

Diagnostik und Leistungsbeurteilung

Die Beurteilung von Schülerinnen und Schülern gehört zu den wichtigsten Aufgaben von Lehrerinnen und Lehrern. Diese Beurteilung ist Ergebnis eines diagnostischen Prozesses. Entsprechend liegen bei der Beurteilung im Unterricht ganz ähnliche Herausforderungen zugrunde wie in der Diagnostik. In diesem Kapitel werden Grundbegriffe der Diagnostik eingeführt und es wird ihre Relevanz für die Arbeit von Lehrerinnen und Lehrern deutlich gemacht. Zudem beschreiben wir den Prozess der Urteilsbildung sowie aktuelle Forschungsarbeiten, die sich mit der Genauigkeit von Urteilen und möglichen Verzerrungen bei der Beurteilung (Urteilsfehler) beschäftigen. Abschließend werden mit Noten, Berichtszeugnissen und Kompetenzrastern Möglichkeiten der Bewertung als Ergebnis eines diagnostischen Prozesses dargestellt.

1 Grundbegriffe der Diagnostik

Die Beurteilung von Schülerinnen und Schülern gehört zum Kerngeschäft von Lehrkräften. Hierzu sammeln sie Informationen aus Klassenarbeiten, dem Unterricht und weiteren relevanten Quellen, um zu einer Einschätzung der Leistung, der Motivation oder des Verhaltens von Schülerinnen und Schülern zu gelangen. Mit anderen Worten: Lehrkräfte betreiben Diagnostik. (Pädagogisch-psychologische) Diagnostik hat die Aufgabe, Informationen über Merkmalsträger zu beschaffen und zu bewerten, um eine möglichst akkurate Einschätzung von Merkmalen treffen zu können, sodass präzise Vorhersagen künftigen Verhaltens und Erlebens möglich werden. Merkmalsträger können Personen, Gruppen, Institutionen, Situationen und Gegenstände sein. Die Fähigkeit zu diagnostizieren wird in der Forschung zur professionellen Kompetenz von Lehrkräften (vgl. den ersten Beitrag in diesem Band) häufig als diagnostische Kompetenz beschrieben. Diese wird als die Fähigkeit definiert, Personen und Aufgaben zutreffend zu beurteilen. Nicht nur bei der Bewertung von Schülerleistungen, sondern auch bei der Gestaltung von Unterricht, etwa bei der Anpassung des Unterrichts an die Lernvoraussetzungen der Schülerinnen und Schüler, ist die diagnostische Kompetenz von zentraler Bedeutung.

1.1 Merkmale und ihre Operationalisierung

Eine wesentliche Herausforderung der Diagnostik besteht darin, dass die interessierenden Merkmale in der Regel nicht direkt beobachtbar sind. So gibt es zwar (wissenschaftlich fundierte) Vorstellungen davon, was z. B. unter Motivation, Intelligenz oder Mathematikkompetenz zu verstehen ist. Wie ausgeprägt solche Merkmale bei einer bestimmten Person sind, ist aber nicht offensichtlich. Solche nicht direkt beobachtbaren (in der Sprache der Statistik „latenten“) Merkmale nennen wir **Konstrukte**. Diese Konstrukte messbar zu machen ist eine der Kernfragen der Diagnostik. Hierzu (und häufig auch im Alltag) werden beobachtbare (in der Sprache der Statistik „manifeste“) Indikatoren herangezogen, die Rückschlüsse auf die Ausprägung eines Konstrukts erlauben. Will etwa eine Mathematiklehrkraft das Verständnis der Prozentrechnung messen, wird sie mithilfe konkreter Aufgaben in der Klassenarbeit versuchen, das Konstrukt „Verständnis der Prozentrechnung“ zu messen. Diese „Übersetzung“ eines Konstrukts in beobachtbare Indikatoren wird **Operationalisierung** genannt.

Lehrkräfte lassen aber nicht nur Klassenarbeiten schreiben, sie verwenden auch Verhaltensbeobachtungen (etwa bei Referaten oder der Beteiligung im Unterricht, Hausaufgaben) oder Interviews (z. B. mündliche Prüfungen). Zudem sammeln sie durch Vergleichsarbeiten wie etwa VERA-3 oder VERA-8 (vgl. Kapitel zu Schulleistungsuntersuchungen und Bildungsmonitoring, S. 189 ff.) Erfahrungen mit standardisierten Leistungstests. Sie diagnostizieren nicht nur Leistungen, sondern auch das Verhalten und die Motivation von Schülerinnen und Schülern. Verhaltensbeobachtungen, Interviews und standardisierte Tests sind neben Fragebögen übliche Verfahren in der Diagnostik. In der pädagogisch-psychologischen Diagnostik werden dabei vor allem Fragebögen und Tests eingesetzt.

Im Folgenden soll die Operationalisierung von Konstrukten am Beispiel eines Fragebogens zur Erfassung der intrinsischen Lesemotivation erklärt werden. In Abbildung 1 ist das wissenschaftliche Konstrukt und seine Operationalisierung mittels eines Fragebogens von Möller und Bonnerad (2007) dargestellt. Intrinsische Lesemotivation bedeutet, dass eine Person gern liest, weil sie das Lesen selbst als befriedigend empfindet (vgl. Kapitel zur Motivation, S. 121 ff.). Im linken Teil der Abbildung sieht man in der Ellipse dargestellt das Konstrukt, also die intrinsische Lesemotivation. In den Kästchen sind die beobachtbaren Verhaltensweisen, im Beispiel die Antworten in den Fragebogenitems, dargestellt. Die Richtung der Pfeile zeigt an, dass die Ausprägung des Konstrukts die Antworten in den Items mitbestimmt. Je höher die intrinsische Lesemotivation ausgeprägt ist, desto größer sollte die Wahrscheinlichkeit sein, in den Items 1 bis 4 (siehe Tabelle rechts in Abbildung 1) „trifft voll zu“ und in Item 5 „trifft nicht zu“ anzukreuzen. Item 5 stellt insofern eine Besonderheit dar, als es negativ formuliert ist, d. h. eine hohe Zustimmung zu diesem Item bedeutet eine geringe Ausprägung der Lesemotivation (daher mit „u“ wie umkodiert gekennzeichnet). Entsprechend der Antworten in allen fünf Items zieht man dann Rückschlüsse auf die Ausprägung

der intrinsischen Lesemotivation einer Person – hierzu werden die einzelnen Antworten zusammengefasst, sodass sich ein Gesamtwert für die intrinsische Lesemotivation ergibt. Ähnlich funktioniert im Prinzip auch eine Klassenarbeit. Hier ist es das Ziel, das Wissen oder die Fähigkeiten von Schülerinnen und Schülern in einem bestimmten Bereich zu messen (z. B. Essays in Englisch zu schreiben). Auch diese Fähigkeit ist nicht direkt beobachtbar. Die Aufgaben, aus denen sich Klassenarbeiten zusammensetzen, stellen eine Auswahl beobachtbarer Indikatoren dar, die dieses Fähigkeiten repräsentieren sollen, ähnlich wie die einzelnen Items im Fragebogen zur Lesemotivation in Abbildung 1.

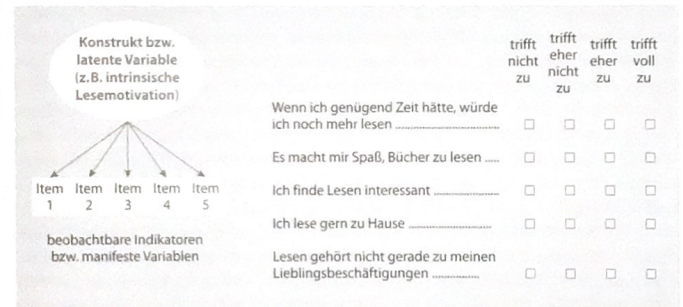


Abb. 1: Erfassung eines Konstrukts mittels Fragebogen am Beispiel intrinsischer Lesemotivation nach Möller und Bonnerad (2007)

Eine wichtige Frage bei der Konstruktion von Messverfahren oder der Suche nach Indikatoren ist die Frage, welches die „richtigen“ Indikatoren sind. Die Auswahl von Indikatoren bzw. die Konstruktion von Fragebogenitems ergibt sich aus der Definition eines Konstrukts. Diese wiederum sollte sich aus vorheriger Forschung oder theoretischen Überlegungen ableiten lassen. Die Definition eines Konstrukts bietet einen Rahmen, aus dem sich dessen zentrale Merkmale ergeben, die dann in Indikatoren erfasst werden. Dabei sollten Indikatoren spezifisch auf das jeweilige Konstrukt zugeschnitten sein und nicht zusätzlich anderes erfassen. In dem Beispiel zur intrinsischen Lesemotivation wurde aus der Definition des Konstrukts abgeleitet, dass Lesen für intrinsisch motivierte Personen etwas Freudvolles ist, das sich etwa in dem Item „Es macht mir Spaß, Bücher zu lesen“ widerspiegelt. Aus diesen Überlegungen folgt unter anderem, dass die Bedeutung eines Konstrukts sich abhängig von den verwendeten Items eines Fragebogens ändern kann. Zudem wird ein Konstrukt – gerade wenn es eher breit definiert ist – mit mehr Items besser abgedeckt. Ähnlich wird man bei einer Klassenarbeit in der Mathematik verschiedenartige Aufgaben wählen, die das Verständnis des Stoffs breit prüfen.

Auch wenn es nicht möglich ist, ein Konstrukt perfekt durch Indikatoren zu repräsentieren, ist es doch ein Ziel bei der Entwicklung diagnostischer Instrumente, bestimmte Fehler, die entstehen können zu minimieren. Zur Prüfung der Güte eines diagnostischen Instruments gibt es entsprechende Kriterien zur Beurteilung. Diese werden im folgenden Abschnitt vorgestellt.

1.2 Gütekriterien und Itemkennwerte

In der Diagnostik werden Haupt- und Nebengütekriterien zur Untersuchung der Qualität von Messinstrumenten unterschieden. Die Hauptgütekriterien sind die Objektivität, Reliabilität und Validität; zu den Nebengütekriterien zählen die Normierung, Ökonomie, Nützlichkeit, Zumutbarkeit, Unverfälschbarkeit und Testfairness. Neben diesen Kriterien, die sich meist auf einen gesamten Fragebogen beziehen, werden auch Itemkennwerte zur Beurteilung der Qualität einzelner Items herangezogen. Zentral sind hier die Itemschwierigkeit und die Trennschärfe. In diesem Abschnitt sollen nur die Hauptgütekriterien sowie Itemkennwerte vorgestellt werden.

Objektivität. Ein Messinstrument ist dann objektiv, wenn das Ergebnis einer Messung nur von dem zu messenden Merkmal abhängt und nicht von der messenden Person. In unserem Beispiel mit der Lesemotivation sollte das Ergebnis im Fragebogen also nur davon abhängen wie hoch die Lesemotivation einer Person ist und nicht davon, ob eine Lehrkraft, eine studentische Mitarbeiterin bzw. ein studentischer Mitarbeiter oder eine andere Person den Fragebogen verteilt und auswertet. Genauso sollten Ergebnisse eines Leistungstests – etwa bei Vergleichsarbeiten – nicht davon abhängen, wer die Testungen durchführt und auswertet. Lediglich das Wissen bzw. die Kompetenz der Schülerinnen und Schüler sollten das Testergebnis bestimmen. Unterschieden werden drei Formen der Objektivität. Erstens bedeutet Durchführungsobjektivität, dass gleiche Bedingungen bei der Bearbeitung eines Messinstruments vorliegen (z. B. einheitliche Instruktionen, gleiche Bearbeitungszeit). Zweitens bedeutet Auswertungsobjektivität, dass eindeutige Regeln zur Quantifizierung der Antworten vorliegen, dass also die Antworten unabhängig davon, wer sie auswertet, in die gleichen Zahlenwerte übersetzt werden. Drittens bedeutet Interpretationsobjektivität, dass verschiedene Personen dieselben Schlüsse aus den Ergebnissen ziehen (z. B. durch eindeutige Interpretationshilfen oder klare Kriterien). Objektivität ist eine Voraussetzung für Reliabilität und Validität. Sie ist bei Lehrerurteilen gefährdet, wie z. B. die Forschungen zum *Big-Fish-Little-Pond-Effekt* (Marsh 1987) zeigen (vgl. Kapitel zur Motivation, S. 121 ff.), der besagt, dass Lehrkräfte oft einen klasseninternen Maßstab verwenden und Lernende folglich schlechtere Noten in einer leistungsstarken Klasse bekommen als wenn sie mit identischer Leistung in einer leistungsschwachen Klasse benotet würden.

Reliabilität. Die Reliabilität ist die Zuverlässigkeit oder Genauigkeit eines Messinstruments. Keine Messung in der psychologischen Diagnostik ist perfekt, da jede Messung mit einem Messfehler behaftet ist oder überspitzt gesagt: „Wer misst, misst Mist.“ Formal schlägt sich dies in einem Axiom (Setzung oder Definition) der

klassischen Testtheorie nieder: Jeder gemessene Wert setzt sich aus einem wahren Wert und einem Messfehler zusammen. Formal ist dieses Axiom in folgender Formel dargestellt.

$$x_{jt} = w_j + e_{jt} \quad (1)$$

In Formel 1 ist x der beobachtete oder gemessene Wert eines Merkmalsträgers j in einem Messinstrument t . Dieser setzt sich additiv zusammen aus dem wahren Wert w eines Merkmalsträgers j in einem Messinstrument t und einem Messfehler e . Dieser Messfehler überdeckt also in gewisser Weise den wahren Wert und führt somit zu Abweichungen des Messergebnisses von diesem wahren Wert. Zustande kommt dieser Messfehler zum einen durch Einflüsse während einer Messung, die sich unkontrolliert auf die Messung auswirken und deren Ergebnis beeinflussen. Zum anderen durch die nicht perfekte Passung zwischen Indikatoren (Aufgaben, Testitems) und Konstrukten (Verständnis eines Stoffes, Lesemotivation). Ein Ziel psychologischer Diagnostik ist es nun, diesen Fehler möglichst klein zu halten. Die Reliabilität erlaubt hierbei Angaben darüber, wie gut das gelingt bzw. welcher Anteil einer Messung auf Messfehler zurückgeht. Ein Messinstrument gilt folglich dann als reliabel, wenn es das, was es messen soll, möglichst fehlerfrei misst. Insbesondere die Wiederholungsreliabilität ist bei schulischen Leistungsbeurteilungen oft recht hoch, die Schülerinnen und Schüler bekommen über die Zeit meist recht ähnliche Noten in den einzelnen Fächern.

Validität. Die Validität ist die Gültigkeit eines Tests. Sie gilt als das wichtigste Gütekriterium und befasst sich mit der Frage, ob ein Messinstrument das, was es vorgibt zu messen, auch tatsächlich misst. In unserem Beispiel stellt sich hier also die Frage, ob der Fragebogen zur intrinsischen Lesemotivation tatsächlich intrinsische Lesemotivation misst und nicht etwa allgemeine Schulfreude. Auch die Frage, inwieweit bei der Bearbeitung von Klassenarbeiten in einem bestimmten Fach fachfremde Fähigkeiten einfließen, ist eine Frage der Validität. Ist die korrekte Bearbeitung einer Aufgabe in einer Klassenarbeit zur Physik sehr vom Leseverständnis der Schülerinnen und Schüler abhängig, weil die Aufgaben zu kompliziert formuliert sind, ist die Validität der Klassenarbeit eingeschränkt. Ein solches Ergebnis fanden Härtig, Heitmann und Retelsdorf (2015) für die Bearbeitung von Schulleistungstests zur Messung der Physikkompetenz – hier zeigten sich deutliche Zusammenhänge mit sprachlichen Fähigkeiten.

Bei der Validität werden verschiedene Aspekte unterschieden. Die Inhaltsvalidität meint eine inhaltliche Analyse eines Messverfahrens hinsichtlich der Frage, ob die Inhalte der Aufgaben oder Items eines Messinstruments das theoretische Konstrukt in seiner Breite abdecken. Diese Analyse erfolgt per Augenschein oder durch Expertenbefragungen. In der pädagogisch-psychologischen Diagnostik geht es hierbei häufig um die curriculare Validität als Spezialfall, also die Frage, ob Testaufgaben in einem Leistungstest dem jeweiligen Curriculum der Zielgruppe entsprechen (z. B. ob die Aufgaben eines Mathematiktests für die

zweite Klasse tatsächlich Inhalte umfassen, die in der zweiten Klasse unterrichtet werden). Die Kriteriumsvalidität ist gegeben, wenn das Ergebnis eines Messinstruments mit einem vorgegeben Kriterium (einem anderen Messinstrument, einem objektiven Maß) zusammenhängt, für das entsprechende Zusammenhänge theoretisch zu erwarten sind. So wäre es etwa plausibel zu erwarten, dass Personen mit mathematischer Kompetenz, gemessen mit einem Leistungstest, auch bessere Noten in Mathematik erzielen. Dies kann man dann empirisch prüfen – ein Hinweis auf die hohe Kriteriumsvalidität des Tests wäre gefunden. Ein Spezialfall dieser Kriteriumsvalidität ist die prognostische Validität, bei der es um die Frage geht, ob ein Messinstrument eine Vorhersage eines Merkmals zu einem späteren Zeitpunkt erlaubt. Dies ist vor allem wichtig bei Selektionsentscheidungen wie z. B. der Schullaufbahnempfehlung. Die Empfehlung für eine bestimmte Schulform ist dann prognostisch valide, wenn die Kinder an den weiterführenden Schulen tatsächlich erfolgreich abschneiden.

Itemkennwerte. Itemkennwerte geben Auskunft über die Güte einzelner Items eines Messverfahrens. Die Itemschwierigkeit bezieht sich hierbei auf Lösungs- bzw. Zustimmungsraten. Bei einem Test mit eindeutig richtigen oder falschen Antworten entspricht die Lösungsrate einfach dem prozentualen Anteil der korrekten Antworten. Lösen also von 100 Personen 50 Personen eine Aufgabe korrekt, beträgt die Itemschwierigkeit 0,5 (oder 50 %). Itemschwierigkeiten können Werte zwischen 0 und 1 erreichen. Wünschenswert sind Werte zwischen 0,2 und 0,8, wobei gleichzeitig ein möglichst großer Anteil der Items Schwierigkeiten um 0,5 erreichen sollte. Man sollte also immer auch schwierige und einfache Aufgaben und **viele mittelschwierige Aufgaben in einem Leistungstest** anbieten.

Neben der Itemschwierigkeit gibt die Trennschärfe Auskunft über die Güte von Items. Die Trennschärfe gibt an, wie gut ein einzelnes Item das Gesamtergebnis eines Messinstruments repräsentiert. Die Trennschärfe ist definiert als die Korrelation der Beantwortung eines Items mit dem Gesamtwert in einem Messinstrument. Trennschärfen können zwischen -1 und 1 liegen und sollten > 0,3 sein. Solche Werte wären in einem Leistungstest ein Hinweis darauf, dass die einzelne Aufgabe (das einzelne Item) gut zum Gesamtergebnis passt.

2 Leistungsbeurteilung in der Schule

Wie eingangs erwähnt, gehört die Beurteilung von Schülerinnen und Schülern zum Kerngeschäft von Lehrkräften. Dabei geht es nicht nur um die Benotung von Schülerinnen und Schülern in Klassenarbeiten oder Zeugnissen, sondern Leistungsbeurteilungen spielen auch eine Rolle bei der Entscheidung für oder gegen Unterrichtsmethoden oder -inhalte. So müssen Lehrkräfte ihren Unterricht jeweils angepasst an die Voraussetzungen ihrer Schülerinnen und Schüler gestalten. Entsprechend zeigt sich die Relevanz der Genauigkeit von Lehrerurteilen auch in ihrem Zusammenhang mit der Leistungsentwicklung der Schülerinnen

und Schüler (Anders et al. 2010). Im Folgenden sollen Prozesse der (alltäglichen) Urteilsbildung, die Genauigkeit von Lehrerurteilen und mögliche Quellen für Verzerrungen von Urteilen dargestellt werden. Zunächst soll aber erst einmal dargestellt werden, wie sich die Urteilsgenauigkeit in das Modell professioneller Kompetenz, wie es im ersten Kapitel (S. 9ff.) dargestellt ist, integrieren lässt.

2.1 Leistungsbeurteilung als Bestandteil professioneller Kompetenz

Die Urteilsgenauigkeit bezüglich Personen- und Aufgabenmerkmalen beschreibt den Kern diagnostischer Kompetenz (Schrader 2010). Nicht nur bei der Bewertung von Schülerleistungen, sondern auch bei der Gestaltung von Unterricht, etwa bei der **Anpassung des Unterrichts an die Lernvoraussetzungen der Schülerinnen und Schüler**, ist die diagnostische Kompetenz von zentraler Bedeutung. Diagnostische Kompetenz wird von Weinert, Schrader und Helmke (1990) neben dem Wissen über Klassenführung, dem didaktischen Wissen und dem fachwissenschaftlichen Wissen zu den vier Wissensbereichen gezählt, die die kognitive Komponente der Lehrkraftexpertise ausmachen. Neben der Kenntnis diagnostischer Grundlagen zählen zum diagnostischen Wissen zum einen das generelle Wissen über Lernende in einem gewissen Altersbereich und deren Fähigkeiten und Lernschwierigkeiten, zum anderen Wissen über bestimmte Stärken und Schwächen einer Klasse insgesamt und auch einzelner Schülerinnen und Schüler. Auch Baumert und Kunter (2006) greifen die diagnostische Kompetenz als Bestandteil professioneller Handlungskompetenz von Lehrkräften auf. Grundsätzlich geht man davon aus, dass Kompetenz veränderbar und damit prinzipiell erlernbar ist. Die bisherige Forschung zur diagnostischen Kompetenz von Lehrkräften beschäftigte sich vor allem mit der Genauigkeit diagnostischer Urteile (s. Abschnitt 3). In Studien zur Urteilsgenauigkeit wurden Lehrerurteile bezüglich verschiedener Schülermerkmale (vor allem Leistung, aber auch Motivation, Intelligenz und Leistungsängstlichkeit) und die tatsächlichen Merkmalsausprägungen, die anhand standardisierter Tests erhoben wurden, miteinander verglichen. Bevor wir noch einmal detaillierter auf Befunde zur Urteilsgenauigkeit eingehen, wird im folgenden Abschnitt zunächst der Prozess der Urteilsbildung beschrieben.

2.2 Urteilsbildung

Wir können Urteilen definieren als „psychologischen Prozess, der zugrunde liegt, wenn Menschen einem Urteilsobjekt einen Wert auf einer Urteilsdimension zuordnen und das daraus resultierende Urteil explizit zum Ausdruck bringen“ (Betsch, Funke & Plessner 2011, S. 12). In dieser recht allgemeinen Definition kann das Urteilsobjekt alles Mögliche sein, etwa Personen, Objekte, Aussagen, Situationen, abstrakte Konstrukte usw. Im schulischen Kontext betrifft das Urteil meist Schülerinnen und Schüler bzw. bestimmte Merkmale derselben, etwa ihre Leistung in einem bestimmten Schulfach. Diesem Urteilsobjekt wird dann ein bestimmter Wert, in der Schule oft in Form einer Note, zugeordnet, wobei dieser Wert im Alltag auch in der natürlichen Sprache der Urteilernden ausge-

drückt werden kann. Zum Beispiel wird eine Person, der hohe Intelligenz unterstellt wird, im Alltag wie folgt beschrieben „Sie ist schlau“. In der professionellen Sprache würde hier ein konkreter Skalenwert auf der IQ-Skala verwendet werden: „Sie hat einen IQ von 128“. Schließlich meint explizit in der Definition nur, dass sich die urteilende Person bewusst ist, dass sie ein Urteil trifft. Der Prozess, der zum Urteil führt, muss nicht zwangsläufig explizit oder bewusst ablaufen.

Um das Zustandekommen von Urteilen zu verstehen, bedarf es Kenntnissen über die psychologischen Grundlagen der Informationsverarbeitung. An dieser Stelle sei dies aber nur sehr oberflächlich angerissen (s. z. B. Betsch, Funke & Plessner 2011). In Abbildung 2 ist ein vereinfachtes Modell der Urteilsbildung dargestellt, in dem deutlich wird, dass Urteile auf zwei Wegen oder in zwei Prozessen entstehen können, die sich in der Art der Informationsverarbeitung unterscheiden. Wie in der Abbildung dargestellt, werden Merkmale entweder heuristisch (automatisch/heuristisch/intuitiv) oder (systematisch/kontrolliert/systematisch/reflexiv) verarbeitet. Die Prozesse unterscheiden sich in der Verarbeitungstiefe der Informationen. Bei der heuristischen Verarbeitung werden Informationen unbewusst, nicht beabsichtigt und unkontrollierbar verarbeitet. Dieser Prozess benötigt nur wenige Kapazitäten. Bei der systematischen Verarbeitung handelt es sich um eine umfassende Analyse aller zur Verfügung stehenden Informationen. Dies erfolgt bewusst und kontrolliert und beansprucht ein hohes Maß an Anstrengung. Urteilende können bei der Urteilsbildung stets beide Pfade verfolgen. Welcher Prozess dabei im Vordergrund steht, hängt ab von der Motivation der urteilenden Person, der Relevanz des Themas usw.

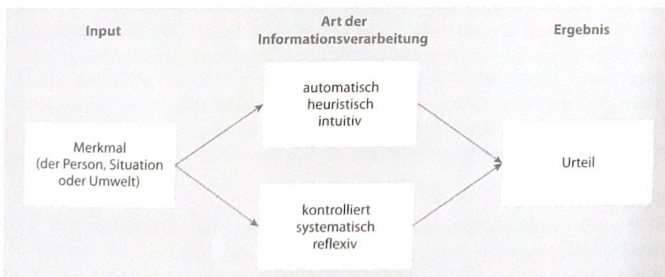


Abb. 2: Zwei-Prozess-Modell der Urteilsbildung

2.3 Urteilsgenauigkeit

Wie genau Urteile von Lehrkräften über ihre Schülerinnen und Schüler sind, hängt demnach auch von der Art der Informationsverarbeitung ab. In der Forschung zur Genauigkeit der Beurteilung von Schülerinnen und Schülern durch

ihre Lehrkräfte stehen diese Prozesse aber nicht im Vordergrund. Hier beschäftigt sich ein Großteil der Forschung zunächst einmal mit der Frage, wie akkurat Lehrerurteile überhaupt sind und wodurch Lehrerurteile beeinflusst werden.

Südkamp, Kaiser und Möller (2012) haben ein heuristisches Modell aufgestellt, anhand dessen die Forschung zu Faktoren, die die Urteilsgenauigkeit beeinflussen, systematisiert werden kann. In dem Modell ist Urteilsgenauigkeit beschrieben als Übereinstimmung von Schülerleistung und Lehrerurteil (siehe Abbildung 3). Beeinflusst wird das Lehrerurteil von den Merkmalen der Lehrkraft (z. B. Berufserfahrung, fachdidaktisches Wissen) und den Merkmalen des abzugebenden Urteils (z. B. Art des Zeugnisses, Anzahl der Notenstufen). Die Schülerleistung wiederum wird beeinflusst von Schülermerkmalen (z. B. Motivation, Vorwissen) und Testmerkmalen (z. B. Länge der Klassenarbeit, Fach). Alle vier Merkmalsbereiche (Lehrer-, Schüler-, Urteils- und Testmerkmale) sowie die Übereinstimmung zwischen Lehrer- und Schülermerkmalen (z. B. Geschlecht, Ähnlichkeit der Persönlichkeit) und die Übereinstimmung zwischen Urteils- und Testmerkmalen (z. B. zwischen Testdomäne und Leistungsurteil) wirken auf die Urteilsgenauigkeit.

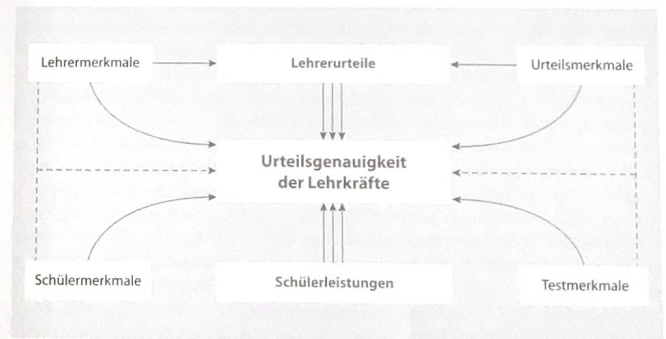


Abb. 3: Heuristisches Modell zu Moderatoren der Urteilsgenauigkeit (nach Südkamp, Kaiser & Möller 2012)

Zusammengefasst gelingt es Lehrkräften recht gut, die Leistung von Schülerinnen und Schülern zu beurteilen, wenngleich es bedeutende Unterschiede zwischen Lehrkräften gibt (Südkamp, Kaiser & Möller 2012, s. Infokasten S. 32). Solche Unterschiede in der Urteilsgenauigkeit können als Ergebnis der Berücksichtigung einer Kombination aus urteilsrelevanten und nicht urteilsrelevanten, ablenkenden Informationen erklärt werden. Sollen beispielsweise Schülerleistungen akkurat eingeschätzt werden, können im Schuljahr gezeigte mündliche und schriftliche Leistungen als urteilsrelevant gelten. Andere Informationen soll-

ten nicht in ein reines Leistungsurteil (im Gegensatz zu einem pädagogisch motivierten, umfangreicheren Urteil) einfließen. Lehrermerkmale, die ein Urteil beeinflussen können, wurden bisher selten untersucht. Diese Lücke in der Forschung scheint erstaunlich, handelt es sich doch um die zentralen Akteure im Beurteilungsprozess. Allerdings scheinen eine hohe inhaltliche Expertise (Fachwissen, fachdidaktisches Wissen und pädagogisch-psychologisches Wissen) sowie hohe kognitive Grundfähigkeiten der Lehrkräfte förderlich zu sein (Kaiser et al. 2012; Leucht et al. 2012). Unterschiede in der Urteilsgenauigkeit können zudem durch Urteilsverzerrungen zustande kommen, wie sie im folgenden Abschnitt ausführlicher besprochen werden.

Info

Meta-Analyse zur Urteilsgenauigkeit von Lehrerinnen und Lehrern

Meta-Analysen werden Studien genannt, die Ergebnisse aus sehr vielen vorliegenden Studien zum gleichen Thema nutzen. Für Meta-Analysen werden also keine neuen Daten erhoben, sondern vorhandene Studienergebnisse zusammengefasst. Beispielsweise existieren viele Studien zum Zusammenhang zwischen Lehrerurteilen zu Schülerleistungen und Schülerleistungen, die mit objektiven Testverfahren gemessen werden. Damit will man die Genauigkeit von Lehrerurteilen messen. Bedingt durch unterschiedliche Messmethoden, Stichproben und andere Faktoren ergeben sich in den unterschiedlichen Studien unterschiedliche Ergebnisse. Wenn man wissen will, wie der Zusammenhang insgesamt aussieht, analysiert man alle Daten, die Forscher weltweit zu dieser Fragestellung zusammengetragen haben. Solche Meta-Analysen nutzen dann beispielsweise die Korrelationen zwischen der Schülerleistung im Test und der Lehrereinschätzung zur Leistung des Schülers aus vielen Studien und verrechnen diese, um – vereinfacht ausgedrückt – den „mittleren“ Zusammenhang zu ermitteln. Diese Art „Zweitverwertung“ von Daten dient dem Überblick über ein Forschungsfeld.

Südkamp et al. (2012) fassen den Forschungsstand zur Urteilsgenauigkeit von Lehrkräften in einer Metaanalyse zusammen. Mit einer mittleren Korrelation von 0,63 finden sie in den 75 berücksichtigten Feldstudien mit mehr als 30.000 Lehrerurteilen eine mittelhohe Urteilsgenauigkeit der Lehrerbewertungen. Dabei berichten sie große Schwankungen in der Genauigkeit der Urteile der Lehrkräfte.

Für die Beurteilung der Schülerleistung kann somit im Durchschnitt von einer zufriedenstellenden Urteilsgenauigkeit der Lehrkräfte gesprochen werden. Dabei liegen in den Beurteilungen große Unterschiede vor, die nicht durch unterschiedliche Schülerleistungen zustande kommen. Einigen Lehrkräften gelingt die akkurate Beurteilung sehr gut, bei anderen Lehrkräften gibt es quasi keinen Zusammenhang zwischen ihrer Beurteilung und der tatsächlich gezeigten Schülerleistung. Für andere Variablen als die Schulleistung wie Motivation, Intelligenz oder Kreativität ist die Urteilsgenauigkeit geringer als für die Schulleistungen.

2.4 Urteilsverzerrungen

Abhängig von den oben genannten Prozessen der Informationsverarbeitung, kann es zu Urteilsverzerrungen kommen. So ist der systematische Weg der Informationsverarbeitung vor allem anfällig für fehlerhafte Beobachtungen oder fehlerhafte Erinnerungen (also letztlich eine Beurteilung auf Basis fehlerhafter Informationen). Die übrigen Urteilsverzerrungen sind vor allem relevant, wenn Urteile eher automatisiert oder heuristisch zustande kommen. Dazu gehören Lehrererwartungen und Stereotype, Halo-Effekte, Ankereffekte, stabile Urteilstendenzen und Sympathie-Effekte. Bezüglich der Sympathie-Effekte zeigte sich etwa, dass beliebte Schülerinnen und Schüler im Vergleich zu Leistungstestergebnissen eher zu positiv benotet werden, während unbeliebte Schülerinnen und Schüler eher zu negativ benotet werden (Hadley 1989). Stabile Urteilstendenzen zeigen Lehrkräfte, die eine Tendenz zur Milde oder Strenge aufweisen, also im Zweifel eher zu besseren oder schlechteren Noten kommen. Ankereffekte hingegen meinen eine Verzerrung eines Urteils in Richtung eines zuvor gesetzten Ankers, d. h. Lehrkräfte urteilen auf Basis vorangegangener Leistungen und sind nicht wirklich offen für aktuelle unerwartete Leistungen. Halo-Effekte schließlich beschreiben das Phänomen, dass eigentlich unabhängige oder nur gering korrelierte Personenmerkmale als zusammenhängend wahrgenommen werden (Nisbett & Wilson 1977; Thorndike 1920). Mit andere Worten: Lehrkräfte können bei der Beurteilung von Schülerinnen und Schülern unzulässige Generalisierungen vornehmen, indem sie ein besonders auffälliges Merkmal oder den Gesamteindruck auf das Urteil über ein anderes Merkmal übertragen. So können die physische Attraktivität (Ritts, Patterson & Tubbs 1992) oder die Intelligenz (Schrader & Helmke 1990) einer Schülerin oder eines Schülers ein Urteil beeinflussen. Carr und Kurtz-Costes (1994) drücken dies in ihrer Vermutung aus, dass Lehrkräfte eine holistische Sicht auf Schülerinnen und Schüler haben und sich ihr Gesamteindruck in jeglichem Einzelurteil über die Schülerin oder den Schüler niederschlägt. Nehmen Lehrkräfte beispielsweise von einer Schülerin an, sie sei hochbegabt, wird ihre Leistung besser beurteilt als die gleiche Leistung einer Schülerin, für die diese Vermutung nicht vorliegt. Auch motivationale Merkmale können Urteile von Lehrkräften beeinflussen: Unter Kontrolle der Leistung führten höhere Selbstkonzepte schulischer Fähigkeiten der Schülerinnen und Schüler zu positiveren Leistungsurteilen (Schrader & Helmke 1990). In ähnlicher Art und Weise zeigten Kaiser et al. (2013), dass Urteile von Lehrkräften über die Motivation und Leistung von Schülerinnen und Schülern deutlich höher zusammenhängen, als Motivation und Leistung selbst. Verzerrungen aufgrund von Lehrererwartungen und Stereotypen sollen im Folgenden etwas ausführlicher dargestellt werden.

Lehrererwartungen und Stereotype. Effekte von Lehrererwartungen waren und sind seit der klassischen Studie „Pygmalion im Klassenzimmer“ von Rosenthal und Jacobsen (1968) immer wieder Gegenstand empirischer Forschung (im Überblick z. B. Jussim & Harber 2005; Ludwig 2010). Im Kern geht es dabei

um die Frage, ob Lehrkräfte die Entwicklung ihrer Schülerinnen und Schüler entsprechend eigener Erwartungen über deren Entwicklungspotenzial beeinflussen, sich also die Schülerinnen und Schüler nach ihrem eigenen Bild erschaffen (daher Pygmalion als Verweis auf eine Figur der griechischen Mythologie). Erwartungseffekte funktionieren dabei im Sinne einer selbsterfüllenden Prophezeiung (Merton 1948), indem Lehrkräfte sich bewusst oder unbewusst entsprechend ihrer Erwartungen gegenüber den Schülerinnen und Schülern verhalten und somit deren Reaktionen beeinflussen. In der klassischen Studie von Rosenthal und Jacobson (1968) zeigte sich, dass sich falsche positive Erwartungen über das Leistungspotenzial individueller Schülerinnen und Schüler positiv auf die tatsächliche Entwicklung des IQs dieser Schülerinnen und Schüler auswirken. Dabei wurde Lehrkräften bezüglich bestimmter Schülerinnen und Schüler gesagt, dass diese ein besonders positives Entwicklungspotenzial hätten. Tatsächlich zeigten zufällig ausgewählte Schülerinnen und Schüler dann später bessere Leistungen. Die Originalstudie wurde aufgrund verschiedener methodischer Aspekte immer wieder massiv kritisiert und kontrovers diskutiert (für eine Zusammenfassung s. Jussim & Harber 2005). In der Folge dieser klassischen Arbeit sind zahlreiche weitere Forschungsarbeiten entstanden, die insgesamt den Schluss zulassen, dass Erwartungseffekte real und wirksam sind. Dabei zeigte sich, dass neben den ursprünglich berichteten positiven Effekten zu hoher Erwartungen auch schädliche Auswirkungen auf die Leistung von Schülerinnen und Schülern entstehen, denen Lehrkräfte nur ein geringes Potenzial unterstellen. In manchen Untersuchungen waren die schädlichen Effekte negativer Erwartungen größer als die förderlichen Effekte positiver Erwartungen (z. B. Alvidrez & Weinstein 1999). Zudem sind von solchen schädigenden Erwartungseffekten vor allem Schülerinnen und Schüler bestimmter Gruppen betroffen. So konnten McKown und Weinstein (2002) etwa zeigen, dass Ethnizität und Geschlecht der Schülerinnen und Schüler eine Rolle spielen. Akademisch stigmatisierte Gruppen (z. B. Mädchen in Zusammenhang mit Mathematik, Afroamerikaner) leiden mit einer höheren Wahrscheinlichkeit an Unterschätzungen durch Lehrkräfte als sie von Überschätzungen profitieren. Ähnlich zeigten sich in Untersuchungen Einflüsse der sozialen Herkunft auf die Leistungsbeurteilung, unabhängig von der tatsächlichen Leistung. So berichten Ready und Wright (2011) einen Zusammenhang des familiären Hintergrunds mit der Leistungsbeurteilung: Lehrkräfte unterschätzten die Leistung der Schülerinnen und Schüler, die aus bildungsferneren Schichten stammten. Hinnant, O'Brien und Ghazarian (2009) fanden eine Überbewertung der Leistungen der Lesefähigkeit der Mädchen durch die Lehrkräfte und eine Unterschätzung der Leistungen der Jungen in der ersten Klasse. Als Konsequenz dieser Urteilsverzerrungen zeigten sich zudem vor allem für Jungen aus Minoritäten Effekte auf die Leistungsentwicklung: Bei Unterschätzung zu Beginn ergaben sich die größten Leistungseinbußen und bei Überschätzung die größten -gewinne. Erklärt werden könnten solche gruppenbezogenen Unterschiede durch Überzeugungen von Lehrkräften wie z. B. ihren Stereotypen

über unterschiedliche Leistungspotenziale von Jungen und Mädchen in einem bestimmten Fach. Dabei muss für die Richtung der Stereotypisierung im schulischen Bereich stets das jeweilige Schulfach berücksichtigt werden. So werden in der Regel Jungen im verbalen Bereich und Mädchen im mathematischen Bereich negativ stereotypisiert (z. B. Plante et al. 2013). Für solche Geschlechterstereotype konnten Retelsdorf, Schwartz und Asbrock (2015) zeigen, dass sich Geschlechterstereotype über das Lesen bei Kindern (wenn Lehrkräfte glauben, dass Mädchen Jungen überlegen sind im Lesen) negativ auf das verbale Selbstkonzept der Jungen auswirken.

Abschließend sei zu diesem Abschnitt angemerkt, dass Erwartungen und Stereotype nicht „abnorm“ oder verwerflich sind. Vielmehr hat jeder Mensch solche Überzeugungen, die sich aus früheren Erfahrungen speisen und uns dabei helfen, die komplexe Realität zu vereinfachen und damit überhaupt die Menge an Informationen, die auf uns einströmen, zu verarbeiten. Problematisch sind Erwartungen und Stereotype vor allem dann, wenn sie starr und unangemessen sind und wenn, im schulischen Kontext, Lehrkräfte diesen Überzeugungen unreflektiert folgen, ohne die tatsächlichen Leistungen und Fähigkeiten ihrer Schülerinnen und Schüler ausreichend zu berücksichtigen. Ein mögliches Mittel für Lehrkräfte, um ihre Urteile möglichst unverzerrt zu fällen, besteht darin, möglichst viele objektive Informationen in ein Urteil einfließen zu lassen und somit ein professionelles diagnostisches Urteil zu fällen.

2.5 Noten, Berichtszeugnisse und Kompetenzraster

Die Erstellung von Zeugnissen gehört zu den zentralen Aufträgen der Schule. Meist werden drei Zeugnisformate unterschieden: Als Notenzeugnis gilt jedes Zeugnis, das den Lernstand von Schülerinnen und Schülern mit Ziffern bewertet. Als Berichtszeugnis gilt jedes Zeugnis, das den Lernstand von Schülerinnen und Schülern in Form eines Fließtexts beschreibt und bewertet. Als Kompetenzraster gilt jedes Zeugnis, das den Lernstand in tabellarischer oder grafischer Form beschreibt und bewertet, wobei das Format einzelne Kompetenzen vorgibt, deren Erreichungsgrad dann markiert wird. Bereits im Jahr 1970 wurde von der KMK die Einführung von Verbalbeurteilungen beschlossen, sodass es zu einer vermehrten Verwendung von qualitativen Berichten kam – vor allem an der Grundschule, wo die Notenzeugnisse weitgehend von den Berichtszeugnissen abgelöst wurden. Die Berichtszeugnisse mit ihrer Aufgabe mittels allgemeiner Aussagen die Leistung eines Kindes, seine Fähigkeiten und Schwächen, sein Sozial- und Arbeitsverhalten und seine besonderen Interessen zu beschreiben, entsprachen somit der von der KMK initiierten Individualisierung des Unterrichts. Sämtliche Zeugnisformate werden auch in der wissenschaftlichen Diskussion kontrovers diskutiert (Lissman 2008). Es liegt allerdings nahe, zu vermuten, dass alle Bewertungssysteme ähnliche Vor- und Nachteile bezüglich der Gütekriterien haben. Auch die Problematik des erwähnten Big-Fish-Little-Pond-Effekts (s. Kapitel zur Motivation, S. 121) verschwindet vermutlich nicht einfach aufgrund eines ande-

ren Bewertungsformats. Überraschend ist, dass es **nur wenige Studien gibt**, die solche Effekte der unterschiedlichen Zeugnisformate vergleichen. Eher gibt es Befragungen zur Akzeptanz der Zeugnisformate: In einer Befragung von Hamburger Schülerinnen und Schülern, Lehrkräften und Eltern unmittelbar nach der Vergabe der Zeugnisse wurden Berichtszeugnisse und Kompetenzrasterzeugnisse deutlich positiver bewertet als Notenzeugnisse (s. Tchoudova et al. 2014). Vor allem Lehrkräfte präferierten die Kompetenzraster; in fast allen Fragen ergaben sich für die Kompetenzraster signifikant höhere Akzeptanzwerte als für die Noten- und Berichtszeugnisse. Lehrkräfte sahen Berichtszeugnisse gegenüber Notenzeugnissen vor allem bei der Frage nach zukünftigen Leistungsentwicklungen der Schülerinnen und Schüler im Vorteil.

3 Fazit

Diagnostik von Schülerleistungen und die Entwicklung geeigneter Instrumente wie z. B. guter Klassenarbeiten ist ein komplexer Prozess, den viele Lehrerinnen und Lehrer im Unterrichtsalltag gut beherrschen. Dennoch kann dieses Kapitel die Professionalität von Lehrkräften bei der Beurteilung von Schülerinnen und Schülern weiter erhöhen, indem die hier besprochenen Grundlagen der Diagnostik dazu genutzt werden die eigenen diagnostischen Methoden und Urteile kritisch zu reflektieren. Dazu können Lehrkräfte zum einen prüfen, ob verwendete Aufgabenstellungen, Klassenarbeiten, Bewertungskriterien etc. objektiv, reliabel und valide sind. Oft werden diese Gütekriterien bereits recht gut erfüllt, was sich in der erwähnten hohen Stabilität von Noten, aber auch in der insgesamt zufriedenstellenden Urteilsgenauigkeit der Lehrkräfte niederschlägt.

Neben diesen allgemeinen Grundlagen der Diagnostik haben wir darauf hingewiesen, wie Urteile zustande kommen, und dass mögliche Verzerrungen häufig auf Einflüsse eigentlich urteilsirrelevanter Merkmale (z. B. Geschlecht, familiärer Hintergrund, Sympathie, Erwartungen) zurückzuführen sind. Die Berücksichtigung solcher Merkmale bei der Beurteilung von Schülerinnen und Schülern, gefährdet vor allem die Objektivität. Das Wissen um solche unerwünschten und oft unbewusst stattfindenden Prozesse kann Lehrkräfte dazu ermuntern, selbst Maßnahmen zu ergreifen, die zu einer **Objektivierung ihrer Urteile** beitragen. Woolfolk (2008) schlägt verschiedene Maßnahmen vor, um zu objektiveren Einschätzungen von Schülerinnen und Schülern zu gelangen. Dazu gehört es zunächst, manche Informationen über Schülerinnen und Schüler (z. B. Aussagen von Kolleginnen und Kollegen etwa bei der Übernahme einer neuen Klasse) mit Vorsicht zu behandeln. Zudem sollten Lehrkräfte vor allem bei der Beurteilung schwächerer Schülerinnen und Schüler aufmerksam für fehlerhafte Beurteilungen sein und insgesamt versuchen, diskriminierende Unterrichtshandlungen zu vermeiden (bei schwierigen Mathematikaufgaben nicht nur Jungen aufrufen). Disziplinarische Maßnahmen sollten unabhängig von der Leistung sein, d. h. bei

einem Fehlverhalten eines leistungsstarken Schülers sollte nicht ein Auge zuge-drückt werden, wenn dieses Verhalten bei einem schwächeren Schüler sanktioniert wird. Zuletzt empfiehlt sie Lehrkräften, sich so **oft wie möglich Notizen zu** urteilsrelevanten Informationen zu machen, um Erinnerungsfehler zu vermeiden und ein hohes Maß an Objektivität zu erzielen. Schließlich könnte es eine gute Idee sein, vor dem Schreiben einer Klassenarbeit die Ergebnisse der Schülerinnen und Schüler einmal zu prognostizieren und die Prognosen dann mit dem tatsächlichen Ergebnis zu vergleichen.

Literatur

- Alvidrez, J. & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. In: *Journal of Educational Psychology*, 91, S. 731–746.
- Anders, Y., Kunter, M., Brunner, A., Krauss, S. & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. In: *Psychologie in Erziehung und Unterricht*, 57, S. 175–193.
- Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. In: *Zeitschrift für Erziehungswissenschaft*, 9, S. 469–520.
- Betsch, T., Funke, J. & Plessner, H. (2011). *Denken – Urteilen, Entscheiden, Problemlösen*. Heidelberg: Springer.
- Carr, M. & Kurtz-Costes, B. E. (1994). Is being smart everything? The influence of student achievement on teachers' perceptions. In: *British Journal of Educational Psychology*, 64, S. 263–276.
- Hadley, S. T. (1989). Feststellungen und Vorurteile in der Zensurierung. In: K.-H. Ingenkamp (Hrsg.), *Die Fragwürdigkeit der Zensurung* (S. 159–166). Weinheim: Beltz.
- Härtig, H., Heitmann, P. & Retelsdorf, J. (2015). Analyse der Aufgaben zur Evaluation der Bildungsstandards in Physik – Differenzierung von schriftsprachlichen Fähigkeiten und Fachlichkeit. *Zeitschrift für Erziehungswissenschaft*, 18, S. 763–779. <doi:10.1007/s11618-015-0646-2>
- Hinnant, J. B., O'Brien, M. & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. In: *Journal of Educational Psychology*, 101, S. 662–670.
- Jussim, L. & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. In: *Personality and Social Psychology Review*, 9, S. 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. In: *Zeitschrift für Pädagogische Psychologie*, 26, S. 251–261.
- Kaiser, J., Retelsdorf, J., Südkamp, A. & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. In: *Learning and Instruction*, 28, S. 73–84.
- 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. In: *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 44, S. 163–177.
- Lissmann, U. (2008). *Leistungsmessung und Leistungsbeurteilung. Eine Einführung*. Landau: Verlag Empirische Pädagogik.
- Ludwig, P. H. (2010). Erwartungseffekt. In: D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 144–150). Weinheim: Beltz.
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. In: *Journal of Educational Psychology*, 79, S. 280–295.
- McKown, C. & Weinstein, C. E. (2002). Modeling the role of child ethnicity and gender in children's differential response to teacher expectations. *Journal of Applied Social Psychology*, 32, S. 159–184.

- Merton, R. K. (1948). The self-fulfilling prophecy. In: *The Antioch Review*, 8, S. 193–210.
- Möller, J. & Bonerad, E.-M. (2007). Fragebogen zur habituellen Lesemotivation. In: *Psychologie in Erziehung und Unterricht*, 54, S. 259–267.
- Nisbett, R. E. & Wilson, T. D. (1977). The halo effect: evidence for unconscious alteration of judgments. In: *Journal of Personality and Social Psychology*, 35, S. 250–256.
- Plante, I., De la Sablonnière, R., Aronson, J. M. & Théorêt, M. (2013). Gender stereotype endorsement and achievement-related outcomes: The role of competence beliefs and task values. In: *Contemporary Educational Psychology*, 38, S. 225–235.
- Ready, D. D. & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: the role of child background and classroom context. In: *American Educational Research Journal*, 48, S. 335–360.
- Retelsdorf, J., Schwartz, K. & Asbrock, F. (2015). "Michael can't read!" – Teachers' gender stereotypes and boys' reading self-concept. In: *Journal of Educational Psychology*, 107, S. 186–194.
- Ritts, V., Patterson, M. L. & Tubbs, M. E. (1992). Expectations, impressions and judgments of physically attractive students: A review. In: *Review of Educational Research*, 62, S. 413–426.
- Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom*. New York, NY: Holt, Rinehart & Winston.
- Schrader, F.-W. (2010). Diagnostische Kompetenz von Eltern und Lehrern. In: D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 102–108). Weinheim: Beltz.
- Schrader, F.-W. & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile. In: *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 22, S. 312–324.
- Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. In: *Journal of Educational Psychology*, 104, S. 743–762.
- Tchoudova, N., Schroeter, B., Harms, U., Möller, J., Paulick, I. & Hahn, C. (2014). Abschlussbericht der Evaluation des Hamburger Schulversuchs alles»können. Senat der Hansestadt Hamburg.
- Thorndike, E. L. (1920). A constant error in psychological ratings. In: *Journal of Applied Psychology*, 4, S. 25–29.
- Weinert, F. E., Schrader, F.-W. & Helmke, A. (1990). Educational Expertise. In: *School Psychology International*, 11, S. 163–180.
- Woolfolk, A. E. (2008). *Pädagogische Psychologie*. München: Pearson Studium.