibt aber auch zu bedenken, dass überall dort, wo es um wichtige Entscheidungen geht, Lehrerurteile nicht allein den Ausschlag geben dürften, da insbesondere keine gute Vergleichbarkeit von in verschiedenen Klassen erteilten Noten gewährleistet sei. Nicht wenige Autoren fordern daher eine Reform der Lehr-

Lern-Kultur, die weniger das Feststellen einer Leistungsreihenfolge von Schülern zum Ziel hat sondern die Förderung jedes Einzelnen aufgrund fundierter Diagnosen (Winter, 2006). Die Begründungen hierfür leuchten unmittelbar ein: Schulleistungsüberprüfungen konzentrieren sich auf Statusdiagnostik und können aufgrund der ermittelten Ergebnisse nur sehr allgemeine Aussagen zur Leistung machen und den betroffenen Personen kaum Hilfe zur Verbesserung zur Verfügung stellen. Am Ende von Unterrichtseinheiten steht die Überprüfung des neuen Stoffes, wobei den Ergebnissen Noten zugewiesen werden. Allen Schülern ist dabei bewusst, dass nur die Noten am Ende zählen, was Auswirkungen auf die Lernkultur der Schüler hat. Schüler lernen bevorzugt das, was überprüft wird, und sogar die Art, wie geprüft wird, kann sich auf die Art auswirken, in der Schüler lernen. In der Folge wird kurzfristiges und wissensakzentuiertes Lernen begünstigt. Möglichkeiten zur Verbesserung der Lehr-Lern-Kultur sind seit einiger Zeit bekannt. Prioritär ist dafür die Beteiligung der Schüler an der Aufklärung und Optimierung der Lernprozesse (Winter, 2006), was jedoch nicht nur spezielle Gelegenheiten und Zeit, sondern auch entsprechende Instrumente wie zum Beispiel das Portfolio für die ständige Reflexion der Arbeit erfordert (Brunner, Häcker & Winter, 2006).

Die in dieser wie in vielen anderen Arbeiten angewendete Methode, Lehrerurteile Testleistungen von Schülern gegenüberzustellen, mag als realitätsfern angesehen werden, da es sich eher um eine künstlich herbeigeführte Situation handelt, die so im Unterricht kaum vorkommt. Dies liegt einerseits an den standardisierten Testverfahren, die sich durchaus von Klassenarbeiten, Testaten und Klausuren unterscheiden. Es liegt auch andererseits daran, dass bei einem ‚von außen' administrierten Test, bei dem es um nichts geht, der also nicht notenrelevant ist, sowohl die Motivation der Schüler als auch die der Lehrer bei der Schülereinschätzung als niedriger einzuschätzen ist. Dennoch ist zu Recht darauf verwiesen worden, dass derlei Testsituationen durchaus einen hohen Nutzen für die Verbesserung der diagnostisch-methodischen Kompetenzen haben kann (Helmke et al., 2004). Durch die empirische Wende der Bildungspolitik, die mit zunehmender Output-Orientierung einhergeht, ist zu erwarten, dass die Anforderungen an die diagnostische Kompetenz der Lehrer in Zukunft weiter zunehmen werden (Helmke & Hosenfeld, 2004). Um sich mit Modellen schulischer Kompetenzen, geeigneter Erfassungsmethoden vertraut zu machen sowie die zugrunde liegenden Leistungen und Aufgaben (z.B. Leistungsvoraussetzungen oder schwierigkeitsgenerierende Aufgabenmerkmale) besser analysieren zu können, genügt nicht der bloße Erwerb von entsprechenden Kenntnissen, son-

dern es ist deren Transfer in praktisches Handeln nötig (Schrader & Helmke, 2005). Die diagnostischen Fähigkeiten können zum Beispiel dadurch selbst geschult werden, indem jeder Lehrer regelmäßig Situationen herstellt, in denen er die Leistungen einzelner oder mehrerer Schüler zunächst vorhersagen und später mit tatsächlichen - zum Beispiel in Tests erfassten - Leistungen der Schüler abgleichen kann (Wahl et al., 1997). Auch eignet sich beispielsweise der Vergleich der eigenen Urteile mit anderen Quellen Einschätzungen durch Kollegen und auch die die Analyse der dabei auftretenden Unterschiede (Schrader, 2008). Durch diese Art der Selbstreflexion können die im Unterrichtsalltag oft nur implizit vorhandenen Theorien und Hypothesen explizit gemacht werden, was noch dadurch verstärkt werden kann, dass im Falle von abweichenden Einschätzungen nach Gründen gesucht wird und ein Abgleich mit weiteren zusätzlichen Informationen erfolgt. Insbesondere ergeben sich durch derartige Tests, aber auch durch z.B. Vergleichs- oder Orientierungsarbeiten, günstige Möglichkeiten, einem der größten Probleme bei der Leistungsbeurteilung zu begegnen, nämlich der uneinheitlichen Bewertungsmaßstäbe in verschiedenen Klassen oder Schulen. Zu den weiteren Vorteilen des Einsatzes standardisierter Testverfahren im Unterricht gehört nicht zuletzt auch die Vergleichbarkeit der Leistungen der eigenen Klasse mit einer repräsentativen Vergleichsstichprobe. Aus der dadurch möglichen Standortbestimmung können nützliche Maßnahmen für die eigene Unterrichtsentwicklung abgeleitet werden. Die reguläre Verankerung derartiger Selbstevaluation zur Verbesserung der diagnostischen Fähigkeiten steckt hierzulande jedoch noch in den Kinderschuhen.

Eine Vorbildrolle für Deutschland könnte aus Skandinavien kommen. In Schweden beispielsweise existiert die Forderung nach hoher diagnostischer Kompetenz der Lehrer nicht nur als Appell oder allgemeine Lehrplanformulierung, sondern die Nationale Behörde für das Bildungswesen (Skolverket) stellt Lernbedarfsdiagnosen und Lernstandserhebungen als zentrale und im (Grund-)Schulsystem, das von der ersten bis zur neunten Klasse reicht, verankerte Elemente bereit (vgl. Eikenbusch, 2006). Mit den offiziell als „diagnostisches Material“ bezeichneten Aufgaben soll Lehrern geholfen werden zu analysieren und zu prüfen, welche Stärken und Schwächen die Fähigkeiten und Kenntnisse der Schüler aufweisen und - als Prozessmerkmal - welche Lösungswege die Schüler gewählt haben. Durch spezielle Beobachtungs- und Auswertungsbögen soll die Diagnose systematisiert und kriterienorientiert erfolgen, so dass zum einen der diagnostische Blick der Lehrkräfte geschärft und zum anderen strukturierte Rückmeldungen an Schüler und Eltern ermöglicht werden. Die Tests werden dabei ausdrücklich nicht als Aus-

leseinstrument verstanden und sind daher kein Ersatz für normale Leistungstests. Stattdessen wird streng darauf geachtet, dass Leistungs- und Diagnosesituationen sorgfältig voneinander getrennt werden (Eikenbusch, 2006).

Dieses System, das trotz seiner hohen Fortschrittlichkeit etwa wegen der fehlenden Prozessdiagnostik noch Verbesserungspotenzial aufweist, findet in Deutschland insbesondere durch die anfangs in sieben, seit 2007/2008 in allen Bundesländern durchgeführten Vergleichsarbeiten (VERA; davor auch als Orientierungsarbeiten bekannt) Nachahmung, in denen neben der Förderung der Schüler vor allem auch die Verbesserung der Diagnosefähigkeiten der Lehrer im Vordergrund steht. Dabei schätzen Lehrer im Vorfeld von bundeseinheitlichen Leistungserhebungen die zu erwartenden Ergebnisse ihrer Schüler ab und vergleichen diese im Anschluss an die Leistungserhebung mit den erreichten Ergebnissen in der eigenen Klasse. Schrader und Helmke (2005) empfehlen dafür den folgenden Zyklus:

1. Individuelle Auseinandersetzung

Am Anfang steht die selbstkritische Reflexion der eigenen Diagnoseleistungen. Lehrer sollten nach Erklärungen für mögliche Diskrepanzen zwischen den eigenen Erwartungen und den tatsächlich erzielten Ergebnissen suchen.

2. Austausch zwischen Lehrkräften

Durch gemeinsamen Austausch zwischen beteiligten Lehrern ergeben sich vielfältige Möglichkeiten, voneinander zu lernen. Klassenübergreifende Vergleiche können den eigenen Horizont erweitern, Vermutungen über Faktoren der Schülerleistungen, die möglicherweise nicht beachtet wurden und so zu Fehleinschätzungen geführt haben, können zusammengetragen werden.

3. Evaluation

Regelmäßige Wiederholungen der Leistungseinschätzungen, zum Beispiel am Schuljahresende, die wiederum im Kollegium gemeinsam besprochen werden, können als Überprüfung darüber dienen, ob sich Diagnoseleistungen in der Zwischenzeit verbessert haben.

Trotz aller Anstrengungen wäre es unrealistisch zu erwarten, dass auf diese Weise Lehrer zu perfekten Diagnostikern würden. Vielmehr ist die die Überwachung der eigenen Diagnoseleistung eine Daueraufgabe, und der kompetente Diagnostiker zeichnet sich gerade dadurch aus, dass er seine Kompetenzen ständig kritisch überprüft und versucht weiterzuentwickeln

(vgl. Helmke, 2004) und dies als wichtigen Bestandteil seiner Professionalität ansieht. Wenn die Forschung zur diagnostischen Kompetenz ihren Teil dazu beitragen kann, Lehrern die vorhandenen Mängel der Beurteilung aufzuzeigen und dies als konstruktive Hilfe statt als Kritik oder Bevormundung aufgegriffen wird, werden sich in Zukunft sicher weitere Fortschritte auf diesem Gebiet erzielen lassen.

Die Forschung zur diagnostischen Kompetenz ist noch längst nicht abgeschlossen. Die Mehrheit der Studien weist große Limitationen auf, so dass bis heute zwar schon viele Erkenntnisse gewonnen werden konnten, viele Fragen aber nach wie vor offen sind. Dies betrifft zuvorderst die Bedingungen diagnostischer Kompetenz. Solange man nicht genau weiß, was einen guten und was einen schlechten Diagnostiker (in bestimmten Bereichen) ausmacht oder warum bestimmte Schüler ungenauer eingeschätzt werden können als andere, wird man auch von der Lehrerausbildung keine wesentliche Verbesserung erwarten können. Dabei ist „die gezielte Qualifizierung der Lehrkräfte zur Erkennung von Entwicklungsdefiziten [...] ohne Frage eine der vordringlichen Aufgaben“ (Ditton, 2010, S. 269). Nur dann, wenn Lehrer sich der Unterschiede zwischen und Fähigkeiten von Schülern objektiv gewahr sind, können sie im Unterricht Möglichkeiten bieten, mit Leistungsheterogenität gewinnbringend umzugehen, bestehende Defizite auszugleichen, angemessene Urteile zu fällen und somit jeden einzelnen Schüler entsprechend optimal zu fördern und ihn auf den für ihn geeigneten Weg bringen.

9 Abbildungsverzeichnis

10 Tabellenverzeichnis

11 Literaturverzeichnis

Alexander, K. L., Entwisle, D. R. & Thompson, M. S. (1987). School performance, status relations, and the structure of sentiment: Bringing the teacher back in. *American Sociological Review, 51*, 665-682.

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

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (Bd. 4.). Heidelberg: Springer.

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, 3*, 175-193.

Anderson, L. W. (2000). Why should reduced class size lead to increased student achievement? In M. C. Wang & J. D. Finn (Hrsg.), *How small classes help teachers do their best*. (S. 3-24). Philadelphia, PA: Temple University Center for Research in Human Development.

Arnold, K. H. (1999). Diagnostische Kompetenz erwerben. Wie das Beurteilen zu lernen und zu lehren ist. *Pädagogik, 51*, 73-77.

Arnold, K. H., Bos, W., Richert, P. & Stubbe, T. C. (2007). Schullaufbahnpräferenzen am Ende der vierten Klassenstufe. In W. Bos, S. Hornberg, K. H. Arnold, G. Faust, L. Fried, E. M. Lankes, K. Schwippert & R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 271-297). Münster: Waxmann.

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

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

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 - Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. (S. 67-137). Opladen: Leske + Budrich.

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

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U. et al. (2008). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institut für Bildungsforschung.

Baumert, J., Bos, W., Brockmann, J., Gruehn, S., Klieme, E., Köller, O. et al. (2000). *TIMSS/III–Deutschland. Der Abschlussbericht: Zusammenfassung ausgewählter Ergebnisse der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie zur mathematischen und naturwissenschaftlichen Bildung am Ende der Schullaufbahn*. Berlin

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W. et al. (Hrsg.). (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., Maaz, K., Gresch, C., McElvany, N., Anders, Y., Jonkmann, K. et al. (2010). Der Übergang von der Grundschule in die weiterführende Schule - Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten: Zusammenfassung zentraler Befunde. In K. Maaz, J. Baumert, C. Gresch & N. McElvany (Hrsg.), *Bildungsforschung Band 34. Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten*. Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Baumert, J., Schnabel, K. & Lehrke, M. (1998). Learning Math in School: Does Interest Really Matter? In L. Hoffmann, A. Krapp, K. A. Renninger & J. Baumert (Hrsg.), *Interest and Learning. Proceedings of the Seeon Conference on Interest and Gender* (S. 327-336). Kiel: IPN.

Baumert, J. & Schümer, G. (2001). Familiäre Lebensverhältnisse. Bildungsbeteiligung und Kompetenzerwerb. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. (S. 323-407). Opladen: Leske + Budrich.

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In J. Baumert, C. Artelt, E. Klieme, J. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

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*. Opladen: Leske + Budrich.

Baurmann, J. (1995). Der Einfluss von Auswertungsbedingungen, Vorinformation und Persönlichkeitsmerkmalen auf die Benotung von Deutschaufsätzen. In K. Ingenkamp (Hrsg.), *Die Fragwürdigkeit der Zensurengebung*. Weinheim: Beltz.

Bayerisches Staatsministerium für Unterricht und Kultus. (2000). *Lehrplan für die bayerische Grundschule*. München: Verlag J. Maiß GmbH.

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*, 43-55.

Begle, E. G. (1972). *Teacher knowledge and student achievement in algebra. School mathematics study group reports number 9.* Standford, CA: Stanford University, Calif. School Mathematics Study Group.

Bennett, N. (1996). Class size in primary schools: Perceptions of head teachers, chairs of governors, teachers and parents. *British Educational Research Journal, 22*, 33-55.

Bennett, R. E., Gottesmann, R. L., Rock, D. A. & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers´ judgments of students´ academic skills. *Journal of Educational Psychology, 85*, 347-356.

Berliner, D. C. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 15*, 5-13.

Berliner, D. C. (1992). The nature of expertise in teaching. In F. K. Oser, A. Dick & J.-L. Patry (Hrsg.), *Effective and responsible teaching.* San Francisco: Jossey-Bass.

Berliner, D. C. (1994). Teacher Expertise. In T. Husen & T. Neville Postlethwaite (Hrsg.), *The International Encyclopedia Of Education* (2 ed., Bd. 10, S. 6020-6026): Pergamon.

Berufsverband Deutscher Psychologinnen und Psychologen (BDP). (2008). *Lehrer brauchen mehr psychologische und diagnostische Kompetenz*Verfügbar unter: http://www.bdp-verband.org/bdp/presse/2008/07_lehrer-kompetenz.html [21.01.2010

Besser, M. & Krauss, S. (2009). Zur Professionalität als Expertise. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrerprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung.* (S. 71-82). Weinheim: Beltz.

Betts, J. R. & Shkolnik, J. L. (1999). The behavioral effects of variations in class size: The case of math teachers. *Educational Evaluation and Policy Analysis, 21*, 193-213.

Beutel, S.-I. (2007). Kinder und ihr Lernen anerkennen: Lerndiagnose und Leistungsbeurteilung. In U. Graf & E. Moser Opitz (Hrsg.), *Diagnostik und Förderung im Elementarbereich und Grundschulunterricht.* (Band 4 ed., S. 15-29). Baltmannsweiler: Schneider Verlag Hohengehren.

Blatchford, P., Russell, A., Bassett, P., Brown, P. & Martin, C. (2007). The effect of class size on the teaching of pupils aged 7 - 11 years. *School Effectiveness and School Improvement, 18*, 147-172.

Blossfeld, H.-P., Bos, W., Lenzen, D., Hannover, B., Müller-Böling, D., Prenzel, M. et al. (2009). *Geschlechterdifferenzen im Bildungssystem - die Bundesländer im Vergleich. Fakten und Daten zum Jahresgutachten 2009.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Böhnke, K., Silbereisen, R. K., Reynolds, C. R. & Richmond, B. O. (1986). What I think and feel: German experience with the revised form of the Children's Manifest Anxiety Scale. *Personality and Individual Differences, 7*, 553-560.

Boland, T. (1993). The importance of being literate: Reading development in primary school and its consequences for the school career in secondary education. *European Journal of Psychology of Education, 8*, 289-305.

Borko, H., Cone, R., Russo, N. A. & Shavelson, R. J. (1979). Teachers' decision making. In P. L. Peterson & H. J. Wahlberg (Hrsg.), *Research on Teaching. Concepts, Findings and Implications*. (S. 136-160). Berkeley: McCutchan Publishing Corporation.

Borko, H. & Putnam, R. T. (2004). Learning to teach. In D. C. Berliner & R. C. Calfee (Hrsg.), *Handbook of educational psychology* (S. 673-708). London: MacMillan Reference Library.

Bortz, J. (2005). *Statistik für Sozial- und Humanwissenschaftler* (6). Heidelberg: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (Bd. 4.). Heidelberg: Springer Verlag.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Valtin, R. & Walther, G. (Hrsg.). (2005). *IGLU. Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien*. Münster: Waxmann.

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

Bos, W., Schwippert, K. & Stubbe, T. C. (2007). Die Koppelung von sozialer Herkunft und Schülerleistung im internationalen Vergleich. In W. Bos, S. Hornberg, K. H. Arnold, G. Faust, L. Fried, E. M. Lankes, K. Schwippert & R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 225-247). Münster: Waxmann.

Bos, W., Voss, A., Lankes, E. M., Schwippert, K., Thiel, O. & Valtin, R. (2004). Schullaufbahnempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe. In W. Bos (Hrsg.), *IGLU - einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 191-228). Münster: Waxmann.

Bressoux, P., Kramarz, F. & Prost, C. (2008). *Teachers' training, class size and students' outcomes: Learning from administrative forecasting mistakes*. Bonn: Forschungsinstitut zur Zukunft der Arbeit.

Bromme, R. (1992). *Der Lehrer als Experte: Zur Psychologie des professionellen Wissens*. Bern: Huber.

Bromme, R. (1995). Was ist 'pedagogical content knowledge'? Kritische Anmerkungen zu einem fruchtbaren Forschungsprogramm. *Zeitschrift für Pädagogik, Beiheft 33*, 105-115.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (Themenbereich D, Serie I, Band 3 ed., S. 177-212). Göttingen: Hogrefe.

Bromme, R. (2008). Lehrerexpertise als Forschungsprogramm der Lehrerforschung. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie*. (S. 160-167). Göttingen: Hogrefe.

Brophy, J. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75*, 631-661.

Bruner, J. S. & Postman, L. (1951). An approach to social perception. In W. Dennis & R. Lippitt (Hrsg.), *Current trends in social psychology*. (S. 71-118). Pittsburg: Unitersity of Pittsburg Press.

Brunner, I., Häcker, T. & Winter, F. (Hrsg.). (2006). *Handbuch Portfolioarbeit. Konzepte und Erfahrungen aus Schule und Lehrerbildung*. Seelze: Kallmeyer.

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*. Münster: Waxmann.

Brunner, M., Kunter, M., Krauss, S., Baumert, J., Blum, W., Dubberke, T. et al. (2006). Welche Zusammenhänge bestehen zwischen dem fachspezifischen Professionswissen von Mathematiklehrkräften und ihrer Ausbildung sowie beruflichen Fortbildung? *Zeitschrift für Erziehungswissenschaft, 9*, 521-544.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. (Bd. 2., aktualisierte Auflage). München: Pearson Studium.

Carter, K. & Doyle, W. (1987). Teachers' knowledge structures and comprehension processes. In J. Calderhead (Hrsg.), *Exploring teachers' thinking*. London: Cassel.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*.

Charter, R. A. & Larsen, B. S. (1983). Fisher's Z to r. *Educational and Psychological Measurement, 43*, 41-42.

Chi, M. T. H., Feltovich, P. J. & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science., 5*, 121-152.

Chi, M. T. H., Glaser, R. & Farr, M. J. (1988). *The nature of expertise*. Hillsdale: Erlbaum.

Clark, C. M. & Peterson, P. L. (1986). Teachers thought processes. In M. Wittrock (Hrsg.), *Handbook of research on teaching* (Bd. 3. ed., 9. print., S. 255-296). New York: Macmillan [u.a.].

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

Colker, L. (1984). Teachers' interactive thoughts about pupil cognition.,*Paper presented at the annual meeting of the American Educational Research Association*. New Orleans.

Corno, L. & Snow, R. (1986). Adapting teaching to individual differences among learners. In M. Wittrock (Hrsg.), *Handbook of research on teaching* (S. 605-629). New York: Macmillan.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177-183.

Czerwenka, K., Nölle, K., Pause, G., Schlotthaus, W., Schmidt, H.-J. & Tessloff, J. (1990). *Schülerurteile über die Schule. Bericht über eine internationale Untersuchung*. Frankfurt: Lang.

Deci, E. L. (1992). The relation of interest to the motivation of behavior: A self-determination theory perspective. In K. A. Renninger, S. Hidi & A. Krapp (Hrsg.), *The role of interest in learning and development*. (S. 43-47). Hillsdale, NJ: Erlbaum.

Demaray, M. K. & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quaterly, 13*, 8-24.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.

DESI-Konsortium. (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI)*. Frankfurt/M.: Deutsches Institut für Internationale Pädagogische Forschung.

Deutscher Bildungsrat. (1970). *Strukturplan für das Bildungswesen*. Bonn: Autor.

Ditton, H. (1992). *Ungleichheit und Mobilität. Theorie und Empirie über sozialräumliche Aspekte von Bildungsentscheidungen*. Weinheim: Juventa-Verlag

Ditton, H. (2004). Schule und sozial-regionale Ungleichheit. In W. Helsper & J. Böhme (Hrsg.), *Handbuch der Schulforschung* (S. 605-624). Wiesbaden: VS Verlag für Sozialwissenschaftler.

Ditton, H. (2010). Der Beitrag von Schule und Lehrern zur Reproduktion von Bildungsungleichheit. In R. Becker & W. Lauterbach (Hrsg.), *Bildung als Privileg? Erklärungen und Befunde zu den Ursachen der Bildungsungleichheit*. (S. 247-275). Wiesbaden: VS-Verlag für Sozialwissenschaften.

Ditton, H. & Krüsken, J. (2006). Sozialer Kontext und schulische Leistungen - zur Bildungsrelevanz segregierter Armut. *Zeitschrift für Soziologie der Erziehung und Sozialisation, 26*, 135-157.

Ditton, H. & Krüsken, J. (2009). Denn wer hat, dem wird gegeben werden?: Eine Längsschnittstudie zur Entwicklung schulischer Leistungen und den Effekten der sozialen Herkunft in der Grundschulzeit. *Journal for Educational Research Online, 1*, 33-61.

Ditton, H., Krüsken, J. & Schauenberg, M. (2005). Bildungsungleichheit - Der Beitrag von Familie und Schule. *Zeitschrift für Erziehungswissenschaft, 8*, 285-304.

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*, 41-60.

Druva, C. A. & Anderson, R. D. (1983). Science teacher characteristics by teacher behavior and by student outcome: A meta-analysis of research. *Journal of Research in Science Teaching, 20*, 467-479.

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

Dusek, J. B. & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75*, 327-346.

Eckert, T. L. & Arbolino, L. A. (2005). 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: Guilford Press.

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

Ehrenberg, R. G., Brewer, D. J., Gamoran, A. & Willms, D. J. (2001). Class size ans student achievement. *Psychological Science in the Public Interest, 2*, 1-30.

Eikenbusch, G. (2006). "Macht richtige Lerndiagnose!": Erfahrungen und Tendenzen aus Schweden. *Friedrich Jahresheft XXIV*, 20-21.

Epstein, J. (1981). Patterns of classroom participation, student attributes, and achievements. In J. Epstein (Hrsg.), *Quality of school life* (S. 271-288). Lexington, MA: Heath (D.C.) and Co.

Faber, G. (2001). Das Verhalten rechtschreibängstlicher Grundschulkinder im Lehrerurteil: Empirische Untersuchungsergebnisse zur Problematik informeller Alltagsdiagnosen. *Heilpädagogische Forschung, 27*, 58-65.

Faber, G. (2006). Die Erfassung rechtschreibängstlicher Besorgtheit und Aufgeregtheit. Zur Bedeutung ausgewählter Forschungsergebnisse für lerntherapeutische Diagnose- und Interventionskonzepte. *Sprachrohr Lerntherapie, 2*, 1-14.

Federer, M., Stüber, S., Margraf, J., Schneider, S. & Herrle, J. (2001). Selbst- und Fremdeinschätzung der Kinderängstlichkeit. *Zeitschrift für Differentielle und Diagnostische Psychologie, 22*, 194-205.

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

Fiedler, K. (1983). Beruhen Bestätigungsfehler nur auf einem Bestätigungsfehler? Eine Replik auf Gadenne. *Psychologische Beiträge, 25*, 280-286.

Finn, J. D., Pannozzo, G. M. & Achilles, C. M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research, 73*, 321-368.

Fisher, C. W., Filby, N., Marliave, R., Cahen, L. S., Dishaw, M. M. & Moore, J., et al. (1978). *Teaching behaviors, academic learning time, and student achievement: Final report of phase III-B, Beginning Teacher Evaluation Study*. San Francisco: Far West Laboratory.

Flett, G. L., Hewitt, P. L. & Hallett, C. J. (1995). Perfectionism and job stress in teachers. *Canadian Journal of School Psychology, 11*, 32-42.

Gage, N. (1978). *Unterrichten - Kunst oder Wissenschaft?* München: Urban & Schwarzenberg.

Gage, N. & Berliner, D. (1977). *Pädagogische Psychologie*. München: Urban & Schwarzenberg.

Gerber, M. M. & Semmel, M. I. (1984). Teacher as imperfect test: reconceptualizing the referral process. *Educational Psychologist, 19*, 137-148.

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*, 321-331.

Gobet, F. (1996). Expertise und Gedächtnis. In H. Gruber & A. Ziegler (Hrsg.), *Expertiseforschung* (S. 58-79). Opladen: Westdeutscher Verlag.

Gobet, F. (2001). Cognitive psychology of chess expertise. In N. J. Smelser & P. B. Baltes (Hrsg.), *International encyclopedia of the social and behavioral sciences.* (Bd. Vol. 8, S. 1663-1667). Amsterdam: Elsevier.

Goldhaber, D. & Brewer, D. J. (1997). Evaluating the effect of teacher degree level on educational performance. In J. W. Fowler (Hrsg.), *Developments in School Finance 1996.* (S. 197-210). Washington.

Goldhaber, D. & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*, 129-145.

Gölitz, D., Roick, T. & Hasselhorn, M. (2006). *DEMAT 4. Deutscher Mathematiktest für vierte Klassen*. Göttingen: Hogrefe.

Graham, J. W., Cumsille, P. E. & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Hrsg.), *Handbook of psychology: Research methods in psychology*. (Bd. 2, S. 87-114). New York: John Wiley & Sons.

Greb, K. & Lipowsky, F. (2009). Kompetenzmodellierung des diagnostischen Urteils bei Grundschullehrern., *Frühjahrstagung der Arbeitsgemeinschaft für empirische pädagogische Forschung (AEPF)*. Landau.

Gresham, F. M., Reschly, D. J. & Carey, M. P. (1987). Teachers as "tests": Classification accuracy and concurrent validation in the identification of learning disabled children. *School Psychology Review, 16*, 543-553.

Grube, D. (2004). Entwicklung des Rechnens im Grundschulalter. In M. Hasselhorn, H. Marx & W. Schneider (Hrsg.), *Diagnostik von Mathematikleistungen* (Bd. 4). Göttingen: Hogrefe.

Gruber, H. (2006). Expertise. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3 ed., S. 175-180). Weinheim, u.a.: Beltz.

Grund, M., Haug, G. & Naumann, C. L. (2003). *DRT 4. Diagnostischer Rechtschreibtest für 4. Klassen*. Göttingen: Hogrefe.

Guskey, T. R. (2000). Grading policies that work against standards ... and how to fix them. *NASSP Bulletin, 84*, 20-29.

Haffner, J., Baro, K., Parzer, P. & Resch, F. (2005). *Heidelberger Rechentest (HRT 1-4)*. Göttingen: Hogrefe.

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*, 228-240.

Hannover, B. (1998). The Development of Self-Concept and Interests. In L. Hoffmann, A. Krapp, K. A. Renninger & J. Baumert (Hrsg.), *Interest and Learning. Proceedings of the Seeon Conference on Interest and Gender* (S. 105-125). Kiel: IPN.

Hartig, J. (2006). Kompetenzen als Ergebnisse von Bildungsprozessen. *dipf informiert, 10*.

Hascher, T. (2005). Diagnostizieren in der Schule. In A. Bartz, C. Kloeft, J. Fabian, S. Huber, H. Rosenbusch & H. Sassenscheidt (Hrsg.), *PraxisWissen SchulLeitung* (S. 1-8). Bonn: WoltersKluwer.

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.

Heckhausen, H. (1974). Lehrer-Schüler-Interaktion. In F. E. Weinert, C. F. Graumann, H. Heckhausen & M. Hofer (Hrsg.), *Funkkolleg Pädagogische Psychologie*. (S. 547-573). Frankfurt: Fischer.

Heckhausen, H. (1982). Task-irrelevant cognitions during an exam: Incidence and effects. In H. W. Krohne & L. Laux (Hrsg.), *Achievement, stress, and anxiety*. (S. 247-274). Washington: Hemisphere Publishing Corp.

Heckhausen, H. (1989). Leistungsmotivation. In *Motivation und Handeln* (2 ed., S. 231-278). Berlin: Springer-Verlag.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Heller, K. & Geisler, H. J. (1983). *Kognitiver Fähigkeitstest für 1. bis 3. Klassen (KFT 1-3)*. Weinheim: Beltz.

Helmke, A. (1980). *Angst in der Schule. Zwischenbericht über Ergebnisse der 1. Hauptuntersuchung für die untersuchten Schulen*. Konstanz: Universität Konstanz, Sonderforschungsbereich 23, Längsschnittprojekt "Entwicklung im Jungendalter".

Helmke, A. (1983a). Prüfungsangst - Ein Überblick über neuere theoretische Entwicklungen und empirische Ergebnisse. *Psychologische Rundschau, Band XXXIV*, 193—211.

Helmke, A. (1983b). *Schulische Leistungsangst - Erscheinungsformen und Entstehungsbedingungen*. Frankfurt: Lang.

Helmke, A. (1993). Die Entwicklung der Lernfreude vom Kindergarten bis zur 5. Klassenstufe. *Zeitschrift für Pädagogische Psychologie, 7*, 77-86.

Helmke, A. (1997). Entwicklung lern- und leistungsbezogener Motive und Einstellungen: Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 59-76). Weinheim: Beltz.

Helmke, A. (2003). *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze: Kallmeyersche Verlagsbuchhandlung.

Helmke, A. (2004). Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *Seminar, Heft 2/2004*, 90-112.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität* (Bd. 1). Seelze-Velber: Kallmyer i.V.m. Klett.

Helmke, A. & Fend, H. (1981). Wie gut kennen Eltern ihre Kinder und Lehrer ihre Schüler? In G. Zimmer (Hrsg.), *Persönlichkeitsentwicklung und Gesundheit im Schulalter* (S. 341-360). Frankfurt: Campus.

Helmke, A. & Hosenfeld, I. (2004). Bildungsstandards und Unterrichtsqualität: Notwendigkeit einer empirischen Wende für die Schulpraxis. *Pädagogische Führung, 04-2004*, 1-6.

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

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (Themenbereich D, Serie I, Band 3 ed., S. 71-176). Göttingen: Hogrefe.

Helwig, R., Anderson, L. & Tindal, G. (2001). Influence of elementary student gender on teachers´ perceptions of mathematics archievement. *Journal of Educational Research, 95*, 93-102.

Herppich, S., Wittwer, J., Nückles, M. & Renkl, A. (2010). Do tutors' content knowledge and beliefs about learning influence their assessment of tutees' understanding? In S. Ohlsson & R. Catrambone (Hrsg.), *Proceedings of the 32th Annual Conference of the Cognitive Science Society*. (S. 314-319). New York: Erlbaum.

Hertel, S., Bruder, S. & Schmitz, B. (2009). Beratungs- und Gesprächsführungskompetenzen von Lehrkräften. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrerprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung*. (S. 117-128). Weinheim: Beltz.

Hesse, I. & Latzko, B. (2009). *Diagnostik für Lehrkräfte*. Opladen: Verlag Barbara Budrich.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice-performance. *Journal of Experimental Psychology: Applied, 5*, 205-221.

Hinnant, J. B., O'Brien, M. & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school year. *Journal of Educational Psychology, 101*, 662-670.

Hofer, M. (1986). *Sozialpsychologie erzieherischen Handelns*. Göttingen: Hogrefe.

Hoge, R. D. & Butcher, R. (1984). Analysis of teacher judgements of pupil achievement levels. *Journal of Educational Psychology, 76*, 777-781.

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

Holtappels, H. G. & Leffelsend, S. (2003). *Entwicklung von Methodenkompetenz durch Schülertrainings und Unterrichtsentwicklung. Ergebnisse einer Schülerbefragung als Teil der Abschlussevaluation des Projekes "Schule & Co."*. Gütersloh: Bertelsmann Stiftung.

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

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*. (45. Beiheft ed.). 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*, 115-144.

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

Ingenkamp, K. (1989). *Diagnostik in der Schule. Beiträge zu Schlüsselfragen der Schülerbeurteilung*. Weinheim: Beltz.

Ingenkamp, K. (1991). Pädagogische Diagnostik. In L. Roth (Hrsg.), *Pädagogik. Handbuch für Studium und Praxis*. (S. 760-785). München: Ehrenwirth.

Ingenkamp, K. (1995a). Erfassung und Rückmeldung des Lernerfolgs. In D. Lenzen (Hrsg.), *Enzyklopädie Erziehungswissenschaft, Band 3*. Stuttgart.

Ingenkamp, K. (2005). *Lehrbuch der pädagogischen Diagnostik* (5). Weinheim: Beltz.

Ingenkamp, K. (Hrsg.). (1995b). *Die Fragwürdigkeit der Zensurengebung*. Weinheim: Beltz.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik. (6. Auflage)*. Weinheim: Beltz.

Itskowitz, R., Navon, R. & Strauss, H. (1988). Teachers' accuracy in evaluating students' self-image: effect of perceived closeness. *Journal of Educational Psychology, 80*, 337-341.

Jäger, R. S. (2000). *Von der Beobachtung zur Notengebung - Ein Lehrbuch: Diagnostik und Benotung in der Aus-, Fort- und Weiterbildung*. Landau: Verlag Empirische Pädagogik.

Janke, N. (2006). Wahrnehmung des Deutschunterrichts durch Schülerinnen und Schüler als Teil des Unterrichtsklimas. In W. Bos & M. Pietsch (Hrsg.), *KESS 4 - Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen* (Bd. 1, S. 165-179). Münster, u.a.: Waxmann.

Jerusalem, M. & Mittag, W. (1999). Selbstwirksamkeit, Bezugsnormen, Leistung und Wohlbefinden in der Schule. In M. Jerusalem & R. Pekrun (Hrsg.), *Emotion, Motivation und Leistung* (S. 223-245). Göttingen: Hogrefe.

Judd, C. M. & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review, 100*, 109-128.

Jude, N. & Klieme, E. (2007). Sprachliche Kompetenzen aus Sicht der pädagogisch-psychologischen Diagnostik. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messungen. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. Weinheim: Beltz.

Jürgens, E. (2005). *Leistung und Beurteilung in der Schule: Eine Einführung in Leistungs- und Bewertungsfragen aus pädagogischer Sicht*. Sankt Augustin: Academia.

Jussim, L. & Eccles, J. (1995). Are teacher expectations biased by students´ gender, social class, or ethnicity? In L. Jussim, J. Eccles & C. R. McCauley (Hrsg.), *Stereotype accuracy. Toward appreciating group differences*. Washington: American Psychological Association.

Jussim, L., Eccles, J. & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology, 28*, 281-388.

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*, 131-155.

Jussim, L., McCauley, C. & Lee, Y. T. (1995). Why study stereotype accuracy and inaccuracy? In Y. T. Lee, L. Jussim & C. McCauley (Hrsg.), *Stereotype accuracy: Toward an appreciation of group differences* (S. 1-23). Washington, DC: American Psychological Association.

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

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

Kebeck, G. (1994). *Wahrnehmung. Theorien, Methoden und Forschungsergebnisse der Wahrnehmungspsychologie*. Weinheim: Juventa.

Kennedy, E. (1995). Contextual effects on academic norms among elementary school students. *Educational Research Quarterly, 18*, 5-13.

Kenny, D. T. & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessment. *Journal of Learning Disabilities, 4*, 227-236.

Klauer, K. J. (1989). Zensierungsmodelle und ihre Konsequenzen für die Notengebung. In R. S. Jäger, R. Horn & K. Ingenkamp (Hrsg.), *Tests und Trends. 7. Jahrbuch der pädagogischen Diagnostik*. (S. 40-68). Weinheim: Beltz.

Kleber, E. W. (1976). *Beurteilung und Beurteilungsprobleme. Eine Einführung in Beurteilungs- und Bewertungsfragen in der Schule*. Weinheim: Beltz.

Kleber, E. W. (1992). *Diagnostik in pädagogischen Handlungsfeldern: Einführung in Bewertung, Beurteilung, Diagnose und Evaluation*. Weinheim, u.a.: Juventa Verlag.

Klieme, E. (2004). Was sind Kompetenzen und wie lassen sie sich messen? *Pädagogik, 56*, 10-13.

Kluge, F. (1975). *Etymologisches Wörterbuch der deutschen Sprache*. Berlin, New York: Walter de Gruyter.

Köller, O., Baumert, J. & Schnabel, K. (2000). Zum Zusammenspiel von schulischem Interesse und Lernen im Fach Mathematik: Längsschnittanalysen in den Sekundarstufen I und II. In U. Schiefele & E. Wild (Hrsg.), *Interesse und Lernmotivation. Untersuchungen zu Entwicklung, Förderung und Wirkung* (S. 163-180). Münster: Waxmann.

Krapp, A. (1989). Neuere Ansätze einer pädagogisch orientierten Interessenforschung. *Empirische Pädagogik, 3*, 233-255.

Krapp, A. (2002). Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and Instruction, 12*, 383-409.

Krapp, A. (2006). Interesse. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3 ed., S. 280-290). Weinheim, u.a.: Beltz.

Krapp, A., Hidi, S. & Renninger, K. A. (1992). Interest, learning, and development. In K. A. Renninger, S. Hidi & A. Krapp (Hrsg.), *The role of interest in learning and development*. (S. 3-25). Hillsdale, NJ: Erlbaum.

Krauss, S., Kunter, M., Brunner, M., Baumert, J., Blum, W., Neubrand, M. et al. (2004). COACTIV: Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz. In J. Doll & M. Prenzel (Hrsg.), *Die Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung*. (S. 31-53). Münster: Waxmann.

Krems, J. F. (1996). Expertise und Flexibilität. In H. Gruber & A. Ziegler (Hrsg.), *Expertiseforschung* (S. 80-91). Opladen: Westdeutscher Verlag.

Kristen, C. (2002). Hauptschule, Realschule oder Gymnasium? Ethnische Unterschiede am ersten Bildungsübergang. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 54*, 534-552.

Kristen, C. (2006). Ethnische Diskriminierung in der Grundschule? Die Vergabe von Noten und Bildungsempfehlungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 58*, 79-97.

Krumboltz, J. D. & Yeh, C. J. (1996). Competitive grading sabotages good teaching. *Phi Delta Kappan, 78*, 324-326.

Kultusministerkonferenz. (2001). *296. Plenarsitzung der Kultusministerkonferenz am 05./06. Dezember 2001 in Bonn.*Verfügbar unter: http://www.kmk.org/presse-und-aktuelles/pm2000/pm2001/296plenarsitzung.html [21.01.2010]

Kultusministerkonferenz. (2004a). *Bildungsstandards im Fach Deutsch für den Primarbereich. Beschluss vom 15.10.2004*. Neuwied: Luchterhand.

Kultusministerkonferenz. (2004b). Standards für die Lehrerbildung: Bildungwissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004. Bonn: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Kunter, M., Kleickmann, T., Klusmann, U. & Richter, D. (2011). Die Entwicklung professioneller Kompetenz von Lehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Langfeldt, H. P. (1984). Die klassische Testtheorie als Grundlage normorientierter (standardisierter) Schulleistungstests. In K. A. Heller (Hrsg.), *Leistungsdiagnostik in der Schule*. (Bd. 4, S. 65-98). Bern, Stuttgart, Toronto: Verlag Hans Huber.

Langfeldt, H. P. (2006). Diagnosekompetenz von Lehrerinnen und Lehrern. In H. P. Langfeldt (Hrsg.), *Psychologie für die Schule* (S. 195-211). Weinheim und Basel: BELTZ.

Lehmann, R., Peek, R. & Gänsfuß, R. (1997). *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klasse an Hamburger Schulen. Bericht über die Untersuchung im September 1996*. Hamburg: Behörde für Schule, Jugend und Berufsbildung, Amt für Schule.

Lehmann, R., Peek, R., Gänsfuß, R., Lutkat, S., Mücke, S. & Barth, I. (2000). *Qualitätsuntersuchungen an Schulen zum Unterricht in Mathematik (QuaSUM)*. Potsdam: Ministerium für Bildung, Jugend und Sport des Landes Brandenburg (MBJS).

Lehmann, R., Peek, R., Gänsfuß, R., Lutkat, S., Mücke, S. & Barth, I. (2004). *QuaSUM. Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik. Ergebnisse einer repräsentativen Untersuchung im Land Brandenburg*. Schulforschung in Brandenburg (Heft 1).

Leinhardt, G. (1983). Novice and expert knowledge of individual student's achievement. *Educational Psychologist, 18*, 165-179.

Leinhardt, G. (1987). Development of an expert explanation: An analysis of a sequence of subtraction lessons. *Cognition and instruction, 4*, 225-282.

Leinhardt, G. (2001). Instructional explanations: A commonplace for teaching and location for contrast. In V. Richardson (Hrsg.), *Handbook of research on teaching*. (S. 333-357). Washington.

Leinhardt, G. & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology, 78*, 75-95.

Leinhardt, G. & Smith, D. (1985). Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology, 77*, 247-271.

Lenhard, W. & Schneider, W. (2006). *ELFE 1-6. Ein Leseverständnistest für Erst- bis Sechstklässler*. Göttingen: Hogrefe.

Liebert, R. M. & Morris, L. W. (1967). Cognitive and emotional components of test anxiety. *Psychological Reports, 20*, 975-978.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse (6. Aufl.)*. Weinheim: Psychologie Verlags Union.

Lingelbach, H. (1995). *Unterrichtsexpertise von Grundschullehrkräften* (Bd. Band 35): Verlag Dr. Kovac.

Lipowsky, F. (2006). Auf den Lehrer kommt es an. Empirische Evidenz für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. *Zeitschrift für Pädagogik, 52*, 47-70.

Livingston, C. & Borko, H. (1989). Expert-novice differences in teaching: A cognitive analysis and implications for teacher education. *Journal of Teacher Education, 40*, 36-42.

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

Lorenz, C. & Artelt, C. (2010). Spezifische vs. globale Lehrerurteile: Unterschiede in Bezug auf Sozialstatus und Geschlecht der Schüler, *Vortrag auf der 74. Tagung der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF)*. 13. - 15. September 2010, Jena.

Lüders, M. (2001). Probleme von Lehrerinnen und Lehrern mit der Beurteilung von Schülerleistungen. *Zeitschrift für Erziehungswissenschaft, 4*, 457-474.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen. *Psychologische Rundschau, 58*, 103-117.

Lukesch, H. (1998). *Einführung in die pädagogisch-psychologische Diagnostik*. Regensburg: Roderer Verlag.

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*, 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 Review, 24*, 1304-1318.

Margies, D., Gampe, H. & Knapp, R. (2001). *Allgemeine Schulordnung in Nordrhein-Westfalen (ASchO). Kommentar*. Neuwied: Luchterhand.

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2*, 77-172.

Marsh, H. W. & Craven, R. G. (1991). Self-other agreement on multiple dimensions of preadolescent self-concept: Inferences by teachers, mothers, and fathers. *Journal of Educational Psychology, 83*, 393-404.

Martinek, D. (2007). *Die Ungewissheit im Lehrberuf. Orientierungsstil, Motivationsstrategie und Bezugsnorm-Orientierung bei Lehrer/innen*. Hamburg: Verlag Dr. Kovač.

McCauley, C. (1995). Are stereotypes exaggerated? A sampling of racial, gender, academic, occupational and political stereotypes. In Y. T. Lee, L. Jussim & C. McCauley (Hrsg.), *Stereotype accuracy: Toward appreciating group differences*. (S. 215-243). Washington, DC: American Psychological Association.

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W. et al. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von

Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern. *Zeitschrift für Pädagogische Psychologie, 23*, 223-235.

McNair, K. (1978). Capturing inflight decisions: Thoughts while teaching. *Educational Research Quarterly, 3*, 26-42.

Medley, D. M. (1979). The Effectiveness of Teachers. In P. L. Peterson & H. J. Wahlberg (Hrsg.), *Research on Teaching. Concepts, Findings and Implications* (S. 11-27). Berkeley: McCutchan Publishing Corporation.

Metzig, W. & Schuster, M. (2001). Prüfungsangst - Leistungsangst: eine neue Sichtweise. *Unterrichten, erziehen, 20*, 287-290.

Minnameier, G. (2005). Wissen und Können im Kontext inferentiellen Denkens. In H. Heid & C. Harteis (Hrsg.), *Verwertbarkeit. Ein Qualitätskriterium (erziehungs-)wissenschaftlichen Wissens?* Wiesbaden: VS Verlag für Sozialwissenschaften.

Möller, J. & Köller, O. (1997). Kontexteffekte in Berichtszeugnissen. *Psychologie in Erziehung und Unterricht, 44*, 187-196.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A. & Ehrle, K. (1999). Evaluating the SAGE program: A pilot program in targeted pupil-teacher reduction in Wisconsin. *Educational Evaluation and Policy Analysis, 21*, 165-177.

Moosbrugger, H. & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer-Verlag.

Moser, U. & Rhyn, H. (2000). *Lernerfolg in der Primarschule. Eine Evaluation der Leistungen am Ende der Primarschule*. Aarau: Sauerländer.

Mulholland, L. A. & Berliner, D. C. (1992). Teacher experience and the estimation of student achievement, *Paper presented at the Annual Meeting of the American Educational Research Association*. San Francisco, CA.

Müller, R. (2003). *DRT 3. Diagnostischer Rechtschreibtest für 3. Klassen*. (Bd. 4). Göttingen: Hogrefe.

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child Development, 49*, 800-814.

Nisbett, R. & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement*. Englewood Cliffs: Prentice Hall.

OECD (Hrsg.). (2005). *School factors related to quality and equity. Results from PISA 2000*. Bonn: UNO-Verlag.

Oevermann, U., Kieper, M., Rothe-Bosse, S., Schmidt, M. & Wienskowiski, P. (1976). Die sozialstrukturelle Einbettung von Sozialisationsprozessen: Empirische Ergebnisse zur Ausdifferenzierung des globalen Zusammenhangs von Schichtzugehörigkeit und gemessener Intelligenz sowie Schulerfolg. *Zeitschrift für Soziologie 2*, 167-199.

Pate-Bain, H., Achilles, C. M., Boyd-Zaharias, J. & McKenna, B. (1992). Class size makes a difference. *Phi Delta Kappan, 74*, 253-256.

Pekrun, R. & Helmke, A. (1991). Schule und Persönlichkeitsentwicklung: Theoretische Perspektiven und Forschungsstand. In R. Pekrun & H. Fend (Hrsg.), *Schule und Persönlichkeitsentwicklung. Ein Resümee der Längsschnittforschung*. Stuttgart: Enke.

Pietsch, M. (2007). Soziale Herkunft und Schulleistung Hamburger Kinder am Ende der Grundschulzeit. In W. Bos, C. Gröhlich & M. Pietsch (Hrsg.), *KESS 4 - Lehr- und Lernbedingungen in Hamburger Grundschulen. Hamburger Schriften zur Qualität im Bildungswesen, Band 2.* (S. 1-46). Münster: Waxmann.

Prenzel, M., Lankes, E.-M. & Minsel, B. (2000). Interessenentwicklung in Kindergarten und Grundschule: Die ersten Jahre. In U. Schiefele & E. Wild (Hrsg.), *Interesse und Lernmotivation. Untersuchungen zu Entwicklung, Förderung und Wirkung* (S. 11-30). Münster: Waxmann.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

Rauer, W. & Schuck, K. D. (2003). *FEESS 3-4. Fragebogen zur Erfassung emotionaler und sozialer Schulerfahrungen von Grundschulkindern dritter und vierter Klassen*. Göttingen: Beltz Test GmbH.

Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation*. Bochum: Verlag für Psychologie.

Rheinberg, F. (1982). Bezugsnorm - Orientierung angehender Lehrer im Verlauf ihrer praktischen Ausbildung. In F. Rheinberg (Hrsg.), *Bezugsnorm zur Leistungsbewertung: Analyse und Intervention* (S. 235-248). Düsseldorf: Pädagogischer Verlag Schwann.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen*. (S. 59-71). Weinheim: Beltz.

Rheinberg, F. (2006). Bezugsnormorientierung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3 ed., S. 55-62). Weinheim, u.a: Beltz.

Rheinberg, F., Bromme, R., Minsel, B., Winteler, A. & Weidenmann, B. (2001). Die Erziehenden und Lehrenden. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (S. 271-356). Weinheim: Beltz PVU.

Rieder, K. (1990). Leistungsbeurteilung und Notengebung. In R. Olechowski & K. Rieder (Hrsg.), *Motivieren ohne Noten*. Wien/München: Jugend und Volk Verlagsgesellschaft.

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

Roeder, P. M., Baumert, J., Sang, F. & Schmitz, B. (1986). Über Zusammenhänge zwischen Zensur und Testleistung. In H. Petillon, J. W. L. Wagner & B. Wolf (Hrsg.), *Schülergerechte Diagnose. Theoretische und empirische Beiträge zur Pädagogischen Diagnostik* (S. 31-59). Weinheim, u.a.: Beltz Verlag.

Rogalla, M. & Vogt, F. (2008). Förderung adaptiver Lehrkompetenz. *Unterrichtswissenschaft, 36*, 17-36.

Rohrmann, T. (2007). Jungen und Mädchen in der Schule. In T. Fleischer, N. Grewe, B. Jötten, K. Seifried & B. Sieland (Hrsg.), *Handbuch Schulpsychologie* (S. 221-229). Stuttgart: Kohlhammer.

Roick, T., Gölitz, D. & Hasselhorn, M. (2004). *DEMAT 3+. Deutscher Mathematiktest für dritte Klassen*. Göttingen: Hogrefe.

Rolff, H.-G., Holtappels, H.-G., Klemm, K., Pfeiffer, H. & Schulz-Zander, R. (Hrsg.). (2002). *Jahrbuch der Schulentwicklung, Band 12. Daten, Beispiele und Perspektiven*. Weinheim: Juventa.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal, R. & Jacobson, L. (1971). *Pygmalion im Unterricht*. Weinheim: Beltz.

Rost, D. H. & Schermer, F. J. (2007). Leistungsängstlichkeit. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie*. Weinheim: Beltz.

Roth, H. (1971). *Pädagogische Anthropologie*. (Bd. Band 2). Hannover.

Rowan, B., Correnti, R. & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *Teachers College Record, 104*, 1525-1567.

Rubin, D. B. (1987). Introduction. In *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Sabers, D. S., Cushing, K. S. & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. . *American Educational Research Journal, 28*, 63-88.

Sacher, W. (1996). *Prüfen - Beurteilen - Benoten. Grundlagen, Hilfen und Denkanstöße für alle Schularten* (2). Bad Heilbrunn: Verlag Julius Klinkhardt.

Sacher, W. (2009). *Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primar- und Sekundarstufe (5., überarbeitete und erweiterte Aufl.)*. Bad Heilbrunn/Obb.: Klinkhardt.

Salvia, J. & Ysseldyke, J. E. (2004). *Assessment* (9th). New York: Houghton Mifflin.

Schaarschmidt, U., Kieschke, U. & Fischer, A. W. (1999). Beanspruchungsmuster im Lehrerberuf. *Psychologie in Erziehung und Unterricht, 4*, 244-268.

Schiefele, U. (1996). *Motivation und Lernen mit Texten*. Göttingen: Hogrefe.

Schiefele, U., Krapp, A. & Winteler, A. (1992). Interest as predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi & A. Krapp (Hrsg.), *The role of interest in learning and development* (S. 183-212). Hillsdale, NJ: Erlbaum.

Schneider, S. (2005). Lernfreude und Schulangst. Wie es 8- bis 9-jährigen Kindern in der Grundschule geht. In C. Alt (Hrsg.), *Kinderleben - Aufwachsen zwischen Familie, Freunden und Institutionen* (S. 199-230). Wiesbaden: VS Verlag.

Schneider, T. (2009). *Die Bedeutung der sozialen und ethnischen Herkunft für Lehrerurteile am Beispiel der Grundschulempfehlung.*

Schneider, W. & Stefanek, J. (2007). Entwicklung der Rechtschreibleistung vom frühen Schul- bis zum frühen Erwachsenenalter. Längsschnittliche Befunde der Münchner LOGIK-Studie. *Zeitschrift für Pädagogische Psychologie, 21*, 77-82.

Scholz, G. (1993). Statusdiagnostik vs. Prozeßdiagnostik? In C. Tarnai (Hrsg.), *Beiträge zur empirischen pädagogischen Forschung*. (S. 124-143). Münster: Waxmann.

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

Schrader, F.-W. (2006). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3 ed., S. 95-100). Weinheim, u.a.: Beltz.

Schrader, F.-W. (2008). Diagnoseleistungen und diagnostische Kompetenzen von Lehrkräften. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie*. (S. 168-177). Göttingen: Hogrefe.

Schrader, F.-W. (2009). Anmerkungen zum Themenschwerpunkt Diagnostische Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 23*, 237-245.

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

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

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

Schrader, F.-W. & Helmke, A. (2005). Überprüfte Vermutungen. *Friedrich Jahresheft XXIII*, 120-121.

Schwippert, K. (2007). Migrationsbedingte Heterogenität von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in Hamburg. In W. Bos, C. Gröhlich & M. Pietsch (Hrsg.), *KESS 4 - Lehr- und Lernbedingungen in Hamburger Grundschulen. Hamburger Schriften zur Qualität im Bildungswesen, Band 2.* (S. 35-46). Münster: Waxmann.

Schwippert, K., Bos, W. & Lankes, E. M. (2004). Heterogenität und Chancengleichheit am Ende der vierten Jahrgangsstufe in den Ländern der Bundesrepublik Deutschland und im internationalen Vergleich. In W. Bos, E. M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 165-190). Münster: Waxmann.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2010). *Übergang von der Grundschule in Schulen des Sekundarbereichs I und Förderung, Beobachtung und Orientierung in den Jahrgangsstufen 5 und 6 (sog. Orientierungsstufe)*. München.

Shapiro, E. S. & Kratochwill, T. R. (2000). Introduction: Conducting a multidimensional behavioral assessment. In E. S. Shapiro & T. R. Kratochwill (Hrsg.), *Conductiong*

school-based assessments of child and adolescent behavior. (S. 1-20). New York: Guilford Press.

Shapson, S. M., Wright, E. N., Eason, G. & Fitzgerald, J. (1980). An experimental study of the effects of class size. *American Educational Research Journal, 17*, 144-152.

Sharpley, C. F. & Edgar, E. (1986). Teachers rating vs standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23*, 106-111.

Shavelson, R. J. & Stern, P. (1981). Research on teachers´ pedagogical thoughts, judgements, decisions, and behavior. *Review of Educational Research, 51*, 455-498.

Shuell, T. (2004). Teaching and learning in a classroom context. In D. C. Berliner & R. C. Calfee (Hrsg.), *Handbook of Educational Psychology*. (S. 726-764). New York: Simon & Schuster Macmillan.

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

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

Shulman, L. S. & Sherin, M. G. (2004). Fostering communities of teachers as learners: disciplinary perspectives. *Journal of Curriculum Studies, 36*, 135-140.

Sparfeldt, J. R., Buch, S. R., Schwarz, F., Jachmann, J. & Rost, D. H. (2009). "Rechnen ist langweilig" - Langeweile in Mathematik bei Grundschülern. *Psychologie in Erziehung und Unterricht, 56*, 16-26.

Sparfeldt, J. R., Schilling, S. R., Rost, D. H., Stelzl, I. & Peipert, D. (2005). Leistungsängstlichkeit: Facetten, Fächer, Fachfacetten? Zur Trennbarkeit nach Angstfacette und Inhaltsbereich. *Zeitschrift für Pädagogische Psychologie, 19*, 225-236.

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

Stahl, N. (2007). Kapitel 7: Schülerwahrnehmung und -beurteilung durch Lehrkräfte In H. Ditton (Hrsg.), *Kompetenzaufbau und Laufbahnen im Schulsystem*. Münster: Waxmann.

Stanat, P. & Kunter, M. (2001). Geschlechterunterschiede in Basiskompetenzen. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 - Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. (S. 249-269). Opladen: Leske + Budrich.

Statistisches Bundesamt (Hrsg.). (2009). *Statistisches Jahrbuch 2009 für die Bundesrepublik Deutschland*. Wiesbaden: Statistisches Bundesamt.

Steinkamp, G. (1967). Die Rolle des Volksschullehrers im schulischen Selektionsprozeß. Ergebnisse einer empirisch-soziologischen Untersuchung. In H. D. Ortlieb (Hrsg.), *Hamburger Jahrbuch für Wirtschafts- und Gesellschaftspolitik* (Bd. 12, S. 302-324). Tübingen: Mohr.

Stern, E. (1997). Erwerb mathematischer Kompetenzen: Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 158-170). Weinheim: Beltz.

Stern, E. (1998). *Die Entwicklung des mathematischen Verständnisses im Kindesalter*. Lengerich: Pabst.

Sternberg, R. J. & Horvath, J. A. (1995). A prototype view of expert teaching. *Educational Researcher, 24*, 9-17.

Strahan, D. B. (1989). How experienced and novice teachers frame their views of instruction: An analysis of semantic ordered trees. *Teaching and teacher education, 5*, 53-67.

Stürzer, M. (2003). Geschlechtsspezifische Schulleistungen. In M. Stürzer, H. Roisch, A. Hunze & W. Cornelißen (Hrsg.), *Geschlechterverhältnisse in der Schule*. Opladen: Leske & Budrich.

Südkamp, A., Kaiser, J. & Möller, J. (eingereicht). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis.

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

Swanson, B. B. (1985). Teachers judgments of first-graders' reading enthusiasm. *Reading Research and Instruction, 25*, 41-46.

Tent, L. (2006). Zensuren. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3 ed., S. 873-880). Weinheim, u.a.: Beltz.

Tent, L., Fingerhut, W. & Langfeldt, H.-P. (1976). *Quellen des Lehrerurteils: Untersuchungen zur Aufklärung der Varianz von Schulnoten*. Weinheim: Beltz.

Tent, L. & Stelzl, I. (1993). *Pädagogisch-psychologische Diagnostik*. Göttingen: Hogrefe.

Terhart, E. (2002). *Standards für die Lehrerbildung - Eine Expertise für die Kultusministerkonferenz*. Münster: Institut für Schulpädagogik und Allgemeine Didaktik, Westfälische Wilhelmsuniversität Münster.

Terhart, E. (Hrsg.). (2000). *Perspektiven der Lehrerbildung in Deutschland. Abschlussbericht der von der Kultusministerkonferenz eingesetzten Kommission*. Weinheim: Beltz.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of applied psychology, 4*, 25-29.

Tiedemann, J. (1995). Gender-specific expectancies in elementary school mathematics. *Zeitschrift für Pädagogische Psychologie, 9*, 153-161.

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of childrens' concept of their mathematical ability in elementary school. *Journal of Educational Psychology, 92*, 144-151.

Tiedemann, J. & Billmann-Mahecha, E. (2004). Kontextfaktoren der Schulleistung im Grundschulalter. Ergebnisse aus der Hannoverschen Grundschulstudie. *Zeitschrift für Pädagogische Psychologie, 18*, 113-124.

Tiedemann, J. & Billmann-Mahecha, E. (2007). Zum Einfluss von Migration und Schulklassenzugehörigkeit auf die Übergangsempfehlung für die Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft, 10*, 108-120.

Trautwein, U. & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind. Referenzgruppeneffekte bei Übertrittsentscheidungen. *Zeitschrift für Pädagogische Psychologie, 21*, 119-133.

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

von Maurice, J., Artelt, C., Blossfeld, H.-P., Faust, G., Roßbach, H.-G. & Weinert, S. (2007). Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor- und Grundschulalter: Überblick über die Erhebungen in den Längsschnitten BiKS-3-8 und BiKS-8-12 in den ersten beiden Projektjahren. *PsyDok (Online), 1008*.

Vygotsky, L. S. (1978). Interaction between Learning and Development. In L. S. Vygotsky (Hrsg.), *Mind in society: The development of higher psychological processes*. Cambridge, M.A.: Harvard University Press.

Wahl, D., Weinert, F. E. & Huber, G. L. (1997). *Psychologie für die Schulpraxis*. München: Kösel.

Wayne, A. J. & Youngs, P. (2006). Die Art der Ausbildung von Lehrern und die Lerngewinne ihrer Schüler. *Zeitschrift für Pädagogik, 52*, 71-96.

Weinert, F. E. (1996). 'Der gute Lehrer', 'die gute Lehrerin' im Spiegel der Wissenschaft. Was macht Lehrende wirksam und was führt zu ihrer Wirksamkeit? *Beiträge zur Lehrerbildung, 14*, 141-151.

Weinert, F. E. (1999a). *Concepts of competence*. München: Max-Planck-Institut für Psychologische Forschung.

Weinert, F. E. (1999b). *Konzepte der Kompetenz*. Paris: OECD.

Weinert, F. E. (2001). Concept of competence. A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies*. Göttingen: Hogrefe.

Weinert, F. E. (2002). Vergleichende Leistungsmessungen in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen*. Weinheim: Beltz.

Weinert, F. E., Helmke, A. & Schrader, F.-W. (1992). Research on the model teacher and the teaching model. In F. K. Oser, A. Dick & J.-L. Patry (Hrsg.), *Effective and responsible teaching. The new synthesis* (S. 249-260). San Francisco: Jossey-Baß.

Weinert, F. E. & Schrader, F.-W. (1986). Diagnose des Lehrers als Diagnostiker. In H. Petillon, J. W. L. Wagner & B. Wolf (Hrsg.), *Schülergerechte Diagnose. Theoretische und empirische Beiträge zur Pädagogischen Diagnostik* (S. 11-29). Weinheim, u.a.: Beltz Verlag.

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

Weinert, F. E., Schrader, F.-W. & Helmke, A. (1990b). Unterrichtsexpertise - ein Konzept zur Verringerung der Kluft zwischen zwei theoretischen Paradigmen. In L.-M. Alisch, J. Baumert & K. Beck (Hrsg.), *Professionswissen und Professionalisierung* (S. 173-206). Braunschweig: Technische Universität.

Weiß, R. (1965). *Zensur und Zeugnis (Beiträge zu einer Kritik der Zuverlässigkeit und Zweckmäßigkeit der Ziffernbenotung)*: Haslinger.

Weiß, R. H. (1998). *CFT 20 - Grundintelligenztest Skala 2 mit Wortschatztest (WS) und Zahlenfolgetest (ZF)*. (Bd. 4). Göttingen: Hogrefe.

Weiß, R. H. (2006). *CFT 20-R Grundintelligenztest Skala 2* (Bd. 1). Göttingen: Hogrefe.

Wigfield, A., Eccles, J., Harold, R., Freedman-Doan, C. & Blumenfeld, P. C. (1997). Change in childrens competence belief and subjective task values across the elementary school years: a 3-year study. *Journal of Educational Psychology, 89*, 451-469.

Wigfield, A., Galper, A., Denton, K. & Seefeldt, C. (1999). Teachers' beliefs about former Head Start and non-Head Start first-grade children's motivation, performance, and future educational prospects. *Journal of Educational Psychology, 91*, 98-104.

Wigfield, A. & Harold, R. (1992). Teacher beliefs and children's achievement self-perceptions: A developmental perspective. In D. Schunk & J. Meece (Hrsg.), *Student perceptions in the classroom*. (S. 95-121). Hillsdale, NJ: Erlbaum Associates.

Wild, K.-P. & Rost, D. H. (1995). Klassengröße und Genauigkeit von Schülerbeurteilungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, XXVII*, 78-90.

Wilson, S. M., Shulman, L. S. & Richert, A. E. (1987). 150 different ways of knowing: Representations of knowledge in teaching. In J. Calderhead (Hrsg.), *Exploring teachers' thinking*. London: Cassel.

Wilson, S. M. & Youngs, P. (2005). Research on accountability processes in teacher education. In M. Cochran-Smith & K. M. Zeichner (Hrsg.), *Studying teacher education*. (S. 591-643). Washington.

Winter, F. (2006). Diagnosen im Dienst des Lernens. *Friedrich Jahresheft XXIV*, 22-25.

Wissenschaftlicher Rat der Dudenredaktion (Hrsg.). (1997). *Duden. Das Fremdwörterbuch*. (Bd. 6. Auflage). Mannheim: Bibliographisches Institut & F.A. Brockhaus AG.

Woolfolk, A. (2008). Erfassung von Leistungen und Notengebung. In A. Woolfolk (Hrsg.), *Pädagogische Psychologie* (Bd. 10). München: Pearson Studium.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *Acer ConQuest. Version 2.0*. Melbourne: ACER Press.

Zeuch, W. (1973). Was spricht gegen die Anwendung von Testverfahren? *Die deutsche Schulwarte, 65*, 340-348.

Ziegenspeck, J. W. (1999). *Handbuch Zensur und Zeugnis in der Schule. Historischer Rückblick, allgemeine Problematik, empirische Befunde und bildungspolitische Implikationen*. Bad Heilbrunn: Verlag Julius Klinkhardt.

Zimbardo, P. G. & Gerrig, R. J. (1999). *Psychologie* (7. Auflage). Berlin: Springer.

UNIVERSITY OF BAMBERG PRESS

In der vorliegenden Dissertation wird sich mit der diagnostischen Kompetenz von Grundschullehrern beschäftigt, wobei deren Struktur und ihre Bedingungen im Zentrum der Betrachtungen stehen. Unter diagnostischer Kompetenz wird bei Lehrern deren Fähigkeit verstanden, Schülerleistungen und -merkmale sowie die Schwierigkeit von Aufgaben korrekt einzuschätzen. Diese Fähigkeit gilt als Schlüsselkompetenz in Lehr- und Lernkontexten, da ihr eine hohe Bedeutung für adäquate Unterrichtsgestaltung sowie faire und objektive Beurteilungen beigemessen wird.

Eine Vielzahl an Forschungsbefunden belegt, dass Lehrkräfte zwar im Mittel gute Diagnostiker sind, dass jedoch große interindividuelle Unterschiede bestehen. Dabei waren die bisherigen Untersuchungen überwiegend querschnittlich angelegt und auf einzelne oder wenige Leistungsbereiche bezogen. Aussagen dazu, wie bereichsspezifisch und stabil die Güte von Lehrerurteilen ist, waren somit bislang kaum möglich. Ebenso erfolgte die Suche nach den Ursachen für die Unterschiedlichkeit zwischen Lehrern in aller Regel nur anhand weniger Lehrer- oder Klassenmerkmale, ohne dass jedoch erklärende Variablen gefunden wurden.

An diese Desiderata wird in dieser Arbeit angeknüpft, indem quer- und längsschnittlich und unter Einbezug einer Vielzahl potentiell erklärender Merkmale die Urteilsgüte in mehreren kognitiven und emotional-motivationalen Bereichen erhoben wird. Zentrale Fragestellungen beziehen sich dabei auf strukturelle Aspekte wie jenen der Bereichshomogenität und Stabilität der Urteilsgüte sowie der Reliabilität der Urteilskomponenten. Bedingungen der Urteilsgenauigkeit werden auf Ebene der Lehrer, der Klassen und der individuellen Schüler vermutet und untersucht. Darüber hinaus werden auch Zeugnisnoten als eine besonders bedeutungsvolle Form der Lehrerurteile betrachtet.

ISBN 978-3-86309-056-2
ISSN 1866-8674
20,00 €