Zur Urteilsgenauigkeit von Mathematiklehrkräften: Genauigkeitsbeeinflussende Faktoren, Stabilität und Auswirkungen

Inauguraldissertation
zur Erlangung des Doktorgrades der Philosophie
im Fach Psychologie an der Universität Passau

vorgelegt von
Justine Stang

August 2016

Erstgutachter: Prof. Dr. Detlef Urhahne

Zweitgutachterin: Prof. Dr. Jutta Mägdefrau

Danksagung

Mein besonderer Dank gilt all jenen, die mich auf verschiedene Art und Weise in der Promotionszeit begleitet und zur Entstehung dieser Arbeit beigetragen haben.

Insbesondere möchte ich meinem Doktorvater, Herrn Prof. Dr. Detlef Urhahne, für die Betreuung, Unterstützung und konstruktive Rückmeldung während des Entstehungsprozesses der Arbeit sowie für die verschiedenen Lernerfahrungen in den vergangenen Jahren danken.

Ebenso gilt mein herzlicher Dank Frau Prof. Dr. Jutta Mägdefrau für ihr Interesse an dieser Arbeit und ihre Bereitschaft, sie zu begutachten.

Ein weiterer Dank geht an die Studierenden, die bei der Datenerhebung geholfen haben.

Auch bedanke ich mich bei allen Lehrkräften sowie Schülerinnen und Schülern, die an den Studien teilgenommen haben.

Ein ganz großes, herzliches Dankeschön gilt meiner Familie und meinem Freund. Meinen Eltern für ihre fortwährende Unterstützung und ihren Rückhalt. Meinem Freund für sein offenes Ohr und für die Versorgung mit kulinarischen Köstlichkeiten zur Stärkung zwischendurch.

Zusammenfassung

Der Schulalltag erfordert es, dass Lehrkräfte nicht nur die Leistung, sondern auch andere Schülermerkmale zutreffend beurteilen müssen. Die vorliegende Arbeit, welche drei Studien umfasst, beschäftigt sich mit der Urteilsgenauigkeit von Mathematiklehrkräften zu verschiedenen Schülermerkmalen sowie mit deren Stabilität und Auswirkungen. Die Urteilsgenauigkeit stellt den Kern der diagnostischen Kompetenz von Lehrkräften dar, unter der man die Fähigkeit von Lehrkräften versteht, insbesondere Schülermerkmale zutreffend einschätzen zu können.

Zunächst wird in der Einleitung die Urteilsgenauigkeit von Lehrkräften theoretisch verortet und auf die Bedeutung von Lehrkrafturteilen eingegangen. Des Weiteren werden theoretische Grundlagen sowie empirische Befunde, welche für die einzelnen Fragestellungen der verschiedenen Teilstudien bedeutsam sind, ausführlicher als in diesen dargestellt.

Im Rahmen der ersten Studie wurde die Urteilsgenauigkeit von Mathematiklehrkräften in der Einschätzung von Mathematikleistung, Konzentration, Arbeits- und Sozialverhalten untersucht. Auch wurde überprüft, ob die Urteilsgenauigkeit mit soziodemografischen Lehrkraftmerkmalen zusammenhängt und wodurch Lehrkrafturteile vorhergesagt werden. An der Untersuchung nahmen 357 Realschulfünftklässler sowie deren 17 Mathematiklehrkräfte teil. Die Lernenden bearbeiteten einen Mathematikleistungstest, einen Konzentrationstest und füllten Fragebögen zum Arbeits- und Sozialverhalten aus, während die Lehrkräfte die verschiedenen Schülermerkmale beurteilten. Lehrkräfte schätzten die Rangfolge der Schülerleistung mit mittlerer Genauigkeit ein, wohingegen ihnen die Beurteilung von Konzentration, Arbeits- und Sozialverhalten schwerer fiel. Lehrkräfte überschätzten das Leistungsniveau ihrer Schülerinnen und Schüler. Weder Alter, Geschlecht noch Berufserfahrung der Lehrkräfte hingen mit der Urteilsgenauigkeit zusammen. Lehrkrafturteile zur Schülerleistung und zu nicht-leistungsbezogenen Schülermerkmalen wurden durch sachfremde Urteilsmerkmale verzerrt.

Die zweite Studie baut auf der ersten auf und erweitert diese um einen Messzeitpunkt. Da über Stabilität, genauigkeitsbeeinflussende Faktoren und differenzielle Wirkungen des Lehrkrafturteils zur Schülerleistung relativ wenig bekannt ist, wurden in dieser Studie diese Aspekte analysiert. An der Untersuchung nahmen 294 Realschülerinnen und -schüler sowie deren 17 Mathematiklehrkräfte teil. Die Fünftklässlerinnen und Fünftklässler

bearbeiteten standardisierte Mathematikleistungstests und machten Angaben zum wahrgenommenen Lehrkraftverhalten. Parallel dazu schätzten die Lehrkräfte die Testleistungen ein und beantworteten Fragen zur Bezugsnormorientierung. Die Rangkomponente, Übereinstimmung des Lehrkrafturteils mit den Schülerleistungen, verbesserte sich im Zeitraum eines halben Schuljahres signifikant. Sie korrelierte signifikant positiv mit der kriterialen Bezugsnormorientierung. Die Genauigkeit des Lehrkrafturteils zum Schulhalbjahr, indiziert als Leistungsüber- und -unterschätzung, war prädiktiv für die Schülerleistung zu Schuljahresende. In der Leistung überschätzte Lernende wiesen im Vergleich zu unterschätzten Lernenden den größeren Leistungszuwachs auf und nahmen das Lehrkraftverhalten anders wahr.

Die dritte Studie ist ebenfalls längsschnittlich angelegt und umfasst zwei Messzeitpunkte, weswegen die Stabilitäten von Lehrkrafteinschätzungen und Schülermerkmalen sowie die differenziellen Wirkungen von Lehrkrafteinschätzungen auf verschiedene Schülergruppen untersucht werden konnten. Im Abstand eines Jahres wurden in der dritten und vierten Klasse Daten von 152 Grundschülerinnen und -schülern sowie von deren zehn Mathematiklehrkräften gesammelt. Die Grundschülerinnen und -schüler bearbeiteten jeweils einen standardisierten Mathematikleistungstest und einen Fragebogen zum motivational-affektiven Erleben in der Schule. Zeitgleich schätzten die Lehrkräfte verschiedene Schülermerkmale auf Skalen ein. Die Rangreihungen der Lehrkräfte zu Testleistung, Lernfreude, Schuleinstellung und Anstrengungsbereitschaft verbesserten sich über den einjährigen Untersuchungszeitraum signifikant. Das Lehrkrafturteil zur Testleistung sowie zur Erfolgserwartung und die Schülermerkmale waren stabil. Der Zusammenhang zwischen der Urteilsgenauigkeit und den Schülermerkmalen spiegelte sich in Unterschieden zwischen den Schülergruppen wider: In der vierten Klasse hatten überschätzte Lernende eine höhere Testleistung, Erfolgserwartung und Lernfreude sowie ein höheres Fähigkeitsselbstkonzept und eine niedrigere Leistungsangst als unterschätzte Lernende.

In der Gesamtdiskussion werden die Einzelergebnisse der drei Studien zusammenfassend diskutiert, so dass deutlich wird, an welchen Forschungsergebnissen die Studien ansetzten, welche Ergebnisse bestätigt und welche Erkenntnisse gewonnen werden konnten. Abschließend werden weiterführende Forschungsmöglichkeiten vorgestellt und es wird auf Implikationen für die Praxis verwiesen.

Inhaltsverzeichnis

1. Einleitung und theoretischer Hintergrund

1.1. Einleitung

Zum Schulalltag einer jeden Lehrkraft gehört die Kernaufgabe der Schüler- und Aufgabenbeurteilung. Die Fähigkeit von Lehrkräften, Merkmale von Schülerinnen und Schülern sowie von Aufgaben akkurat einzuschätzen ist als *diagnostische Kompetenz* definiert (Schrader, 2009). Neben der Schülerleistung müssen Lehrkräfte auch in der Lage sein, Merkmale, welche leistungs- wie lernrelevant sind, entsprechend beurteilen zu können (Schrader, 2009). Demnach sollten Lehrkräfte nicht nur die Schülerleistung zutreffend einschätzen, sondern auch Schülermerkmale wie das Fähigkeitsselbstkonzept, die Leistungsangst oder das Arbeitsverhalten, welche in Zusammenhang mit der Schülerleistung stehen (Credé & Kuncel, 2008; Helmke & Schrader, 2010; Schunk, Pintrich & Meece, 2008).

Lehrkraftfähigkeiten sind in den letzten Jahren verstärkt in den Fokus der Bildungsforschung und Bildungspolitik gerückt (Baumert & Kunter, 2006; Schrader, 2009). Einen Auslöser des verstärkten Interesses an den Fähigkeiten von Lehrkräften wie der diagnostischen Kompetenz stellen die Ergebnisse der PISA-Studie aus dem Jahr 2000 dar (Artelt, Stanat, Schneider & Schiefele, 2001). Diese zeigten, dass viele der befragten Hauptschullehrkräfte Schwierigkeiten hatten, förderungsbedürftige Schülerinnen und Schüler als solche zu erkennen. Genauer konnten 90% der schwachen Leserinnen und Leser nicht entsprechend identifiziert werden (Artelt et al., 2001). Als Konsequenz der Studienergebnisse formulierte die Kultusministerkonferenz (KMK) auf der 296. Plenarsitzung am 05. und 06. Dezember 2001 diverse Maßnahmen zur Steigerung der Schulbildung (KMK, 2001). Eine zentrale Rolle dabei spielen die „Maßnahmen zur Verbesserung der Professionalität der Lehrertätigkeit, insbesondere im Hinblick auf diagnostische und methodische Kompetenz als Bestandteil systematischer Schulentwicklung“ (S. 1; KMK, 2001). Seither stellt der Kompetenzbereich *Beurteilen* einen zentralen Aspekt in der bildungspolitischen Debatte dar, in welchem die kompetente und verantwortungsbewusste Schülerbeurteilung betont wird (KMK, 2004). Einhergehend mit dem bildungspolitischen Interesse an der diagnostischen Kompetenz von Lehrkräften, führten die PISA-Ergebnisse auch zu umfassenden, vielschichtigen Forschungsarbeiten (Artelt & Gräsel, 2009; Urhahne et al., 2010).

Das vermehrte Interesse an der diagnostischen Kompetenz von Lehrkräften, sowohl in der Bildungspolitik als auch in der Forschung, erfordert es, diese theoretisch zu verorten, um sie detailliert beschreiben und ihre Bedeutung aufzeigen zu können. In den letzten Jahren wurden verschiedene Modelle aufgestellt, welche die diagnostische Kompetenz

veranschaulichen. Eines dieser Modelle ist das Modell professioneller Handlungskompetenz von Baumert und Kunter (2006). Baumert und Kunter (2011) zufolge befähigt das Zusammenspiel von motivationalen Orientierungen, Werten und Überzeugungen, Selbstregulation sowie von Professionswissen eine Lehrkraft zu professionellem Handeln. Diese vier Kompetenzaspekte setzen sich jeweils aus mehreren Kompetenzbereichen zusammen, welche wiederum in verschiedene Facetten unterteilt werden. In dem heuristischen Rahmenmodell professioneller Handlungskompetenz stellen das Wissen und Können einer Lehrkraft den Kern ihrer Professionalität dar (Baumert & Kunter, 2006, 2011). Das professionelle Wissen und Können einer Lehrkraft wird, in Anlehnung an Shulman (1986), in verschiedene Wissensbereiche untergliedert. Das allgemeine pädagogisch-psychologische Wissen, das Fachwissen sowie das fachdidaktische Wissen stellen die drei zentralen Bereiche dar, welche um das Organisations- und Beratungswissen ergänzt werden (vgl. Abbildung 1.1). Die einzelnen Wissensbereiche unterteilen sich zusätzlich in verschiedene Wissensfacetten. Baumert und Kunter (2006) fassen die folgenden Kompetenzen einer Lehrkraft unter den Bereich des allgemeinen pädagogisch-psychologischen Wissens, welcher für diese Arbeit relevant ist, zusammen: das konzeptuelle bildungswissenschaftliche Grundlagenwissen, das allgemeindidaktische Konzeptions- und Planungswissen, die Unterrichtsführung und Orchestrierung von Lerngelegenheiten sowie die fachübergreifenden Prinzipien des Diagnostizierens, Prüfens und Bewertens. Die diagnostische Kompetenz von Lehrkräften lässt sich demnach in diesen Bereich verorten. Um z. B. die Schülerleistung im Fach Mathematik kompetent diagnostizieren zu können, sind zusätzlich Kompetenzen im fachdidaktischen Wissen, welches u. a. die Facetten Wissen über mathematische Aufgaben und Wissen über das mathematische Denken von Lernenden beinhaltet, erforderlich (Brunner, Anders, Hachfeld & Krauss, 2011). Die Bedeutung der diagnostischen Kompetenz für das professionelle Handeln von Lehrkräften wird im Folgenden dargestellt.

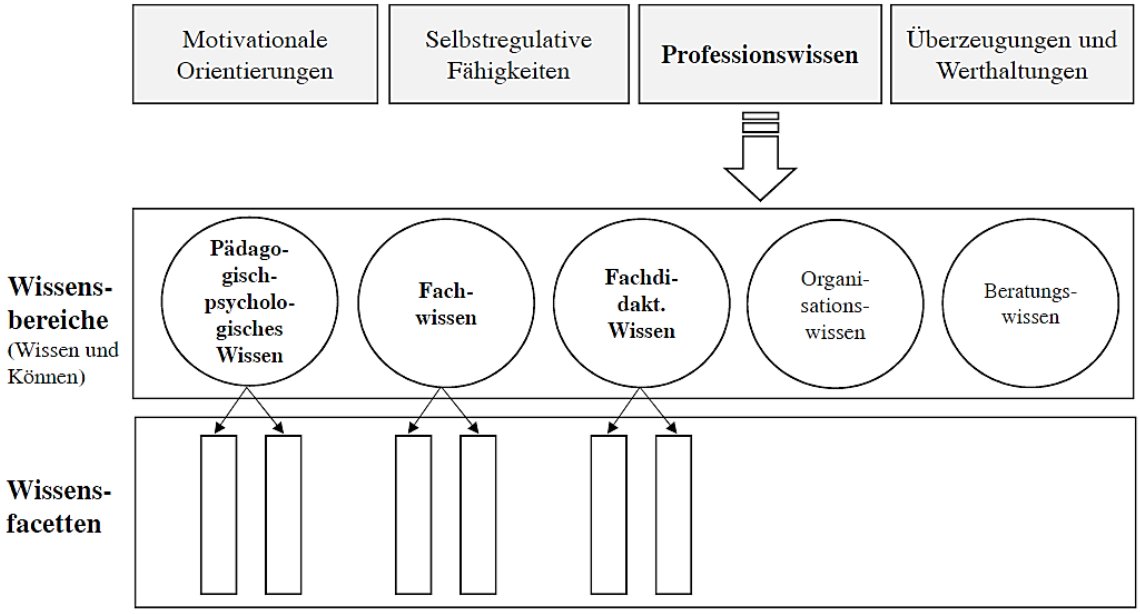

Abbildung 1.1: Modell professioneller Handlungskompetenz (in Anlehnung an Baumert & Kunter, 2006, 2011).

1.2. Bedeutung der diagnostischen Kompetenz

Der diagnostischen Kompetenz von Lehrkräften kommt im schulischen Kontext eine besondere Bedeutung zu. Sie trägt dazu bei, dass Lehrkräfte ihren Beruf erfolgreich ausüben, weswegen sie oftmals als wichtiger Faktor erfolgreichen Unterrichtens genannt wird (Helmke, 2009). Ihre Bedeutung spiegelt sich in den vielseitigen Funktionen von Lehrkrafturteilen wider, welche im Folgenden erläutert werden.

Lehrkrafturteile stellen die Grundlage instruktionaler Entscheidungen dar (Alvidrez & Weinstein, 1999; Hoge, 1983; Hoge & Coladarci, 1989). Dies bedeutet, dass Lehrkrafturteile Einfluss auf die Unterrichtsplanung und -gestaltung nehmen (Südkamp, Kaiser & Möller, 2012). Mit guten diagnostischen Fähigkeiten kann Unterricht demnach so geplant werden, dass eine optimale Passung zwischen den Unterrichtsanforderungen und den individuellen Lernvoraussetzungen der Schülerinnen und Schüler besteht (Rogalla & Vogt, 2008). Die Lehrkrafteinschätzungen bestimmen weiterhin die Auswahl der Aufgabenschwierigkeit sowie die Auswahl der einzusetzenden Unterrichtsstrategien und -materialien (Südkamp et al., 2012). Die Unterrichtsplanung und -gestaltung wird also dadurch mitbestimmt, wie Lehrkräfte die verschiedenen Schülermerkmale wahrnehmen und einstufen (Alvidrez & Weinstein, 1999; Clark & Peterson, 1986). Daher kann die Feststellung

einer Veränderung von Schülermerkmalen zur Reflexion über die bisher eingesetzte Unterrichtsstrategie und zur Adaptation selbiger führen (Shavelson & Stern, 1981).

Lehrkrafturteile sind zudem hilfreich, wenn es darum geht, den Förderbedarf von Schülerinnen und Schülern zu identifizieren (Bailey & Drummond, 2006; Kenny & Chekaluk, 1993). Auf ihrer Grundlage können, in einem weiteren Schritt, Empfehlungen für bestimmte Maßnahmen ausgesprochen werden (Neber, 2004). Die Identifikation des Förderbedarfs beschränkt sich allerdings nicht auf Schülerinnen und Schüler, welche z. B. Anzeichen von Lernschwierigkeiten zeigen, sondern es werden auch Hochbegabte identifiziert und speziellen Förderprogrammen zugewiesen (Gear, 1976; Rost & Hanses, 1997).

Des Weiteren können Lehrkrafturteile die schulische sowie die sich anschließende berufliche Karriere von Schülerinnen und Schülern entscheidend mitbeeinflussen (Begeny, Eckert, Montarello & Storie, 2008; Feinberg & Shapiro, 2003; Harlen, 2005; Trautwein & Baeriswyl, 2007). So haben beim Übergang vom Primar- zum Sekundarbereich die durch Lehrkräfte vergebenen Schulnoten immer noch eine bedeutende Rolle. Im Rahmen der Übergangsempfehlung entscheiden sie darüber, welcher weiterführende Schultyp besucht werden kann (Bos et al., 2004; Pit-ten Cate, Krolak-Schwerdt & Glock, 2016). Lehrkrafturteile in Form von Schulnoten können daher weitreichende Konsequenzen für die berufliche Laufbahn von Schülerinnen und Schülern haben (Feinberg & Shapiro, 2003). Auch bei der Vergabe von Schülerstipendien, welche zum Ziel haben, begabte Schülerinnen und Schüler gesondert zu fördern, wird auf das Lehrkrafturteil zur Schülerleistung, aber auch auf die Beurteilung nicht-kognitiver Schülermerkmale wie des Arbeits- und Sozialverhaltens vertraut. Daraus ergibt sich die Bedeutung der Lehrkrafturteile bei Zugangsentscheidungen. Aufgrund dessen wird oft gefordert, dass Lehrkrafturteile akkurat und fair sein sollen (Brookhart, 1993; Gerber & Semmel, 1984; Helmke, 2009; Hoge, 1983).

Lehrkrafturteile in Form von Noten beeinflussen nicht nur die Laufbahnentscheidungen von Schülerinnen und Schülern (Trautwein & Baeriswyl, 2007), sondern stellen für Schülerinnen und Schüler sowie für deren Eltern auch eine wichtige Informationsquelle dar (Hoge & Coladarci, 1989). Durch die Lehrkrafturteile erhalten sie Feedback über z. B. das aktuelle Leistungsniveau (Hoge & Coladarci, 1989), über die Begabung (Hoge & Cudmore, 1986) oder über Lernschwierigkeiten (Bates & Nettelbeck, 2001).

Auch können Lehrkrafturteile zur Schülerleistung die Erwartungen der Lehrkraft über das Leistungsniveau von Schülerinnen und Schülern beeinflussen (Brophy & Good, 1986).

Die Lehrkrafterwartungen wirken sich wiederrum auf die Schülerleistung aus, aber auch auf die Schülermotivation (Hinnant, O'Brien & Ghazarian, 2009; Jussim, 1989; Kuklinski & Weinstein, 2001). Die sich selbsterfüllende Prophezeiung (Jussim, 1986) erklärt dabei, wie es zum Einfluss der Lehrkrafterwartung auf die Schülermerkmale kommt.

1.3. Genauigkeit von Lehrkrafturteilen

Aufgrund der vielseitigen Bedeutung von Lehrkrafturteilen ist es wichtig, dass sie akkurat ausfallen. Die Genauigkeit, mit der Lehrkräfte verschiedene Schülermerkmale beurteilen, stellt dabei den Kern der diagnostischen Kompetenz dar (Kaiser, Helm, Retelsdorf, Südkamp & Möller, 2012; Pit-ten Cate, Krolak-Schwerdt, Glock & Markova, 2014), weswegen in der Forschungstradition der diagnostischen Kompetenz oftmals der Frage nachgegangen wurde, wie akkurat Lehrkrafturteile zu verschiedenen Schülermerkmalen ausfallen.

Bei der Bestimmung der Urteilsgenauigkeit werden drei Komponenten herangezogen (Schrader & Helmke, 1987): die *Rang*-, die *Niveau*- und die *Differenzierungskomponente*. Schrader und Helmke (1987) leiteten diese drei Komponenten anhand einer Arbeit von Cronbach (1955) ab. In dieser Arbeit kritisierte Cronbach die damalige Vorgehensweise, die Genauigkeit von Urteilen nur anhand globaler Differenzmaße zu bestimmen. Die von Schrader und Helmke (1987) formulierten Komponenten werden im Folgenden genauer beschrieben.

Die Rangkomponente gilt als der zentrale Indikator der Genauigkeit von Lehrkrafturteilen (Südkamp, Möller & Pohlmann, 2008). Sie ist die klassenweise berechnete Korrelation zwischen Lehrkrafturteil und Schülermerkmal und gibt darüber Auskunft, wie gut die Lehrkraft in der Lage ist, eine Rangreihe des einzuschätzenden Schülermerkmals vorherzusagen (Praetorius, Lipowsky & Karst, 2012). Metaanalysen berichten gemittelte Korrelationskoeffizienten von $r \geq .60$ zwischen Lehrkrafturteil und Schülerleistung (Hoge & Coladarci, 1989; Südkamp et al., 2012). Diese lassen darauf schließen, dass Lehrkräfte in der Regel die Schülerleistung, in verschiedenen Domänen, relativ genau einzuschätzen vermögen. Allerdings verweist die Spannweite von $r = -.03$ bis $r = .92$ der in den Studien berichteten Korrelationsmittelwerte darauf hin, dass nicht jede Lehrkraft im Stande ist, die Schülerleistung akkurat einzuschätzen (Hoge & Coladarci, 1989; Südkamp et al., 2012). Im Gegensatz dazu fallen die Korrelationen zwischen Lehrkrafturteil und nicht-kognitiven Schülermerkmalen geringer aus (Spinath, 2005; Urhahne et al., 2010; Urhahne & Zhu,

2015b). Zwar liegen für den nicht-kognitiven Bereich deutlich weniger Studien vor (Urhahne & Zhu, 2015a), doch diese zeigen einheitlich, dass es Lehrkräften schwerer fällt, andere Merkmale als die Schülerleistung einzuschätzen (Bilz, Steger & Fischer, 2016; Karing, Dörfler & Artelt, 2015; Machts, Kaiser, Schmidt & Möller, 2016; Praetorius, Berner, Zeinz, Scheunpflug & Dresel, 2013; Spinath, 2005; Urhahne, Chao, Florineth, Luttenberger & Paechter, 2011; Urhahne et al., 2010; Urhahne & Zhu, 2015b). In diesen Studien wurden das Fähigkeitsselbstkonzept und die Leistungsangst mit am häufigsten untersucht. Für das Fähigkeitsselbstkonzept wurden im Mittel Korrelationen zwischen $r = .19$ und $r = .43$ berichtet, für Leistungsangst deutlich kleinere Korrelationen zwischen $r = .07$ und $r = .15$ (Praetorius et al., 2013; Spinath, 2005; Urhahne et al., 2011; Urhahne et al., 2010).

Die Niveaukomponente ergibt sich aus der Differenz zwischen Lehrkrafturteil und Schülermerkmal. Sie veranschaulicht, inwiefern die Lehrkraft dazu tendiert, das einzuschätzende Schülermerkmal zu über- oder zu unterschätzen (Praetorius et al., 2012). Bei einer perfekten Übereinstimmung zwischen Lehrkrafturteil und Schülermerkmal würde der Wert 0 resultieren. Dementsprechend weisen Werte größer 0 auf eine Über- und Werte kleiner 0 auf eine Unterschätzung des Schülermerkmals hin. Aus der Forschung ist bekannt, dass Lehrkräfte dazu tendieren, die Schülerleistung zu positiv einzuschätzen (Bates & Nettelbeck, 2001; Begeny et al., 2008; Demaray & Elliott, 1998; Feinberg & Shapiro, 2003, 2009; Urhahne et al., 2010).

Die Differenzierungskomponente, welche sich aus der Streuung der Lehrkrafturteile dividiert durch die der Schülerwerte ergibt, bildet ab, ob eine Lehrkraft Merkmalsunterschiede zwischen Lernenden über- oder unterschätzt (Praetorius et al., 2012). Bei exakter Einschätzung der Merkmalsvariation resultiert ein Wert von 1. Werte größer 1 weisen auf eine Über- und Werte kleiner 1 auf eine Unterschätzung der Streuung hin. Die Forschungsergebnisse zur Differenzierungskomponente sind heterogen. Einige Ergebnisse deuten auf eine Überschätzung, eine Unterschätzung oder genaue Einschätzungen der Variabilität verschiedener Schülermerkmale hin (Schrader & Helmke, 1987; Spinath, 2005; Südkamp, et al., 2008; Urhahne, Timm, Zhu & Tang, 2013; Urhahne et al, 2010).

Bei der Analyse der Genauigkeit von Lehrkrafturteilen sollten alle Komponenten berechnet werden, da jede einzelne mit Einschränkungen behaftet ist (Praetorius et al., 2012). Bei der Rangkomponente besteht das Problem, dass sie hoch ausfallen kann, obwohl das

Schülermerkmal systematisch über- oder unterschätzt wurde. Des Weiteren kann die Niveaukomponente Werte nahe des Idealwertes annehmen, auch wenn einige Lernende von ihrer Lehrkraft stark über- oder unterschätzt wurden. Auch bei der Differenzierungskomponente besteht ein Problem. Sie kann ebenfalls noch Werte nahe des Idealwertes annehmen. Hierzu reicht es, wenn die Lehrkraft die Variabilität des Schülermerkmals relativ gut einschätzt, aber die Mittelwerte der Lehrkrafturteile und Schülerangaben nicht nahe beieinander liegen (Praetorius et al., 2012). Auf eine alleinige Analyse nur einer Komponente, meist der Rangkomponente, wie es in früheren Arbeiten der Fall war, sollte daher verzichtet werden (Demaray & Elliott, 1998; Eckert, Dunn, Codding, Begeny & Kleinmann, 2006; Feinberg & Shapiro, 2003; Ohle & McElvany, 2015).

Zudem werden zwei Messmethoden unterschieden, welche zur Erfassung der Urteilsgenauigkeit eingesetzt werden können. Unterschieden wird zwischen der direkten und der indirekten Messung. Bei der direkten Messung stehen den Lehrkräften sowie den Lernenden die gleichen Skalen zur Verfügung, wohingegen sich bei der indirekten Messung die Schätzskalen unterscheiden. Sagt die Lehrkraft vorher, wie viele Aufgaben eines Tests Schülerinnen und Schüler richtig lösen können, handelt es sich um ein direktes Urteil (Demaray & Elliott, 1998; Hoge & Coladarci, 1989; Urhahne et al., 2010). Ein indirektes Urteil liegt hingegen vor, wenn Lehrkräfte ein Schülermerkmal auf einer mehrstufigen Ratingskala einschätzen (DuPaul, Rapport & Perriello, 1991; Hoge & Coladarci, 1989; Urhahne et al., 2010). Die Auswahl der Messmethode determiniert, welche der Komponenten der Urteilsgenauigkeit berechnet werden können. Bei direkter Messung können alle drei Komponenten berechnet, bei indirekter Messung kann hingegen nur die Rangkomponente bestimmt werden. Im Vergleich zur indirekten Messung fällt die direkte Messung etwas genauer aus (Feinberg & Shapiro, 2003, 2009; Hoge & Coladarci, 1989).

1.4. Einflussvariablen der Urteilsgenauigkeit

In den Studien zur Urteilsgenauigkeit von Lehrkräften wurde deutlich, dass Lehrkräfte unterschiedlich genau beurteilen (Schrader & Helmke, 1987; Urhahne et al., 2013). Dies veranlasste dazu, Einflussvariablen der Urteilsgenauigkeit zu suchen. Südkamp et al. (2012) entwickelten das heuristische Modell der Urteilsgenauigkeit von Lehrkräften, in welchem verschiedene Einflussvariablen kategorisiert werden. Das Modell basiert auf empirischen Ergebnissen sowie auf theoretischen Überlegungen. Kern des Modells stellt die Urteilsgenauigkeit dar. Dem Modell entsprechend nehmen Urteilsmerkmale, Testmerk-

male, Lehrkraftmerkmale sowie Schülermerkmale Einfluss auf die Urteilsgenauigkeit (vgl. Abbildung 1.2). Zusätzlich kann das Zusammenspiel von Urteils- und Testmerkmalen sowie von Lehrkraft- und Schülermerkmalen einen Einfluss auf die Urteilsgenauigkeit haben (gestrichelte Linien).

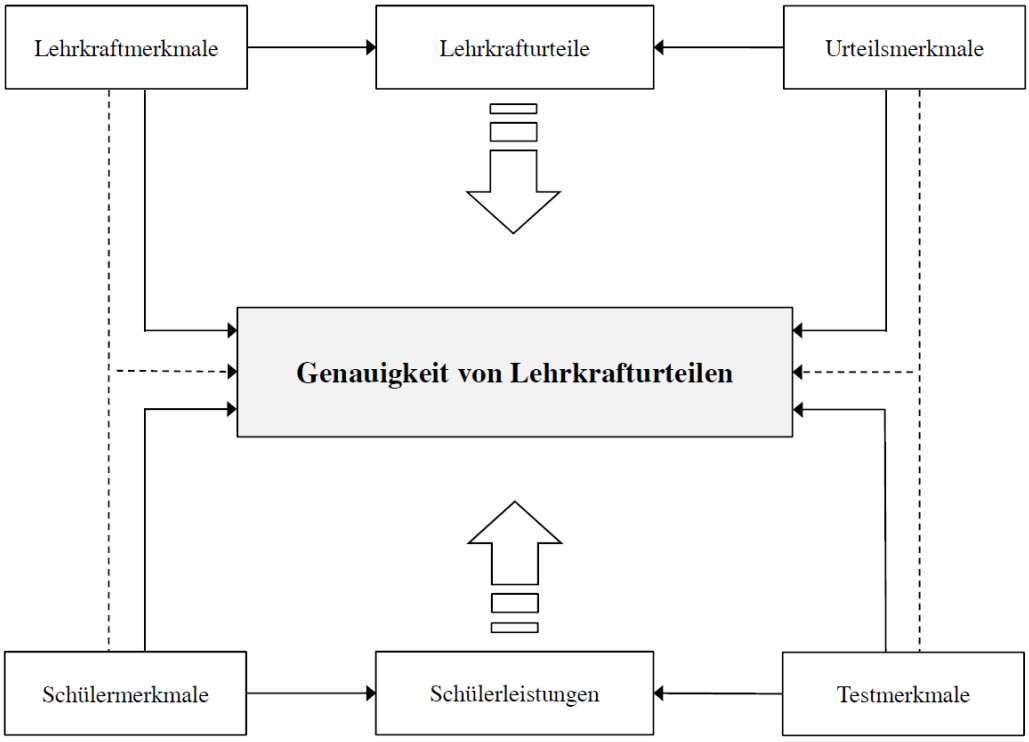

Abbildung 1.2: Modell zu Einflussfaktoren der Urteilsgenauigkeit (in Anlehnung an Südkamp et al., 2012).

Zu den Urteilsmerkmalen, welche als einflussnehmend erachtet werden, zählen z. B. die Messmethode oder die Urteilsspezifität (Südkamp, Kaiser & Möller, 2014). Wie bereits erwähnt, konnten Hoge und Coladarci (1989) sowie Feinberg und Shapiro (2003, 2009) feststellen, dass die Genauigkeit des Lehrkrafturteils bei direkter Messung höher ausfällt als bei indirekter Messung. Zur Urteilsspezifität konnten Hoge und Coladarci (1989) herausfinden, dass spezifischere Urteile, bei denen z. B. Noten zur Einschätzung des Schülermerkmals verwendet wurden, zu einer höheren Urteilsgenauigkeit führen als unspezifische Urteile, welche zur Einschätzung des Schülermerkmals eine Ratingskala heranziehen.

Als ein Testmerkmal wird die Domänenspezifität eines Tests genannt (Südkamp et al., 2014). Bei standardisierten Tests, welche zur Erfassung verschiedener Fähigkeiten ein-

gesetzt werden, wird zwischen spezifischen und unspezifischen Tests unterschieden. So können Tests eingesetzt werden, die entweder spezifische Fähigkeiten, z. B. geometrisches Verständnis im Speziellen, oder aber verschiedene Aspekte von Fähigkeiten, z. B. mathematische Fähigkeiten im Allgemeinen wie Arithmetik, Sachrechnen und Geometrie, messen. Südkamp et al. (2012) konnten in ihrer Metaanalyse allerdings keinen Effekt der Domänenspezifität auf die Urteilsgenauigkeit ausmachen. Ebenso wenig konnten Effekte von Testmerkmalen wie Schulfach oder curriculumbasierte Messverfahren vs. standardisierte Leistungstests auf die Urteilsgenauigkeit gefunden werden (Südkamp et al., 2012).

Bei den Lehrkraftmerkmalen wurde hauptsächlich die Berufserfahrung betrachtet (Demaray & Elliott, 1998; Impara & Plake, 1998; Mulholland & Berliner, 1992). Weitere Lehrkraftmerkmale wie Geschlecht und Alter wurden weniger stark beforscht (Südkamp et al., 2012). Insgesamt betrachtet ist die bestehende Forschungslage zu diesen Lehrkraftmerkmalen sehr heterogen (McElvany et al., 2009; Schrader, 1989; Südkamp et al., 2012). So konnte in einigen Studien ein Zusammenhang zwischen der Berufserfahrung von Lehrkräften und der Urteilsgenauigkeit gefunden werden, in anderen wiederum nicht (McElvany et al., 2009; Praetorius, Karst, Dickhäuser & Lipowsky, 2011; Schrader, 1989). Zum Geschlecht und zum Alter von Lehrkräften können keine Schlussfolgerungen gezogen werden, da diese Merkmale erst in wenigen Arbeiten untersucht wurden (Dicke, Lüdtke, Trautwein, Nagy & Nagy, 2012; Südkamp et al., 2012). Neben soziodemografischen Lehrkraftmerkmalen wurden psychologische Lehrkraftmerkmale bislang vernachlässigt, ebenso wie Urteilsfehler oder Stereotype, welche ebenfalls einen Einfluss haben könnten (Südkamp et al., 2012). In einer der wenigen Studien, die u. a. auf psychologische Lehrkraftmerkmale fokussierte, konnte herausgefunden werden, dass die Motivation und die Einstellung zum Diagnostizieren sowie das Wissen über Diagnostik positive Prädiktoren der mittels eines Testszenarios gemessenen diagnostischen Kompetenz von Lehrkräften zum Lernverhalten von Schülerinnen und Schülern sind (Klug, Bruder & Schmitz, 2016).

Aufseiten der Schülermerkmale wurden verschiedene Aspekte untersucht, welche die Urteilsgenauigkeit von Lehrkräften beeinflussen können. Untersucht und diskutiert wurde vor allem das Geschlecht der Lernenden, welches allerdings keinen Einfluss zu haben scheint (Demaray & Elliott, 1998; Helwig, Anderson & Tindal, 2001; Hoge & Butcher, 1984). Des Weiteren wurden Schülermerkmale analysiert, welche als urteilsirrelevant

eingestuft werden und das Lehrkrafturteil verzerren können. Als urteilsirrelevant gelten solche Schülermerkmale, welche in keinem direkten Zusammenhang mit der Schülerleistung stehen. So konnten Schrader und Helmke (1990) feststellen, dass Lehrkrafturteile zur Schülerleistung von Schülermerkmalen wie der Intelligenz und dem Fähigkeitsselbstkonzept abhängen. Auch zeigte sich in einer weiteren Studie, dass die Beurteilung der Schülerleistung durch Informationen zur Deutschleistung und Intelligenz systematisch beeinflusst wird (Kaiser, Möller, Helm & Kunter, 2015).

Beim Zusammenspiel von Urteils- und Testmerkmalen diskutieren Südkamp et al. (2014) u. a. den Einfluss der Zeitspanne zwischen der Bearbeitung des Leistungstests durch die Schülerinnen und Schüler und der vorgenommenen Einschätzungen durch die Lehrkräfte auf die Urteilsgenauigkeit. Angenommen wird, dass es zu einer höheren Urteilsgenauigkeit kommt, wenn Lehrkräfte zeitgleich mit der Testung der Schülerinnen und Schüler die Merkmale beurteilen statt nacheinander (Südkamp et al., 2014). In ihrer Metaanalyse unterscheiden Südkamp et al. (2012) zwischen gleichzeitiger Bearbeitung, Bearbeitung des Testes vor der Lehrkrafteinschätzung und Lehrkrafteinschätzung vor Bearbeitung des Testes durch die Schülerinnen und Schüler. Allerdings konnten Südkamp et al. (2012) keinen statistisch signifikanten Effekt der Zeitspanne auf die Urteilsgenauigkeit ausmachen.

Beim Zusammenspiel von Lehrkraft- und Schülermerkmalen nennen Südkamp et al. (2014) verschiedene Merkmale, welche in Kombination einen Einfluss auf die Urteilsgenauigkeit haben können. Hierzu gehören z. B. die Ethnizität sowie der sozioökonomische Status von Lehrkräften und von Schülerinnen und Schülern (Alexander, Entwisle & Thompson, 1987). So konnten Alexander et al. (1987) herausfinden, dass Schülerinnen und Schüler, die einer ethnischen Minorität angehörten und einen geringen sozioökonomischen Status hatten, schlechter beurteilt wurden als Schülerinnen und Schüler, die einer ethnischen Majorität angehörten und einen hohen sozioökonomischen Status hatten. Insbesondere fiel die Beurteilung so aus, wenn die Lehrkraft einen hohen sozioökonomischen Status hatte.

Eine weitere Erklärungsmöglichkeit, warum manche Urteile genauer ausfallen als andere, liefert das Realistic Accuracy Model von Funder (1995, 2012). In diesem Modell wird die Genese eines akkuraten Urteils beschrieben (vgl. Abbildung 1.3), wodurch die Komplexität des Urteilsprozesses bzw. die Komplexität akkurater Urteile deutlich wird.

Grundlegende Annahme ist, dass das Merkmal, welches beurteilt werden soll, beobachtbar ist. Insgesamt müssen vier essentielle Schritte erfolgreich vollzogen werden, um zu akkuraten Urteilen gelangen zu können: Zunächst müssen relevante Hinweisreize für das zu beurteilende Merkmal ausgesendet werden, welche für den Beurteiler verfügbar sind. Der Beurteiler muss diese Reize wiederum angemessen wahrnehmen und zur Urteilsgenerierung nutzen. Ist eine der genannten Bedingungen nicht erfüllt, so kann die Lehrkraft keine akkurate Einschätzung vornehmen.

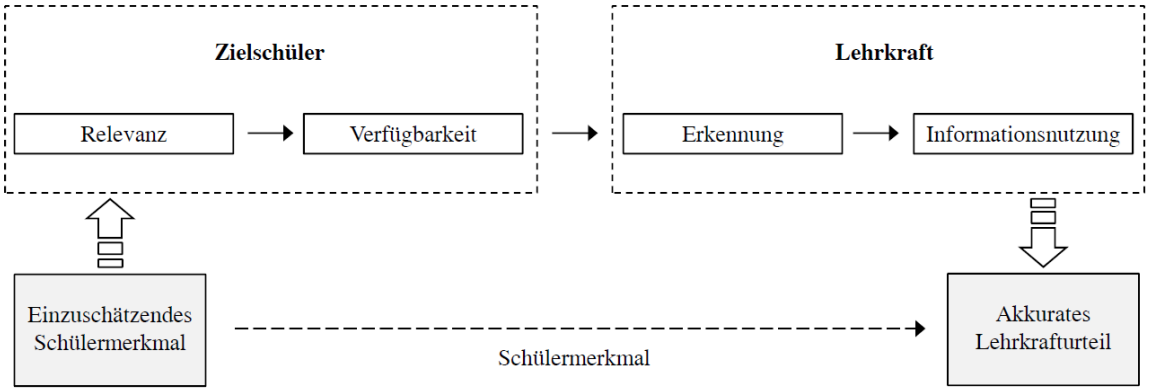

Abbildung 1.3: Genese eines akkuraten Urteils (in Anlehnung an Funder, 1995, 2012).

Anhand des Modells kann erläutert werden, warum Lehrkrafturteile zur Testleistung genauer ausfallen als die zu nicht-kognitiven Schülermerkmalen (Spinath, 2005; Urhahne et al., 2010; Urhahne & Zhu, 2015b). Bei der Leistungsbeurteilung stehen den Lehrkräften mehr Informationen zur Verfügung. Das Leistungsniveau ist direkt, durch Tests oder Hausaufgaben, feststellbar. Demgegenüber stehen motivational-affektive Schülermerkmale wie z. B. Leistungsangst, welche für Lehrkräfte nicht direkt erfassbar und nicht durch Schultests quantifizierbar sind. Bei diesen Merkmalen besteht zudem die Problematik, dass sich das Erleben der Schülerinnen und Schüler nicht in direktem, beobachtbarem Verhalten zeigen muss (Kenny & West, 2010), weswegen die relevanten Hinweisreize für Lehrkräfte nicht verfügbar sind und somit nicht zur Urteilsgenerierung herangezogen werden können.

1.5. Variabilität und Veränderung der Urteilsgenauigkeit

Um die Variabilität und Veränderung der Genauigkeit von Lehrkrafturteilen über die Zeit näher beschreiben zu können, ist zunächst eine Klärung des Kompetenzbegriffs notwendig. Zum Kompetenzbegriff existieren mehrere Annahmen, da er nicht einheitlich definiert ist

(Jürgens & Lissmann, 2015). Eine gängige, weit verbreitete Definition des Begriffs stammt von Weinert. Nach Weinert (2014) handelt es sich bei Kompetenzen um „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“ (S. 27f.).

Zwei zentrale Ansichten, welche auf der Auslegung des Kompetenzbegriffes basieren, sind für diese Arbeit interessant. So wird einerseits angenommen, dass es sich bei der Kompetenz um ein stabiles Fähigkeitskonstrukt, demnach um ein Persönlichkeitsmerkmal, handelt. Das Verständnis von Kompetenz als eine zeitstabile Disposition kann auf eine Definition von Chomsky (1968) zurückgeführt werden. Der Annahme entsprechend, müsste eine Lehrkraft immer gut diagnostizieren, ungeachtet des Erhebungszeitpunktes (Lorenz & Artelt, 2009). Andererseits steht dem die Auffassung gegenüber, dass Kompetenzen durch Erfahrung erlernbar, trainierbar sowie veränderbar sind. Hascher (2008) nimmt an, dass dies auf die diagnostische Kompetenz von Lehrkräften zutrifft. Dieser Annahme entsprechend könnte eine Lehrkraft ihre diagnostische Kompetenz durch Erfahrung in relevanten Situationen und durch Training steigern (Hascher, 2008).

Erkenntnisse, die eine der beiden Annahmen stützen, sind rar. Dies liegt daran, dass die meisten Arbeiten zur diagnostischen Kompetenz rein querschnittlich angelegt waren (Oerke, McElvany, Ohle, Ullrich & Horz, 2015, Südkamp et al., 2012). Aus diesen Arbeiten lässt sich daher nur ableiten, wie genau das Lehrkrafturteil zu einem bestimmten Zeitpunkt ist, jedoch nicht, wie stabil es ist oder ob es sich mit der Zeit verändert. Bislang liegen nur vereinzelt Studien vor, die die Stabilität von Lehrkrafturteilen betrachtet haben. In einer Arbeit von Spinath (2005) erwies sich die Genauigkeit der Lehrkrafturteile zur Fähigkeitsselbstwahrnehmung, Intelligenz, Lernmotivation und Leistungsangst als zeitlich stabil. In dieser Arbeit wurde die Urteilsgenauigkeit innerhalb von sechs Monaten zweimal erfasst. Die berichteten diachronen Zusammenhänge lagen im Bereich von r_{tt} = .50 bis .72. Bei Lorenz und Artelt (2009) zeigten sich Stabilitäten in Höhe von r_{tt} = .38 bis .58 für die Einschätzungen von Arithmetik, Wortschatz und Textverstehen. In zwei experimentellen Arbeiten von Südkamp et al. (2008) wurde nicht der diachrone Zusammenhang betrachtet, sondern es wurden die Rang- und die Differenzierungskomponente per Fisher *Z*-Test bzw. *t*-Test auf Unterschiede zwischen den Testzeitpunkten geprüft. Die Probanden schätzten in

je zwei Durchgängen die Schülerleistung der virtuellen Schülerinnen und Schüler ein. In der ersten Arbeit wurde die Schülerleistung experimentell variiert, in der zweiten Arbeit hingegen nicht. In beiden Arbeiten bestanden zwischen den Rangkomponenten keine signifikanten Unterschiede. Allerdings fiel in der ersten Arbeit die Rangkomponente im zweiten Durchgang höher aus als im ersten Durchgang. Beim Vergleich der Differenzierungskomponenten zeigte sich nur in der ersten Arbeit in signifikanter Unterschied. Im ersten Durchgang wurde die Streuung der Schülerleistung von den Probanden signifikant stärker unterschätzt als im zweiten Durchgang.

1.6. Auswirkungen von Lehrkrafturteilen

Die Arbeiten zum Pygmalioneffekt von Rosenthal und Jacobson (1966, 1968) stellen den Ausgangspunkt der pädagogisch-psychologischen Forschung zu Lehrkrafterwartungen dar. Überprüft wurde, ob die Lehrkrafterwartungen zur Intelligenzentwicklung der Schülerinnen und Schüler deren tatsächliche kognitive Entwicklung beeinflussen können. Hierzu wurden den Lehrkräften Namen von Schülerinnen und Schülern mitgeteilt, welche sich im kommenden Jahr besonders gut im kognitiven Bereich entwickeln sollten. Allerdings wurden diese Kinder zufällig ausgewählt. Im Vergleich zur Kontrollgruppe wiesen die Kinder, an die die Lehrkräfte aufgrund der Vorhersage hohe Erwartungen stellten, zu Schuljahresende einen signifikant stärkeren Zuwachs in der kognitiven Leistungsfähigkeit auf (Rosenthal & Jacobson, 1971). Dieses Phänomen ist als die sich selbsterfüllende Prophezeiung bekannt (Merton, 1948).

Beim Vergleich der Konzepte Lehrkrafterwartung und Lehrkrafturteil fallen Unterschiede und Gemeinsamkeiten auf. Ein Unterschied der zwischen den beiden Konzepten besteht, liegt darin, dass sich die Lehrkrafterwartung auf die zukünftige Schülerleistung, wohingegen sich das Lehrkrafturteil auf die aktuelle Schülerleistung bezieht (Brophy, 1998; Jussim, 1989; Rubie-Davies, 2010; Südkamp et al., 2012). Gemeinsam ist den beiden Konzepten jedoch, dass sie auf eine ähnliche Art und Weise gemessen werden. Sowohl in der Forschung zu Lehrkrafterwartungen als auch in der Forschung zu Lehrkrafturteilen werden Lehrkräfte gebeten, die verschiedenen Merkmale ihrer Schülerinnen und Schüler einzuschätzen (Jussim, 1989; Rubie-Davies, 2010; Urhahne et al., 2010). Da der Unterschied nicht allzu groß ist, erscheint es plausibel, die bestehenden Forschungsergebnisse und die bestehenden Modelle zu Lehrkrafterwartungen auch zur Erklärung und Vorhersage der Auswirkungen von Lehrkrafturteilen heranzuziehen (Urhahne, 2015).

Lehrkrafturteile können die zukünftige Schülerleistung, aber auch die Schülermotivation beeinflussen (Hinnant et al., 2009; Jussim, 1989; Kuklinski & Weinstein, 2001; Peterson, Rubie-Davies, Osborne & Sibley, 2016; Rubie-Davies, 2008; Rubie-Davies et al., 2014). Damit es zu einem Einfluss kommt, müssen die folgenden Schritte durchlaufen werden: Lehrkräfte bilden sich ein Urteil über das Leistungsvermögen ihrer Schülerinnen und Schüler. Dieses Urteil kann sich in eine Lehrkrafterwartung wandeln, welche mit kongruentem Verhalten aufseiten der Lehrkräfte einhergeht. Das Lehrkraftverhalten beeinflusst wiederum das Schülerverhalten, so dass sich die Schülerinnen und Schüler entsprechend verhalten. Dadurch wird, in einem nächsten Schritt, das Lehrkrafturteil bzw. die Lehrkrafterwartung bestätigt. Dieser Prozess ist bekannt als die sich selbst erfüllende Prophezeiung, welche von Jussim (1986) beschrieben wurde (vgl. Abbildung 1.4).

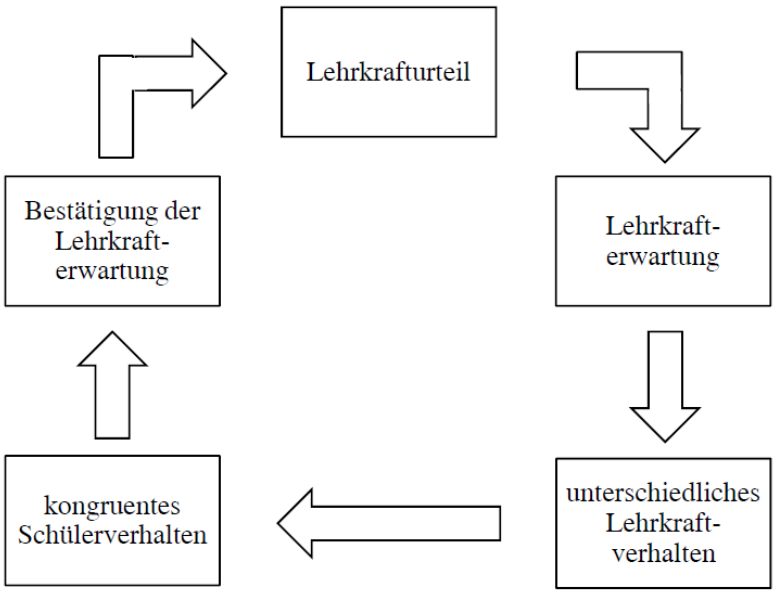

Abbildung 1.4: Sich selbst erfüllende Prophezeiung (in Anlehnung an Jussim, 1986).

Der Einfluss auf die zukünftige Schülerleistung konnte in vielen Studien aufgezeigt werden (Cooper, Findley & Good, 1982; Kuklinski & Weinstein, 2001; Peterson et al., 2016; Rubie-Davies et al., 2014). So fanden Kuklinski und Weinstein (2001) heraus, dass die Lehrkrafterwartung einen direkten Einfluss auf die Schülerleistung ausübt. Dieses Ergebnis steht im Einklang mit dem von Cooper et al. (1982), welche ebenfalls einen direkten Einfluss ausmachen konnten. De Boer, Bosker und van der Werf (2010) zeigten,

genau wie Babad, Inbar und Rosenthal (1982), dass eine negative Lehrkrafterwartung die Schülerleistung negativ beeinflusst, wohingegen eine positive Lehrkrafterwartung die Schülerleistung positiv beeinflusst. Genauer gesagt kann eine Leistungsüberschätzung, welche eine hohe Erwartung widerspiegelt, leistungsförderlich und eine Leistungsunterschätzung, welche eine niedrige Erwartung widerspiegelt, leistungshinderlich sein.

Im Gegensatz zum Einfluss auf die Schülerleistung, wurde der Einfluss auf nicht-kognitive Schülermerkmale weniger stark beforscht. Rubie-Davies (2010) konnte allerdings herausfinden, dass Schülerinnen und Schüler einer Klasse, an die hohe Erwartungen gestellt werden, von der Lehrkraft als motivierter, engagierter und interessierter wahrgenommen werden. Zusätzlich zeigte sich in diversen Arbeiten (Urhahne et al., 2011; Urhahne et al., 2010; Zhu & Urhahne, 2015), dass zwischen Schülerinnen und Schülern, deren Leistung über- oder unterschätzt wurde, Unterschiede bestehen. Signifikante Unterschiede bestanden in den Merkmalen Erfolgserwartung, Fähigkeitsselbstkonzept und Leistungsangst, wobei überschätzte im Vergleich zu unterschätzten Lernenden die signifikant bessere Ausprägung der Merkmale aufwiesen. Die Ergebnisse verweisen darauf, dass sich auch im motivational-affektiven Bereich hohe Lehrkrafterwartungen positiv auf Schülermerkmale wie Motivation, Interesse und Leistungsangst auszuwirken scheinen (Jussim, 1989; Jussim & Harber, 2005).

Eine entscheidende Rolle bei der sich selbst erfüllenden Prophezeiung spielt das Lehrkraftverhalten (Jussim, 1986). Urhahne (2015) stellte fest, dass das erlebte Lehrkraftverhalten die Beziehung zwischen Leistungsurteil und motivational-affektiven Schülermerkmalen mediiert. Des Weiteren zeigte sich, dass sich Lehrkräfte den Schülerinnen und Schülern gegenüber, an die sie hohe Erwartungen stellen, herzlicher und unterstützender verhalten (Harris, Rosenthal & Snodgrass, 1986; Jussim & Harber, 2005). Zudem geben sie ihnen mehr Feedback und stellen mehr herausfordernde Aufgaben für sie bereit (Brophy, 1983; Jussim & Harber, 2005). Auch werden Lehrkräfte von in der Leistung unterschätzten Schülerinnen und Schülern als weniger zugänglich wahrgenommen (Urhahne, 2015).

1.7. Ziele dieser Arbeit

Anhand des dargestellten theoretischen Hintergrundes zur Akkuratheit von Lehrkrafturteilen wird deutlich, dass zahlreiche Fragen noch gänzlich unbeantwortet oder nur unzureichend beantwortet sind. Hierzu gehören die folgenden Fragen: Wie gut können Lehrkräfte nicht-kognitive Schülermerkmale einschätzen? Wodurch werden Lehrkraft-

urteile beeinflusst? Verändert sich die Urteilsgenauigkeit über die Zeit? Unterscheiden sich über- und unterschätzte Lernende und wie wirkt sich die Urteilsgenauigkeit auf die verschiedenen Schülergruppen aus? Aus diesen offenen Fragen leitet sich die Notwendigkeit ab, weiterführende Arbeiten durchzuführen, welche auf die weniger stark beforschten Aspekte fokussieren, um zu neuen Erkenntnissen zu gelangen.

Im Fokus der ersten Studie stand daher die Frage, wie gut Realschullehrkräfte neben der Mathematikleistung, auch die Schülermerkmale Konzentration, Arbeits- und Sozialverhalten einschätzen können. Des Weiteren wurden Einflussfaktoren der Urteilsgenauigkeit betrachtet. Genauer wurde der Frage nachgegangen, ob die Lehrkrafturteile zur Schülerleistung und zu den anderen Schülermerkmalen mit anderen, von der Lehrkraft vorgenommenen Merkmalseinschätzungen oder mit anderen Schülermerkmalen zusammenhängen. Auch wurde aufseiten der Lehrkräfte nach Moderatoren der Urteilsgenauigkeit gesucht.

An die erste Studie schließt sich die zweite an. Das längsschnittliche Vorgehen erlaubt es, zusätzlich zur Akkuratheit von Lehrkrafturteilen nach Veränderungen der Urteilsgenauigkeit über die Zeit und nach Auswirkungen der Urteilsgenauigkeit zu fragen. Dementsprechend stand die Klärung folgender Fragen im Fokus der zweiten Studie: Verändert sich die Urteilsgenauigkeit über die Zeit? Wirkt sich die Urteilsgenauigkeit unterschiedlich auf verschiedene Schülergruppen aus? Nehmen diese Schülergruppen, der sich selbst erfüllenden Prophezeiung entsprechend, das Lehrkraftverhalten unterschiedlich wahr? Zusätzlich wurde in der zweiten Studie untersucht, ob die Bezugsnormorientierung der Lehrkräfte mit den Komponenten der Urteilsgenauigkeit zusammenhängt.

Abschließend wurde in der dritten Studie überprüft, wie gut Grundschullehrkräfte die Schülerleistung und motivational-affektive Schülermerkmale einschätzen können. Neben einer relativ genauen Leistungsbeurteilung ist es für Lehrkräfte ebenso wichtig motivational-affektive Schülermerkmale treffend einschätzen zu können. Dies liegt darin begründet, in Lage zu sein, möglichen Leistungsdefiziten, welche Folge eines negativen motivational-affektiven Erlebens sind, präventiv entgegenwirken zu können. Aufgrund des längsschnittlichen Vorgehens konnte auch in dieser Studie der Frage nachgegangen werden, inwiefern sich die Genauigkeit der Lehrkrafturteile über die Zeit verändert. Ebenfalls im Fokus standen Unterschiede zwischen über- und unterschätzten Schülerinnen und Schülern. Zusätzlich wurde betrachtet, ob sich die motivational-affektiven Schülermerkmale über den Zeitraum eines Jahres verändern.

2. Studie 1: Wie gut schätzen Lehrkräfte Leistung, Arbeits- und Sozialverhalten ihrer Schülerinnen und Schüler ein? Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften

Stang, J. & Urhahne, D. (2016a). Wie gut schätzen Lehrkräfte Leistung, Konzentration, Arbeits- und Sozialverhalten ihrer Schülerinnen und Schüler ein? Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. *Psychologie in Erziehung und Unterricht, 63*, 204–219.

Die erste Teilarbeit ist in der Fachzeitschrift *Psychologie in Erziehung und Unterricht* erschienen. Sie liegt als Druck- und als Online-Ausgabe, doi: 10.2378/peu2016.art18d, vor. Daher befindet sich auf der nachfolgenden Seite nur eine kurze Zusammenfassung des Studieninhaltes.

2.1. Zusammenfassung Studie 1

Der Schulalltag erfordert es, dass Lehrkräfte nicht nur die Leistung der Lernenden, sondern auch deren Arbeits- und Sozialverhalten beurteilen müssen. Bemerkungen oder Bewertungen über Konzentration, Mitarbeit und Verhalten können einerseits in Zeugnisse aufgenommen werden, andererseits sind sie fester Bestandteil in Empfehlungsschreiben für Schülerstipendien. In der ersten Studie wurde daher neben der Genauigkeit der Einschätzung der Schülerleistung auch untersucht, wie gut Lehrkräfte imstande sind das Konzentrationsvermögen, Arbeits- und Sozialverhalten der Lernenden einzuschätzen. Des Weiteren wurde der Frage nachgegangen, welche Faktoren in einem Zusammenhang mit der Urteilsgenauigkeit stehen.

An der Untersuchung, die an Realschulen durchgeführt wurde, nahmen insgesamt 357 Fünftklässler sowie deren 17 Mathematiklehrkräfte teil. Die Lernenden bearbeiteten einen standardisierten Mathematikleistungs- und Konzentrationstest und beantworteten Fragen zum Arbeits- und Sozialverhalten. Währenddessen machten Lehrkräfte soziodemografische Angaben und schätzten für jeden Lernenden die Merkmale ein.

Es zeigte sich, dass die Lehrkräfte die Schülerleistung mit mittlerer Genauigkeit vorhersagen konnten. Die Einschätzung des Konzentrationsvermögens, Arbeits- und Sozialverhaltens fiel ihnen deutlich schwerer. Zudem überschätzten die Lehrkräfte die Schülerleistung. Zwischen den Lehrkräften ergaben sich starke interindividuelle Unterschiede in der Urteilsgenauigkeit. Die soziodemografischen Lehrkraftmerkmale Alter, Geschlecht und Berufserfahrung erwiesen sich als unabhängig von der Urteilsgenauigkeit. Allerdings zeigte sich, dass die Lehrkrafturteile zur Schülerleistung und zu den nichtleistungsbezogenen Schülermerkmalen durch weitere Lehrkrafteinschätzungen und durch urteilsirrelevante Schülermerkmale verzerrt waren.

3. Studie 2: Stabilität, Bezugsnormorientierung und Auswirkungen der Urteilsgenauigkeit

Stang, J. & Urhahne, D. (2016b). Stabilität, Bezugsnormorientierung und Auswirkungen der Urteilsgenauigkeit. *Zeitschrift für Pädagogische Psychologie, 30*, 251–262.

Die zweite Teilarbeit ist in der Fachzeitschrift *Zeitschrift für Pädagogische Psychologie* erschienen. Sie liegt als Druck- und als Online-Ausgabe, doi: 10.1024/1010-0652/a000190, vor. Daher befindet sich auf der nachfolgenden Seite nur eine kurze Zusammenfassung des Studieninhaltes.

3.1. Zusammenfassung Studie 2

Die zweite Studie knüpft an die erste Studie an und erweitert diese um einen zweiten Messzeitpunkt. Im Fokus dieser Arbeit stand die Urteilsgenauigkeit zur Schülerleistung. Aufgrund der Hinzunahme des zweiten Messzeitpunktes konnten bisher kaum beforschte Aspekte der Urteilsgenauigkeit genauer untersucht werden. Daher wurden die Veränderung der Urteilsgenauigkeit über die Zeit und der Zusammenhang zwischen der Urteilsgenauigkeit und der Leistungsentwicklung der Lernenden analysiert. Auch wurde der Frage nachgegangen, ob die Bezugsnormorientierung der Lehrkräfte in einem Zusammenhang mit der Urteilsgenauigkeit steht.

Für beide Messzeitpunkte lagen übereinstimmende Daten von 294 Fünftklässlerinnen und Fünftklässlern vor. Zum zweiten Messzeitpunkt bearbeiteten die Lernenden wieder den standardisierten Mathematikleistungstest. Zusätzlich machten sie Angaben zum wahrgenommenen Lehrkraftverhalten. Parallel dazu beantworteten Lehrkräfte erneut soziodemografische Fragen sowie Fragen zur Bezugsnormorientierung und schätzten für jeden Lernenden die Schülerleistung ein.

Zum zweiten Messzeitpunk, am Ende des Schuljahres, konnten die Lehrkräfte genauere Leistungsvorhersagen treffen. Die Rangkomponente als Übereinstimmung des Lehrkrafturteils mit den ermittelten Testleistungen der Schülerinnen und Schüler stieg innerhalb des Schulhalbjahres signifikant an. Am höchsten war die Rangkomponente ausgeprägt, wenn Lehrkräfte eine kriteriale Bezugsnormorientierung verfolgten. Die Urteilsgenauigkeit des ersten Messzeitpunktes am Ende des ersten Schulhalbjahres, indiziert als Leistungsüber- und Leistungsunterschätzung, war prädiktiv für die Testleistungen der Lernenden zu Schuljahresende. Die Lernenden, die in der Leistung überschätzt wurden, wiesen im Vergleich zu jenen, die unterschätzt wurden, einen größeren Leistungszuwachs auf. Zusätzlich nahmen sie die Lehrkraft als unterstützender und zugänglicher wahr und waren der Meinung, gerechtere Noten zu erhalten.

4. Studie 3: Stabilität und Auswirkungen der Urteilsgenauigkeit von Grundschullehrkräften

4.1. Theoretischer Hintergrund

Das Forschungsgebiet der diagnostischen Kompetenz von Lehrkräften hat seit den Untersuchungen im Rahmen von PISA 2000 in der Bundesrepublik Deutschland einen starken Aufschwung erfahren. In den Testergebnissen von PISA zeigte sich, dass die meisten schwachen Leserinnen und Leser von den Lehrkräften nicht als solche identifiziert werden konnten und damit Gefahr liefen, nicht in ausreichendem Maße gefördert zu werden (Artelt, Stanat, Schneider & Schiefele, 2001). Seitdem wurden zahlreiche Projekte initiiert, um mehr über die Genauigkeit der Einschätzungen von Lehrkräften zu erfahren (Artelt & Gräsel, 2009). Gleichwohl bestehen nach wie vor Forschungslücken und es ergeben sich Fragen, zu denen noch keine evidenzbasierten Antworten möglich sind.

So wurde in den meisten Studien die diagnostische Kompetenz von Lehrkräften nur zu einem Messzeitpunkt erfasst (Oerke, McElvany, Ohle, Ullrich & Horz, 2015). Daraus lässt sich ableiten, wie genau das Lehrkrafturteil zu einem bestimmten Zeitpunkt ist. Es lässt sich aber keine Aussage darüber treffen, ob sich das Lehrkrafturteil mit der Zeit verändert und ob es an Genauigkeit gewinnt, wenn die Lehrkraft mit der Klasse besser vertraut ist. Deshalb ist es sinnvoll, die Stabilität und Akkuratheit des Lehrkrafturteils auf der Grundlage eines Längsschnittdesigns zu bemessen.

Des Weiteren werden in Studien zur diagnostischen Kompetenz in erster Linie Lehrkrafturteile über die kognitiven Fähigkeiten von Schülerinnen und Schülern erhoben. Es liegen bereits zwei Metaanalysen zum Zusammenhang von Lehrkrafturteil und Schülerleistung vor (Hoge & Coladarci, 1989; Südkamp, Kaiser & Möller, 2012). Wiederholt konnte dadurch gezeigt werden, dass Lehrkräfte gute Diagnostiker von Schülerleistungen sind und zu weitgehend akkuraten Urteilen gelangen. Allerdings sind auch motivationale und emotionale Faktoren in der Schule von Bedeutung. Mit welcher Lernmotivation und Anstrengungsbereitschaft Schülerinnen und Schüler ihren Aufgaben nachgehen, mit welcher Lernfreude oder ängstlichen Befangenheit sie Neues aufnehmen, sollte auch Lehrkräften nicht verborgen bleiben. Deshalb erscheint es sinnvoll, neben Lehrkrafturteilen über kognitive Komponenten auch Einschätzungen über motivationale und affektive Schülermerkmale einzuholen.

Lehrkräfte urteilen nicht über alle Schülerinnen und Schüler gleich und trotz angestrebter Fairness kann es zu Urteilsfehlern kommen. So werden sozial angepasste und sympathisch erscheinende Lernende in ihrer Leistungsstärke eher überschätzt. Bei

verhaltensauffälligen und unbeliebten Schülerinnen und Schülern wird das Leistungsvermögen dagegen zu gering eingeschätzt (Hinnant, O'Brien & Ghazarian, 2009; Itskowitz, Navon & Strauss, 1988). Diese Dichotomie von unterschiedlich eingeschätzten Schülerinnen und Schülern geht mit verschiedenen motivational-affektiven Unterschieden einher. Überschätzte Lernende haben beispielsweise ein höheres Fähigkeitsselbstkonzept und weniger Leistungsangst als unterschätzte Lernende (Urhahne et al., 2010). Mit der Möglichkeit des zeitlichen Längsschnitts kann nun geprüft werden, ob die Urteilsgenauigkeit als Ursache für die Unterschiede zwischen über- und unterschätzten Schülergruppen in Frage kommt.

Komponenten und Messung der diagnostischen Kompetenz

Bei der diagnostischen Kompetenz wird zwischen der Rang-, Niveau- und Differenzierungskomponente unterschieden (Schrader & Helmke, 1987):

Der zentrale und in den Arbeiten zur diagnostischen Kompetenz am meisten berichtete Indikator, die *Rangkomponente*, gibt darüber Auskunft, wie gut Lehrkräfte Schülerinnen und Schüler in eine Rangreihe bringen können. Studien zeigen, dass die Höhe der Rangkomponente in Abhängigkeit vom zu beurteilenden Merkmal unterschiedlich ausfällt. Bei der Einschätzung kognitiver Merkmale ist sie höher ausgeprägt als bei motivational-affektiven Merkmalen. So verweisen Metaanalysen darauf, dass Lehrkräfte die Schülerleistung relativ genau einschätzen können und üblicherweise sind Korrelationswerte höher als $r = .60$ (Hoge & Coladarci, 1989; Südkamp et al., 2012). Deutlich kleinere Korrelationswerte sind dagegen für nicht-kognitive Schülermerkmale wie Lernmotivation, Anstrengungsbereitschaft oder Leistungsangst zu verzeichnen (Spinath, 2005; Urhahne, Chao, Florineth, Luttenberger & Paechter, 2011; Urhahne, Timm, Zhu & Tang, 2013; Urhahne et al., 2010).

Die *Niveaukomponente* veranschaulicht, ob Lehrkräfte dazu tendieren, das zu beurteilende Schülermerkmal zu über- oder zu unterschätzen. Sie wird als Differenz aus Lehrkraft- und Schülerangabe gebildet und nimmt im besten Falle den Wert 0 an. Verschiedenen Studien zufolge neigen Lehrkräfte dazu, die Schülerleistung zu positiv einzuschätzen (Urhahne et al., 2010; Zhu & Urhahne, 2015).

Die *Differenzierungskomponente* zeigt, ob Lehrkräfte dazu neigen, die Streuung des zu beurteilenden Schülermerkmals zu über- oder zu unterschätzen. Sie resultiert aus dem

Verhältnis der Streuungen von Lehrkraft- zu Schülerangaben und nimmt im Optimalfall den Wert 1 an. Die zugehörigen Studienergebnisse sind recht heterogen. Es wird wechselweise von einer Unterschätzung, einer Überschätzung oder einer relativ genauen Einschätzung der Variabilität der Schülerleistung berichtet (Schrader & Helmke, 1987; Stang & Urhahne, 2016b; Südkamp, Möller & Pohlmann, 2008).

Um die Genauigkeit des Lehrkrafturteils zu beurteilen, muss auch die Messmethode berücksichtigt werden. Bei *direkter* Messung erhalten Lehrkräfte die gleichen Items wie die Schülerinnen und Schüler. Bei *indirekter* Messung kommen bei Lehrkräften und Lernenden unterschiedliche Maße zum Einsatz. Insgesamt betrachtet fallen direkte Messungen etwas genauer aus (Hoge & Coladarci, 1989).

Veränderung der Urteilsgenauigkeit

Zur Veränderung der Komponenten der Urteilsgenauigkeit über die Zeit liegen bislang nur wenige Erkenntnisse vor, da Querschnittserhebungen den Großteil der Arbeiten bilden. In einer Studie von Stang und Urhahne (2016b) zur Einschätzung von Schülerleistungen in der Realschule wurden die Komponenten der Urteilsgenauigkeit im Abstand eines Schulhalbjahres zweimal gemessen und miteinander verglichen. Die Rangkomponente stieg innerhalb des Untersuchungszeitraums signifikant an. Zudem ergab sich ein Unterschied auf der Niveaukomponente. Zum ersten Messzeitpunkt wurde die Schülerleistung signifikant überschätzt, zum zweiten Messzeitpunkt hingegen signifikant unterschätzt. Kein Unterschied zeigte sich bei der Differenzierungskomponente. In zwei experimentellen Arbeiten von Südkamp et al. (2008) bestanden hingegen bei der Rangkomponente keine signifikanten Unterschiede, wohingegen sich in der ersten der beiden Arbeiten ein Unterschied in der Differenzierungskomponente abzeichnete. Die Variabilität der Schülerleistung wurde im ersten Durchgang signifikant stärker unterschätzt als im zweiten Durchgang.

Bei Betrachtung der Stabilität der Urteilsgenauigkeit als diachroner Zusammenhang über eine sechsmonatige Zeitspanne sind die Ergebnisse uneinheitlich. In der Arbeit von Stang und Urhahne (2016b) ergab sich ein mittelhoher, allerdings nicht signifikanter Stabilitätswert für die Rangkomponente zur Testleistung, die sich in der Studie von Lorenz und Artelt (2009) noch als relativ zeitstabil erwies. Die Stabilitätswerte der Niveau- und Differenzierungskomponente fielen niedriger aus und waren ebenfalls nicht signifikant (Stang & Urhahne, 2016b). In einer Untersuchung von Spinath (2005) zeigte sich die

Genauigkeit des Lehrkrafturteils zur Fähigkeitsselbstwahrnehmung, Lernmotivation und Leistungsangst der Schülerinnen und Schüler als zeitlich stabil.

Stabilität motivational-affektiver Schülermerkmale

Damit Lehrkräfte auch über einen gewissen Zeitraum hinweg, bei wiederholter Einschätzung der gleichen Schülermerkmale, zu akkuraten Urteilen kommen, sollte das jeweilige Schülermerkmal verhältnismäßig stabil sein und keinen allzu großen, zufälligen Veränderungen unterworfen sein.

Eine Studie von Spinath (2005) deutet darauf hin, dass motivational-affektive Schülermerkmale von Grundschülern wie Fähigkeitsselbstkonzept, Lernmotivation und Leistungsängstlichkeit über einen sechsmonatigen Zeitraum relativ stabil sind. Aus der Forschung ist jedoch auch bekannt, dass es im Schulverlauf zu einem Absinken von Motivation und Emotion kommt (Gottfried, Fleming & Gottfried, 2001; Helmke, 1993; Jerusalem & Mittag, 1999). Dieser Trend setzt bereits in der Grundschulzeit ein (Helmke, 1993; Reindl & Hascher, 2013). Eine Erklärungsmöglichkeit bietet in diesem Zusammenhang die Stage-Environment-Fit-Theorie, welche von einer sich verschlechternden Passung von Schulbedingungen und Schülerbedürfnissen ausgeht (Eccles et al., 1993).

Zusammenhang zwischen der Urteilsgenauigkeit und Schülermerkmalen

Zwischen Lehrkrafturteilen und Schülermerkmalen besteht ein psychologischer Zusammenhang. Lehrkrafturteile spiegeln die Erwartungen einer Lehrkraft wider, welche zu einer erwartungskongruenten Entwicklung der Schülerinnen und Schüler führen können. Durch die Lehrkrafterwartung werden z. B. die Schülerleistung, -intelligenz und -motivation beeinflusst (Jussim, 1989; Jussim & Harber, 2005). Bei einer erwartungskongruenten Entwicklung der Schülerinnen und Schüler spricht man von einer sich selbsterfüllenden Prophezeiung.

Jussim (1986) geht davon aus, dass die sich selbst erfüllende Prophezeiung eine dreistufige Entwicklung nimmt. In der ersten Phase fällt die Lehrkraft ein Urteil und bildet eine Erwartung aus. Die Erwartung führt in der zweiten Phase dazu, dass sich die Lehrkraft gemäß ihrer Erwartung den Schülerinnen und Schülern gegenüber unterschiedlich verhält. In der dritten Phase reagieren die Schülerinnen und Schüler auf das differenzielle Lehrkraftverhalten, wodurch sich die Lehrkrafterwartung selbst bestätigt.

Aus der Forschung ist bekannt, dass die Schülerleistung durch das Lehrkrafturteil beeinflusst wird. Eine niedrige Lehrkrafterwartung, die sich in einer Leistungsunterschätzung äußern kann, hat einen negativen Einfluss auf die Schülerleistung. Eine hohe Lehrkrafterwartung, die sich in einer Leistungsüberschätzung zeigen kann, hat hingegen einen positiven Einfluss auf die Schülerleistung (Babad, Inbar & Rosenthal, 1982). Auch Stang und Urhahne (2016b) konnten einen Einfluss der Lehrkrafterwartung in der Realschule verzeichnen. Zwischen über- und unterschätzten Schülerinnen und Schülern bestanden Unterschiede in der Leistungsentwicklung. Überschätzte wiesen im Vergleich zu unterschätzten Schülerinnen und Schülern den größten Leistungszuwachs auf (Stang & Urhahne, 2016b).

Im Vergleich zu Effekten auf die Schülerleistung, wurden Auswirkungen auf motivational-affektive Schülermerkmale weniger stark beforscht. Es darf angenommen werden, dass hohe Erwartungen einen positiven Effekt auf die Schülermotivation und andere lernrelevante Merkmale haben (Jussim, 1989; Jussim & Harber, 2005). Bei der Analyse von Unterschieden zwischen über- und unterschätzten Schülerinnen und Schülern konnte festgestellt werden, dass diese vor allem in der Erfolgserwartung, dem Fähigkeitsselbstkonzept und in der Leistungsangst zu Ungunsten von unterschätzten Lernenden bestehen (Urhahne et al., 2011; Urhahne et al., 2010; Zhu & Urhahne, 2015).

Fragestellung und Hypothesen

In dieser Studie wird nicht nur die Urteilsgenauigkeit betrachtet (Schrader, 2009). Aufgrund des längsschnittlichen Vorgehens können zusätzlich die Stabilität der Lehrkrafturteile und Schülermerkmale analysiert werden sowie Zusammenhänge des diagnostischen Lehrkrafturteils mit der Schülerleistung und motivational-affektiven Schülermerkmalen. Zudem werden Erkenntnisse über die Merkmale über- und unterschätzter Schülerinnen und Schüler gewonnen.

1. Verändert sich die Genauigkeit der Lehrkrafturteile über den Zeitraum eines Jahres?

Es wird vermutet, dass Lehrkräfte nach einem Jahr, im Vergleich von dritter zu vierter Klasse, die Rangfolge von Schülerinnen und Schülern in der Testleistung signifikant besser einschätzen können (Lorenz & Artelt, 2009; Südkamp et al., 2008), weil die Leistungseinschätzung zu ihren Kernaufgaben in der Schule gehört und sie die Klassen besser kennen. Lehrkrafteinschätzungen von Motivation und Emotion der Schülerinnen und Schüler

sollten mit der Zeit genauer werden. Aufgrund der unzureichenden resp. heterogenen Forschungslage werden zur Veränderung der Niveau- und Differenzierungskomponenten resp. zur Stabilität der Urteilsgenauigkeit keine Hypothesen formuliert.

2. Verändern sich motivational-affektive Schülermerkmale über die Zeit?

Es wird angenommen, dass die motivational-affektiven Schülermerkmale relativ stabil sind (Spinath, 2005). Es sollten sich keine signifikanten Unterschiede zwischen den beiden Messzeitpunkten ergeben. Schülerleistungen werden aufgrund des Einsatzes unterschiedlicher Tests in der dritten und vierten Jahrgangsstufe nicht miteinander verglichen.

3. Bestehen Unterschiede zwischen über- und unterschätzten Schülerinnen und Schülern?

Es wird erwartet, dass zwischen der Genauigkeit des Lehrkrafturteils und den Schülermerkmalen ein Zusammenhang besteht. Unterschätzte sollten eine niedrigere Testleistung und ein ungünstigeres motivational-affektives Erleben aufweisen. Überschätzte sollten von der Leistungsüberschätzung profitieren. Insbesondere sollten sie eine höhere Erfolgserwartung, ein höheres Fähigkeitsselbstkonzept und weniger Leistungsangst haben (Urhahne et al., 2011; Urhahne et al, 2010; Zhu & Urhahne, 2015).

4.2. Methode

Stichprobe

An der Längsschnittstudie nahmen zehn Grundschulklassen teil, die über den Zeitraum eines Jahres zweimal getestet wurden. Für den ersten Messzeitpunkt lagen Daten von 189 Drittklässlern und für den zweiten Messzeitpunkt von 179 Viertklässlern vor. Für beide Messzeitpunkte fanden sich übereinstimmende Daten von 152 Schülerinnen und Schülern (39.5% weiblich), auf die sich die nachfolgenden Analysen beziehen. Die Schülerinnen und Schüler waren zum ersten Messzeitpunkt im Mittel 8.46 Jahre alt (SD = 0.55) und zum zweiten Messzeitpunkt durchschnittlich 9.47 Jahre alt (SD = 0.55). Deren zehn Mathematiklehrkräfte (80% weiblich) waren zum ersten Messzeitpunkt im Mittel 50.60 Jahre alt (SD = 6.92) und verfügten über eine Berufserfahrung von durchschnittlich 23.70 Jahren (SD = 7.50). Zum ersten Messzeitpunkt unterrichteten sie die Klassen im Schnitt 18.50 Stunden die Woche (SD = 2.42) und zum zweiten Messzeitpunkt im Mittel 22.26 Stunden in der Woche (SD = 2.88). Bei allen erhobenen Variablen waren höchstens 1.3% fehlende Werte zu verzeichnen. Diese wurden in den Analysen nicht ersetzt. Da die Daten komplett zufällig fehlen, konnte die Methode listenweiser Fallausschluss angewandt werden.

Material

Testleistung

Zur Erfassung der mathematischen Fähigkeiten der Drittklässler wurde der Deutsche Mathematiktest für dritte Klassen (DEMAT 3+; Roick, Gölitz & Hasselhorn, 2004) mit 31 Aufgaben verwendet. Zur Messung der Mathematikleistung der Viertklässler wurde der Deutsche Mathematiktest für vierte Klassen (DEMAT 4; Gölitz, Roick & Hasselhorn, 2006) mit 40 Testaufgaben herangezogen. Beide Tests basieren auf den Lehrplänen der deutschen Bundesländer und lassen sich in die Bereiche Arithmetik (z. B. Additionen), Sachrechnen (z. B. Sachrechnungen) und Geometrie (z. B. Spiegelzeichnungen) aufteilen. Zum ersten Messzeitpunkt wurden die mathematischen Fähigkeiten mit einer Reliabilität von Cronbachs $\alpha = .77$[1] erfasst, zum zweiten Messzeitpunkt betrug Cronbachs $\alpha = .81$.

Motivational-affektive Schülermerkmale

Die *Erfolgserwartung* wurde durch die erwartete Note in der nächsten Mathematikarbeit abgebildet. Das *Anspruchsniveau* bemaß sich aus der gerade noch zufriedenstellenden Note in der nächsten Mathematikarbeit. Die genauen Items der Ulmer Motivationstest-batterie lauteten (UMTB; Ziegler, Dresel, Schober & Stöger, 2005): „Was denkst du, welche Note wirst du in der nächsten Matheprobe erhalten?“ resp. „Mit welcher Note in der nächsten Matheprobe wärst du gerade noch zufrieden?“. Als Antwortmöglichkeiten waren ganze Schulnoten von 1 bis 6 möglich. Items zur Lernfreude, Schuleinstellung und Anstrengungsbereitschaft wurden dem Fragebogen zur Erfassung emotionaler und sozialer Schulerfahrungen von Grundschulkindern dritter und vierter Klassen (FEESS 3-4; Rauer & Schuck, 2003) entnommen. Die *Lernfreude* spiegelt das Ausmaß der erlebten Freude der Schülerinnen und Schüler an schulbezogenen Aufgaben wider. Ein Beispielitem für Lernfreude lautet: „Ich lerne gern in der Schule.“ Die Lernfreude wurde zum ersten Mess-zeitpunkt mit einer Reliabilität von Cronbachs $\alpha = .84$ und zum zweiten Messzeitpunkt mit einer Reliabilität von Cronbachs $\alpha = .86$ erfasst. Die *Schuleinstellung* beschreibt, wie sehr sich Schülerinnen und Schüler in der Schule wohlfühlen. Ein Beispielitem für Schul-einstellung lautet: „Ich fühle mich in der Schule wohl.“ Schuleinstellung wurde zum ersten Messzeitpunkt mit einer Reliabilität von Cronbachs $\alpha = .94$ und zum zweiten Messzeit-punkt mit einer Reliabilität von Cronbachs $\alpha = .93$ erfasst. Die *Anstrengungsbereitschaft*

[1] Bei Testverlängerung auf 40 Items entspricht Cronbachs α einem Wert von $\alpha = .81$.

gibt darüber Auskunft, wie gut die Schülerinnen und Schüler schulische Anforderungen durch eigenes Bemühen bewältigen können. Ein Beispielitem für Anstrengungsbereitschaft heißt: „Ich gebe mein Bestes in der Schule." Anstrengungsbereitschaft wurde zum ersten und zweiten Messzeitpunkt mit einer Reliabilität von Cronbachs α = .74 erfasst. Das *Fähigkeitsselbstkonzept* indiziert, wie die Schülerinnen und Schüler ihre Fähigkeiten in Mathematik einschätzen. Ein Beispielitem lautet: „Ich bin gut in Mathe." Das Fähigkeitsselbstkonzept wurde zum ersten Messzeitpunkt mit einer Reliabilität von Cronbachs α = .91 und zum zweiten Messzeitpunkt mit einer Reliabilität von Cronbachs α = .93 erfasst. Die Items zur Erfassung des Fähigkeitsselbstkonzeptes wurden der UMTB (Ziegler et al., 2005) und der Studie von Dickhäuser und Galfe (2004) entnommen. Die *Leistungsangst* gibt an, wie viel Angst die Schülerinnen und Schüler vor dem Fach Mathematik haben. Ein Beispielitem der Leistungsangst ist: „Ich habe Angst vor einer Matheprobe." Die Leistungsangst wurde zum ersten Messzeitpunkt mit einer Reliabilität von Cronbachs α = .84 und zum zweiten Messzeitpunkt mit einer Reliabilität von Cronbachs α = .85 erfasst. Die Items zur Erfassung der Leistungsangst stammen aus der UMTB (Ziegler et al., 2005) und einem Bericht von Hanisch (2004). Alle Variablen, außer Erfolgserwartung und Anspruchsniveau, wurden anhand von neun Items auf einer vierstufigen Likert-Skala von *0 – stimmt gar nicht* bis *3 – stimmt genau* gemessen. Die Items wurden so adaptiert, dass sie auf das Unterrichtsfach Mathematik bezogen und der besseren Verständlichkeit halber für die Grundschulkinder positiv formuliert waren.

Lehrkraftdaten

Die Lehrkräfte beantworteten soziodemografische Fragen zu Alter, Geschlecht und Berufserfahrung. Zudem gaben sie an, welche Fächer und wie viele Stunden sie pro Woche in der Klasse unterrichten. Zu beiden Messzeitpunkten erhielten die Lehrkräfte jeweils eine Kopie des standardisierten Mathematikleistungstests sowie des Fragebogens, um sich mit den eingesetzten Instrumentarien vertraut zu machen. Nach Beantwortung des kurzen Lehrkraftfragebogens schätzten sie die mathematischen Leistungen und motivational-affektiven Schülermerkmale für jede Schülerin bzw. jeden Schüler ein: Testleistung („Wie viele der 31/40 Aufgaben des Mathematiktests löst der Schüler richtig?"), Erfolgserwartung resp. Anspruchsniveau („Bitte schätzten Sie ein, wie der Schüler folgende Fragen beantwortet: „Was denkst du, welche Note wirst du in der nächsten Matheprobe erhalten?" resp. „Mit welcher Note in der nächsten Matheprobe wärst du

gerade noch zufrieden?“), Fähigkeitsselbstkonzept („Wie schätzt der Schüler seine Fähigkeiten in Mathematik ein“), Leistungsangst („Wie viel Angst hat der Schüler vor Mathematik“), Lernfreude („Wie positiv erlebt der Schüler im Allgemeinen schulische Aufgaben“), Schuleinstellung („Wie wohl fühlt sich der Schüler im Allgemeinen in der Schule“) und Anstrengungsbereitschaft („Wie gut bewältigt der Schüler im Allgemeinen Anforderungen durch eigenes Bemühen“). Erfolgserwartung und Anspruchsniveau wurden mit Noten von 1 bis 6 beantwortet. Die anderen Schülermerkmale wurden im Vergleich zu anderen Schülern desselben Alters auf einer neunstufigen Likert-Skala von z. B. *0 – sehr viel weniger* bis *8 – sehr viel mehr* eingeschätzt.

Durchführung

Die Studie wurde von der zuständigen Behörde genehmigt. Geschulte Testleiterinnen und Testleiter führten die Untersuchung im Klassenzimmer durch. Die Studienteilnahme war freiwillig und nur mit Einverständniserklärung der Eltern möglich. Die erste Testung fand gegen Ende des ersten Halbjahres in der dritten Klasse statt. Die zweite Testung erfolgte ein Jahr später zum Ende des ersten Halbjahres in der vierten Klasse. Zu beiden Messzeitpunkten dauerte die Untersuchung ca. 1.5 Schulstunden. Zuerst bearbeiteten die Schülerinnen und Schüler die Aufgaben des standardisierten Leistungstests. Nach einer kurzen Pause wurden ihnen die Items des Fragebogens langsam vorgelesen, um möglichen Verständnisschwierigkeiten direkt begegnen und ein gleichmäßiges Antworttempo gewährleisten zu können. Schülerinnen und Schüler, die an der Testung nicht teilnehmen durften, waren in der Zwischenzeit mit Stillarbeit z. B. Malen oder Lesen beschäftigt. Lehrkräfte erhielten ihre Materialen zur gleichen Zeit wie die Schülerinnen und Schüler. Den Lehrkräften war es freigestellt, ihre Fragebögen im Lehrerzimmer oder in der Klasse zu beantworten. Die Einschätzungen dauerten im Mittel 45 Minuten. Nach jeder Erhebungswelle wurde den Lehrkräften Rückmeldung über Fragebogen- und Testergebnisse sowie die Akkuratheit ihrer Einschätzungen auf Klassenebene gegeben.

Statistische Analysen

Die Komponenten der diagnostischen Kompetenz wurden wie in der Forschung üblich berechnet (vgl. Praetorius, Lipowsky & Karst, 2012). Um zwischen in der Leistung über- und unterschätzten Schülerinnen und Schülern differenzieren zu können, wurde eine Regressionsanalyse mit der Schülerleistung als Prädiktor und der Leistungseinschätzung der Lehrkraft als Kriterium vorgenommen (Alvidrez & Weinstein, 1999). Die standardi-

sierten Residualwerte dienten der Einteilung in überschätzte (oberes Drittel) und unterschätzte (unteres Drittel) Lernende. Aufgrund der Stichprobengröße wurde diese Vorgehensweise gewählt.

4.3. Ergebnisse

Im Hinblick auf die Variabilität der Urteilsgenauigkeit über die Zeit erbringt der Fisher *Z*-Test für die Rangkomponenten zur Testleistung (z = -2.10, $p < .05$), Lernfreude (z = -2.43, $p < .05$), Schuleinstellung (z = -2.08, $p < .05$) und Anstrengungsbereitschaft (z = -2.23, $p < .05$) signifikante Unterschiede zwischen den beiden Messzeitpunkten. Bei den Variablen Erfolgserwartung (z = -0.15, *ns*), Anspruchsniveau (z = -1.43, *ns*), Fähigkeitsselbstkonzept (z = -0.13, *ns*) und Leistungsangst (z = -0.45, *ns*) bestehen keine signifikanten Unterschiede in den Rangkomponenten der beiden Messzeitpunkte. Die *z*-standardisierten Niveau- bzw. Differenzierungskomponenten der Testleistung des ersten und zweiten Messzeitpunktes unterscheiden sich nicht signifikant voneinander, $t(9)$ = -2.14 resp. $t(9)$ = -0.43, beide *ns*. Des Weiteren bestehen keine signifikanten Unterschiede in den Niveaukomponenten der Erfolgserwartung resp. des Anspruchsniveaus des ersten und zweiten Messzeitpunktes, $t(9)$ = 0.17 resp. $t(9)$ = -1.13, beide *ns*, sowie keine signifikanten Unterschiede in den Differenzierungskomponenten dieser Variablen, $t(9)$ = 0.59 resp. $t(9)$ = -1.39, beide *ns*.

Die diachronen Zusammenhänge ergeben nur für die Rangkomponenten zur Testleistung und Erfolgserwartung signifikante Stabilitätskennwerte in Höhe von r_{tt} = .73 resp. r_{tt} = .75, beide $p < .05$. Die Stabilitätswerte der Rangkomponenten der weiteren motivational-affektiven Schülermerkmale sind nicht signifikant und schwanken zwischen r_{tt} = -.59 und r_{tt} = .40 (vgl. Tabelle 4.1). Zudem ergeben sich Stabilitätswerte in Höhe von r_{tt} = .22, *ns*, für die Niveau- und in Höhe von r_{tt} = .64, $p < .05$, für die Differenzierungskomponente der Testleistung. Für die Niveau- und Differenzierungskomponente der Erfolgserwartung ergeben sich Werte in Höhe von r_{tt} = .33 bzw. von r_{tt} = -.26, beide *ns.* Werte in Höhe von r_{tt} = .76, $p < .01$, ergeben sich für die Niveau- und in Höhe von r_{tt} = -.24, *ns*, für die Differenzierungskomponente des Anspruchsniveaus.

Tabelle 4.1: Stabilitäten der Genauigkeit der Lehrkrafturteile und Schülermerkmale.

	Lehrkrafturteile		Schülermerkmale	
	r_{tt}	p	r_{tt}	p
Testleistung	.73	< .05	.64	< .01
Erfolgserwartung	.75	< .05	.41	< .01
Anspruchsniveau	-.01	*ns*	.23	< .01
Fähigkeitsselbstkonzept	.02	*ns*	.50	< .01
Leistungsangst	.40	*ns*	.47	< .01
Lernfreude	.12	*ns*	.40	< .01
Schuleinstellung	-.59	*ns*	.51	< .01
Anstrengungsbereitschaft	.14	*ns*	.23	< .01

In Tabelle 4.2 wird hinsichtlich der Akkuratheit der Lehrkrafteinschätzungen ersichtlich, dass zum ersten und zweiten Messzeitpunkt das Lehrkrafturteil zur Testleistung und die Schülerleistung am höchsten miteinander korrelieren. Niedrigere Korrelationen zwischen Lehrkrafteinschätzung und Schülermerkmal im Vergleich zur Schülerleistung bestehen für beide Messzeitpunkte für die Schülermerkmale Erfolgserwartung (z_{t1} = 0.79, *ns,* z_{t2} = 2.70, p < .05), Anspruchsniveau (z_{t1} = 5.62, z_{t2} = 6.13, beide p < .001), Fähigkeitsselbstkonzept (z_{t1} = 2.18, p < .05, z_{t2} = 4.07, p < .001), Leistungsangst (z_{t1} = 3.62, z_{t2} = 5.15, beide p < .001), Lernfreude (z_{t1} = 5.62, z_{t2} = 5.15, beide p < .001), Schuleinstellung (z_{t1} = 6.70, z_{t2} = 6.55, beide p < .001) und Anstrengungsbereitschaft (z_{t1} = 6.22, z_{t2} = 5.92, beide p < .001). Die Spannweiten der Korrelationen indizieren, dass zwischen den Lehrkräften interindividuelle Differenzen in der Genauigkeit der Einschätzung der Schülerleistung und motivational-affektiver Schülermerkmale bestehen. Demnach ist nicht jede Lehrkraft im Stande, die Schülerleistung oder die anderen Schülermerkmale genau einzuschätzen. Die Niveaukomponenten in Tabelle 4.2 verdeutlichen, dass bei der Messung in der dritten Klasse lediglich eine Tendenz zur Überschätzung der Schülerleistung besteht, $t(9)$ = 1.34, *ns.* Ein Jahr später, in der vierten Klasse, schätzen Lehrkräfte die Schülerleistung hingegen signifikant zu positiv ein, $t(9)$ = 4.16, p < .01. Die Erfolgserwartung wird zu beiden Messzeitpunkten signifikant unterschätzt, $t(9)$ = -2.63, p < .05 bzw. $t(9)$ = -3.26, p < .01. Das Anspruchsniveau wurde weder signifikant über- noch

unterschätzt, $t(9) = -0.11$ bzw. $t(9) = 1.12$, beide *ns*. Die Differenzierungskomponenten in Tabelle 4.2 zeigen, dass zum ersten Messzeitpunkt die Streuung der Schülerleistung nahezu realistisch eingeschätzt wird, $t(9) = 1.31$, *ns*, wohingegen sie beim zweiten Messzeitpunkt signifikant überschätzt wird, $t(9) = 2.82$, $p < .05$. Zudem wurden zu beiden Messzeitpunkten die Streuungen der Erfolgserwartung nahezu realistisch eingeschätzt, $t(9) = 0.71$ bzw. $t(9) = -.12$, ebenso wie die des Anspruchsniveaus, $t(9) = -.83$ bzw. $t(9) = 1.32$, alle *ns*.

Tabelle 4.2: Genauigkeit der Lehrkrafturteile ($N = 10$) in der dritten und vierten Klasse.

	3. Klasse				4. Klasse			
Variable	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Rangkomponente								
Testleistung	.65**	0.30	-.19	.85	.76**	0.10	.58	.84
Erfolgserwartung	.60**	0.32	-.26	.86	.61**	0.29	-.08	.88
Anspruchsniveau	.19	0.22	-.23	.42	.33**	0.17	.06	.66
Fähigkeitsselbstkonzept	.50**	0.33	-.25	.82	.51**	0.11	.38	.72
Leistungsangst	.38**	0.28	-.23	.64	.42**	0.23	.09	.71
Lernfreude	.19	0.30	-.38	.52	.42**	0.31	-.37	.79
Schuleinstellung	.08	0.25	-.19	.50	.29**	0.26	-.01	.71
Anstrengungsbereitschaft	.13	0.37	-.40	.56	.35**	0.30	-.41	.66
Niveaukomponente								
Testleistung	1.52	3.59	-4.92	6.22	4.32**	3.28	-2.39	10.38
Erfolgserwartung	-0.24*	0.29	-0.71	0.16	-0.26**	0.25	-0.64	0.20
Anspruchsniveau	-0.02	0.61	-1.11	0.93	0.13	0.36	-0.46	0.82
Differenzierungskomponente								
Testleistung	1.15	0.37	0.52	1.87	1.19*	0.21	0.85	1.47
Erfolgserwartung	1.07	0.32	0.57	1.44	0.99	0.20	0.56	1.19
Anspruchsniveau	0.93	0.26	0.59	1.36	1.13	0.32	0.76	1.78

Anmerkungen. *$p < .05$; **$p < .01$. Zur Berechnung der Niveau- und Differenzierungskomponenten der Erfolgserwartung und des Anspruchsniveaus wurden die Werte umgepolt.

Im Hinblick auf die Veränderungen in den motivational-affektiven Schülermerkmalen über die Zeit bestehen signifikante niedrige bis mittlere Stabilitätskennwerte von $r = .23$ bis $r = .51$ (vgl. Tabelle 4.1). Motivation und Emotion ändern sich im Mittel nur wenig. Wie *t*-Tests zeigen, ist lediglich ein signifikantes Absinken der Lernfreude über den Zeitraum eines Jahres zu verzeichnen. Alle anderen Merkmale verändern sich nicht signifikant (vgl. Tabelle 4.3).

Tabelle 4.3: Veränderung motivationaler und emotionaler Schülermerkmale über die Zeit.

	3. Klasse		4. Klasse				
Variable	*M*	*SD*	*M*	*SD*	df	*t*	*p*
Erfolgserwartung	2.03	.89	2.12	.94	151	-1.15	.251
Anspruchsniveau	2.62	.72	2.70	.68	149	-1.21	.227
Fähigkeitsselbstkonzept	2.23	.69	2.14	.69	149	1.60	.111
Leistungsangst	1.15	.78	1.03	.76	149	1.88	.063
Lernfreude	2.08	.67	1.89	.67	149	3.03	.003
Schuleinstellung	1.85	.92	1.85	.81	149	0.02	.984
Anstrengungsbereitschaft	2.46	.50	2.41	.57	149	1.01	.312

Bei genauerer Betrachtung zeigen sich differenzielle Wirkungen der Urteilsgenauigkeit auf unterschiedliche Schülergruppen. In multivariaten Varianzanalysen wurden unter- und überschätzte Schülerinnen und Schüler miteinander verglichen. In der dritten Klasse fällt die multivariate Varianzanalyse signifikant aus, Wilks-Λ = .74, $F(8, 91) = 4.07$, $p < .001$, $\eta^2 = .26$. Wie aus Tabelle 4.4 ersichtlich, unterscheiden sich über- von unterschätzten Schülerinnen und Schülern nicht in der Testleistung. Allerdings bestehen signifikante Unterschiede zu Gunsten der überschätzten Schülerinnen und Schüler in Erfolgserwartung, Fähigkeitsselbstkonzept und Leistungsangst. Überschätzte erwarten in der nächsten Mathematikarbeit eine bessere Note, haben ein höheres Fähigkeitsselbstkonzept und erleben weniger Leistungsangst.

In der vierten Klasse erbringt die multivariate Varianzanalyse ebenfalls ein signifikantes Ergebnis, Wilks-Λ = .79, $F(8, 88) = 2.91$, $p < .01$, $\eta^2 = .21$. Wie aus Tabelle 4.5 ersichtlich, unterscheiden sich die in der dritten Klasse über- und unterschätzten Schülerinnen und Schüler nun auch in der Testleistung signifikant voneinander. Des Weiteren bestehen signifikante Unterschiede zu Ungunsten der unterschätzten Schülerinnen und Schüler bei den Variablen Erfolgserwartung, Fähigkeitsselbstkonzept, Leistungsangst und Lernfreude. Unterschätzte haben eine signifikant niedrigere Testleistung sowie eine niedrigere Erfolgserwartung, ein geringeres Fähigkeitsselbstkonzept, weniger Lernfreude und erleben stärkere Leistungsangst.

Tabelle 4.4: Unterschiede zwischen über- und unterschätzten Schülerinnen und Schülern in der dritten Klasse.

	Überschätze (n = 51)		Unterschätzte (n = 49)				
Variable	*M*	*SD*	*M*	*SD*	$F(1, 98)$	η^2	1-ß
Testleistung	15.14	4.21	14.57	4.37	0.44	-	.10
Erfolgserwartung	1.67	0.66	2.33	0.89	17.85***	.15	-
Anspruchsniveau	2.57	0.66	2.71	0.71	1.13	-	.18
Fähigkeitsselbstkonzept	2.45	0.59	1.95	0.79	12.79**	.12	-
Leistungsangst	0.82	0.61	1.41	0.81	16.58***	.15	-
Lernfreude	2.14	0.63	1.95	0.77	1.68	-	.25
Schuleinstellung	1.84	0.89	1.88	0.91	0.04	-	.06
Anstrengungsbereitschaft	2.53	0.50	2.36	0.54	2.60	-	.36

Anmerkungen. Noten von 1 bis 6 bei Erfolgserwartung und Anspruchsniveau. ** $p < .01$; *** $p < .001$.

Tabelle 4.5: Unterschiede zwischen über- und unterschätzten Schülerinnen und Schülern in der vierten Klasse.

	Überschätze (n = 50)		Unterschätzte (n = 47)				
Variable	*M*	*SD*	*M*	*SD*	*F*(1, 95)	η²	1-ß
Testleistung	19.44	6.04	16.09	6.22	7.26*	.07	-
Erfolgserwartung	1.85	0.70	2.58	1.08	15.72***	.14	-
Anspruchsniveau	2.60	0.57	2.84	0.70	3.43	-	.45
Fähigkeitsselbstkonzept	2.24	0.62	1.87	0.66	8.00*	.08	-
Leistungsangst	0.86	0.72	1.29	0.77	7.88*	.08	-
Lernfreude	2.00	0.64	1.74	0.65	4.10*	.04	-
Schuleinstellung	1.89	0.78	1.81	0.83	0.21	-	.07
Anstrengungsbereitschaft	2.40	0.45	2.39	0.43	0.01	-	.05

Anmerkungen. Noten von 1 bis 6 bei Erfolgserwartung und Anspruchsniveau. * $p < .05$; *** $p < .001$.

Regressionen ergeben, dass die Urteilsgenauigkeit des ersten Messzeitpunktes, indiziert als Leistungsüber- bzw. -unterschätzung, die Schülerleistung, $R^2 = .07$, $B = 3.29$, $SE = 1.21$, $p < .01$, die Erfolgserwartung, $R^2 = .15$, $B = -.76$, $SE = .18$, $p < .001$, das Fähigkeitsselbstkonzept, $R^2 = .08$, $B = .39$, $SE = .13$, $p < .01$, die Leistungsangst, $R^2 = .08$, $B = -.44$, $SE = .15$, $p < .01$ sowie die Lernfreude, $R^2 = .05$, $B = .29$, $SE = .13$, $p < .05$, des zweiten Messzeitpunktes signifikant vorhersagt.

4.4. Diskussion

Eine adäquate Beurteilung der Schülerleistung und motivational-affektiver Schülermerkmale ist essentielle Voraussetzung der Vergabe gerechter Noten, weshalb in dieser Arbeit auf die Akkuratheit der Einschätzung der Schülerleistung und motivational-affektiver Schülermerkmale gleichermaßen fokussiert wurde. Aufgrund des noch unbefriedigenden Forschungsstandes zur diagnostischen Kompetenz (Artelt & Gräsel, 2009), wurden neben der Genauigkeit der Lehrkrafteinschätzungen auch die Veränderung der Urteilsgenauigkeit und in dem Kontext auch Veränderungen motivational-affektiver Schülermerkmale über einen einjährigen Zeitraum betrachtet. Zudem wurden Unterschiede zwischen über- und unterschätzten Schülerinnen und Schülern analysiert.

Im Einklang mit metaanalytischen Ergebnissen und früheren Forschungsbefunden wiesen Schülerleistung und Lehrkrafturteil eine hohe Übereinstimmung auf (Hoge & Coladarci, 1989; Südkamp et al., 2012; Urhahne et al., 2011; Urhahne et al., 2013). Lehrkräfte können demnach die Schülerleistung relativ genau einschätzen. Ursache hierfür könnte in der Besonderheit der Stichprobe liegen: Grundschullehrkräfte wissen besonders gut über ihre Schützlinge Bescheid. Sie sind dazu angehalten auch die Hausaufgaben der Schülerinnen und Schüler zu korrigieren. Ihr Wissen über die Schülerleistung beziehen Grundschullehrkräfte daher nicht nur aus Tests und Klassenarbeiten. Insgesamt zeigte sich jedoch, vor allem bei Messung in der 3. Klasse, dass nicht jede Lehrkraft eine genaue Einschätzung der Schülerleistung abgeben konnte. Dieses Ergebnis stimmt mit anderen bekannten Befunden überein (Schrader & Helmke, 1987; Urhahne et al., 2013).

Ebenfalls im Einklang mit vorherigen Forschungsergebnissen konnten Lehrkräfte motivational-affektive Schülermerkmale weniger genau einschätzen (Spinath, 2005; Urhahne et al., 2011; Urhahne et al., 2010). Lehrkrafturteil und Schülerleistung korrelierten zu beiden Messzeitpunkten höher miteinander als die Lehrkrafteinschätzungen mit motivational-affektiven Schülermerkmalen. Dies kann daran liegen, dass Lehrkräfte

nicht geschult sind motivational-affektive Schülermerkmale einzuschätzen. Zudem ist die Leistungsbewertung deren Profession. Des Weiteren wurde die Schülerleistung direkt gemessen, wohingegen die anderen Merkmale, außer Erfolgserwartung und Anspruchsniveau, indirekt erfasst wurden. Dabei fallen laut Hoge und Coladarci (1989) direkte Messungen etwas genauer aus. Außerdem wurden die motivational-affektiven Schülermerkmale per subjektiver Selbsteinschätzung gemessen und basieren nicht auf objektiven Testungen. Daher könnte ein Problem der sozialen Erwünschtheit bestehen. Auch stehen den Lehrkräften nur sichtbare Informationen zur Verfügung, welche mit dem wahren Erleben der Schülerinnen und Schüler nicht übereinstimmen müssen. Nicht immer äußert sich das Erleben des Schülers in sichtbarem Verhalten bzw. es liegen nicht immer eindeutige Indikatoren vor (Kenny & West, 2010) oder aber nur subtil, weswegen das Erleben kaum wahrnehmbar ist (Givvin, Stipek, Salmon & MacGyvers, 2001). Daher kann es für Lehrkräfte schwierig sein, die wahre Ausprägung motivational-affektiver Schülermerkmale genau einzuschätzen. Insgesamt fielen die Zusammenhänge zwischen Lehrkrafteinschätzung und motivational-affektiven Schülermerkmalen am höchsten für solche Merkmale aus, welche auf Schülerebene stärker mit der Schülerleistung zusammenhängen.

In der dritten Klasse bestand eine Tendenz der Lehrkräfte zur Überschätzung der Schülerleistung, in der vierten Klasse war die Überschätzung signifikant (Urhahne et al., 2011; Urhahne et al., 2013; Zhu & Urhahne, 2015). Da der standardisierte Leistungstest für die Lehrkräfte ein relativ unbekanntes Instrument darstellt, könnten sie die Aufgabenschwierigkeit unterschätzt und somit Fehlermöglichkeiten übersehen haben. Ähnliches zeigte sich bei der Differenzierungskomponente: Zum ersten Messzeitpunkt bestand lediglich eine Tendenz, zum zweiten Messzeitpunkt wurden Unterschiede hingegen signifikant überschätzt (Schrader & Helmke, 1987; Urhahne et al., 2010). In Einklang mit Urhahne et al. (2013) zeigte sich zudem, dass die Erfolgserwartung unterschätzt und das Anspruchsniveau tendenziell eher überschätzt wurde.

Nach einem Jahr, bei Messung in der vierten Klasse, konnten die Lehrkräfte die Schülerleistung signifikant besser einschätzen als bei Messung in der dritten Klasse. Ebenfalls vermochten sie die Lernfreude, Schuleinstellung und Anstrengungsbereitschaft der Schülerinnen und Schüler nach einem Jahr signifikant besser einzuschätzen. Zum zweiten Messzeitpunkt kennen die Lehrkräfte die Schülerinnen und Schüler länger und besser. Zudem befinden sich die Schülerinnen und Schülerinnen in der heißen Phase des

Übertritts, weswegen Lehrkräfte die Schülerinnen und Schüler intensiver beobachten und testen müssen, um eine adäquate Übertrittsempfehlung abgeben zu können. Daher verfügen Lehrkräfte zu diesem Messzeitpunkt über mehr Wissen. Eine weitere Ursache könnte in der differenzierten Rückmeldung liegen, welche die Lehrkräfte nach Ende des ersten Messzeitpunktes erhielten. Diese könnte die Lehrkräfte motiviert haben, die Schülerleistung und die anderen Schülermerkmale noch genauer einzuschätzen bzw. ihren Blick für motivational-affektive Schülermerkmale zu schulen. Allerdings erhielten die Lehrkräfte die Rückmeldung zeitnah nach der ersten Testung, so dass ein Effekt zum zweiten Messzeitpunkt, ein Jahr später, eher ausgeschlossen werden kann. Dass bei den anderen Merkmalen die Anstiege nicht signifikant waren, kann darin begründet liegen, dass es schwierig ist, diese Merkmale einzuschätzen. Hinzukommt, dass die Korrelationen wie bei der Erfolgserwartung bereits hoch waren, weshalb eine signifikante Steigerung schwieriger zu erreichen ist.

Bei Betrachtung des diachronen Zusammenhangs der Rangkomponente erwiesen sich nur das Lehrkrafturteil zur Testleistung (Lorenz & Artelt, 2009) und Erfolgserwartung als stabil. Die anderen Lehrkrafteinschätzungen waren nicht stabil, was nicht in Einklang mit anderen Einzelbefunden steht (Spinath, 2005). Allerdings nahmen an der Studie von Spinath (2005) deutlich mehr Lehrkräfte teil. Ursache der nicht stabilen Ergebnisse könnte sein, dass manche Lehrkräfte zum ersten Messzeitpunkt die Merkmale besser einschätzen konnten als beim zweiten Mal und andersherum. Hinzu kommt, dass sich das wahre Schülererleben nicht in kongruentem, beobachtbarem Verhalten äußern muss, weswegen es schwieriger ist, diese Merkmale mehrfach akkurat einzuschätzen. Eine weitere Erklärungsmöglichkeit stellt der Stichprobenumfang dar, weswegen die statistische Signifikanz kritisch zu betrachten ist. Daher sollte zur Bewertung, ob die Urteilsgenauigkeit stabil ist, ein anderes Kriterium als die statistische Signifikanz herangezogen werden.

Um Schülermerkmale über einen längeren Zeitraum mehrfach genau einschätzen zu können, müssen Lehrkräfte ihre Urteile immer an die aktuellen Gegebenheiten anpassen. Dies bedeutet, dass die Schülermerkmale keinen allzu großen, zufälligen Veränderungen unterliegen dürfen. Daher wurden die Stabilitätswerte der Schülermerkmale ebenfalls berechnet. Die Stabilitätskoeffizienten der Schülermerkmale fielen alle signifikant aus (Spinath, 2005). Es kann davon ausgegangen werden, dass die nicht signifikanten Stabilitätskoeffizienten der Lehrkrafteinschätzungen nicht instabilen Schülermerkmalen

geschuldet sind. Im Vergleich zu den Stabilitätswerten der Lehrkrafteinschätzungen fallen die der Schülermerkmale teilweise gleich hoch und höher aus. Dies lässt darauf schließen, dass die Schülermerkmale in Wirklichkeit stabiler sind als Lehrkräfte annehmen.

Die Stabilität der motivational-affektiven Schülermerkmale zeichnete sich auch im *t*-Test ab. Hier gab es nur ein signifikantes Absinken der selbstberichteten Lernfreude vom ersten zum zweiten Messzeitpunkt, welches in Einklang mit der Forschung steht (Jerusalem & Mittag, 1999). Die Abnahme der Lernfreude könnte dadurch erklärt werden, dass zum Übertritt hin mehr und intensiver gelernt wird. Dies kann zu Lasten von Freizeitaktivitäten gehen und somit zu einer verminderten Lernfreude führen. Dieser hypothetische Zusammenhang müsste allerdings in weiteren Studien geprüft werden. Bei den anderen Merkmalen zeichnete sich ebenfalls eine negative Tendenz ab (Gottfried et al., 2001). Das tendenzielle Absinken von schulbezogener Motivation und Emotion im Verlauf der gesamten Schulzeit kann mit der Stage-Environment-Fit-Theorie erklärt werden (Eccles et al., 1993).

Bei genauerer Betrachtung der Schülerebene zeigen sich differenzielle Effekte der Leistungsüber- und -unterschätzung durch die Lehrkraft. Wesentliche Forschungsergebnisse konnten repliziert werden: In der dritten Klasse unterschieden sich die über- und unterschätzten Schülerinnen und Schüler in der Erfolgserwartung, dem Fähigkeitsselbstkonzept und der Leistungsangst (Urhahne et al., 2010; Urhahne et al., 2011; Zhu & Urhahne, 2015). Alle drei Merkmale fielen zu Ungunsten der unterschätzten Lernenden aus. Bisher wurden Unterschiede zwischen Über- und Unterschätzten nur im Querschnitt betrachtet (Urhahne et al., 2011; Urhahne et al., 2013; Urhahne et al., 2010; Zhu & Urhahne, 2015). Durch die längsschnittliche Herangehensweise zeigten sich differenzielle Befunde der dritten Klasse auch in den Ausprägungen der Schülermerkmale der vierten Klasse. In der vierten Klasse bestehen zusätzlich Unterschiede in den Variablen Testleistung und Lernfreude, welche bei den Überschätzten im Vergleich zu den Unterschätzten signifikant höher ausfielen. Zwischen Lehrkrafteinschätzungen und Schülermerkmalen besteht also ein Zusammenhang, welcher durch das Eintreten einer sich selbsterfüllenden Prophezeiung erklärt werden kann (Jussim, 1986). Das Lehrkrafturteil, welches die Lehrkrafterwartung spiegelt, kann die Schülerleistung, aber auch die Schülermotivation beeinflussen (Jussim, 1989; Jussim & Harber, 2005). In einer Folgestudie sollte das erlebte

Lehrkraftverhalten der Grundschülerinnen und Grundschüler miterhoben werden, weil davon auszugehen ist, dass dieses einen mediierenden Faktor darstellt (Urhahne, 2015).

In der dritten wie vierten Klasse unterschieden sich über- und unterschätzte Schülerinnen und Schüler. Überschätzte wiesen günstigere Ausprägungen der motivational-affektiven Merkmale und eine höhere Testleistung auf. Eine leichte Überschätzung scheint demnach pädagogisch-psychologisch sinnvoll, da sie einen positiven Effekt auf Schülermerkmale hat (Stang & Urhahne, 2016b; Urhahne & Zhu, 2015a). Im Sinne der Zone der proximalen Entwicklung (Vygotsky, 1978) wird durch eine leichte Überschätzung ein förderlicher Entwicklungsraum für Schülerinnen und Schüler geschaffen. Einer Unterschätzung sollte hingegen präventiv entgegengewirkt werden, da diese mit einer negativeren Ausprägung motivational-affektiver Merkmale und einer geringeren Testleistung einhergehen kann.

Eine akkurate Erfassung der motivational-affektiven Schülermerkmale ist in vielerlei Hinsicht essentiell. Motivational-affektive Schülermerkmale stellen wichtige Determinanten und somit auch Stellschrauben schulischer Leistung dar (Helmke & Schrader, 2010). Werden die Merkmale akkurat von der Lehrkraft eingeschätzt, so kann diese potenziellen Problemen, welche mit negativen oder geringen motivational-affektiven Merkmalsausprägungen zusammenhängen, entgegenwirken. In Studienveranstaltungen, aber auch in Weiterbildungsmaßnahmen sollten daher bspw. Verhaltensindikatoren erarbeitet und besprochen werden, welche auf die Ausprägung der Schülermerkmale hindeuten (Lohaus, 2009). Zudem sollten standardisierte Tests vorgestellt und eruiert werden, welche im Klassenzimmer von der Lehrkraft eingesetzt werden können, um das motivationale wie emotionale Erleben der Schülerinnen und Schüler besser abbilden zu können. Auch sollten Lehrkräfte ermutigt werden, standardisierte Tests im Unterrichtsgeschehen einzubinden. Des Weiteren sollte stärker thematisiert werden, wie sich die Schülerleistung entwickelt sowie wie und warum sich motivational-affektive Schülermerkmale im Schulverlauf verändern. Hierdurch könnte die Bedeutung hervorgehoben werden, diese Merkmale akkurat einzuschätzen, um nötigenfalls präventiv einschreiten zu können.

Wünschenswert wären demnach Folgestudien, welche die Ergebnisse, vor allem zur Stabilität, anhand einer größeren Stichprobe replizieren können. Ein Desiderat wären mehr als nur zwei Messzeitpunkte. Bei Hinzunahme zusätzlicher Messzeitpunkte könnten zum

einen differenziertere Aussagen getroffen werden, wie stabil Lehrkrafturteile sind bzw. bleiben. So könnten belastbarere, weiterführende Erkenntnisse zur diagnostischen Kompetenz als Persönlichkeitsmerkmal, also als zeitstabile Fähigkeit oder aber zur diagnostischen Kompetenz als über verschiedene Situationen variabel und zeitinstabil, gewonnen werden, da Zufallsmessungen ausgeglichen werden könnten. Zum anderen könnte überprüft werden, wie stark und weitreichend Leistungsüber- und -unterschätzungen wirken. Es könnte erforscht werden, ob sich die Auswirkungen sogar auf der weiterführenden Schule manifestieren und in einer konstant niedrigeren Leistung sowie in einer negativeren schulbezogenen Motivation und Emotion kulminieren. Sollte dies der Fall sein, so könnten per qualitativer Fragebogen- oder Interviewstudie Zusammenhänge zwischen einer Über- und Unterschätzung durch Lehrkräfte, dem erlebten Lehrkraftverhalten und dem höchsten erworbenen Bildungsabschluss oder Berufserfolg aufgedeckt werden.

Von besonderer Bedeutung ist, dass den Lehrkräften ihre Wirkung auf Schülerleistungen, Motivation und Emotion stärker bewusst wird. Lehrkräfte sollten demnach ihre eigenen Urteile gelegentlich kritisch reflektieren (Helmke, 2009). Zudem wäre es wünschenswert, das Schulerleben von Zeit zu Zeit an objektiven, standardisierten Tests festzumachen, da sich das Erleben nicht unbedingt in Verhalten manifestieren muss (Kenny & West, 2010). Insbesondere im Grundschulbereich sollte aufgrund der Wirkung von negativen Lehrkrafterwartungen auf die Schülerleistung darauf geachtet werden, da dort der Grundstein der Bildungskarriere gelegt wird.

Dem Zitat „Lernen ist Erfahrung. Alles andere ist einfach nur Information.“ (Albert Einstein) folgend, sollten Informationen über die diagnostische Kompetenz in Weiterbildungsveranstaltungen oder Seminaren den Lehrkräften nicht einfach nur dargeboten werden. Dies wäre fatal. In Weiterbildungsveranstaltungen sollte die Informationsgabe daher mit interaktiven Elementen und Übungsaufgaben verbunden sein. Zudem sollten Lehramtsstudierende bereits durch Erfahrung lernen, die Schülerleistung und motivational-affektive Schülermerkmale einzuschätzen. Zum einen könnte dies im Rahmen von Praktika geschehen, in denen sie verschiedene Schülerinnen und Schüler kennenlernen, um dann bei Tests und Prüfungen, die von der Lehrkraft gestellt werden, einen begründeten Tipp abzugeben, wie der Lernende abschneiden wird. Zum anderen könnte eine Simulationssoftware entwickelt werden, welche ein Training der diagnostischen Fähigkeiten bereits im Vorfeld ermöglicht sowie differenziertes Feedback zum abgegebenen Urteil bereithält.

Durch die so gewonnenen Erfahrungswerte bei der Einschätzung virtueller Schülerinnen und Schüler, könnte die angehende Lehrkraft lernen und zwar ohne dass eine starke Leistungsüber- oder -unterschätzung eine gravierende Auswirkung auf reale Lernende hätte.

5. Gesamtdiskussion

5.1. Zusammenfassung und allgemeine Diskussion der Ergebnisse

Im Fokus dieser Arbeit steht die Urteilsgenauigkeit von Lehrkräften, welche den Kern der diagnostischen Kompetenz ausmacht (Kaiser et al., 2012; Pit-ten Cate et al., 2014). Da die diagnostische Kompetenz eine wichtige Facette des Professionswissens von Lehrkräften (Baumert & Kunter, 2011) und einen wichtigen Faktor erfolgreichen Unterrichtens darstellt (Helmke, 2009), ist es bedeutsam, mehr über diese in Erfahrung zu bringen.

In dieser Arbeit wurden, neben der Untersuchung der Urteilsgenauigkeit (Schrader, 2009), auch Einflussvariablen der Urteilsgenauigkeit, die Veränderung der Urteilsgenauigkeit über die Zeit sowie Auswirkungen der Genauigkeit von Lehrkrafturteilen betrachtet. In allen drei Studien wurde die Urteilsgenauigkeit von Mathematiklehrkräften untersucht. Die ersten beiden Arbeiten fokussierten auf Realschullehrkräfte, wohingegen in der dritten Arbeit Grundschullehrkräfte analysiert wurden.

Im Folgenden werden die Ergebnisse der Studien zusammenfassend diskutiert. Dabei wird darauf eingegangen, an welchen Forschungsergebnissen die Studien ansetzten und welche Forschungsergebnisse bestätigt werden konnten. Des Weiteren wird darauf verwiesen, welche Aspekte im Rahmen der verschiedenen Studien zusätzlich untersucht wurden und welche Erkenntnisse gewonnen werden konnten.

5.1.1. Studie 1: Wie gut schätzen Lehrkräfte Leistung, Arbeits- und Sozialverhalten ihrer Schülerinnen und Schüler ein? Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften

Im Rahmen der ersten Studie, welche querschnittlich angelegt ist, wurde zum einen die Urteilsgenauigkeit betrachtet. Insbesondere wurde der Frage nachgegangen, wie genau Lehrkrafturteile zu Schülermerkmalen wie der Konzentration, dem Arbeits- und Sozialverhalten ausfallen. Diese Schülermerkmale wurden in dieser Studie erstmals explizit im Kontext der diagnostischen Urteilsgenauigkeit von Lehrkräften betrachtet. Zum anderen wurden Variablen betrachtet, die mit der Urteilsgenauigkeit zusammenhängen.

Im Einklang mit früheren Studienergebnissen zeigte sich, dass Lehrkräfte andere Schülermerkmale als die Schülerleistung schlechter in eine Rangreihe bringen konnten (Spinath, 2005; Urhahne et al., 2010; Urhahne & Zhu, 2015b). Die Differenzen zwischen den Korrelationskoeffizienten der Urteilsgenauigkeit der Mathematikleistung und der Konzentration sowie des Arbeits- und Sozialverhalten wurden in dieser Arbeit zusätzlich auf

Signifikanz untersucht. Lehrkräfte vermögen es demnach, die Schülerleistung signifikant besser einzuschätzen als die anderen untersuchten Schülermerkmale. Dieses Ergebnis deutet darauf hin, dass ein Training zur Steigerung der Urteilsgenauigkeit wünschenswert ist, damit Kopfnoten aber auch Empfehlungsschreiben fair sind.

Ebenfalls konnte bestätigt werden, dass Lehrkräfte dazu tendieren, die Schülerleistung zu überschätzen (Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003, 2009; Urhahne et al., 2010). Des Weiteren konnte auch in dieser Studie kein Zusammenhang zwischen Lehrkraftmerkmalen und der Urteilsgenauigkeit gefunden werden (Praetorius et al., 2011; Schrader, 1989; Südkamp et al., 2012).

Anknüpfend an die Arbeit von Schrader und Helmke (1990) wurde in dieser Studie untersucht, ob das Lehrkrafturteil mit urteilsirrelevanten Schülermerkmalen zusammenhängt. Allerdings lag der Fokus in dieser Arbeit auf in dem Kontext bisher noch nicht untersuchten Schülermerkmalen. Zusätzlich wurde überprüft, ob andere von der Lehrkraft vorgenommene Merkmalseinschätzungen einen Einfluss auf die Urteilsgenauigkeit haben. Es zeigte sich, dass die Lehrkrafturteile sowohl durch urteilsirrelevante Schülermerkmale als auch durch andere vorgenommene Lehrkrafteinschätzungen mitbestimmt wurden.

Abschließend ist festzuhalten, dass sich die Ergebnisse zur Genauigkeit des Lehrkrafturteils zur Mathematikleistung und zum Zusammenhang der Lehrkraftmerkmale mit der Urteilsgenauigkeit in die bestehenden Forschungsergebnisse einreihen lassen. Allerdings lieferte die Studie auch neue Erkenntnisse. Im Vordergrund stehen hier die Ergebnisse zur Genauigkeit des Lehrkrafturteils zu den Schülermerkmalen Konzentration, Arbeits- und Sozialverhalten sowie die Ergebnisse zu den Variablen, die die Lehrkrafturteile beeinflussen können.

5.1.2. Studie 2: Stabilität, Bezugsnormorientierung und Auswirkungen der Urteilsgenauigkeit

In der zweiten Studie, welche auf der ersten aufbaut, wurde wieder die Urteilsgenauigkeit betrachtet. Durch die Hinzunahme eines weiteren Messzeitpunktes stand in dieser Arbeit allerdings die Frage im Fokus, ob zwischen den Daten der beiden Messzeitpunkte Unterschiede bestehen, es also zu einer Veränderung in der Urteilsgenauigkeit kommt. Des Weiteren wurde im Rahmen der zweiten Arbeit untersucht, ob die Bezugsnormorientierung der Lehrkräfte in Zusammenhang mit der Urteilsgenauigkeit steht. Zusätzlich wurde nach

den Auswirkungen der Urteilsgenauigkeit gefragt und in dem Kontext ebenfalls das wahrgenommene Lehrkraftverhalten mituntersucht.

Anknüpfend an die erste Studie zeigte sich auch in der zweiten, dass die Rangkomponente, Übereinstimmung zwischen Lehrkrafturteil und Schülerleistung, niedriger als in Metaanalysen berichtet ausfiel (Hoge & Coladarci, 1989; Südkamp et al., 2012).

Neue Erkenntnisse konnten durch die Hinzunahme des zweiten Messzeitpunktes gewonnen werden. Bisher wurde die diagnostische Urteilsgenauigkeit von Lehrkräften fast ausschließlich querschnittlich betrachtet. Es existieren nur wenige Studien, welche zwei Messzeitpunkte umfassen (Lorenz & Artelt, 2009; Spinath, 2005; Südkamp et al., 2008). Die Hinzunahme des zweiten Messzeitpunktes ermöglichte es, Aussagen zur Veränderung über die Zeit und zur Auswirkung der Urteilsgenauigkeit zu treffen.

Zur Veränderung und Stabilität der Urteilsgenauigkeit über die Zeit ergab sich, dass die Rangkomponente zum zweiten Messzeitpunkt signifikant höher ausfiel als zum ersten Messzeitpunkt. Zusätzlich ergab sich eine Differenz in der Niveaukomponente, denn zum zweiten Messzeitpunkt wurde anders als zum ersten Messzeitpunkt die Schülerleistung signifikant unterschätzt. Dieses Ergebnis ist konträr zu den bisherigen Forschungsergebnissen, welche von einer Überschätzung der Schülerleistung ausgehen (Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003, 2009; Urhahne et al., 2010). Ebenfalls konträr zu bisherigen Forschungsergebnissen fielen die diachronen Zusammenhänge der Rangkomponente nicht signifikant aus (Lorenz & Artelt, 2009; Spinath, 2005). Zusätzlich wurde in dieser Arbeit die Stabilität der Niveau- und Differenzierungskomponente berechnet, welche ebenfalls nicht signifikant ausfielen.

In dieser Studie wurde zudem zum ersten Mal die Bezugsnormorientierung der Lehrkräfte in Zusammenhang mit der Urteilsgenauigkeit gesetzt. Es zeigte sich, dass ein Zusammenhang besteht. Legten die Lehrkräfte eher eine kriteriale Bezugsnormorientierung zu Tage, so war die Rangkomponente am höchsten ausgeprägt.

In den Arbeiten zur diagnostischen Kompetenz wurde selten nach den Auswirkungen einer Über- resp. Unterschätzung der Schülerleistung auf selbige gefragt. Ursächlich dafür ist, dass die meisten Arbeiten querschnittlich angelegt sind und hauptsächlich die Urteilsgenauigkeit untersuchten (Oerke et al., 2015; Schrader, 2009). Erste Arbeiten zu möglichen Auswirkungen von Lehrkrafturteilen liegen daher nur im Querschnitt vor. Diese

betrachten vornehmlich Unterschiede zwischen über- und unterschätzten Schülerinnen und Schülern und nehmen als ursächlich für die Unterschiede die Genauigkeit der Lehrkrafteinschätzung an, welche sich – vermittelt durch die sich selbst erfüllende Prophezeiung – auf die Schülerinnen und Schüler auswirkt (Urhahne et al., 2011; Urhahne et al., 2013; Urhahne et al., 2010). In dieser Arbeit zeigte sich im Längsschnitt, dass die Urteilsgenauigkeit zu Messzeitpunkt eins, indiziert als Leistungsüber- oder -unterschätzung, einen Einfluss auf die Schülerleistung zu Messzeitpunkt zwei hatte. Dies bedeutet, dass sich eine Leistungsüberschätzung der Lehrkraft positiv auf die Schülerleistung auswirkt. Eine leichte Überschätzung der Schülerleistung scheint demnach pädagogisch positiv zu bewerten sein (McElvany et al., 2009; Weinert & Schrader, 1986).

Anknüpfend an die Forschung zur sich selbsterfüllenden Prophezeiung (Jussim, 1986) wurde auch in dieser Arbeit das wahrgenommene Lehrkraftverhalten mitbetrachtet. Übereinstimmend mit der Forschung zeigten sich Unterschiede in der Wahrnehmung des Lehrkraftverhaltens zwischen über- und unterschätzten Schülerinnen und Schülern. So nahmen Überschätzte die Lehrkraft als unterstützender, zugänglicher und gerechter wahr (Harris et al., 1986; Urhahne, 2015).

Die Ergebnisse zur Auswirkung der Urteilsgenauigkeit und zum wahrgenommenen Lehrkraftverhalten stimmen mit den Ergebnissen aus der Forschungstradition der Lehrkrafterwartung überein (Babad et al., 1982; Cooper et al., 1982; de Boer et al., 2010; Jussim, 1986; Peterson et al., 2016). Die Ergebnisse dieser Studie indizieren, dass die Theorien und Forschungsergebnisse zu Lehrkrafterwartungen auf Lehrkrafturteile übertragbar sind.

Zusammenfassend konnten auch in dieser Studie bestehende Forschungsergebnisse bestätigt werden. Hierzu zählen die Ergebnisse zur Urteilsgenauigkeit sowie zum wahrgenommen Lehrkraftverhalten. Zur Veränderung und Stabilität der Urteilsgenauigkeit konnten sowohl ergänzende als auch neue Erkenntnisse gewonnen werden. Des Weiteren lieferte die Studie erste Ergebnisse zum Zusammenhang zwischen der Bezugsnormorientierung der Lehrkräfte und der Urteilsgenauigkeit. Die Ergebnisse zur Auswirkung der Urteilsgenauigkeit ergänzen die Erkenntnisse zur Lehrkrafterwartung und verdeutlichen, dass diese sich auch bei Lehrkrafturteilen zeigen. Zusätzlich erweitern die Ergebnisse die bestehenden Forschungsergebnisse zur Auswirkung von Lehrkrafturteilen, welche nur

einen Messzeitpunkt umfassen. Die Ergebnisse dieser Studie sind daher etwas aussagekräftiger.

5.1.3. Studie 3: Stabilität und Auswirkungen der Urteilsgenauigkeit von Grundschullehrkräften

Die Urteilsgenauigkeit, welche am häufigsten in der Forschungstradition der diagnostischen Kompetenz untersucht wurde (Schrader, 2009), wurde auch in der dritten Studie betrachtet. Da Lehrkräfte allerdings auch motivational-affektive Schülermerkmale einzuschätzen vermögen müssen, welche lern- und leistungsrelevant sind (Schrader, 2009), wurde analysiert wie genau das Urteil zu diesen Merkmalen ausfällt. Aufgrund dessen, dass auch diese Arbeit zwei Messzeitpunkte umfasst, konnte zusätzlich untersucht werden, ob sich die Genauigkeit des Lehrkrafturteils sowie die Schülermerkmale über die Zeit verändern. Des Weiteren wurden die Auswirkungen der Genauigkeit des Lehrkrafturteils untersucht.

Übereinstimmend mit den bisherigen Forschungsergebnissen zeigte sich auch in dieser Arbeit, dass es Lehrkräften leichter fällt die Schülerleistung als motivational-affektive Schülermerkmale einzuschätzen (Spinath, 2005; Urhahne et al., 2011; Urhahne et al., 2010). Als neues einzuschätzendes Schülermerkmal wurde in dieser Arbeit die Schuleinstellung aufgenommen. Des Weiteren konnte festgestellt werden, kongruent zu bisherigen Studienergebnissen, dass Lehrkräfte zur Überschätzung der Schülerleistung tendieren (Bates & Nettelbeck, 2001; Feinberg & Shapiro, 2003, 2009; Urhahne et al., 2010).

Zur Stabilität und Veränderung der Urteilsgenauigkeit über die Zeit zeigte sich in dieser Arbeit, dass die Rangkomponenten der Mathematikleistung und Erfolgserwartung zeitstabil sind (Lorenz & Artelt, 2009). Auch wurde in dieser Arbeit die Stabilität der Niveau- und Differenzierungskomponente berechnet. Die Differenzierungskomponente der Testleistung sowie die Niveaukomponente des Anspruchsniveaus waren zeitstabil. Allerdings veränderte sich die Urteilsgenauigkeit über die Zeit. Lehrkräfte konnten die Schülerleistung (Stang & Urhahne, 2016b) sowie die Schülermerkmale Lernfreude, Schuleinstellung und Anstrengungsbereitschaft zum zweiten Messzeitpunkt signifikant besser einschätzen. Zwischen den Niveau- und Differenzierungskomponenten ergaben sich keine signifikanten Differenzen.

Die Stabilität der Genauigkeit des Lehrkrafturteils zur Mathematikleistung in dieser Arbeit steht nicht in Einklang mit dem Teilergebnis der zweiten Studie, in welcher die Genauigkeit des Lehrkrafturteils zur Mathematikleistung nicht zeitstabil ausfiel. Allerdings handelt es sich um Einzelbefunde (Lorenz & Artelt, 2009; Stang & Urhahne, 2016b), welche schwer miteinander zu vergleichen sind. Dies liegt daran, dass sowohl die Stichprobengrößen als auch die Stichproben selbst, Grundschullehrkräfte bei Spinath (2005), Lorenz und Artelt (2009) sowie in dieser Teilstudie vs. Realschullehrkräfte bei Stang und Urhahne (2016b), stark variieren. Ein entsprechender Vergleich der Ergebnisse würde eine Messinvarianzanalyse voraussetzen, welche jedoch, aufgrund der Größe des Datensatzes, nicht durchführbar ist. Insgesamt ist es daher bedeutsam, die Ergebnisse zu replizieren, um zu generalisierbaren Ergebnissen zu kommen.

Zur Stabilität der Schülermerkmale konnte herausgefunden werden, dass diese über den Zeitraum eines Jahres relativ stabil sind (Spinath, 2005). Lediglich sank die Lernfreude in dem Zeitraum signifikant ab, was in Einklang mit der Forschung zur Entwicklung von schulbezogener Emotion und Motivation steht (Jerusalem & Mittag, 1999).

Da bisher erst wenige Forschungsergebnisse zur Auswirkung der Genauigkeit von Lehrkrafturteilen existieren, fokussierte auch diese Arbeit auf diesen Aspekt. Festgestellt wurde, dass sich überschätzte von unterschätzten Schülerinnen und Schülern unterscheiden. Während in der dritten Klasse keine Unterschiede in der Schülerleistung bestehen, zeigten sich in der vierten Klasse signifikante Unterschiede. Des Weiteren wiesen überschätzte Schülerinnen und Schüler ein positiveres motivational-affektives Muster auf (Urhahne et al., 2011; Urhahne et al., 2010; Zhu & Urhahne, 2015). All den Studien und dieser Arbeit ist gemein, dass insbesondere in den Variablen Erfolgserwartung, Fähigkeitsselbstkonzept und Leistungsangst Unterschiede zwischen Über- und Unterschätzten bestehen. Auch diese Arbeit deutet darauf hin, dass es pädagogisch positiv zu sein scheint, wenn die Schülerleistung überschätzt wird (McElvany et al., 2009; Weinert & Schrader, 1986).

Die Ergebnisse zur Auswirkung der Urteilsgenauigkeit auf die Schülerleistung und auf motivational-affektive Schülermerkmale stehen im Einklang mit den Ergebnissen aus der Forschungstradition der Lehrkrafterwartung (Babad et al., 1982; de Boer et al., 2010; Jussim, 1986; Jussim, 1989; Jussim & Harber, 2005). Die Ergebnisse dieser Studie zeigen, dass die Theorien und Forschungsergebnisse zu Lehrkrafterwartungen auf Lehrkrafturteile

anwendbar sind. Das Modell der sich selbst erfüllenden Prophezeiung eignet sich zur Erklärung der Befunde.

Abschließend ist festzuhalten, dass die dritte Studie sowohl Ergebnisse brachte, welche die bisherigen Forschungsergebnisse bestätigen, als auch neue Erkenntnisse lieferte. So reihen sich die Ergebnisse zur Urteilsgenauigkeit in bereits bestehende Forschungsergebnisse ein, ebenso wie die Ergebnisse zur Veränderung von Schülermerkmalen innerhalb eines Schuljahres. Weitere Erkenntnisse lieferte die Arbeit zur Veränderung und Stabilität der Urteilsgenauigkeit. Durch das längsschnittliche Vorgehen konnten neue Erkenntnisse zur Auswirkung der Genauigkeit von Lehrkrafturteilen auf die Schülerleistung und auf motivational-affektive Schülermerkmale erbracht werden.

5.2. Ausblick und Implikationen

Aus den dargestellten Ergebnissen wird deutlich, dass die Urteilsgenauigkeit von Lehrkräften noch nicht vollumfassend untersucht wurde. So konnten bestehende Forschungsfragen noch nicht abschließend beantwortet werden. Auch verweisen die berichteten Ergebnisse auf bestehende Forschungslücken und liefern Anregungen für weiterführende Forschungsarbeiten.

In allen drei Arbeiten zeigten sich, in Übereinstimmung mit bisherigen Forschungsarbeiten, starke interindividuelle Unterschiede in der Genauigkeit der Lehrkrafturteile (Schrader & Helmke, 1987; Urhahne et al., 2013). Die zum Teil niedrigen Korrelationskoeffizienten lassen eine Verbesserung der Urteilsgenauigkeit wünschenswert und die Frage nach Einflussvariablen auf die Urteilsgenauigkeit wichtiger denn je erscheinen. Außerdem stellt sich in dem Kontext die Frage, wie Lehrkräfte ihre Urteilsfähigkeit selber wahrnehmen. Daher wird im Folgenden auf Möglichkeiten der Förderung der Urteilsgenauigkeit von Lehrkräften, auf potenzielle Untersuchungsansätze zu Einflussvariablen der Urteilsgenauigkeit sowie auf eine mögliche Selbsteinschätzung der Urteilsfähigkeit eingegangen.

5.2.1. Förderung der Urteilsgenauigkeit von Lehrkräften

Da die Genauigkeit von Lehrkrafturteilen u. a. essentiell für das Lernergebnis von Schülerinnen und Schülern ist, ist es wichtig, über die Entwicklung und Förderung der Urteilsgenauigkeit nachzudenken (Ohle & McElvany, 2015). Zur Förderung der Urteilsgenauigkeit stehen mehrere Möglichkeiten zur Verfügung. Einige Optionen wurden bereits in der

Praxis erprobt, andere sind noch hypothetischer Natur. In den folgenden Abschnitten werden die verschiedenen Möglichkeiten vorgestellt.

Training

Eine Möglichkeit, die Urteilsgenauigkeit zu steigern ist die Entwicklung und Implementation eines Trainings. Trainings lassen sich durch systematisches Durchführen von Einheiten und Übungen charakterisieren, die der Steigerung oder dem Erwerb einer spezifischen Fähigkeit oder Fertigkeit dienen (Altmann, 2014). So werden z. B. in der Personalpsychologie Beurteiler trainiert, um die Leistungen von Mitarbeiterinnen und Mitarbeitern eines Unternehmens adäquat einschätzen zu können (Lohaus, 2009).

Urteile können auf verschiedene Art und Weise verzerrt sein (Lohaus, 2009; Südkamp & Möller, 2009; Urhahne et al., 2013). Zu den Ursachen dieser Verzerrungen zählen Fehler und Tendenzen wie der Halo-Effekt, der logische Fehler oder die Tendenz zur Mitte (Helmke, 2009). Um diesen Urteilsverzerrungen effizient entgegen zu treten, bieten sich Trainingsformen an, dessen Wirkungen bereits untersucht wurden.

Zur Verringerung von Urteilsfehlern und -tendenzen haben sich insbesondere zwei Trainingsformen als wirksam erwiesen (Pulakos, 1984; Roch, Woehr, Mishra & Kieszczynska, 2012; Woehr & Huffcutt, 1994). Bei dem *Rater-Error-Training* werden Urteilsfehler direkt thematisiert. Verschiedene Fehler werden dabei erläutert und die Entstehungsbedingungen beleuchtet. Wichtige Trainingsbausteine sind Urteilsfehler und -tendenzen zu erklären und zu diskutieren, nach Möglichkeit eigene Einschätzungen zu visualisieren, Beispiele zu besprechen sowie Diskussionen darüber zu führen, warum eine Person wie bewertet wurde (Cellar, Curtis, Kohlepp, Poczapski & Mohiuddin, 1989; Pulakos, 1984). Bei dem *Frame-of-Reference-Training* erarbeiten die Urteilenden gemeinsam einen Bezugsrahmen zur Bewertung des einzuschätzenden Merkmals und Indikatoren, welche für die verschiedenen Merkmalsabstufungen kennzeichnend sind. Die erarbeiteten Indikatoren und Bezugsrahmen können dann zu einem späteren Zeitpunkt zur Einschätzung diverser Merkmale herangezogen werden, so dass den Urteilenden gleiche Maßstäbe zur Verfügung stehen. Wichtige Trainingselemente sind die Herausarbeitung der Indikatoren, Übungs- und Feedbackeinheiten und die Erarbeitung des gemeinsamen Bezugsrahmens (Athey & McIntyre, 1987; Bernardin & Buckley, 1981; Day & Sulsky, 1995; McIntyre, Smith & Hassett, 1984).

Eine Kombination beider Trainingsformen könnte positiv sein. Die Thematisierung der Urteilsverzerrungen kann diese minimieren und dafür sensibilisieren. Somit kann der Problematik, dass Lehrkrafturteile verzerrt sein können (Südkamp & Möller, 2009; Urhahne et al., 2013), entgegen gewirkt werden. Die Erarbeitung von Indikatoren ist insbesondere dann angezeigt, wenn es darum geht, andere Schülermerkmale als die Schülerleistung zu bewerten. Es scheint für Lehrkräfte schwierig zu sein, valide Indikatoren für nicht-leistungsbezogene Schülermerkmale zu identifizieren (Helmke & Fend, 1981). Ein Training könnte daher Abhilfe schaffen, so dass nach Absolvierung des Trainings die relevanten, verfügbaren Schülerinformationen von der Lehrkraft erkannt und zur Urteilsgenerierung herangezogen werden könnten, da sie mit entsprechenden Indikatoren vertraut ist (Funder, 1995, 2012).

Trittel, Gerich und Schmitz (2014) entwickelten ein Trainingsprogramm, welches in Seminarform durchgeführt wurde. Ziel des Programmes war die Steigerung der diagnostischen Urteilsfähigkeit von angehenden Grundschullehrkräften. Dementsprechend wurden sie im diagnostischen Bereich geschult. Thematisiert wurden u. a. Urteilsfehler und das Sammeln von diagnostisch relevanten Informationen. Am Ende der Schulungseinheit schnitt die Experimental- im Vergleich zur Kontrollgruppe, welche kein Training erhielt, signifikant besser ab. Des Weiteren bestand nur bei der Experimentalgruppe ein signifikant positiver Zusammenhang zwischen dem angewandten Wissen und der diagnostischen Kompetenz (Trittel et al., 2014). Demgemäß scheint es lohnenswert, ein Training zu entwickeln oder ein bestehendes weiterzuentwickeln, welches Elemente sowohl des Rater-Error- als auch des Frame-of-Reference-Trainings beinhaltet. Dieses könnte dann integraler Bestandteil der Lehramtsausbildung und in Weiterbildungsveranstaltungen für Lehrkräfte implementiert werden.

Tagebuch

Eine weitere Option zur Steigerung der Urteilsgenauigkeit könnte im Einsatz von Tagebüchern liegen. Tagebücher dienen dazu, sowohl das Verhalten als auch das Erleben von Personen in bestimmten Situationen zu erfassen (Nohe, Peters & Sonntag, 2014). Tagebücher sind wirkungsvoll, da sie sich das Self-Monitoring, die absichtsvolle Eigenüberwachung (Lan, 1996), zu Nutze machen (Klug, Gerich, Bruder & Schmitz, 2012). So werden z. B. im pädagogisch-psychologischen Kontext Tagebücher eingesetzt, um das selbstregulierte Lernen von Schülerinnen und Schülern zu messen (Schmitz & Wiese, 2006).

Klug et al. (2012) setzten ein standardisiertes Tagebuch zur Förderung der diagnostischen Urteilsgenauigkeit von Hauptschullehrkräften ein. Es zeigte sich jedoch, dass das Tagebuch keinen Effekt auf die Kompetenz hatte. Dies mag an dem hohen Standardisierungsgrad der Fragen gelegen haben. Dementsprechend wäre die Entwicklung eines optimierten Tagebuches zur Steigerung der Urteilsgenauigkeit wünschenswert.

Der von Klug et al. (2012) entwickelte standardisierte Fragenkatalog könnte dabei um Fragen mit freiem Antwortformat oder gar um Reflexionsaufgaben ergänzt werden. Offene Fragen erlauben mehr Reflexionsprozesse (Klug et al., 2012). Reflexionsaufgaben regen des Weiteren dazu an, sich im Sinne eines *reflective practitioners* (Schön, 1983) mit den eigenen Stärken und Schwächen auseinander zu setzen. Zusätzlich könnte als Bestandteil des Tagebuchs der von Helmke (2009) weiterentwickelte Zyklus zur Förderung der diagnostischen Fähigkeiten aufgenommen werden. Dieser sieht vor, ein Schülermerkmal auszuwählen, dieses zu beurteilen sowie zu messen. In einem weiteren Schritt kann so die eigene Einschätzung mit der Messung abgeglichen werden. Zusätzliche Reflexionsaufgaben helfen, mögliche Diskrepanzen zwischen Einschätzung und Messung zu analysieren (Helmke, 2009). Auch können so die der Urteilsgenerierung zugrundeliegenden kognitiven Prozesse analysiert und gestärkt werden, was wiederum zu einer Erhöhung der Akkuratheit führen kann (Pit-ten Cate et al., 2014).

Simulierter Klassenraum

Auch der simulierte Klassenraum stellt eine Möglichkeit dar, die Urteilsgenauigkeit von Lehrkräften zu steigern. Bei dem simulierten Klassenraum handelt es sich um eine Computersimulationssoftware. In dieser Simulation agiert der Proband als Lehrkraft und interagiert mit virtuellen Lernenden (Südkamp et al., 2008). Der simulierte Klassenraum wird dazu genutzt, die diagnostische Urteilsgenauigkeit von Lehrkräften experimentell und systematisch zu untersuchen (Südkamp & Möller, 2009; Südkamp et al., 2008).

In der Lehreraus- und Weiterbildung könnte diese Software effizient eingesetzt werden, um zukünftige sowie sich bereits im Beruf befindende Lehrkräfte in der Genauigkeit der Schülereinschätzung zu trainieren. An virtuellen Schülerinnen und Schülern könnte die Beurteilung von nicht-kognitiven Schülermerkmalen geübt werden, welches eine Steigerung der Akkuratheit mit sich bringen würde. Des Weiteren könnten anhand dieser Software auch Ergebnisse eines Trainings evaluiert werden.

Clicker im Klassenzimmer

Learner-Response-Systeme, auch Clicker genannt, stellen ebenfalls eine denkbare Möglichkeit dar, die Urteilsgenauigkeit von Lehrkräften zu fördern. Bei der Clicker-Methode können die Lernenden mithilfe eines kleinen elektronischen Geräts direkt auf von der Lehrkraft gestellte Fragen und Aufgaben antworten. Die Schülerantworten werden auf dem Laptop der Lehrkraft gesammelt und gespeichert, so dass die Lehrkraft detailliert Rückmeldung über den Leistungsstand der einzelnen Schülerinnen und Schüler erhält. Clicker werden seit den 1960er Jahren im Schulkontext eingesetzt. Studien zeigen, dass der Einsatz der Clicker-Methode sich lern- und leistungsförderlich auswirkt (Cohn & Fraser, 2016; Hunsu, Adesope & Bayly, 2016; Landrum, 2015; Mayer et al., 2009).

Beim Einsatz von Clickern im Klassenzimmer erhält die Lehrkraft Rückmeldung über das Leistungsniveau ihrer Schülerinnen und Schüler. Diese Information erleichtert es der Lehrkraft, auf Schwächen einiger Schülerinnen und Schüler gezielter einzugehen und diese mit geeigneten didaktischen Mitteln abzubauen. Ein mehrmaliger wöchentlicher Einsatz, gestreckt über mehrere Wochen, versorgt Lehrkräfte demnach mit wichtigen Informationen und könnte zudem die Leistungsvorhersage erleichtern, da Lehrkräfte über mehr Wissen verfügen. Dies würde bedeuten, dass die Urteilsgenauigkeit gesteigert werden könnte.

Neben der Rückmeldung der Schülerleistung könnten auch Fragen zu motivational-affektiven Schülermerkmalen gestellt werden. Da einige Schüleremotionen wie z. B. Leistungsangst mit Scham behaftet sind, könnte es Schülerinnen und Schülern leichter fallen, solche Ängste über ein Learner-Response-System rückzumelden. Auch hier könnte die den Lehrkräften zur Verfügung gestellte Rückmeldung über das motivational-affektive Schülererleben die Merkmalseinschätzung erleichtern und zu genaueren Urteilen führen. Zusätzlich könnte versucht werden, das subjektive Wohlbefinden der Schülerinnen und Schüler zu fördern, welches Einfluss auf die Schülerleistung haben kann (Hascher, 2005). Des Weiteren könnte auch der Frage nachgegangen werden, ob mehr Informationen über die Schülerinnen und Schüler den Lehrkräften helfen, die Motivation und das Interesse der Schülerinnen und Schüler genauer einzuschätzen und ob eine höhere Urteilsgenauigkeit dazu führen kann, motivational-affektive Schülermerkmale zu fördern oder bei Problemen besser intervenieren zu können.

Feedback

Die Gabe von gezieltem und detailliertem Feedback könnte eine weitere Möglichkeit sein, die Genauigkeit von Lehrkrafturteilen zu steigern. Unter Feedback versteht man das Übermitteln von Informationen, welche ggf. zur Korrektur eines Verhaltens genutzt werden können. Basis des Feedbacks ist dabei ein Vergleich zwischen Ist- und Soll-Zustand (Woolfolk, 2014). In verschiedenen Bereichen und Kontexten wird Feedback gegeben, wie z. B. im schulischen Kontext, wo Feedback – je nach Art und Umfang – positive wie negative Folgen auf die Schülerleistung haben kann (Hattie & Timperley, 2007; Narciss, 2014).

In den hier vorgestellten Arbeiten erhielten die Lehrkräfte nach den Erhebungen Rückmeldung über die Ausprägung ihrer Rang- und Niveaukomponente sowie über die Schülerleistung und die erfassten Schülermerkmale. Rückmeldungen unterscheiden sich allerdings von Feedback in verschiedenen Kernaspekten. Im Vergleich zu Feedback sind Rückmeldungen z. B. eher verhaltens- und zeitfern (Müller & Ditton, 2014). An dieser Stelle müsste angesetzt werden, um zu prüfen, ob sich Rückmeldungen auf die Urteilsgenauigkeit auswirken.

Ebenso sollte im Kontext der diagnostischen Urteilsgenauigkeit auch die Gabe von zeit- und verhaltensnahem Feedback untersucht werden, um die Frage beantworten zu können, ob sich Feedback positiv auf die Urteilsgenauigkeit auswirken kann. Bei zeitnaher Gabe von Feedback, ist es womöglich einfacher, eine mögliche Diskrepanz zwischen Ist- und Soll-Zustand zu analysieren und zu reflektieren. Vorausgesetzt, das Feedback regt auch zur Reflexion an. Zudem sollte das Feedback nach Möglichkeit detailliert resp. schülerbezogen sein, so dass die Lehrkraft darüber gewahr wird, wen sie über-, unter- oder richtig einschätzt. Reflektorisch sollte dann erarbeitet werden, warum es zu einer Leistungsüber- oder -unterschätzung kommt, um z. B. den Effekt der sich selbsterfüllenden Prophezeiung zu unterbinden (Pit-ten Cate et al., 2014).

5.2.2. Einflussvariablen der Urteilsgenauigkeit

Im Rahmen von Studien zur diagnostischen Kompetenz von Lehrkräften wurden bereits einige Einflussvariablen der Urteilsgenauigkeit wie z. B. Test-, Urteils-, Schüler- und Lehrkraftmerkmale untersucht (Südkamp et al., 2012). Allerdings ist die Forschungslage zu den untersuchten Variablen recht heterogen (Südkamp et al., 2012), weshalb auch andere Einflussvariablen in Betracht gezogen werden sollten.

Zur Untersuchung von Einflussvariablen böte sich eine Computersimulationssoftware wie das simulierte Klassenzimmer an. Der Einsatz einer Computersimulationssoftware bietet mehrere Vorteile. Zum einen können mehr Personen in einem kürzeren Zeitraum getestet werden. Zum anderen erfolgt die Testung anhand von virtuellen Schülerinnen und Schülern, so dass z. B. Leistungsunterschätzungen keine Auswirkungen haben. Hieran schließt sich allerdings die Frage der Generalisierbarkeit solcher im Labor gewonnen Ergebnisse an, welche hier allerdings nicht näher beleuchtet werden soll.

Eine potenzielle und interessante Einflussvariable stellt der Zeitdruck dar. Lehrkräfte werden tagtäglich mit sehr verschiedenen Anforderungen und Verpflichtungen konfrontiert. So müssen sie u. a. ihre Stunden vorbereiten, Prüfungen und Tests erstellen sowie korrigieren, aber auch verschiedene Angelegenheiten von Schülerinnen und Schülern klären. Aus verschiedenen Forschungsarbeiten ist allerdings bekannt, dass es unter Zeitdruck häufiger zu Urteilsverzerrungen kommt. Insbesondere ist das Auftreten des Halo-Effekts auf Zeitdruck zurückzuführen (Klauer & Schmeling, 1990). Mit der Computersimulationssoftware könnte Zeitdruck bei der Bearbeitung verschiedener diagnostischer Urteilsaufgaben variiert und der Zusammenhang zwischen Zeitdruck und Urteilsgenauigkeit analysiert werden.

Des Weiteren könnte auch Feedback systematisch im Rahmen einer Computersimulationssoftware variiert und auf ökonomische Art und Weise der Zusammenhang zwischen Feedback und Urteilsgenauigkeit untersucht werden, so dass Aussagen über die verschiedenen Feedbackarten und deren Wirkung auf die Urteilsgenauigkeit erbracht werden könnten.

5.2.3. Selbsteinschätzung der Urteilsfähigkeit

Im Kontext der Urteilsgenauigkeit wurde bisher noch nicht die subjektive Selbsteinschätzung der eigenen diagnostischen Fähigkeiten untersucht. Studierende, Referendare sowie Lehrkräfte könnten z. B. in einer querschnittlich angelegten Studie befragt werden (Anhang C). Die Reflexion über die eigenen Fähigkeiten ist dabei auf der Metaebene zu verankern. Um die subjektive Selbsteinschätzung bewerten zu können, sollte diese an einer objektiven Überprüfung relativiert werden. Neben der Relativierung der Selbsteinschätzung an realen Testergebnissen von Schülerinnen und Schülern bietet sich z. B. die Einschätzung der Schwierigkeit von Aufgaben eines standardisierten Tests an (Anhang C). Überprüft werden könnte, ob die Selbsteinschätzung mit der Urteilsgenauigkeit zusammenhängt und ob die Selbsteinschätzung bei zunehmender Praxiserfahrung besser wird.

Die Fragen zur Selbsteinschätzung der eigenen diagnostischen Fähigkeiten (Anhang C) sind dabei so konstruiert, dass sie in die verschiedenen Kompetenzfacetten, welche im Modell von Baumert und Kunter (2006, 2011) als essentiell für die diagnostischen Fähigkeiten von Lehrkräften beschrieben werden, fallen. So lassen sich die Fragen z. T. in die Bereiche *Wissen über das mathematische Denken von Schülerinnen und Schülern*, *Wissen über mathematische Aufgaben* und *Wissen um Leistungsbeurteilung* einordnen.

Die Beurteilung der Aufgabenschwierigkeit stellt ebenfalls eine Anforderung dar, die an Lehrkräfte gestellt wird. Im Schulalltag müssen Lehrkräfte in der Lage sein, die Schwierigkeit von Aufgaben richtig einzuschätzen, um entsprechende Aufgaben für das Leistungsniveau ihrer Schülerinnen und Schüler auswählen zu können (Südkamp et al., 2012). Die Einschätzung der Aufgabenschwierigkeit wurde im Rahmen der Forschung zur Urteilsgenauigkeit weitaus weniger untersucht als die Einschätzung von Schülermerkmalen. Dementsprechend liegen zur Beurteilung der Aufgabenschwierigkeit weniger Studien vor (z. B. Hoffmann & Böhme, 2014; McElvany et al., 2009), weswegen es bedeutsam ist, auch in diesem Bereich Forschung zu betreiben.

5.2.4. Schlussbemerkung

Einhergehend mit dem sogenannten Theorie-Praxis-Dilemma, welches die Problematik beschreibt, dass zwischen Wissenschaftlern und Praktikern eine Kluft besteht, widergespiegelt als Lücke zwischen empirischen Erkenntnissen und der Umsetzung derer in der Praxis, stellt sich die Frage, wie wissenschaftliche Erkenntnisse in den schulischen Alltag oder in die Aus- und Weiterbildung von Lehrkräften transferiert werden können. Für ein erfolgreiches Transferieren sind Stark und Mandl (2007) zufolge entsprechende, zweckgerichtete Förderprogramme erforderlich, welche zuvor vorgestellt wurden.

Um die Wirksamkeit der verschiedenen vorgestellten Möglichkeiten zur Verbesserung der Urteilsgenauigkeit zu prüfen bietet sich ein Pretest-Posttest-Kontrollgruppenplan an. Beim Vergleich der Ergebnisse von Interventions- und Kontrollgruppe können z. B. moderierende Effekte wie der Zeitraum des Kennens einer Klasse herausgerechnet werden. Die Ergebnisse können dann in einem weiteren Schritt für die Praxis aufbereitet werden.

Wünschenswert wäre es, dass mit der Förderung der Urteilsgenauigkeit bereits im Lehramtsstudium begonnen werden sollte, da dort das Fundament für gute diagnostische Urteilsfähigkeiten und den weiteren Ausbau derer gelegt werden kann. Im Rahmen der Lehramtsausbildung sollte daher stärker auf die Schülerbeurteilung fokussiert werden – auch anhand des Einsatzes von Computersimulationssoftware, so dass an virtuellen Lernenden geübt werden kann und es im Berufsleben nicht zu negativen Effekten einer sich selbsterfüllenden Prophezeiung kommt – etwa durch Unterschätzung der Schülerleistung.

Eine genauere Einschätzung der Schülerleistung sowie der nicht-leistungsbezogenen Schülermerkmale ist bedeutsam und erstrebenswert. Einerseits um negative Auswirkungen einer Unterschätzung, andererseits um die schulbezogene Motivation und Emotion der Schülerinnen und Schüler besser fördern zu können, was dem für den Primarstufenbereich geforderten Bildungsziel der Entwicklung und Beibehaltung einer positiven motivational-affektiven Schuleinstellung dienlich ist. Dadurch, dass motivational-affektive Schülermerkmale wichtige Determinanten der schulischen Leistung sind, ergibt sich ebenfalls die Bedeutsamkeit der akkuraten Einschätzung dieser Merkmale.

6. Literaturverzeichnis

Hinweis: Das Literaturverzeichnis umfasst auch die in den in Fachzeitschriften veröffentlichen Studien enthaltene Literatur.

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

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Altmann, T. (2014). Training. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (17. Aufl., S. 1678). Bern: Huber.

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

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

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

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In D. PISA-Konsortium (Hrsg.), *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 67–137). Opladen: Leske + Budrich.

Athey, T. R. & McIntyre, R. M. (1987). Effect of rater training in rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72*, 567–572.

Babad, E. (1993). Teachers' differential behavior. *Educational Psychology Review, 5*, 347–376.

Babad, E., Avni-Babad, D. & Rosenthal, R. (2003). Teachers' brief nonverbal behaviors in defined instructional situations can predict students' evaluations. *Journal of Educational Psychology, 95*, 553–562.

Babad, E., Inbar, J. & Rosenthal, R. (1982). Pygmalion, Galatea and the Golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology, 74*, 459–474.

Baeriswyl, F., Wandeler, C. & Trautwein, U. (2011). Auf einer anderen Schule oder bei einer anderen Lehrkraft hätte es für's Gymnasium gereicht. *Zeitschrift für Pädagogische Psychologie, 25*, 39–47.

Bailey, A. L. & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment, 11*, 149–178.

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

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

Baumert, J. & Kunter, M. (2011). Das Kompetenzmodell von COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 30–53). Münster: Waxmann.

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.

Bernardin, H. J. & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*, 205–212.

Bilz, L., Steger, J. & Fischer, S. M. (2016). Die Genauigkeit des Lehrerurteils bei der Identifikation von an Mobbing beteiligten Schülerinnen und Schülern. *Psychologie in Erziehung und Unterricht, 63*, 122–136.

Blöschl, L. (1966). BTS, HAWIK und schulisches Arbeitsverhalten. *Diagnostica, 12*, 47–52.

Bos, W., Voss, A., Lankes, E.-M., Schwippert, K., Thiel, O. & Valtin, R. (2004). Schullaufbahnempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangs-

stufe. 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. 191–228). Münster: Waxmann.

Brickenkamp, R. & Karl, G. A. (1986). Geräte zur Messung von Aufmerksamkeit, Konzentration und Vigilanz. In R. Brickenkamp (Hrsg.), *Handbuch apparativer Verfahren in der Psychologie* (S. 195–211). Göttingen: Hogrefe.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*, 123–142.

Brophy, J. E. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75*, 631–661.

Brophy, J. E. (1998). Introduction. In J. E. Brophy (Ed.), *Expectations in the classroom: Vol. 7. Advances in research on teaching* (pp. ix–xvii). Greenwich, CT: Jai Press.

Brophy, J. E. & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 328–375). New York, NY: McMillan.

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

Caldarella, P. & Merrell, K. W. (1997). Common dimensions of social skills of children and adolescents: A taxonomy of positive behaviours. *School Psychology Review, 26*, 264–278.

Cellar, D. F., Curtis, J. R., Kohlepp, K., Poczapski, P. & Mohiuddin, S. (1989). The effect of rater training, job analysis format and congruence of training on job evaluation ratings. *Journal of Business and Psychology, 3*, 387–401.

Chomsky, N. (1968). *Language and mind*. Cambridge: Cambridge University Press.

Clark, C. M. & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 255–296). New York, NY: Macmillian.

Cohn, S. T. & Fraser, B. J. (2016). Effectiveness of student response systems in terms of learning environment, attitudes and achievement. *Learning Environments Research, 19*, 153–167.

Cooper, H., Findley, M. & Good, T. (1982). Relations between student achievement and various indexes of teacher expectations. *Journal of Educational Psychology, 74*, 577–579.

Credé, M. & Kuncel, N. R. (2008). Study habits, skills, and attitudes. The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science, 3*, 425–453.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177–193.

Dalbert, C., Schneidewind, U. & Saalbach, A. (2007). Justice judgments concerning grading in school. *Contemporary Educational Psychology, 32*, 420–433.

Day, D. V. & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*, 158–167.

de Boer, H., Bosker, R. J. & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology, 102*, 168–179.

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

Dicke, A.-L., Lüdtke, O., Trautwein, U., Nagy, G. & Nagy, N. (2012). Judging students' achievement goal orientations: Are teacher ratings accurate? *Learning and Individual Differences, 22*, 844–849.

Dickhäuser, O. & Galfe, E. (2004). Besser als…, schlechter als…: Leistungsbezogene Vergleichsprozesse in der Grundschule. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 36*, 1–9.

Düker, H., Lienert, G. A., Lukesch, H. & Mayrhofer, S. (2001). *Konzentrations-Leistungs-Test – Revidierte Fassung* (KLT–R). Göttingen: Hogrefe.

DuPaul, G. J., Rapport, M. D. & Perriello, L. M. (1991). Teacher ratings of academic skills: The development of the Academic Performance Rating Scale. *School Psychology Review, 20*, 284–300.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C. & Mac Iver, D. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in school and in families. *American Psychologist, 48*, 90–101.

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.

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

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

Frey, K. A. (2013). *Soziale Kompetenz. Eine Fragebogenerfassung in der Grundschule*. Münster: Waxmann.

Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K. & Trautwein, U. (2015). Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement. *Contemporary Educational Psychology, 41*, 1–12.

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

Funder, D. C. (2012). Accuracy of personality judgment. *Current Directions in Psychological Sciences, 21*, 177–182.

Gear, G. H. (1976). Accuracy of teacher judgment in identifying intellectually gifted children: A review of the literature. *The Gifted Child Quarterly, 20*, 478–490.

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.

Gölitz, D., Roick, T. & Hasselhorn, M. (2006). *DEMAT 4 – Deutscher Mathematiktest für vierte Klassen*. Göttingen: Beltz.

Götz, L., Lingel, K. & Schneider, W. (2013). *DEMAT 5 – Deutscher Mathematiktest für fünfte Klassen*. Göttingen: Beltz.

Gottfried, A. E., Fleming, J. S. & Gottfried, A. W. (2001). Continuity of academic intrinsic motivation from childhood through late adolescence: A longitudinal study. *Journal of Educational Psychology, 93*, 3–13.

Hanisch, G. (2004). Messung von Schulangst. *Erziehung und Unterricht, 154*, 897–902.

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

Harris, M. J. & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects – 31 meta-analyses. *Psychological Bulletin, 97*, 363–386.

Harris, M. J., Rosenthal, R. & Snodgrass, S. E. (1986). The effects of teacher expectations, gender and behavior on pupil academic performance and self-concept. *Journal of Educational Research, 79*, 173–179.

Hascher, T. (2005). Emotionen im Schulalltag: Wirkungen und Regulationsformen. *Zeitschrift für Pädagogik, 51*, 610–625.

Hascher, T. (2008). Diagnostische Kompetenzen im Lehrerberuf. In C. Kraler & M. Schratz (Hrsg.), *Wissen erwerben, Kompetenzen entwickeln. Modelle zur kompetenzorientierten Lehrerbildung* (S. 71–86). Münster: Waxmann.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Helmke, A. (1993). Die Entwicklung der Lernfreude vom Kindergarten bis zur 5. Klassenstufe. *Zeitschrift für Pädagogische Psychologie, 7*, 77–86.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität*. Seelze-Velber: Klett-Kallmeyer.

Helmke, A. & Fend, H. (1981). Wie gut kennen Eltern ihre Kinder und Lehrer ihre Schüler? Ergebnisse der Konstanzer Untersuchungen zu Bedingungen der Diagnosegenauigkeit bei Eltern und Lehrern. In G. Zimmer (Hrsg.), *Persönlichkeitsentwicklung und Gesundheit im Schulalter: Gefährdungen und Prävention* (S. 341–360). Frankfurt am Main: Campus.

Helmke, A. & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education, 3*, 91–98.

Helmke, A. & Schrader, F.-W. (2010). Determinanten der Schulleistung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl., S. 90–102). Weinheim: Beltz.

Helwig, R., Anderson, L. & Tindal, G. (2001). Influence of elementary student gender on teachers' perceptions of mathematics achievement. *Journal of Educational Research, 95*, 93–102.

Hinnant, J. B., O'Brien, M. O. & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*, 662–670.

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

Hoffmann, L. & Böhme, K. (2014). Wie gut können Grundschullehrkräfte die Schwierigkeit von Deutsch- und Mathematikaufgaben beurteilen? Eine Untersuchung zur Genauigkeit aufgabenbezogener Lehrerurteile auf Klassenebene. *Psychologie in Erziehung und Unterricht, 61*, 42–55.

Hoge, R. D. (1983). Psychometric properties of teacher-judgment measures of pupil aptitudes, classroom behaviors, and achievement levels. *The Journal of Special Education, 17*, 401–429.

Hoge, R. D. & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76*, 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.

Hoge, R. D. & Cudmore, L. (1986). The use of teacher-judgment measures in the identification of gifted pupils. *Teaching and Teacher Education, 2*, 181–196.

Holz-Ebeling, F. (2010). Arbeitsverhalten und Arbeitsprobleme. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (4. Aufl., S. 29–38). Weinheim: Beltz.

Hox, J. J. & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*, 157–174.

Hunsu, N. J., Adesope, O. & Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education, 94*, 102–119.

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.

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

Jerusalem, M. & Mittag, W. (1999). Selbstwirksamkeit, Bezugsnormen, Leistung und Wohlbefinden in der Schule. In M. Jerusalem & R. Pekrun (Hrsg.), *Emotion, Motivation und Leistung* (S. 221–245). Göttingen: Hogrefe.

Jürgens, E. & Lissmann, U. (2015). *Pädagogische Diagnostik: Grundlagen und Methoden der Leistungsbeurteilung in der Schule*. Weinheim: Beltz.

Jurkowski, S. & Hänze, M. (2010). Soziale Kompetenz, transaktives Interaktionsverhalten und Lernerfolg: Experimenteller Vergleich zweier unterschiedlich gestalteter Gruppenunterrichtsbedingungen und Evaluation eines transaktivitätsbezogenen Kooperationsskriptes. *Zeitschrift für Pädagogische Psychologie, 24*, 241–257.

Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review, 93*, 429–445.

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology, 57*, 469–480.

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.

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

Kaiser, J., Möller, J., Helm, F. & Kunter, M. (2015). Das Schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift für Erziehungswissenschaft*, *18*, 279–302.

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

Karing, C., Dörfler, T. & Artelt, C. (2015). How accurate are teacher and parent judgements of lower secondary school children's test anxiety? *Educational Psychology, 35*, 909–925.

Karing, C., Pfost, M. & Artelt, C. (2011). Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen? *Journal for Educational Research Online, 2*, 119–147.

Keller, G. (1993a). Das Lern- und Arbeitsverhalten leistungsstarker und leistungsschwacher Schüler. *Psychologie in Erziehung und Unterricht, 40*, 125–129.

Keller, G. (1993b). Veränderungen im Lern- und Arbeitsverhalten von Kindern und Jugendlichen. *Pädagogische Welt, 47*, 259–261.

Keller, G. & Thiel, R.-D. (1998). *Lern- und Verhaltensinventar* (LAVI). Hogrefe: Göttingen.

Kenny, D. A. & West, T. V. (2010). Similarity and agreement in self- and other perception: A meta-analysis. *Personality and Social Psychology Review, 14*, 196–213.

Kenny, D. T. & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessments. *Journal of Learning Disabilities, 26*, 227–236.

Klauer, K. C. & Schmeling, A. (1990). Sind Halo-Fehler Flüchtigkeitsfehler? *Zeitschrift für experimentelle und angewandte Psychologie, 37*, 594–607.

Klug, J., Bruder, S. & Schmitz, B. (2016). Which variables predict teachers' diagnostic competence when diagnosing students' learning behavior at different stages of a teacher's career? *Teachers and Teaching: Theory and Practice, 22*, 461–484.

Klug, J., Gerich, M., Bruder, S. & Schmitz, B. (2012). Ein Tagebuch für Hauptschullehrkräfte zur Unterstützung der Reflektionsprozesse beim Diagnostizieren. *Empirische Pädagogik, 26*, 292–311.

Kuklinski, M. R. & Weinstein, R. S. (2001). Classroom and developmental differences in a path model of teacher expectancy effects. *Child Development, 72*, 1554–1578.

Kultusministerkonferenz (2001). *296. Plenarsitzung der Kultusministerkonferenz am 05./06. Dezember 2001 in Bonn*. Zugriff am 30.01.2014 13:48 unter http://www.kmk.-org/index.php?id=1032&type=123

Kultusministerkonferenz (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004*. Zugriff am 30.01.2014 10:40 unter http://www.kmk.org/fileadmin/veroeffentlichungen_-beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf.

Kultusministerkonferenz. (2005). *Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung*. München: Luchterhand.

Laidra, K., Allik, J., Harro, M., Merenäkk, L. & Harro, J. (2006). Agreement among adolescents, parents, and teachers on adolescent personality. *Assessment, 13*, 187–196.

Lan, W. Y. (1996). The effects of self-monitoring on students' course performance, use of learning strategies, attitude, self-judgement ability, and knowledge representation. *The Journal of Experimental Education, 64*, 101–115.

Landrum, R. E. (2015). Teacher-ready research review: Clickers. *Scholarship of Teaching and Learning in Psychology, 1*, 250–254.

Leucht, M., Tiffin-Richards, S., Vock, M., Pant, H. A. & Köller, O. (2012). Diagnostische Kompetenz von Englischlehrkräften bei der Bewertung von Schülerleistungen mit Hilfe des Gemeinsamen Europäischen Referenzrahmens für Sprachen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44*, 163–177.

Lohaus, D. (2009). *Leistungsbeurteilung*. Göttingen: Hogrefe.

Lohbeck, A., Nitkowski, D., Petermann, F. & Petermann, U. (2014). Erfassung von Schülerselbsteinschätzungen zum schulbezogenen Sozial- und Lernverhalten – Validierung der Schülereinschätzliste für Sozial- und Lernverhalten (SSL). *Zeitschrift für Erziehungswissenschaft, 17*, 701–722.

Lohbeck, A., Petermann, F. & Petermann, U. (2015). Selbsteinschätzungen zum Sozial- und Lernverhalten von Grundschulkindern der vierten Jahrgangsstufe. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*, 1–13.

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.

Maas, C. J. M. & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86–92.

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

Malecki, C. K. & Elliot, S. N. (2002). Children's social behaviors as predictors of academic achievement: A longitudinal analysis. *School Psychology Quarterly, 17*, 1–23.

Malti, T., Bayard, S. & Buchmann, M. (2008). Mitgefühl, soziales Verstehen und prosoziales Verhalten: Komponenten sozialer Handlungsfähigkeit in der Kindheit. In T. Malti & S. Perren (Hrsg.), *Soziale Kompetenz bei Kindern und Jugendlichen: Entwicklungsprozesse und Förderungsmöglichkeiten* (S. 52–69). Stuttgart: Kohlhammer.

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A. & Zhang, H. (2009). Clickers in college classrooms:

Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 34*, 51–57.

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

McIntyre, R. M., Smith, D. & Hassett, C. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*, 147–156.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*, 20–32.

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review, 8*, 193–210.

Moosbrugger, H. (2012). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 103–117). Heidelberg: Springer.

Müller, A. & Ditton, H. (2014). Feedback: Begriff, Formen und Funktionen. In A. Müller & H. Ditton (Hrsg.), *Rückmeldungen und Feedback: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S.11–28). Münster: Waxmann.

Mulholland, L. A. & Berliner, D. C. (1992, April). *Teacher experience and the estimation of student achievement.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Muthén, L. K. & Muthén, B. O. (2015). *Mplus 7.31* [computer software]. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K. & Muthén, B. O. (2016). *Mplus 7.4* [computer software]. Los Angeles, CA: Muthén & Muthén.

Narciss, S. (2014). Modelle zu den Bedingungen und Wirkungen von Feedback in Lehr-Lernsituationen. In A. Müller & H. Ditton (Hrsg.), *Rückmeldungen und Feedback: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (S.43–82). Münster: Waxmann.

Neber, H. (2004). Lehrernominierungen für ein Enrichment-Programm als Beispiel für die Talentsuche in der gymnasialen Oberstufe. *Psychologie in Erziehung und Unterricht, 51*, 24–39.

Nohe, C., Peters, A. & Sonntag, K. (2014). Tagebuch. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (17. Aufl., S. 1636). Bern: Huber.

Oerke, B., McElvany, N., Ohle, A., Ullrich, M. & Horz, H. (2015). Verbessert sich die diagnostische Urteilsgenauigkeit von Lehrkräften bei längerem Kontakt mit der Klasse? *Psychologie in Erziehung und Unterricht, 63*, 34–47.

Ohle, A. & McElvany, N. (2015). Teachers' diagnostic competences and their practical relevance. *Journal for Educational Research Online, 7*, 5–10.

Peterson, E. R., Rubie-Davies, C., Osborne, D. & Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: Relations with student achievement and the ethnic achievement gap. *Learning and Instruction, 42*, 123–140.

Pit-ten Cate, I., Krolak-Schwerdt, S. & Glock, S. (2016). Accuracy of teachers' tracking decisions: short- and long-term effects of accountability. *European Journal of Psychology of Education, 31*, 225–243.

Pit-ten Cate, I., Krolak-Schwerdt, S., Glock, S. & Markova, M. (2014). Improving teachers' judgments: Obtaining change through cognitive processes. In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 45–61). Rotterdam: Sense.

Praetorius, A.-K., Berner, V.-D., Zeinz, H., Scheunpflug, A. & Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *The Journal of Educational Research, 106*, 64–76.

Praetorius, A.-K., Karst, K., Dickhäuser, O. & Lipowsky, F. (2011). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften. *Psychologie in Erziehung und Unterricht, 58*, 81–91.

Praetorius, A.-K., Lipowsky, F. & Karst, K. (2012). Diagnostische Kompetenz von Lehrkräften: Aktueller Forschungsstand, unterrichtspraktische Umsetzbarkeit und Bedeutung für den Unterricht. In R. Lazarides & A. Ittel (Hrsg.), *Differenzierung im*

mathematisch-naturwissenschaftlichen Unterricht (S. 115–146). Bad Heilbrunn: Klinkhardt.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*, 581–588.

Raudenbush, S. W., Bryk, A. S. & Congdon, R. (2013). *HLM 7.01 for Windows* [computer software]. Skokie, IL: Scientific Software International, Inc.

Rauer, W. & Schuck, K. D. (2003*). FEESS 3-4 – Fragebogen zur Erfassung emotionaler und sozialer Schulerfahrungen von Grundschulkindern dritter und vierter Klassen*. Göttingen: Beltz.

Reindl, S. & Hascher, T. (2013). Emotionen im Mathematikunterricht in der Grundschule. *Unterrichtswissenschaft, 41*, 268–288.

Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation*. Göttingen: Hogrefe.

Rheinberg, F. (2005). Trainings auf der Basis eines kognitiven Motivationsmodells. In F. Rheinberg & S. Krug (Hrsg.), *Motivationsförderung im Schulalltag* (S. 36–52). Göttingen: Hogrefe.

Rheinberg, F. (2014). Bezugsnorm. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (17. Aufl., S. 296). Bern: Huber.

Rheinberg, F. & Fries, S. (2010). Bezugsnormorientierung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 61–68). Weinheim: Beltz.

Rindermann, H. (2014). Emotionale Kompetenz. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (17. Aufl., S. 438). Bern: Huber.

Roch, S. G., Woehr, D. J., Mishra, V. & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*, 370–395.

Rogalla, M. & Vogt, F. (2008). Förderung adaptiver Lehrkompetenz: eine Interventionsstudie. *Unterrichtswissenschaft, 36*, 17–36.

Roick, T., Gölitz, D. & Hasselhorn, M. (2004*). DEMAT3+ – Deutscher Mathematiktest für dritte Klassen*. Göttingen: Beltz.

Rosenthal, R. (1991). Teacher expectancy effects: A brief update 25 years after the Pygmalion experiment. *Journal of Research in Education, 1*, 3–12.

Rosenthal, R. & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports, 19*, 115–118.

Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York, NY: Holt, Rinehart & Winston.

Rosenthal, R. & Jacobson, L. (1971). *Pygmalion im Unterricht*. Weinheim: Beltz.

Rosenthal, R. & Rubin, D. B. (1978). Issues in summarizing the 1st 345 studies of interpersonal expectancy effects. *Behavioral and Brain Sciences, 1*, 410–415.

Rost, D. H. & Hanses, P. (1997). Wer nichts leistet, ist nicht begabt? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 29*, 167–177.

Rubie-Davies, C. M. (2008). Teacher expectations. In T. Good (Ed.), *21st century education: A reference handbook* (pp. 254 – 262). Thousand Oaks, CA: Sage.

Rubie-Davies, C. M. (2010). Teacher expectations and perceptions of student attributes: Is there a relationship? *British Journal of Educational Psychology, 80*, 121–135.

Rubie-Davies, C. M., Weinstein, R. S., Huang, F. L., Gregory, A., Cowan, P. A. & Cowan, C. P. (2014). Successive teacher expectation effects across the early school years. *Journal of Applied Developmental Psychology, 35*, 181–191.

Schmitz, B. & Wiese, B. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology, 31*, 64–96.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

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

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.), *Leistungsmessungen in Schulen* (S. 45–58). Weinheim: Beltz.

Schunk, D. H., Pintrich, P. R. & Meece, J. L. (2008). *Motivation in education. Theory, research, and applications*. Upper Saddle River, NJ: Pearson Education.

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

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

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

Stang, J. & Urhahne, D. (2016a). Wie gut schätzen Lehrkräfte Leistung, Konzentration, Arbeits- und Sozialverhalten ihrer Schülerinnen und Schüler ein? Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. *Psychologie in Erziehung und Unterricht, 63*, 204–219.

Stang, J. & Urhahne, D. (2016b). Stabilität, Bezugsnormorientierung und Auswirkungen der Urteilsgenauigkeit. *Zeitschrift für Pädagogische Psychologie, 30*, 251–262.

Stark, R. & Mandl, H. (2007). Bridging the gap between basic and applied research by an integrative research approach. *Educational Research and Evaluation, 13*, 249–261.

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

Südkamp, A., Kaiser, J. & Möller, J. (2014). Teachers' judgments of students' academic achievement. Results from field and experimental studies. In S. Krolak-Schwerdt,

S. Glock & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 5–25). Rotterdam: Sense.

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

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

Ter Laak, J. F., DeGoede, M. & Brugman, G. (2001). Teacher's judgements of pupils: Agreement and accuracy. *Social Behavior and Personality, 29*, 257–270.

Thorndike, E. L. (1920). A constant error on psychological rating. *Journal of Applied Psychology, 4*, 25–29.

Trautwein, U. & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind. Referenzgruppeneffekte bei Übertrittsentscheidungen. *Zeitschrift für Pädagogische Psychologie, 21*, 119–133.

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

Urban, D. & Mayerl, J. (2006). *Regressionsanalyse: Theorie, Technik und Anwendung*. Wiesbaden: VS.

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

Urhahne, D., Chao, S.-H., Florineth, M. L., Luttenberger, S. & Paechter, M. (2011). Academic self-concept, learning motivation, and test anxiety of the underestimated student. *British Journal of Educational Psychology, 81*, 161–177.

Urhahne, D., Timm, O., Zhu, M. & Tang, M. (2013). Sind unterschätzte Schüler weniger leistungsmotiviert als überschätzte Schüler? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 45*, 34–43.

Urhahne, D., Zhou, J., Stobbe, M., Chao, S.-H., Zhu, M. & Shi, J. (2010). Motivationale und affektive Merkmale unterschätzter Schüler. Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 24*, 275–288.

Urhahne, D. & Zhu, M. (2015a). Teacher judgement and student motivation. In C. M. Rubie-Davies, J. M. Stephens & P. Watson (Eds.), *International Handbook of Social Psychology of the Classroom* (pp. 304–315). London, UK: Routledge.

Urhahne, D. & Zhu, M. (2015b). Accuracy of teachers' judgments of students' subjective well-being. *Learning and Individual Differences, 43*, 226–232.

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Weinert, F. E. (2014). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

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. Festschrift zum 60. Geburtstag von Karlheinz Ingenkamp* (S. 11–29). Weinheim: Beltz.

Welsh, M., Parke, R. D., Widaman, K. & O'Neil, R. (2001). Linkages between children's social and academic competence: A longitudinal analysis. *Journal of School Psychology, 39*, 463–481.

Wirtz, M. A. (2014). Logischer Fehler. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (17. Aufl., S. 975). Bern: Hans Huber.

Woehr, D. J. & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189–205.

Woolfolk, A. (2014). *Pädagogische Psychologie*. Hallbergmoos: Pearson.

Zhu, M. & Urhahne, D. (2015). Teachers' judgement of students' foreign-language achievement. *European Journal of Psychology of Education, 30*, 21–39.

Ziegler, A., Dresel, M., Schober, B. & Stöger, H. (2005). *Ulm Motivational Test Battery (UMTB): Documentation of items and scales (Ulm Educational Research Report, No. 15)*. Ulm: Ulm University, Department of Educational Psychology.

7. Abbildungsverzeichnis

8. Tabellenverzeichnis

9. Anhang

Anhang A: Materialien der ersten und zweiten Studie

Anhang B: Materialien der dritten Studie

Anhang C: Materialien für weiterführende Studien

Anhang A: Materialien der ersten und zweiten Studie

Die Materialien befinden sich auf den folgenden Seiten.

Fragebogen A-1: Soziodemografische Fragen für Lehrkräfte

Unterrichtete Klasse: _____

Alter: _____ Jahre

Geschlecht: ☐ männlich ☐ weiblich

Welche Fächer unterrichten Sie in dieser Klasse?

_____

Wie viele Stunden pro Woche unterrichten Sie in der Klasse?

_____ Stunden

Wie lange kennen Sie diese Klasse?

_____ Monate

Wie viele Jahre Berufserfahrung haben Sie?

_____ Jahre

Fragebogen A-2: Lehrkrafteinschätzungen für jeden Schüler

Schülernummer (Bitte nummerieren Sie in einer Klassenliste die Schülerinnen und Schüler fortlaufend durch, so dass Ihre Einschätzungen den Schülerfragebögen zugeordnet werden können)

Schülernummer _____

1. Wie viele der 35 Aufgaben des Mathematiktests löst der Schüler richtig?

_____ Aufgaben

2. Bitte schätzen Sie folgende Merkmale des Schülers im Vergleich zu anderen Schülern im selben Alter ein:

Konzentration (wie konzentriert bearbeitet der Schüler Mathematikaufgaben)

sehr viel schlechter	deutlich schlechter	etwas schlechter	gleich	etwas besser	deutlich besser	sehr viel besser
❍	❍	❍	❍	❍	❍	❍

Arbeitsverhalten (wie gewissenhaft und pflichtbewusst verhält sich der Schüler)

sehr viel weniger	deutlich weniger	etwas weniger	gleich	etwas stärker	deutlich stärker	sehr viel stärker
❍	❍	❍	❍	❍	❍	❍

Empathie (wie gut kann der Schüler die Gefühle anderer verstehen)

sehr viel schlechter	deutlich schlechter	etwas schlechter	gleich	etwas besser	deutlich besser	sehr viel besser
❍	❍	❍	❍	❍	❍	❍

Prosoziales Verhalten (wie freundlich, freigiebig und hilfsbereit verhält sich der Schüler)

sehr viel schwächer	deutlich schwächer	etwas schwächer	gleich	etwas stärker	deutlich stärker	sehr viel stärker
❍	❍	❍	❍	❍	❍	❍

Anhang B: Materialien der dritten Studie

Die Materialien befinden sich auf den folgenden Seiten.

Fragebogen B-1: Soziodemografische Fragen für Lehrkräfte

Unterrichtete Klasse: _____

Alter: _____ Jahre

Geschlecht: ☐ männlich ☐ weiblich

Welche Fächer unterrichten Sie in dieser Klasse?

_____

Wie viele Stunden pro Woche unterrichten Sie in der Klasse?

_____ Stunden

Wie viele Jahre an Berufserfahrung haben Sie?

_____ Jahre

Fragebogen B-2: Lehrkrafteinschätzungen für jeden Schüler

Schülernummer: _____

1. Wie viele der 31/40 Aufgaben des Mathematiktests löst der Schüler richtig?

_____ Aufgaben

2. Bitte schätzen Sie ein, wie der Schüler folgende Fragen beantwortet:

a. Was denkst du, welche Note wirst du in der nächsten Matheprobe erhalten? _____ (1 bis 6)

b. Mit welcher Note in der nächsten Matheprobe wärst du gerade noch zufrieden? _____(1 bis 6)

3. Bitte schätzen Sie folgende Merkmale des Schülers im Vergleich zu anderen Schülern im selben Alter ein:

Fähigkeitsselbstkonzept (wie schätzt der Schüler seine Fähigkeiten in **Mathematik** ein)

sehr viel geringer	deutlich geringer	geringer	etwas geringer	gleich	etwas größer	größer	deutlich größer	sehr viel größer
❍	❍	❍	❍	❍	❍	❍	❍	❍

Leistungsangst (wie viel Angst hat der Schüler vor **Mathematik**)

sehr viel weniger	deutlich weniger	weniger	etwas weniger	gleich	etwas mehr	mehr	deutlich mehr	sehr viel mehr
❍	❍	❍	❍	❍	❍	❍	❍	❍

Lernfreude (wie positiv erlebt der Schüler **im Allgemeinen** schulische Aufgaben)

sehr viel schwächer	deutlich schwächer	schwächer	etwas schwächer	gleich	etwas stärker	stärker	deutlich stärker	sehr viel stärker
❍	❍	❍	❍	❍	❍	❍	❍	❍

Schuleinstellung (wie wohl fühlt sich der Schüler **im Allgemeinen** in der Schule)

sehr viel schlechter	deutlich schlechter	schlechter	etwas schlechter	gleich	etwas besser	besser	deutlich besser	sehr viel besser
❍	❍	❍	❍	❍	❍	❍	❍	❍

Anstrengungsbereitschaft (wie gut bewältigt der Schüler **im Allgemeinen** Anforderungen durch eigenes Bemühen)

sehr viel schlechter	deutlich schlechter	schlechter	etwas schlechter	gleich	etwas besser	besser	deutlich besser	sehr viel besser
❍	❍	❍	❍	❍	❍	❍	❍	❍

Anhang C: Materialien für weiterführende Studien

Die Materialien befinden sich auf den folgenden Seiten.

Fragebogen C-1: Fragebogen zur Selbsteinschätzung der eigenen diagnostischen Urteilsfähigkeiten

		stimme über-haupt nicht zu	stimme nicht zu	stimme eher nicht zu	stimme eher zu	stimme zu	stimme voll und ganz zu
1.	Ich kann Schüler in Bezug auf Leistung in eine Rangreihe bringen.	❒	❒	❒	❒	❒	❒
2.	Ich weiß, wer ein guter Schüler ist.	❒	❒	❒	❒	❒	❒
3.	Ich kann die Schwierigkeit von Mathematikaufgaben gut einschätzen.	❒	❒	❒	❒	❒	❒
4.	Ich weiß, wer gut rechnen kann.	❒	❒	❒	❒	❒	❒
5.	Ich weiß, was meine Schüler in Arithmetik können.	❒	❒	❒	❒	❒	❒
6.	Ich überschätze die Leistung der Schüler.	❒	❒	❒	❒	❒	❒
7.	Ich weiß, wie viele Aufgaben ein Schüler in Arithmetik lösen kann.	❒	❒	❒	❒	❒	❒
8.	Meine Beurteilungen sind streng.	❒	❒	❒	❒	❒	❒
9.	Ich schätze die Leistung der Schüler genau ein.	❒	❒	❒	❒	❒	❒
10.	Ich weiß, was meine Schüler in Geometrie können.	❒	❒	❒	❒	❒	❒
11.	Ich kann die Schwierigkeit von Sachrechenaufgaben gut einschätzen.	❒	❒	❒	❒	❒	❒
12.	Ich weiß, wie viele Aufgaben ein Schüler in Geometrie lösen kann.	❒	❒	❒	❒	❒	❒
13.	Meine Beurteilungen sind mild.	❒	❒	❒	❒	❒	❒
14.	Ich schätze Leistungsunterschiede zwischen Schülern genau ein.	❒	❒	❒	❒	❒	❒
15.	Ich weiß, wer schlecht rechnen kann.	❒	❒	❒	❒	❒	❒
16.	Ich weiß, was meine Schüler in Sachrechnen können.	❒	❒	❒	❒	❒	❒
17.	Ich unterschätze die Leistung der Schüler.	❒	❒	❒	❒	❒	❒
18.	Ich kann die Schwierigkeit von Arithmetikaufgaben gut einschätzen.	❒	❒	❒	❒	❒	❒
19.	Ich weiß, wie viele Aufgaben ein Schüler in Sachrechnen lösen kann.	❒	❒	❒	❒	❒	❒
20.	Ich überschätze Leistungsunterschiede zwischen Schülern.	❒	❒	❒	❒	❒	❒
21.	Ich weiß, wer ein schlechter Schüler ist.	❒	❒	❒	❒	❒	❒
22.	Ich kann die Schwierigkeit von Geometrieaufgaben gut einschätzen.	❒	❒	❒	❒	❒	❒
23.	Ich weiß, wie viele Mathematikaufgaben ein Schüler im Test lösen kann.	❒	❒	❒	❒	❒	❒
24.	Ich unterschätze Leistungsunterschiede zwischen Schülern.	❒	❒	❒	❒	❒	❒
25.	Meine Beurteilungen sind gerecht.	❒	❒	❒	❒	❒	❒
26.	Ich weiß, was meine Schüler in Mathematik können.	❒	❒	❒	❒	❒	❒

Fragebogen C-2: Einschätzung der Aufgabenschwierigkeit

Liebe Lehrkräfte,
bitte schätzen Sie für jede der 35 Aufgaben des Mathematiktests ein, wie leicht oder schwierig diese für Ihre Schülerinnen und Schüler zu lösen ist. Kreuzen Sie die Antwort an, die am besten zutrifft und kreuzen Sie pro Aussage nur ein Kästchen an. Bitte nehmen Sie dazu den Mathematiktest zur Hand.

Aufgaben	sehr leicht	leicht	eher leicht	mittel	eher schwie-rig	schwie-rig	sehr schwie-rig
Teil A							
1 Zahlenstrahl	❒	❒	❒	❒	❒	❒	❒
2 Bestimmung eines Anteils	❒	❒	❒	❒	❒	❒	❒
3a Umwandlung von Brüchen	❒	❒	❒	❒	❒	❒	❒
3b Umwandlung von Brüchen	❒	❒	❒	❒	❒	❒	❒
4a Umwandlung von Brüchen	❒	❒	❒	❒	❒	❒	❒
4b Umwandlung von Brüchen	❒	❒	❒	❒	❒	❒	❒
5a Kürzen von Brüchen	❒	❒	❒	❒	❒	❒	❒
5b Kürzen von Brüchen	❒	❒	❒	❒	❒	❒	❒
6a Umwandlung von Maßeinheiten	❒	❒	❒	❒	❒	❒	❒
6b Umwandlung von Maßeinheiten	❒	❒	❒	❒	❒	❒	❒
7a Grundrechenarten mit Brüchen	❒	❒	❒	❒	❒	❒	❒
7b Grundrechenarten mit Brüchen	❒	❒	❒	❒	❒	❒	❒
8a Bruchrechnen mit unbekannter Variable	❒	❒	❒	❒	❒	❒	❒
8b Bruchrechnen mit unbekannter Variable	❒	❒	❒	❒	❒	❒	❒
9 Aufstellen einer Gleichung	❒	❒	❒	❒	❒	❒	❒
10 Lösung einer Gleichung	❒	❒	❒	❒	❒	❒	❒
Teil B							
1 Symmetrie	❒	❒	❒	❒	❒	❒	❒
2 Flächenberechnung	❒	❒	❒	❒	❒	❒	❒
3 Winkelbestimmung	❒	❒	❒	❒	❒	❒	❒
4 Volumenberechnung	❒	❒	❒	❒	❒	❒	❒
Teil C							
1 Proportionalität Dezimalzahl	❒	❒	❒	❒	❒	❒	❒
2 Aufstellen von Termen Subtraktion	❒	❒	❒	❒	❒	❒	❒
3a Proportionalität	❒	❒	❒	❒	❒	❒	❒
3b Proportionalität	❒	❒	❒	❒	❒	❒	❒
4 Proportionalität Bruchzahl	❒	❒	❒	❒	❒	❒	❒
5 Aufstellen von Termen Division	❒	❒	❒	❒	❒	❒	❒
6 Proportionalität Bruchzahl	❒	❒	❒	❒	❒	❒	❒
7a Datenbearbeitung Tabelle	❒	❒	❒	❒	❒	❒	❒
7b Datenbearbeitung Tabelle	❒	❒	❒	❒	❒	❒	❒
8 Flächenanteil	❒	❒	❒	❒	❒	❒	❒
9a Datenbearbeitung Diagramm	❒	❒	❒	❒	❒	❒	❒
9b1 Datenbearbeitung Diagramm	❒	❒	❒	❒	❒	❒	❒
9b2 Datenbearbeitung Diagramm	❒	❒	❒	❒	❒	❒	❒
9b3 Datenbearbeitung Diagramm	❒	❒	❒	❒	❒	❒	❒
9b4 Datenbearbeitung Diagramm	❒	❒	❒	❒	❒	❒	❒

Die einzuschätzenden Aufgaben stammen aus dem Deutschen Mathematiktest für sechste Klassen (Götz, L., Lingel, K. & Schneider, W. (2013). *DEMAT 6+. Deutscher Mathematiktest für sechste Klassen*. Göttingen: Hogrefe.).

Eidesstattliche Erklärung

Hiermit erkläre ich, dass die vorliegende Dissertationsschrift von mir selbständig angefertigt wurde und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet wurden. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen sind, habe ich als solche kenntlich gemacht. Die Arbeit wurde nicht in derselben oder einer ähnlichen Fassung an einer anderen Universität zur Erlangung eines akademischen Grades eingereicht.

_____ _____

Ort, Datum Unterschrift