

# Publikations-Freigabe

Hiermit bestätige ich, dass der nachfolgende Artikel für die Zeitschrift: \_\_\_\_\_

Beitrag: \_\_\_\_\_

von mir auch im Namen aller Mitautoren zur Veröffentlichung genehmigt wird, nachdem die angezeigten Korrekturen ausgeführt worden sind.

Datum

Unterschrift

Name (Bitte in Großbuchstaben)

Bitte überprüfen Sie den beigefügten Korrekturabzug und korrigieren Sie nur die Fehler, die unbedingt vor dem Druck behoben werden müssen. Vom Manuskript abweichende Ergänzungen oder Änderungen sind nach Möglichkeit zu vermeiden. Sollten erhebliche Korrekturen anfallen, so können diese dem Autoren in Rechnung gestellt werden. Bitte senden Sie das unterschriebene Formular zusammen mit Ihrem Korrekturabzug an die unten angegebene Adresse zurück. Informationen über Hans Huber OpenMind, unser neues Open Access-Publikationsmodell, finden Sie unter [www.verlag-hanshuber.com/openmind](http://www.verlag-hanshuber.com/openmind).

## Urheberrechte: Richtlinien für Autoren

Mit Unterzeichnung dieses Formulars bestätigen und garantieren die Autoren, dass sie uneingeschränkt über sämtliche Urheberrechte an ihrem Beitrag einschließlich eventueller Bildvorlagen, Zeichnungen, Pläne, Karten, Skizzen und Tabellen verfügen und dass der Beitrag keine Rechte Dritter verletzt.

Die Autoren räumen dem Verlag räumlich und mengenmäßig unbeschränkt für die Dauer des gesetzlichen Urheberrechts das ausschließliche Recht der Vervielfältigung und Verbreitung bzw. der unkörperlichen Wiedergabe des Beitrages ein. Dies gilt auch für die Verwertung des Beitrages außerhalb der Zeitschrift, für die der Artikel eingereicht wurde, und unabhängig von deren Veröffentlichung. Die Autoren räumen dem Verlag ferner die folgenden ausschließlichen Nutzungsrechte ein:

1. Das Recht zum ganzen und teilweisen Vorabdruck oder Nachdruck – auch in Form eines Sonderdrucks, zur Übersetzung in andere Sprachen, zu sonstiger Bearbeitung und zur Erstellung von Zusammenfassungen (Abstracts);
2. das Recht zur Veröffentlichung einer Mikrokopie-, Mikrofilm- und Mikroformausgabe, zur Nutzung im Weg von Bildschirmtext, Videotext und ähnlichen Verfahren, zur Aufzeichnung auf Bild- und/oder Tonträger und zu deren öffentlicher Wiedergabe – auch multimedial – sowie zur öffentlichen Wiedergabe durch Radio- und Fernsehsendungen;
3. das Recht zur maschinenlesbaren Erfassung und elektronischen Speicherung auf einem Datenträger (z.B. Diskette, CD-Rom, Magnetband) und in einer eigenen oder fremden Online-Datenbank, zum Download in einem eigenen oder fremden Rechner, zur Wiedergabe am Bildschirm – sei es unmittelbar oder im Wege der Datenfernübertragung –, sowie zur Bereithaltung in einer eigenen oder fremden Online-Datenbank zur Nutzung durch Dritte;
4. das Recht zu sonstiger Vervielfältigung, insbesondere durch fotomechanische und ähnliche Verfahren (z.B. Fotokopie, Fernkopie) und zur Nutzung im Rahmen eines sogenannten Kopienversands auf Bestellung; das Recht zur Vergabe der vorgenannten Nutzungsrechte an Dritte in In- und Ausland sowie die von der Verwertungsgesellschaft WORT wahrgenommenen Rechte einschließlich der entsprechenden Vergütungsansprüche;
5. Online-Rechte für Zeitschriftenbeiträge  
Autoren von Beiträgen in Zeitschriften der Hogrefe Verlagsgruppe sind berechtigt:
  - zur Archivierung eines sogenannten Pre-Print (Manuskriptfassung vor dem Peer Review-Verfahren) jederzeit auf ihre eigene Homepage, auf die Homepage ihres Arbeitgebers oder in ihr institutionelles Repositorium, ausschließlich für nicht-kommerzielle Zwecke
  - zur Archivierung eines sogenannten Post-Print (die letzte Manuskript-Fassung nach dem Peer Review-Verfahren) auf ihre eigene Homepage, auf die Homepage ihres Arbeitgebers oder in ihr institutionelles Repositorium 12 Monate nach der Veröffentlichung im entsprechenden Zeitschriftenheft ausschließlich für nicht-kommerzielle Zwecke

- zur Archivierung auf Anfrage eines sogenannten Post-Print (die letzte Manuskript-Fassung nach dem Peer Review-Verfahren) in ein Repositorium, das von ihrer Forschungseinrichtung bzw. Geldgeber verpflichtend vorgegeben wird, (1) 12 Monate nach Veröffentlichung im entsprechenden Zeitschriftenheft oder (2) um gesetzlichen Verpflichtungen nachzukommen

Voraussetzungen hierfür sind:

- Pre-Print: Die Manuskriptfassung sollte mit dem Erstellungsdatum versehen werden, sowie mit einem Hinweis, dass das Manuskript in dieser Form noch nicht für eine Veröffentlichung angenommen wurde.
- Post-Print: Es darf ausschließlich die letzte Manuskript-Fassung des Artikels (Post-Refereeing) verwendet werden, nicht die Verlagsfassung in der veröffentlichten Form (Verlags-PDF, -XML). Die Post-Print-Fassung darf zeitlich erst 12 Monate nach dem Erscheinen des Artikels im entsprechenden Zeitschriftenheft zur Verfügung gestellt werden. Sie muss auf die DOI der Verlagsfassung verlinkt werden und mit dem Copyright-Vermerk des Verlages (Titel der Zeitschrift, Band-/Heft-Nr., © [Jahr] by [Verlag]) versehen werden. Darüber hinaus ist der folgende Vermerk anzubringen: „Diese Artikelfassung entspricht nicht vollständig dem in der Zeitschrift veröffentlichten Artikel. Dies ist nicht die Originalversion des Artikels und kann daher nicht zur Zitierung herangezogen werden“.
- Der Verlag stellt keine elektronischen Daten der veröffentlichten Verlagsversion zur Verfügung. Die Anfertigung einer elektronischen Kopie der veröffentlichten Verlagsversion zum Zwecke der Verbreitung ist nicht erlaubt.
- Der Verlag Hans Huber bietet ein optionales Open Access-Publikationsmodell für Autoren bzw. deren Institutionen/Geldgeber an, falls eine vorgezogene Open Access-Veröffentlichung vorgeschrieben oder gewünscht wird. Nähere Informationen unter: <http://www.verlag-hanshuber.com/openmind>.

(31. Januar 2010)

**Verlag Hans Huber**  
Hogrefe AG  
c/o Satzspiegel  
Postfach 1103  
DE-37171 Nörten-Hardenberg  
Tel. +49 5503 997670  
Fax +49 5503 997679  
HHZeitschriften@aol.com  
[www.verlag-hanshuber.com](http://www.verlag-hanshuber.com)

HUBER



# Publikations-Freigabe

(Open Access-Publikation)

Hiermit bestätige ich, dass für die Zeitschrift \_\_\_\_\_  
der nachfolgende Artikel \_\_\_\_\_

von mir auch im Namen aller Mitautoren zur Veröffentlichung genehmigt wird, nachdem die angezeigten Korrekturen ausgeführt worden sind.

Ich stimme den Bedingungen und Konditionen zu einer Hogrefe Open Access-Publikation zu und zahle die Artikelgebühr für diesen Open Access-Artikel. Bitte belasten Sie meine Kreditkarte mit dem Betrag und stellen Sie mir eine entsprechende Rechnung aus:

Gebühr: € 2,500\*, CHF 3,750\* (Europe) or US \$3,000 (RdW)  
\*zzgl. MwSt. für Kunden in Deutschland und der Schweiz

Rechnungsanschrift

(sofern sie von der Korrespondenzanschrift des Artikels abweicht)

Kreditkarten (bitte unterstreichen): Visa, Mastercard, AmE:

Kreditkarten-Nr. \_\_\_\_\_

Sicherheits-Code: \_\_\_\_\_

Gültig bis: \_\_\_\_\_

Kreditkarteninhaber: \_\_\_\_\_

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

Datum

Unterschrift

Name (Bitte in Großbuchstaben)

Bitte überprüfen Sie den beigegefügtten Korrekturabzug und korrigieren Sie nur die Fehler, die unbedingt vor dem Druck behoben werden müssen. Vom Manuskript abweichende Ergänzungen oder Änderungen sind nach Möglichkeit zu vermeiden. Sollten erhebliche Korrekturen anfallen, so können diese dem Autoren in Rechnung gestellt werden. Bitte senden Sie das unterschriebene Formular zusammen mit Ihrem Korrekturabzug an die unten angegebene Adresse zurück.

## „Hogrefe OpenMind“ Urheber- und Lizenzrechte: Richtlinien für Autoren

Mit Unterzeichnung dieses Formulars bestätigen und garantieren die Autoren, dass sie uneingeschränkt über sämtliche Urheberrechte an ihrem Beitrag einschließlich eventueller Bildvorlagen, Zeichnungen, Pläne, Karten, Skizzen und Tabellen verfügen und dass der Beitrag keine Rechte Dritter verletzt.

- Mit der Imprimatur des Artikels räumt der Autor auch im Namen aller Mitautoren dem Verlag die uneingeschränkten Verwertungsrechte an dem Artikel ein, wie z. B. Verkauf von Nachdrucken an kommerzielle Kunden oder die Lizenzierung als Teil einer Sammlung elektronischer Zeitschriftenartikel. Im Copyright-Vermerk des Artikels wird der Name des Verlags angegeben, solange dies vom Gesetzgeber nicht anders vorgesehen wird.
- Der Autor stimmt auch im Namen aller Mitautoren der Hogrefe OpenMind-Lizenz zu <http://dx.doi.org/10.1026/a000002>, die identisch ist mit der Creative Commons Attribution-Noncommercial 3.0 Unported License. Hier eine Zusammenfassung der Creative Commons-Lizenz:

Es ist Ihnen gestattet:

- Das Werk zu vervielfältigen, zu verbreiten und öffentlich zugänglich zu machen
- Abwandlungen bzw. Bearbeitungen des Inhaltes anzufertigen

Zu den folgenden Bedingungen:

- **Namensnennung.** Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.

- **Keine kommerzielle Nutzung.** Dieses Werk darf nicht für kommerzielle Zwecke verwendet werden.
- Im Falle einer Verbreitung müssen Sie anderen die Lizenzbedingungen, unter die dieses Werk fällt, mitteilen. Am Einfachsten ist es, einen Link auf diese Seite einzubinden.

Jede der vorgenannten Bedingungen kann aufgehoben werden, sofern Sie die Einwilligung des Rechteinhabers dazu erhalten. Diese Lizenz lässt die Urheberpersönlichkeitsrechte unberührt. Die Vollversion der Lizenzvereinbarung ist verfügbar unter: <http://dx.doi.org/10.1026/a000002>

Verlag Hans Huber

Hogrefe AG

c/o Satzspiegel

Postfach 1103

DE-37171 Nörten-Hardenberg

Tel. +49 5503 997670

Fax +49 5503 997679

HHZeitschriften@aol.com

[www.verlag-hanshuber.com](http://www.verlag-hanshuber.com)

HUBER



# Zeitschrift für Pädagogische Psychologie

Zusätzliche Sonderdrucke für Autorinnen und Autoren

## Preisliste

Als Autor(in) erhalten Sie 25 kostenlose Sonderdrucke. Wenn Sie noch zusätzliche Sonderdrucke benötigen, so gelten dafür folgende Preise:

Format	20 Ex.	50 Ex.	100 Ex.	150 Ex.	200 Ex.
21 x 27.7 cm					
Bis 4 Seiten	€ 12	€ 30	€ 48	€ 72	€ 96
Bis 8 Seiten	€ 24	€ 60	€ 96	€ 144	€ 192
Bis 12 Seiten	€ 36	€ 90	€ 144	€ 216	€ 288
Bis 16 Seiten	€ 48	€ 120	€ 192	€ 288	€ 384
Bis 24 Seiten	€ 72	€ 180	€ 288	€ 432	€ 576

+ Versandkostenanteil  
+ 7.6 % MWST (Schweiz)

### Bestellung:

Von meinem Artikel: .....

in der *Zeitschrift für Pädagogische Psychologie*, Ausgabe .....

bestelle ich: ..... zusätzliche Sonderdrucke

Liefer-/  
Rechnungsadresse .....

Datum: .....

Unterschrift: .....

Bitte senden Sie die Bestellung an:

VERLAG HANS HUBER  
Zeitschriftenherstellung  
c/o Joseph A. Smith  
Bachstrasse 16  
D-37176 Nörten-Hardenberg

Originalartikel

# Motivationale und affektive Merkmale unterschätzter Schüler

## Ein Beitrag zur diagnostischen Kompetenz von Lehrkräften

Detlef Urhahne<sup>1</sup>, Ji Zhou<sup>1</sup>, Martin Stobbe<sup>1</sup>, Sheng-Han Chao<sup>1</sup>,  
Mingjing Zhu<sup>2</sup> und Jiannong Shi<sup>2</sup>

<sup>1</sup>Ludwig-Maximilians-Universität München, <sup>2</sup>Chinesische Akademie der Wissenschaften, Peking

**Zusammenfassung.** In diesem Beitrag werden zwei Untersuchungen zur diagnostischen Kompetenz von Lehrkräften und den motivational-affektiven Merkmalen unterschätzter und überschätzter Schüler präsentiert. An den Untersuchungen in Deutschland und China nahmen jeweils acht Grundschulklassen sowie deren Mathematiklehrer teil. Die Viertklässler bearbeiteten einen Mathematikleistungstest und einen Selbstbeschreibungsfragebogen. Die Lehrkräfte schätzten die Schüler in Bezug auf die Testleistung und motivational-affektive Merkmale ein. Auf Grundlage der geschätzten und der tatsächlichen Testleistung erfolgte eine Einteilung in unterschätzte und überschätzte Schüler. Beide Untersuchungen kommen zu übereinstimmenden Ergebnissen. Die Lehrkräfte konnten Testleistungen mit guter Genauigkeit prognostizieren, aber hatten Schwierigkeiten damit motivational-affektive Schülermerkmale richtig einzuschätzen. Unterschätzte Schüler zeigten die gleiche Testleistung und Lernmotivation wie überschätzte Schüler, aber hatten eine niedrigere Erfolgserwartung, ein niedrigeres Fähigkeitsselfkonzept und eine höhere Leistungsangst. Die Lehrer erwarteten, dass unterschätzte Schüler in der nächsten Mathematikarbeit schlechter abschneiden werden und mit schlechteren Noten zufrieden sind als überschätzte Schüler. Der Beurteilungsfehler der Lehrer beschränkte sich nicht allein auf die Testleistung, sondern generalisierte auf motivational-affektive Schülermerkmale.

**Schlüsselwörter:** Genauigkeit des Lehrurteils, diagnostische Kompetenz, Motivation, Fähigkeitsselfkonzept, Leistungsangst

**Motivational and Affective Characteristics of Underestimated Students: A Contribution to the Diagnostic Competence of Teachers**

**Abstract.** In this contribution, two investigations on teachers' diagnostic competence and the motivational-affective characteristics of underestimated and overestimated students are presented. Each time, eight elementary school classes and their mathematics teachers participated in investigations in Germany and China. The fourth graders worked on a mathematics achievement test and a self-description questionnaire. Teachers estimated students' test performance as well as motivational-affective characteristics. The distinction between underestimated and overestimated students was made on basis of the estimated and the real test performance. Both investigations have concurrent results. Teachers could predict test achievement with good accuracy, but had difficulties to assess students' motivational-affective characteristics. Underestimated students showed the same test performance and learning motivation as overestimated students, but had lower expectancy for success, lower self-concept of ability and higher test anxiety. Teachers expected that underestimated students would get worse grades on the next mathematics test and would be satisfied with lower grades than overestimated students. Teachers' judgment error did not confine to test performance but generalized to students' motivational-affective characteristics.

**Keywords:** accuracy of teacher judgment, diagnostic competence, motivation, self-concept of ability, test anxiety

## 1 Einführung

Gerechtigkeit spielt in der Schule eine große Rolle. Schüler wollen von ihrem Lehrer gerecht beurteilt werden und Lehrer sind darum bemüht, möglichst gerecht zu urteilen (Brookhart, 1993; Dalbert, Schneidewind & Saalbach, 2007). Dennoch ergeben sich Fälle, in denen Lehrer die

Leistungsfähigkeit von Schülern über- oder unterschätzen. Studien zur Akkuratheit des Lehrurteils zeigen, dass Lehrkräfte eine Tendenz aufweisen, Schülerleistungen zu positiv zu beurteilen (Artelt, Stanat, Schneider & Schiefele, 2001; Bates & Nettelbeck, 2001; Begeny, Eckert, Montarello & Storie, 2008; Demaray & Elliott, 1998; Feinberg & Shapiro, 2003, 2009; Hamilton & Shinn, 2003). Im Allge-

meinen wird kaum ein Schüler etwas dagegen haben, wenn der Lehrer ein zu positives Urteil fällt. Doch wie fühlen sich die Schüler, die vom Lehrer in ihrer Leistungsfähigkeit unterschätzt werden? Ist die Fehleinschätzung von Schülerleistungen nur eine Frage der Beurteilergenauigkeit oder unterscheiden sich unterschätzte und überschätzte Schüler in ihren Eigenschaften voneinander?

In diesem Beitrag werden zwei Untersuchungen zu den motivationalen und affektiven Merkmalen unterschätzter Schüler im Vergleich zu überschätzten Schülern vorgestellt. Die Studien in vierten Grundschulklassen in Deutschland und China untersuchen, welches Fähigkeitsselbstkonzept unterschätzte Schüler aufweisen, welche Lernmotivation sie zeigen und mit welchen Ansprüchen, Erwartungen und Gefühlen sie Tests und Prüfungen entgegensehen. Darüber hinaus wird berichtet, welches Bild Lehrkräfte von dem Fähigkeitsselbstkonzept, der Motivation und der Leistungsangst von unterschätzten und überschätzten Schülern haben.

## 2 Theoretischer Hintergrund

Als diagnostische Kompetenz wird die Fähigkeit von Lehrkräften zur genauen Beurteilung von Schülermerkmalen bezeichnet (Artelt & Gräsel, 2009; Karing, 2009). Die diagnostische Kompetenz von Lehrkräften dient der angemessenen Planung, Gestaltung und Evaluation von Unterrichtsprozessen (Schrader, 2009) und ist insbesondere dann gefordert, wenn über die schulische oder berufliche Zukunft von Heranwachsenden entschieden wird (Ophuysen, 2006; Trautwein & Baeriswyl, 2007). Alarmiert durch die geringen Diagnoseleistungen von Lehrkräften bei der Erkennung förderungsbedürftiger Schüler in den PISA-Untersuchungen (Artelt et al., 2001) hat sich die Forschung zur diagnostischen Kompetenz in den letzten Jahren verstärkt (Artelt & Gräsel, 2009). Experimentelle, fächervergleichende und längsschnittlich angelegte Studien erlauben es, detailliertere Kenntnisse über Bedingungen, Spezifität und zeitliche Stabilität von Diagnoseleistungen zu erlangen (Hinnant, O'Brien & Ghazarian, 2009; Krolak-Schwerdt, Böhmer & Gräsel, 2009; Lorenz & Artelt, 2009; Südkamp & Möller, 2009; Südkamp, Möller & Pohlmann, 2008).

Zur Messung der diagnostischen Kompetenz von Lehrkräften hat sich ein einfaches Verfahren eingebürgert (Hoge & Coladarsi, 1989). Üblicherweise wird mit einer Gruppe von Schülern ein standardisierter Leistungstest durchgeführt. Die Lehrer sollen einschätzen, wie gut die Schüler darin abschneiden werden. Die Korrelation zwischen der vorhergesagten und der tatsächlichen Schülerleistung bildet ein Maß für die Genauigkeit des Lehrerurteils. Anstelle eines standardisierten Leistungstests können auch stärker unterrichtsbezogene Maße zum Einsatz kommen (Feinberg & Shapiro, 2003, 2009; Hamilton & Shinn, 2003).

Verschiedene Messmethoden wurden entwickelt, um die Genauigkeit des Lehrerurteils näher zu untersuchen. Bei

direkter Messung schätzt der Lehrer die Itemanzahl, die jeder Schüler im standardisierten Leistungstest voraussichtlich lösen wird (Helmke & Schrader, 1987). Im Gegensatz dazu wird bei indirekter Messung eine Ratingskala verwendet, was die Einschätzungen erleichtert, aber mit einer geringeren Variabilität der Einschätzungsmöglichkeiten einhergeht (Begeny et al., 2008). Eine andere Unterscheidung ist dadurch gegeben, dass die Lehrkräfte Schüler im Vergleich zu ihren Klassenkameraden auf einem peer-abhängigen Maß oder ohne Bezug zu anderen auf einem peer-unabhängigen Maß beurteilen. Die Forschung zeigt, dass es keine generellen Vorteile für direkte oder indirekte, peer-abhängige oder peer-unabhängige Messverfahren gibt. Hoge und Coladarsi (1989) berichten in ihrer Metaanalyse, dass der Korrelationsmedian für Studien mit einer peer-abhängigen Messung bei .64 liegt, wohingegen Messungen mit einem peer-unabhängigen Maß einen Korrelationsmedian von .68 aufweisen. Indirekte Messungen führen mit einem Korrelationsmedian von .62 zu einer vergleichbaren Genauigkeit wie direkte Messungen, bei denen der Korrelationsmedian bei .69 liegt. Eine neue Metaanalyse zur Genauigkeit des Lehrerurteils weist allerdings signifikant höhere Korrelationen bei direkter als bei indirekter Messmethode aus (Südkamp, Kaiser & Möller, submitted).

Zur Bewertung der diagnostischen Kompetenz von Lehrkräften können verschiedene Indikatoren herangezogen werden, die als Rang-, Niveau- und Differenzierungs-komponente bezeichnet werden (Schrader & Helmke, 1987; Spinath, 2005). Die Rangkomponente zeigt, inwiefern die Lehrkraft die Rangfolge der Schüler hinsichtlich eines Merkmals wie Leistung, Motivation oder Intelligenz richtig vorhersagen kann. Die Rangkomponente wird über eine Pearson-Korrelation zwischen Lehrerurteil und Schülermerkmal ermittelt und gilt als das zentrale Maß für die Akkuratheit von Lehrerurteilen (Südkamp et al., 2008). Die Niveauelemente misst, inwiefern Lehrer zu Unter- oder Überschätzungen von Schülermerkmalen neigen. Die Niveauelemente ergibt sich aus der Differenz zwischen Lehrerurteil und Schülermerkmal. Die Forschung zeigt, dass Lehrkräfte eine Tendenz zur Überschätzung des Leistungsniveaus von Schulklassen haben (Bates & Nettelbeck, 2001; Begeny et al., 2008; Demaray & Elliott, 1998; Hamilton & Shinn, 2003; Helmke, Hosenfeld & Schrader, 2004). Durch die Differenzierungskomponente wird abgebildet, inwiefern Lehrkräfte zur Über- oder Unterschätzung der Streuung eines Schülermerkmals neigen. Die Differenzierungskomponente wird selten berichtet, doch deuten die Forschungsergebnisse von Schrader und Helmke (1987) auf eine Überschätzung der Leistungsstreuung hin.

Lehrer sind also relativ gute Beurteiler von Schülerleistungen – zumindest wenn es darum geht, Schüler entsprechend ihrer Testleistung in eine Rangfolge zu bringen. In einer Metaanalyse zum Lehrerurteil über schulische Leistungen wird für die Rangkomponente ein Median der Korrelationen von .66 berichtet (Hoge & Coladarsi, 1989). Neuere Untersuchungen weichen in ihren Ergebnissen nicht wesentlich von dem berichteten Korrelationswert ab



(Bates & Nettelbeck, 2001; Demaray & Elliott, 1998; Karing, 2009). Allerdings ist nicht jeder Lehrer ein guter Diagnostiker. So berichten z. B. Helmke und Schrader (1987) in einer Untersuchung mit 32 Lehrern einen Median der Rangkomponente von .67, aber eine Schwankungsbreite der Korrelationen von .03 bis .90.

Die Frage der Urteilsgenauigkeit von Lehrkräften im motivational-affektiven Bereich gilt als weniger gut untersucht (Karing, 2009). Im Allgemeinen fallen hier die Korrelationen jedoch geringer aus als im Leistungsbereich. So berichten Marsh und Craven (1991) für verschiedene schulische Selbstkonzepte Korrelationen zwischen .39 und .62. Eine neuere Untersuchung an Grundschulen (Praetorius, Greb, Dickhäuser & Lipowsky, in Druck) erbrachte Rangkomponenten von .55 für das Selbstkonzept in Mathematik, .52 für das Selbstkonzept im Lesen und .25 für das Selbstkonzept im Schreiben. In einer längsschnittlichen Studie wurde nicht nur die Urteilsgenauigkeit für das Fähigkeitsselbstkonzept, sondern auch für die Lernzielorientierung der Schüler analysiert (Givvin, Stipek, Salmon & MacGyvers, 2001). Zum Anfang und Ende des Schuljahres zeigten sich für beide Dimensionen stabile, aber geringe Korrelationen von Lehrer- und Schülerurteil zwischen .26 und .30. Ebenso schwierig wie die Beurteilung von Motivation scheint für Lehrkräfte auch die Einschätzung von Leistungsangst zu sein. Helmke und Fend (1982) konnten keine substanziellen Korrelationen bei der Beurteilung von Leistungsangst finden. Boehnke, Silbereisen, Reynolds und Richmond (1986) berichten über eine nur schwache Korrelation von .21 zwischen Lehrerurteil und der manifesten Angst von Schülern. Die geringe diagnostische Güte des Lehrerurteils erklärt Faber (2001) damit, dass die Einschätzung der Leistungsangst maßgeblich vom aktuellen Leistungsniveau der Schüler abhängig erscheint. Die Leistungseinschätzung der Lehrkräfte hat eine Leitfunktion für die Beurteilung anderer Schülermerkmale – ein Phänomen, das in der Psychologie als Halo-Effekt bekannt ist (Thorndike, 1920).

Eine umfassende Untersuchung zu motivational-affektiven Schülermerkmalen, an die sich auch die eigene Studie anlehnt, stammt von Spinath (2005). Die Autorin untersuchte mittels Selbstbericht das Fähigkeitsselbstkonzept, die Lernmotivation und die Leistungsängstlichkeit von Grundschulkindern. Die Klassenlehrer wurden gebeten, über einzelne Items das schulische Selbstkonzept, die Lernfreude und die Schulängstlichkeit eines jeden Schülers im Vergleich zu anderen Kindern im gleichen Alter einzuschätzen. Die gefundenen Korrelationen sprechen für eine mittlere Genauigkeit der Einschätzung des Fähigkeitsselbstkonzepts ( $r = .39$ ), aber eine niedrige Genauigkeit bei der Beurteilung von Lernmotivation ( $r = .20$ ) und Leistungsängstlichkeit ( $r = .15$ ).

Während die meisten Untersuchungen zur diagnostischen Kompetenz von Lehrkräften sich auf die Rangkomponente fokussieren, hat die Niveauelemente bislang nur wenig Aufmerksamkeit erfahren. Zwar wird oft berichtet, dass Lehrkräfte zu Überschätzungen von Schülerleis-

tungen neigen (Artelt et al., 2001; Bates & Nettelbeck, 2001; Begeny et al., 2008; Demaray & Elliott, 1998; Feinberg & Shapiro, 2003, 2009; Hamilton & Shinn, 2003), doch wie kommen diese überhaupt zustande? Was sind die Gründe dafür, dass Lehrkräfte zu verzerrten Urteilen über Schüler gelangen?

Antworten auf diese Fragen lassen sich in der Forschung zur sich selbst erfüllenden Prophezeiung finden (Jussim & Harber, 2005; Jussim, Robustelli & Cain, 2009): Erstens ist als Erklärung für ein mangelhaftes Lehrerurteil zu sagen, dass Schüler sich mit der Zeit in ihren Fähigkeiten verändern. Sie nehmen neue Dinge auf und vergessen andere. Das macht es für den Lehrer schwierig, immer das richtige Urteil zu fällen. Zweitens kann auch die Erinnerung der Lehrkräfte an Leistungen der Schüler falsch sein (Jussim et al., 2009). Über einige Schüler hat der Lehrer mehr und bessere Informationen zur Verfügung als über andere und sein Urteil wird dementsprechend unterschiedlich genau sein. Drittens kann eine Lehrkraft auch von äußeren Merkmalen beeinflusst sein und bestimmten Gruppen, die sich z. B. in Hinblick auf Geschlecht, ethnische Zugehörigkeit oder soziale Herkunft voneinander unterscheiden, weniger zutrauen als anderen (Madon et al., 1998).

Zusammengenommen führen die oben genannten Gründe dazu, dass sich auf Grundlage der ungenauen Leistungsprognose von Lehrkräften Gruppen von unterschätzten und überschätzten Schülern voneinander unterscheiden lassen. Bislang ist jedoch noch nicht untersucht worden, ob mit diesem Umstand weitergehende Konsequenzen verbunden sind. An dieser Stelle setzen die eigenen Untersuchungen an, die sich den motivational-affektiven Merkmalen von über- und unterschätzten Schülern zuwenden. Wie fühlen sich zwei leistungsmäßig gleich starke Schüler, von denen der Lehrer die Leistung des einen zu positiv und die Leistung des anderen zu negativ beurteilt? Cooley (1902, S. 175) sagt dazu: «In the presence of one whom we feel to be of importance, there is a tendency to enter into and adopt, by sympathy, his judgment of ourself.» Insofern würde Unterschätzung des Lehrers zu einer Unterschätzung der eigenen Fähigkeiten beitragen, während Überschätzung des Lehrers die Wahrnehmung eigener Kompetenzen erhöht. Dieses galt es zu überprüfen.

### 3 Hypothesen

In zwei Untersuchungen in Deutschland und China stand die Untersuchung von unterschätzten und überschätzten Schülern im Mittelpunkt. Es wurden drei Hypothesen formuliert, die Lehrerurteil und Selbstwahrnehmung der Schüler miteinander verbinden:

1. Lehrer können Schülerleistungen relativ genau einschätzen (Hoge & Coladarci, 1989), haben aber Schwierigkeiten damit, Fähigkeitsselbstkonzept, Motivation und Leistungsangst der Schüler zu beurteilen (Spinath, 2005).

2. Unterschätzte Schüler zeigen ein niedrigeres Fähigkeitsselbstkonzept, eine geringere Motivation und mehr Leistungsangst als überschätzte Schüler.
3. Lehrer beurteilen unterschätzte und überschätzte Schüler nicht nur hinsichtlich der Testleistung falsch, sondern auch in Bezug auf Fähigkeitsselbstkonzept, Motivation und Leistungsangst. Es zeigt sich ein Halo-Effekt von der Testleistung auf motivational-affektive Schülermerkmale. Lehrer nehmen für unterschätzte Schüler an, dass diese auch über ein geringeres Fähigkeitsselbstkonzept, eine geringere Motivation und mehr Leistungsangst verfügen als überschätzte Schüler.

## 4 Studie 1

### 4.1 Stichprobe

An der ersten Studie nahmen acht Klassen mit insgesamt 144 Grundschulern der vierten Jahrgangsstufe aus dem Raum München teil. Die Klassengröße schwankte zwischen 12 und 23 Schülern. Die 78 männlichen und 64 weiblichen Schüler (2 Schüler ohne Geschlechtsangabe) waren im Durchschnitt 9.93 Jahre alt ( $SD = .61$ ). Die Schüler wurden in Mathematik, Deutsch und anderen Fächern von acht Klassenlehrerinnen unterrichtet, welche im Mittel 42.13 Jahre ( $SD = 10.99$ ) alt waren. Die Lehrerinnen verfügten über eine hohe Berufserfahrung von durchschnittlich 14.38 Jahren ( $SD = 9.43$ ) und unterrichteten die Klassen fast zwanzig Stunden in der Woche ( $M = 19.63$ ,  $SD = 3.11$ ).

### 4.2 Material

#### 4.2.1 Schülerdaten

Im Verlauf der Untersuchung wurden die Grundschüler gebeten, einen standardisierten Mathematikleistungstest und einen Schülerfragebogen zu bearbeiten. Die verwendeten Materialien sind im Folgenden näher beschrieben.

#### 4.2.2 Mathematikleistung

Zur Messung der mathematischen Fähigkeiten wurde der Deutsche Mathematiktest für vierte Klassen (DEMAT 4; Gölit, Roick & Hasselhorn, 2006) eingesetzt. Die Itemzusammenstellung des DEMAT 4 orientiert sich an den Lehrplänen aller deutschen Bundesländer. Insgesamt beinhaltet der Mathematiktest vierzig Aufgaben aus den Bereichen Arithmetik (Zahlenstrahlen, Additionen, Subtraktionen, Multiplikationen und Divisionen), Algebra (Größenvergleiche und Sachrechnungen) und Geometrie (Lagebeziehungen und Spiegelzeichnungen). Der Test ist in zwei Parallelförmigen verfügbar, die beide in der Studie eingesetzt

wurden, um das Abschreiben von Lösungen zu verhindern. Von den Testautoren wird eine interne Konsistenz des DEMAT 4 zum Ende der vierten Klasse von Cronbachs  $\alpha = .85$  berichtet. Zur Konstruktvalidität wird eine Korrelation von  $.70$  mit den Mathematiknoten zum Ende der vierten Klasse angegeben (Gölit et al., 2006). Um den Lehrkräften das Einschätzen von Schülerleistungen zu erleichtern, wurden einzelne Teilaufgaben zu einer Aufgabe zusammengefasst. Zum Beispiel sind in einem Busfahrplan vier Abfahrtszeiten zu ergänzen. Diese Aufgabe wurde nur dann als gelöst gewertet, wenn alle Zeitangaben richtig waren. Dadurch verringerte sich die Gesamtaufgabenzahl des DEMAT 4 von 40 auf 36 Aufgaben.

Zur Messung motivationaler und affektiver Schülermerkmale wurden Items und Skalen aus der Ulmer Motivationstestbatterie (Ziegler, Dresel, Schober & Stöger, 2005) und der Dokumentation der Erhebungsinstrumente zu PISA 2003 (PISA-Konsortium Deutschland, 2006) entnommen. Alle Items, falls nicht anders angegeben, wurden auf einer vierstufigen Likertskala von 1 – «trifft nicht zu» bis 4 – «trifft genau zu» bewertet. Auf diese Weise wurden folgende Schülermerkmale erfasst:

#### 4.2.3 Mathematisches Selbstkonzept

Beim mathematischen Selbstkonzept handelt es sich um die Einschätzung der eigenen Fähigkeiten und Kompetenzen für Mathematik. Nach dem hierarchischen Selbstkonzeptmodell von Shavelson, Hubner und Stanton (1976) ist das mathematische Selbstkonzept Teil des schulischen Selbstkonzepts, welches wiederum einen Teil des allgemeinen Selbstkonzepts darstellt (s. a. Möller & Trautwein, 2009). Das mathematische Selbstkonzept wurde über fünf Items mit einer Reliabilität von  $\alpha = .86$  erfasst. Ein Beispielitem aus der Ulmer Motivationstestbatterie lautet: «Im Mathematikunterricht verstehe ich sogar die schwierigsten Aufgaben.»

#### 4.2.4 Lernmotivation

Es lassen sich verschiedene Arten der Lernmotivation voneinander unterscheiden (Urhahne, 2008). Ein sehr populäres und zugleich dem Alltagsverständnis von Lernmotivation sehr nahe kommendes Konzept ist die Lernzielorientierung (Dweck, 1986). Schüler mit einer ausgeprägten Lernzielorientierung versuchen sich selbst zu verbessern, neue Fähigkeiten zu entwickeln, etwas Herausforderndes zu meistern sowie Einsicht und Verständnis zu erlangen (Schunk, Pintrich & Meece, 2007). Die Lernzielorientierung wurde über fünf Items mit einer Reliabilität von  $\alpha = .75$  abgefragt. Ein Beispielitem der Ulmer Motivationstestbatterie heißt: «Es ist mir sehr wichtig, dass ich den Stoff in Mathematik wirklich verstehe.»

#### 4.2.5 Leistungsangst

Wenn Schüler ein Ungleichgewicht zwischen persönlichen Fähigkeiten und schulischen Herausforderungen wahrnehmen, kann Leistungsangst entstehen (Rost & Schermer, 2006). Im Fragebogen wurde die Leistungsangst der Schüler mit vier Items aus der Ulmer Motivationstestbatterie und zwei Items aus PISA 2003 mit einer Reliabilität von  $\alpha = .83$  gemessen. Ein Beispielitem der Skala lautet: «Vor Klassenarbeiten in Mathematik bin ich sehr nervös.»

Die eingesetzten motivational-affektiven Skalen weisen nur wenige fehlende Werte auf. Um die Power der statistischen Analysen noch zu erhöhen, wurde der Estimation Maximization-Algorithmus angewendet. Dieses Verfahren wird empfohlen, um fehlende Daten von Ratingskalen zu ersetzen (Lüdtke, Robitzsch, Trautwein & Köller, 2007). Der Prozentsatz der geschätzten Werte beträgt für das mathematische Selbstkonzept 1.8 %, für Lernmotivation 2.8 % und für Leistungsangst 1.4 %.

#### 4.2.6 Erfolgserwartung

Die Überzeugung von Schülern, bei einer gegebenen Aufgabe erfolgreich zu sein, wird als Erfolgserwartung bezeichnet (Schunk et al., 2007). Die Erfolgserwartung der Grundschüler in Bezug auf das Fach Mathematik wurde über ein Item aus der Ulmer Motivationstestbatterie erfasst: «Was denkst du, welche Note wirst du in der nächsten Mathematikarbeit erhalten?» Die Schüler konnten Noten von 1 bis 6 angeben.

#### 4.2.7 Anspruchsniveau

Sobald Kinder die multiplikative Verbindung von Erfolgserwartung und Erfolgsanreiz verstehen, sind sie in der Lage ein Anspruchsniveau zu setzen (Heckhausen & Heckhausen, 2008). Das Anspruchsniveau bildet einen Standard zur Beurteilung der eigenen Leistung bei einer Aufgabe (Schunk et al., 2007). Entsprechend der Vorgabe aus der Ulmer Motivationstestbatterie wurden die Schüler gefragt: «Mit welcher Note in der nächsten Mathematikarbeit wärest du gerade noch zufrieden?» Wiederum konnten die Schüler Noten von 1 bis 6 vergeben.

#### 4.2.8 Lehrerdaten

Den Grundschullehrerinnen wurden zunächst soziodemographische Fragen zu Alter, Geschlecht, Ausbildung und Berufserfahrung gestellt, bevor sie dazu übergingen, die Leistungen und Einstellungen der Schüler zu beurteilen. Um mit den Aufgaben des Mathematiktests vertraut zu werden, erhielten die Lehrer eine Kopie des DEMAT 4. Anschließend wurden die Lehrer gebeten, für jeden Schü-

ler in der Klasse sechs Items bezogen auf das Fach Mathematik zu beantworten:

- Testleistung (Wie viele der 36 Aufgaben des Mathematiktests löst der Schüler korrekt?)
- Fähigkeitsselbstkonzept (Wie schätzt der Schüler seine eigenen mathematischen Fähigkeiten ein?)
- Lernmotivation (Wie interessiert und zielorientiert lernt der Schüler?)
- Leistungsangst (Wie viel Angst hat der Schüler vor Prüfungen?)
- Erfolgserwartung (Schätzen Sie ein, wie der Schüler die folgende Frage beantwortet: Was denkst du, welche Note wirst du in der nächsten Mathematikarbeit erhalten?)
- Anspruchsniveau (Schätzen Sie ein, wie der Schüler die folgende Frage beantwortet: Mit welcher Note in der nächsten Mathematikarbeit wärest du gerade noch zufrieden?)

Zur peerabhängigen Messung von Fähigkeitsselbstkonzept, Lernmotivation und Leistungsangst wurde eine neunstufige Ratingskala mit den Ankerpunkten 1 – «sehr viel geringer», 5 – «gleich» und 9 – «sehr viel größer» verwendet, um eine ausreichende Differenzierung in der Beurteilung der Schüler zu ermöglichen. Für Erfolgserwartung und Anspruchsniveau sollten die Lehrerinnen Schulnoten von 1 bis 6 angeben.

### 4.3 Untersuchungsablauf

Die Untersuchung im Klassenraum nahm zwei Schulstunden in Anspruch. In der ersten Stunde wurde der standardisierte Mathematikleistungstest durchgeführt. In der zweiten Stunde bearbeiteten die Schüler die Skalen und Items zur Messung von Motivation und Affekt. Dabei wurden die Items von einem Versuchsleiter im Klassenzimmer laut vorgelesen, um ein gleichmäßiges Bearbeitungstempo aller Schüler zu gewährleisten.

Den Lehrerinnen wurde es selbst überlassen, ob sie den Lehrerfragebogen in oder außerhalb der Klasse ausfüllen. Für gewöhnlich konnten die Lehrkräfte diese Aufgabe in 45 Minuten bewältigen. Nach einigen Wochen erhielten die Lehrerinnen Rückmeldungen über das Abschneiden der Klasse im Mathematikleistungstest und über die Genauigkeit ihrer Einschätzungen.

## 5 Ergebnisse

### 5.1 Akkuratheit des Lehrerurteils

Durchschnittlich wurde die Testleistung der Schüler um  $M = 3.07$  Punkte ( $SD = 5.43$ ) überschätzt, bei einer Schwankungsbreite von  $-16$  bis  $+14$  Punkten. Die Niveauelemente der Testleistung liegt bei  $M = 3.51$  ( $SD = 2.99$ ) und für die Differenzierungskomponente ergibt sich ein



Tabelle 1  
Mittelwertvergleiche der Schülermerkmale und Lehrerurteile der deutschen Stichprobe

Variable	Fehleingeschätzte Schüler				Deutlich fehleingeschätzte Schüler			
	Schülermerkmal		Lehrerurteil		Schülermerkmal		Lehrerurteil	
	unterschätzt	überschätzt	unterschätzt	überschätzt	unterschätzt	überschätzt	unterschätzt	überschätzt
Testleistung	20.94 (6.50)	21.02 (5.01)	16.64 (7.58)	27.04 (4.88)	19.85 (6.12)	20.76 (5.14)	13.75 (6.67)	27.68 (4.77)
Erfolgserwartung	2.76 (.98)	2.18 (.79)	3.09 (.84)	2.46 (.82)	2.90 (.91)	2.12 (.78)	3.35 (.75)	2.40 (.84)
Anspruchsniveau	2.96 (.55)	2.81 (.65)	3.27 (.72)	2.73 (.83)	2.95 (.54)	2.75 (.64)	3.50 (.76)	2.71 (.84)
Fähigkeitsselbstkonzept	2.69 (.67)	3.07 (.74)	4.88 (1.29)	5.13 (1.21)	2.60 (.67)	3.15 (.70)	4.70 (1.30)	5.19 (1.23)
Lernmotivation	3.57 (.44)	3.58 (.47)	4.73 (2.33)	5.64 (1.89)	3.60 (.42)	3.62 (.46)	3.90 (2.13)	5.85 (1.89)
Leistungsangst	2.64 (.72)	2.23 (.84)	5.15 (.80)	5.04 (1.11)	2.77 (.77)	2.19 (.84)	5.30 (.87)	5.02 (1.15)

Wert von  $M = 1.35$  ( $SD = .28$ ), welcher auf eine Überschätzung der Variabilität der Schülerleistungen hindeutet. Drei- und dreißig Schüler, 16 Jungen und 17 Mädchen, wurden im Mittel um  $M = 4.30$  Punkte ( $SD = 3.68$ ) unterschätzt, wobei die Anzahl unterschätzter Schüler in den einzelnen Klassen zwischen 0 und 11 Schülern schwankte. Sieben- und neunzig Schüler wurden im Mittel um  $M = 6.02$  Punkte ( $SD = 3.25$ ) überschätzt. Bei 14 Schülern wurde das Testergebnis von den Lehrkräften exakt vorhergesagt. Die fair beurteilten Schüler werden bei den Vergleichen von unterschätzten und überschätzten Schülern nicht berücksichtigt.

Zusätzlich wurde zur Festlegung unterschätzter und überschätzter Schülergruppen ein weiteres, strengeres Kriterium eingeführt. Es kann schlichtweg nicht erwartet werden, dass Lehrer immer genau richtige Einschätzungen der Leistungen ihrer Schüler vornehmen. Psychologisch relevant könnte eine Fehleinschätzung vielleicht erst dann sein, wenn Lehrerurteil und Schülerleistung um ein bestimmtes Maß voneinander abweichen. Als ein solches Maß wurde ein halbe Standardabweichung vom Mittelwert des Leistungstests ( $M = 21.00$ ,  $SD = 5.78$ ) gewählt. Damit werden als deutlich fehleingeschätzte Schüler all jene bezeichnet, bei denen Lehrerurteil und Testleistung mindestens drei Punkte weit auseinander liegen. Die Gruppe der deutlich unterschätzten Schüler umfasst bei Anwendung dieses engeren Kriteriums zwanzig Personen, während die Gruppe der deutlich überschätzten Schüler sich aus 81 Personen zusammensetzt.

Zur Untersuchung der ersten Hypothese wurden Pearson-Korrelationen berechnet, um Zusammenhänge zwischen Lehrerurteilen und Schülermerkmalen zu prüfen. Lehrerurteil und Schülermerkmal korrelieren am höchsten in Bezug auf die Testleistung ( $r = .69$ ,  $p < .001$ ). Auch die Erwartung der Schüler für die nächste Mathematiknote können die Lehrkräfte recht sicher einschätzen ( $r = .61$ ,  $p < .001$ ). Hinsichtlich des Anspruchsniveaus ( $r = .29$ ,  $p < .01$ ) und des mathematischen Selbstkonzepts ( $r = .19$ ,  $p < .05$ ) zeigen sich deutlich geringere Korrelationen zwischen Lehrer- und Schülerurteil. Die Validität des Lehrerurteils in Hinblick auf Lernmotivation ( $r = .07$ ,  $ns$ ) und Leistungsangst ( $r = .07$ ,  $ns$ ) ist sogar ausgesprochen gering. Zusammengefasst bestätigt sich die Vermutung, dass Lehr-

kräfte leistungsabhängige Parameter wie Testleistung oder erwartete Schulnoten gut vorhersagen, motivational-affektive Schülermerkmale aber nur schwer einschätzen können.

## 5.2 Vergleich der Schülergruppen

Zur Untersuchung der zweiten Hypothese wurden multivariate Varianzanalysen berechnet, um Unterschiede zwischen der Selbstwahrnehmung unterschätzter und überschätzter Schüler aufzudecken. In der Tabelle 1 sind die zugehörigen deskriptiven Statistiken dargestellt. Die Vergleiche zwischen unterschätzten und überschätzten Schülergruppen wurden auf Basis klassenweise standardisierter Variablen vorgenommen, die zu den gleichen Ergebnissen führen wie die unstandardisierten Messwerte.

Eine erste multivariate Varianzanalyse zum Vergleich der unterschätzten und überschätzten Schüler in Bezug auf die sechs abhängigen Variablen fällt signifikant aus (Hotelling's  $T^2 = .11$ ,  $F(6, 123) = 2.18$ ,  $p < .05$ ,  $\eta^2 = .096$ ). Die weiteren statistischen Kennwerte des Vergleichs von unterschätzten und überschätzten Schülern sind im oberen linken Viertel der Tabelle 2 dargestellt. Es fällt auf, dass sich unterschätzte und überschätzte Schüler in der Testleistung nicht voneinander unterscheiden. Klare Unterschiede zu Ungunsten der unterschätzten Schüler zeigen sich jedoch in Hinblick auf die Erfolgserwartung und die Leistungsangst. Dagegen haben unterschätzte Schüler ein ähnliches Anspruchsniveau und eine vergleichbare Lernmotivation wie überschätzte Schüler. Der Vergleich des Fähigkeitsselbstkonzepts unterschätzter und überschätzter Schüler wird nur knapp nicht signifikant ( $p = .051$ ) und deutet auf ein niedrigeres mathematisches Selbstkonzept unterschätzter Schüler hin.

Eine zweite multivariate Varianzanalyse wurde für den Vergleich deutlich unterschätzter Schüler und deutlich überschätzter Schüler durchgeführt (Hotelling's  $T^2 = .17$ ,  $F(6, 94) = 2.58$ ,  $p < .05$ ,  $\eta^2 = .14$ ). Aus dem oberen rechten Viertel der Tabelle 2 wird ersichtlich, dass auch bei eindeutiger Separierung der fehleingeschätzten Gruppen sich die Resultate kaum verändern. Keine Unterschiede zwischen den

Tabelle 2

*Ergebnisse der Varianzanalysen für fehleingeschätzte und deutlich fehleingeschätzte Schüler der deutschen Stichprobe*

Variable	Fehleingeschätzte Schüler		Deutlich fehleingeschätzte Schüler	
	$F(1, 128)$	$\eta^2$	$F(1, 99)$	$\eta^2$
	Haupteffekt unterschätzte vs. überschätzte Schüler			
Testleistung	.80	.006	1.07	.011
Erfolgserwartung	8.27**	.061	11.40**	.103
Anspruchsniveau	.69	.005	.96	.010
Fähigkeitsselbstkonzept	3.87	.029	8.50**	.079
Lernmotivation	.03	.000	.27	.003
Leistungsangst	6.13*	.046	7.53**	.071
	Interaktionseffekt Schülermerkmal vs. Lehrerurteil			
Testleistung	56.23***	.305	67.59***	.406
Erfolgserwartung	2.02	.016	4.02*	.039
Anspruchsniveau	6.13*	.046	7.40**	.070
Fähigkeitsselbstkonzept	.50	.004	2.59	.026
Lernmotivation	6.32*	.047	7.82**	.073
Leistungsangst	7.58**	.056	7.14**	.067

Anmerkungen. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

Schülergruppen ergeben sich in Hinblick auf die Testleistung, die Lernmotivation und das Anspruchsniveau. Dagegen zeigen deutlich unterschätzte Schüler eine niedrigere Erfolgserwartung, ein geringeres Fähigkeitsselbstkonzept und mehr Leistungsangst als deutlich überschätzte Schüler. Insgesamt stützen die Ergebnisse die zweite Hypothese, dass unterschätzte Schüler – bei gleicher Leistungsfähigkeit – ungünstigere motivational-affektive Merkmale aufweisen als überschätzte Schüler.

### 5.3 Beziehungen von Lehrerurteilen und Schülermerkmalen

Zur Überprüfung der dritten Hypothese wurden Varianzanalysen mit Messwiederholung gerechnet. Lehrerurteile und zugehörige Schülermerkmale fungierten dabei als Innersubjektvariable und Unterschätzung versus Überschätzung der Testleistung als Zwischensubjektfaktor. Die Varianzanalysen mit Messwiederholung haben einen Haupteffekt und einen Interaktionseffekt. Der Haupteffekt liefert Informationen über Unterschiede zwischen Lehrer- und Schülerurteil. Durch die klassenweise Standardisierung der Variablen verliert der Haupteffekt jedoch seine Aussagekraft. Er wird daher nicht berichtet. Der Interaktionseffekt zeigt, ob Unterschätzung versus Überschätzung den Unterschied zwischen Lehrer- und Schülerurteil moderiert. Über den Interaktionseffekt kann die dritte Hypothese getestet werden. Wenn ein Interaktionseffekt auftritt, haben unterschätzte und überschätzte Schüler eine andere Wahrnehmung ihrer motivational-affektiven Merkmale als sie der Lehrer hat.

Die Ergebnisse der Varianzanalysen mit Messwiederholung sind in der unteren Hälfte der Tabelle 2 gelistet. Die Resultate für fehleingeschätzte und deutlich fehleinge-

schätzte Schüler fallen wiederum recht ähnlich aus. Bei der Testleistung zeigt sich ein starker Interaktionseffekt, da unterschätzte Schüler bei gleicher Leistung von den Lehrkräften schlechter als überschätzte Schüler beurteilt werden. Bei der Erfolgserwartung und beim Fähigkeitsselbstkonzept ergeben sich keine Interaktionseffekte. Die Lehrer trauen den unterschätzten Schülern genau so wenig zu wie diese sich selbst. Die Lehrer vermuten, dass unterschätzte Schüler ein geringeres mathematisches Selbstkonzept haben und in der nächsten Mathematikarbeit etwa eine halbe Note schlechter abschneiden werden als überschätzte Schüler. Bei den deutlich fehleingeschätzten Schülergruppen beträgt der Unterschied im Lehrerurteil sogar fast eine volle Note (s. Tabelle 1). In letzterem Fall wird auch die Interaktion von Schülermerkmal und Lehrerurteil bei der Erfolgserwartung signifikant (s. Tabelle 2). Beim Anspruchsniveau, der Lernmotivation und der Leistungsangst zeigen sich bei den Gruppen der fehleingeschätzten und der deutlich fehleingeschätzten Schüler signifikante Interaktionseffekte. Die Lehrerinnen glauben, dass unterschätzte Schüler ein niedrigeres Anspruchsniveau und eine niedrigere Lernmotivation haben, was aber tatsächlich nicht der Fall ist. Bei der Leistungsangst nehmen die Lehrkräfte die stärkere Angst der unterschätzten Schüler nicht ausreichend wahr. Insgesamt fällt in keinem der angestellten Vergleiche das Lehrerurteil zugunsten der Gruppe der unterschätzten Schüler aus.

## 6 Diskussion

In der deutschen Grundschuluntersuchung konnte insgesamt eine Überschätzung der Testleistung durch die Lehr-

kräfte festgestellt werden. Dieses Ergebnis ordnet sich in den Kontext von Studien ein, die auf ein zu positives Lehrerurteil über Schülerleistungen verweisen (Bates & Nettelbeck, 2001; Begeny et al., 2008; Demaray & Elliott, 1998; Helmke et al., 2004). Der Grund dafür könnte sein, dass Lehrkräfte nicht die Performanz, sondern die Kompetenz der Schüler beurteilen und so zu positiv gefärbten Urteilen gelangen. Insgesamt muss den beteiligten Grundschullehrkräften aber eine gute diagnostische Kompetenz für den Leistungsbereich bescheinigt werden (s. a. Karing, 2009). Die positive Verzerrung des Lehrerurteils ist nicht zu stark ausgeprägt und auch die Rangfolge der Schüler im Test kann mit hinreichender Genauigkeit vorhergesagt werden.

Weniger genau fallen die Lehrerurteile über motivational-affektive Komponenten aus. Auch dieses Ergebnis steht in Einklang mit bisherigen Forschungsergebnissen (Spinath, 2005). Solange die einzuschätzende Variable einen Bezug zum kognitiven Bereich zeigt, wie bei der Erfolgserwartung, dem Anspruchsniveau und dem mathematischen Selbstkonzept, hat das Lehrerurteil noch Erklärungskraft. Beunruhigend ist allerdings, dass Lehrer- und Schülerurteil über Lernmotivation und Leistungsangst so weit auseinanderklaffen. Den Lehrkräften ist nicht klar, welches Kind in der Klasse einer besonderen motivationalen oder emotionalen Unterstützung bedarf. Freilich ist die Erschließung motivational-affektiver Zustände keine leichte Aufgabe, dennoch wäre mehr diagnostische Kompetenz der Lehrkräfte im nicht-kognitiven Bereich durchaus wünschenswert.

Der Schwerpunkt der Untersuchung lag auf der Unterscheidung von unterschätzten und überschätzten Schülern. Dieses Schülermerkmal stellt einen neuartigen Forschungsaspekt dar. Bisherige Studien zum Lehrerurteil nahmen lediglich die Fehleinschätzung von Schülerleistungen in den Blick, nicht aber die motivational-affektiven Besonderheiten der beurteilten Schüler. Hervorzuheben ist an den Ergebnissen, dass unterschätzte und überschätzte Schüler sich in Testleistung und selbstberichteter Lernmotivation nicht voneinander unterscheiden. Gleichwohl haben die unterschätzten Schüler weniger Selbstvertrauen in das eigene Leistungsvermögen, sind stärker prüfungsängstlich und erwarten deutlich schlechtere Noten als die überschätzten Schüler. Der Erwartungsunterschied von mehr als einer halben Note bei gleicher Leistungsstärke der Schülergruppen ist beachtlich.

Hinzu kommt, dass die Lehrkräfte die vergleichsweise negativere Auffassung über unterschätzte Schüler teilen. Auch sie erwarten ein um mehr als eine halbe Note schlechteres Abschneiden der unterschätzten Schüler in der nächsten Mathematikarbeit und schreiben ihnen ferner ein geringeres Anspruchsniveau zu. Dabei wollen die unterschätzten Schüler genau so gute Noten erzielen wie ihre überschätzten Klassenkameraden.

Es stellt sich die Frage, wie die Unterschiede zwischen unterschätzten und überschätzten Schülern in den motivational-affektiven Variablen zu erklären sind. Eine mögliche

Erklärung könnte in Erwartungseffekten bestehen, die dazu führen, dass sich Schüler nach den Vorstellungen ihrer Lehrkräfte entwickeln (Rosenthal & Jacobson, 1968). Eine andere Erklärung könnte darin bestehen, dass unterschätzte Schüler in der Vergangenheit schlechtere Lernleistungen gezeigt haben als überschätzte Schüler und folgerichtig schlechtere Noten erwarten. Leider wurden in dieser Studie wie in Studie 2 keine Schulnoten erhoben, so dass diese Hypothese erst in Folgeuntersuchungen getestet werden kann. Bei der Fragwürdigkeit der Zensurengebung (Ingenkamp, 1995) ist jedoch nicht davon auszugehen, dass die Fehleinschätzung der Lehrer im Leistungstest vollkommen durch vorherige Schulnoten erklärt werden kann.

Zusammenfassend lässt sich festhalten, dass für alle drei Hypothesen Belege gesammelt werden konnten. Die Grundschullehrkräfte erwiesen sich als gute Diagnostiker kognitiver, aber als unzureichende Diagnostiker motivational-affektiver Schülermerkmale. Leistungsmäßig unterschätzte Schüler wurden von den Lehrkräften auch auf motivational-affektiven Dimensionen schlechter beurteilt.

## 7 Studie 2

In der zweiten Studie sollte überprüft werden, wie verlässlich die erzielten Ergebnisse sind und ob Unterschiede zwischen unterschätzten und überschätzten Schülern auch in einer anderen Population beobachtet werden können. Zu diesem Zweck wurde die Studie in der Volksrepublik China mit Grundschulern vierter Klassen unter Beibehaltung der Hypothesen wiederholt.

Die Struktur des chinesischen Bildungssystems ist der des deutschen Systems nicht unähnlich. In China kommen die Kinder im Alter von sechs Jahren in die Grundschule, die sechs Jahre dauert. Danach folgen drei Jahre Unterstufe in der Mittelschule, womit die verpflichtende Schulzeit in China endet. Anschließend findet wie in Deutschland eine Selektion statt. Nur die besten 50 % eines Jahrgangs besuchen die dreijährige Oberstufe der Mittelschule (Ludwig, 2005). Eine weitere Gemeinsamkeit zwischen dem deutschen und dem chinesischen Bildungssystem besteht darin, dass dem Lehrerurteil eine hohe Bedeutung zukommt. In Bayern, dem Bundesland der ersten Studie, erfolgt nach der vierten Klasse der Übertritt auf weiterführende Schulen im Wesentlichen nach Maßgabe des Lehrerurteils. In China genießt der Lehrberuf ein hohes Ansehen (Ludwig, 2005) und das Lehrerurteil hat ein entsprechendes Gewicht.

## 8 Methode

In der zweiten Studie wurden die gleichen Messinstrumente verwendet und der gleiche Untersuchungsablauf gewählt, so dass an dieser Stelle nur die Unterschiede zur ersten Studie hervorgehoben werden.

## 8.1 Stichprobe

Die Untersuchung wurde in acht Grundschulklassen mit insgesamt 272 Viertklässlern in der chinesischen Hauptstadt Peking durchgeführt. Die Klassengröße in den chinesischen Schulen lag zwischen 26 und 46 Schülern, was die Lehrkräfte vor die schwierige Aufgabe stellt, gerechte Beurteilungen für alle Schüler vorzunehmen. Die 150 Jungen und 122 Mädchen hatten ein durchschnittliches Alter von 9.87 Jahren ( $SD = .65$ ). Die Lehrer, drei Männer und fünf Frauen, waren im Durchschnitt 34.14 Jahre alt ( $SD = 8.55$ ). Sechs der teilnehmenden Lehrer unterrichteten fünf Stunden Mathematik in der Woche, zwei Lehrer unterrichteten sieben Stunden wöchentlich in den untersuchten Klassen. Sie arbeiteten seit durchschnittlich 12.75 Jahren ( $SD = 6.43$ ) als Mathematiklehrer. Die hohe Berufserfahrung der relativ jungen Grundschullehrer kommt durch eine verkürzte Ausbildungsphase zustande. Nach Absolvierung der Unterstufe der Mittelschule erfolgt eine berufsqualifizierende Ausbildung von drei bis vier Jahren Dauer für lediglich ein Unterrichtsfach.

## 8.2 Material

Alle Fragebögen wurden von einer Koautorin vom Deutschen ins Chinesische übersetzt. Weitere Koautoren halfen die Items sprachlich an das Verständnis von Viertklässlern anzupassen. Ein gebürtiger Chinese, der seit vielen Jahren in Deutschland lebt, übersetzte den Lehrer- und den Schülerfragebogen zurück ins Deutsche. Die Rückübersetzung der Items zeigte, dass diese sprachlich äquivalent mit den ursprünglichen Items waren.

### 8.2.1 Schülerdaten

Zur Messung der Mathematikleistung wurde der DEMAT 4 eingesetzt, der nach Aussagen der Lehrkräfte gut mit dem chinesischen Mathematikcurriculum übereinstimmt. Zur Erfassung des mathematischen Selbstkonzepts (Cronbachs  $\alpha = .81$ ) und der Leistungsangst ( $\alpha = .79$ ) wurden die gleichen Messinstrumente wie in der ersten Studie genutzt. Zur Messung von Erfolgserwartung und Anspruchsniveau wurde vom chinesischen Notensystem Gebrauch gemacht. Dadurch konnten die Kinder auf die Fragen, welche Note sie in der nächsten Mathematikarbeit erwarten und mit welcher Note sie gerade noch zufrieden wären, Zahlen zwischen 0 (schlechteste Note) und 100 (beste Note) angeben. Zur besseren Einordnung: In der Grundschule werden wie in Deutschland häufig noch sehr gute Noten vergeben. Die Note 60 steht für bestanden. Um den weiteren Besonderheiten des chinesischen Bildungssystems Rechnung zu tragen, wurde anstelle von Lernmotivation die Bereitschaft zur Lernanstrengung erfragt.

### 8.2.2 Lernanstrengung

Chinesen legen einen hohen Wert auf Bildung und betonen die Rolle von Anstrengung für das Zustandekommen exzellenter Leistungen. Die Lehren des Konfuzius fördern den Glauben an die Veränderbarkeit des menschlichen Verhaltens und fleißiges Lernen wird als der Weg zur Zielerreichung gesehen (Chen & Stevenson, 1995). Die Items zur Lernanstrengung wurden der Dokumentation der Erhebungsinstrumente zu PISA 2003 (PISA-Konsortium Deutschland, 2006) entnommen. Die Skala mit fünf Items erreicht eine interne Konsistenz von  $\alpha = .69$ . Ein Beispielitem zur Lernanstrengung lautet: «Ich gebe nicht auf, auch wenn die Mathematikaufgaben sehr schwierig und umfangreich sind.»

Erneut wurden Ersetzungen für fehlende Werte mittels des EM-Algorithmus vorgenommen. Dabei wurden für das mathematische Selbstkonzept 1.0 %, für Leistungsangst 0.9 % und für Lernanstrengung 1.0 % der Daten mathematisch geschätzt.

### 8.2.3 Lehrerdaten

Es wurden die gleichen Daten wie in Studie 1 erhoben. Anstelle von Lernmotivation wurde nun nach der Lernanstrengung der Schüler gefragt. Das chinesische Notensystem von 0 bis 100 wurde verwendet, um den Lehrern differenzierte Aussagen über die Schüler ihrer Klasse zu ermöglichen. Auf diese Weise wurden von den Lehrkräften nicht nur die Erfolgserwartung und das Anspruchsniveau der Schüler, sondern auch die übrigen motivational-affektiven Schülermerkmale eingeschätzt.

## 9 Ergebnisse

### 9.1 Akkuratheit des Lehrerurteils

Die Testleistungen der chinesischen Schüler wurden im Mittel um  $M = 1.76$  Punkte ( $SD = 4.55$ ) überschätzt, wobei die Abweichungen von -13 bis + 13 Punkten schwankten. Die Niveauelemente der Testleistung beträgt  $M = 1.96$  ( $SD = 1.82$ ) und die Differenzierungskomponente der Testleistung hat einen Wert von  $M = 1.12$  ( $SD = .18$ ). Fünfundachtzig Schüler, 52 Jungen und 33 Mädchen, wurden um  $M = 3.46$  Punkte ( $SD = 2.53$ ) unterschätzt, wobei die Anzahl unterschätzter Schüler in den einzelnen Klassen zwischen 1 und 20 variierte. Bei 172 Schülern wurde die Testleistung um  $M = 4.50$  Punkte ( $SD = 3.02$ ) überschätzt. Für 15 Schüler wurde das Testergebnis korrekt vorhergesagt. Die fair beurteilten Schüler finden beim Vergleich unterschätzter und überschätzter Schüler keine Berücksichtigung. Zusätzlich wurde auch hier als ein strengeres Kriterium der Fehleinschätzung eine halbe Standardabweichung der Testleistung ( $M = 25.24$ ,  $SD = 5.60$ ) zugrunde gelegt.



Tabelle 3  
Mittelwertvergleiche der Schülermerkmale und Lehrerurteile der chinesischen Stichprobe

	Fehleingeschätzte Schüler				Deutlich fehleingeschätzte Schüler			
	Schülermerkmal		Lehrerurteil		Schülermerkmal		Lehrerurteil	
	unterschätzt	überschätzt	unterschätzt	überschätzt	unterschätzt	überschätzt	unterschätzt	überschätzt
Testleistung	25.95 (5.98)	24.78 (5.24)	22.49 (6.85)	29.28 (4.40)	24.77 (6.11)	23.44 (4.90)	19.51 (6.62)	29.17 (4.35)
Erfolgserwartung	93.19 (9.01)	95.94 (4.71)	82.37 (12.96)	89.88 (9.15)	92.11 (9.48)	95.41 (5.08)	78.88 (13.16)	89.25 (9.27)
Anspruchsniveau	95.95 (5.65)	97.65 (3.01)	86.69 (10.89)	92.84 (7.80)	94.88 (5.89)	97.41 (3.14)	83.86 (11.99)	92.48 (7.82)
Fähigkeitsselbstkonzept	2.98 (.68)	3.25 (.65)	74.95 (16.51)	83.17 (12.58)	2.82 (.69)	3.19 (.64)	71.51 (17.81)	82.50 (12.88)
Lernanstrengung	3.47 (.48)	3.57 (.44)	67.92 (18.58)	81.81 (15.00)	3.45 (.48)	3.55 (.44)	61.63 (19.33)	80.94 (15.67)
Leistungsangst	2.52 (.72)	2.19 (.80)	35.36 (18.42)	24.84 (19.40)	2.55 (.77)	2.23 (.79)	36.74 (17.96)	25.48 (19.73)

Dadurch umfasst die Gruppe der deutlich unterschätzten Schüler 43 Personen und die Gruppe der deutlich überschätzten Schüler 122 Personen.

Zur ersten Hypothese wurden Pearson-Korrelationen zwischen Lehrerurteilen und Schülermerkmalen ermittelt. Trotz der hohen Schüleranzahl in den Klassen die Lehrkräfte eine gute Vorhersage der Testleistung ( $r = .71, p < .001$ ) treffen können. Mittlere Korrelationen ergeben sich für die Zusammenhänge von Lehrer- und Schülerurteil bei den Variablen Erfolgserwartung ( $r = .55, p < .001$ ) und Anspruchsniveau ( $r = .39, p < .001$ ). Dagegen zeigen sich für das fachspezifische Selbstkonzept ( $r = .23, p < .001$ ), die Lernanstrengung ( $r = .24, p < .001$ ) und die Leistungsangst ( $r = .18, p < .01$ ) nur geringe Korrelationen. Wie in Hypothese 1 vermutet, wird die Testleistung von den Lehrkräften zutreffender eingeschätzt als die motivational-affektiven Schülermerkmale.

## 9.2 Vergleich der Schülergruppen

In der zweiten Hypothese wurde nach Unterschieden zwischen unterschätzten und überschätzten Schülern gefragt. Die deskriptiven Statistiken für fehleingeschätzte und deutlich fehleingeschätzte Schüler sind in Tabelle 3 enthalten. Alle multivariaten Vergleiche zwischen überschätzten und unterschätzten Schülern sowie zwischen Lehrerurteil und Schülermerkmal beziehen sich auf klassenweise z-transformierte Werte. Die Transformation der Messwerte führt zu keiner Veränderung der Ergebnisse, ist aber erforderlich, um Lehrer- und Schülerurteil für alle Variablen miteinander in Bezug setzen zu können.

Eine erste multivariate Varianzanalyse zum Vergleich unterschätzter und überschätzter Schüler wird signifikant (Hotelling's  $T^2 = .20, F(6, 249) = 8.13, p < .001, \eta^2 = .164$ ). Auch eine zweite multivariate Varianzanalyse, bei der deutlich unterschätzte und deutlich überschätzte Schüler einander gegenübergestellt werden, fällt signifikant aus (Hotelling's  $T^2 = .27, F(6, 157) = 6.97, p < .001, \eta^2 = .210$ ). Die Ergebnisse sind in beiden Fällen recht ähnlich, wie sich aus der oberen Hälfte der Tabelle 4 ablesen lässt. In der Testleistung zeigt sich erneut kein Unterschied zwi-

schen unterschätzten und überschätzten Schülern. Die unterschätzten Schüler sind im Schnitt sogar um gut einen Punkt besser als die überschätzten Schüler. Bei den deutlich unterschätzten Schülern ist der Punktabstand zu den deutlich überschätzten Schülern ähnlich und wird hier sogar signifikant. Das Verhältnis von unterschätzten und überschätzten Schülern kehrt sich bei den motivational-affektiven Variablen jedoch um. Unterschätzte Schüler zeigen eine geringere Erfolgserwartung, ein geringeres Fähigkeitsselbstkonzept und eine höhere Leistungsangst als überschätzte Schüler. Beim Anspruchsniveau und bei der Lernanstrengung unterscheiden sich die Schülergruppen dagegen nicht statistisch bedeutsam. Insgesamt gesehen wird die zweite Hypothese durch die Forschungsbefunde gestützt. Trotz gleicher Leistungsfähigkeit und gleicher Lernanstrengung haben die unterschätzten Schüler ein weniger positives Erleben in Bezug auf das Fach Mathematik als ihre überschätzten Klassenkameraden.

## 9.3 Beziehungen von Lehrerurteilen und Schülermerkmalen

Um die dritte Hypothese zu überprüfen, wurde ein Vergleich von Lehrer- und Schülerurteilen vorgenommen. Die Ergebnisse zum Interaktionseffekt befinden sich in der unteren Hälfte der Tabelle 4. Es wird offenbar, dass die Befunde für fehleingeschätzte und deutlich fehleingeschätzte Schüler wiederum in die gleiche Richtung gehen. Bei der Testleistung erbringt die Varianzanalyse mit Messwiederholung einen signifikanten Interaktionseffekt, der aus der Aufteilung in unterschätzte und überschätzte Schüler herrührt. Bei der Erfolgserwartung zeigt sich ebenfalls ein signifikanter Interaktionseffekt. Er macht deutlich, dass die chinesischen Lehrer den überschätzten Schülern wesentlich bessere Noten zutrauen als den unterschätzten Schülern. Der signifikante Interaktionseffekt beim Anspruchsniveau dokumentiert, dass die Lehrkräfte eher bei den überschätzten als bei den unterschätzten Schülern hohe Erfolgsansprüche vermuten. Beim mathematischen Selbstkonzept und der Leistungsangst ergeben sich keine signifikanten Interaktionseffekte. Unterschätzte Schüler beurtei-



Tabelle 4  
*Ergebnisse der Varianzanalysen für fehleingeschätzte und deutlich fehleingeschätzte Schüler der chinesischen Stichprobe*

Variable	Fehleingeschätzte Schüler		Deutlich fehleingeschätzte Schüler	
	$F(1, 255)$	$\eta^2$	$F(1, 163)$	$\eta^2$
	Haupteffekt unterschätzte vs. überschätzte Schüler			
Testleistung	3.63	.014	4.34*	.026
Erfolgserwartung	12.89***	.048	9.62**	.056
Anspruchsniveau	3.23	.013	3.62	.022
Fähigkeitsselbstkonzept	12.86***	.048	14.88***	.084
Lernanstrengung	3.88	.015	3.49	.021
Leistungsangst	15.53***	.058	7.85**	.046
	Interaktionseffekt Schülermerkmal vs. Lehrerurteil			
Testleistung	293.33***	.535	323.69***	.665
Erfolgserwartung	8.59**	.033	7.48**	.044
Anspruchsniveau	17.22***	.063	10.78**	.062
Fähigkeitsselbstkonzept	3.46	.013	2.49	.015
Lernanstrengung	12.20**	.046	9.63**	.056
Leistungsangst	1.43	.006	.31	.002

Anmerkungen. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

len sich auf diesen Variablen deutlich negativer als überschätzte Schüler und das Lehrerurteil dazu geht in die gleiche Richtung. Bei der Lernanstrengung hingegen besteht ein signifikanter Effekt. Die Lehrer beurteilen die Anstrengungsbereitschaft der unterschätzten Schüler negativer und die der überschätzten Schüler positiver, während die Schülergruppen selbst keinen Unterschied ihrer Lernanstrengung wahrnehmen.

## 10 Diskussion

Die chinesische Grundschuluntersuchung gleicht in ihren Ergebnissen in dreierlei Hinsicht der deutschen Grundschulstudie:

Erstens zeigt das Lehrerurteil erneut die höchste Genauigkeit in Bezug auf die Testleistung. Demgegenüber fallen die Rangkomponenten der motivational-affektiven Variablen geringer aus. Schülermerkmale mit einer direkten Verbindung zu Noten wie Erfolgserwartung, Anspruchsniveau und Fähigkeitsselbstkonzept können mit einer größeren Zuverlässigkeit richtig taxiert werden als Lernanstrengung und Leistungsangst, die nur einen indirekten Bezug zur Notengebung aufweisen.

Zweitens zeigen unterschätzte Schüler eine vergleichbare Testleistung, berichten über die gleiche Lernanstrengung und haben das gleiche Anspruchsniveau wie überschätzte Schüler. Allerdings haben unterschätzte Schüler eine niedrigere Erfolgserwartung, ein niedrigeres Fähigkeitsselbstkonzept und eine höhere Leistungsangst. Auch dieses Ergebnis stimmt im Wesentlichen mit der deutschen Untersuchung überein.

Drittens generalisiert das Lehrerurteil von der Testleis-

tung auf motivational-affektive Schülereigenschaften. Unterschätzte Schüler werden durch die Bank schlechter beurteilt als überschätzte Schüler. Teilweise konvergieren die negativeren Beurteilungen der Lehrkräfte mit den unvoreilhaftesten Selbstwahrnehmungen der unterschätzten Schüler. So zeigen unterschätzte Schüler der chinesischen Stichprobe ein niedrigeres Fähigkeitsselbstkonzept als überschätzte Schüler und Lehrkräfte nehmen dieses auch so wahr. Ein entsprechendes Ergebnis hatte sich auch für die deutsche Stichprobe gezeigt.

Im Vergleich der Schätzleistungen von deutschen und chinesischen Lehrern fällt auf, dass die chinesischen Lehrer etwas genauere Beurteilungen vornehmen und sich insgesamt über den Leistungsstand der Schüler besser informiert zeigen. Das mag damit zusammenhängen, dass in China jede Hausaufgabe vom Lehrer schriftlich korrigiert wird (Ludwig, 2005). Dadurch ist es dem Lehrer selbst bei großen Klassenstärken möglich, die Leistung des Einzelnen richtig zu taxieren. Dieses Ergebnis kann auch in Hinblick auf die Förderung der diagnostischen Kompetenz von Lehrkräften von Bedeutung sein.

Im Großen und Ganzen kann die zweite Studie die Befunde der ersten Studie untermauern. Die Ergebnisse beider Studien zeigen die gleiche Tendenz: Unterschätzte Schüler haben bei gleicher Lernleistung und gleicher Lernbereitschaft eine negativere Selbstwahrnehmung und weniger Vertrauen in die eigene Leistungsfähigkeit als überschätzte Schüler.

## 11 Abschließende Diskussion

Die Untersuchungen in Deutschland und China unterstreichen die Anfälligkeit des Lehrerurteils für Fehleinschät-

zungen. Besonders betroffen ist davon die in der Regel kleinere Gruppe von Schülern, die von den Lehrern in ihrer Leistungsfähigkeit unterschätzt wird. Unterschätzte Schüler haben kein Performanz- und auch kein Motivationsdefizit, aber ihnen fehlt Vertrauen in die eigene Leistungsfähigkeit und die Zuversicht, das Gleiche erreichen zu können, wie überschätzte Schüler. Bedauerlicherweise urteilen Lehrkräfte über unterschätzte Schüler ähnlich und attestieren ihnen in Bezug auf Motivation und Emotion ein weniger positives Erscheinungsbild als anderen Schülern im Klassenverband.

Die beiden Studien demonstrieren die Verlässlichkeit der erzielten Ergebnisse – auch über kulturelle Grenzen hinweg. Entscheidend dürfte dafür sein, dass in Deutschland wie in China dem Lehrerurteil ein großes Gewicht zukommt und sich Schüler an den Aussagen des Lehrers orientieren, um sich selbst zu beurteilen. Dabei lassen sich zwei unterschiedliche Effekte voneinander trennen (Madon et al., 2001): Zum einen können Lehrer im Sinne einer sich selbsterfüllenden Prophezeiung Einfluss auf die Motivation und das Selbstkonzept der Schüler ausüben. Zum anderen können Schüler im Sinne einer Selbstbewahrheitung sich an die Vorstellungen der Lehrer anpassen. Für beide Effekte lassen sich empirische Belege finden (Madon et al., 2001). Auf diese Weise würde erklärlich, warum unterschätzte Schüler über ihre Fähigkeiten, Emotionen und Erfolgsaussichten deutlich ungünstiger urteilen als überschätzte Schüler.

Aus der Forschung zum Pygmalion-Effekt ist weiterhin bekannt, dass Lehrererwartungen sich in unterschiedlichen Verhaltensweisen niederschlagen (Cooper, 1979; Rosenthal, 1974). Lehrer schaffen für Schüler, an die sie hohe Erwartungen richten, ein wärmeres, sozioemotionales Klima, unterhalten sich häufiger und länger mit ihnen, geben ihnen mehr Zeit zum Antworten, loben sie mehr und erklären ihnen ihre Fehler gründlicher (Cooper, 1979). Diese unterschiedlichen Verhaltensweisen könnten auch bei unterschätzten und überschätzten Schülern zum Tragen kommen und den Schülern Informationen darüber liefern, wie der Lehrer über sie denkt. Die Wirkungskette ist noch ungeklärt, doch kann vermutet werden, dass unterschätzte und überschätzte Schüler über das Interaktionsverhalten des Lehrers Kenntnis von seinem Urteil erlangen und sich dann im Sinne einer sich selbsterfüllenden Prophezeiung entwickeln (Jussim et al., 2009).

In Hinblick auf weitere Untersuchungen sollen einige Verbesserungen angesprochen werden, die zu einer größeren Validität der Ergebnisse beitragen könnten. Die Messung verschiedener Konstrukte wurde wie bei Spinath (2005) nur mit jeweils einem Item vorgenommen, um den Bearbeitungsaufwand für die Lehrkräfte in einem erträglichen Rahmen zu halten. Die Ergebnisse von Lorenz und Artelt (2009) zeigen zwar, dass eine Messung des Lehrerurteils durch mehrere Aussagen zu keiner wesentlichen Verbesserung der Schätzqualität führt, doch sollte generell eine Mehrfachmessung von Konstrukten

angestrebt werden. Ferner sind die Stichproben der Schüler zwar ausreichend groß, die Stichproben der Lehrkräfte aber relativ klein. Dadurch können weiterführende statistische Verfahren wie Mehrebenenanalysen nicht zur Anwendung gelangen. Cohen (1998) berichtet, dass für hinreichende Stabilitäten bei der Schätzung von Regressionskoeffizienten etwa vierzig Einheiten auf der zweiten Ebene, also vierzig Lehrkräfte, erforderlich sind. Ein weiteres Desiderat geht in Richtung einer Längsschnittstudie. Nur bei längsschnittlicher Betrachtung könnte ermittelt werden, ob die Lehrereinschätzungen einen Einfluss auf das Selbstbild der Schüler ausüben und als Ursache für die negativere Selbstwahrnehmung unterschätzter Schüler in Frage kommen. Ein negatives Lehrerurteil über die Schülerleistung kann zu einem niedrigen Fähigkeitsselbstkonzept beitragen, doch ist es ebenso möglich, dass ein niedriges Fähigkeitsselbstkonzept Lehrer erst zu einem negativen Urteil über den Schüler veranlasst. Über eine Längsschnittstudie könnte auch geklärt werden, ob mit der Fehldiagnose der Lehrkräfte Leistungsveränderungen der Schüler einhergehen. Die hier präsentierten querschnittlichen Befunde zeigen eine Leistungsäquivalenz der Schülergruppen, die aber angesichts der negativeren Selbstperzeption unterschätzter Schüler nicht erhalten bleiben muss. Des Weiteren sollten in Folgeuntersuchungen mehr Kontrollvariablen einbezogen werden, um etwas über die Hintergründe der Fehleinschätzung von Schülerleistungen zu erfahren. Es ist bekannt, dass Verhaltensauffälligkeiten in der Klasse zu einer schlechteren Beurteilung der schulischen Leistung führen (Bennett, Gottesman, Rock & Cerullo, 1993). Ebenso können der sozioökonomische Hintergrund (Alvidrez & Weinstein, 1999) oder physische Attraktivität und persönliche Ausstrahlung (Wang, Treat & Brownell, 2008) zu einer Beeinflussung des Lehrerurteils über Lernleistungen beitragen.

Es ist wichtig, dass alle Schüler die Möglichkeit haben, sich in der Schule nach ihren Fähigkeiten und Interessen zu entwickeln. Unterschätzte Schüler erwarten schlechtere Noten und trauen sich weniger zu, als es ihrem Potenzial entspricht. Daher ist es geboten, diesen Schülern mehr Aufmerksamkeit zu Teil werden zu lassen und sie motivational und emotional besser zu unterstützen. Den beteiligten Lehrkräften kommt in diesem Prozess eine Schlüsselrolle zu.

## Danksagung

Die vorliegende Studie wurde aus DFG-Mitteln im Rahmen des Projekts «Selbstkonzept, Lernmotivation, Leistungsangst und Kreativität des unterschätzten Schülers» (AZ: UR 170/3–1) gefördert. Die Autoren danken vier anonymen Gutachtern für die zahlreichen wertvollen Hinweise zur Verbesserung der Manuskriptqualität.

## Literatur

- Alvidrez, J. & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*, 731–746.
- 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.
- Bates, C. & Nettelbeck, T. (2001). Primary school teachers' judgments of reading achievement. *Educational Psychology, 21*, 177–187.
- 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.
- Bennett, R. E., Gottesman, R. L., Rock, D. A. & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85*, 347–356.
- Boehnke, K., Silbereisen, R. K., Reynolds, C. R. & Richmond, B. O. (1986). What I think and feel – German experience with the revised form of the children's manifest anxiety scale. *Personality and Individual Differences, 7*, 553–560.
- Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*, 123–142.
- Chen, C. & Stevenson, H. W. (1995). Motivation and mathematics achievement: A comparative study of Asian-American, Caucasian-American, and East-Asian High school students. *Child Development, 66*, 1215–1234.
- Cohen, M. P. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics, 14*, 267–275.
- Cooley, C. H. (1902). *Human nature and the social order*. New York: Charles Scribner's Sons.
- Cooper, H. M. (1979). Pygmalion grows up: A model for teacher expectation communication and performance influence. *Review of Educational Research, 49*, 389–410.
- Dalbert, C., Schneidewind, U. & Saalbach, A. (2007). Justice judgments concerning grading in school. *Contemporary Educational Psychology, 32*, 420–433.
- 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.
- Dweck, C. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040–1048.
- Faber, G. (2001). Das Verhalten rechtschreibängstlicher Grundschulkinder im Lehrerurteil: Empirische Untersuchungsergebnisse zur Problematik informeller Alltagsdiagnosen. *Heilpädagogische Forschung, 27*, 58–65.
- 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.
- 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 Test.
- Hamilton, C. & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32*, 228–240.
- Heckhausen, J. & Heckhausen, H. (2008). Motivation and development. In J. Heckhausen & H. Heckhausen (Eds.), *Motivation and action* (pp. 384–443). Cambridge: Cambridge University Press.
- Helmke, A. & Fend, H. (1982). Diagnostic sensitivity of teachers and parents with respect to the test anxiety of students. In R. Schwarzer, H. M. van der Ploeg & C. D. Spielberger (Eds.), *Advances in test anxiety research* (pp. 115–128). Hillsdale, NJ: Erlbaum.
- 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., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), *Schulleitung und Schulentwicklung* (S. 119–144). Hohengehren: Schneider.
- Hinnant, J. B., O'Brien, M. & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*, 662–670.
- Hoge, R. D. & Coladarci, T. (1989). Teacher-based judgment of academic achievement: A review of literature. *Review of Educational Research, 59*, 297–313.
- Ingenkamp, K. (Hrsg.). (1995). *Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte*. Weinheim: Beltz.
- Jussim, L. & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*, 131–155.
- Jussim, L., Robustelli, S. L. & Cain, T. R. (2009). Teacher expectations and self-fulfilling prophecies. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 349–380). New York: Routledge.
- Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Pädagogische Psychologie, 23*, 197–209.
- Krolak-Schwerdt, S., Böhmer, M. & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess. Der Lehrer als «flexibler Denker». *Zeitschrift für Pädagogische Psychologie, 23*, 175–186.
- 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.
- Ludwig, M. (2005). Schule und Mathematik in China. In H.-W. Henn (Hrsg.), *Beiträge zum Mathematikunterricht: Vorträge auf der 37. Tagung für Didaktik der Mathematik vom 3. bis 7. März 2003 in Dortmund* (S. 425–428). Hildesheim: Franzbecher.

- Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen. *Psychologische Rundschau*, 58, 103–117.
- Madon, S., Smith, A., Jussim, L., Russell, D. W., Eccles, J., Palumbo, P. & Walkiewicz, M. (2001). Am I as you see me or do you see me as I am? Self-fulfilling prophecies and self-verification. *Personality and Social Psychology Bulletin*, 27, 1214–1224.
- Madon, S. J., Jussim, L., Keiper, S., Eccles, J., Smith, A. & Palumbo, P. (1998). The accuracy and power of sex, social class and ethnic stereotypes: Naturalistic studies in person perception. *Personality and Social Psychology Bulletin*, 24, 1304–1318.
- Marsh, H. & Craven, R. (1991). Self-other agreement on multiple dimensions of preadolescent self-concept: Inferences by teachers, mothers, and fathers. *Journal of Educational Psychology*, 83, 393–404.
- Möller, J. & Trautwein, U. (2009). Selbstkonzept. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 179–203). Heidelberg: Springer-Verlag.
- Ophuysen, S. v. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahneempfehlung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 38, 154–161.
- PISA-Konsortium Deutschland. (Hrsg.). (2006). *PISA 2003. Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.
- Praetorius, A.-K., Greb, K., Dickhäuser, O. & Lipowsky, F. (in Druck). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften. *Psychologie in Erziehung und Unterricht*.
- Rosenthal, R. (1974). *On the social psychology of the self-fulfilling prophecy: Further evidence for Pygmalion effects and their mediating mechanisms*. New York: MSS Modular Publications.
- Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and student intellectual development*. New York: Holt, Rinehart & Winston.
- Rost, D. H. & Schermer, F. J. (2006). Leistungsängstlichkeit. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 404–416). Weinheim: BeltzPsychologie Verlags Union.
- 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.
- Schunk, D. H., Pintrich, P. R. & Meece, J. L. (2007). *Motivation in education. Theory, research, and applications*. Upper Saddle River, NJ: Pearson Education.
- Shavelson, R. J., Hubner, J. J. & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, 46, 407–441.
- Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, 19, 85–95.
- Südkamp, A. & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum: Direkte und indirekte Einschätzungen von Schülerleistungen. *Zeitschrift für Pädagogische Psychologie*, 23, 161–174.
- Südkamp, A., Kaiser, J. & Möller, J. (submitted). *Accuracy of teachers' judgments of students' academic achievement: A meta-analysis*.
- Südkamp, A., Möller, J. & Pohlmann, B. (2008). Der simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, 22, 261–276.
- 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.
- Urhahne, D. (2008). Sieben Arten der Lernmotivation. Ein Überblick über zentrale Forschungskonzepte. *Psychologische Rundschau*, 59, 150–166.
- Wang, S., Treat, T. & Brownell, K. (2008). Cognitive processing about classroom-relevant contexts: Teachers' attention to and utilization of girls' body size, ethnicity, attractiveness, and facial affect. *Journal of Educational Psychology*, 100, 473–489.
- 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, Germany: Ulm University, Department of Educational Psychology.

Prof. Dr. Detlef Urhahne

---

Ludwig-Maximilians-Universität München  
 Psychology of Excellence in Business and Education  
 Martiusstr. 4  
 D-80802 München  
 Tel. +49 89 2-180-9791  
 Fax +49 89 2-180-4814  
 E-Mail urhahne@lrz.uni-muenchen.de