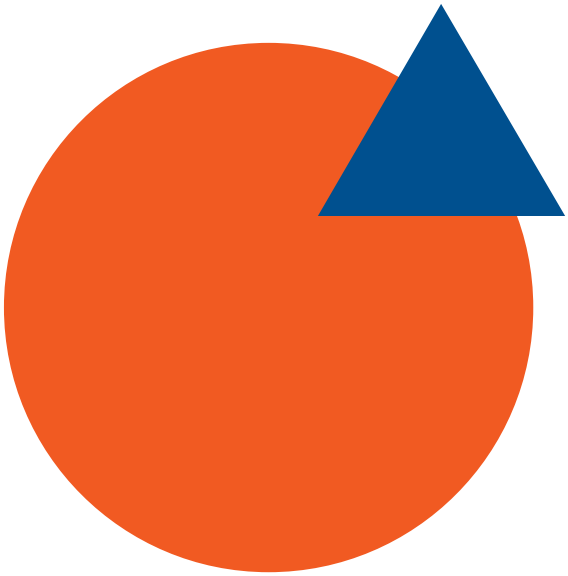
Brügelmann, Hans

# Sind Noten nützlich - und nötig? Ziffernzensuren und ihre Alternativen im empirischen Vergleich. Eine wissenschaftliche Expertise des Grundschulverbandes

3. aktualisierte Auflage

*Frankfurt am Main : Grundschulverband e.V. 2014, 72 S.*

Eine wissenschaftliche Expertise
des Grundschulverbandes



# Sind Noten nützlich und nötig?

## Ziffernzensuren und ihre Alternativen im empirischen Vergleich

3. aktualisierte Auflage 2014 · Grundschulverband · Niddastraße 52 · 60329 Frankfurt/Main

Grundschul
verband

## Vorwort zur 3. aktualisierten Auflage

Sieben Jahre sind seit Ersterscheinen dieser Expertise vergangen. Im Schulalltag hat sich in dieser Zeit allerdings nichts grundlegend verändert. Dies verwundert nicht, sind die Befunde, die in dieser Expertise genutzt wurden - ohne seit ihrem Entstehen entkräftet worden zu sein - teilweise auch bald 50 und mehr Jahre alt und haben schon damals kaum Wirkung gezeigt. Wir drucken das Gutachten deshalb unverändert nach, um die kritische Diskussion in Gang zu halten.

Gleichwohl ist in diesem Feld weiter gearbeitet worden. Unter den zwischenzeitlich erschienen Veröffentlichungen ist ein Band zur Funktion und Geschichte von Zeugnissen erhellend (Urabe 2009).

Die Datenlage selbst hat sich aber nicht verändert (vgl. etwa die Überblicke bei Tent 2006; Faber/Billmann-Mahecha 2010). Im Gegenteil: Neuere Untersuchungen und Analysen bestätigen die in diesem Band zusammengetragenen *Forschungsbefunde* zur Problematik von Ziffernnoten:

▲ Noten sind nicht objektiv und nicht vergleichbar, sondern in hohem Maße abhängig von der beurteilenden Person und zusätzlich ist die Leistung der SchülerInnen abhängig von ihrer Beziehung zur Lehrperson (Deimel 2007; Bergin/Bergin 2009; Birkel 2009); insbesondere soziale Herkunft und Geschlecht tragen zu einer ungerechten Beurteilung bei (Maaz u.a. 2011), aber ebenso die Leistungsstärke der Klasse (Bezugsgruppeneffekt, vgl. Tiedemann/Billmann-Mahecha 2007; Baumert u.a. 2010);

▲ Noten erlauben keine zureichend genaue Vorhersage der weiteren Leistungsentwicklung, um z.B. Übergangsentscheidungen (Baumert u.a. 2010, 13; Trautwein u.a. 2008; Perleth/Sen 2010) oder Auswahlentscheidungen in Betrieben (Herrmann/Bohn 2009) zu fundieren;

▲ Noten gehören weiterhin zu den verbreitetsten Angstauslösern unter Kindern (vgl. die »Kinderbarometer« in LBS 2007 ff.; und die Kinderstudien von World Vision 2010; 2013).

Die *bildungspolitischen* Entwicklungen seit Veröffentlichung dieser Expertise (2006) sind widersprüchlich. Immer noch werden trotz der vorliegenden Forschungsbefunde neue »Modellversuche« eingerichtet. Selbst wenn diese - wie z.B. in Nordrhein-Westfalen (vgl. Bos u.a. 2010) - zu positiven Ergebnissen geführt haben, legen andere Bundesländer ihre eigenen Versuche auf, z.B. 2013 Baden-Württemberg.

Die im Grundschulbereich in der Regel ab Ende der zweiten Klasse, spätestens ab Ende Klasse 3 vorgeschriebenen Ziffernnoten wurden in einigen Bundesländern wie Bremen und Hamburg für die neue Schulformen auf der Sekundarstufe ab Klasse 5 wieder abgeschafft. In Baden-Württemberg führt das - wie auch in einigen anderen Bundesländern - zu der absurden Situation, dass nach Notenfreiheit bis Mitte

zweiter Klasse Ziffernnoten verpflichtend, in der anschließenden Gemeinschaftsschule aber für die ersten Jahrgänge nicht mehr vorgesehen sind …

Bis 2007 waren in sieben Bundesländern Kopfnoten neu eingeführt worden (vgl. Bartnitzky 2008), schon 2008 hat Bayern die Ziffern (bzw. Buchstaben) wieder abgeschafft (SPIEGELonline 2008), allerdings nur, um sie durch notengleiche Satzbausteine zu ersetzen. Eine interne Erhebung des Grundschulverbands 2011 bei den Schulverwaltungen der 16 Bundesländer ergab einen bunten Fleckenteppich von Varianten für die Bewertung des Arbeits- und Sozialverhaltens - von Ziffernnoten über verbale Beurteilungen bis hin zu gar keinen Aussagen

In der *Unterrichtspraxis* haben sich vielerorts neue Formen der Lernbeobachtung und Leistungsbeurteilung durchgesetzt, wie sie der Grundschulverband in seinen Bänden »Pädagogische Leistungskultur« (Bartnitzky u.a. 2005, 2006, 2007) vorgeschlagen hat. Je nach Bundesland sind die Freiräume dafür unterschiedlich groß, die von manchen KollegInnen kaum, von anderen dagegen voll ausgenutzt werden (vgl. etwa die Berichte von Czerny 2010; Leppert 2011).

So müssen beispielsweise in Baden-Württemberg zwar Noten in den Zeugnissen, aber nicht für einzelne Leistungen, z.B. Klassenarbeiten, vergeben werden. Entsprechend heißt es in § 7 der Noten-Verordnung: *»Die Bildung der Note in einem Unterrichtsfach ist eine pädagogisch-fachliche Gesamtwertung der vom Schüler im Beurteilungszeitraum erbrachten Leistungen.«*

Andererseits bewahrt selbst die Einhaltung der schon 1968 von der KMK festgelegten Kriterien-Orientierung nicht vor repressiven Maßnahmen. Dies musste etwa eine Kollegin in Bayern feststellen, als sie wegen guter Lernerfolge aller Kinder in ihrer Klasse entsprechend gute Noten vergeben hatte. Selbst die Bestätigung dieser Beurteilungen durch überdurchschnittliche Ergebnisse in einem Vergleichstest bewahrten sie nicht vor einer Versetzung wegen *»Störung des Schulfriedens«* (Bleher 2008).

Um die negativen Nebenwirkungen vergleichender Noten abzumildern, verbinden inzwischen viele Schulen die Übergabe der Zeugnisse mit dem Elternsprechtag und beziehen möglichst auch die Kinder in den Rückblick und in Absprachen über die weitere Arbeit ein.

Angesichts der klaren politischen Vorgabe, dass die Schulen inklusiv werden müssen, werden sich Ziffernzensuren als vergleichende Leistungsbeurteilung auf Dauer nicht halten lassen. Andererseits ist dieses Feld immer noch konfliktträchtig und deshalb selbst für einsichtige BildungspolitikerInnen ein (zu) »heißes Eisen«. Ein pragmatischer Weg könnte eine übergangsweise Öffnungsklausel sein, die Schulen erlaubt, auf Ziffernnoten zu verzichten, wenn sie ein Alternativkonzept entwickeln, das von der Schulkonferenz verabschiedet wird. Diese Ernstnahme der allerorten propagierten »selbstständigen Schule« würde einen schrittweisen Übergang überall dort ermöglichen, wo LehrerInnen sich andere Formen der Leistungsbeurteilung zutrauen und die Elternschaft für diesen Weg gewinnen können.

Dass Schulen dazu in der Lage und gewillt sind, zeigen nicht nur Schulen in Schulversuchen, sondern auch die Bensberger Erklärung des Schulverbunds »Blick über den Zaun«. Die Schulen des Verbunds fordern darin *»einen veränderten Umgang mit Schülerleistungen. Wir brauchen differenziertere Instrumente als Zensuren. Verbesserte und unterstützende Formen der staatlichen Evaluation von Unterricht. Wir brauchen differenziertere Instrumente als standardisierte Tests und Prüfungen.«*

Sie fordern aber nicht nur, sondern bieten gleichzeitig aus ihrer Praxis gelungene *»Beispiele für eine veränderte Schul- und Lernkultur, mit einer konsequenten Individualisierung und Freiräumen, in denen Kinder und Jugendliche, eingebunden in eine verlässliche Gemeinschaft, Verantwortung für ihr Lernen und ihre persönliche Entwicklung übernehmen.*

*Beispiele von Schulen, die zeigen, wie Kinder ohne Noten und ohne Selektion gemeinsam lernen und dadurch individuell bestmögliche Leistungen erreichen können.*

*Beispiele für einen veränderten Umgang mit Leistungen, für eine prozessorientierte und transparente Leistungsrückmeldung.«*

Da wie schon 2006 gilt, dass Zensuren weder nötig, noch nützlich, informationsreich oder lernförderlich sind, sollte die Bildungspolitik den Erfahrungen und Forderungen dieser Schulen Raum geben.

*Hans Brügelmann und Axel Backhaus*

---


Bartnitzky, H. (2008): Zur Renaissance der »Kopfnoten« - Anmerkungen zur Umfrage bei den Schulministerien. In: Grundschule aktuell, Nr. 101, 24-25.

Bartnitzky, H., u.a. (Hrsg.): Pädagogische Leistungskultur. Beiträge zur Reform der Grundschule. Bde. 119, 121, 124. Grundschulverband: Frankfurt.
Bd. 119 (2005): Materialien für Klasse 1/2 (Deutsch, Mathematik, Sachunterricht)
Bd. 121 (2006): Materialien für Klasse 3/4 (Deutsch, Mathematik, Sachunterricht)
Bd. 124 (2007): Ästhetik, Sport, Englisch - Arbeits-/Sozialverhalten.

Baumert, J. u.a. (2010): Der Übergang von der Grundschule in die weiterführende Schule - Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten: Zusammenfassung der zentralen Befunde. In: Maaz u.a. (2010, 5-21).

Bergin, C.A./Bergin, D.A. (2009). Attachment in the classroom. In: Educational Psychology Review, Vol. 21, 141-170.

Birkel, P. (2009): Rechtschreibleistung im Diktat - eine objektiv beurteilbare Leistung? In: Didaktik Deutsch, 15. Jg., H. 27, 5-32.

Bleher, C. (2008): Bloß nicht zu viele Einser, bitte! Lehrer, deren Schüler zu gute Noten schreiben, werden systematisch ausgebremst. In: Süddeutsche Zeitung v. 28.7.2008. Download: http://www.christianbleher.de/texte/bildung-schule/noten/

Bos, W., u.a. (2010): LUZI. Leistungsbeurteilung ohne Ziffernzeugnisse. Abschlussbericht der wissenschaftlichen Begleitforschung. Institut für Schulentwicklung der Universität: Dortmund.

Czerny, S. (2010): Was wir unseren Kindern in der Schule antun: ... und wie wir das ändern können. Südwest Verlag: München.

Deimel (2007): Über die Unmöglichkeit, objektiv zu urteilen - Zur Klärung eines Paradoxons. Download: http://www.aba-fachverband.org/index.php?id=1257 [Abruf: 28.11.2013].

Faber, G./Billmann-Mahecha, E. (2010): Notengebung im Spiegel wissenschaftlicher Untersuchungen. Probleme, Erfordernisse und Möglichkeiten aus pädagogisch-psychologischer Sicht. In: Lernchancen, 13. Jg., H. 74, 30-33.

Herrmann, U./Bohn, H. (2009): Leistungsbeurteilung, Selbsteinschätzung und Bildungsstandards - nicht nur in der Berufsausbildung. In: Lehren und Lernen, 35. Jg., H. 2, 30-37.

LBS-Initiative Junge Familie (Hrsg.) (2007): LBS-Kinderbarometer Deutschland 2007. Stimmungen, Meinungen, Trends von Kindern in sieben Bundesländern. Ergebnisse des Erhebungsjahres 2006/07. PROSOZ ProKids-Institut: Herten.

LBS-Initiative Junge Familie (Hrsg.) (2011): LBS-Kinderbarometer Deutschland 2011. Stimmungen, Trends und Meinungen von Kindern aus Deutschland. PROSOZ Institut für Sozialforschung: Herten.

Lin-Klitzing, S., u.a. (Hrsg.) (2010): Übergänge im Schulwesen. Chancen und Probleme aus sozialwissenschaftlicher Sicht. Julius Klinkhardt: Bad Heilbrunn.

Maaz, K., u.a. (Hrsg.) (2010): Der Übergang von der Grundschule in die weiterführende Schule. Bundesministerium für Bildung und Forschung: Berlin. Download: http://www.bmbf.de/pub/bildungsforschung_band_vierunddreissig.pdf [Abruf: 25.11.2013].

Maaz, K., u.a. (2011): Herkunft zensiert? Leistungsdiagnostik und soziale Ungleichheiten in der Schule. Vodafone Stiftung Deutschland: Düsseldorf.

Perleth, C./Sen M.A. (2010): Zuverlässigkeit von Schulnoten, kognitiven Fähigkeitstests und Begabungseinschätzung von Eltern für die weitere Schullaufbahn. In: Lin-Klitzing u.a. (2010, 105-126).

Rost, D.H. (Hrsg.) (2006): Handwörterbuch Pädagogische Psychologie. Weinheim: Beltz (3. Aufl.).

Schlemmer, E./Gerstberger, H. (Hrsg.): Ausbildungsfähigkeit im Spannungsfeld zwischen Wissenschaft, Politik und Praxis. Wiesbaden: Verlag für Sozialwissenschaften.

Schulverbund ›Blick über den Zaun‹: Bensberger Erklärung. Download: www.blickueberdenzaun.de/publikationen/112-bensbergererklaerung.html [Abruf: 27.11.2013].

SPIEGELonline (2008): Benimm-Zeugnisse. Bayern schafft die Kopfnoten ab. Download: http://www.spiegel.de/schulspiegel/wissen/0,1518,530847,00.html [Abruf: 26.11.2013].

Tent, L. (2006). Zensuren. In: Rost (2006, 873-880).

Tiedemann, J./Billmann-Mahecha, E. (2007): Zum Einfluss von Migration und Schulklassenzugehörigkeit auf die Übergangsempfehlung für die Sekundarstufe I. Zeitschrift für Erziehungswissenschaft, 10. Jg., H. 1, 108-120.

Trautwein, U., u.a. (2008): Die Sekundarstufe I im Spiegel der empirischen Bildungsforschung: Schulleistungsentwicklung, Kompetenzniveaus und die Aussagekraft von Schulnoten. In: Schlemmer/Gerstberger (2008, 91-107).

Urabe, M. (2009): Funktion und Geschichte des deutschen Schulzeugnisses. Klinkhardt: Bad Heilbrunn.

World Vision (Hrsg.) (2010): Kinder in Deutschland 2010. 2. World Vision Kinderstudie. Download: http://www.worldvision-institut.de/_downloads/allgemein/Kinderstudie2010_Zusammenfassung. pdf [Abruf: 28.11.2013].

World Vision (Hrsg.) (2013): Kinder in Deutschland 2013. 3. World Vision Kinderstudie. Download:http://www.stern.de/panorama/infografiken-zur-world-vision-kinderstudie-was-kindern-wirklich-wichtig-ist-2071288-341bf8048e0ad9c4.html

# Kurzfassung
# für eilige LeserInnen

# Sind Noten nützlich - und nötig?

## Ziffernzensuren und ihre Alternativen im empirischen Vergleich

*Eine wissenschaftliche Expertise*
*des Grundschulverbandes*
*erstellt von der Arbeitsgruppe Primarstufe*
*an der Universität Siegen*

*von Hans Brügelmann mit*
*Axel Backhaus, Erika Brinkmann (Gast)*
*Hendrik Coelen, Thomas Franzkowiak*
*Simone Knorre, Barbara Müller-Naendrup*
*Elisabeth Oser, Sara Roth*

## Das Wichtigste auf einen Blick[1]

Ein erstes Problem in der Diskussion ist die unklare Begrifflichkeit. In diesem Gutachten verstehen wir - sofern nicht ausdrücklich etwas anderes gesagt wird - unter »Noten« bzw. »Zensuren« *Ziffern*noten, die zur *formellen* Beurteilung verwendet werden, z.B. bei Klassenarbeiten oder in Zeugnissen.
»Leistungsbeurteilung« verwenden wir als Obergriff für die *Beschreibung* und die *Bewertung* von Leistungen - zwei unterschiedliche Formen ihrer Rückmeldung, die auch in der Umsetzung sorgfältig zu trennen sind.

1 Der folgende Text ist eine Zusammenfassung der ausführlichen *Expertise,* in dem die einschlägigen Publikationen, insbesondere die empirischen Studien, differenziert ausgewertet und belegt sind. Verweise im Text beziehen sich auf die entsprechenden Kapitel der *Lang*fassung. Unsere Analysen legen grundlegende Probleme einer pädagogischen Leistungsbeurteilung offen, die konkreten Folgerungen beziehen sich aber vor allem auf die Grundschule.

## Zwei zentrale Erträge unserer Analysen vorweg

▲ Leistungsbeurteilungen haben in unserem Schulsystem nicht nur unterschiedliche, sondern oft widersprüchliche Funktionen zu erfüllen: als Beschreibungen *orientieren* sie über den individuellen Leistungsstand und über Möglichkeiten zu dessen gezielter Verbesserung; sie sind damit ein pädagogisches Medium zur Förderung des Lernens. Als Bewertungen dienen sie der *Disziplinierung* und *Selektion.* Spätestens seit der UN-Kinderrechtskonvention erweist sich ein *hierarchisches* Verständnis von Leistungsbeurteilung als nicht mehr zeitgemäß. Nicht Anpassung und Gehorsam, sondern Mitbestimmung und Selbstverantwortung sind vorrangige Erziehungsziele einer demokratischen Schule. Schärfere Selektion führt im Übrigen nicht zu besseren Leistungen wie die internationalen Leistungsstudien gezeigt haben.

*Empfehlung:*
Eine demokratische Schule hat die Persönlichkeit der SchülerInnen durch Formen der Dokumentation und der Bewertung von Leistung zu achten, die ihre Selbstständigkeit fördern statt Abhängigkeiten zu verstärken. Einem solchen Verständnis von Schule sind Noten als Belohnungs-/-Bestrafungssystem nicht mehr angemessen. Vielmehr ist die Fähigkeit zur Selbsteinschätzung und zum konstruktiven Umgang mit Kritik zu fördern. Hierfür ist eine sachliche Information der SchülerInnen über den individuellen Stand ihrer Lern- und Leistungsentwicklung unerlässlich.

▲ Ziffernnoten sind immer noch die häufigste Form formeller Leistungsbewertung in der Schule. Aber die Forschung zeigt seit langem: Noten sind nicht in der behaupteten Weise für das Lernen nützlich und sie sind erst recht nicht nötig. Sie betonen einseitig die Bewertungsfunktion - können aber auch diese wegen ihrer mangelnden Aussagekraft, Vergleichbarkeit und Objektivität nicht angemessen erfüllen. Es gibt deshalb keinen Grund, auf ihnen zu beharren, zumal sie darüber hinaus etliche unerwünschte Nebenwirkungen haben.

*Empfehlung:*
Ziffernoten sind zu ersetzen durch differenziertere Formen der Dokumentation und der Bewertung von Leistungen. Rückmeldung und Bewertung sind klar zu trennen. Beschreibungen sollen den Leistungsstand bezogen auf konkrete Lernziele und die individuelle Entwicklung darstellen. Das lernförderliche Potenzial differenzierter Rückmeldungen wird in der Praxis aber nur dann zur Geltung gebracht werden können, wenn die entsprechenden Rahmenbedingungen geschaffen werden: vor allem durch eine Verringerung des Selektionsdrucks im Bildungssystem und durch eine fachliche Qualifizierung der LehrerInnen.

## Die Ergebnisse unserer Analysen im Einzelnen

▲ Noten sind *informationsarm.* Dieselbe Punktzahl in einer Probe kann Ausdruck ganz unterschiedlicher Leistungen sein. Entsprechend werden unterschiedliche Leistungsprofile mit derselben Ziffer belegt.

*Empfehlung:*
Leistungen sollten nicht nur bewertet, sondern zunächst differenziert beschrieben werden. Die individuellen Stärken und Entwicklungsmöglichkeiten verdienen eine besondere Beachtung. Wo Noten vorgeschrieben bleiben, sind sie schriftlich oder im Gespräch inhaltlich zu kommentieren.

▲ Noten sind *nicht vergleichbar,* da die Bewertung in der Regel auf den Durchschnitt einer Klasse bezogen wird. Je nach Leistungsniveau der einzelnen Klasse wechseln die Noten für dieselbe Leistung. Zudem sind die Maßstäbe je nach Fach und Altersstufe unterschiedlich.

*Empfehlung:*
Soweit Leistungen überhaupt vergleichend beurteilt werden, sollten *Bewertungen* auf klassenübergreifende Stichproben bezogen werden (z.B. in Form von Prozentrangplätzen in normierten Tests). Allerdings muss bedacht werden, dass Tests nur bestimmte Arten von Leistungen erfassen können. Sie dürfen deshalb nicht zum heimlichen Curriculum werden - zum Beispiel über zentrale Lernstandserhebungen.

▲ Vorrangig orientiert sich die Leistungsbewertung immer noch am Vergleich mit einer Bezugsgruppe. Die Dominanz des sozialen Vergleichs bei der Notengebung *widerspricht* allerdings den *rechtlichen* Vorgaben. Sie hat zudem *negative* Auswirkungen auf die *Lernmotivation* von leistungsschwächeren SchülerInnen, und sie beschädigt die Kraft *intrinsischer Motivation* auch bei den leistungsstärkeren.

*Empfehlung:*
Die Bewertung von Leistungen muss sich deshalb stärker an Lernzielen und in der Grundschule vor allem am individuellen Lernfortschritt (Entwicklungsnorm) orientieren.

▲ Zensuren sind Urteile von Lehrpersonen. Sie basieren in der Regel auf informellen Leistungsproben und Beobachtungen. Diese Daten und ihre Bewertung in Form von Noten haben sich als *nicht zureichend gültig* (»valide«), *personunabhängig* (»objektiv«) und *verlässlich* (»reliabel«) erwiesen. Soziale und ethnische Herkunft, Geschlecht, aber auch Verhaltensauffälligkeiten und persönliche Sympathie führen zu systematischen Verzerrungen der Beurteilung. Deren Fehleranfälligkeit verliert erst an Bedeutung, wenn sie nicht zu Selektionszwecken eingesetzt werden. Für lernförderliche Rückmeldungen sind Empathie und eine persönliche Beziehung sogar von Vorteil. Im Übrigen kann die Nutzung von standardisierten Tests zwar die Datengrundlage von Beurteilungen erweitern; ersetzen können die - ebenfalls fehleranfälligen - Tests das Lehrerurteil aber nicht.

*Empfehlung:*
Leistungen sind möglichst zu mehreren Zeitpunkten und in unterschiedlichen Aufgaben/Situationen zu *erfassen.* Vorstrukturierte Portfolios bieten eine gute Möglichkeit, Leistungen differenzierter, aus verschiedenen Perspektiven und in ihrer Entwicklung über die Zeit hinweg zu dokumentieren. *Bewertet* werden sollten Leistungen möglichst von mehreren Personen, die den Kontext der Leistung und ihrer Entwicklung kennen.

▲ Als Alternative zu Noten werden Verbalbeurteilungen vorgeschlagen. Da sie in der Regel wie Noten auf den Beobachtungen und Bewertungen von LehrerInnen basieren, unterliegen sie aber denselben Einschränkungen, was ihre Validität, Objektivität und Reliabilität angeht. Ihr Vorzug gegenüber Noten: Zumindest vom Anspruch her erfassen sie Leistungen differenzierter, ihre Aussagen lassen individuelle Besonderheiten besser erkennen und sie orientieren sich stärker am Lernfortschritt; darüber hinaus machen sie die Maßstäbe der Lehrperson und die Lernbedingungen deutlicher erkennbar. In der Realität werden Verbalgutachten diesen Anforderungen in vielen Fällen aber nicht gerecht.

*Empfehlung:*
Um die Vorteile einer verbalen Dokumentation und entwicklungsbezogenen Bewertung von Leistungen stärker zur Geltung zu bringen, sind vier Maßnahmen erforderlich:
- Sensibilisierung von LehrerInnen für die Schwierigkeiten bzw. Fallen von Beurteilungen sowie für die Erwartungen und Lesarten der Zielgruppen;
- eine gezielte Aus-/Fortbildung ihrer Kompetenzen zur Erfassung, Interpretation, Bewertung und differenzierten Darstellung von Leistungen sowie von deren Entwicklung;
- Entwicklung von fachdidaktisch begründeten Kriterien für die Beurteilung von Leistungen - und zwar immer wieder neu in Zusammenarbeit mit den LehrerInnen vor Ort;
- Organisation eines kontinuierlichen kollegialen Austausches über die Maßstäbe und über ihre Anwendung in kritischen Fällen.

▲ Gegen eine ausschließlich verbale Begutachtung unter völligem Verzicht auf Noten haben viele Eltern, LehrerInnen und SchülerInnen immer noch *Vorbehalte.* Diese Skepsis hat aber in der Grundschule und dort besonders bei Personen, die eigene Erfahrungen mit dieser Praxis haben, deutlich

abgenommen. Empirisch *widerlegt* sind die Befürchtungen, Verbalbeurteilungen hätten einen *negativen* Einfluss auf die Leistungsbereitschaft. Bei konsequenter Umsetzung einer ziel- und entwicklungsorientierten Bewertung von Leistungen lassen sich im Gegenteil sogar *positive* Effekte auf das Lernklima in der Klasse sowie auf die Einstellungen und die Motivation der SchülerInnen nachweisen.

*Empfehlung:*
Wie auch in vielen beruflichen Bereichen sollten zunehmend dialogische Formen einer Verbindung von Selbst- und Fremdbeurteilungen erprobt werden. Die Fähigkeit zur Wahrnehmung und Einschätzung der eigenen Leistung ist gezielt zu entwickeln und in der alltäglichen Anwendung zu unterstützen. Noten »von oben« fördern weder diese Fähigkeit noch die Bereitschaft zur Selbstkritik, sondern provozieren eher Abwehr- oder Ausweichverhalten. Eine symmetrische Beziehung schließt außerdem ein, dass die SchülerInnen nicht nur ihre eigene Leistung, sondern auch die Bedeutung von Lernbedingungen einzuschätzen lernen.

▲ Trotz der durchgängig negativen Befunde über Nutzen und Nebenwirkungen von Ziffernnoten dürfte deren Abschaffung schwierig werden. Dies hängt vor allem mit der frühen und starken *Selektionsorientierung* des deutschen Schulsystems zusammen. Eine rein »technische« Verbesserung des Beurteilungswesens wird deshalb in der Praxis nicht viel bewirken, wenn sich die institutionellen Bedingungen nicht ändern: Verlängerung der gemeinsamen Schulzeit; Abschaffung von Zurückstellungen am Schulanfang, der Wiederholung von Klassen und der Aussonderung in Sonderschulen.

*Empfehlung:*
Der Förderauftrag der Schule muss bildungspolitisch, in den Schulprogrammen und in der täglichen Arbeit vor Ort Vorrang vor der Selektion gewinnen. PISA-Spitzenreiter wie Schweden - oder im deutschsprachigen Raum: Südtirol - kommen seit vielen Jahren ohne vergleichende Noten aus. Werden dagegen Sanktionen an die Bewertung von Leistungen geknüpft, ist mit einem Rückschlag in dem Bemühen um Verbesserungen zu rechnen. Ranking sowie Selektion *in* und *von* Schulen haben sich vor allem in den angelsächsischen Ländern als pädagogisch kontraproduktiv erwiesen.

## Fazit:
## Vier Resümees aus vier Perspektiven

Wie bei allen pädagogischen Fragen (und sozialen Phänomenen generell) ist die Befundlage zu Noten nicht auf einen einfachen Nenner zu bringen. Formen der Leistungsbeurteilung wirken unterschiedlich, je nachdem *wie* und in welchem *Kontext* sie eingesetzt werden. Für Folgerungen aus dem Forschungsstand kommt es deshalb darauf an, von welcher Basisannahme man ausgeht: Wer die Beweislast für Veränderungen bei den Reformern sieht, kann zu einer anderen Einschätzung kommen als jemand, der normativ die Förderung des Einzelnen als zentrale Norm und noch uneingelöste Aufgabe der Schule sieht. Vor diesem Hintergrund lassen sich als Ergebnis unserer Analysen vier Folgerungen formulieren:

▲ Wer an Ziffernnoten festhalten will, weil sie angeblich objektiv und vergleichbar seien bzw. erforderlich, damit SchülerInnen sich auf die Anstrengungen des Lernens einlassen, findet in der Empirie keine stützenden Belege für seine Position.

▲ Auch diejenigen, die Verbalgutachten ablehnen, weil sie negative Auswirkungen auf die Lernbereitschaft und den fachlichen Lernerfolg der SchülerInnen befürchten, können sich auf keine empirischen Daten stützen.

▲ Wer andererseits hofft, ohne zusätzliche Maßnahmen, d.h. allein durch die Verordnung von Verbalgutachten Lernbereitschaft und Lernerfolg von SchülerInnen verbessern zu können, wird durch die Befunde zur bisherigen Beurteilungspraxis und ihre Wirkungen ernüchtert. Ohne eine pädagogische und didaktische Öffnung des Unterrichts und ohne die Sicherung bestimmter Rahmenbedingungen bleibt eine Veränderung der Bewertung meist erfolglos.

▲ Diejenigen aber, die mit dem Verzicht auf Ziffernnoten pädagogische Ziele verfolgen, können mit einer Verbesserung der Unterrichtssituation und der Motivation der SchülerInnen sowie ihres Lernerfolgs rechnen, sofern sie bereit sind,
- als LehrerInnen sich auf den höheren, aber lohnenden Aufwand einzulassen,
- als Schulverwaltung die für Evaluation verfügbaren Ressourcen gezielter in die Fortbildung und Unterstützung der LehrerInnen zu investieren und
- als BildungspolitikerInnen den Selektionsdruck im System zu verringern und Rahmenbedingungen wie die Schüler-Lehrer-Relation zu verbessern.

# Langfassung

# Sind Noten nützlich - und nötig?

## Ziffernzensuren und ihre Alternativen im empirischen Vergleich

*Eine wissenschaftliche Expertise*
*des Grundschulverbandes*
*erstellt von der Arbeitsgruppe Primarstufe*
*an der Universität Siegen*

*von Hans Brügelmann mit*
*Axel Backhaus, Erika Brinkmann (Gast)*
*Hendrik Coelen, Thomas Franzkowiak*
*Simone Knorre, Barbara Müller-Naendrup*
*Elisabeth Oser, Sara Roth*

## Inhalt

*Notabene*
Ein zentrales Problem in der Diskussion ist die unklare Begrifflichkeit.
In dieser Expertise verstehen wir - sofern nicht ausdrücklich etwas anderes gesagt wird - unter »Noten« bzw. »Zensuren« *Ziffern*noten, die zur *formellen* Beurteilung verwendet werden, z.B. bei Klassenarbeiten oder in Zeugnissen.
»Leistungsbeurteilung« verwenden wir als Obergriff für die *Beschreibung* und die *Bewertung* von Leistungen - zwei unterschiedliche Formen ihrer Rückmeldung, die auch in der Umsetzung sorgfältig zu trennen sind.

## Vorweg: Resümees aus vier Perspektiven[1]

Wie bei allen pädagogischen Fragen (und sozialen Phänomene generell) ist die Befundlage zu Noten nicht auf einen einfachen Nenner zu bringen. Formen der Leistungsbeurteilung wirken unterschiedlich, je nachdem *wie* und in welchem *Kontext* sie eingesetzt werden. Für Folgerungen aus dem Forschungsstand kommt es deshalb darauf an, von welcher Basisannahme man ausgeht: Wer die Beweislast für Veränderungen bei den Reformern sieht, kann zu einer anderen Einschätzung kommen als jemand, der normativ die Förderung des Einzelnen als zentrale Norm und noch uneingelöste Aufgabe der Schule sieht. Vor diesem Hintergrund lässt sich als Ergebnis unserer Analysen festhalten:

▲ Wer an Ziffernnoten festhalten will, weil sie angeblich objektiv und vergleichbar seien bzw. erforderlich, damit SchülerInnen sich auf die Anstrengungen des Lernens einlassen, findet in der Empirie keine stützenden Belege für seine Position.

▲ Auch diejenigen, die Verbalgutachten ablehnen, weil sie angeblich negative Auswirkungen auf die Lernbereitschaft und den fachlichen Lernerfolg der SchülerInnen haben, können sich auf keine empirischen Daten stützen.

▲ Wer andererseits hofft, ohne zusätzliche Maßnahmen, d.h. allein durch die Verordnung von Verbalgutachten, Lernbereitschaft und Lernerfolg von SchülerInnen verbessern zu können, wird durch die Befunde zur bisherigen Beurteilungspraxis und ihre Wirkungen ernüchtert. Ohne eine pädagogische und didaktische Öffnung des Unterrichts und ohne die Sicherung bestimmter Rahmenbedingungen bleibt eine Veränderung der Bewertung meist erfolglos.

▲ Diejenigen aber, die mit dem Verzicht auf Ziffernnoten pädagogische Ziele verfolgen, können mit einer Verbesserung der Unterrichtssituation und des Lernerfolgs, vor allem der schwächeren SchülerInnen, rechnen - sofern sie bereit sind, als LehrerInnen einen höheren Aufwand zu leisten, als Schulverwaltung mehr in die Fortbildung und Unterstützung der LehrerInnen zu investieren und als BildungspolitikerInnen den Selektionsdruck im System zu verringern.

1 Diese Expertise geht zurück auf Vorarbeiten in Seminaren an der Universität Siegen (zum Teil auch publiziert) von Erika Brinkmann (2004; 2006), Hans Brügelmann (1980; 2000a+b; 2003a+b; 2005a, Kap. 27, 29, 56-60) und Barbara Müller-Naendrup (2005). Hilfreich waren auch die aktuellen Überblicke in: Valtin (2002a); Jachmann (2003, Kap.2); Bartnitzky/Speck-Hamdan (2004); Beutel (2005).

## 0

## Auftrag und Kontext der Expertise

»Rund 400 Millionen Zensuren werden jährlich in der Bundesrepublik Deutschland von den etwa 500.000 Lehrern in über 300.000 Schulklassen an die knapp zehn Millionen Schüler vergeben; jeder Schüler wird hierzulande also pro Jahr etwa 40-mal offiziell zensiert. In jeder Unterrichtsstunde ergehen an deutschen Schulen fast 300.000 und in jeder Minute an die 5.000 Noten.«[2]

Diese schon 20 Jahre alte Schätzung signalisiert unmissverständlich die hohe Bedeutung des Themas. Noten sind sowohl unter den Betroffenen als auch in Fachkreisen ein viel diskutiertes Thema. Ihr Nutzen war und ist heftig umstritten[3]. Angesichts der ungebrochen harten Auseinandersetzungen ein erstes frappierendes Ergebnis unserer Literaturrecherche: Zentrale empirische Befunde zur Problematik von Noten liegen seit 50 Jahren, zum Teil noch länger vor. Seit den 1970er Jahren sind diese Studien im deutschen Sprachraum vor allem von Ingenkamp (1971, 1975, 1981, 1989, 1991) in systematisierenden Überblicken publiziert worden. Ihre Befunde sind in der Zwischenzeit durch weitere Studien bestätigt und erneut mehrfach zusammengefasst worden[4]. Trotz dieser empirisch fundierten Kritik hat sich in der Wahrnehmung durch die Betroffenen (durch SchülerInnen, Eltern und LehrerInnen, aber auch durch die Öffentlichkeit) und erst recht im Schulalltag nur wenig verändert. Soweit es in den letzten Jahren Veränderungen gab (vor allem im Grundschulbereich[5]) sind eher Tendenzen zu beobachten, das Rad der Entwicklung zurückzudrehen, ja, die Noten über den Leistungsbereich hinaus auszuweiten[6].

Vor diesem Hintergrund hat der Grundschulverband die vorliegende Expertise in Auftrag gegeben. Es soll die Forschungsergebnisse zu Ziffernnoten und alternativen Formen der Leistungsbeurteilung sichten und bewerten. Im Fokus des Gutachtens steht die Grundschule. Viele der ausgewerteten Studien und auch viele unserer Überlegungen beziehen sich aber auf grundsätzliche Fragen der Leistungsbeurteilung und reichen deshalb über diese Schulstufe hinaus.

## 0.1

## Ansatz und Aufbau des Gutachtens

Es gab für uns zwei Optionen, diese Expertise zu erstellen[7]: entweder über eine statistische Meta- oder durch eine interpretative Sekundäranalyse der vorliegenden empirischen Studien.

Bei einer Metaanalyse werden die Daten verschiedener Studien zu übergreifenden Kennwerten *verrechnet.* Die Komplexität der Notenproblematik hätte in unserem Fall mehrere getrennte Metaanalysen erforderlich gemacht, um differenziertere Aussagen zu den verschiedenen Teilaspekten treffen zu können. Und selbst innerhalb solch eng gefasster Bereiche ist das Datenmaterial sehr heterogen[8]. Eine bloß statistische Verdichtung der Daten wäre kaum möglich gewesen. Ohne theoriebezogene Interpretationen wäre sie sehr oberflächlich und damit missverständlich geblieben.

Zudem täuscht der Anschein, als handele es sich bei Metaanalysen um rein technische Verrechnungen, generell. Zwar ist dieses Instrument methodisch inzwischen gut etabliert[9]. Aber es gibt eine Reihe von Einschränkungen, die bei seiner Nutzung zu bedenken sind: »In Metaanalysen werden die statistischen Daten verschiedener Studien nach explizit definierten Kriterien miteinander verrechnet. Damit wird der Anspruch erhoben, den Einfluss der jeweils besonderen Kontextbedingungen in den Einzelstudien sowie den persönlichen Einfluss der AuswerterInnen zu reduzieren. Ausschalten lässt sich das subjektive Moment aber auch hier nicht. Es kommt bereits zum Tragen bei der Entscheidung über die anzulegenden Kriterien, wenn die Frage ansteht, welche Studien überhaupt als forschungsmethodisch adäquat in die geplante Metaanalyse einbezogen werden sollen. Denn für die ›methodische Qualität‹ von Studien gibt es unterschiedliche Maßstäbe. Auch für die Alternativen, ob man die berücksichtigten Studien einzeln zählt, also ihre Ergebnisse gleichgewichtig verrechnet, oder ob man die Kennwerte nach der Zahl der jeweils in der Studie untersuchten Fälle gewichtet, gibt es jeweils gute Gründe. Am stärksten kommt die persönliche Position der WissenschaftlerInnen, die die Metaanalyse durchführen, in der verbalen Zusammenfassung der Rechenergebnisse zum Ausdruck. An dieser Stelle wird notwendigerweise fokussiert, gewichtet, geglättet, gedeutet - denn Zahlen sprechen nicht für sich.«[10]

Angesichts dieser Einschränkungen und der begrenzten zeitlichen und finanziellen Ressourcen haben wir uns dafür entschieden, die vorliegenden Studien in Form einer Sekundäranalyse *interpretativ* zusammenzufassen. Für dieses

2 Mreschar (1985, 41).

3 Vgl. u.a. die Pro & Contra-Diskussionen von Ramseger (1993a+b) vs. Schröter (1993); Einsiedler vs. Schöll (1995); Herrmann (2003) vs. Brügelmann (2003); Brügelmann (2005) vs. CDU-Bremen (2005), vgl. dazu auch Wolschner (2005).

4 Vgl. u.a. Zielinski (1974a+b); Becker/Hentig (1983); Bartnitzky/Portmann (1992); Oelkers (2001) und die deutliche Kritik von PädagogInnen aus der Grundschulpraxis, etwa bei Bolscho u.a. (1979); Bartnitzky/Christiani (1987); Schmitt (1999, 137 ff.).

5 Da aber auch schon viel früher, vgl. Petersen (1974) zum »Jenaplan« von 1927, die Abschaffung der Noten in den Waldorfschulen - sogar bis Klasse 12 - und bei den Freineit-PädagogInnen ebenfalls in den 1920er Jahren.

6 S. zur Debatte über Kopfnoten die Hinweise > Kap. 0.4.

7 Vgl. zum Folgenden ausführlicher: Brügelmann/Heymann (2006).

8 S. die Aufschlüsselung in > Kap.1 bis 4.

9 Vgl. Glass (1976; 1977); Hunter u.a. (1982/2004); für den deutschen Sprachraum: Fricke/Treinies (1985).

10 Brügelmann/Heymann (2006, 2-3).

Vorgehen spricht auch der unterschiedliche Status der einbezogenen Untersuchungen. Sie reichen von Laborversuchen über Feldexperimente bis hin zu Beobachtungen und Befragungen unter nicht kontrollierten Bedingungen.

Das Problem einer solchen *research synthesis* ist, dass die Auswahl, Ordnung und Deutung der Forschungsergebnisse in noch höherem Maße von den Personen abhängt, die die Sichtung vornehmen, als bei einer Metaanalyse. Wir haben uns bemüht, den Prozess der Verdichtung durchsichtig und nachvollziehbar zu halten[11] - weshalb neben der Kurzform mit den zentralen Ergebnissen des Gutachtens zusätzlich diese sehr ausführliche Darstellung der Befunde und ihrer Würdigung publiziert wird. Außerdem haben wir versucht, innerhalb des Gutachtens Analyse (> Kap. 1 bis 5) und Folgerungen (> Kap. 6 und 7) möglichst deutlich zu trennen. Im analytischen Teil sind deshalb auch widersprüchliche Befunde repräsentiert. Durch die Beteiligung von neun Personen an ihrer Sichtung und Bewertung und durch die intensiven team-internen Diskussionen wurde schon ein hohes Maß an sozialer Kontrolle persönlicher Sichtweisen erreicht. Zusätzlich haben wir eine Vorfassung dieses Gutachtens auswärtigen ExpertInnen zur Kritik vorgelegt. In beiden Fällen sind substanzielle Differenzen im Gutachten selbst dokumentiert. Als Validierung unseres Vorgehens werten wir den hohen Deckungsgrad unseres Resümees der deutschsprachigen Literatur mit den Ergebnissen und Folgerungen neuerer *reviews* aus dem angelsächsischen Bereich[12].

Menschliche Erkenntnis- und Urteilskraft ist immer begrenzt. Insofern ist es einfach, die Schwächen einer jeden Form von Leistungsbeurteilung nachzuweisen, wenn man sie nur für sich betrachtet. In unserem Gutachten haben wir deshalb Potenzial und Grenzen von Ziffernnoten im Vergleich untersucht.

Noten sind seit langem umstritten. Als Alternative wurden und werden Verbalzeugnisse empfohlen. In der Gegenüberstellung dieser beiden Formate werden allerdings verschiedene Argumentationsebenen vermischt. Damit wird die Klärung der Titelfrage erschwert. So werfen Ziffernnoten sehr unterschiedliche Probleme auf. Drei Entscheidungsfragen mit je besonderen Problemen sind zu unterscheiden:

▲ die Wahl der Verfahren zur *Feststellung* des Lernerfolgs (informelle vs. standardisierte Aufgaben, offene vs. strukturierte Beobachtung)
▲ die Wahl der Bezugsnorm zur *Bewertung* des Lernerfolgs (nach Annäherung an das Lernziel und/oder individuellem Lernforschritt und/oder relativer Leistungsposition in einer Gruppe)
▲ die Wahl der *Darstellungsform* in der Rückmeldung (Beschreibung vs. Bewertung, freie Formulierung vs. Ziffern).

Diese drei Aspekte werden im Schul- und Berufsalltag in unterschiedlichen Kombinationen realisiert. Dabei sind die gängigen Muster nicht sachlich zwingend. Insofern sind im Folgenden mehrere Teilfragen sorgfältig zu trennen:

*In welcher Funktion werden Leistungen beschrieben und bewertet*
*(> Kap. 0.3 und 7)*
Leistungen können im Blick auf einen festzustellenden Förderbedarf beurteilt werden oder auch, um den Unterricht zu verbessern. Im deutschen Schulsystem dominieren dagegen die Selektions- und Disziplinierungsfunktion. Dieser institutionelle Kontext prägt die Wirkung von Noten - und schränkt die Möglichkeiten alternativer Beurteilungsformen ein.

*Über welche Verfahren werden Leistungen erfasst?*
*(> Kap. 1)*
Noten wird vorgeworfen, sie seien nicht objektiv, nicht valide und nicht zuverlässig. Diese Probleme haben aber auch Verbalzeugnisse. Beider Datengrundlage ist an die Person der Beurteilenden und ihre Auswahl der Instrumente zur Erhebung von Leistungen gebunden. Insofern sind als Alternative zu Klassenarbeiten und informellen Beobachtungen standardisierte Tests und strukturierte Beobachtungen zu diskutieren.

*Anhand welcher Maßstäbe werden Leistungen bewertet?*
*(> Kap. 2)*
Noten wird eine einseitige Orientierung an der sozialen Bezugsnorm - mit der jeweiligen Schulklasse als dominierendem Maßstab - vorgeworfen. Diese Verbindung ist aber nicht zwingend. Noten können sich auch am Lernfortschritt oder an den Anforderungen orientieren (und sollen dies sogar, vgl. bereits KMK 1968). Umgekehrt orientieren sich auch Verbalzeugnisse nicht zwangsläufig an der individuellen Entwicklung. Die Bedeutung und die Wirkungen unterschiedlicher Maßstäbe für die Bewertung von Leistungen sind also übergreifend zu klären.

*In welchen Formen werden Leistungsbeurteilungen dargestellt?*
*(> Kap. 3)*
Erst auf dieser Stufe geht es um Ziffernnoten vs. sprachliche Formulierungen. Dabei interessieren vom Gutachtenauftrag her zwei Fragen:

▲ Werden die Ansprüche der beiden Zeugnisformen in der praktischen Umsetzung tatsächlich eingelöst? (> Kap. 3.1)
▲ Welche Wirkungen haben verschiedene Rückmeldeformate auf den Unterricht bzw. auf die Entwicklung der SchülerInnen (Erfüllung verschiedener Erwartungen/ Funktionen und etwaige negative Nebenwirkungen)? (> Kap. 3.2)

---

11 Allerdings war es uns angesichts des zeitlich und finanziell knappen Rahmens auch nicht möglich, die einzelnen Studien ähnlich systematisch zu bewerten, wie das etwa nach den *Guidelines des* »Evidence for Policy and Practice Information and Coordinating Centre (EPPI- Centre)« gefordert wird, vgl. u.a. Harlen/Deckin Crick (2002, 19-29); Harlen (2004, 22-32); Newman u.a. (2004).

12 Vgl. Harlen/Deakin Crick (2002); Harlen (2004a+b).

*Wie werden verschiedene Zeugnisformen wahrgenommen? (> Kap. 4)*
Wer Beurteilungsformen ändern will, muss deren Akzeptanz und etwaige politische Vorbehalte bzw. persönliche Bedenken und Ängste kennen. Unabhängig von den empirisch festgestellten Stärken und Schwächen verschiedener Formate geht es darum, wie die Beteiligten selbst die Leistungsfähigkeit unterschiedlicher Darstellungsformen einschätzen. Im Vordergrund steht die Frage, wie gut die Formate die unterstellten Funktionen - nach Einschätzung verschiedener Gruppen - erfüllen (insbesondere Informationsgehalt und Verständlichkeit der Rückmeldung).

*In welchem Verhältnis stehen Aufwand und Ertrag verschiedener Darstellungsformen? (> Kap. 5)*
Die Chancen von Reformen hängen schließlich auch davon ab, dass sie von den Beteiligten nicht nur als inhaltlich wichtig, sondern auch als praktikabel, zumindest aber nicht als unergiebige zusätzliche Belastung wahrgenommen werden.

## 0.2
## Datengrundlage des Gutachtens

Unsere Literatursuche in internationalen Datenbanken zu Stichworten wie »assessment«, »marking« und »grading« war wenig ergiebig[13]. Auf den ersten Blick überrascht dies, werden doch in den angelsächsischen Ländern alle denkbaren Aspekte von Unterricht immer wieder empirisch untersucht. Zu bedenken ist aber, dass die Notendiskussion nicht überall den gleichen Stellenwert (mehr) hat[14]. In vielen westlichen Industrieländern besteht eine langjährige Tradition, für die Leistungsbeurteilung standardisierte Tests einzusetzen. Zum Teil ist dies bereits eine Folge der empirisch begründeten Kritik an Noten vor und nach dem zweiten Weltkrieg[15]. Die Probleme sind mit dem Einsatz von standardisierten Verfahren allerdings nicht weniger geworden, wie die vehemente Testkritik der letzten Jahre, vor allem in den USA[16] zeigt. Inzwischen gibt es sogar wieder die Tendenz, den Beobachtungen und Einschätzungen von LehrerInnen ein stärkeres Gewicht bei der Leistungsbeurteilung einzuräumen[17]. Eine gezielte Suche über SchlüsselautorInnen in dieser kritischen Diskussion über *assessment* hat dann auch einige interessante Überblicke, die sich auf die Beurteilung durch LehrerInnen einerseits und durch standardisierte Tests andererseits beziehen, zu Tage gefördert[18].

Vor allem angesichts der Unterschiede zwischen den Schulsystemen (> Kap. 0.5) und zwischen den kulturellen Normen verschiedener Länder bedürfen die Ergebnisse ausländischer Studien sowieso einer sorgfältigen Interpretation und lassen sich nur mit Einschränkungen von einem nationalen Kontext auf einen anderen übertragen. Eine zentrale Differenz ist etwa die unterschiedliche Dauer der gemeinsamen Schulzeit (mit entsprechend frühen oder späten Selektionsanforderungen). Aber auch die Häufigkeit von Zurückstellungen am Schulanfang, von Überweisungen in Sonderschulen und von Klassenwiederholungen variiert erheblich zwischen verschiedenen Ländern (> Kap. 0.4).

Abgesehen von einigen grundlegenden Untersuchungen aus Großbritannien und den USA und den bereits erwähnten Metaanalysen und *reviews* konzentrieren wir uns in diesem Gutachten deshalb auf Studien aus den deutschsprachigen Ländern mit ihren noch vergleichsweise ähnlichen Schulsystemen. Erfreulicherweise gibt es aus den letzten fünf bis zehn Jahren eine Reihe relevanter und methodisch fundierter Vergleichsuntersuchungen zu Ziffernnoten und Berichtszeugnissen.

Wegen ihrer breiten Anlage innerhalb des Bereichs Leistungsbeurteilung sind die folgenden Studien besonders bedeutsam:

▲ Das Projekt NOVARA[19] (s. zum Projektdesign Valtin 1999; 2002d) untersuchte in der Umbruchphase nach der Wende im Vergleich von 41 Ost- und West-Berliner Klassen
- die Akzeptanz der Verbalbeurteilung bei LehrerInnen, Kindern und Eltern;
- die Realisierung der Ansprüche in den Beurteilungen und
- die Auswirkungen auf zentrale Persönlichkeitsmerkmale und ausgewählte Fachleistungen der SchülerInnen.

Der Längsschnitt wurde im Projekt SABA[20] fortgeführt bis zur sechsten Klasse, in Berlin Abschluss der Grundschulzeit. In einer Teilstichprobe (NOVUS[21]) wurde darüber hinaus der Zusammenhang zwischen den Beurteilungsformen und dem Unterricht selbst untersucht.

---

13 So liefert das europäische Datenbanksystem *eurydice* zwar einen vergleichenden Bericht zur Evaluation von Schulen, aber nicht zur Leistungsbeurteilung von SchülerInnen > http://www.eurydice.org/Doc_intermediaires/analysis/de/frameset_analysis.html [Abruf: 19.2.2006]; auch die Datenbank des Deutschen Instituts für Internationale Pädagogische Forschung erbrachte kaum verwertbares Material, vgl. http://www.dipf.de/datenbanken/ines/IZB_bildungweltweit_ines.htm [Abruf: 23.2.2006].

14 Die letzte Metaanalyse stammt von Fraser u.a. (1987).

15 Vgl. etwa die Übersetzungen klassischer Studien in Ingenkamp (1971).

16 Vgl. vor allem Kohn (1999; 2000), der sowohl die Wirkung von Noten als auch von extern verordneten Tests kritisiert; s.a. Nichols u.a. (2006).

17 S. zur Diskussion in den USA etwa Hiebert/Davinroy (1993, 1-4) und in Großbritannien Black/Wiliam (1998); Freitag (2001); Harlen (2004b, 4-7, 33-71).

18 Sehr hilfreich waren die *reviews* in: Fuchs/Fuchs (1986); Crooks (1988); Weston (1991); Black/Wiliam (1998); Kohn (1999; 2000); Deci u.a. (1999); Stiggins (1999); Linn (2000); Harlen/Deakin Crick (2002); Harlen (2004a+b).

19 NOVARA = »Noten- oder Verbalbeurteilung? Akzeptanz, Realisierung und Auswirkungen«; vgl. zu einzelnen Aspekten der Studie die Beiträge zu Valtin (2002a) und die Einzelarbeiten von Schmude (2001); Wagener (2003); Thiel (2005).

20 SABA = »Schulische Adaptation und Bildungsaspiration«.

21 NOVUS = »Noten oder Verbalbeurteilung: Unterrichtsorgansiation und Sanktionsverhalten von Lehrkräften in Ost- und Westberliner Grundschulen«.

▲ Eine Projektgruppe[22] um Lütgert und Tillmann hat das Hamburger Projekt »LeiHS«[23], zum Teil im Vergleich mit Untersuchungen in Thüringen (»KomThü«)[24], durchgeführt und ausgewertet. Beiden Projekten ging es darum, »diagnostisch anspruchsvolle Formen der Rückmeldung schulischer Leistungen an die Lernenden und ihre Eltern entwickeln oder ausdifferenzieren zu wollen«[25]. In Hamburg wurde die Anwendung bestehender Instrumente (z.B. von Notenzeugnissen mit Kommentarbogen und Notenzeugnissen mit Bemerkungen zum Arbeits und Sozialverhalten) durch Befragungen von 1.476 SchülerInnen der Sekundarstufe, 61 Kindern der Grundschule sowie 1.328 Eltern und 637 LehrerInnen beider Schulstufen evaluiert. In die Thüringer Studie gingen Fragebögen von 925 Schülerinnen und Schülern, 1019 Eltern, 295 LehrerInnen und eine qualitative Befragung von 235 Grundschulkindern sowie eine Dokumentenanalyse ausgewählter Zeugnisse ein. Die beiden oben skizzierten Studien bieten mit ihrem Datenmaterial und Einzugsgebieten den Forschungskontext für die eigentliche Frage einer umfassenden Studie von Iris Beutel: Können Kinder Experten ihrer eigenen Leistung sein?[26]

▲ Im Modellversuch »Lern- und Spielschule« in Rheinland-Pfalz wurden u.a. Verbalzeugnisse bis Klasse 4 erprobt, am Ende ergänzt um ein Ziffernzeugnis als Anlage. Aus dieser Längsschnittstudie wurden SchülerInnen, Eltern und LehrerInnen in 15 Versuchs- und 7 Kontrollklassen (329 bzw. 157 Kinder) zu ihren Erfahrungen und Einschätzungen befragt; ergänzend wurden 468 Zeugnisse am Ende der 3. und der 4. Klasse, also von 234 SchülerInnen, analysiert[27].

Spezifische Untersuchungen zur Kontroverse um die Aussagekraft von Ziffern- vs. Verbalbeurteilung haben - insbesondere in Form von Zeugnisanalysen - Schmidt (1981), Benner/Ramseger (1985), Scheerer u.a. (1985), Freese (1990), Ulbricht (1993), Haenisch (1996a+b), Lübke (1996), Jürgens (1997; 1998b) und Döpp u.a. (2002) vorgelegt.

Daneben konnten wir auf eine Fülle von Untersuchungen zurückgreifen, die zu spezifischen Aspekten wie der Aufsatzbeurteilung, mündlichen Prüfungen oder den Erwartungen und Einschätzungen verschiedener Zielgruppen durchgeführt worden sind, wobei für viele Fragen die Aufsatzsammlung von Ingenkamp (1971) nach wie vor grundlegend ist.

Am Rande spielt die Frage der Leistungsbeurteilung, der Aussagekraft von Noten und der Organisation von Prüfungen auch eine Rolle in Studien, die einen anderen Fokus hatten, wie LAU in Hamburg, KILIA in Bayern und LUST in Nordrhein-Westfalen sowie in den internationalen Vergleichsstudien TIMSS, PISA und IGLU. Vor allem aus diesen Untersuchungen lassen sich Anhaltspunkte über die Bedeutung spezifischer Kontextfaktoren gewinnen.

## 0.3

## Historischer Rückblick und gesellschaftlicher Kontext[28]

Die Verwendung von Ziffernoten ist eng verknüpft mit der Einführung von Zeugnissen, die sich zunächst an den weiterführenden Schulen und erst später an den Volksschulen etabliert haben, und zwar teilweise mit unterschiedlichen Intentionen[29]:

»Als Entstehungszeit des Schulzeugnisses ergibt sich das 16. Jahrhundert, als schulischer Ursprungsort die höhere Schule. In der Elementarschule findet das Schulzeugnis erst Aufnahme nach Einführung der allgemeinen Schulpflicht. Die Urfunktion des Schulzeugnisses der höheren Schule ist die Auslesefunktion, diejenige des Zeugnisses der Elementarschule die Kontrollfunktion, und zwar im Hinblick auf den Schulbesuch und damit die Erfüllung der Schulpflicht.«[30]

Dabei steht diese Entwicklung in engem »Zusammenhang mit der Säkularisierung und Verstaatlichung des Schulwesens sowie der Ausdehnung von Schule auf breite Bevölkerungskreise. Erst im Laufe des 19. Jahrhunderts hatten sich Ziffernzensuren als unhinterfragbarer Maßstab und als Ausdruck gängiger Leistungsbeurteilung etabliert.«[31]

Zeugnisse sollten Fähigkeiten ausweisen, um die Vergabe von Berufspositionen an Leistung statt an Herkunft zu binden. Das Leistungsprinzip stellte somit einen großen Fortschritt dar gegenüber dem Abstammungsprinzip der feudalen Gesellschaft - zumindest galt das für das Bürgertum gegenüber dem Adel. Denn die Bindung an Zeugnisse warf zwei neue Probleme auf, die normalerweise nur von privilegierten Schichten zu überwinden waren:

▲ Es reichte nicht mehr, etwas zu können - dieses Können musste auch durch Prüfungen nachgewiesen werden. Diese bedeuteten nicht nur eine zusätzliche Hürde auf dem Weg in den Beruf; es stellte sich auch die Frage nach ihrer Aussagekraft für die spätere Bewährung im Beruf[32].

---

22 Vgl. die Beiträge zu Beutel u.a. (1999; 2000); Beutel/Vollstädt (2000); Jachmann (2003); Tillmann/Vollstädt (1999).

23 LeiHS = »Leistungsbeurteilung und Leistungsrückmeldung an Hamburger Schulen«.

24 »KomThü« = »Einschätzung zur Kompetenzentwicklung«, vgl. Beutel (2000; 2002; 2004/2005).

25 Beutel (2005, 118); Lütgert/Tillmann (2000, 8); Jachmann/Tillmann (2000).

26 Beutel (2004/2005).

27 Vgl. Maier (2001; 2003); Petillon (2001).

28 Dieses Kap. geht auf einen Eigenbeitrag von Barbara Müller-Naendrup zurück. Vgl. auch Arzberger (1988) und die Überblicke bei Dohse (1967); Breitschuh (1979); Ziegenspeck (1999); Fiegert (2001); Maier (2001, 17-18); Huber (2002) und Jung (2005, 63-66).

29 Vgl. Breitschuh (1979, 58).

30 Dohse (1967, 40).

31 Vgl. Beutel (2005, 44, 46).

32 Vgl. Breitschuh (1979, 49), der in diesem Zusammenhang auch von der »Schule als Statuszuweiser« spricht.

▲ Prüfungen waren zudem an das Absolvieren institutioneller Bildungswege gebunden. Diese Hürde verursachte mehrfache Kosten: früher (und teilweise heute noch) ein Schulgeld, darüber hinaus den Unterhalt für den Schüler und seine Lernmittel, vor allem aber den Verzicht auf seinen Beitrag zum Familieneinkommen.

Das Prüfungs- und Zeugniswesen konnte vom Bürgertum auch als Mittel zur Ausgrenzung durch Selektion genutzt werden - jetzt gegen die Arbeiterschicht. Gleichzeitig legitimierte (und legitimiert) das Leistungsprinzip gesellschaftliche Ungleichheit, kann sie doch als Folge unterschiedlicher Fähigkeiten gerechtfertigt werden. Diese Deutung aber ist Ideologie[33]. Aktuell belegen dies Studien zur Elitenbildung in der Bundesrepublik Deutschland[34]. Sie zeigen, dass AbsolventInnen aus höheren sozialen Schichten deutlich bessere Chancen haben, in Führungspositionen zu gelangen als BewerberInnen aus Mittel- oder Unterschicht - bei gleichen Abschlussnoten.

Mit ihrem Anspruch, als »standesunabhängige Beurteilungsgröße schulischen Lernens« zu dienen, haben sich Zensuren erst Mitte des 19. Jahrhunderts etabliert. Ihre Durchsetzung in den Schulen ist eng verbunden mit der Realisierung des Jahrgangsklassenprinzips:[35] »Das Jahrgangsklassensystem, das um 1840 mit seinen wichtigsten Merkmalen, der jahrgangsweisen Einschulung, der jährlichen Versetzung nach dem Leistungsstand in allen Fächern, dem verbindlichen Fächerkanon, der Festlegung von Wochenstundenzahlen und Stoffverteilungen, für die höheren Schulen in Preußen gegen viele Widerstände ministeriell verordnet wurde, muß sogar als unentbehrliche Voraussetzung für den Ausbau des Berechtigungswesen angesehen werden.«[36]

Leistungsbeurteilungen in dieser Form wurden durch die bürokratischen Abläufe nachvollziehbarer und durchschaubarer. Schon kurz nach der Einführung des Zensurensystems entwickelt sich eine Kritik an diesem Beurteilungsverfahren, die mit unterschiedlichen Nuancen bis heute währt. Verbale Beurteilungsformen, die in der historischen Entwicklung der Leistungsbeurteilung weitaus früher etabliert waren als Ziffernnoten, werden vielfach als alternative Beurteilungsformen vorgeschlagen und eingesetzt. »Das Instrument der Zensurengebung war nicht dafür entwickelt worden, den in der Weimarer Verfassung festgelegten Auftrag zu erfüllen, dass für die weiterführende Schulbildung eines Kindes, seine Anlagen und Neigungen, nicht die wirtschaftliche oder gesellschaftliche Stellung oder das Religionsbekenntnis seiner Eltern maßgebend' sein sollten.«[37]

In diesem Zusammenhang ist auf die kritischen Impulse und Gegenvorschläge der reformpädagogischen Bewegung zu Beginn des 20. Jahrhunderts hinzuweisen, die sich bei aller Unterschiedlichkeit der Ansätze einig war in der »Ablehnung der Ziffernzensuren« und in der Unterstützung für eine verbale Beurteilung, die den Lernprozess des Individuums ins Zentrum rückt.[38]

Insofern sind Lernberichte »keine Erfindung der neueren Schulreform der 70er Jahre des 20. Jahrhunderts«, sondern »eine Frucht einer Stärkung des Lernenden zugewandten Pädagogik« in den 1920er Jahren[39].

In den 1970er Jahren wurde die Einführung des Berichtszeugnisses[40] auch von kultusadministrativer Seite vehement vorangetrieben[41], ja zum Teil sogar »von oben« verordnet. In dieser Zeit ging es in der pädagogischen Diskussion um diese Beurteilungsform »vorrangig um die Technik des Berichtsschreibens«[42]. Diese Entwicklung ist, wie eine Analyse der Forschungslage zeigt (> Kap. 1 ff.) in vielerlei Hinsicht kritisch zu betrachten. »Das erklärte Ziel der Reformer der Ziffernzensur, an die Stelle der nüchternen Zahl das erklärende Wort zu setzen, erwies sich in jeder Hinsicht als voraussetzungsreich.«[43]

Denn Leistungsbeurteilungen haben verschiedene Funktionen zu erfüllen, differierend vor allem in (extern-)gesellschaftlicher und (intern-)pädagogischer Perspektive[44]. Daraus ergeben sich unterschiedliche Anforderungen - und jeweils spezifische Probleme[45]:

a) *Motivationsfunktion:* Durch Beurteilungen sollen SchülerInnen angehalten werden, sich den schulischen Anforderungen zu stellen (»Erhöhung der Lernbereitschaft«) und dadurch bessere Leistungen zu erbringen (»Steigerung des Lernerfolgs«). Vor allem den Noten wird vorgeworfen, dass die unterstellte Normalverteilung die Hälfte der Kinder von vornherein zum Verlieren verurteilt und damit zumindest diese Gruppe demotiviert. Aber auch leistungsstarke SchülerInnen könnten beim Kampf um Notenzehntel unter einen leistungsmindernden Stress geraten - z.B. bei Übergangsprüfungen für die weiterführenden Schulen. Verbalgutachten stehen dagegen im Verdacht, sie beschönigten Leistungs-

33 Vgl. Herrlitz u.a. (1998, 36), die auf die schon im Zuge der preußischen Bildungsreform »›systematisch‹ produzierte(n) Ungleichheiten neuer Qualität« hinweisen und die »Durchsetzung von ›Zensuren‹ als legitime Ordnungsschema für abgestufte Teilhabechancen und für den sozialen Ausschluß« bezeichnen.

34 Vgl. vor allem Hartmann (2002).

35 Vgl. Dohse (1963; 1971, 39); Ingenkamp (1995, 49); Beutel (2005, 13).

36 Vgl. Ingenkamp (1995, 49).

37 Vgl. Ingenkamp (1995, 50).

38 Vgl. z.B. Key (1992, 179-180) und Beutel (2005, 44-51). Beutel zieht hier als Beispiele Bertold Otto, Hugo Gaudig, Peter Petersen und Rudolf Steiner heran.

39 Vgl. Beutel (2005, 41, 26).

40 In der Sekundarstufe waren es »Diagnosebögen« mit differenziert aufgeschlüsselten Unterkategorien, die zu einer differenzierteren Beurteilung von Leistungen verhelfen sollten.

41 Vgl. Deutscher Bildungsrat (1970); Beutel (2005, 56-57); Rodehüser (1987, 661). Rodehüser (1987, 31) verweist in einem Schaubild auf den Beginn einer differenzierten und individualisierenden Leistungsbeurteilung ab etwa 1968 (mit der Verselbständigung der Grundschule). Vgl. dazu auch KMK (1970, 33) und die Kritik an Noten von Ingenkamp (1969) auf dem ersten Bundesgrundschulkongress.

42 Beutel (2005, 234).

43 Beutel (2005, 234).

44 Vgl. ausführlicher die Zusammenfassung bei Tillmann/Vollstädt (1999).

45 Vgl. für viele Zielinski (1974b, 881-882).

schwierigkeiten, so dass SchülerInnen der Antrieb fehle, an ihren Schwächen zu arbeiten.

b) *Rückmelde- und Berichtsfunktion:* Über die Beurteilungen sollen SchülerInnen und ihren Eltern Hinweise auf den Lernstand der Kinder bzw. Jugendlichen und evtl. Probleme erhalten. Genau dies könnten Noten nicht leisten, behaupten KritikerInnen, weil sie unterschiedliche Teilleistungsprofile pauschal in einer Ziffer zusammenfassen und weil dieselbe Leistung ganz verschiedene Ursachen haben könne (hohe/niedrige Ausprägung von z.B. Begabung, Vorwissen, Fleiß oder externer Unterstützung). Verbalgutachten dagegen wird vorgehalten, sie seien für SchülerInnen und Eltern oft nicht verständlich und vor allem im Gegensatz zu Noten nicht eindeutig.

c) *Ausweisfunktion:* Analog zur internen Rückmeldung sollen Beurteilungen auch Außenstehenden helfen, ein zuverlässiges Bild von den Fähigkeiten einer Bewerberin oder eines Bewerbers zu gewinnen. Insofern gelten hier dieselben Einwände wie unter (b) - verstärkt durch den Vorwurf, dass dieselbe Leistung in verschiedenen Klassen ganz unterschiedlich bewertet werde.

d) *Selektions- und Zuteilungsfunktion:* Innerhalb des Bildungssystems sollen Noten durch die Klassifikation nach Leistung Auswahlentscheidungen stützen - z.B. bei Versetzungen (vs. Sitzenbleiben), bei der Feststellung eines sonderpädagogischen Förderbedarfs oder bei der Zuweisung zu den Schulformen der Sekundarstufe I. Vorbehalte beziehen sich auf die diagnostische Aussagekraft und auf die prognostische Sicherheit von Leistungsbeurteilungen.

e) *Sozialisierungs- und Disziplinierungsfunktion:* In der Schule begegnen die Kinder anderen Anforderungen an ihr Leistungs- und Sozialverhalten als in der Familie und in den informellen Interaktionen des Alltags. Zumindest von ihrem Anspruch her sehen Leistungsbeurteilungen von persönlichen Besonderheiten ab und suggerieren eine sachbezogene, neutrale Bewertung von Wissen und Können. Dies sei wichtig, um Kinder und Jugendliche auf entsprechende Anforderungen im stärker formalisierten öffentlichen Raum einzustellen. Andererseits würden Leistungsbeurteilungen missbraucht, um schulische Anforderungen durchzusetzen und abweichendes Verhalten zu sanktionieren.

Schon diese kurze Skizze lässt vermuten, dass die Anforderungen verschiedener Funktionen leicht in Konflikt miteinander geraten und dass verschiedene Formen der Beurteilung die eine oder die andere Funktion besser erfüllen können. Diese Probleme werden vor allem in den > Kap. 2 (Bezugsnormen der Bewertung), 3 (Wirkungen auf Unterricht und SchülerInnen) und 4 (Einschätzung durch die Betroffenen) genauer untersucht werden müssen.

Fazit: Aus historischer und soziologischer Perspektive ist der Anspruch von Noten, das Leistungsprinzip im gesellschaftlichen Wettbewerb um attraktive Positionen durchzusetzen, ambivalent einzuschätzen. Dennoch: Die traditionellen Formen der Leistungsurteilung (vor allem Ziffernzensuren) haben sich bis heute durchgesetzt und nicht nur in unserem Schulsystem, sondern auch in den »Köpfen der Gesellschaft« fest etabliert.[46]

Seit einiger Zeit wird in der Diskussion um Formen der Leistungsbeurteilung allerdings die wechselseitige Beziehung zwischen der Gestaltung des Unterrichts und den für sie notwendigen und möglichen Beurteilungsformen thematisiert. Dabei werden auch Verbalbeurteilungen als unzureichend kritisiert. Alternativ sollten Formen der Evaluation etabliert werden, »die einer neuen Lernkultur dienlich« sein können; das Spektrum an Erhebungs- und Bewertungsverfahren müsse durch Elemente wie das Portfolio, das Lerntagebuch, die Präsentation usw.[47] bereichert werden.

## 0.4

## Die Situation in den Bundesländern: ein Überblick[48]

Ziffernoten sind selbst in der Grundschule weithin Standard. Verbalgutachten beschränken sich in der Regel auf Klasse 1 und 2. Die Möglichkeit, ihre Anwendung durch Beschluss der Klassen- oder Schulkonferenz auf Klasse 3 und 4 auszudehnen, ist schon in der Vergangenheit nur von einer Minderheit der LehrerInnen und Eltern genutzt worden (s. dazu > Kap. 3.1 und 4). Gegenwärtig werden diese Ausnahmen auch rechtlich weiter eingeschränkt. Außerdem wird der Beginn der Notengebung in mehreren Bundesländern auf Klasse 2 vorverlagert; Ausnahmeregelungen werden zunehmend restriktiv gehandhabt, wie aktuell das Beispiel Bremen zeigt[49].

Auch am Beispiel Frühenglisch lässt sich studieren, wie die anfänglich geplante Notenfreiheit mehr und mehr abgeschafft wurde und teilweise - in Nordrhein-Westfalen 2007 - sogar die Versetzungsrelevanz eingeführt wird[50]. Und dies, obwohl Kinder, Eltern und LehrerInnen in Bundesländern, die an Verbalgutachten festgehalten haben, Noten zu jeweils zwei Dritteln bis drei Vierteln ablehnen, wie eine aktuelle Studie von Gompf/Henrich (2005) zeigt.

Die aktuellste Übersicht über die Situation in den Bundesländern stammt von Müller (2005, 94-97)[51]. Sie zeigt für alle Bundesländer, wie sich die verschiedenen Zeugnisformate auf die Jahrgangsstufen verteilen:

---

46 Vgl. dazu auch die Entwicklung in der ehemaligen DDR nach der Wende 1989/90, dargestellt bei Döbert/Geißler (2000).

47 So z.B. Winter (2004, 30 und 185 ff.).

48 Vgl. den Überblick bei Müller (2005); s. zum Vergleich Reimers (1991).

49 Im Jahr 2005 wurden 26 Anträge gestellt ( 24 Bremen/2 Bremerhaven). Von der Behörde wurden im September 17 (15+2) für akzeptabel erklärt. Der Bildungdsdeputation wurden am 22.12.05 7 (5+2) Schulen vorgeschlagen. Keine Schule wurde akzeptiert. Allerdings wurde beschlossen, das Thema weiterhin zu behandeln (pers. Mitteilung des Zentralelternbeirats Bremen per Mail v. 5.1.2006).

50 Vgl. Minker (2005); Gompf/Henrich (2005).

51 Vgl. als ein Beispiel für die rechtliche Durchregulierung der Notengebung: Kultusministerium Baden-Württemberg (2004).

| | Jahrgangsstufe 1 · Erstes Halbjahr | Jahrgangsstufe 1 · Zweites Halbjahr |
|---|---|---|
| Baden-Württemberg | | Schulbericht |
| Bayern[2] | Berichtszeugnis | Berichtszeugnis |
| Berlin | | Berichtszeugnis |
| Brandenburg | Elterngespräch | Lernentwicklungsbericht |
| Bremen | | Lernentwicklungsbericht<br>oder mündliche Information<br>(Beschluss durch Mehrheit der Schulkonferenz) |
| Hamburg | Information der Eltern in »geeigneter« Weise | Lernentwicklungsbericht |
| Hessen | | Berichtszeugnis und Elterngespräch |
| Mecklenburg-Vorpommern | Lernentwicklungsbericht | Lernentwicklungsbericht |
| Niedersachsen | Berichtszeugnis | Berichtszeugnis |
| Nordhein-Westfalen | | Berichtszeugnis |
| Rheinland-Pfalz | | Berichtszeugnis |
| Saarland | Elterngespräch | Berichtszeugnis |
| Sachsen | Berichtszeugnis | Berichtszeugnis |
| Sachsen-Anhalt | Berichtszeugnis (kann aber ab Jg. 1 auch bereits durch Notenzeugnis ersetzt werden durch Beschluss der Gesamtkonferenz) | Berichtszeugnis |
| Schleswig-Holstein | Elterngespräch | Berichtszeugnis |
| Thüringen | Wortgutachten | Wortgutachten |

| | Jahrgangsstufe 2 · Erstes Halbjahr | Jahrgangsstufe 2 · Zweites Halbjahr |
|---|---|---|
| Baden-Württemberg | Elterngespräch auf Beschluss der Schulkonferenz möglich, statt Schulbericht | Schulbericht und Noten in Deutsch und Mathematik |
| Bayern | Berichtszeugnis | - Benotung in den einzelnen Fächern<br>- ab Jg. 2: Bewertung des Arbeits- und Sozialverhaltens in standardisierter Form |
| Berlin | Berichtszeugnis oder Elterngespräch | Berichtszeugnis |
| Brandenburg | Elterngespräch | Lernentwicklungsbericht<br>oder Notenzeugnis (Beschluss durch Mehrheit der Klassenkonferenz und Elternversammlung) |
| Bremen | | Lernentwicklungsbericht |
| Hamburg | Information der Eltern in »geeigneter« Weise | Lernentwicklungsbericht |
| Hessen | Berichtszeugnis | Ziffernzeugnis<br>- ab Jg. 2-4: Beurteilung des Arbeits- und Sozialverhaltens durch Noten oder in verbalisierter Form (Beschluss durch Mehrheit der Gesamtkonferenz) |
| Mecklenburg-Vorpommern | Lernentwicklungsbericht | Lernentwicklungsbericht und Notenzeugnis |
| Niedersachsen | Berichtszeugnis | Berichtszeugnis |
| Nordhein-Westfalen | | Berichtszeugnis |
| Rheinland-Pfalz | Berichtszeugnis | Berichtszeugnis |
| Saarland | Berichtszeugnis und Noten in Deutsch und Mathematik | Notenzeugnis |
| Sachsen | Berichtszeugnis | - Berichtszeugnis und Noten in Deutsch und Mathematik<br>- Bewertung des Arbeits- und Sozialverhaltens ab Jg. 2 durch Noten |
| Sachsen-Anhalt | Berichtszeugnis | - Berichtszeugnis und Noten in Deutsch und Mathematik<br>- Bewertung des Arbeits- und Sozialverhaltens ab Jg. 2 durch Noten |
| Schleswig-Holstein | Berichtszeugnis | Berichtszeugnis |
| Thüringen | Wortgutachten | Wortgutachten |

| | Jahrgangsstufe 3 und 4 |
|---|---|
| Baden-Württemberg | - Notenzeugnis<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |
| Bayern | Notenzeugnis |
| Berlin | - Notenzeugnis oder Berichtszeugnis<br>(Beschluss durch 2/3-Mehrheit der Erziehungsberechtigten)<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |
| Brandenburg | - Notenzeugnis oder Lernentwicklungsbericht<br>(Beschluss durch Mehrheit der Klassenkonferenz und der Elternversammlung)<br>- Bewertung des Arbeits- und Sozialverhaltens ab Klasse 3 als schriftliche Information, die getrennt vom Ziffernzeugnis ausgegeben wird |
| Bremen | - ab 2. Halbjahr Jg. 3:<br>Lernentwicklungsbericht oder Notenzeugnis<br>(Beschluss durch Mehrheit der Schulkonferenz)<br>- Ende der Jg. 3 kann statt schriftlicher Information eine mündliche Information erfolgen<br>(Beschluss durch Mehrheit der Schulkonferenz)<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |
| Hamburg | - ab Halbjahr Jg. 3: Notenzeugnis mit ergänzenden Berichten<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |
| Hessen | Notenzeugnis |
| Mecklenburg-Vorpommern | - Notenzeugnis<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |
| Niedersachsen | - Notenzeugnis<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |
| Nordhein-Westfalen | - Jg. 3 Bericht und Notenzeugnis<br>- Halbjahreszeugnis nur in Klasse 4<br>- verbale Beurteilung in Jg. 3 möglich<br>(Beschluss durch Schulkonferenz)<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |
| Rheinland-Pfalz | - Notenzeugnis<br>- Bewertung des Arbeits- und Sozialverhaltens ab Halbjahr Jg. 3: durch Noten |
| Saarland | - Notenzeugnis<br>- Bewertung des Arbeits- und Sozialverhaltens ab Halbjahr Jg. 3: durch Noten |
| Sachsen | Notenzeugnis |
| Sachsen-Anhalt | Notenzeugnis |
| Schleswig-Holstein | - für Jg. 3 kann die Schulkonferenz Noten beschließen<br>- für Jg. 4 Notenzeugnis<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |
| Thüringen | - ab Jg. 3 Wortgutachten und Noten in Deutsch, Mathematik und Heimat- und Sachkunde<br>- Bewertung des Arbeits- und Sozialverhaltens in Form einer Verbalbeurteilung |

Abb. 1:
Zeugnisbestimmungen in den Bundesländern
(nach: Karin Müller 2005[1]).

1 Die - von der Veröffentlichung (S. 94-97) etwas abweichende - Datei wurde von der Autorin freundlicherweise für dieses Gutachten zur Verfügung gestellt.
2 Die Auswertung für Bayern bezieht sich auf die Informationen des Kultusministeriums im Internet und auf eine telefonische Auskunft des Kultusministeriums im September 2004.

*Baden-Württemberg*

- Verordnung des Kultusministeriums über die Notenbildung in der Fassung v. 05.02.2004.
- Verwaltungsvorschrift in der Fassung v. 24.06.2003.
- Verordnung des Kultusministeriums über die Schülerbeurteilung in Grundschulen und Sonderschulen in der Fassung v. 05.02.2004.

*Bayern*

- Schulordnung für die Volksschulen in Bayern in der Fassung v. 18.11.2002.
- Reform der Notengebung. In: http://www.km.bayern.de/km/schule/schularten/allgemein/grundschule/notengebung/index.shtml (02.12.2004).
- Telefonisches Gespräch mit Hr. Jörg Maier (Kultusministerium) (Sept. 2004).

*Berlin*

- Schulgesetz für das Land Berlin in der Fassung v. 26.01.2004.
- Ausführungsvorschrift über Noten und Zeugnisse in der Fassung v. 21.07.2002.

*Brandenburg*

- Gesetz über Schulen im Land Brandenburg.
- Verordnung über den Bildungsgang der Grundschule in der Fassung v. 02.08.2001.
- Verwaltungsvorschriften zu Informationen über das Arbeits- und Sozialverhalten in den Jahrgangsstufen 3 bis 10 in der Fassung v. 29.07.2004.
- Amtsblatt des Ministeriums für Bildung, Jugend und Sport Nr. 14 v. 23.12.2002.

*Bremen*

- Verordnung für Zeugnisse und Lernentwicklungsberichte und über die Abschlüsse an öffentlichen Schulen in der Fassung v. 08.07.2002.

*Hamburg*

- Hamburgisches Schulgesetz in der Fassung v. 27.06.2003.
- Ausbildungs- und Prüfungsordnung für die Klassen 1 bis 10 der allgemeinbildenden Schulen in der Fassung v. 27.06.2003.

*Hessen*

- Hessisches Schulgesetz in der Fassung v. 30.06.1999.
- Verordnung zur Gestaltung des Schulverhältnisses in der Fassung v. 01.09.2000.
- Verordnung zur Ausgestaltung der Bildungsgänge und Schulformen der Grundstufe (Primarstufe) und der Mittelstufe (Sekundarstufe I) und der Abschlussprüfungen in der Mittelstufe.

*Mecklenburg-Vorpommern*

- Schulgesetz für das Land Mecklenburg-Vorpommern in der Fassung v. 07.07.2003.
- Verwaltungsvorschrift des Kultusministeriums in der Fassung v. 08.09.1998.

*Niedersachsen*

- Die Arbeit in der Grundschule Erlasse des MK in der Fassung v. 03.02.2004.
- Zeugnisse in den allgemein bildenden Schulen RdErl. D. MK in der Fassung v. 24.05.2004.

*Nordrhein-Westfalen*

- Allgemeine Schulordnung in der Fassung v. 08.04.2003.
- Verordnung über den Bildungsgang in der Grundschule.

*Rheinland-Pfalz*

- Schulordnung für die öffentlichen Grundschulen in der Fassung v. 21.07.2003.

*Saarland*

- Zeugnis- und Versetzungsordnung - Schulordnung für die Grundschulen im Saarland in der Fassung v. 04.07.2003.

*Sachsen*

- Verordnung des Sächsischen Staatsministeriums für Kultus über Grundschulen im Freistaat Sachsen in der Fassung v. 02.08.2004.

*Sachsen-Anhalt*

- Leistungsbewertung in der Grundschule RdErL des MK in der Fassung v. 30.06.2004.

*Schleswig-Holstein*

- Landesverordnung über Notenstufen und andere Angaben in Zeugnissen in der Fassung v. 15.06.2004.
- Landesverordnung über Aufnahme und Aufsteigen nach Klassenstufen an der Grundschule in der Fassung v. 08.03.1999.

*Thüringen*

- Thüringer Schulgesetz in der Fassung v. 30.04.2003.
- Thüringer Schulordnung für die Grundschule, die Regelschule, das Gymnasium und die Gesamtschule in der Fassung v. 07.04.2004.

Angesichts der aktuellen Bewegungen in mehreren Ländern haben wir auf eine erneute Bestandsaufnahme verzichtet. Sie wäre vermutlich schon in wenigen Monaten veraltet.

Festhalten lassen sich aber zwei Trends, die über alle Bundesländer hinweg zu beobachten sind:

▲ eine Ausdehnung des Zeugnisinhalts vom Leistungs- auf den Verhaltensbereich - und damit eine Rückkehr zu den in den 1960er Jahren (westliche Bundesländer) bzw. nach 1989 (östliche Bundesländer) verbannten Kopfnoten[52];
▲ eine Vorverlagerung der Notengebung auf frühere Jahrgangsstufen sowie eine Einschränkung von Ausnahmeregelungen für Berichtszeugnisse - und damit eine Rückkehr zu den ab etwa 1970 in den westlichen und ab 1990 in den östlichen Bundesländern auf Klasse 3/4 aufgeschobenen Ziffernzeugnissen.

Gegenläufig zu diesen Entwicklungen finden alternative Formen der Leistungsbeurteilung auf der Sekundarstufe[53] und im tertiären Bereich wachsende Aufmerksamkeit - bis hin zu neuen Beurteilungsverfahren in der Berufswelt. Zielvereinbarungen, inhaltliche Beurteilungen, Selbsteinschätzungen und regelmäßige Mitarbeitergespräche ermöglichen differenziertere Bewertungen. Insofern ist das Argument, Ziffernoten seien notwendig, um SchülerInnen auf den »Ernst des Lebens« vorzubereiten, überholt. Andererseits sind auch in Arbeitszeugnissen Tendenzen zu beobachten, die aus schlechten Verbalbeurteilungen in der Schule bekannt sind[54]: die bloß verbale Umschreibung von Ziffernnoten durch Textbausteine ohne Bezug auf Kriterien[55].

**0.5**

## Blicke über den Zaun: Die internationale Situation[56]

»Als empirisches Argument gegen Notenzeugnisse wird gerne auf die skandinavischen Länder verwiesen, die in Schulleistungsvergleichen regelmäßig sehr gut abschneiden und die bis zur achten Jahrgangsstufe auf Noten und Ziffernzeugnisse verzichten. Damit ist zwar kein Kausalzusammenhang bewiesen, wohl aber, dass Gesamtschulen ohne Noten effizient sein *können.*

Als empirische Gegenbeispiele werden jedoch einige asiatische Länder genannt, die bei den Vergleichen vor allem im mathematisch-naturwissenschaftlichen Bereich überdurchschnittlich gut abschneiden. Kaum übersehbar ist aber, dass viele asiatische Kulturen einen erheblich höheren Wert auf die Bildung und Ausbildung ihrer Kinder legen, wodurch u.a. eine deutlich größere Leistungsbereitschaft schon in den jüngeren Schülern vorhanden ist. Einige betrachten es allerdings in negativem Sinne als erhöhten Leistungsdruck.

Fraglich ist in diesem Zusammenhang, ob Notenzeugnisse überhaupt eine wichtige Rolle bei den Ergebnissen dieser Tests spielen.«[57]

Die Situation in anderen Ländern ist sehr heterogen. Insbesondere variiert der institutionelle Kontext. Im Vergleich zu Deutschland sind allerdings in den meisten Ländern (bis auf den ehemaligen Ostblock) Selektionsentscheidungen wie Zurückstellung, Sitzenbleiben, Überweisung in Sonderschulen seltener, dauert der gemeinsame Unterricht länger und setzt auch eine Benotung von Leistungen später ein. Viele dieser Länder schneiden bei internationalen Vergleichen besser ab als Deutschland[58].

Übersichtliche Muster und einfache Abhängigkeiten gibt es aber nicht. Dazu sind die Konstellationen zu komplex und vielfältig, auch innerhalb einzelner Länder, wie schon das Beispiel Schweiz[59] anschaulich macht:
»Die gegenwärtigen Tendenzen auf Volksschul- und Sekundarschulstufe zeigen im übrigen, dass das Notenprinzip längst nicht mehr in der Absolutheit gilt, die die Kritik unterstellt (Vögeli-Mantovani 1999, S. 89 ff.). Inzwischen gibt es nicht nur ›Noten‹,
- sondern Noten mit und ohne explizite Bezugsnormen,
- Lernberichte,
- fakultative wie nicht-fakultative Beurteilungsgespräche,
- Orientierungsarbeiten zur Standortbestimmung,
- Selbstbeurteilungen der Schülerinnen und Schüler,
- Zeugnisse mit lernzielbezogenen Wortetiketten,
- Zeugnisse mit lernzielbezogenen Wortetiketten für Beurteilung des Lernprozesses und der Leistung.

In Basel gibt es ›prognostische Noten‹ erst ab der sechsten Klasse, reguläre Noten erst nach dem Übertritt in der achten Klasse. In Baselland gibt es reguläre Noten ab der sechsten Klasse, zuvor wahlweise Noten oder Lernberichte. Im Aargau gibt es reguläre Noten ab dem zweiten Beurteilungszeitraum der ersten Klasse, in Bern werden bis zur sechsten Klasse Beurteilungsgespräche geführt, Lernberichte erstellt und in der dritten sowie sechsten Klasse lernzielorientierte Noten erteilt. In Solothurn sind die ersten drei Jahrgänge notenfrei, in Freiburg gibt es Verbalzeugnisse, Lernberichte und Beurteilungsgespräche von der ersten Klasse an. Diese

52 Vgl. zur Debatte über Kopfnoten: Pro: Matthias Rößler. Contra: Ulrich Herrmann, in: Pädagogik H. 10/2000, 60-61; Arnold/Vollstädt (2001); Solzbacher (2001); Thomas (2001); Landtag NRW 2003; Kirsten 2003; Becker/Ramseger 2003; Bayerisches Kultusministerium 2004; s.a. dazu auch die Übersicht über die Bundesländer in Bohl (2003).
53 Vgl. etwa Grunder/Bohl (2001); Winter (2004).
54 S. unten > Kap. 3.1.
55 Vgl. etwa > www.arbeitszeugnis.de/ [Abruf: 12.3.2006] und Weuster/Scheer (2005).
56 Dieses Kap. geht auf einen Eigenbeitrag von Axel Backhaus zurück. Nur bedingt hilfreich waren die einzelnen Länder-Übersichten in der Datenbank www.eurydice.org, weitere Berichte finden sich in den Beiträgen zu Weston (1991) und Vergleiche bei Vögeli-Mantovani (1999, Kap. 4); Schmitt (2001, Teil I).
57 > http://de.wikipedia.org/wiki/Leistungsbeurteilung_(Schule) [Abruf: 20.1.2006].
58 Vgl. OECD (1995, 89).
59 Vgl. Birkhäuser (1999).

| Kriterium | | IGLU Lesen | IGLU Lesen BRD | IGLU Mathematik | IGLU Mathematik BRD | PISA 2003 Lesen | PISA 2003 Lesen BRD | PISA 2003 Mathematik | PISA 2003 Mathematik BRD |
|---|---|---|---|---|---|---|---|---|---|
| Frühe Selektion | über wiegend KEINE Noten | Es gibt innerhalb Europas keine Länder, die die Schüler früh auf verschiedene Schulformen schicken und die gleichzeitig (überwiegend) auf Noten verzichten, sofern man nicht 6 Jahre noch als kurz wertet (z.B. Irland) | | | | | | | |
| | über wiegend Noten | Deutschland 4; 539 Rang 11 | BW 4; 539 Rang 5[1] | Österreich 4; 559 Rang 7 | BW 4; 565 Rang 6 | - | Bayern 4; 518 Rang 6 | Zum Vergleich Deutschland 4; 513; Rang 13 | Bayern 4; 534 Rang 4 |
| | | - | Bremen 4; 507 Rang 22 | Zum Vergleich Deutschland 4; 545; Rang 12 | Bremen 4; 512 Rang 18 | Deutschland 4; 491 Rang 18 | Bremen 4; 467 Rang 27 | Österreich 4; 506 Rang 16 | Bremen 4; 491 Rang 21 |
| Späte Selektion | über wiegend KEINE Noten | Schweden 9; 561 Rang 1 | - | Niederlande 8; 577 Rang 5 | - | Finnland 9; 543 Rang 1 | - | Finnland 9; 548 Rang 2 | - |
| | | Norwegen 10; 499 Rang 24 | - | Norwegen 10; 502 Rang 20 | - | Dänemark 9; 492 Rang 17 | - | Norwegen 10; 490 Rang 20 | - |
| | über wiegend Noten | Bulgarien 8; 550 Rang 4 | - | Tschechien 9; 567 Rang 6 | - | - | - | Tschechien 9; 516 Rang 12 | - |
| | | Türkei 8; 440 Rang 27 | - | Island 10; 474 Rang 25 | - | Türkei 8; 441 Rang 28 | - | Türkei 8; 408 Rang 27 | - |

1 Einsortierung der Bundesländer im Ranking der internationalen Vergleichsgruppe.

Abb. 2: Backhaus 2006 (erstellt für diese Expertise)

Entwicklung zeigt, dass in Zukunft eher die Wahrung des Notenschemas das Problem ist, weil Teile der Schulreformdiskussion die Meinung von Karlheinz Ingenkamp übernommen hat, wonach Noten und Zeugnisse an sich höchst fragwürdige Phänomene seien und daher abgeschafft oder ersetzt werden müssten.«[60]

Die Konstellationen werden noch komplexer und vielfältiger, wenn man weitere Rahmenbedingungen berücksichtigt, z.B. den Beginn der Schulpflicht, die Dauer der gemeinsamen Schulzeit, System/e der Leistungsbeurteilung, Formen und Umfang der Selektion - und wenn man diese auf unterschiedliche Erfolgsmaße bezieht (z.B. auf das Abschneiden der Länder in verschiedenen Fächern und bei IGLU einerseits vs. PISA andererseits[61]). Festhalten lässt sich aber[62]:

▲ Es gibt mehrere Länder und Regionen mit späterer Benotung und weniger Selektion, die in den Leistungsvergleichen deutlich besser abschneiden als Deutschland (vgl. etwa Schweden, Finnland oder Südtirol).

▲ In dieser Gruppe finden sich aber auch Länder, die nicht besser oder sogar schlechter abschneiden als Deutschland (z.B. Norwegen).

▲ Schließlich gibt es innerhalb Deutschlands auch Länder mit früherer Benotung und/oder stärkerer Selektion (z.B. Bayern und Baden-Württemberg), die zumindest oberflächlich erfolgreicher zu sein scheinen[63] als Bundesländer, die bis vor kurzem noch stärker integrative Ansätze favorisiert haben (z.B. Nordrhein-Westfalen oder Bremen).

60 Oelkers (2001, o.S.).

61 Die unten folgende Tabelle (Abb. 2) ist ein Auszug aus einer umfassenden Übersicht (50 S.), die Axel Backhaus für die Arbeitsgruppe aufgrund folgender Quellen erstellt hat: Schmitt u.a. (1992); Artelt u.a. (2001b); Schmitt (2001); Bos u.a. (2004);Prenzel u.a. (2005b); Eurydice (o.J.). Ausführliche Beschreibungen von Notenstufen von 30 Staaten der Welt finden sich in http://en.wikipedia.org/wiki/Grade_%28education%29 [Abruf: 12.02.06]. Ein deutscher Text mit 6 Ländern ist unter http://de.wikipedia.org/wiki/Schulnoten erhältlich [Abruf: 12.02.06].

62 Kommentierung der Grafik in Abb. 2:

- Sowohl die internationalen als auch die innerdeutschen Rangplätze bei PISA und IGLU sind nur näherungsweise als Erfolgskriterien nutzbar, da Besonderheiten in den Populationen nicht berücksichtigt werden. Dazu zählen etwa die Arbeitslosen- und Sozialhilfequote und die Anteile an MigrantInnen, aber auch die unterschiedliche sozio-ökonomische und sozio-kulturelle Zusammensetzung der MigrantInnengruppen, selbst wenn ihr Gesamtanteil quantitativ übereinstimmt (vgl. zur Kritik an den publizierten Auswertungen: Brügelmann 2002; Block/Klemm 2005; 2006).
- Die deutschen Bundesländer sind nicht ganz sauber einzuordnen: Bezieht man die Kriterien nur auf die Grundschule, so gibt es einige Bundesländer mit »später(er)« Selektion und Noten. Aber bezogen auf den internationalen Vergleich sind alle Länder durch das Muster »frühe Selektion (Klasse 4-6) und Noten« charakterisierbar. Diese Besonderheit ist bei Vergleichen in der Tabelle zu bedenken.

63 Vgl. zur Kritik an zu vereinfachten Deutungen der Länderdifferenzen etwa Block/Klemm (2006).

Fazit:
Strukturelle Systemmerkmale garantieren keine pädagogischen Erfolge[64]. Die Abschaffung von Noten ist kein »Selbstläufer«. Dieses Ergebnis deckt sich mit dem später zu erörternden Befund, dass auch deutschlandintern die Wirkungen von Noten vs. Verbalgutachten *innerhalb* der beiden Ansätze stärker streuen als zwischen ihnen. Allerdings machen die Befunde von IGLU und PISA ganz deutlich: Ein Verzicht auf Benotung und Selektion in den ersten Schuljahren und über die Grundschulzeit hinaus ist kein Hindernis für eine erfolgreiche pädagogische Arbeit - auch im fachlichen Leistungsbereich.

Eindrucksvoll belegt wird dies durch das Beispiel Südtirol[65]: Die bei PISA 2003 erfolgreichste deutschsprachige Region liegt - in Italien. Im Lesen noch einen Punkt besser als der »Weltmeister« Finnland und satte 26 Punkte, das ist rund ein halbes Schuljahr, vor dem deutschen Spitzenreiter Bayern. Seit 1977 gibt es keine Ziffernnoten mehr. »1993 wurden individuelle Bewertungsbogen eingeführt. Die Bewertungsstufen ›ausgezeichnet‹, ›sehr gut‹, ›gut‹, ›genügend‹, ›nicht genügend‹ dienen dazu, das Kind mit sich selbst zu vergleichen und seinen eigenen Lernfortschritt und seine eigene Anstrengung zu bewerten. Sie bedeuten keinen Rangplatz des Kindes in der Klasse. So kann ein ›sehr gut‹ bei dem einen Kind etwas völlig anderes bedeuten als ein ›sehr gut‹ bei einem anderen Kind. Ganz offensichtlich kommt die Südtiroler (und die italienische) Gesellschaft mit einem nichtvergleichenden Bewertungssystem ohne weiteres klar, wir haben keine Kritik daran gehört.«[66]

**1**

## Mit welchen Verfahren werden Leistungen erfasst?[67]

Die Fundiertheit von Beurteilungen hängt von ihrer Datengrundlage, diese von den eingesetzten Instrumenten ab. Deren Qualität wiederum wird üblicherweise über drei Gütekriterien bestimmt[68]:

- ▲ Gültigkeit (»Validität« > Kap. 1.1)
- ▲ Personunabhängigkeit (»Objektivität« > Kap. 1.2)
- ▲ Verlässlichkeit (»Reliabilität« > Kap. 1.3)

Bewertungen im Schulalltag stützen sich auf eine Vielfalt von Informationen: Tests, Klassenarbeiten, mündliche Beiträge, informelle Beobachtungen. Die Erhebung dieser Daten und ihre Auswertung muss sich an den drei Gütekriterien messen lassen. Allerdings stellt ihre Auslegung in der Testtheorie eine Verkürzung dar, die andere Standards der Evaluation von Lernen gefährdet. Zu wenig Beachtung finden bisher Kriterien aus der weiteren Evaluationsdiskussion wie Fairness, Glaubwürdigkeit, Stimmigkeit, Ökonomie, Nützlichkeit.[69]

**1.1**

## Wie gut erfassen Leistungsbeurteilungen, was sie erfassen sollen? (Validität)

Die Grundfrage: Misst ein Instrument wirklich das, was es zu messen vorgibt? Leistungen sind beobachtbare Verhaltensweisen. Ihre Beurteilung zielt aber nicht nur auf das beobachtete Verhalten (»Performanz«), sondern auch auf die zugrunde liegenden Fähigkeiten (»Kompetenz«). Die Gültigkeit solcher Schlüsse ist nur schwer zu begründen, da die psychologischen Modelle und die pädagogischen Maßstäbe selbst umstritten sind. Außerdem können die Form der Aufgabe bzw. die Bedingungen, unter denen sie zu bewältigen ist, die zu erbringende Leistung verändern.

Ein Beispiel: Werden über Diktate tatsächlich wesentliche Aspekte der Rechtschreibkompetenz erfasst oder haben (auch) andere Faktoren Einfluss für die beobachtete Leistung? Verschiedene Studien[70] zeigen, dass die Leistung in Diktaten im zweiten Teil des Textes schwächer ausfällt als im ersten Teil der Aufgabe. Dieser Leistungsabfall ist allein durch eine Fehlerzunahme in der Gruppe der schwächeren RechtschreiberInnen bedingt. Vermutlich hängt deren abnehmende Leistung damit zusammen, dass sie im Verlauf des Diktats zunehmend unter Stress geraten. Insofern werden die Fehlerhäufigkeit und damit die Leistung nicht nur durch die Rechtschreibkompetenz der SchülerInnen, sondern auch durch ihre (objektiv) unterschiedliche Belastung und ihre

64 Dies ist ein genereller Befund der Schulforschung, belegt an so unterschiedlichen Strukturkontrasten wie Gesamtschule vs. dreigliedriges Schulsystem oder Koedukation vs. Jungen- und Mädchenschulen, vgl. Brügelmann (2005, Kap. 30, 31).

65 Vgl. zum Abschneiden bei PISA und zu den Konzepten und Bedingungen des Unterrichts: Höllrigl/Meraner (2005); Leitzgen (2005); Meraner (2005); Ratzki (2005; 2006).

66 Ratzki (2006, 25).

67 Einen leichten Zugang zu den verschiedenen Studien bietet Ammann (2002). Immer noch empfehlenswert: Ingenkamp (1971a), gute neuere Zusammenfassung bieten Jachmann (2003, Kap. 2.1) und Wagener (2003, Kap. 1.1.1 und 1.1.2).

68 Fundierte und verständliche Einführungen finden sich bei: Diekmann (1995, Kap. VI.3); Jachmann (2003, Kap. 2.1.1); Brügelmann (2005a, Kap. 56).

69 Vgl. Winter (2004, 91-94) und ausführlicher House (1980), der Kriterien wie Gerechtigkeit, Glaubwürdigkeit, Unparteilichkeit und Fairness hervorhebt. Nisbet (1978) formuliert genereller für Verfahren der Rechenschaft (Hervorhebungen durch die AutorInnen), »... that they
- must operate in a way that is *fair* to all concerned;
- should be valid and *relevant* to current concerns;
- should provide feedback for decision-making and encourage *wider involvement* in decisions;
- should either be objective or make *subjectivity explicit;*
- should be verifiable, i.e. *open to checking;*
- should *not distort the processes of teaching and learning;*
- should be understandable and the results communicable;
- hould be comprehensive and take account of the wide variety of aspects of education.«

70 Vgl. Schneider (1985); Brügelmann (1994a, 206-207).

(subjektiv) unterschiedliche Stressresistenz und Konzentrationsfähigkeit beeinflusst[71].

Man kann die Validität von Methoden und Instrumenten auf verschiedene Weise bestimmen und überprüfen. Drei dieser Wege sind für Verfahren der Leistungsbeurteilung besonders bedeutsam:

▲ die Analyse von Aufgaben mit Bezug auf vorgegebene Inhalte bzw. Kriterien, z.B. Abstimmung von Testaufgaben auf die Ziele und Inhalte von Lehrplänen (> Kap. 1.1.1);

▲ der Vergleich der Ergebnisse mit denen, die durch ein anderes etabliertes Verfahren gewonnen wurden, z.B. durch einen Abgleich von Noten mit Testwerten (> Kap. 1.1.2);

▲ die Vorhersage zukünftiger aus aktuellen Leistungen und die Überprüfung der Prognosegenauigkeit (> Kap. 1.1.3).

### 1.1.1

### Wie gut sind die Kriterien für Leistungsbeurteilungen inhaltlich abgesichert?

Inhalte für Unterricht und Kriterien für den Lernerfolg finden sich in Richtlinien bzw. Fachlehrplänen. Auf sie bezieht sich beispielsweise die Überprüfung der »curricularen Validität« von Tests in den großen Leistungsstudien wie PISA und IGLU[72]. Die Diskussion über die Bildungsstandards zeigt aber ein hohes Maß an Uneinigkeit, was als Mindest- oder Regelleistung eingefordert werden kann. Die Kritik an den Aufgaben der landesweiten Lernstandserhebungen und internationalen Leistungsvergleiche hat offen gelegt, wie umstritten die Annahmen zu den angeblich erfassten »Fähigkeiten« sind.

Ratzka (2003, Kap. 4.5.6) hat für den Bereich Mathematik gezeigt, wie wichtig die Auswahl des konkreten Tests für die Ergebnisse und damit für die Einstufung individueller Leistungen ist. Sie hat auch die Autorität der Befunde aus den internationalen Leistungsstudien in Frage gestellt, die in der bildungspolitischen Diskussion als zentraler Qualitätsausweis des Schulwesens gehandelt werden. Bezogen auf die TIMS-Studie sind zwei Befunde aus dem Vergleich mit zwei weiteren Tests bedeutsam[73]:

▲ 58% der SchülerInnen erreichen in anderen Tests als TIMSS andere Ergebnisse. Selbst zwischen zwei verschiedenen Tests des gleichen Grundtyps (»Textaufgaben« im TIMSS- und im AMI-Test) kann es erhebliche Unterschiede geben.

▲ Unter Zeitdruck (»Speed-Test«) ergeben sich auch innerhalb von TIMSS, also bei denselben Aufgaben, teilweise andere Ergebnisse als ohne Zeitdruck (»Power-Test«). Das gilt für die deutschen SchülerInnen vor allem bei komplexeren Aufgaben bzw. bei unbekannten Aufgabenformaten und generell für Mädchen im Vergleich zu Jungen.

Andere AutorInnen haben einzelne Aufgaben in den internationalen und landesweiten Tests genauer untersucht und dabei die Angemessenheit der Aufgaben mit bedenkenswerten Argumenten in Frage gestellt[74]. Die unterschiedlichen Einschätzungen der Lesefähigkeit deutscher SchülerInnen nach PISA und DESI[75] machen darauf aufmerksam, wie vorsichtig die Ergebnisse von Leistungstests interpretiert werden müssen. Ihre Aussagen beschränken sich auf spezifische Inhalte und Aufgabenformen, die nur in Kenntnis dieser Ausschnitthaftigkeit als Indikatoren für übergreifende Kompetenzen genommen werden können. Das gilt nicht nur für bildungspolitische Folgerungen, sondern ebenso für individuelle Bewertungen. Besonders deutlich geworden ist dies bei der Diagnose sog. »LegasthenikerInnen«, deren Besonderheit durch die Diskrepanz zwischen IQ und Lese-/Rechtschreibleistung definiert wurde. Je nach eingesetztem Intelligenz- und Lese- oder Rechtschreibtests fielen einzelne Kinder in diese Rubrik - oder auch nicht[76].

Aber auch für das Lehrerurteil, das die Noten bestimmt, haben empirische Studien Validitätsprobleme aufgezeigt: Bei Aufsätzen beeinflussen sowohl der Umfang[77] als auch die Zahl der orthografischen Fehler und die Qualität der Handschrift die Note[78].

Es geht aber nicht nur um die Übereinstimmung der Inhalte und die Form der Aufgabe. Zu bestimmen ist auch, welches Niveau der Aneignung verlangt werden soll. Die Diskussion über die Bildungsstandards zeigt ein hohes Maß an Uneinigkeit, was als Mindest- oder Regelleistung eingefordert werden kann. Die Kritik an den Aufgaben der landesweiten Lernstandserhebungen und internationalen Leistungsvergleiche hat offen gelegt, wie umstritten die Annahmen zu den angeblich erfassten »Fähigkeiten« sind.

Die Validität von Noten wird in den letzten Jahren oft mit dem Hinweis kritisiert, dass Noten nicht hinreichend mit den

71 Durch Auslösung von Angst sinkt das Leistungsniveau im Vergleich zur tatsächlichen Kompetenz (Moeller 1972).

72 Vgl. etwa zur Lehrplangültigkeit der am *literacy*-Konzept orientierten PISA-Tests: Baumert u.a. (2003, Kap.2).

73 Ratzka ließ in ihrer Studie dieselben Kinder verschiedene Tests bearbeiten. Im einzelnen liegen die linearen Korrelationen zwischen den Tests nur bei .48** (TIMSS - AMI), .37** SCHOLASTIK - TIMSS) bzw. .07 (AMI - SCHOLASTIK), d. h. die Leistungen erklären nur 0.4% bis maximal 23% gemeinsamer Varianz. Vgl. zum Lesen die unterschiedlich hohen Korrelationen verschiedener (Teil-)Tests in den LUST-Teilstudien: Backhaus (2005).

74 Vgl. für Mathematik u.a.: Bender (2004); Scheerer (2004); Selter (2005); für Sprache Bartnitzky (2005a+b); Benholz u.a. (2005) und zur Diskussion der KritikerInnen mit dem VERA-Team Heft 90/2005 von »Grundschule aktuell«.

75 Vgl. zu den Ergebnissen der Studie »Deutsch Englisch Schülerleistungen International« (DESI) Klieme u.a. (2006).

76 Vgl. Scheerer-Neumann (1996, Kap. 2); zur grundsätzlichen Problematik des Legasthenie-Konstrukts: Brügelmann (2005a, Kap. 19).

77 Kürzere Aufsätze werden generell schlechter benotet (Baurmann 1977).

78 Dazu zitiert Ammann (2002) eine norwegische Studie von Osnes (1972); s. zu Rechtschreibfehlern auch Birkel (2003).

Ergebnissen von Leistungstests in den entsprechenden Fächern übereinstimmten. Damit ist aber ein problematischer Maßstab gesetzt[79], unterstellt dieses Vorgehen doch, dass Tests eher die »wahre« Fähigkeit einer Person erfassen. Andererseits wird die inhaltliche Gültigkeit von Tests in der Regel damit begründet, dass ihre Ergebnisse in der Normierungsphase »gut« mit den Lehrerurteilen übereinstimmen. Damit entstehen Kreisschlüsse, bei denen kein Verfahren beanspruchen kann besser zu sein als das andere[80]. Das einzige unabhängige Kriterium ist ihre Vorhersagekraft, bezogen auf zukünftige Leistungen. Diese aber erweist sich als sehr begrenzt (s. dazu > Kap. 1.1.3).

Schließlich ist noch eine weitere Schwäche sowohl des Lehrerurteils als auch von Tests festzuhalten - wenn ihre Ergebnisse in Form einer Ziffernnote oder eines Summenwerts verdichtet werden. Deren Validität als Pauschalbewertung eines Lernbereichs wird den differenzierten Leistungsprofilen (Geometrie vs. Sachrechnen vs. Arithmetik, schriftliches vs. Kopfrechnen) nicht gerecht. Die fehlende Differenziertheit einer einzelnen Ziffer kann Stärken und Schwächen in den Teildimensionen eines Leistungsbereichs nicht zureichend darstellen. Hier liegt ein Potenzial von Lernberichten, auch wenn es oft nicht zureichend ausgeschöpft wird (vgl. > Kap. 3.1).

### 1.1.2

### Wie gut stimmen Beurteilungen aus verschiedenen Quellen überein?

Nicht nur die Ergebnisse von verschiedenen Tests desselben Bereichs, auch Fachnoten und Tests stimmen nur begrenzt überein. Problematisch an der Diskussion »nach PISA« ist, dass Tests dabei fast selbstverständlich als Maßstab für die »wahre« Leistung von SchülerInnen gesetzt werden[81]. Vergleiche mit weiteren Kriterien zeigen aber, dass Lehrerurteil und Tests unterschiedliche Aspekte fachlicher Leistungen erfassen. Es macht deshalb keinen Sinn, die Qualität des einen Verfahrens allein durch den Grad der Übereinstimmung mit den Ergebnissen des anderen zu bestimmen.

Eine viel zitierte[82] Studie von Ingenkamp (1975) zeigte, dass die Zensuren in 37 sechsten Berliner Klassen stark von den »lehrplangültigen« Testergebnissen der Kinder abwichen: Bei gleichem Testergebnis hatten SchülerInnen unterschiedliche Noten erhalten. In der IGLU-Studie stellten Bos u.a. (2004b, 205) über verschiedene Schulen hinweg eine breite Streuung der Testleistungen *innerhalb* einer Notenstufe und damit eine starke Überlappung *zwischen* den Notenstufen fest - ein Phänomen, das sich auch in vielen anderen Untersuchungen nachweisen ließ[83]. Bei PISA-2000 lag die Korrelation zwischen Mathematiknote und curriculumnahem Mathematiktest bei .32 über die verschiedenen Schulformen hinweg und bei .43 innerhalb der Bildungsgänge[84].

Für diese Abweichungen sind verschiedene Erklärungen denkbar:

▲ LehrerInnen erkennen schlechter als Tests, wo Kinder in ihrer Entwicklung stehen, welche Schwierigkeiten sie bei der Auseinandersetzung mit dem jeweiligen Gegenstand haben, was als fehlende diagnostische Kompetenz interpretiert werden könnte.

Oder:

▲ LehrerInnen verschiedener Klassen ordnen den richtig erkannten Leistungsstand unterschiedlichen Notenstufen zu, so dass sie lediglich abweichende Bewertungsmaßstäbe anlegen.

Die vorliegenden Studien sprechen eindeutig für die zweite Sicht. So korrelieren Noten und Testwerte *innerhalb* von Klassen sehr viel höher miteinander als über verschiedene Klassen hinweg[85]. LehrerInnen differenzieren unterschiedliche Lernstände also weitgehend zutreffend[86], aber sie setzen den Bezugspunkt für die anschließende Benotung unterschiedlich an. Für diese Deutung spricht auch der engere Zusammenhang, wenn man nicht Noten mit dem Lernerfolg korreliert, sondern von den LehrerInnen qualitative Urteile über die voraussichtliche Entwicklung ihrer SchülerInnen erfragt und diese mit Tests zur kognitiven Leistungsfähigkeit abgleicht[87].

Und damit sind wir bei einem dritten Grund für die Abweichungen, der positive wie negative Seiten hat: In das Urteil der LehrerInnen gehen Informationen aus einer kontinuierlichen Beobachtung der SchülerInnen in vielfältigen Situationen ein. Die Urteile von LehrerInnen, z.B. ihre Noten, sind breiter fundiert als die Ergebnisse punktueller Tests. Mit der Berücksichtigung von »Randbedingungen« werden sie aber auch stärker abhängig von den persönlichen Kriterien und Wahrnehmungsfiltern der einzelnen Lehrperson.

Soll man vor diesem Hintergrund die Aussagekraft von Tests am breiter fundierten Lehrerurteil oder die des Lehrerurteils am stärker kontrollierten Test überprüfen?

---

79 So auch im Forschungsüberblick bei Hoge/Coladarci (1989), auch wenn er zu einer positiven Einschätzung der Validität und Genauigkeit des Lehrerurteils kommt.

80 S. > Kap. 1.1.2 .

81 Vgl. etwa Lehmann (1999).

82 U.a. bei Mreschar (1985, 51).

83 Vgl. u.a. Ingenkamp (1971c); Thiel/Valtin (2002); Brügelmann (2003); Pietsch (2005).

84 Vgl. Baumert u.a. (2003, 325).

85 Allerdings streuen die Korrelationen zwischen Testergebnis und Note in verschiedenen Klassen breit, z.B. in den Klassen der Siegener LUST-Studie von .02 bis .94 (Brügelmann 2003c, Kap. 9). Während also einige LehrerInnen den Leistungsstand ihrer SchülerInnen sehr ähnlich einschätzen wie die eingesetzten Tests, gibt es bei anderen erhebliche Differenzen in der Rangfolge. US-amerikanische Studien berichten mit .28 bis .92 eine ähnlich breite Streuung der Korrelationen zwischen dem Lehrerurteil und den Schülerleistungen in standardisierten Tests (vgl. Hoge/Coladarci 1989, 303).

86 ... unter der Prämisse, dass man die Testergebnisse als Maßstab anerkennt.

87 Vgl. Merkens (2004).

Dieses Dilemma ist zu bedenken, wenn man zum Beispiel die Validität von Verbalzeugnissen einzuschätzen versucht, wie dies Maier (2001, 211 ff.) in einer differenzierten Studie getan hat. Seine Außenkriterien waren:

- Ziffernnoten des Abschlusszeugnisses in der 4. Klasse
- Ergebnisse des Schulleistungstests (AST 4)
- Einschätzung der Eltern und LehrerInnen bezüglich der Schulleistungen.

Seine Ergebnisse[88]: »Die zur Analyse der Übereinstimmungsvalidität zwischen Verbalzeugnis und Schulleistungsvariablen durchgeführten Korrelationsanalysen belegen insgesamt einen schwachen Zusammenhang, d.h. eine geringe Übereinstimmungsvalidität. Mit der Regressionsanalyse wird ein gemeinsamer Varianzanteil zwischen Verbalzeugnis und Außenkriterien von 13% ermittelt: Dabei trägt der Schulleistungstest am meisten zur Varianzaufklärung bei, gefolgt von den Zensuren, der Leistungseinschätzung durch die Lehrkraft und der Leistungseinschätzung durch die Eltern. Lediglich der Prädiktor Schulleistungstest leistet einen signifikanten Beitrag zur Varianzaufklärung der Kriteriumsvariablen Verbalzeugnis.«

Die Befunde bestätigen die Vermutung, dass Verbalzeugnisse und Ziffernzeugnisse verschiedene Informationen übermitteln, u.a. weil sie sich auf unterschiedliche Bezugsnormen beziehen und »...dass eine ›Übersetzung‹ der Verbalzeugnisse in Noten und umgekehrt keinen Sinn macht, *›da beiden Berichtsformen letztlich ein unterschiedliches Verständnis von Leistungen und ihrer Beurteilung zugrunde liegt‹ (Portmann (1997, 239) [...] Ziffernnoten die Ranginformationen liefern, zeigen nur schwache Korrelationen mit den verbalen Bewertungen, denen in hohem Ausmaß die individuelle und kriteriale Bezugsnorm zu Grunde liegt.«* (Maier 2001, 228, 230).

Dies ist ein gewichtiges Argument gegen den Vorwurf, den u.a. Schröter (1981a) Verbalbeurteilungen macht[89], sie seien nicht aussagekräftig, denn sie ließen sich nicht in Ziffern (rück)übersetzen. Analog sind auch Test und Beobachtung als unterschiedliche Zugänge zur Dokumentation der Leistung zu sehen, deren Ergebnisse sich ergänzen, aber nicht ersetzen können.

### 1.1.3

### Wie genau lässt sich aus der Beurteilung von Leistungen deren zukünftige Entwicklung vorhersagen (prognostische Validität)

Ein Problem einer jeden Prüfung[90] ist der Grad ihrer *externen* Validität. Damit ist die Schwierigkeit gemeint, aus der Prüfungsleistung in einer künstlich arrangierten Aufgabe auf erwartbare Leistungen in Alltagssituationen zu schließen. Ein echtes Außenkriterium stellt der spätere Schul-, Ausbildungs- oder Berufserfolg dar (»prognostische Validität«). Leistungen - als beobachtbare Verhaltensweisen - werden zwar rückblickend beurteilt. Die Beurteilung soll aber die zugrunde liegende Fähigkeit erfassen und damit auch Aufschluss geben über zukünftig zu erwartende Leistungen. Die Vorhersagekraft von Noten ist in verschiedenen Phasen der Bildungslaufbahn untersucht worden.

#### 1.1.3.1

#### Kindergarten > Schulerfolg

Vor Schulbeginn gibt es keine Noten. Über viele Jahre stand aber die Frage der Zurückstellung vom altersgemäßen Schulbeginn zur Diskussion. Als Grundlage für diese Entscheidung wurden vielfach standardisierte Tests herangezogen - bis die hohe Fehlerquote der Prognosen ihren Einsatz zunehmend fragwürdig werden ließ. Eine der wichtigsten Untersuchungen stammt von Krapp/Mandl (1977)[91]. Danach blieben von den Kindern, die nach einschlägigen Tests als »nicht schulreif« eingestuft und die deshalb nicht eingeschult wurden, bis zum 9. Schuljahr immerhin 13% sitzen. Aus der Kontrollgruppe, die trotzdem eingeschult wurde, waren es mit 28% zwar doppelt so viele. Individuell bedeutsamer aber ist der Kehrwert: Mit 72% schaffte die große Mehrheit die Pflichtschulzeit ohne Wiederholung einer Klasse, wenn sie entgegen der Testempfehlung eingeschult wurden. Der Schulreifetest produzierte also fast drei Viertel Fehlprognosen. Deshalb sind Schuleingangstests weitgehend abgeschafft worden.

Auch Klassifikationsversuche mit Hilfe fachbezogener Verfahren haben eine zu hohe Fehlerquote. Im Bereich der Schriftsprache schwankt sie zum Beispiel für die fonologische Bewusstheit - je nach Verfahren, Zeitspanne der Prognose und vor allem Art des zwischenzeitlichen Unterrichts - zwischen 20% und 80%[92]. Bei derart hohen Fehlprognosen lassen sich keine Fördermaßnahmen, erst recht aber keine Selektionsentscheidungen rechtfertigen - ein Befund, der auch beim Einsatz von Sprachstandserhebungen vor der Schule zu beachten ist.

Demgegenüber stellte Röhr (1978, 259) fest, dass die Einschätzung der KindergartenpädagogInnen eine hohe Trefferquote hatte: 74% der Kinder, denen sie »(sehr) große Schwierigkeiten« in der Schule voraussagten, hatten tatsächlich Probleme - dagegen weniger als 10% derjenigen, für die sie »gar keine« Schulschwierigkeiten vermuteten. Ein Grund für die Stärke des Urteils von PädagogInnen liegt darin, dass sie das Kind über einen längeren Zeitraum und in verschiedenen Situationen beobachten konnten und dass sie oft auch die Schulsituation kennen, in die die Kinder kommen werden.

---

88 Maier (2001, 228).

89 Vgl. Mreschar (1985, 65).

90 ... und generell von Schule als sozialem Raum, der bewusst aus dem Leben herausgelöst wurde (vgl. grundsätzlich dazu: Brügelmann 2005, Kap. 2 und 39-41).

91 Hier zusammengefasst nach Brügelmann (2005a, 167).

92 Vgl. zusammenfassend zum Prognoserisiko von Risikoprognosen: Brügelmann (2005c).

Damit können sie ein grundsätzliches Problem von Leistungsprognosen entschärfen: Begriffe wie »Schulreife« und »Schulfähigkeit« suggerieren nämlich, dass Schwierigkeiten in der ersten Klasse allein auf persönliche Merkmale des Kindes zurückzuführen seien. Mit seinem »ökologischen Modell« hat Nickel[93] die Erfahrung aufgenommen, dass Kinder mit gleichen Voraussetzungen in der einen Schulklasse scheitern, in der anderen aber erfolgreich sind. Je nach Anspruch, aber auch Kompetenz der Lehrperson, je nach Zusammensetzung der Lerngruppe und nach den institutionellen Rahmenbedingungen kommt ein Kind zurecht oder nicht. Entwicklung ist also die Folge einer Interaktion zwischen persönlichen Merkmalen und Kontextbedingungen. Entwicklungsprobleme lassen sich nicht einseitig auf Eigenschaften des Kindes zurückführen, sondern sind als Passungsproblem zu verstehen[94].

#### 1.1.3.2

#### Schule > Fachleistungen über die Schuljahre hinweg

Innerhalb der Schulzeit kann man die Leistungspositionen der SchülerInnen von Jahr zu Jahr vergleichen. In der SCHOLASTIK-Studie des Münchener Max-Planck-Institut für Psychologische Forschung ergaben sich folgende Korrelationen, wenn man die Ergebnisse von Fachtests miteinander vergleicht[95]:

| | Von Jahr zu Jahr (wachsend von Klasse 1-5) |
|---|---|
| .60 bis .78 | für Rechtschreibung |
| .60 bis .70 | für Mathematik |

Für Schulnoten ergeben sich ähnliche Werte; mittelt man sie über verschiedene Fächer, kommt man sogar auf Werte von .80 und höher[96].

Auf den ersten Blick sprechen diese Werte für eine hohe Stabilität und damit Prognostizierbarkeit von Leistungen. An einer Klasse aus der LUST-Studie konnten wir beispielhaft zeigen, dass selbst bei einer Korrelation von .66 über die Gruppe hinweg auf der Einzelfallebene noch mit erheblichen Verschiebungen zu rechnen ist[97]. Für Klassifikationen hat Ingenkamp (1993, 70f) in einer Modellrechnung deutlich gemacht, dass selbst bei einer Korrelation von .70 mit fast 20% Fehlentscheidungen zu rechnen ist.

Verlängert man den Prognosezeitraum, so nehmen die Korrelationen von Noten zudem drastisch ab, z.B. auf .20 bei einer Vorhersage vom ersten bis zum achten Schuljahr[98]. Zielinski (1974b, 889) resümiert: »Die Zusammenhänge zwischen Zensuren der verschiedenen Schulstufen, die bei aufeinander folgenden Jahrgängen noch zufriedenstellend hoch sind, nehmen mit zunehmender zeitlicher Distanz laufend ab, wobei sich ein Wechsel des Schulsystems besonders gravierend bemerkbar macht. Sie liegen die Korrelationskoeffizienten für den Zusammenhang zwischen dem Grundschulzeugnis und dem Erfolg auf weiterführenden Schulen nach 3 bis 6 Jahren im Durchschnitt nur etwa bei .30 ...«

Eine der wichtigsten Nutzungsformen von Beurteilungen betrifft die Übergangsentscheidung nach Klasse 4[99]. In manchen Bundesländern (z.B. Bayern) hängt der Zugang zu einer höheren Schulform unmittelbar vom Notendurchschnitt ab. In anderen bestimmen Noten den Wechsel zumindest indirekt - über die Empfehlung der Schule, an der sich viele Eltern bei der Wahl der weiterführenden Schule orientieren (z.B. in Hamburg). Insofern ist die Klassifikationsleistung der Noten zu prüfen[100].

Roeder (1997, 410) wertet als Erfolg der Prognose, dass unter den Schulformwechslern etwa doppelt so viele *ohne* Gymnasialempfehlung sind wie solche *mit* Empfehlung für das Gymnasium[101]. Dieser Bezug verzerrt aber die Berechnungsbasis, wie Thiel (2005, 255) überzeugend zeigt. Ausgangspunkt muss die Art der Empfehlung sein. Dann stellt man zunächst fest: 1.4% *mit* Gymnasialempfehlung wechseln später auf eine niedrigere Schulform, aber 5-6mal so viele, nämlich 7.6%, von denen ohne Gymnasialempfehlung. Ist die Prognose also doch gut? Nein, denn 92.4% derjenigen, die *keine* Gymnasialempfehlung bekommen haben, schaffen es trotzdem - und das sind mehr als 12mal so viele wie die Abgänger. Der Anteil falscher Prognosen beträgt - auf die Gesamtgruppe bezogen - immerhin 29% (Thiel 2005, 256) - ein schwer zumutbares Risiko für die Betroffenen.

Aufgrund von Befunden aus der Hamburger Studie zur Lernausgangslage in 5. Klassen (LAU) kommentieren Lehmann u.a. (1997, 94) die Prognosevalidität des Urteils von LehrerInnen so, dass im Vergleich zur freien Elternwahl

---

93 Vgl. Nickel (1982).

94 Damit verändert sich auch der Blick auf die Ursachen von Lernschwierigkeiten und Formen der Förderung. So zeigt das Modell des »Teufelskreis Lernstörungen« eindrucksvoll, wie sich punktuelle Lernschwierigkeiten aufgrund geringfügiger Fehlpassung von Leistungsanforderungen und Lernvoraussetzungen zunächst zu übergreifenden Lern*störungen* ausweiten und später als individuelle Lern*schwächen* stabilisieren können - in denen manche Diagnostiker dann die eigentliche Ursache aktueller Leistungsprobleme sehen (vgl. Betz/Breuninger 1987, zusammengefasst nach Brügelmann 2005a, 224).

95 Vgl. Weinert/Helmke (1997b, 467).

96 Vgl. Tent (1998, 583) und zur Prognosekraft von Noten zusammenfassend: Ziegenspeck (1999, 156 ff.).

97 Brügelmann (2005c, 150).

98 Vgl. Tent (1998, 583) - allerdings verschlechtern sich die Werte mit dem Übergang in die Sekundarstufe auch deshalb, weil die Noten in den verschiedenen Schulformen eine unterschiedliche Wertigkeit haben, d.h. ihre Prognosekraft wird vermutlich systematisch unterschätzt.

99 Vgl. dazu Heller (1995; 1997); Lehmann u.a. (1997); Hartinger u.a. (2003); Bos u.a. (2004a, Kap. IX); Faust (2005, 164-167); Thiel (2004). Vgl. zur Verzerrung der Empfehlungen durch die Zugehörigkeit der Eltern zu höheren bzw. niedrigeren sozialen Schichten > Kap.1.2.1 und 7.

100 Vgl. zusammenfassend zu den Studien zum Prognoseerfolg von Empfehlungen: Ingenkamp (1967; 1993); Sauer/Gamsjäger (1996); Thiel (2005, 255 ff.).

101 Auch Heller (1999) wertet die Möglichkeiten einer zutreffenden Zuordnung von SchülerInnen am Ende der vierten Klasse positiv, warnt aber an anderer Stelle selbst davor, »... die Erwartungen an die Gültigkeit von Schulerfolgsprognosen nicht zu hoch ...« anzusetzen (1997, 986).

Entscheidungen durch die Schulen die Zusammensetzung von Klassen in der Sekundarstufe I nicht stärker homogenisieren würden. Anhand der PISA-Daten bestätigt Block (2006, 2): »Jugendliche, die in ihrer Schullaufbahn von einer höheren auf eine niedrigere Schulform wechseln mussten, weisen zum überwiegenden Teil Grundschulempfehlungen für die Schulformen auf, an denen sie letztlich gescheitert sind. [...] 73% aller 15-jährigen Realschüler, die von einem Gymnasium gewechselt sind, haben seinerzeit eine Grundschulempfehlung für das Gymnasium erhalten. Das relative Risiko für Realschüler, einer falschen (zu hohen) Schulform zugewiesen zu werden, ist aufgrund einer unzutreffenden Grundschulempfehlung rund 24 Mal größer als aufgrund falscher (überhöhter) elterlicher Bildungsansprüche. Bei den Hauptschülern, die einen Schulformabstieg hinter sich haben, sind es bundesweit wiederum rund 69%, denen seinerzeit von der Grundschule die Fähigkeit für eine höhere Bildungslaufbahn prognostiziert wurde. Das Risiko eines Hauptschülers, aufgrund der Grundschulempfehlung einer falschen, nämlich zu hohen Schulform zugewiesen zu werden, ist 8 bis 9 Mal größer als die falsche Schulwahl aufgrund übersteigerter Bildungsaspiration der Eltern.«[102]

Und er kommentiert: »… Alle relevanten Studien der letzten Jahre - zuletzt die internationale Grundschulstudie IGLU (Bos, W. u.a. 2004) - zeigen aber, dass in die Beurteilungen der Grundschulen nicht nur rein leistungsbezogene Aspekte Eingang finden: Denn weder Testleistungen noch die von den Lehrkräften vergebenen Noten können die Unterschiede in den Übergangsempfehlungen von Schülern hinreichend erklären. In der Praxis orientieren sich die Grundschulempfehlungen häufig auch an sozialen Kriterien wie z.B. dem Bildungsniveau der Elternhäuser.«[103]

Für das dreigliedrige Schulsystem stellt sich die Legitimationsfrage, wenn die Zuordnung durch Empfehlungen auf der Basis von Noten so fehlerhaft - und kein überzeugender Ersatz in Sicht ist. So hat Thiel[104] festgestellt, dass die Durchschnittsnote sogar noch eine bessere Prognose erlaubt als Schulleistungstests. Insofern warnen auch Bos u.a. (2004b, 225) vor der Hoffnung, das Problem durch die Einführung standardisierter Tests überwinden zu können: »Wenig zielführend wäre vermutlich der Versuch, durch bessere Testverfahren eine normorientierte Verteilung auf die Schulformen zu versuchen. Auf Individualebene gibt es solche Tests nicht und neue zu entwickeln, um eine unanfechtbare langfristige Prognosesicherheit zu gewinnen, dürfte nur schwerlich gelingen. Deshalb muss die Durchlässigkeit der Bildungsgänge weiter ausgebaut werden.«

Auch die Hoffnung, die Prognosevalidität des Lehrerurteils durch ergänzende Informationen »erheblich« verbessern zu können, stellen Sauer/Gamsjäger (1996, 201) in Frage: »Das heißt, zusätzliche Intelligenz- und Motivationstests sowie Ursachenerklärungen von schulischem Erfolg bzw. Misserfolg bringen über die Einschätzung des Lehrers hinaus keine zusätzlichen Informationen.«

Ähnlich resümiert Hopf (1994, 340) für gesonderte Prüfungen beim Übergang von der Grundschule in die Sekundarstufe I: »... die genannten Nachteile ließen sich allenfalls in Kauf nehmen, wenn die Ausleseverfahren für die weiterführenden Schulen - zentral gestellte Normarbeiten, Prüfungen, Beurteilungen durch die Lehrer usw. - den gewünschten Erfolg hätten. Gerade dies ist aber fraglich, wie mehrere empirische Untersuchungen über die Zuverlässigkeit und Genauigkeit der Übergangsauslese ergeben haben. [...] Zensurendurchschnitt und Resultate von Probearbeiten spiegeln ohnehin höchstens einen kleinen Teil der für den Erfolg auf weiterführenden Schulen wichtigen Fähigkeiten wider.«

Andere Studien zeigen zudem, dass auch die in einer bestimmten Schulform erfolgreichen SchülerInnen keine homogene Gruppe darstellen[105]. In verschiedenen Schulklassen können unterschiedliche Schüler»typen« erfolgreich sein. Wie im »ökologischen Modell« von Nickel (1982) für den Schulanfang ist also die Wechselwirkung von individuellen Voraussetzungen sowie institutionellen und didaktischen Bedingungen ([Mindest-]Passung oder nicht) für den Erfolg entscheidend.

#### 1.1.3.3

#### Schule > Studien-/Ausbildungserfolg

Die genannten Probleme verschärfen sich, wenn die Beurteilung schulischer Leistungen die Bewährung in außerschulischen Situationen vorhersagen soll.

Verschiedene ForscherInnen fanden in verschiedenen Ländern[106] Korrelationen von .30 bis .50 für die Vorhersage des Studienerfolgs und Schuler (1998) nennt .41 als Mittelwert[107] diverser Untersuchungen des Zusammenhangs von Schulnoten und Ausbildungserfolg. Dabei wird der theoretische Prüfungsteil der beruflichen Abschlussprüfung besser vorhergesagt als der praktische. »Ähnliches gilt für das Abiturzeugnis, dessen prognostische Gültigkeit für Studien- und Berufserfolg ebenfalls als unzureichend angesehen werden muss. So ergaben sich z.B. zwischen Abiturdurchschnitt

102 Zu Recht wird eingewandt, dass LehrerInnen in manchen Fällen mit ihren Empfehlungen dem Druck der Eltern nachgeben, so dass man ihnen die Fehlprognose nicht anlasten könne. Es bleiben aber auch dann die oben genannten Fehlprognosen in umgekehrter Richtung: Nicht empfohlene SchülerInnen, die *trotzdem* erfolgreich sind.

103 Block (2006, 3).

104 Vgl. Thiel (2005, 238) und seiner Kritik an Klassifikationsversuchen über Tests a.a.O., 54-64.

105 Vgl. die kritische Zusammenfassung der Versuche von Rosemann (1978) und Sauer/Gamsjäger (1996) bei Thiel (2005, 57-62).

106 Vgl. die Zusammenfassungen bei Weingardt (1971b); Schlattmann (1978); Schuler (1998, 370); Trost u.a. (1998, 67).

107 ... korrigiert um methodische Artefakte; eine Korrelation von .3 bedeutet übrigens: Es werden genauso viele nicht geeignete Bewerber aufgenommen, wie geeignete abgewiesen (vgl. Ammann 2002).

und 1. Lehramtsprüfung an Pädagogischen Hochschulen Korrelationen zwischen .29 und .49, zwischen Abiturdurchschnitt und Vordiplom in Physik ein Koeffizient von .37...«[108]

Wer stattdessen auf Tests setzt, sollte aber vorsichtig sein. Gemittelte Abiturnoten sind vorhersagekräftiger als eignungsdiagnostische Verfahren[109]. Und in den USA hat die University of California in Los Angeles nach langen Jahren den fest etablierten SAT als Auswahlinstrument aufgegeben[110]: »The University of California's own research has shown that the SAT I - the widely used ›reasoning‹ test of math and verbal abilities - was the least predictive indicator of freshman academic success, ranking behind high school grades and scores on the so-called ›SAT II‹ achievement tests in various academic subjects.«[111]

Auch im deutschen System wären keine besseren Ergebnisse zu erwarten, wenn man Noten durch Tests ersetzt, wie der folgende Vergleich von Korrelationen zur Vorhersage des Erfolgs im Physikum zeigt[112]:

| Korrelation mit Physikum | Gesamtschule | Gymnasium | Fach-Gymnasium |
|---|---|---|---|
| Abiturnoten | .33 | .48 | .59 |
| Zulassungstest | .48 | .49 | .49 |

Die Korrelation zwischen Abiturnote und dem späteren Berufserfolg fällt allerdings auf .10 und liegt damit schon fast im Zufallsbereich[113]. In einer Schweizer Untersuchung wurden zum Beispiel subjektive Zufriedenheit und objektive Indikatoren für Berufserfolg zu einem Gesamt-Index verrechnet und auf die Maturanoten bezogen. In einer - allerdings kleinen - Stichprobe von 49 (aus 95 befragten) Personen fanden sich kaum Unterschiede zwischen den SchülerInnen verschiedener Notengruppen und sogar eher negative Korrelationen zwischen Maturanoten und Erfolgsindikatoren wie dem Einkommen[114]. Erstaunlicherweise verdienten auch die *ohne* Studienabschluss mehr als die ohne.

Was Schulabschlüsse für den späteren beruflichen und privaten Lebenserfolg bedeuten, lässt sich nicht auf einen einfachen Nenner bringen. Dies zeigen eindrucksvoll auch die Ergebnisse einer gerade veröffentlichten Längsschnittstudie des Züricher Bildungsforschers Helmut Fend (2006). In seiner LifE-Studie wurden ca. 2.000 SchülerInnen der Jahrgänge 1966 und 1967, von der 6. bis zur 10. Klasse begleitet und dann zwanzig Jahre später wieder befragt. Sein Resümee[115]: »In vielfacher Hinsicht ist die Zugehörigkeit der Jugendlichen aus der LifE-Studie zu verschiedenen Schulformen im 9. Schuljahr jedoch nicht lebensbestimmend. Endrucksvoll hat sich gezeigt, dass von den Schulformen zu den Abschlüssen und zur beruflichen Eingliederung noch sehr viel ›Bewegung‹ zu beobachten ist. Schließlich sind viele Bereiche der Lebenszufriedenheit nicht von den Abschlüssen betroffen.«

Damit sind wir beim nächsten Punkt.

### 1.1.3.4

### Studium/Ausbildung > Berufserfolg

Obwohl Studium und Ausbildung stärker auf ein bestimmtes Berufsbild fokussiert sind, verschlechtert sich die Prognosekraft von Noten noch einmal, wenn man Vorhersagen aus dem Bildungserfolg auf den Berufserfolg[116] versucht.

Die Korrelation zwischen Examensnote im Studium und Berufserfolg liegt bei .32, wobei sie von .45 nach einem Jahr auf .11 nach sechs Jahren abnimmt[117]. Mit zunehmender Dauer der Berufstätigkeit werden also andere Faktoren relevant als die durch Noten ausgewiesenen Fachleistungen des Studiums.

Seel (2002, 77) verweist dazu auf Unterschiede zwischen verschiedenen Prüfungsformen. In seiner Follow-up Studie von AbsolventInnen drei bis vier Jahre nach der Diplomprüfung fand er, dass Klausurennoten kaum einen Vorhersagewert für den Berufserfolg haben, wohl aber mündliche Prüfungen, die auf Verständnis prüfen. Aber auch deren Korrelation liegt je nach Erfolgskriterium bei nur .12 bis .35 .

108 Zielinski (1974b, 889).

109 Schuler (1998, 373); auch Hell u.a. (2005) sehen in ihrer Metaanalyse zur Vorhersage des Studienerfolgs im Durchschnitt der Abiturnoten noch den besten Prädiktor, verweisen zugleich aber darauf, dass bisher lediglich die Studiennoten als Erfolgskriterium einbezogen wurden, Studiendauer, -abbruch usw. dagegen nicht.

110 Ich verdanke diesen Hinweis dem Konstanzer Bildungsinfo (25.1.2005) meines Kollegen Georg Lind, der ergänzend dazu schreibt: »Wenn Schulnoten und Testleistungen nicht übereinstimmen, wer hat dann recht? Klarer Fall, sagen die Verkäufer von Tests: die Noten sind unzuverlässig und invalide. Nun, das müssen Test-Verkäufer sagen. Eine Studie der University of California kommt eher zu dem gegenteiligen Schluss: die Noten scheinen valider; sie erlauben eine bessere Vorhersage des Studienerfolgs. Wenn man bei Noten die soziale Herkunft berücksichtigt, verbessert sich die Vorhersagekraft noch, während sie sich bei Testwerten verschlechtert. Das heißt, Tests helfen Kindern reicher Eltern, Einlass zu bekommen in renommierte Universitäten, aber sie sagen weniger über deren Studierfähigkeit aus als die Noten.
Die Universität von Kalifornien (UC) hat ihre Konsequenzen daraus gezogen und den Vertrag mit dem Educational Testing Service (ETS), dem Vertreiber des SAT (College Aufnahme-Tests) gekündigt. ETS will dieses Jahr einen ›validen‹ SAT vorstellen. Die UC ist inzwischen dazu übergangenen, den Spruch eines US-Bundesgerichts umzusetzen, jeden Bewerber individuell zu beurteilen und nach individueller Prüfung über seine Aufnahme zu entscheiden.«

111 Sacks (2004, 7).

112 Klieme (o.J.), zit. nach Köller u.a. (1999, 413).

113 Vgl. Schuler (1998, 372) und die Forschungssynthese von Samson (1984) sowie die Metaanalyse von Roth u.a. (1996). Zu bedenken ist allerdings auch die Schwierigkeit, das Kriterium »Berufserfolg« angemessen zu erfassen: Position in der Stellenhierarchie? Einkommen? Zufriedenheit? ...

114 Vgl. Oberholzer (2002, 17-19).

115 Fend (2006, 53).

116 S. dazu auch die vorhergehende Anmerkung.

117 Vgl. Schuler (1998, 372).

Gebert (1983) wertete 53 Personalbeurteilungen mit acht Dimensionen[118] aus und korrelierte sie mit dem IHK-Abschluss 10 bzw. 20 Jahre vorher. Auch er fand unterschiedlich starke Zusammenhänge, je nach dem gewählten Erfolgskriterium:

.50+ berufl. Fachkenntnisse (sowohl Theorie als auch Praxis)
.40+ Arbeitsgüte, Sorgfalt/Zuverlässigkeit
.20+ Auffassung, Eigeninitiative

Gegenüber diesen - schon an sich geringen - Korrelationen wurde die Prognosen von Arbeitstempo und Führung nicht einmal statistisch signifikant. Bei der Auswahl von BewerberInnen für berufliche Aufgaben haben sich qualitative lernbiografische Daten meist als aussagekräftiger erwiesen als punktuelle Prüfungen[119].

Schon diese wenigen Hinweise zeigen, wie schwierig es generell ist, »Erfolg« aus »Voraussetzungen« vorherzusagen - unabhängig davon, welche Merkmale man mit welchem Verfahren erfasst. Neben der fachlichen Kompetenz spielen persönliche Faktoren wie z.B. die Motivation eine wichtige Rolle. Zum anderen sind die Anforderungen und die Leistungsmöglichkeiten in beruflichen Situationen so unterschiedlich, dass die breite Streuung der »Erfolge« nicht verwundern sollte - selbst wenn die aktuelle Leistung einer Person zutreffend erfasst und bewertet wurde.

### 1.1.4

### Zwischenbilanz zu »Validität«

Lehrerurteile basieren in der Regel auf informellen Leistungsproben und beiläufigen Beobachtungen. Die auf ihnen basierenden Bewertungen haben nur eine eingeschränkte Validität. Denn verschiedene LehrerInnen bewerten nach unterschiedlichen Kriterien: Sie betonen unterschiedliche Aspekte der Leistung und sie orientieren sich zudem an unterschiedlichen Schwellenwerten (z.B. »Welche Leistung entspricht welcher Ziffernnote?«). Diese Probleme treten bei Ziffernoten wie bei verbalen Beurteilungen auf. In letzteren werden sie allerdings sichtbarer als in den Ziffern der Notenskala und damit auch leichter kritisierbar.

Als Möglichkeit, die Validität von Urteilen zu verbessern, wird immer wieder die inhaltliche Präzisierung der Anforderungen und Beurteilungskriterien genannt[120]. In diesem Kontext ist auch die Diskussion um verbindliche »Bildungsstandards« zu sehen. So erhofft man sich von expliziten Kriterien für die Benotung von Aufsätzen eine stärkere Fokussierung der Bewertung. Dies ist in der Tat der Fall - allerdings auch nur begrenzt[121], wie das folgende Kapitel zeigt.

Die Sicherung von Validität ist aber auch eine Schwierigkeit bei der Entwicklung standardisierter Tests. Ihr Vorteil: Die Frage wird ausdrücklich thematisiert und damit werden die Annahmen des Tests für Außenstehende nachprüfbar. Der Einsatz standardisierter Tests bringt aber auch einer Reihe von Problemen mit sich:

▲ Einengung der in einer ökonomischen Erhebung verlässlich erfassbaren Ausschnitte/ Aspekte einer Kompetenz auf ausgewählte Teilleistungen;
▲ Künstlichkeit der Testsituation mit begrenzter Aussagekraft für Alltagsanforderungen.

Bei Tests wird häufig eine Validierung über Expertenurteile angestrebt[122], die aber nicht immer verlässlich sind, wie z.B. die Ergebnisse der deutschen SchülerInnen in den PISA-Aufgaben gezeigt haben: Die vorher befragten ExpertInnen hatten für die einzelnen Aufgaben wesentlich höhere Lösungsquoten vermutet[123].

Zudem konnte weder für Tests noch für das Lehrerurteil eine überzeugende Prognose-Validität nachgewiesen werden. Die Entwicklung von Personen ist nicht berechenbar - und variiert vor allem in Wechselwirkung mit den Lernbedingungen. Damit wird vor allem die Selektionsfunktion von beiden Verfahren nachdrücklich in Frage gestellt.

## 1.2

## Wie unabhängig sind Beurteilungen von persönlichen Einflüssen? (Objektivität)

Aus dem Prinzip der Chancengleichheit folgt, dass die Bewertung von Leistungen nicht davon abhängig sein darf, unter welchen Bedingungen sie zustande kommen (> Kap. 1.3) und wer sie bewertet. Vor allem zu Noten gibt es eine Fülle von Untersuchungen, die diesen Anspruch untersuchen.

### 1.2.1

### Objektivität des Lehrerurteils

Es überrascht auch Laien wenig, wenn Ulshöfer (1949) feststellt, dass 42 DeutschlehrerInnen denselben Aufsatz unterschiedlich bewerten. Wohl aber erstaunt, dass die Noten über das ganze Spektrum von 1 bis 6 streuen. Schröter (1981a) hat den Versuch erweitert und besonders problematische Aufsätze von 11.000 Grund und HauptschullehrerInnen beurteilen lassen[124]. In mehr als 10% der Aufsätze streuten auch hier die Noten über fünf oder gar sechs Stufen. Und auch bei sieben Aufsätzen, die von 72 GymnasiallehrerInnen beurteilt werden sollten, wurden in keinem Fall nur dieselbe Note oder nur benachbarte Noten vergeben.

Nun gelten Aufsätze als besonders anfällig für subjektive Einschätzungen. Aber auch bei anderen Leistungen ergeben sich ähnliche Bilder.

---

118 Auf einer 5er Skala jeweils von 1 bis 5 bewertet.
119 Vgl. Landmesser u.a. (2003) und > Kap 4.4 .
120 Vgl. u.a. Harlen (2004a, 6-7).
121 S. dazu die Studien im > Kap. 1.2.3 »Objektivität«.
122 Baumert u.a. (2001, 43); Artelt u.a. (2001a, 97-101).
123 Artelt u.a. (2001a, 100 vs. 102).
124 Vgl. die Zusammenfassung bei Mreschar (1985, 47). Vgl. zur fehlenden Objektivität von Aufsatzzensuren auch Faigel (1973).

In einer Studie von Weiss (1965)[125] sollten 92 LehrerInnen nur die Rechtschreibung in zwei kleinen Aufsätzen von ViertklässlerInnen benoten. Auch hier streuen die Bewertungen über fünf Notenstufen:

| Note > | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Rechtschreibung Aufsatz A | 10% | 18% | 41% | 24% | 7% | - |
| Rechtschreibung Aufsatz B | 7% | 28% | 39% | 22% | 4% | - |

Er ließ weitere 153 LehrerInnen eine Mathematikarbeit (ebenfalls 4. Klasse) beurteilen, und selbst hier streuten die Noten breit[126]:

| Note > | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Mathematikarbeit | 7% | 41% | 42% | 9% | 1% | - |

In den vorgestellten Studien handelte es sich jeweils um ausgewählte Einzelarbeiten, die den LehrerInnen vorlagen. Aber die Ergebnisse waren nicht anders bei ganzen Klassensätzen (Klink 1964): Verschiedene LehrerInnen legen an dieselbe Arbeit unterschiedliche Maßstäbe an. Ein Grund können Differenzen in der Gewichtung fachlicher Kriterien sein, ein anderer unterschiedliche Erfahrungen mit dem, was man von SchülerInnen einer bestimmten Altersgruppe erwarten kann (> Kap. 2.1).

Gründe für die berichteten Abweichungen gibt es viele. Oelkers (2001) hat die wichtigsten »subjektiven« Fehlerquellen übersichtlich zusammengefasst:

»▲ Halo-Effekt: Ein globaler Allgemeineindruck bestimmt die Wahrnehmung einzelner Merkmale
▲ Beharrlichkeitstendenz: Lehrkräfte rücken von einem bereits gefällten Urteil bei späteren Beurteilungen nicht ab
▲ Reihungseffekt: Unter dem Eindruck, ›es können doch nicht alle gleich schlecht sein‹ werden bessere Noten gegeben
▲ Kontrasteffekt: Nach einer Serie von sehr guten Leistungen wird eine mittelmässige Leistung tendenziell als schlecht bewertet
▲ Beurteilungstendenzen: Milde oder Strenge, ›zentrale Tendenz‹ (Vermeidung von Extremwerten) und ›motivierende‹ versus ›selektive‹ Notengebung
▲ Wissen-um-die-Folgen-Fehler: Mildere Beurteilung bei absehbar negativen Folgen für die Schüler, nicht umgekehrt.«

Diese Fehler wirken generell auf Beurteilungen ein - schon unabhängig von der bewerteten Person. Das Problem verschärft sich aber, wenn man den Einfluss sachfremder Bedingungen systematischer untersucht.

In einer Studie von Hadley (1954) wurden SchülerInnen getestet und parallel von den LehrerInnen nach Beliebtheit eingeschätzt. Diese Daten wurden mit den Zensuren verglichen, die die LehrerInnen den SchülerInnen gegeben hatten. Sie verteilten sich wie folgt[127]:

| | Note besser als Testleistung | Note wie Testleistung | Note schlechter als Testleistung |
|---|---|---|---|
| Beliebteste SchülerInnen | 50 % | | 16 % |
| Durchschnitt | 31 % | | 34 % |
| Unbeliebteste SchülerInnen | 19 % | | 50 % |

Systematische Verzerrungen wurden auch für die Merkmale: Verhalten, Alter, soziale Herkunft, Geschlecht und ethnische Zugehörigkeit nachgewiesen[128].

So veränderte die Vorgabe von Schichtprofilen für die VerfasserInnen von Aufsätzen und Rechenarbeiten (!) die Bewertung derselben Arbeit nach oben bzw. unten - im Durchschnitt um immerhin eine ganze Note[129]. Einflussreich wird dieser Schichteffekt besonders bei den Übergangsempfehlungen von LehrerInnen. Als ein Ergebnis der LAU-Untersuchung stellten Lehmann u.a. (1997) fest: Gemessen an den Testleistungen benachteiligen GrundschullehrerInnen in ihren Empfehlungen für die Sekundarstufe SchülerInnen aus unteren Bildungsschichten (Schwellenwerte bei Vätern mit Abitur 65 Testpunkte, bei Vätern ohne Schulabschluss 97,5 Testpunkte).

Dieser Befund ist in den internationalen Leistungsstudien PISA[130] und IGLU aktuell bestätigt worden: »Untersucht man den Einfluss der Sozialschicht (EGP-Klassen) der Kinder auf ihre Schullaufbahnempfehlungen, so wird deutlich, dass selbst bei Kontrolle der kognitiven Grundfähigkeiten und der Lesekompetenz Kinder aus oberen Schichten eine 2,68- bzw. 1,76-fache größere Chance haben, eine Gymnasialempfehlung zu erhalten als ein Kind aus einem Haushalt aus unteren Schichten«[131].

125 Zusammengefasst bei (Zielinski 1974a, 889).
126 Für Geometrie und andere Fächer fanden schon Starch/Elliot (1913) ähnliche Verteilungen.
127 Zit. nach Zielinski (1974a, 887), der allerdings darauf hinweist, dass die Korrelation zwischen Note und Beliebtheit mit .02 bis .92 über die Klassen hinweg erheblich schwankt. Es gibt also LehrerInnen, bei denen eine enge Beziehung zwischen beiden Faktoren besteht, und andere, bei denen die Noten unabhängig von der Leistung vergeben werden.
128 Vgl. außer den im Folgenden zitierten Studien: Baurmann (1971); Bennett u.a. (1993).
129 Weiss (1965b; 1971, 98-101); Stallmann (1990, 253) hat diesen Befund erneut bestätigt.
130 Vgl. Baumert/Schümer (2001, 357).
131 Bos u.a., (2004, 213).

Auch die ethnische Zugehörigkeit beeinflusst das Lehrerurteil. So stellte Stallmann (1999, 254) fest, dass Migrantenkinder bei gleicher Leistung in Tests schlechtere Noten bekommen. Ditton u.a. (2005, 298-299) haben diese Benachteiligung auch für Empfehlungen von GrundschullehrerInnen beim Übergang zur Sekundarstufe nachgewiesen.

Schließlich spielt das Geschlecht ein bedeutsame Rolle. Nach Carter (1971) bekommen Mädchen[132] bessere Noten und geben Lehrerinnen bessere Noten. Dieser Befund ist allerdings zu differenzieren. So fand Klauer (1992, 56) in Rechenarbeiten, dass Mädchen im Vergleich zu ihren Testleistungen eher schlechter beurteilt wurden. Im Bereich der Schriftsprache erreichen Mädchen zwar bessere Noten - aber sie erbringen auch in Tests bessere Leistungen[133]. Allerdings fanden Bos u.a. (2005, 190-191), dass die Mädchen in Deutsch und im Sachunterricht auch dann noch einen Notenvorteil haben, wenn man die Unterschiede in den Testleistungen berücksichtigt[134]. Der Grund könnte darin liegen, dass LehrerInnen bei Jungen in diesen Bereichen genauer hingucken - oder dass sie deren Leistungen strenger bewerten bzw. sich durch andere Auffälligkeiten beeinflussen lassen.

In einer Sonderauswertung des Schreibvergleichs Bundesrepublik-DDR ging Brügelmann (1994, 31) deshalb von der Bewertungsebene eine Stufe zurück auf die Wahrnehmungsebene und untersuchte in einer Schweizer Stichprobe, ob es schon beim Auszählen von Rechtschreibfehlern geschlechtsspezifische Verzerrungen gibt. Das Ergebnis spricht gegen eine einseitige Bevorzugung eines Geschlechts: Zwar wurden in freien Texten Ende erster Klasse bei Mädchen mehr Rechtschreibfehler übersehen als bei Jungen (3.5 vs. 8.0 Prozentpunkte der Fehlerquote). Im Diktat war es aber genau umgekehrt: Bei den Jungen wurden 13.8 Prozentpunkte der tatsächlichen Fehlerquote nicht angestrichen, bei Mädchen dagegen nur 10.2 Prozentpunkte. In den Texten und Diktaten der zweiten bis vierten Jahrgangsklassen fanden sich nur geringe Unterschiede - und das einmal zugunsten der Mädchen vs. viermal zugunsten der Jungen.

Nimmt man beide Untersuchungsstränge zusammen, so ist die Situation also differenziert zu betrachten: Es gibt zwar Wahrnehmungsunterschiede - diese sind aber nicht geschlechtsspezifisch. Die geschlechtsspezifische Sicht schlägt systematisch erst auf der Bewertungsebene durch.

Insgesamt ist aber festzuhalten: Noten und andere Formen der Einschätzung von Leistungen sind in hohem Maße personabhängig. Als bewusste Empathie hat dies Vorteile für förderorientierte Rückmeldungen. Subjektivität ist insofern die Basis einer ermutigenden Rückmeldung. Denn diese setzt die Bereitschaft und Fähigkeit voraus, sich in die Probleme einer Person, die weniger Kompetenz als der Beurteilende hat, einzufühlen, und ist insofern Ausdruck päagogischen Taktes im Umgang mit ihrer besonderen Verletzlichkeit. Fatal wirken sich dagegen unterschiedliche Maßstäbe und persönliche Sympathie oder der Einfluss von sachfremden Informationen bei Selektionsentscheidungen aus.

### 1.2.2

### Kann der Einsatz standardisierter Tests das Objektivitätsproblem lösen?

Mit der Standardisierung von Aufgaben, ihrer Durchführung und Auswertung soll der Einfluss persönlicher Eigenheiten auf die Leistungsbewertung ausgeschlossen, zumindest kontrollierbar und somit deren Ausweis vergleichbar gemacht werden.

Oberflächlich wird dadurch eine Eindeutigkeit der Bewertung erreicht - allerdings auf Kosten eines neuen Problems: Menschliches Verhalten ist mehrdeutig und deshalb immer interpretationsbedürftig. Dieses Problem stellt sich bei allen Formen der Leistungsbeurteilung, macht sich aber verschärft bei standardisierten Tests bemerkbar. Denn das möglichst eindeutig bestimmte Oberflächenverhalten (z.B. beim Ankreuzen von Auswahlantworten) kann Ausdruck ganz unterschiedlicher Intentionen, Konzepte und Strategien sein. Aufgrund der kontextfreien Kommunikation zwischen Testentwicklern, getesteten Personen und AuswerterInnen lassen sich Interpretationsdifferenzen nicht auflösen: SchülerInnen deuten die Fragen anders, als sie von den AutorInnen gemeint waren[135], und sie kreuzen Antworten aus anderen Gründen an, als die Auswertungsschemata unterstellen. Sprachliche Äußerungen und damit sowohl die Aufgaben als auch die Antworten sind mehrdeutig[136]. Das ist offenkundig bei Übersetzungen, wie Untersuchungen zu PISA belegen. Die Leistungen von SchülerInnen differieren nämlich je nachdem, ob eine Aufgabe aus dem Testpool des betreffenden Landes stammt oder in deren Sprache übersetzt worden ist[137].

Aber auch die oben (> Kap. 1.1.1) referierte Kritik an den Aufgaben von VERA macht deutlich, dass Aufgaben mehrdeutig und auch »falsche« Lösungen je nach Blickwinkel »richtig« sein können. Wie Prüflinge eine Aufgabe gedeutet und wie sie ihre Antworten gemeint haben, ist aber durch die Ausblendung persönlicher Interaktionen nicht mehr verhandelbar. Damit wird nicht Objektivität gesichert, sondern die Subjektivität der TestentwicklerInnen und -auswerterInnen über die der beurteilten Personen privilegiert.

---

132 So auch Hadley (1954), der zugleich feststellte, dass Mädchen auch eher als »sympathisch« eingestuft wurden (s. oben).

133 Vgl. zusammenfassend: Richter/Brügelmann (1994) und Richter (1996).

134 So auch in der Berliner Studie Thiel/Valtin (2002, 72). In Mathematik, wo die Jungen in den Tests besser abschneiden, haben auch sie einen Notenvorteil, aber dieser ist deutlich geringer als die Vorteile der Mädchen, so dass er statistisch nicht signifikant wird (Bos u.a. 2005, 190).

135 Und dies zum Teil mit guten Gründen, vgl. etwa Bartnitzky (2005a).

136 Vgl. zum Problem der »Operationalisierung« ausführlicher Brügelmann (1977).

137 Vgl. Baumgarten u.a. (2005, 101-102).

### 1.2.3

### Wie weit lässt sich das Lehrerurteil objektivieren?

Verschiedene Formen der Objektivierung sind denkbar: methodisch-technisch durch die inhaltliche Präzisierung von Kriterien und Maßstäben bzw. sozial durch die wechselseitige Kontrolle mehrerer PrüferInnen. Beide Maßnahmen können die Streubreite der Urteile reduzieren.

Seit der Veröffentlichung von Ingenkamp (1971a) werden die in > Kap. 1.2.1 referierten Probleme in der Ausbildung immer wieder thematisiert. Birkel (2003) stellt aber fest, dass sich die Situation bei einer Wiederholung der damaligen Versuche nicht verändert hat. Er resümiert verschiedene Studien[138], die für die Sekundarstufe zeigen, dass die Verwendung von Kriterienkatalogen in einigen Fällen die Übereinstimmung von Urteilen über Aufsätze so weit steigern konnte, dass sie in die Nähe der für Tests geforderten Werte kommt. Eher skeptisch stimmen dagegen die Befunde aus einer Studie, in der 30 LehrerInnen eine Stichprobe von Aufsätzen nach 17 Kriterien beurteilt haben. Danach führt der Einsatz solcher Kriteriensätze zwar zu einer Ausdifferenzierung des Urteils, aber weder bei einer Wiederholung der Beurteilung durch dieselben PrüferInnen noch im Vergleich verschiedener PrüferInnen ergaben sich befriedigende Übereinstimmungen: »Die enttäuschend niedrige Korrelation um .50, die den amerikanischen Erfahrungen voll entspricht, besagt, dass in nur 25% aller Fälle das Urteil zweier Beurteiler übereinstimmt. Damit muss die Hoffnung aufgegeben werden, durch den Gebrauch von Kriterien eine Urteilsgerechtigkeit zu erzielen, die die Form des ganz oder zumindest weitgehend übereinstimmenden Urteils aller Beurteiler besitzt.«[139]

In dieser Studie wurden allerdings Kriterien vorgegeben und die BeurteilerInnen nicht speziell in ihrer Anwendung geschult. Für die Auswertung offener Antworten wurde bei PISA ein mehrstufiges Programm entwickelt, um die auf konkrete Aufgaben bezogenen Raster zu optimieren und die BeurteilerInnen zu schulen. Auf diese Weise wurde erreicht, dass 92% der Kodierungen übereinstimmten (Baumert u.a 2001, 42). In anderen Forschungsprojekten mit ähnlich aufwändigen Schulungsmaßnahmen wurde eine Übereinstimmung der Kodierung sprachlicher Äußerungen von 75-85% erreicht (vgl. Diekmann 1995, 493). Solche Formen der Qualitätssicherung sind jedoch für die Anwendung von Auswertungsschemata im jedoch Schulalltag nicht möglich, erst recht nicht für die Bewertung von Leistungen generell, also ohne Verständigung auf spezifische Aufgaben. Insofern sind selbst bei Vorgabe von Beobachtungs- oder Auswertungsrastern zwar eine bessere Übereinstimmung der Urteile[140], aber immer noch deutliche Differenzen zu erwarten.

Das zeigt sich bei der Beurteilung von pädagogischen Prozessen generell. Metz (1982)[141] stellt sogar fest, dass die Schulung von Beobachterinnen mit Hilfe vorgegebener Kriterien die Streuung der Bewertungen eines Videos nicht reduzierte:»Die Urteile von 85 Schulleitern zu einer gemeinsam visionierten Unterrichtseinheit streuten wie eine perfekte Gauß-Kurve über die ganze Breite der Skala. Nach einer intensiven Unterrichtsbeobachtungs-Schulung derselben Personen wechselte zwar ein Großteil der Probanden ihre Einschätzung; nur nahm die Streuung keineswegs ab!«

Die Vorgabe von Kriterien allein reicht also nicht. Angelsächsische Studien verweisen auf die Notwendigkeit, drei Elemente zu kombinieren[142]:

- ▲ klar definierte Kriterien,
- ▲ die möglichst gemeinsam mit den AnwenderInnen erarbeitet und
- ▲ von ihnen während der Anwendung im wechselseitigen Austausch verfeinert werden.

Eine Metaanalyse von mehr als 40 kontrollierten Studien zeigt, dass sich der Aufwand lohnt. Eine Verbesserung der Leistungsbeurteilung *im* Unterricht führte in der Regel dazu, dass auch die Leistungen der SchülerInnen deutlich besser werden[143], und zwar profitieren vor allem leistungsschwächere SchülerInnen von einer differenzierteren Rückmeldung[144].

Unter diesen Bedingungen ist eine stärkere Übereinstimmung der Urteile erwartbar, wie sich auch in einer deutschen Pilotstudie zeigte. Brinkmann (2006) hat in einem Seminar zur Leistungsbewertung ein dreistufiges Verfahren erprobt. In einem ersten Schritt haben Studierende einen Aufsatz spontan beurteilt. Danach wurden diese Bewertungen verglichen, die impliziten Kriterien intensiv diskutiert und in Form eines Beurteilungsrasters zusammengefasst. Anschließend beurteilten die Studierenden einen zweiten Aufsatz. Wie die folgende Tabelle zeigt, haben sich Noten unter der Gruppen-Bedingung (Abstimmung im Team) im Vergleich zur Ausgangserhebung stärker konzentriert. Dennoch bleibt bei der Einzelbewertung eine breite Streuung über mehrere Notenstufen erhalten:

| 1,5 | 2,0 | 2,5 | 3,0 | 3,5 | 4,0 | aM | SD | N |
|---|---|---|---|---|---|---|---|---|
| 1 | 14 | 13 | 10 | 3 | | 2,5 | 5.0 | 41 Gruppen/ (~100 Personen) |
| 4 | 30 | 17 | 5 | 3 | 5 | 2,4 | 6.5 | 64 Personen |
| | 12 | 1 | 3 | 1 | | 2,3 | 5.0 | 17 Gruppen (64 Personen) |

138 U.a. Lehmann (1990; 1994); Beck/Hofen (1991).

139 Grzesik/Fischer (1984, 193; s.a. 184-185, 215).

140 So fanden Meisels u.a. (2001) beim Einsatz von Checklisten eine hohe Übereinstimmung mit externen Kriterien. Lehmann (1990, 92) sieht ebenfalls Vorteile in einer Ausdifferenzierung von Kriterien - aber auch nur in begrenztem Umfang. In den Vordergrund rückt er die Mehrfachbeurteilung.

141 Ref. bei Strittmatter (2003, 11).

142 Hargreaves u.a. (1996); Frederiksen/White (2004).

143 Vgl. Black/Wiliam (1998a+b). Statistisch ausgedrückt beträgt der Zuwachs 0.4 bis 0.7 Standardabweichungen, d.h. ein durchschnittlicher Schüler (d.h. mit ursprünglichem Prozentrang 50) steigt in Vergleichtests immerhin auf einen Prozentrang zwischen etwa 65 und 75.

144 Vgl. Stiggins (1999, 193).

Wichtig ist also die Abstimmung von Urteilen. So könnten die Doppelkorrektur von schriftlichen Arbeiten und Kollegial- statt Einzelprüfungen im Mündlichen Einseitigkeiten entgegenwirken. Allerdings scheint diese Korrektur die Schwankungsbreite nur begrenzt zu dämpfen. Brügelmann (2000b) berechnete - getrennt für die Bereiche Klausuren, mündlichen Prüfungen und Hausarbeiten - im ersten Staatsexamen aus den Bewertungen Durchschnittsnoten bezogen auf die jeweils beteiligten PrüferInnen. Die Bandbreiten der Noten schwankten - bezogen auf die beteiligten PrüferInnen - *innerhalb* der Fächer je nach Prüfungsform zwischen 0,5 und 1,2 Stufen. Trotz der Korrektur durch ZweitgutachterInnen konnten sich also Milde- und Strenge-Effekte immer noch durchsetzen, d.h. die schon gemittelten Noten unterschätzen die Spreizung der einzeln gegebenen Noten noch. Selbst in den gemeinsam durchgeführten und beratenen mündlichen Prüfungen bleibt eine Differenz von 0,5 bis 0,9 Notenstufen - je nach Zusammensetzung der Prüfungsteams. Vergleicht man die Notendurchschnitte *über Fächergrenzen hinweg* erweitert sich die Bandbreite auf 0,9 Notenstufen bei Hausarbeiten, 1,0 bei mündlichen Prüfungen und 2,3 bei Klausuren.

### 1.2.4

### Zwischenbilanz zu »Objektivität«

Unterschiedliche Maßstäbe, aber auch sachfremde Gesichtspunkte wie Sprachstil oder Sozialverhalten des Schülers bzw. persönliche Sympathien der Lehrperson beeinflussen das fachbezogene Urteil und schränken deshalb die Objektivität sowohl von Noten als auch von Verbalgutachten erheblich ein. Nachgewiesen sind auch systematische Verzerrungen durch Gruppenmerkmale wie Geschlecht, soziale Herkunft und ethnische Zugehörigkeit. In Tests werden deshalb Aufgaben, ihre Durchführung und Auswertung standardisiert. Aber auch dieser Versuch hat seine Probleme. Sprache ist nur kontextbezogen verständlich, ihre Bedeutung muss von den Beteiligten stets neu ausgehandelt werden. Genau das ist aber ohne direkte Kommunikation nicht möglich. Strukturierte Beobachtungs- und Auswertungsbögen versprechen, verbunden mit einer Schulung der BeurteilerInnen eine verbesserte - allerdings immer noch begrenzte - Übereinstimmung der Urteile.

So wichtig das Bemühen darum ist, Willkürlichkeit in der Bewertung auszuschließen - die Bedeutung von Empathie für eine lernförderliche Leistungsbeurteilung darf darüber nicht vergessen werden. Dies gilt zumindest für verbale Lernberichte, wie Bambach (1994, 15) in ihrem Plädoyer für »Ermutigungen. Nicht Zensuren« zu Recht anmahnt: »Die Berichte sind nicht nur ›nicht objektiv‹, sondern bewußt subjektiv; an ihnen lässt sich ablesen, was dem berichtenden Lehrer für die ihm anvertrauten Kinder am Herzen liegt, welche Entwicklungen er besonders schätzt, welche er ändern und welche er verhindern möchte. An den Berichten ist auch ablesbar, welchen Lerngegenständen der Lehrer besonderes Gewicht beimißt, welchen seine Vorliebe gilt und welche er als nachrangig betrachtet. Ich vermute, dies alles spielt bei Noten-Zeugnissen ebenso eine Rolle, erkennen allerdings kann man es dort nicht, und deshalb hält sich bei vielen Menschen so hartnäckig die irrige Vorstellung, dass Noten objektiv seien.«

## 1.3

## Wie verlässlich sind verschiedene Beurteilungsverfahren? (Reliabilität)

Dieses Kriterium zielt auf die Verlässlichkeit von methodischen Verfahren. Eine Beurteilung soll von äußeren Umständen (Tageszeit, Reihenfolge der Prüflinge und ähnlichen Bedingungen) unabhängig sein. Die Reliabilität wird in der Regel festgestellt, indem Messungen wiederholt werden und deren Übereinstimmung geprüft wird. Bei Tests, die eine Kompetenz durch den Durchschnitt von Leistungen über mehrere Aufgaben hinweg zu erfassen suchen, ist auch eine Halbierung des Aufgabensatzes und die Berechnung von zwei Teilsummen möglich, deren Übereinstimmung dann ein Maß für die Verlässlichkeit des Verfahrens abgibt.

### 1.3.1

### Die Zuverlässigkeit des Lehrerurteils

Finlayson (1951/1971) ließ LehrerInnen pro SchülerIn zwei Aufsätze beurteilen. Die Noten für die beiden Aufsätze korrelierten im Durchschnitt mit .70. Auch bei Eells (1930/1971) ergab eine Wiederholung der Beurteilung von Aufsätzen durch dieselbe Lehrperson nach einem Monat bzw. vier Jahren die gleiche Streuung wie bei den Noten verschiedener LehrerInnen zum gleichen Zeitpunkt (s. Kap. 1.2.1). Ammann (2002) zitiert eine Studie Osnes (1972), wonach äußere Faktoren wie die Zahl der Rechtschreibfehler oder Handschrift die Bewertung von Aufsätzen beeinflussen. Die Bewertung von Aufsätzen ist außerdem abhängig von der Situation: in der Reihenfolge spätere erhalten eine bessere Note[145]. Auch der Kontext der Beurteilung spielt eine Rolle: Nach einer guter Arbeit wird eine schlechte noch schlechter beurteilt (Birkel 1978/1984)[146].

Aber es sind nicht nur die Aufsätze, deren Beurteilung für den Einfluss von Randbedingungen anfällig ist. Dicker (1973)[147] ließ dieselben Mathematikarbeiten von 24 HauptschullehrerInnen nach drei Monaten erneut bewerten. Nur acht, also 1/3 der LehrerInnen, gab dieselbe Note, dem entspricht eine Korrelation von .50 . Noch ungünstiger fiel das Ergebnis von 61 LehrerInnen aus, die zwei bzw. drei Arbeiten in Geschichte und Geografie zweimal zu bewerten hatten[148].

145 Baurmann (1975).

146 Man kann solche Reihungs- und Kontrasteffekte auch als Einschränkung der Objektivität interpretieren, s. > Kap. 1.2.1.

147 Zusammengefasst bei Zielinski (1974a, 888).

148 Eells (1930/1971).

Auch in mündlichen Prüfungen streut das Notenniveau nicht zufällig. Vielmehr lässt sich ein Auf- und Absteigen des Durchschnitts beobachten, besonders stark bei einer höheren Zahl von Prüfungen pro Tag (Hartog/Rhodes 1971b).

Festzuhalten ist, dass Schwankungen des Urteils derselben Lehrperson die Verlässlichkeit der Noten und Verbalgutachten gleichermaßen beeinträchtigen.

### 1.3.2

### Die Zuverlässigkeit von Tests

Aber auch bei Tests gibt es Schwierigkeiten mit der Verlässlichkeit. Schon die Wiederholung desselben Tests führt nicht zu denselben Ergebnissen. In unserem Projekt LUST erhielten wir bei einer Reliabilitätsprüfung desselben, sehr robusten Lesetests nicht nur - wie erwartet - beim zweiten Mal deutlich bessere Ergebnisse; die Rangfolgen der Leistungen korrelierten nach einer Woche immerhin noch mit .90[149]. Bei der Durchführung in einer anderen Form (PC vs. Papier-und-Bleistift) sank die Korrelation aber schon deutlich auf .70[150]. Bei Tests spielen auch andere Durchführungsbedingungen eine Rolle, nicht nur die Tagesform der SchülerInnen. Dies wird besonders deutlich in Einzelfallstudien, in denen einzelnen SchülerInnen derselbe Test zweimal oder zwei Tests mit gleichem Schwerpunkt gegeben werden. Dabei zeigt sich, wie riskant die Einstufung einer Person nach einmaliger Testung ist[151]. Das Problem von Tests ist also die breite Schwankung einer punktuell erfassten Testleistung um den »wahren Wert« der eigentlich angezielten Fähigkeit (= hoher Messfehler bei Individualdaten). In Aussagen über größere Gruppen, wie sie für bildungspolitische Entscheidungen genutzt werden, stellt sich dieses Problem in geringerem Umfang, weil sich individuelle Schwankungen in den Kennwerten für die Stichprobe insgesamt ausgleichen. Insofern liefern Studien wie PISA, IGLU und VERA verlässliche Daten für eine schulübergreifende Systemevaluation. Ihre Daten haben aber nur einen begrenzten Stellenwert für die Bewertung individueller Leistungen von SchülerInnen (oder auch LehrerInnen ...).

### 1.3.3

### Zwischenbilanz zu »Reliabilität«

Auf der Individualebene sind sowohl Lehrerurteile als auch Tests sehr unzuverlässig. Punktuelle Leistungsproben bzw. Beobachtungen reichen deshalb in keinem Fall aus, um institutionelle Förder- oder gar Selektionsentscheidungen abzusichern. Je folgenreicher die Entscheidung für die Betroffenen, um so weniger darf man sich auf eine einzige Leistungsprobe verlassen. Außerdem sollten die Aufgabentypen variieren, um Zufallseffekte der Situation zu minimieren (z.B. mündliche vs. schriftliche Aufgaben; offene vs. geschlossene Fragen).

### 1.4

### Fazit

Gemessen an den drei Gütekriterien weisen alle Erhebungsformen Mängel auf. Diese Einsicht relativiert den Status von Bewertungen. Die Diskussion hat aber auch gezeigt, dass die Gütekriterien in ihrem testtheoretischen Verständnis dem Gegenstand nicht voll gerecht werden: Menschliches Verhalten ist kontextabhängig und mehrdeutig. Ohne kognitive und emotionale Empathie kann es oft weder erklärt noch angemessen gewürdigt werden. Es kommt hinzu, dass Beschreibungen und Bewertungen für die Betroffenen nicht nur kognitiv nachvollziehbar, sondern auch sozial annehmbar sein müssen: Damit werden Standards wie Glaubwürdigkeit, Fairness und Verständlichkeit bedeutsam, die hier noch gar nicht bedacht sind[152] (> Kap. 6.5).

## 2

## An welchen Maßstäben sollen Leistungen gemessen werden? (Bezugsnormen)[153]

Die Bewertung einer Leistung kann sich an verschiedenen Maßstäben orientieren[154]:

▲ *Kollektive Norm/Gruppenorientierung:* Vergleich mit anderen Personen einer Bezugsgruppe, z.B. einer Klasse, der Altersgruppe oder des Jahrgangs einer bestimmten Schulform; sie ist verbunden mit dem Anspruch der *Höchstleistung* wie z.B. im sportlichen Wettkampf, in dem es auf die relative (»Sieger«) oder absolute Bestleistung (»Rekord«) ankommt;

▲ *Sachnorm/Kriteriumsorientierung:* Feststellung, wie weit eine Leistung den in Lernzielen definierten Anforderungen entspricht; sie ist bezogen auf eine allgemein geforderte *Mindestleistung,* wie sie für die Sicherung alltagstauglicher Fähigkeiten notwendig ist (z.B. bei »Führerscheinen«, bei dem es nur um das Urteil »bestanden« geht);

▲ *Individualnorm/Entwicklungsorientierung:* Bestimmung des Lernzuwachses, bezogen auf die unterschiedlichen Voraussetzungen einzelner Personen, zum Beispiel in der Rehabilitation nach einem Unfall, die auf eine weitestgehende Förderung der individuell vorhandenen Möglichkeiten zielt (»Fortschritt« von den jeweiligen Voraussetzungen her als Maßstab des Erfolgs).

---

149 Vgl. Brügelmann (2003c, 6).

150 Vgl. Backhaus/Moskopp (2006, 4).

151 Vgl. als ein Beispiel unter vielen Seidel (2005; 2006).

152 Vgl. zu Kritik an einem verkürzten Verständnis von Gütekriterien: House (1980) und Winter (2004, 91-95).

153 Empfehlenswert als Einführung sind die Überblicke bei Klauer (1987); Rheinberg (1998; 2001, 59-68) und Persy (1990).

154 Vgl. zur Erläuterung dieser Bestimmungen: Brügelmann (1998).

Je nach Maßstab wird dieselbe Leistung anders bewertet. In der Praxis dominiert wegen der unterschiedlichen Funktionen mal der eine, mal der andere Maßstab: Für die Auswahl von BewerberInnen auf knappe Stellen in einem Betrieb oder in einer Bildungseinrichtung ist der Vergleich mit anderen angemessen; für die Zulassung zu einer Tätigkeit, deren Folgen andere betreffen, zum Beispiel im »Erste-Hilfe«-Kurs, macht die Überprüfung definierter Mindestanforderungen Sinn; zur Rückmeldung über Effekte des Unterrichts oder den Erfolg individueller Lerntätigkeit ist eher der Ausweis von Leistungsfortschritten angemessen[155].

Die Wahl des Maßstabs ist *unabhängig* von der Wahl des Erhebungsverfahrens - z.B. Test vs. Beobachtung - oder der Entscheidung für eine bestimmte Form der Dokumentation und Rückmeldung (wie Ziffernnoten vs. Verbalbeurteilung). Die unterschiedliche Nutzung von Tests für verschiedene Funktionen kann dies verdeutlichen[156]:

▲ Gruppenorientierte Schulleistungstests sind Tests, bei denen das individuelle Ergebnis mit den Ergebnissen einer relevanten Stichprobe, z.B. Klassenstufe, verglichen werden kann. Als Vergleichswerte und Informationen werden meist Prozentrangplätze benutzt, die auch in Noten umgerechnet werden können. Geeignet sind solche Tests für die Auslese und bei Wettbewerben.

▲ Kriteriumsorientierte Schulleistungstests sind Tests, bei denen das individuelle Ergebnis mit einem vorher gesetztem Kriterium (Lernziel) verglichen wird. Mit Bezug auf verschiedene Kriterien sind unterschiedliche Vergleiche denkbar. Die Erfüllung von Grundlagenanforderungen oder Minimallernzielen (Kriterium) kann geprüft und mitgeteilt werden. Alternativ kann das Erreichen eines bestimmten Lernzielniveaus bzw. unterschiedlicher Anforderungen überprüft und ggf. bestätigt werden.

▲ Diagnostische Schulleistungstests sind kriteriumsorientierte Tests, bei denen das Ergebnis aus dem Vergleich mit dem Kriterium zur Feststellung und Interpretation von Abweichungen und zur Planung und Durchführung von fördernden Maßnahmen verwendet wird.

Die beschriebenen Ziele sind Ansprüche bzw. Erwartungen. Zu prüfen ist aber, wie die Realität aussieht. Dabei geht es an dieser Stelle - wie gesagt - zunächst nicht um eine Untersuchung von *Ziffern*noten und *Verbal*gutachten, sondern grundsätzlicher um die Wirkungen der Bezugsnormen unabhängig von der Darstellungsform.

## 2.1

## Wo steht ein Schüler im Vergleich zu anderen? (kollektive Norm/Gruppenorientierung)

Verfechter von Ziffernnoten betonen als einen ihrer Vorteile ihre angebliche Vergleichbarkeit. In der Tat zeigen verschiedene Studien, dass LehrerInnen innerhalb ihrer Klassen zu ähnlichen Rangfolgen kommen wie Tests. Die Rangfolge von Noten einerseits und von Leistungswerten in Fachtests andererseits stimmen innerhalb einzelner Klassen relativ gut überein[157]. Allerdings dürfen solche Werte nicht überschätzt werden: Individuell kann es auch bei Korrelationen in diesem Bereich erhebliche Rangverschiebungen geben[158]. Über verschiedene Fächer hinweg sinkt die Vergleichbarkeit von Noten noch mehr, da z.B. Mathematik und Rechtschreiben (unter voller Ausnutzung der Notenskala) strenger bewertet werden, Sport und Kunst dagegen im Durchschnitt um eine halbe bis ganze Note milder[159]; und auch auf verschiedenen Klassenstufen[160] gelten unterschiedliche Anforderungen. So halbiert sich z.B. der Anteil der »(sehr) guten« Noten in Rechtschreibung von rund 70% in Klasse 2 auf gut 30% in Klasse 6[161].

Vor allem aber sinken die Korrelationen beim Vergleich über verschiedene Klassen hinweg noch einmal erheblich. So stellte schon Schiefele (1960) fest, dass dieselbe Fehlerquote in demselben Diktat in verschiedenen Klassen unterschiedlich benotet wird. Bekam ein Schüler in der einen Klasse noch mit 12 Fehlern eine »3«, gab es in einer anderen Klasse bereits ab 6 Fehlern eine »4«[162]. Thiel/Valtin (2002, 76) folgern sogar: »Die Klassenzugehörigkeit ist [...] entscheidender als die Testleistungen«.

Auch wenn das Niveau der jeweiligen Klasse einen beachtlichen Einfluss hat, so ist diese Deutung überzogen. So fand Backhaus (2006) in einer Sekundärauswertung der Lese-Studie LUST[163], dass die die Zugehörigkeit zur Klasse zwar 6% bzw. 21% der Unterschiede in den Noten aufklärt, die Testleistung in den Jahrgängen 3 und 4 aber 27% bzw. 34%, Auch die Reanalyse der Daten im Berliner NOVARA-Projekt durch Thiel[164] kommt am Beispiel des Mathematiktests im SL-HAM 6/7 zu dem Schluss, dass die Testleistung mit etwa 40% den größten Teil der Notenunterschiede erklärt, die Zugehörigkeit zur Schulklasse dagegen nur bis zu 10%.

Wie man die Abweichung der Noten von der Testleistung bewertet, hängt davon ab, wie hoch man die Aussagekraft punktueller Tests im Vergleich zur kontinuierlichen Beobachtung durch die Fachlehrerin einschätzt. Unumstritten ist aber der Sachverhalt, dass die Noten *innerhalb* einzelner Klassen wesentlich höher mit den Tests korrelieren als über verschiedene Klassen hinweg. Insofern ist auf alle Fälle belegt, dass

155 Vgl. zur Begründung im Einzelnen die Beiträge von Bartnitzky, Flitner, Schwartz, Röbe und Knauf in Bartnitzky/Portmann (1992, 8-47).
156 Gaude (1989, 192); s.a. Ingenkamp (1992).
157 So berichten beispielsweise Thiel/Valtin (2002, 75) Korrelationen von .50 bis .88; s. ergänzend zu ähnlichen Übereinstimmungen in den Studien IGLU und LUST > Kap. 1.1.2.
158 Vgl. > Kap. 1.1.2 .
159 A.a.O., 69-70.
160 Vor allem in den verschiedenen Schulformen der Sekundarstufe I.
161 Thiel/Valtin (2002, 71); in Berlin Abschlussklasse der dort sechsjährigen Grundschule.
162 Zusammenfassung bei Zielinski (1974a, 884).
163 Vgl. Brügelmann (2003b+c; 2005b),
164 Pers. Mitteilung v. 3.2.06.

die unterschiedlichen Maßstäbe von LehrerInnen eine bedeutsame Rolle für die Benotung von Leistungen spielen.

Diese Abweichungen werden in der Regel als Schwäche des Lehrerurteils ausgelegt, was aber nicht zwingend ist (s. > Kap. 1.2). Die Achillesferse des Lehrerurteils ist der klassenbezogene Maßstab. Insofern ist der punktuelle Einsatz von Tests, die in größeren Stichproben normiert wurden, wichtig, um die eigenen Maßstäbe auf *mögliche* Verzerrungen hin zu überprüfen. In der Beurteilung einzelner SchülerInnen können abweichende Testergebnisse zudem auf blinde Flecke aufmerksam machen. Ersetzen können punktuelle Tests die langfristige Beobachtung jedoch nicht: In beiden Fällen sind abweichende Testergebnisse *Anlass* für eine Überprüfung - in keinem Fall aber eine unbefragt hinzunehmende Autorität.

Ein spezielles Problem der gruppenorientierten Bewertung verdient besondere Beachtung: Die Orientierung an der Gauß'schen Normalverteilung (»Glockenkurve«) verzerrt die inhaltliche Bedeutung von Leistungsunterschieden. An einem Beispiel aus dem Sport lässt sich die Problematik eindrücklich zeigen: Während beim Rodeln in der Freizeit Abstände von mehreren Sekunden über Sieg oder Niederlage entscheiden können, müssen im Vereinssport schon Unterschiede von Zehntel- und bei Olympischen Spielen von Hundertstel- oder Tausendstelsekunden herangezogen werden, um Leistungen differenzieren zu können - auch wenn sie für den Alltag irrelevant sind.

Für das Lehrerurteil wie auch für Tests kann die Unterstellung einer Verteilung nach der Gauß'schen Normalverteilung Fehldeutungen nahe legen, wenn selbst kleine Leistungsunterschiede um der Notendifferenzierung willen überbewertet werden. So können beispielsweise in manchen Diktaten schon zwei Fehler mehr auf hundert Wörter eine ganze Notenstufe - oder bei Tests einen Sprung um zehn oder zwanzig Prozentränge ausmachen.

Ein zweites Problem stellt die Wirkung auf schwächere SchülerInnen dar: Obwohl sie Lernfortschritte machen, können diese nicht honoriert werden, da sich ihr Rangplatz wegen des Lernzuwachses aller SchülerInnen in der Regel nicht verändert. Damit kann ihre Lernmotivation sinken[165].

Die Annahme, dass sich gute und schwache Leistungen entsprechend einer Glockenkurve verteilen, ist nicht notwendigerweise richtig. Zwar kann man durch eine entsprechende Gestaltung von Aufgaben und die Art ihrer Auswertung eine solche Normalverteilung sichern, »[l]ogische Schwierigkeiten, diese ›Normalverteilung‹ beim Zensurengeben anzuwenden, resultieren aber aus diesen Tatsachen:

- Nicht alle Eigenschaften sind normal verteilt.
- Die Natur der Verteilung hängt teilweise vom Messinstrument ab. Das gilt vor allem für Tests, da sie oft keinen definiten Null-Punkt haben und keine absoluten Messeinheiten.
- Auch wenn eine Eigenschaft normal verteilt sein sollte, gilt dies nur für große, unselektierte Gruppen von Menschen. Viele Schüler- und Studierendengruppen sind stark selektiert, womit die Annahme einer Normalverteilung sehr zweifelhaft ist. Wenn diese Personen noch dazu sich in der Ausbildung befinden, um bestimmte Änderungen bei ihnen zu erzeugen, gibt es starke Zweifel, ob die Verteilung ihrer Fähigkeiten ›normal‹ sein wird. Nehmen Sie, zum Beispiel, an wir würden eine solche gedankenlose Anwendung der ›normalen‹ Verteilung auf solche Eigenschaften machen wie die Verteilung der Lehrkompetenz unter Lehrern. ›Offensichtlich‹ müssten dann einige Lehrer als Versager bezeichnet werden. ›Offensichtlich‹ müsste ihre Zahl genau so groß sein, wie die der herausragenden Lehrer. Lehrer werden schnell Gründe finden, um eine solche Annahme bei der Evaluation ihrer Arbeit zurück zu weisen. Diese Einwände gelten gleichermaßen für die Evaluation der Arbeit von Schülern und Studierenden.«[166]

## 2.2

## Wo steht ein Schüler auf dem Weg zum Lernziel? (Sachnorm/Kriteriumsorientierung)

Die Kultusministerkonferenz hat bereits in ihrem Beschluss von 1968 gefordert, die Bewertung von Leistungen nicht am Klassendurchschnitt, sondern an definierten Anforderungen zu orientieren. Wie die in > Kap. 3.1 berichteten Studien zeigen, hat sich dieser Maßstab bei *Ziffernnoten* bisher nicht durchgesetzt. Eine größere Rolle spielt er bei *Verbalgutachten* und insbesondere in den zentralen Leistungsvergleichen auf internationaler Ebene (PISA, IGLU) wie auch in den Bundesländern (VERA)[167].

Bei der Entwicklung und Auswertung von Tests versucht man - orientiert an den sog. »Bildungsstandards«[168] - Kompetenzstufen zu definieren, die eine zunehmende Annäherung an das Lernziel beschreiben. Dabei stellt sich allerdings ein Problem: Lernen wird modelliert als eindimensionaler und linearer Zuwachs von Kompetenz.

Diese Vereinfachung wird der Komplexität von Lernprozessen gerade in der Anfangsphase nicht gerecht: So führt zum Beispiel der Wechsel vom wortweisen Satzlesen zum inhaltsorientierten Textlesen einerseits zu einem wachsendem Tempo und besseren Inhaltsverständnis, aber gleichzeitig - zumindest phasenweise - auch zu mehr Verlesungen auf Wortebene. Lerngewinne lassen sich also nicht immer als bloß quantitative Reduktion von Fehlerquoten messen. Notwendig sind differenziertere Leistungsprofile, deren Ergebnisse vor dem Hintergrund von qualitativen Entwicklungsmodellen inhaltlich gedeutet werden müssen.

---

165 S. zu empirischen Befunden > Kap. 3.2.3.1.

166 Dressel (1957, 7-8; Übersetzung: Georg Lind).

167 Allerdings wurden auch in diesen Studien die Anforderungen der einzelnen Stufen nicht normativ vorweg, sondern erst im Nachhinein auf der Basis der empirischen Ergebnisse formuliert. Vgl. zu den Zweifeln an der ökologischen Validität dieser Festlegungen: Brügelmann (2005, 277) mit Verweis auf Testergebnisse bei Erwachsenen in der LUST-Studie (ausführlicher: Brügelmann 2004).

168 Vgl. grundlegend: Klieme u.a. (2003); zu den Problemen, vor allem bei der Umsetzung: Brügelmann (2005, 46-48).

Krampen (1985, 117) stellt für inhaltliche Kommentare zu Noten, die die Leistung auf Lernziele beziehen, grundsätzlich positive Wirkungen auf Motivation und Leistung fest[169]: »An einem sachlichen (lehrstoff-bezogenen) Gütemaßstab orientierte Kommentare wirken in der Tendenz bei allen Schülern positiv, ohne dass gesagt werden kann, dass eine Leistungsgruppe von ihnen besonders profitiert; die Effekte sind jedoch eher gering.«

In anderen Studien wurden für differenzierte Rückmeldungen, die sich an Lernzielen als Sachkriterium orientierten, unterschiedliche Effekte gefunden: Leistungsschwächere SchülerInnen profitierten von ihnen, während leistungsstärkere bei einer Rückmeldung nach sozialer Bezugsnorm besser abschnitten[170].

Die teilweise nur geringe Ausprägung der Effekte könnte damit zusammenhängen, dass die (vergleichsorientierten) Noten in der Wahrnehmung der SchülerInnen dominieren, der kriteriumsorientierte Kommentar seine Wirkung also nicht voll entfalten kann.

## 2.3

## Welche Fortschritte hat ein Schüler gemacht? (individuelle Norm/Entwicklungsorientierung)

Der Lernfortschritt spielt für die Vergabe von Ziffernnoten praktisch kaum eine Rolle. Nach dem Beschluss der KMK von 1968 sollte sie sich an *Lernzielen* orientieren. Aber selbst diese Kriteriumsorientierung hat sich kaum durchgesetzt. Faktisch ist die Benotung fast ausschließlich am Durchschnitt der Lerngruppe ausgerichtet. Die Orientierung an der Individualnorm wird vor allem für Verbalgutachten gefordert. In ihnen geht es nicht nur um eine differenziertere Beschreibung des erreichten Leistungsstands, als dies durch Ziffernnoten möglich ist. Vor allem können auch die Bedingungen verdeutlicht werden, unter denen SchülerInnen die beschriebene Leistung erbracht haben. In Form eines ›Entwicklungsberichts‹ können Leistungen auf die jeweiligen Ausgangsbedingungen bezogen und damit - selbst bei gleicher Punktzahl in einem Test - als individuell unterschiedlicher Zuwachs ausgewiesen werden. Leistungsbeurteilung zielt außerdem nicht nur auf den Ausweis von erworbenen Kompetenzen (»summative« Bewertung). Sie hat auch eine wichtige Funktion für die Förderung von Lernen (»formative« Bewertung).

Entgegen dem explizit vertretenen Anspruch beziehen sich aber auch Verbalbeurteilungen nur in wenigen Fällen auf die individuelle Entwicklung[171]: Nach Zeugnisanalysen kommt die individuelle Bezugsnorm in maximal 10% der Aussagen zur Geltung. Insofern verwundert es nicht, dass sich die Wirkungen unterschiedlicher Zeugnisformen in Feldstudien des Alltagsunterrichts so wenig unterscheiden: Bei gleichem Maßstab sind keine unterschiedlichen Effekte auf die Selbstwahrnehmung zu erwarten. Dort allerdings, wo die individuelle Bezugsnorm realisiert wird - und sei es nur *neben* den anderen Maßstäben - zeigen sich positive Effekte auf Motivation und Leistung.

Eine Metaanalyse von Kluger/deNisi (1996) hat Studien aus ganz verschiedenen Bereichen ausgewertet. Jacobs (o.J.) resümiert die Ergebnisse: »Die Rückmeldung über eine individuelle Leistungsentwicklung, etwa der Feedbackhinweis auf eine Verbesserung gegenüber vorheriger Leistung [...] wird von mir überwiegend als motivationales Feedback (Leistungsbewertung im Längsschnitt) betrachtet. 50 Effektstärken[172] beziehen sich auf den Vergleich ›Feedback individuelle Leistungsveränderung‹ vs. ›kein Feedback‹ und bestätigen eine leistungssteigernde Wirkung dieser Rückmeldung in Höhe einer durchschnittlichen Effektstärke von d = .55. Sonstige Leistungsstandards im Feedback, wie etwa der ›Vergleich mit den Leistungen anderer Personen‹ bzw. Noten als Rückmeldung waren offenbar nicht so wirksam.«

Da sich unter den ausgewerteten Studien auch Laborexperimente, darunter viele aus ganz unterschiedlichen Bereichen, befinden, deren Bedeutung für den Schulalltag ungeklärt ist, sind die Ergebnisse nur als erster Hinweis zu nehmen. Einschlägiger ist eine kontrollierte Feldstudie von Krampen (1985). Er untersuchte über mehrere Mathematikarbeiten hinweg vier Formen der Rückmeldung:

Ziffernnoten

Ziffernnoten
plus Kommentar,
der sich an der Gruppennorm orientierte

Ziffernnoten
plus Kommentar,
der sich an der Kriteriumsnorm orientierte

Ziffernnoten
plus Kommentar,
der sich an der Entwicklungsnorm orientierte.

169 Generell lernförderliche Effekte von inhaltlichen Kommentaren stellte Page (1992) aufgrund verschiedener US-amerikanischer Untersuchungen fest.

170 Vgl. Lissmann/Paetzold (1987).

171 Vgl. hierzu und zu den anschließenden Kommentaren die Belege > Kap. 3.1.

172 Anm. der Verf.: Die Effektstärke ist ein statistisches Maß für die Bedeutung von Unterschieden zwischen den Mittelwerten zweier Gruppen. Sie ist umso größer, je weniger sich die Verteilungen der beiden Gruppen überlappen. Rechnerisch wird der Wert bestimmt, in dem die Differenz der beiden Mittelwerte durch die Standardabweichung (als Maß für die Streuung) in der Kontrollgruppe dividiert wird. Effektstärken von mehr als .50 gelten als beachtlich.

Krampen stellte positive Auswirkungen der Individualnorm auf die Motivation und die Leistungen aller SchülerInnen fest, wobei sie - ähnlich wie Lissmann/Paetzold (1987; s. Kap. 2.2) - erwartungsgemäß bei den leistungsschwachen SchülerInnen besonders ausgeprägt waren[173]: »Über Lehrerkommentare zu Leistungen ist folgendes bekannt (Krampen 1987):

1.
Sozial orientierte Lehrerkommentare wirken
bei leistungsschwächeren Schülern deutlich negativ,
bei leistungsstärkeren neutral oder leicht positiv.

2.
An einem sachlichen Standard orientierte Lehrerkommentare wirken in der Tendenz bei allen Schülern positiv, ohne dass eine bestimmte Leistungsgruppe deutlich von ihnen profitiert.

3.
Individuell orientierte Lehrerkommentare wirken
ebenfalls bei allen Schülern tendenziell positiv,
am meisten profitieren davon die leistungsschwächeren.«

Auch nach anderen Untersuchungen lassen sich positive Wirkungen der individuellen Norm auf das Selbstwertgefühl, die Motivation und Erfolgszuversicht feststellen[174]. Im Blick auf die fachliche Selbsteinschätzung stellte etwa Rheinberg (2001, 64, 65) fest, dass »... mehr als die Hälfte der Schüler von Lehrern, die sich ausschließlich an sozialen Bezugsnormen orientierten, am Schuljahresende sagten, sie könnten jetzt nur gleichviel oder sogar weniger (!) als zu Schuljahresbeginn [...] Bei Lehrern, die sich nicht nur an sozialen, sondern auch individuellen Bezugsnormen orientierten, gaben immerhin zwei Drittel der Schüler an, sie könnten jetzt am Schuljahresende mehr als zu Schuljahresbeginn«.[175]

Damit relativiert diese Untersuchung einige ältere Untersuchungen, nach denen die individuelle Bezugsnorm den Konkurrenz- und Leistungsdruck nicht gemindert hat (vgl. Lissmann 1981).

Auch die Leistungsentwicklung wird positiv beeinflusst, wenn die individuellen Lernfortschritte bei der Beurteilung stärker berücksichtigt werden. Dies stellte Rheinberg (1998) in der Auswertung verschiedener Studien fest: »Es zeigte sich, dass leistungsschwächere Schüler von der individuellen Bezugsnorm besonders profitieren, ohne dass leistungsstärkere benachteiligt wären. Allerdings ist hier einschränkend zu beachten, dass in (fast) allen Untersuchungen die individuelle Bezugsnorm als *zusätzliche* Beurteilungsperspektive eingeführt war, d.h. in Kombination mit anderen Bezugsnormen auftrat ...« (Rheinberg 2001, 65).

Hartinger/Fölling-Albers (2002, 119) resümieren die Ergebnisse verschiedener Studien: »Individuelle Bezugsnormorientierung von LehrerInnen korreliert positiv mit günstigen Attributionen[176] und einer höheren Leistungsmotivation der Schüler/innen. Daneben zeigen diese Schüler/innen auch weniger Furcht vor Misserfolg und mehr Hoffnung auf Erfolg - wichtige Faktoren für positiv motiviertes Lernverhalten. Dies resultierte dann in mehr Freude am Unterricht und letztendlich auch in besseren Lernleistungen.«

Wichtig ist allerdings, dass die Bewertung sich nicht allein an quantitativen Fortschritten orientiert (Zuwachs an richtigen Lösungen). Da Lernfortschritte sich auch in zunehmender Fehlerzahl ausdrücken können (z.B. bei der Übergeneralisierung neu gelernter Rechtschreibmuster), ist oft eine qualitative Bewertung der Entwicklung erforderlich. Diese kann durch standardisierte Messungen nicht geleistet werden. Eine Interpretation durch die fachkundige Lehrkraft ist immer notwendig.

## 2.4

## Zwischenbilanz zu »Bezugsnormen«

Trotz der Vorgaben der KMK (1968) dominiert bei der Notenvergabe die Gruppennorm - bezogen auf die einzelne Klasse. Aber auch für Verbalgutachten spielt sie neben der Kriteriumsorientierung eine wichtige Rolle: Die individuelle Bezugsnorm kommt nur in einer Minderheit der Aussagen zur Geltung. Grundsätzlich hat eine Rückmeldung von individuellen Lernfortschritten (statt einer Bewertung von Leistungen im Vergleich mit einer Bezugsgruppe) positivere Effekte auf leistungsschwächere SchülerInnen: ihre Motivation ist höher, ihre Selbsteinschätzung ist positiver und ihre Leistungen sind besser. Aber je nach Funktion haben auch die Zielorientierung und der Gruppenvergleich ihre Berechtigung - als ergänzende Information (vgl. > Kap. 6.4).

Oft wird die individuelle Bezugsnorm allein für den Schulanfang als angemessen betrachtet. Interessant ist insofern eine Befragung, die Roos (2000) in einem Modellversuch »Erweiterte Schülerinnen- und Schülerbeurteilung« an Schweizer Gymnasien durchgeführt hat. SchülerInnen, Eltern und LehrerInnen sollten die Bedeutung der verschiedenen Bezugsnormen für die Leistungsbeurteilung bewerten. In allen drei Gruppen wurde der individuellen Entwicklung und dem Grad, zu dem die Lernziele erreicht sind, der Vorrang vor der sozialen Bezugsnorm eingeräumt. Letztere spielt allerdings für SchülerInnen und Eltern - im Vergleich zu LehrerInnen - eine etwas gewichtigere Rolle, wie das folgende Schaubild zeigt.

---

173 So resümiert bei Oelkers (2001, o.S.).

174 Schwarzer u.a. (1982), ref. bei Oerter/Montada (1995, 997); Rheinberg/Peter (1982, 156), Schwarzer u.a. (1982, 171) und Trudewind/Kohne (1982, 182) - alle referiert bei Persy (1990,159-160).

175 Vgl. dazu im Einzelnen die Studie von Rheinberg (1980).

176 ... d.h.: (Miss-)Erfolge werden nicht auf äußere Umstände abgeschoben, sondern von den SchülerInnen sich selbst zugerechnet - und dabei eher der eigenen Anstrengung (die ja veränderbar ist) als einer (stabilen) Begabung (vgl. ebda.)

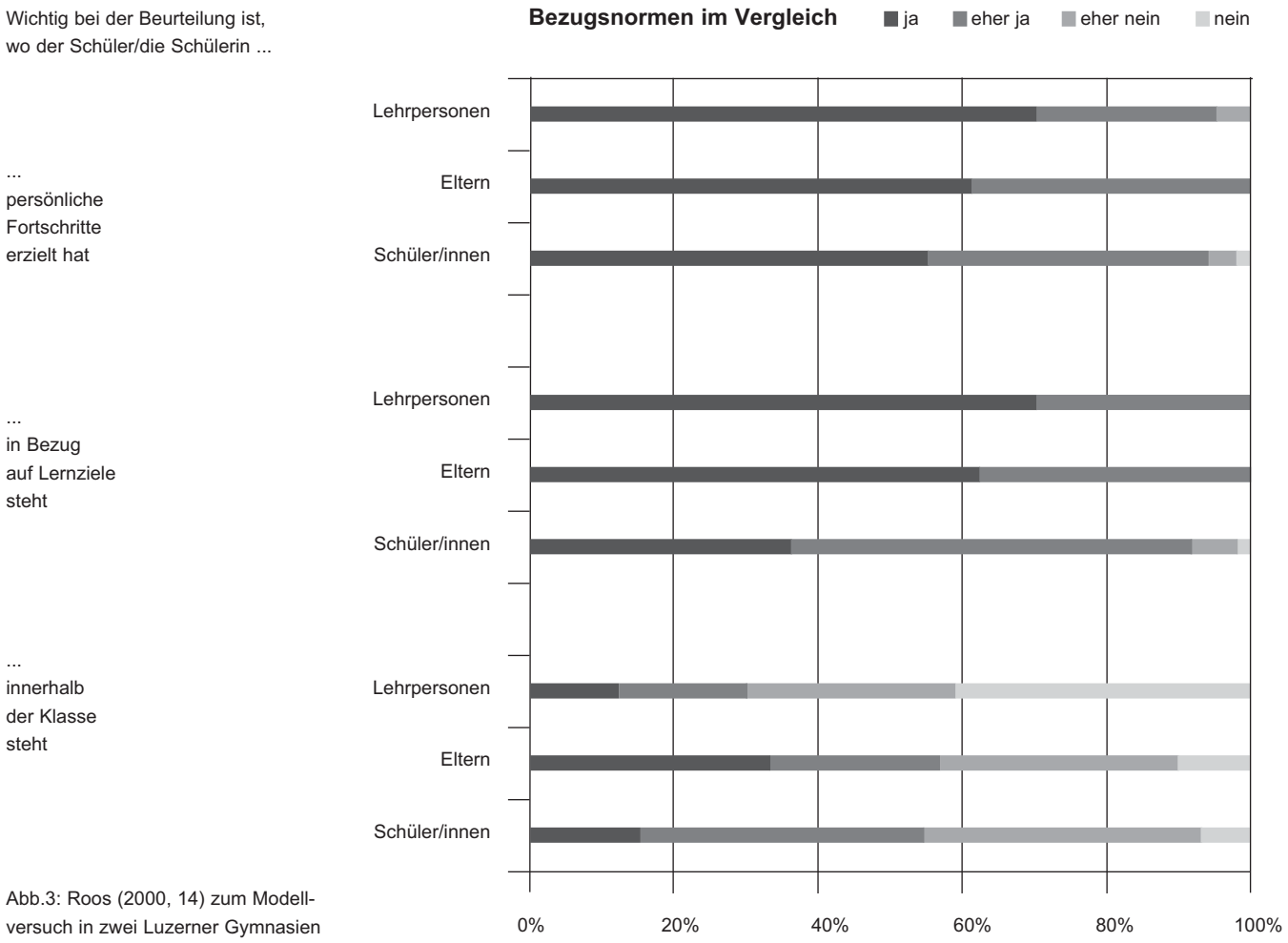



Abb.3: Roos (2000, 14) zum Modellversuch in zwei Luzerner Gymnasien

## 3

## Wie werden verschiedenen Formen der Leistungsbeurteilung umgesetzt, und welche Wirkungen haben sie?[177]

In > Kap. 2 wurden die Effekte unterschiedlicher Bezugsnormen grundsätzlich untersucht. Im Folgenden werden ihre Umsetzung in der Praxis und damit Unterschiede zwischen den im Schulalltag gebräuchlichen Formen der Darstellung analysiert. Wie schon angedeutet greift dabei die Gleichung »Ziffernnoten = soziale Bezugsnorm« bzw. »Verbalgutachten = individuelle Bezugsnorm« zu kurz.

### 3.1

### Wie weit werden Ziffernnoten und Verbalgutachten ihren eigenen Ansprüchen gerecht?[178]

Ziffernnoten werden vor allem zwei Vorteile unterstellt: Verständlichkeit und Vergleichbarkeit (vgl. > Kap. 4). Diese Erwartungen können bislang allerdings kaum erfüllt werden, wie die in Kap. 1 und 2 referierten Untersuchungen gezeigt haben. Die Vergleichbarkeit wird allenfalls im Klassenrahmen erreicht, und die Pauschalität der Ziffer wird den unterschiedlichen Teilleistungen nicht gerecht. Noten wird zu Recht vorgeworfen, dass dieselbe Ziffer sehr Unterschiedliches bedeuten kann. So kommentiert Bambach (1994, 212-213): »Für die Zensuren-Verfechter hat offenbar auch die Note ›befriedigend‹ einen höheren Informationswert als ein ausführlicher Entwicklungsbericht, obwohl diese Note - wie *Andreas Flitner* schon 1966 anmerkte - gleichermaßen ›einen hochbegabten Nichtstuer, einen fleißigen Durchschnittskopf, einen guten Denker, der aber flüchtig arbeitet, einen selbständigen Routinier und noch weiteres andere bedeuten kann.‹«

Eine Differenzierung nach Teilkriterien führt bei jeder Form der Beurteilung zu valideren Aussagen (> Kap. 2.1). Insofern ist bereits die traditionelle Auffächerung etwa der Aufsatznote in »Inhalt/Sprache/Form« oder der Gesamtnote für Deutsch in »Sprachgebrauch/Lesen/Rechtschreiben«

177 Vgl. allgemein vor allem: Haenisch (1996a+b) sowie die Beiträge zu Valtin (2002a) und zu Beutel u.a. (2000).

178 Vgl. vor allem die Zeugnisanalysen von Benner/Ramseger (1985); Elbing/Buschmann (1985); Scheerer u.a. (1985); Haußer (1991); Ulbricht (1993); Lübke (1996); Maier (2001, 137 ff.); Schmude (2001, 129 ff.) und die Zusammenfassung bei Götz (2005, 82-85).

(wie in NRW) aussagekräftiger. Im Vergleich zu den Möglichkeiten einer Verbalbeurteilung bleibt aber selbst diese Differenzierung noch zu grob, um die Komplexität der angestrebten Fähigkeiten angemessen zu erfassen.

Andererseits schöpfen Verbalgutachten diese Möglichkeiten bisher bei weitem nicht aus[179]: Sie werden eher (nur) produkt- statt (auch) prozess-orientiert formuliert, vermitteln nur selten Klarheit über Erfüllung der Lehrplananforderungen und geben zu wenig Förderhinweise[180].

Vor allem nehmen sie kaum Bezug auf die unterschiedlichen Voraussetzungen[181]. Inhaltsanalysen[182] zeigen, dass weniger als 10% der ausformulierten Berichte Aussagen zur Entwicklung der individuellen Leistung und sogar weniger als 5% konkrete Fördervorschläge machen. Auch die mangelnde Verständlichkeit wird moniert[183].

Insofern erfüllen auch Berichtszeugnisse die in sie gesetzten Erwartungen meist nicht. Die in der Berliner Studie festgestellten Schwächen werden durch verschiedene Analysen von Verbalgutachten bestätigt[184]:

- fehlender Bezug auf die individuelle Leistungsentwicklung[185];
- Ungleichgewicht der Fächer und Leistungsdimensionen, d.h. starke Dominanz der Lese-, Rechtschreib- und Rechenleistungen[186];
- fehlende Fördervorschläge[187];
- Beschönigung der Rückmeldungen[188];
- Standardisierung der Aussagen durch Nutzung von Textbausteinen[189].

Diese Kritik wurde bereits in den ersten Evaluationen Anfang und Mitte der 1980er Jahre geäußert, findet sich aber in nur wenig veränderter Form unveränderter Form bis heute. Etwas differenzierter kritisiert Ulbricht (1993, 212): »Die Diskrepanz zwischen der Intention der Zeugnisreformer und den Ergebnissen meiner Zeugnisanalyse macht deutlich, dass die Verbalbeurteilung per se keine Garantie für eine kindgemäße (Grund)-Schule bedeutet. Während die Leistungsstandsbeschreibung in Anlehnung an die Vorgaben des Curriculums zumindest in den Fächern Deutsch und Mathematik bereits differenziert und unter individueller Bezugsnorm […] erfolgt, erweisen sich die Angaben zum Sozialverhalten und zum Lern- und Arbeitsverhalten als eher unsystematisch und hauptsächlich von der Person des Lehrers abhängig.«[190]

Vor diesem Hintergrund ist zu vermuten, dass sich die grundsätzlich positiven Effekte einer entwicklungsorientierten Bewertung von Leistungen (> Kap. 2.3) im Schulalltag nur eingeschränkt wiederfinden werden (vgl. > Kap. 3.2).

Konstruktiv gewendet verweisen die Ergebnisse auf die Notwendigkeit, in der Ausbildung von LehrerInnen mehr Wert auf die Beobachtung von Lernprozessen[191] und auf Kriterien für die Darstellung von Beurteilungen zu legen[192]. Um die Qualität von Verbalbeurteilungen bzw. Lernberichten zu sichern, sollten beim Verfassen bestimmte Schreibstandards erfüllt werden, wie sie Beutel (2005)[193] in differenzierter Weise fordert und entwickelt hat. Das bedeutet auch, dass verschiedene Instrumente genutzt werden, mit deren Hilfe die beobachtende Lehrkraft die entsprechenden Daten und Informationen für den Lernbericht sammelt, z.B. Tests, Lern-Tagebücher, Beobachtungsbögen, Schülerbriefe usw.[194].

179 Vgl. u.a. Schmude (2002a, 78-81). Dies ist auch eine Frage der Erfahrung und Qualifikation. So fand Leffelsand (2003), dass berufserfahrene LehrerInnen Informationen differenzierter nutzen und dass sie widersprüchliche Daten eher aufnehmen als Lehramtsstudierende.

180 Vgl. für das Berliner NOVARA-Projekt: Schmude (2002b) und Valtin (2002c, 145).

181 Immerhin ein zentraler Grund für die Einführung von Berichtszeugnissen, wie Maier (2001, 157) mit Verweis auf Lübke (1996, 41) festhält: »[…] herrscht hinsichtlich der Frage der Bezugsnormorientierung bei reformorientierten Grundschulpädagogen weitgehend darüber Einigung, dass mit der verbalen Beurteilung im Kontext eines individualisierenden Unterrichts die kriteriale und vor allem die individuelle Bezugsnorm realisiert werden sollen, denn unter *›der individuellen Bezugsnorm können sowohl die Leistungsstarken als auch die Leistungsschwachen den Zusammenhang von Anstrengung und Leistung erfahren: Schülerinnen und Schüler aller Leistungsniveaus haben Aussicht auf Erfolg und können ihre Kompetenz vor dem Hintergrund ihres bisherigen Leistungsvermögens jederzeit steigern‹«*.

182 Vgl. Haußer (1991); Ulbricht (1993); Maier (2001); Schmude (2001): Die Zahlen hier sind entnommen aus Schmude (2002a, 79), die in NOVARA für Berlin immerhin 86% der Aussagen der sachlichen Bezugsnorm zuordnet.

183 Vgl. Schaub (1993).

184 Vgl. die übersichtliche Zusammenfassung der einschlägigen Studien bei Jachmann (2003, 64-65), an der sich die auch folgende Übersicht orientiert; s. auch die differenzierte Zusammenfassung der Forschungslage bei Beutel (2005, 62-110).

185 Vgl. zu diesem durchgängigen Befund: Schmidt (1980, 87); Scheerer u.a. (1985); Valtin u.a. (1996, 292); Schmude (2002a, 79); Beutel (2005, 26, 40ff). Lediglich Elbing/Buschmann (1985) fanden in rund 1/4 und Haußer (1991) sogar in rund 3/4 der untersuchten Zeugnisse entwicklungsbezogene Beurteilungen.

186 Vgl. Schmack (1978); Schmidt (1980, 107, 489); Benner/Ramseger (1985, 154); Elbing/Buschmann (1985, 15-16); Ulbricht (1985, 129 ff.); Thiel/Valtin (2002, 69-70); Valtin (2002c, 145).

187 Unter 10% nach Schmude (2002a, 79); ähnlich schon Ulbricht (1993, 203), die vor allem darauf hinweist, dass Gründe für Lernschwierigkeiten eher im Kind als in Unterrichtsbedingungen gesucht werden, und Schmidt (1980); Haußer (1991, 358); Valtin (1996).

188 Vgl. Benner/Ramseger (1985); dagegen stehen aber die Befunde von Jürgens (1998, 188-189) aus seiner späteren Untersuchung.

189 Vgl. zum mangelnden Bezug auf »die individuelle Besonderheit der Schüler« auch Lübke (1996, 66); Scheerer u.a. (1985, 228) kommen zu dem Schluss, dass Lehrerinnen und Lehrer dazu neigen, sich an die Formulierungshilfen der Kultusadministration zu halten.

190 Zitiert nach Beutel (2005, 73).

191 Vgl. die Anm. von Maier (2001, 208) zu seiner Auswertung von Verbalzeugnissen: »Vorab muss darauf hingewiesen werden, dass auch im Rahmen dieser Textanalyse aufgrund der Datenbasis leider unklar bleibt, aus welchen diagnostischen Prozessen die Zeugnistexte resultieren. Informationen darüber, ob die Texte zum Beispiel mit Hilfe subjektiver, unklarer fragmentarischer Erinnerungen entstanden sind oder ob sie auf fundierten systematischen Dokumentationen basieren, würden einen differenzierten Interpretationsspielraum zulassen. Hier ergibt sich ein Ansatz für die weitere Forschung.«

192 So auch Schmude (2002a, 87) und Valtin (2002c, 146), die aufgrund der analogen Befunde in der Berliner Studie ebenfalls eine Ausbildung fordern, die diagnostische und Förderkompetenzen stärkt.

193 Beutel (2005, 42, 110-115).

194 Beutel (2005, 113) und die Beispiele in Winter (2004) und Bartnitzky u.a. (2005).

## 3.2

## Welche (Neben-)Wirkungen haben verschiedene Beurteilungsformen?

Im Folgenden wird in mehreren Schritten untersucht, ob und ggf. wie veränderte Beurteilungsformen überhaupt den Unterricht sowie die Motivation, die Leistung und das Selbstkonzept von SchülerInnen verändern.

### 3.2.1

### Gibt es einen Zusammenhang zwischen Unterrichtskonzept und Beurteilungsform?

Es besteht eine wechselseitige Abhängigkeit zwischen Unterrichtsform und Art der Leistungsbeurteilung[195]. Für die Einführung neuer Formen der Leistungsbeurteilung stellt deshalb die weithin noch unveränderte Unterrichtskultur ein besonderes Problem dar. Individualisierter Unterricht, in dessen Rahmen individuelle Rückmeldungen eine tragende Funktion haben, ist immer noch nicht sehr verbreitet. Und sofern seine Prinzipien umgesetzt werden, geschieht dies meist nur in inhaltlich reduzierten Formen[196].

So stellt Wagener (2002) für die Berliner Studie[197] fest: »Die meisten Anhängerinnen der verbalen Beurteilung waren zwar offen gegenüber Reformen, hatten diese jedoch nur ansatzweise in die Praxis umgesetzt.«
Ergebnis: Der Unterricht bleibt lehrerzentriert, allerdings wurden bei den »verbalorientierten« Lehrkräften mehr »schülerorientierte Unterrichtsmerkmale« beobachtet.

Offener Unterricht, der über eine organisatorische Differenzierung »von oben« hinausgeht, ist also immer noch die Ausnahme. Insofern verwundert die Häufigkeit der in > Kap. 3.1 berichteten Fehlformen nicht. Unter diesem Vorbehalt sind auch die im Folgenden berichteten Ergebnisse zu Wirkungen unterschiedlicher Beurteilungsformen zu sehen: Anders als in > Kap. 2 beziehen sie sich auf die gegenwärtige Praxis im Schulalltag.

Das gut dokumentierte Beispiel der Laborschule Bielefeld[198] zeigt, wie die Veränderung der Leistungsbeurteilung in eine Reform des Unterrichts eingebettet werden kann - und muss. Offenkundig wird aber auch, dass diese ein Prozess mehrjähriger Schulentwicklung und nicht eine einmalige Entscheidung ist[199].

Die Einführung pädagogisch anspruchsvollerer Formen der Leistungsbeurteilung verlangt also eine umfassendere Reform des Unterrichts. Umgekehrt kann aber auch die Einführung oder zumindest das Zulassen differenzierterer Beurteilungsformen Räume öffnen für eben solche Initiativen. Im Blick auf die anspruchsvoll formulierten Ziele der eigenen Richtlinien haben Kultusministerien hier eine Verantwortung, die sie nicht an zufällige Eltern-/Lehrer-Mehrheiten in den Schulkonferenzen abtreten dürfen.

### 3.2.2

### Beeinflusst die gewählte Beurteilungsform das Unterrichtsklima?

Es ist ein allgemeiner Befund der Unterrichtsforschung, dass die Lernfreude vom Kindergarten zur Grundschule hin ansteigt. Schon über die vier Grundschuljahre fällt sie dann aber kontinuierlich ab, während gleichzeitig die Versagensängstlichkeit steigt[200]. Innerhalb dieses Rahmens stellen Olechowski/Rieder (1991) positive Wirkungen einer entwicklungsorientierten Bewertung auf Motivation und Schulfreude der SchülerInnen und Maier (2001, 117 ff.) auf das Sozialklima generell fest[201]. Severinski (1990, 222) beobachtet einen kompensatorischen Effekt einer entwicklungsorientierten Bewertung bei eher konservativ unterrichtenden LehrerInnen: Schulfreude und positives soziales Verhalten der SchülerInnen nehmen zu.

Sowohl in der qualitativ-interpretativen als auch in der standardisiert-quantitativen Evaluation des NRW-Schulversuchs »Zeugnisse ohne Noten in Klasse 3 und 4« berichten LehrerInnen aus verschiedenen Schulen übereinstimmend von einer positiven Veränderung des Klimas in den Klassen[202]: weniger Angst vor Leistungsproben, weniger Rivalität, differenziertere Selbst- und Fremdeinschätzung von Leistungen, mehr Arbeit aus Interesse an der Sache und ein größeres Selbstbewusstsein.

Gegenüber dem Einwand, dass sich ein gutes Sozialklima (fälschlich als »Kuschelecken-Pädagogik« apostrophiert) leistungsmindernd auswirke, ist der Befund aus der Hamburger LAU-Studie wichtig, dass zu Beginn des fünften Schuljahrs Kinder aus Klassen, in denen sie sich überdurchschnittlich wohl gefühlt haben, keine schlechteren Leistungen erbringen als Kinder, die sich in der Grundschule nicht so wohl gefühlt haben (Lehmann u.a. 1997, 46). Eine plausible Erklärung aus der SCHOLASTIK-Studie: Die Lernfreude steigt, wenn Anforderungen als angepasst erlebt werden[203]. LAU hat allerdings ebenso wenig Belege gefunden, dass um-

195 Vgl. zu den Erfahrungen von Reformschulen zusammenfassend: Fiegert/Solzbacher (2001, 289-312).
196 Vgl. die Befunde von Brügelmann (2000); Hanke (2002); Valtin (2002, 143, 146); Wagener (2002); Winter (2004).
197 Im Projekt NOVUS (»Noten- oder Verbalbeurteilung: Unterrichtsorganisation und Sanktionsverhalten«) wurden 138 Unterrichtstunden von 7 Lehrkräften in 3. Klassen Ost- und Westberlins protokolliert und vergleichen - darunter 71 verbalorieniert und 67 notenorientiert.
198 Vgl. Bambach (1994); Groeben/Lenzen (1996; 1997); Lübke (1996); Thurn (1997); Beutel (1998); Döpp u.a. (2002).
199 Vgl. Thurn (1998).
200 Vgl. zu den Befunden in der SCHOLASTIK-Studie: Weinert/Helmke (1997b, 463-464).
201 Schwächer ausgeprägt und weniger klar in der Berliner NOVARA-Studie, vgl. Schmude (2001, 245-246) und Valtin/Wagner (2002, 116-118).
202 Vgl. Haenisch (1996a, 14; 1996b, 23).
203 Vgl. Helmke (1997b, 75).

gekehrt ein besseres Sozialklima automatisch zu *besseren* Leistungen führe.

Zum letzten Punkt ist der Hinweis von Krampen (1985, 118) zu bedenken, dass Veränderungen der Leistungsbeurteilung im Kontext eines unveränderten Schulsystems nur bedingt Wirkung entfalten können. Deshalb verwundern gering ausgeprägte Unterschiede nicht. Immerhin hat die von ihm untersuchte individuumsbezogene Kommentierung von Noten über den Versuch hinaus positive Wirkungen auf die Schulfreude der SchülerInnen gehabt.

Dass umgekehrt eine Verstärkung des Leistungsdrucks negative Auswirkungen auf die Beziehungen zwischen LehrerInnen und (vor allem leistungsschwachen) SchülerInnen hat, lässt sich an den Folgen des *high stakes testing* in den USA beobachten[204]. Dort hängen nicht nur die Schulkarrieren der SchülerInnen, sondern auch die Gehälter der LehrerInnen und finanzielle Zuweisungen an Schulen von den Ergebnissen in Vergleichstests ab. Dies führt u.a. zu höheren *drop-out*-(treffender: *push-out*-)Quoten und generell zu schlechteren Ergebnissen im unteren Leistungsbereich und in den Gruppen der gesellschaftlichen Minderheiten.

Auf der Sekundarstufe variiert die Schulfreude zunächst einmal mit der Zugehörigkeit zu verschiedenen Schulformen. Innerhalb der Schulformen spielt aber die durch Noten definierte Leistungsposition in der Klasse eine wichtige Rolle für die Einstellung zur Schule[205]. Die Situation ist komplex und lässt sich nicht in einfache Ursache-Wirkung-Beziehungen auflösen, wie auch die AutorInnen der Hamburger »LeiHS«-Studie betonen, die darauf hinweisen[206]... »... dass die Schüler(innen) mit einem Berichtszeugnis in der 5. Klasse offenbar mehr Freude an und in der Schule angeben als diejenigen Schüler(innen) mit einem Notenzeugnis. Wenn zudem in die Sichtweise der Sekundarschülerinnen und -schüler Rechnung gestellt wird, dass die Lernkultur und das Unterrichtsklima in den Klassen mit Berichtszeugnissen (am Ende der 5. Klasse) signifikant besser eingeschätzt werden als in Klassen mit Notenzeugnissen, lassen sich durchaus Zusammenhänge zur Zeugnisform herstellen. Einfache Kausalbehauptungen sind dabei allerdings unzulässig. Dennoch lassen unsere Ergebnisse den Schluss zu, dass in den Schulen, in denen es intensive Bemühungen um die Verbesserung der Lernkultur und der Unterrichtsqualität gibt, in denen ein lernstimulierendes, freudvolles soziales Klima vorherrscht, zugleich günstige Bedingungen für eine Entscheidung für Berichtszeugnisse bestehen. Berichtszeugnisse mit ihrem reformpädagogischen Impetus benötigen offenbar ein innovationsfreundliches und um Veränderung bemühtes Kollegium.«

In dieselbe Richtung weisen die Ergebnisse einer internationalen Aufsatzstudie zum Thema »Schule«. Danach ist der Anteil von deutschen SchülerInnen, die sich in der Schule nicht so wohl fühlen, mit 38.4% drei- bis viermal so hoch wie in den USA mit 11.3%. Die AutorInnen konkretisieren die Richtung der Kritik[207]: »Diese [Schülerinnen in den USA; brü] leiden eindeutig seltener unter Notendruck, haben eine deutlich höhere Selbsteinschätzung ihrer Leistungsfähigkeit und sehen einen Zusammenhang zwischen Begabung und Schulerfolg kaum.« Damit sind wir bei den Folgen des Unterrichtsklimas:

### 3.2.3

### Beeinflusst die gewählte Beurteilungsform zentrale Merkmale der Persönlichkeitsentwicklung?

Ob Kinder Leistungen erbringen, hängt von ihren kognitiven Grundfähigkeiten, aber darüber hinaus auch von ihrer Motivation und ihren Emotionen ab. Thiel (2005, 47) verweist auf die Vielzahl nichtkognitiver Persönlichkeitsmerkmale, die in der Forschung als Bedingungen für Lernerfolg untersucht werden, sieht jedoch in der Lernmotivation und der Zurechnung von Erfolgen bzw. Schwierigkeiten (»Kausalattribuierung«) die zentralen Faktoren.

Aber nicht nur wegen dieser »instrumentellen« Bedeutung, sondern auch wegen ihres Eigenwerts sind Wirkungen der Leistungsbeurteilung auf die Persönlichkeit zu beachten. Denn die Grundschule hat nicht nur einen Unterrichts-, sondern auch einen Erziehungsauftrag.

#### 3.2.3.1

#### Beeinträchtigen oder stützen Ziffernnoten bzw. Verbalgutachten die Lernmotivation?

Als ein zentrales Argument für Noten wird immer wieder ihre motivierende Funktion genannt: SchülerInnen würden nur mit der Aussicht auf gute Noten oder aus Angst vor schlechten Noten lernen. Zumindest der zweite Grund erscheint zweifelhaft, wie eine Befragung[208] gezeigt hat, in der HauptschülerInnen angeben konnten, was sie täten, wenn sie in einer Mathematikarbeit eine »5« bekämen. Zwar gab die Mehrheit der SchülerInnen an, sie würden »gute Vorsätze für die Vorbereitung auf die nächste Arbeit fassen«[209], aber fast genauso häufig wurden Ausweichreaktionen genannt.

Eine Erklärung findet sich in dem Hinweis von Valtin (2002a, 16) auf eine Studie von Faust-Siehl/Schweitzer (1992), nach der Kinder der 2. und 4. Klasse Misserfolge, wie sie durch schlechte Noten signalisiert werden, nicht konstruktiv verarbeiten. Nach Ingmar Hosenfeld[210] wirken sich Kausalitätsüberzeugungen dann positiv auf Schulleistungen aus, wenn der Anstrengung im Gegensatz zu Fähigkeit oder externen Bedingungen eine hohe Bedeutung zugemes-

204 Vgl. u.a. Kohn (2000); Linn (2000); Amrein/Berliner (2003) und die Zusammenfassung bei Brügelmann (2005, Kap. 46-48).

205 Fend u.a. (1976, 454) und Czerwenka u.a. (1990, 107), zit. nach Jachmann (2003, 58).

206 Vollstädt/Jachmann (2000, 152-153).

207 Czerwenka u.a (1990, 428 und 422); s. zu den Befunden genauer am Ende von > Kap. 4.2 .

208 Vgl. Krampen/Mory (1982), referiert bei Mreschar (1985, 62-64).

209 Vgl. die Zusammenfassung bei Mreschar (1985, 63).

210 Vgl. Hosenfeld (2002, 165 ff.).

sen wird. Gleichzeitig wurde aber herausgefunden, dass diese Kausalitätsüberzeugungen relativ instabil sind - was damit erklärt wurde, dass Ziffernnoten zum einen eigentlich nur als Halbjahresnoten eine Rückmeldefunktion haben und als Schuljahresendnoten eher der Selektion dienen. Zum anderen böten sie durch die Beschränkung auf sechs Ausprägungen wenig Möglichkeiten, auch kleinere Lernfortschritte widerzuspiegeln und seien somit ungeeignet, die Entwicklung angemessener und tragfähiger Kausalitätsüberzeugungen zu stützen. Für lernförderliche Kausalitätsüberzeugungen seien insofern Noten, die sich an individuellen Bezugsnormen orientieren, förderlicher als solche, die eher aufgrund sozialer Bezugsnormen vergeben werden.

Für eine entwicklungsorientierte Beurteilung stellten Olechowski/Rieder (1991) folgerichtig positive Wirkungen nicht nur auf die Schulfreude der Kinder, sondern auch auf ihre Lernmotivation fest. In dieselbe Richtung weisen die Befunde aus dem Berliner NOVARA-Projekt[211]: »In Bezug auf die Lernmotivation konnten wir Anzeichen für das beobachten, was Sacher (1996, S. 74) als Notenangst und Notengeilheit bezeichnet: eine stärkere Misserfolgsorientierung in der Leistungsmotivation der Kinder mit schlechten Noten sowie eine stärkere externale Motivation bei Kindern mit guten Noten.« Allerdings fielen die Unterschiede nur gering aus. Stärker waren die positiven Effekte, die Persy (1990, 159-162) aus verschiedenen Studien zur Orientierung an der individuellen Bezugsnorm berichtet. Valtin/Wagner (2002) differenzieren sie nach Teilgruppen: »Allgemein kann man sagen, dass die schwächeren und ängstlichen Schüler und Schülerinnen mehr von der verbalen Bewertung profitieren als die leistungsstarken, weniger ängstlichen. [...] ... bei schwachen Kindern, also Kindern mit schlechten Noten, ist eine stärkere Misserfolgsorientierung zu beobachten sowie eine größere Leistungsangst. Bei Schülern mit guten Noten ergab sich eine höhere externale Motivation. Mit Noten beurteilte Kinder erleben die schulischen Anforderungen als schwieriger.«[212] Nach ihrer Metaanalyse von Studien zur Auswirkung von externer Verstärkung auf intrinsische Motivation kommen Deci u.a. (1999, 652, 656-657) zu einem ähnlichen Ergebnis[213]:

▲ Materielle Belohnungen haben generell eine leicht negative Wirkung auf die intrinsische Motivation (Effektstärken d = .30 bis .40).

▲ Eine verbale Belohnung hat zwar generell eine leicht positive Wirkung (Effektstärke d = .30); das gilt aber *nicht* für eine erwartete positive Verstärkung (d = -.40) und nicht in einer Kontrollbeziehung. Im Vergleich zu einem rein informierenden Feedback wirkt sich eine kontrollierende Rückmeldung deutlich negativer aus (d = -.44).

In Schulversuchen, die nach einem veränderten pädagogischen Konzept arbeiten, fallen die Ergebnisse meist positiver aus als bei breiten Erhebungen in Schulen, die noch in traditionellen Formen oder unter herkömmlichen Rahmenbedingungen arbeiten.

Hartinger/Fölling-Albers (2002, 113) fassen die Ergebnisse verschiedener Studien in fünf Punkten zusammen, von denen die ersten drei[214] für unser Thema zentral sind:

»a) Vorhandene intrinsische Motivation kann durch zusätzliche extrinsische Motivation verringert, wenn nicht gar zum Erliegen gebracht werden.

b) Die Gefahr dieses kontraproduktiven Effektes extrinsischer Motivierung ist dann besonders groß, wenn sie den Schüler/innen Kontrolle signalisiert.

c) Weniger Probleme durch extrinsische Motivation gibt es dann, wenn Rückmeldungen vor allem informativ gehalten sind.«

Damit haben selbst *gute* Noten *negative* Nebenwirkungen auf die Entwicklung einer sachorientierten Motivation. Lempp (1971, 65) hat deshalb schon vor über 30 Jahren darauf hingewiesen: »Damit [daß die Leistung eines Kindes immer in Relation zu den Leistungen der Mitschüler gestellt wird] wird beständig eine Leistungshierarchie in der Klassengruppe hergestellt, die für viele Kinder definitiv und aussichtslos erscheinen muß. Das Erlebnis, stets zu den Kindern zu gehören, die eine geforderte Leistung bewältigen können, oder aber zu denen, die dies in der Regel nie können, muß prägend sein für die Einstellung zum Leben, zum Beruf, zur Umwelt überhaupt. Zu fragen wäre, ob nicht unter voller Korrektur der geleisteten Arbeit eine Abstufung nach dem Notensystem besser unterbleiben sollte und könnte.«

Noch frühere Stellungnahmen gegen Prüfungen und Noten finden sich schon Ende des 19. Jahrhunderts, seinerzeit vorgebracht von Ärzten, die vor den schädlichen Folgen von Prüfungsstress warnten[215]. Deshalb die Frage:

---

211 Valtin (2002c, 145); vgl. ausführlicher Valtin/Wagner (2002).

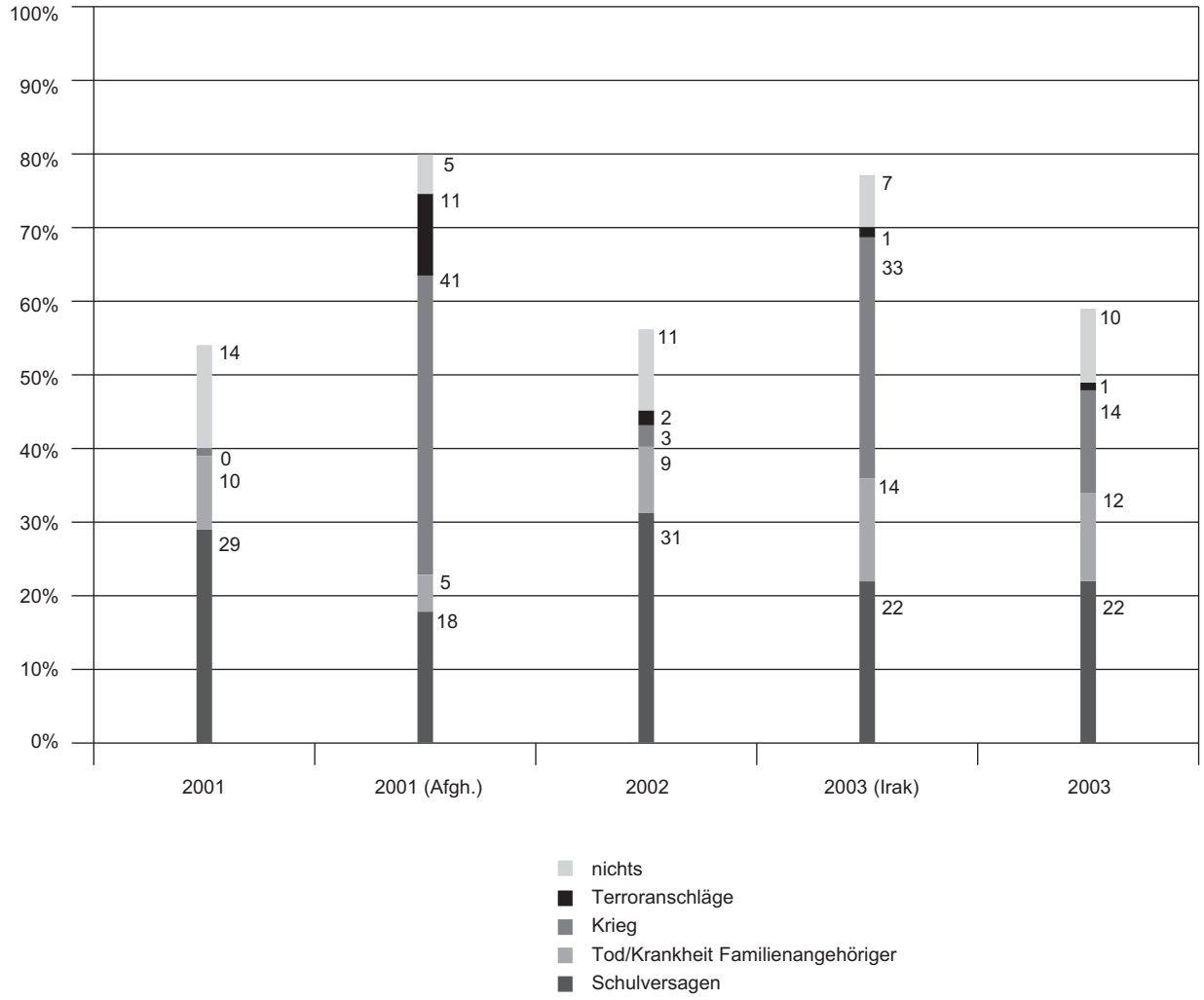
212 Valtin/Wagner (2002, 128, 137); s.a. Valtin (2002b, 15).

213 Ebenso in ihrem Forschungsüberblick: Harlen/Deakin Crick (2002).

214 Die beiden letzten Punkte (a.a.O., 111-114):
»d) Lob und Tadel können gegenteilig interpretiert werden, wenn z.B. Schüler/innen Lob auf große Anstrengung bei mangelnder Begabung oder auf die geringe Erwartungshaltung der Lehrer/innen zurückführen.
e) Es ist aber immer dann wichtig, auch extrinsische Motivationsformen - vor allem Lob und Anerkennung - einzusetzen, wenn bei gehäuften Misserfolgen Selbstwertprobleme entstehen können.«
Ähnlich kritisch fasst Kohn (1999, 202-203) den Stand der US-amerikanischen Forschung zu Noten und Motivation zusammen:
» - Noten unterminieren die intrinsische Motivation zum Lernen; sie sind starke Demotivierer unabhängig davon, wofür sie gebraucht werden.
- Das Ziel, Schüler zu sortieren, passt nicht zu dem Ziel, sie mit Noten zum Lernen zu motivieren.
- Andere Rückmeldungen über den Leistungsstand als Noten und Tests sind weniger bestrafend und mehr informativ.«

215 Vgl. die Hinweise bei Oelkers (2001, o.S.).

**Die größten Ängste der Kinder** (Angaben in Prozent) Abb. 4: pro Kids (2004, 25)




### 3.2.3.2

### Verringern oder vergrößern Ziffernoten bzw. Verbalgutachten die Schul- und Prüfungsangst?

Laut einer Statistik aller Kinder- und Jugendtelefone von 2003 stehen im Bereich Schule Noten als Stressfaktor an erster Stelle. Vor allem wenn es am Ende eines Schuljahres das Abschlusszeugnis gibt, sind sie häufig Anlass zum Streit zwischen Eltern und Kindern[216].

Auch Befragungen zeigen: Die Angst vor Schulversagen, insbesondere vor schlechten Noten, ist für SchülerInnen zwischen neun und zwölf Jahren die beherrschende Sorge, deutlich vor anderen Ängsten: »Fast ein Drittel der Kinder äußert als größte Angst, in der Schule zu versagen (29%). In diese Kategorie fallen beispielsweise Ängste vor dem Ergebnis der Klassenarbeit oder vor schlechten Zeugnisnoten.«[217]

Wenn auch die Werte über die Jahre hinweg zwischen 20% und 30% schwanken, zeigt das Schaubild, dass nur 2001 und 2003 die akuten Kriegsgefahren die Angst vor dem Versagen in der Schule etwas in den Hintergrund gedrängt haben. Diese Versagensängste erhöhen sich generell von der 4. bis zur 7. Klasse auf mindestens das Doppelte. In der NOVARA-Studie waren es vor allem die »Kinder mit schlechten Noten«, die eine stärkere Leistungsangst entwickelten[218]. Unter den Eltern sind es sogar 38%, die bei ihren Kindern »Angst vor schlechten Noten« beobachten, aber nur 11% bzw. 28% nennen unter ihren Reformwünschen explizit eine »Ersetzung von Zeugnisnoten durch verbale Beurteilung«[219].

216 Vgl. die Auswertungen der Kinder- und Jugendärzte v. 23.7.2004 im Netz unter > www.kinderaerzteimnetz.de/bvkj/aktuelles1/show.php3?id=1203&nodeid=26&nodeid=26& [Abruf: 20.1.06].

217 Pro Kids (2002, 21); vgl. auch Huber (2003).

218 Vgl. Valtin/Wagner (2002, 121-125) und Valtin (2002c, 145).

219 Rosenfeld/Valtin (2002, 33, 35).

Auch Selbstaussagen von SchülerInnen auf der Sekundarstufe lassen erkennen, dass Noten Angst auslösen, allerdings nicht durchgängig[220]: Zwar gab ein Drittel an, vor Klassenarbeiten Versagensangst, Herzklopfen und Nervosität zu spüren; und es besteht im Mittel mehr Angst vor Klassenarbeiten mit Noten, aber konkret sind es nur rund 25%, die angaben, weniger Angst vor Klassenarbeiten *ohne* Zensuren zu haben, und sogar nur 10% fanden »Noten in der Schule schlecht, weil sie mir Angst machen«.

Andererseits scheint für die Wirkung der Beurteilungen die Art und Weise, wie LehrerInnen sie präsentieren, eine wichtige Rolle zu spielen - und zwar mehr als der institutionelle Kontext[221]. Dies haben Hartinger u.a. (2003) bei einem Vergleich von Leistungsangst und -motivation festgestellt, als sie ViertklässlerInnen aus Bayern und Niedersachsen befragten. In Bayern bestimmt der Notendurchschnitt den Übergang zur weiterführenden Schule direkt, in Niedersachsen konnten die Eltern nach Beratung durch die LehrerInnen selbst entscheiden, in welche Schulform ihr Kind auf der Sekundarstufe I wechselte: »Zusammenfassend kann man festhalten, dass die niedersächsischen Schüler/innen zwar etwas günstigere Einschätzungen abgaben, dass die Unterschiede zwischen den beiden Bundesländern jedoch recht gering sind. Erklären lässt sich dieser Befund zum einen dadurch, dass in unserem Schulsystem die Zensuren als Indikator von Schulleistungen immer als bedeutsam angesehen werden, unabhängig davon, ob sie die weitere Schullaufbahn direkt bestimmen oder nicht. [...] Einen stärkeren - da direkteren Einfluss haben dann wieder die einzelnen Lehrer/innen, die die Prüfungsangst oder Motivation der Schüler positiv oder negativ beeinflussen können. So zeigte sich in unseren Untersuchungen, dass die Unterschiede zwischen den einzelnen Klassen innerhalb der beiden Bundesländer deutlich größer sind als die Unterschiede zwischen den Bundesländern.« (Hartinger u.a. 2003, 118).

Dazu passt ein Befund aus der KILIA-Studie von Martschinke u.a. (2004): Rund 20% der Kinder hatten Angst vor dem Wechsel zum Notenzeugnis. Wie stark diese Angst ausgeprägt war, hing aber vom »Notenklima« in der Klasse ab, also von der Art und Weise, wie die Lehrperson mit der Bewertung im Unterrichtsalltag umgeht.

Wie bedeutsam moderierende Variablen sind, zeigt auch ein zweiter Befund. Es sind nicht immer die leistungsschwachen SchülerInnen, die besondere Angst vor den Noten haben: »Sechstklässler mit schlechteren Noten zeigten in einer Studie von Jachmann (2003) höhere Schulangst. Aber es gibt auch Hinweise, dass Kinder mit guten Noten (in Mathematik) sich mehr Sorgen um zukünftige Noten machen als Kinder mit schlechteren Noten (Sirsch 2000). Hartinger, Graumann und Grittner (2004) finden ihre Vermutung bestätigt, dass in Bayern besonders Kinder, deren Übertritt auf das Gymnasium aufgrund der Notensituation noch unsicher ist, mit höherer Leistungsangst reagieren.«[222]

Auch wenn es eher Minderheiten sind, die unter Noten und unter Schulangst leiden, sollte nicht übersehen werden, dass es vor allem Unterschichtkinder sind, die von angstbesetzten Situationen aus der Schule berichten[223]. Ob dies direkt durch eine größere Schulfremdheit oder indirekt durch häufigere schlechte Leistungen bedingt ist, kann dahin gestellt bleiben. Prüfungsangst wirkt sich jedenfalls ungünstig auf Leistungsfähigkeit und Lernbereitschaft aus und sollte deshalb gerade für Kindergruppen, die sowieso häufiger Schwierigkeiten in der Schule haben, möglichst gering gehalten werden.

#### 3.2.3.3

#### Schädigen oder stärken Ziffernnoten bzw. Verbalgutachten das Selbstkonzept?

Das Verhältnis von Selbstwertgefühl und Leistung ist wechselseitig: Wer sich etwas zutraut, erbringt bessere Leistungen[224], und positive Rückmeldungen zur eigenen Leistung steigern das Selbstwertgefühl von SchülerInnen.

Wie verschiedene Studien zeigen, ist dabei die unmittelbare Bezugsgruppe entscheidend[225]. Hier steckt ein zentrales Problem von Ziffernnoten. Die unterstellte Normalverteilung der Leistungen verurteilt die Hälfte der SchülerInnen einer Klasse von Vornherein zum Versagen. Die Negativeffekte werden allerdings in den Vergleichsstudien nicht sichtbar: In der Hamburger LAU-Studie war das Selbstkonzept von ViertklässlerInnen mit Berichtzeugnissen zwar etwas günstiger, aber sie kamen tendenziell auch aus sozial besser gestellten Elternhäusern[226]. Im Berliner NOVARA-Projekt zeigten verbal beurteilte Kinder ebenfalls kein besseres Selbstkonzept als diejenigen, die mit Ziffern benotet worden waren[227].

Ein Grund für dieses »Patt« könnte auch in diesem Fall darin liegen, dass Verbalzeugnisse in der Praxis die Individualnorm weitgehend vernachlässigen (vgl. > Kap. 2.3 und 3.1), so dass gar kein Kontrasteffekt erwartet werden

220 Vgl. Vollstädt/Jachmann (2000, 145, 151).

221 Vgl. zum Einfluss von Bedingungen, unter denen die zu beurteilende *Leistung* zu erbringen ist, auf die Prüfungsangst: Olechowski/Sretenovic (1983).

222 Martschinke u.a. (2005, 90).

223 Vgl. Büchner/Koch (20021, 240-241).

224 Allerdings ist der viel zitierte »Pygmalion-Effekt« nicht pauschal zu halten, sondern von einer Reihe spezifischer Bedingungen in der Person und in der Situation abhängig, vgl. Heckhausen (1974). Kritisch zu vereinfachten Deutungen des Pygmalion-Effekts auch Baumeister u.a. (2004.)

225 Vgl. Jachmann (2003, 57), Auf den ersten Blick wirkt das - einschließlich der Sonderschulen - viergliedrige Schulsystem der Sekundarstufe hier entlastend. In der Tat lässt sich nach dem Wechsel in eine niedrigere Schulform oft ein höheres Selbstkonzept feststellen. Dieser Anstieg ist aber auf die leistungsstärkeren SchülerInnen beschränkt und schwindet selbst bei ihnen mit Annäherung an den Schulabschluss, der dann zunehmend im schulartübergreifenden Vergleich wahrgenommen wird, vgl. Zielinski (1980, 104f.).

226 Vgl. Lehmann u.a. (1997, 81f.).

227 Vgl. Valtin u.a. (2000, 12); Valtin/Wagner (2002, 118-221) und zur Entwicklung des Fähigkeitsselbstbilds Schmude (2001, 245 ff.).

kann. Denn grundsätzlich lassen sich positive Wirkungen der individuellen Norm auf das Selbstwertgefühl feststellen[228].

Andererseits ist während der Grundschulzeit generell zu beobachten, dass das fachbezogene Selbstkonzept der SchülerInnen im Durchschnitt schlechter wird, wie Helmke (1998; 1998) unter dem Stichwort »vom Optimisten zum Realisten« anhand von Daten aus der Münchener LOGIK-Studie, aber auch aus anderen Untersuchungen resümiert (1999, 207, 218). Zur Erklärung verweist er auf die parallel zunehmende Bedeutung von Leistungsbeurteilungen, die sich an der sozialen Bezugsnorm orientieren (a.a.O., 206, 218) - in deutschen Schulen verschärft durch die wachsende Bedeutung von Noten nach Klasse 1/2. Diese trägt auch zur Entwicklung fachspezifisch differenzierter Selbstkonzepte bei[229].

### 3.2.4

### Belasten oder fördern Ziffernnoten bzw. Verbalgutachten die Leistungsentwicklung?

Wie bereits in > Kap. 3.3.2 kurz angesprochen, führt ein Verzicht auf Noten nicht zu einem Leistungsabfall. Lehmann u.a. (1997, 81-82) fanden in der LAU-Studie in Hamburg, dass Kinder aus vierten Klassen mit Berichtszeugnissen keine schlechteren Leistungen erbringen als Klassen mit Notenzeugnissen. Dieser Befund spricht gegen die verbreitete Annahme, dass Notendruck leistungssteigernd wirke. Auch im Projekt NOVARA führte der Verzicht auf Noten nicht zu Negativeffekten: »Insgesamt traten [in fachlicher Hinsicht, brü] nur zwei bedeutsame Unterscheide auf: im 2. Schuljahr waren die Notenkinder etwas besser in der Rechtschreibung, im 4. Schuljahr erzielten die verbal beurteilten Kinder im Rechentest einen höheren Wert.«[230]

Ein Schulversuch im Kanton Luzern in der Schweiz bestätigte schon vor Jahren, dass der Verzicht auf Noten nicht zu einem Leistungsabfall führte[231]. In den USA fanden Fraser u.a. (1987) in ihrer Metaanalyse sogar eine negative Korrelation zwischen einer starken Betonung von Noten im Unterricht und den Leistungen der SchülerInnen. Diese fiel mit -.07 allerdings so niedrig aus, dass eher - wie in Deutschland - von einem fehlenden Zusammenhang zu sprechen ist[232].

National[233] wie international[234] gibt es eine Vielfalt von Bewertungssystemen, in denen Ziffernnoten, Testwerte und Verbalbeurteilungen unterschiedlich kombiniert werden, ohne dass sich systematische Zusammenhänge zu den Fachleistungen der SchülerInnen herstellen lassen[235]. Das gilt auch für den Leistungsvergleich von Systemen mit Noten vs. solchen, die nur eine verbale Rückmeldung vorsehen[236].

In einem kontrollierten Vergleich stellten Grolnick/Ryan (1987) allerdings positive Effekte eines Verzichts auf Noten fest[237]. Sie verglichen drei Gruppen, die sich eine Textpassage unter unterschiedlichen Bedingungen erarbeiten sollten[238]:

N = nicht-direktiv: SchülerInnen lesen und berichten anschließend der Lehrerin, was sie an dem Text interessant fanden;
A = autonomieunterstützend: die Lehrerin zeigt ein persönliches Interesse am Leistungsfortschritt der SchülerInnen;
K = kontrollierend: den SchülerInnen wird vorweg eine Überprüfung und Benotung ihrer Leseleistung durch die Lehrperson angekündigt.

Die Effekte wurden an drei Kriterien gemessen:

a) Welche Gruppe erreicht die beste Leistung im *konzeptuellen* Textverständnis?
Ergebnis: A besser als N und deutlich besser als K
b) Welche Gruppe schneidet *kurzfristig* im *auswendig* gelernten Wissen am besten ab?
Ergebnis: A und K besser als N
c) Welche Gruppe schneidet *langfristig* im *auswendig* gelernten Wissen am besten ab?
Ergebnis: A besser als K und N

228 Vgl. Oerter/Montada (1995, 997).
229 A.a.O., 216-217, wobei Helmke darauf aufmerksam macht, dass neben institutionellen Faktoren auch die kognitive Entwicklung der Kinder zu einer differenzierteren Sicht auf die eigene Leistung beitragen dürfte (a.a.O, 219).
230 Valtin/Wagner (2002, 135); s.a. Valtin u.a. (1999, 12).
231 Vgl. Theiler u.a. (1992, 13-14).
232 Speziell für den Sekundarbereich verweist Lind (2003) auf die groß angelegte Studie von Chamberlin u.a. (1942) in den USA, die schon in den 1930er Jahren nachgewiesen hat, dass Absolventen von so genannten »Progressive Schools« (John Dewey), die keine Benotung kannten, im College gleich gut oder sogar besser abschnitten als Absolventen von traditionellen High Schools.
233 Vgl. Reimers (1991) und oben > Kap. 0.4 .
234 Vgl. Schmitt (1992).
235 Nimmt man etwa das Zentralabitur als Beispiel für externe vs. interne Prüfungen, so sind die Effekte einer zentralen Prüfungsorganisation im Vergleich zu einer dezentralen Beurteilung durch die LehrerInnen heterogen - bezogen sowohl auf positive als auch auf negative Erwartungen:
- Zentralabitur sichert nur begrenzt, d.h. nur über wenige Fächer/Kursstufen hinweg ein höheres Leistungsniveau und eine geringere Streuung der Leistungen.
- Bezogen auf die Noten ist die normierende Kraft der zentralen Prüfung ebenfalls begrenzt - auch die Streuung zwischen Schulen wird im Vergleich zu dezentralem Abitur nicht kleiner.
- Zentrale Prüfungen beeinträchtigen andererseits nicht die Fähigkeit zur Lösung anspruchsvoller Probleme.
- Sie lösen ebenfalls nicht mehr Angst bei SchülerInnen aus (Baumert/Watermann 2000, 345-350; vgl. zu TIMSS auch Bos/Baumert 1999; zu PISA: Baumert u.a. 2000, 341-351; zum innerdeutschen Vergleich: Bellenberg u.a. 2004, 140).
236 Vgl. Fadisch/Steinert (2005, 178-180) zur Seltenheit schriftlicher Rückmeldungen über den Leistungsstand der Kinder an ihre Eltern in den IGLU-Spitzenreitern England und Schweden sowie > Kap. 0.5.
237 Dieser Befund wurde in Japan repliziert durch Kage/Namiki (1990) und Kage (1991).
238 Zusammengefasst nach Deci/Ryan (1993, 234).

Damit ist dies ein sehr robuster Befund: Die vielerorts vorgetragene Sorge, SchülerInnen würden nichts mehr lernen, wenn die Schule auf Noten als Lock- und Drohmittel verzichtet, lässt sich empirisch nicht halten. Unter kontrollierten Bedingungen lässt sich durch eine entwicklungsorientierte Beurteilung sogar ein höherer Lernerfolg erreichen.

### 3.2.5

### Zwischenbilanz zu »Wirkungen«

Über alle Untersuchungen hinweg finden sich nur wenige und zudem in der Regel nur schwach ausgeprägte Unterschiede in den Effekten. Dieses Fazit deckt sich im Wesentlichen mit dem Resümee des Berliner NOVARA-Projekts: »Insgesamt ist die Ausbeute an statistisch bedeutsamen Unterschieden bei den zahlreichen Vergleichen von Kindern mit Notengebung und verbaler Beurteilung recht bescheiden. [...] Ein Grund für die geringe Wirksamkeit der verbalen Beurteilung ist sicherlich darin zu sehen, dass die mit dieser Zeugnisform verbundenen Intentionen in der Praxis kaum umgesetzt worden sind, wie die in diesem Buch geschilderten Ergebnisse der Unterrichtsbeobachtungen und der Zeugnisanalysen zeigen.«[239]

Allerdings gibt es aus sehr unterschiedlichen Studien ernst zu nehmende Anhaltspunkte für negative Auswirkungen von Noten auf beachtliche Teilgruppen von SchülerInnen. Umgekehrt lassen sich die Befunde am ehesten auf den folgenden Nenner bringen: Grundsätzlich *kann* eine Veränderung der Bewertungsformen positive Wirkungen haben. Diese werden aber im Schulalltag nur selten beobachtet - sei es, dass Mischformen (z.B. Ziffernnoten kombiniert mit erläuterndem Bericht) verwendet werden, sei es, dass die Intentionen der Verbalgutachten - wie oben festgestellt - nicht (zureichend) umgesetzt werden. Über alle Untersuchungen hinweg finden sich deshalb nur wenige und zudem in der Regel nur schwach ausgeprägte Unterschiede in den Effekten. Diese Befunde aus dem Schulalltag stehen im Widerspruch zu den in > Kap. 2 referierten Studien, die unter kontrollierten Bedingungen positive Ergebnisse aus dem Verzicht auf eine vergleichsorientierte Bewertung berichten. Bei einer Einführung von Verbalzeugnissen sind deshalb bestimmte Voraussetzungen (fachliche Qualifikation der LehrerInnen) bzw. Rahmenbedingungen (Verringerung des Selektionsdrucks) zu sichern.

## 4

## Wie gut erfüllen Ziffernnoten und Verbalgutachtenwichtige Funktionen aus der Sicht der Betroffenen?[240]

Mit der Beurteilung von Leistungen werden verschiedene Erwartungen verbunden (> Kap. 0.3). Im Folgenden soll geprüft werden, bis zu welchem Grad die unterschiedlichen Beurteilungsformen die genannten Erwartungen einlösen - und zwar aus der Sicht der verschiedenen Beteiligten. Wie in den > Kap. 1-3 gezeigt, können Noten diese Ansprüche zwar faktisch nicht erfüllen. Aber auch Verbalgutachten werden den gesetzten Anforderungen im Schulalltag meist nicht gerecht. Für die Entscheidung, ob Noten abgeschafft und durch Verbalbeurteilungen ersetzt werden sollen (und können...), ist deshalb wichtig zu wissen, wie die Betroffenen Vor- und Nachteile der verschiedenen Beurteilungsformen wahrnehmen. In den 1970er und 1980er Jahren hielten rund drei Viertel der SchülerInnen, der Eltern und der LehrerInnen Zensuren für notwendig[241]. Dieses Bild hat sich verändert, wenn man die Ergebnisse neuerer Erhebungen betrachtet.

### 4.1

### Einschätzungen von LehrerInnen

Zu Beginn der 1980er Jahr wurde die (damals neue) Berichtspraxis in niedersächsischen Grundschulen untersucht. In der Lehrerschaft zeigte sich eine eher unsichere Haltung im Umgang mit verbalen Beurteilungen: »Mehr als 2/3 sähen die Reform gern - wenigstens teilweise - rückgängig gemacht«[242].

Nach einer Befragung von 157 GrundschullehrerInnen in Baden-Württemberg wurden Berichte damals einerseits als Ausdruck einer besseren Beurteilungspraxis gesehen; zum anderen wurde die hohe Arbeitsbelastung beklagt, ihre Abfassung wurde als problematisch eingeschätzt und die Kommunikation mit den Eltern als schwierig[243].

Deutlich positiver fällt die Auswertung der 20 Erfahrungsberichte aus dem Schulversuch »Zeugnisse ohne Noten in den Klassen 3 und 4« in NRW aus, in denen die rund 100 Eltern durchweg positive Einschätzungen äußerten: »Die Ergebnisse der Studie sprechen insgesamt dafür, dass zensurenfreie Beurteilungen in den Klassen 3 und 4 erfolgreich eingesetzt werden können. Eltern und Lehrkräfte, die eigene

239 Valtin/Wagner (2002, 136, 137).

240 Vgl. zur Akzeptanz vor allem: Haenisch (1996a+b), Maier (2001, 111 ff.), Jachmann (2003, 103 ff.); Beutel (2004/2005); Pohl/Beekmann (2005a+b) sowie die Beiträge zu Beutel u.a. (2000), zu Döpp u.a.(2002), zu Valtin (2002a) und die Zusammenfassung bei Götz (2005, 86-88).

241 Vgl. u.a. Valtin/Schmude (2002) mit Verweis auf Weiß (1986) u.a.

242 Schmidt (1981, 488).

243 Weiss (1986).

Erfahrungen damit gemacht haben, berichten mehrheitlich von positiven Erfahrungen und sind auch mehrheitlich von der Überlegenheit gegenüber Noten überzeugt«.[244] Allerdings hatten sie - ebenso wie die LehrerInnen - dem Versuch vorher zugestimmt, so dass eine eher positive Voreinstellung angenommen werden muss.

Für Klasse 1 und 2 stellte Jürgens (1998b, 188) inzwischen generell eine positive Resonanz bei den LehrerInnen fest: 98% waren mit der 1979 in NRW beschlossenen Einführung der Verbalgutachten zufrieden, immerhin 43% befürworteten eine Ausweitung auf die dritte Klasse und 10% auf die Sekundarstufe. Die hier erkennbare Tendenz, verbale Beurteilungen auf die unteren Klassenstufen zu beschränken, zeigt sich auch bei Eltern (> Kap. 4.3) und dürfte mit dem von Jahrgangsstufe zu Jahrgangsstufe zunehmenden Selektionsdruck zusammenhängen[245]. Immerhin empfanden Sekundarstufen-LehrerInnen die Berichtszeugnisse aus dem NRW-Modellversuch in Klasse 3 und 4 im Vergleich zu Ziffernnoten als aussagekräftiger; sie haben auch mit den Kindern eher positive Erfahrungen gemacht, zumindest aber keine Probleme beim Übergang zu Noten beobachtet[246].

Nach einer aktuellen Befragung von FORSA[247] stimmen zwar weiterhin 75% der LehrerInnen der Aussage zu »Noten gehören zur Schule dazu«. Aber dieses Urteil bezieht sich vor allem auf die unterstellten Erwartungen von SchülerInnen und Eltern. Denn fast die Hälfte der LehrerInnen hält ausformulierte Beurteilungen für aussagekräftiger.

Dabei fällt das Ergebnis in den verschiedenen Schulformen ganz unterschiedlich aus[248]: »Etwa 60% der Grund- und Sonderschullehrer halten Noten für überflüssig und bevorzugen ausformulierte Bewertungen. Die befragten Realschullehrer stimmen dieser Aussage nur zu 21% zu, Gymnasiallehrer zu 28%.«[249]

Diese auffällige Differenz könnte an der unterschiedlichen Nähe zum Schulabschluss liegen[250], aber auch an den fehlenden Erfahrungen mit Verbalbeurteilungen in den

244 Haenisch (1996b, 51).

245 S. zu dieser zentralen Bedingung für die Möglichkeiten einer veränderten Leistungsbeurteilung ausführlicher > Kap. 7.

246 Vgl. Haenich (1996a, 15).

247 Vgl. Pohl/Beekmann (2005a, 85).

248 Dieselbe Tendenz, wenn auch nicht ganz so deutlich ergab eine Befragung von Kanders u.a. (1998, 170): 37% der GrundschullehrerInnen gegenüber 32% der LehrerInnen weiterführender Schulen hielten Noten in den ersten drei Schuljahren für überflüssig.

249 Pohl/Beekmann (2005a, 167): Interessant auch: 48% der Frauen halten Noten für überflüssig, aber nur 30% der Männer (a.a.O., 90). LehrerInnen aus West (48%) und Ost (29%) unterscheiden sich ebenfalls deutlich in der Bevorzugung von Verbalbeurteilungen (a.a.O., 92).

250 Bei Eltern jedenfalls nimmt die Bedeutung der Ausweisfunktion von Noten zu; zur Abnahme der Zustimmung zu Verbalgutachten von der Hälfte (2. Klasse) auf ein Viertel (4. Klasse) in Hamburg vgl. Wallrabenstein (1992, 120-121).

**Einstellung zu Noten**

Basis: Gesamt; Angaben in Prozent; stimme voll und ganz/überwiegend zu.
»Ich lese Ihnen nun einige Aussagen vor, die Noten betreffen. Bitte geben Sie an, ob sie diesen Aussagen voll und ganz, überwiegend, weniger oder gar nicht zustimmen.«

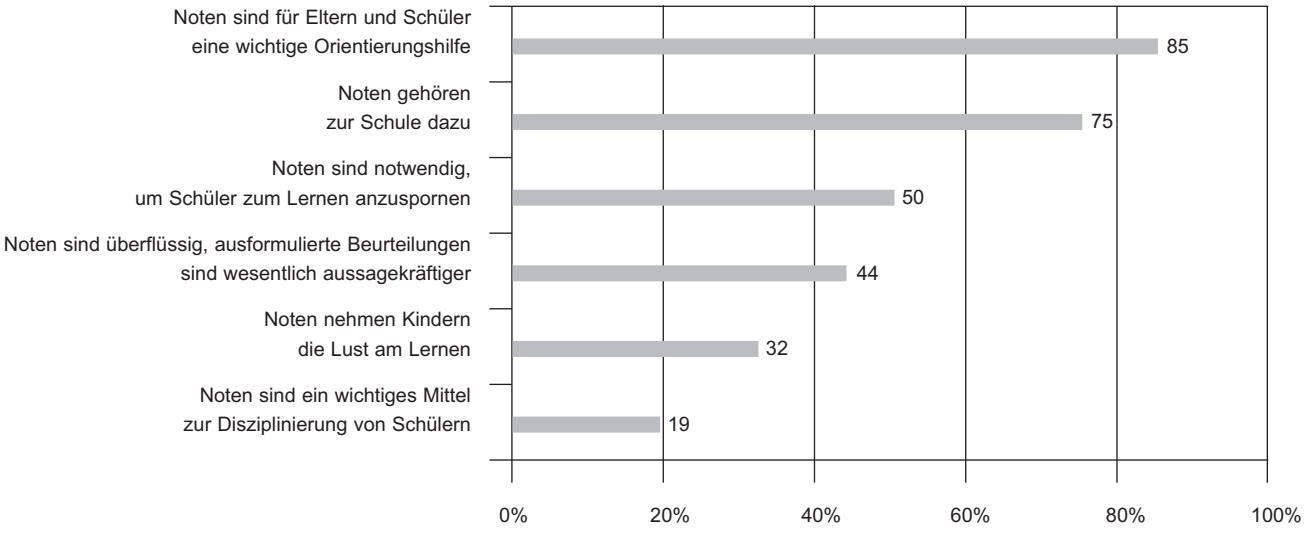



Abb. 5: EFF-Schulbefragung: Ergebnisse der repräsentativen Lehrer-Befragung, September 2005, Pohl/Beekmann, S. 85.

weiterführenden Schulen[251]. Generell finden weder die Motivationsfunktion (50% Zustimmung) noch die Disziplinierungsfunktion (19%) eine Mehrheit. Im Vordergrund steht die Orientierung der SchülerInnen und Eltern. Auch in einer Befragung zur Funktion von Zeugnissen, an der 81 Berliner LehrerInnen teilnahmen, hoben diese die Rückmeldefunktion als wesentlich hervor[252]. Hier zeigt sich ein Dilemma: Ziffernnoten sind SchülerInnen und Eltern vertrauter; deshalb werden sie auch von vielen LehrerInnen als notwendig betrachtet, obwohl diese selbst ihren Informationsgehalt gering einschätzen.

Durchgängig positiv beurteilen LehrerInnen aus Rheinland-Pfalz und Thüringen den Verzicht auf Noten im neuen Fach Englisch in der Grundschule, wie Gompf/Henrich (2005, 10-11) herausfanden: »Von dem Gesamt der Lehrkräfte wird die ›verbale Kommentierung‹ von 76% befürwortet. [...] Als wichtigster Grund zählt für zwei Drittel das Argument: ›Die Kinder beginnen mit Englisch erst ab Klasse 3 und sollten daher in einer unbelasteten Atmosphäre lernen, in gleicher Weise, wie ihnen dies für die Fächer ab 1. Schuljahr ermöglicht wird‹ (73%). Nahezu ebenso wichtig ist diesen Lehrkräften als zweites Argument: ›Auch langsamer lernende beziehungsweise schüchterne Kinder werden ermutigt, sich immer wieder neu zu erproben und ihr Können zu verbessern‹ (71%). Die Hälfte der Lehrkräfte schätzt es, dass sie differenziert auf die Teilbereiche Hörverstehen, Sprechen, Lesen und Schreiben eingehen können (50%). Rund einem Drittel aller Lehrerinnen ist auch der Aspekt wichtig: ›Die Eltern erhalten genauere Informationen über die Stärken beziehungsweise Schwächen ihres Kindes‹.«

Leider haben zwei Bundesländer, in denen Frühenglisch benotet wird, ihre Teilnahme an der Studie widerrufen, so dass sich nicht einschätzen lässt, welche Rolle die Erfahrungen in der eigenen Praxis für die positiven Beurteilungen spielen. Andererseits bestätigt diese Studie den oben genannten allgemeinen Befund, dass die Mehrheit der GrundschullehrerInnen Ziffernnoten ablehnt.

Kritik von LehrerInnen bezieht sich allerdings auf den erheblichen Zeitaufwand, der auf rund drei Stunden pro Kind beziffert wird[253]. So weit manche Lehrkräfte den zusätzlichen Aufwand für die Beobachtung und Dokumentation der Lernprozesse betonen[254], stellt dieser Befund allerdings eher die Validität der Ziffernbenotung in Frage. So weit andererseits das Schreiben aussagekräftiger Berichte gemeint ist, verdient dieser Einwand Beachtung. Wir werden ihn deshalb noch einmal gesondert aufgreifen (> Kap. 5).

Insgesamt zeigt sich, dass LehrerInnen, die über mehr Erfahrungen mit Berichtszeugnissen verfügen, und auch diejenigen, die als Teilzeitkräfte mehr Zeit haben, dieser Darstellungsform positiver gegenüberstehen[255]. Überdies ist die Einstellung zu Noten in Grundschulen durchgängig kritischer als in den weiterführenden Schulen.

## 4.2

## Einschätzungen von SchülerInnen

Befragungen von SchülerInnen erbringen sehr unterschiedliche, zum Teil widersprüchliche Ergebnisse.

Bei Schröter (1982) sprachen sich nur 10% der SchülerInnen dafür aus, Noten abzuschaffen. Auch nach der Befragung von Weiß (1986) hielten in den 1970er Jahren 71% der SchülerInnen Zensuren für notwendig. Und in der Hamburger Studie »LeiHS«[256] stimmten nur 12% der befragten SchülerInnen für eine »Schule ohne Noten«.

Nach einer Umfrage der Zeitschrift ELTERN hielten dagegen 58% der 2.060 befragten SchülerInnen Noten für »... unnütz oder sogar schädlich; denn sie seien ungenau, ungerecht, ohne Aussagekraft über die tatsächliche Leistungsfähigkeit. Die häufigsten Kritikpunkte waren: Gute Noten machen überheblich, schlechte mutlos und - Noten verschärfen den Konkurrenzkampf in der Klasse.«[257]

Das Problem solcher Umfragen sind die geforderten Pauschalurteile. Sie lassen den Befragten wenig Raum, ihre Einschätzungen zu differenzieren, und sie eröffnen den InterpretInnen vielfältige Deutungsmöglichkeiten. Aufschlussreicher sind deshalb Studien, die die Erfahrungen und Meinungen detaillierter erfassen und zusätzlich Untergruppen von SchülerInnen unterscheiden[258].

Generell lässt sich ein Alterseffekt feststellen. So fanden in der von Valtin u.a. durchgeführten Berliner Studie fast alle Kinder die erste verbale Beurteilung »(sehr) gut«, dagegen sprachen sich im 5. Schuljahr schon 50% für Ziffernnoten aus[259]. Diese Tendenz korrespondiert mit entsprechenden Tendenzen bei LehrerInnen und Eltern (> Kap. 4.1 und 4.3).

Maier (2001) hat detailliert erfasst, worin die Kinder die Stärken der beiden Varianten sehen. Gründe der Kinder, die für ein Verbalzeugnis sprechen[260]:

---

251 So auch Jachmann/Tillmann (2000, 68-69). In ihrer Hamburger Studie sprechen sich allerdings nur 34% der GrundschullehrerInnen für eine »Schule ohne Zensuren« aus - damit aber immer noch deutlich mehr als die 4% BefürworterInnen im Gymnasium (a.a.O., 33).

252 Valtin/Schmude (2002, 24).

253 So Valtin (2002b, 14) unter Verweis auf eine unveröffentlichte Studie von Freese (1990).

254 Vgl. Maier (2001, 117).

255 Vgl. Jachmann/Tillmann (2000, 69) und die differenzierten Daten von Jachmann (2003, 122-123, 138-141).

256 Vgl. Vollstädt/Jachmann (2000, 148, 152).

257 Bambach (1994, 9). Auch nach einer Befragung von Czerwenka u.a. (1988) im 10. Schuljahr erleben nahezu 50% Zensuren negativ.

258 Vgl. vor allem Maier (2001), Beutel (2004) und aus dem NOVARA-Projekt: Darge u.a. (2002) und Ostrop u.a. (2002).

259 Valtin (2002c, 140), die ergänzend berichtet, dass sich mehr als drei Viertel der SchülerInnen gerecht beurteilt fühlten (s. dazu auch Ostrop u.a. 2002, 57).

260 Maier (2001, 123).

- verständliche Begründung der Leistungsbeurteilung (29,3%)
- Ausführlichkeit (27,7%)
- Verbesserungsmöglichkeiten werden aufgezeigt (14,9%)
- keine Begründung (12,7%)
- nicht so großer Konkurrenzdruck (7,8%)
- persönliche Einschätzung durch die Lehrkraft (7,6%)

Vorteile des Ziffernzeugnisses[261]:
- bessere Selbsteinschätzung möglich (43,5%)
- höhere Akzeptanz durch familiäre Umwelt[262] (31,6%)
- keine Begründung (11,2%)
- Fortschritt ist besser messbar (9,2%)
- bessere Vergleichsmöglichkeit mit anderen Kindern (4,5%).

Wie für viele LehrerInnen ist also auch für Kinder die unterstellte Erwartung der Eltern ein wesentlicher Grund dafür, dass sie Ziffernnoten bevorzugen. Bei einer beachtlichen Zahl steht dabei der materielle oder emotionale Tauschwert im Vordergrund[263].

Im Übrigen ist eine deutliche Polarisierung wahrnehmbar: Während gut 40% angeben, sich selbst mit Hilfe von Ziffernzeugnissen besser einschätzen zu können, betonen fast 30% die verständlichere Begründung der Verbalgutachten. Die Konsequenz: »Während annähernd ein Viertel der Kinder das Verbalzeugnis bevorzugen, wünscht sich ein Drittel eine Kombination aus Verbalzeugnis und Ziffernzeugnis. Das Ziffernzeugnis wird den anderen Rückmeldeformen mit deutlichem Abstand vorgezogen. [...] Weiterhin zeigt sich ein zwar statistisch nicht signifikanter, aber deutlicher Zusammenhang zwischen der Leistungseinschätzung der Kinder und der Präferenz der Form der Leistungsrückmeldung: Leistungsstärkere Kinder ziehen das Ziffernzeugnis vor und leistungsschwächere Kinder tendieren eher dazu, sich ein Verbalzeugnis zu wünschen.«[264]

Der letzte Befund passt zu den Ergebnissen, die die Studien zu den Auswirkungen der sozialen und der individuellen Bezugsnorm, insbesondere auf die Lernmotivation, erbracht haben (vgl. > Kap. 2 und 3.2.3).

Wie Maier stellen auch Vollstädt/Jachmann (1999) und Valtin (2002c) fest, dass die Mehrheit der Kinder Ziffernzeugnisse bevorzugt, dass sie sich aber schriftliche Kommentare, vor allem als Förderhilfe, wünscht. Dieses inhaltliche Interesse an einer gehaltvollen Rückmeldung hebt auch Beutel (2004) hervor. Sie stellt in ihren Interviews zu Verbalzeugnissen und kommentierten Ziffernnoten fest, dass Kinder sich sehr differenziert zur Qualität der Aussagen äußern können und resümiert[265]: »Berichte geben Kindern eine wichtige Auskunft über fachliches Lernen. Man kann nach Auswertung der Gespräche mit Grundschulkindern die Lehrerinnen und Lehrer nur auffordern, weniger Mühe in das sprachliche Verkleiden von Lerndefiziten und -mängeln zu investieren. Vielmehr scheint man Kindern mehr an Klarheit und Wahrheit zutrauen zu dürfen, als dies im pädagogischen Geschäft bisweilen der Fall ist […] Berichtszeugnisse werden von Kindern als diagnostisch gehaltvoller im Vergleich zu Notenzeugnissen wahrgenommen. Dabei wünschen Kinder ihrer Person angemessene Urteile. Solche Beschreibungen üben einen besonderen Reiz aus. Kinder fordern ein, dass auch Defizite benannt werden. Schwächen sollen ausgesprochen und nicht sprachlich zugedeckt werden.«

Zwar gilt bei den meisten Kindern der erste Blick den Noten[266], aber von 143 Kindern haben 128 ihr Zeugnis mehrmals gelesen, 136 Kinder sagen, das sie ihr Zeugnis auch später noch lesen werden[267].

Der angeblich verbreitete Notenwunsch von Kindern wird auch durch eine Befragung zum Fach Englisch in der Grundschule in Frage gestellt[268]: »Von den 1193 befragten Jungen und Mädchen haben 796 Kinder (67%) das Item: ›Seit dem 3. Schuljahr bekommst du in allen Fächern eine Note in deinem Zeugnis. Im Fach Englisch nicht. Findest du das gut?‹ mit ›JA‹ angekreuzt. [...] Von den 796 Kindern rangiert bei 533 an erster Stelle das Argument: ›Ich mache in Englisch lieber mit, wenn ich dafür keine Note bekomme‹ (45%). [...] Den insgesamt 796 Kindern, die Englischnoten ablehnen, stehen 386 gegenüber, die gerne eine Zeugnisnote hätten. Von diesen Kindern sagen nahezu alle, dass sie sich dann ›mehr anstrengen‹ würden (357; 32%). 122 Kinder begründen ihren Wunsch, in Englisch eine Note zu wollen, ferner mit der Aussage: ›Weil ich bei einer guten Englischnote mehr Taschengeld bekomme‹ (10%).«

Die unterschiedlichen Ergebnisse in den referierten Untersuchungen machen deutlich, dass die Einschätzungen der SchülerInnen von verschiedenen Faktoren abhängen: von ihren eigenen Erfahrungen, vom Kontext der Befragung und wohl auch von der Art, wie die Frage selbst formuliert wird. So lehnten in der Hamburger »LeiHS«-Studie einerseits rund 70% der SchülerInnen einen Ersatz der Noten durch Verbalbeurteilungen ab[269]. Nach ihren Erwartungen an

261 Maier (2001, 124).

262 Das Ziffernzeugnis bietet aus Sicht der Kinder den Vorteil, dass es von den Eltern, Großeltern und Verwandten in höherem Maße akzeptiert wird als ein Verbalzeugnis. Auf dieses verbreitete Missverständnis schulischen Lernens weist auch die Denkschrift der Bildungskommission Nordrheinwestfalen (1995, S.87) hin: *»Schülerinnen und Schüler machen zum Teil früh die Erfahrungen, dass ihre Umgebung, vor allem die eigene Familie, sich weniger für das Lernen selbst, für seine Schwierigkeiten und Inhalte interessiert, als für seine Ergebnisse in Form quantifizierend bewerteter Leistungen«.* (Maier 2001, 126).

263 Vgl. die Rangfolge der Gründe für »Zeugnisse« bei Kirschner (1992, 83). Inzwischen gibt es sogar Schulen, die ihre SchülerInnen für gute Noten bezahlen, vgl. zu einem Beispiel aus Bristol in England Stepanek (2005).

264 Maier (2001, 121, 125).

265 Beutel (2004, 167-168; s.a. 2000, 177).

266 In der Berliner Studie steht schon für Zweitklässler bei Zeugnissen die Auslesefunktion im Vordergrund (vgl. Valtin 2002c, 140, und ergänzend Valtin/Schmude 2002, 18-21).

267 Vgl. Beutel (2004, 197, 199).

268 Gompf/Henrich (2005, 6).

269 Gerundete Werte aus Tabelle 4/17 in Vollstädt/Jachmann (1999, 131).

Zeugnisse befragt, stimmte aber die Mehrheit der SchülerInnen den folgenden Items »ganz/überwiegend« zu[270]:

»Zeugnisse sollen mir sagen,
was ich in einzelnen Fächern kann.« (85%)

»In einem Zeugnis erwarte ich Hinweise,
wie ich mich verbessern kann.« (73%)

»Durch ein Zeugnis möchte ich erfahren,
was ich in dem Schuljahr dazugelernt habe.« (66%).

Dagegen unterstützten nur 44% die Aussage

»Durch ein Zeugnis möchte ich erfahren,
ob ich besser oder schlechter
als andere Schüler(innen) bin.«

Es wäre interessant zu wissen, wie die Ergebnisse ausfallen würden, wenn man Kinder fragte: »In vielen anderen Ländern bekommen die SchülerInnen keine Noten auf den Zeugnissen, sondern Hinweise zu ihren Fortschritten und zu ihren Schwierigkeiten. Findest du das besser oder schlechter als bei uns?«. Einen Hinweis auf die zu erwartenden Ergebnisse liefert die Analyse von 1.212 Aufsätzen aus 4. bis 13. Klassen zum Thema »Schule«, die mit analogen Befragungen in anderen Ländern verglichen wurden[271]. Sowohl die quantitative als auch die qualitative Analyse ergaben, dass Leistungsbeurteilung für deutsche SchülerInnen häufiger ein Thema war als für SchülerInnen in anderen Ländern[272]. Zudem war der Tenor der Aussagen innerhalb der deutschen Gruppe wesentlich häufiger negativ[273]. Die AutorInnen resümieren: »Wir sehen, dass die deutschen SchülerInnen am meisten unter Noten leiden. Amerikanische Schüler betonen sogar häufiger - wenn sie sich überhaupt zu Noten äußern - den positiven Rückmeldeffekt von Zensuren als die belastende Kontrollfunktion. Das hängt sicher mit dem wesentlich geringeren Selektionsdruck zusammen.«[274]

## 4.3

## Einschätzungen von Eltern

Wie unter den LehrerInnen und SchülerInnen so hielten auch unter den Eltern vor 30 Jahren rund drei Viertel Zensuren für notwendig[275]. Schon damals führten differenziertere Befragungen auch zu differenzierteren Einschätzungen, wie die drei folgenden Untersuchungen zeigen[276]:

Schmack (1978) wertete in einer eher informellen Studie Notizen aus Elterngesprächen während der Zeugnisausgabe aus und fand mit über 60% eine hohe Zustimmung und Zufriedenheit der Eltern, obwohl eine Analyse der Berichte zeigte, dass sie noch wenig differenziert formuliert waren.

Schmidt (1981) stellte zwar fest, dass die Eltern die Vorbehalte der LehrerInnen nicht teilten, kommt andererseits dennoch zu dem Schluss[277]: »Die Hoffnungen aber, die sich mit Verbalzeugnissen verbanden, haben sich nach unseren Untersuchungen nicht erfüllt«.

Positiver fielen die Ergebnisse einer Elterbefragung durch Schlottke/Speidel (1981) in Baden-Württemberg aus. Danach haben Eltern qualifizierte Erwartungen an ziffernfreie Zeugnisse: 67% möchten über Entwicklungsfortschritte der Kinder, 42% über Förderungsmöglichkeiten informiert werden und 41% erwarten Ermutigungen des Kindes. Anhand der konkreten Verbalbeurteilungen sahen sich 59% der Eltern besser informiert. Nur 12% der Eltern fehlt die Möglichkeit der sozialen Einordnung des Kindes in die Lerngruppe, 9% befürchten den später auftretenden Notendruck, nur 5% attestieren den Verbalbeurteilungen eine geringe Aussagequalität.

Inzwischen sind Verbalbeurteilungen fester Bestandteil der Grundschularbeit, also über die unvermeidlichen Schwächen der Anfangsphase hinaus und auch den Eltern vertrauter[278]. Andererseits sind es unter den LehrerInnen nicht mehr die Pioniere, die dieses Instrument einsetzen, und in der Breite ist bei jeder pädagogischen Reform mit einer Verwässerung der Intentionen zu rechnen. Wie also gewichten Eltern heute die Vorteile und Schwächen von Ziffernnoten bzw. Verbalbeurteilungen?

Die aktuellste Befragung stammt von FORSA für die Zeitschrift »Eltern for Family« - analog aufgebaut zu den bereits berichteten Befragung von LehrerInnen (s. > Kap. 4.1). Im folgenden Schaubild sind sie nach Jahrgangsstufe der Kinder aufgeschlüsselt. Mit um die 90% wird noch stärker als bei den LehrerInnen die Orientierungsfunktion betont. Nur ein Viertel bis ein Drittel der Eltern hält Noten für überflüssig und ausformulierte Beurteilungen für wesentlich aussagekräftiger. Das sind deutlich weniger als unter den LehrerInnen - vor allem im Grundschulbereich: Auf dieser Schulstufe sind es nur 24% der Eltern gegenüber etwa 60% der LehrerInnen. Insofern gibt es unter den Eltern einen umgekehrten Jahrgangseffekt, was die Kritik an Noten betrifft: Während nur 17% der Grundschuleltern meinen, dass Noten den Kindern die Lust am lernen nehmen, sind es in der Oberstufe schon 30%.

---

270 Die Werte sind gerundet übernommen aus Tabelle 4/19 in Vollstädt/Jachmann (1999, 132).

271 Czerwenka u.a.(1990); vgl. die Zusammenfassung (420-421).

272 Über 50% in Deutschland gegenüber knapp 40% in Frankreich und weniger als 20% in den USA.

273 Über die Hälfte in Deutschland gegenüber einem guten Drittel in Frankreich und weniger als einem Fünftel in den USA.

274 Czerwenka u.a. (1990, 422).

275 Vgl. Weiß (1986).

276 Referiert bei Beutel (2005, 65-68).

277 Schmidt (1981, 488).

278 So stellte Wallrabenstein (1992, 120) schon über die 1980er Jahre hinweg einen Anstieg der Elternzustimmung zu Berichtszeugnissen in Klasse 3 auf fast 50% fest.

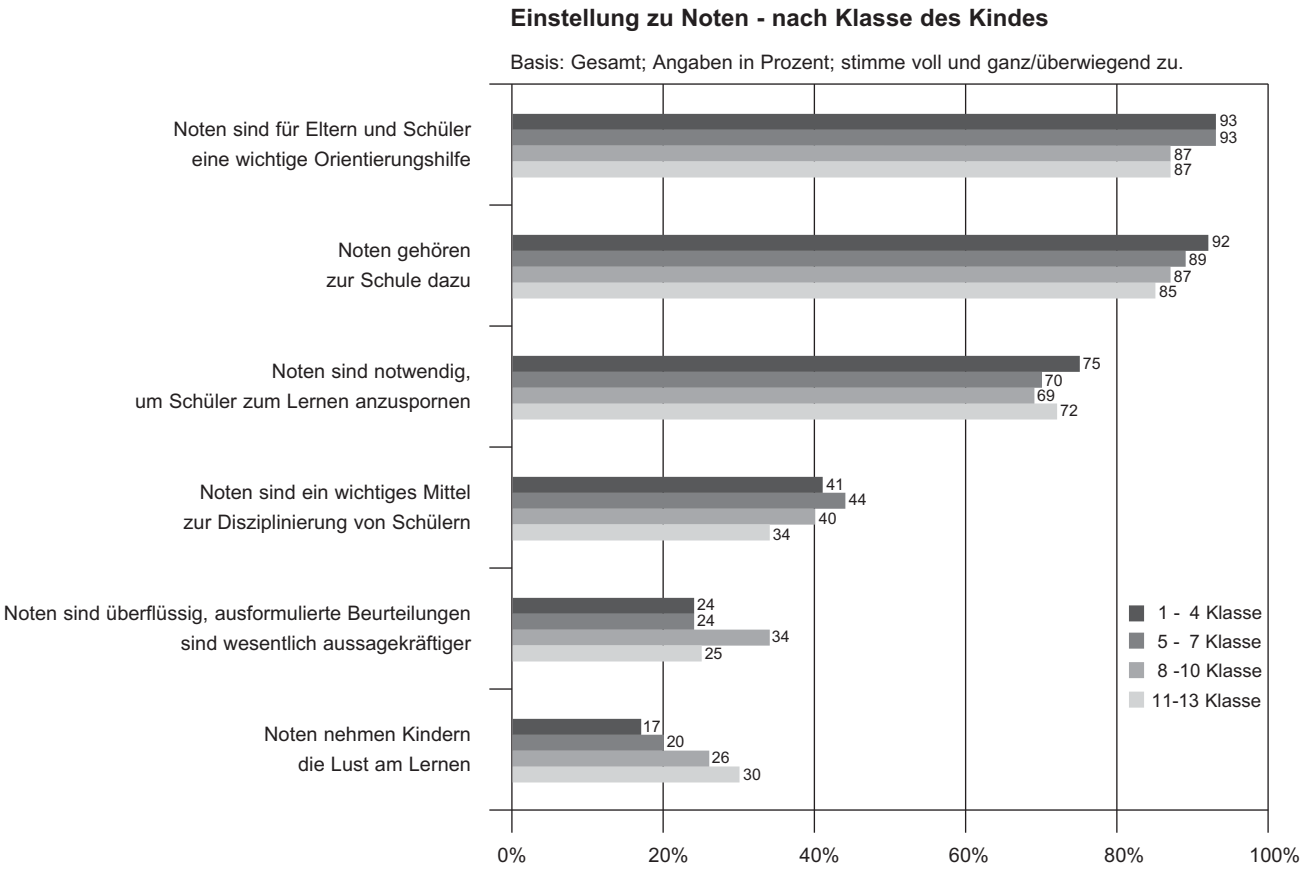



Abb. 6: EFF-Schulbefragung: Ergebnisse der repräsentativen Eltern-Befragung, September 2005, Pohl/Beekmann, S.109.

Aber auch hier lohnt ein genauerer Blick auf die Motive und Begründungen, wie ihn einige detailliertere Befragungen erlauben[279].

In den 20 Schulen des Schulversuchs in NRW, in dem die Notenfreiheit auf Antrag bis zur 4. Klasse ausgedehnt werden konnte, äußerten sich nicht nur die LehrerInnen, sondern auch die Eltern sehr positiv. In einer Befragung stimmten 80-90% den folgenden Aussagen zu[280]:

▲ schriftliche Hinweise unter den Arbeiten und im Zeugnis sind unverzichtbar;

▲ Berichte und Kommentare sind viel aussagekräftiger als Noten:

▲ sie helfen, die richtige Wahl für die weiterführende Schule zu treffen.

Auch die LehrerInnen stellen bei den Eltern nach Ausweitung der Entwicklungsberichte auf Klasse 3/4 ein erhöhtes Interesse an der Lernentwicklung ihrer Kinder fest und berichten, dass die Gespräche zwischen Eltern und LehrerInnen intensiviert wurden.

Da die Eltern dem Versuch schon vorher zugestimmt hatten, dürfte es sich allerdings um eine positive Stichprobe handeln. In der Breite ist mit mehr Vorbehalten zu rechnen, wie beispielsweise die Befunde aus der Berliner NOVARA-Studie zeigen. Valtin fasst die Erfahrungen und Erwartungen der Eltern bezogen auf die Rückmeldung zu den Leistungen ihrer Kinder wie folgt zusammen[281]:

- Eltern sind Informationen zum Lernstand, aber auch konkrete Hinweise für die Förderung besonders wichtig.
- Sie erwarten auch eine detaillierte Rückmeldung zum Arbeits- und Sozialverhalten.
- Den Informationsgehalt und die Verständlichkeit der Verbalbeurteilungen schätzen sie eher »verhalten« ein - zusätzlich mit abnehmender positiver Bewertung über die Grundschulzeit hinweg: »Selbst 80% der Anhänger der Verbalbeurteilung unterstützen die Aussage: ›Bei einem Notenzeugnis weiß man genau, wo das Kind steht.‹«
- Andererseits sehen Eltern Vorteile der Verbalbeurteilungen in den konkreten Hinweisen auf Stärken und Schwächen und Förderhinweisen zu deren Überwindung - auch diejenigen, die für Noten sind.
- Insgesamt bevorzugen Eltern also eine Verbindung beider Darstellungsformen.

Die Hamburger Grundschuleltern urteilen positiver als die Berliner, aber sie bevorzugen ebenfalls eine Kombination

279 Vgl. Haenisch (1996b); Lütgert/Jachmann (2000); Maier (2001); Rosenfeld/Valtin (2002); Valtin/Rosenfeld (2002); Gompf/Henrich (2005).
280 Vgl. Haenisch (1996b, 16, 34).
281 Valtin (2002c, 142-143); vgl. ausführlicher Valtin/Schmude (2002, 21-24).

von Ziffern und Kommentar[282]. Insgesamt stimmen Eltern eher der Zeugnisform zu, die an der Schule ihrer Kinder praktiziert wird[283] - darum sind vermutlich die Hamburger Bewertungen positiver ausgefallen als in Berlin, wo die notenfreien Zeugnisse weniger verbreitet waren.

Auch im Modellversuch »Lern- und Spielschule« in Rheinland-Pfalz decken sich die Wünsche der Eltern mit den Voten aus Berlin und Hamburg[284]: »Hinsichtlich der Präferenz der Form der Leistungsrückmeldung wird deutlich, dass die Kombination von Verbalzeugnis und Ziffernzeugnis von allen Gruppen mit deutlichem Abstand (71,3%) gegenüber den anderen Beurteilungsformen (Verbalzeugnis 11,7%, Ziffernzeugnis 17,0 %) vorgezogen wird. [...] Insgesamt kann nicht davon ausgegangen werden, dass bei den befragten Eltern die Akzeptanz von Zeugnissen ohne Noten sehr hoch ist. Sie streben vielmehr mehrheitlich eine pragmatische Lösung nämlich die Ergänzung beider Formen der Leistungsrückmeldung an, wodurch ihrer ambivalenten Einschätzung Rechnung getragen wird. [...] Für diejenigen Eltern, die das Beurteilungssystem ohne Noten weniger akzeptieren können, scheint durch die verbale Beurteilung keine ›Klarheit‹ bei der Einschätzung der Leistung ihrer Kinder zu bestehen. Noten, die ihnen aus ihrer eigenen schulischen Sozialisation bekannt sind und deshalb ein System bilden, das ihnen vertraut ist, scheinen für sie die einzige Chance zu sein, die Lernentwicklung und Lernsituation richtig einschätzen zu können.«

In der Befragung wurden die Einschätzungen der Eltern sehr differenziert erfasst[285]:

*Vorteile des Verbalzeugnisses:*

- konkrete, auf das Individuum bezogene Leistungsrückmeldung (32,4%)
- detaillierte Information über Leistungsentwicklung (14,9%)
- in einer für das Kind verständlichen Sprache formuliert (13,7%)
- nicht nur Leistung in den Fächern, die gesamte Persönlichkeit wird berücksichtigt (8,1%)
- Verringerung der Schulangst (vor allem bei schwachen Kindern) (8,1%)
- besondere Berücksichtigung des Sozialverhaltens (8,1%)
- Stärkung des Selbstwertgefühls (8,0%)
- Kinder werden besser motiviert (5,4%)
- keine Vorteile gegenüber Ziffernzeugnis (1,3%)

*Nachteile des Verbalzeugnisses:*

- Interpretationsprobleme (Eltern) (42,9%)
- Probleme beim Übergang zur weiterführenden Schule (12,5%)
- Fehlende Objektivität (12,5%)
- Interpretationsprobleme (Kinder) (7,1%)
- Keine Nachteile (7,1%)
- Nicht alle Lernbereiche werden erwähnt (5,4%)
- Euphemistische Formulierungen (5,3%)
- Gespräch mit Lehrkraft ist notwendig (3,6%)
- Negative Eigenschaften des Kindes werden beschrieben (3,6%)

*Vorteile des Ziffernzeugnisses:*

- Vergleich mit anderen Kindern (24,1%)
- klare Leistungseinordnung, exakte Rückmeldung an Eltern (19,0%)
- bessere Einschätzung der Leistung durch Kind (12,7%)
- kein Interpretationsspielraum (10,1%)
- weiterführende Schulen wollen Noten (8,9%)
- jeder Lernbereich wird erwähnt (7,6%)
- Motivation der Kinder (6,3%)
- Kinder gewöhnen sich an späteren Beurteilungsmodus (5,1%)
- Kinder wollen Noten und freuen sich darauf (3,8%)
- keine Vorteile gegenüber Verbalzeugnis (2,4%)

*Nachteile des Ziffernzeugnisses (Maier (2001) Seite 132):*

- Verunsicherung und Stigmatisierung leistungsschwacher Kinder (28,7%)
- keine Rückmeldung über Leistungsentwicklung (25,0%)
- Förderung konkurrenzorientierten Verhaltens (21,2%)
- keine konkreten Informationen über das Kind (11,5%)
- keine Informationen über das Sozialverhalten (7,7%)
- keine Nachteile (5,9%)

*Was fehlt den Eltern ohne Noten?*

- Klarheit über Leistungsstand (38,0%)
- Vergleich mit anderen Kindern (16,8%)
- Rückmeldung an weiterführende Schule (15,8%)
- nichts (9,7%)
- Noten gehören zur Schule (9,4%)
- Hilfe für die Übergangsentscheidung (3,8%)
- Leistungsbeurteilung in jedem Fach (3,8%)
- Motivation für Kinder (2,9%)

282 Vgl. Lütgert/Jachmann (2000, 96-97, 109).

283 Vgl. Lütgert/Jachmann (2000, 109) und Valtin (2002c, 143), die darauf hinweist, dass Eltern mit höherem Bildungsabschluss Verbalbeurteilungen ebenfalls positiver einschätzen als Eltern mit niedrigeren Abschlüssen.

284 Maier (2001, 27, 134, 119). Dabei ergab sich »ein hoch signifikanter Zusammenhang zwischen der von den Eltern für das Kind vorgesehenen Schulart und der Präferenz der Rückmeldeform.« (a.a.O., 27).

285 Maier (2001, 129-134).

Diese detaillierten und zum Teil widersprüchlichen Aussagen lassen sich auf einen knappen Nenner bringen[286]: »Aus der Vielzahl der gewonnenen Befunde ist besonders die Ambivalenz der Befragten zu nennen. So wünschen sich beispielsweise die Eltern einerseits klare Rückmeldungen darüber, wo ihr Kind im Bezugssystem der Schulklasse steht; andererseits sind sie an individuellen Informationen über die Entwicklung ihres Kindes interessiert und beklagen die Konkurrenzorientierung der Noten ...«.

Deutlich positiver fallen die Reaktionen in der Elternbefragung zu Englisch in der Grundschule aus[287]: 66% begrüßen in der betreffenden Untersuchung die Notenfreiheit ausdrücklich. Als zentrales Argument nennen 54 % (wie schon die LehrerInnen und Kinder): »Mein Kind lernt erst seit Klasse 3 Englisch. Es sollte dieses Fach daher in einer unbelasteten Atmosphäre lernen, in gleicher Weise, wie es dies in den Fächern machen darf, die ab erstem Schuljahr beginnen«.

Gefordert wird von den Eltern allerdings eine aussagekräftige Rückmeldung zum Lernstand des Kindes[288]. Dies ist über die verschiedenen Befragungen hinweg das vorrangige Bedürfnis der Eltern. Bisher findet die Mehrheit, dass dieses Bedürfnis am besten durch eine Kombination von Note und Bericht befriedigt wird.

## 4.4

## Einschätzungen von Arbeitgebern

Ob Unternehmen, Verwaltung oder auch Schule, die selbst[289] Personal einstellen: Neben den Noten spielen Verbalbeurteilungen und Aufnahmegespräche oder eigene Eingangstests eine große Rolle für die Auswahl von BewerberInnen. Dass Arbeitgeber eigene Eingangsprüfungen durchführen, zeigt, wie wenig sie den Noten von Schulen und anderen Ausbildungseinrichtungen trauen.

Diese Skepsis wird auch deutlich in Befragungen zur Bedeutung verschiedener Informationsquellen bei Einstellungen.

Das Bundesinstitut für Berufsbildung (1998) hat 1575 Fragebögen aus 805 Betrieben zu der Frage ausgewertet, wie verschiedene Informationsquellen bei der Einstellung von MitarbeiterInnen gewichtet werden. Danach halten nur 20% die Berufsschulnoten für »sehr wichtig«, während über 90% die Eindrücke aus dem Vorstellungsgespräch hoch einschätzen. Weniger als 25% finden in dem IHK Zeugnis »wertvolle Hinweise auf die berufliche Handlungsfähigkeit«. Das Berufsschulzeugnis schätzen 90% mit Blick auf grundlegende Schulkenntnisse als aussagekräftig ein und 70% vertrauen den Aussagen über fachliches Wissen. Aber nur eine Minderheit zieht aus diesen Quellen Informationen für Pünktlichkeit/Zuverlässigkeit, Sorgfalt/Genauigkeit, Planungs-/Organisationsfähigkeit, Kommunikationsfähigkeit, praktische Fertigkeiten, Schnelligkeit, Einfallsreichtum, Kontakt/Teamfähigkeit. In allen diesen Punkten wird das (verbale) Ausbildungszeugnis des Betriebs höher eingeschätzt.

Auch in Prognosestudien findet sich nur eine geringe Korrelation von Schulnoten mit Kennwerten für berufspraktische Bewährung[290]. Ein wesentlich engerer Bezug besteht zwischen Umfang und Niveau einschlägiger außerschulischer Aktivitäten und dem Erfolg in der Berufspraxis[291]. Folgerichtig haben neben persönlichen Einstellungsgesprächen biografische Fragebögen für die Auswahl von BewerberInnen an Bedeutung gewonnen und sich auch bei schwierigen Entscheidungen bewährt[292].

Hier deuten sich interessante Parallelen zwischen den Argumenten für Verbalgutachten in der Schule und dem betrieblichen Beurteilungswesen an, in dem individuelle Zielvereinbarungen, Selbsteinschätzungen und regelmäßige Mitarbeitergespräche in den letzten Jahren zunehmend an Bedeutung gewonnen haben.

## 4.5

## Einschätzungen in der Öffentlichkeit

Aktuelle Repräsentativbefragungen zum Thema »Noten« liegen vom Institut für Schulentwicklung an der Universität Dortmund (IfS)[293] und im »Bildungsbarometer« des Zentrums für empirische pädagogische Forschung an der Universität Koblenz-Landau vor[294].

Das ZEPF (2005) resümiert die Antworten in seinem Bildungsbarometer in dem klaren Fazit: »Mit deutlichem Abstand werden dagegen alle Vorschläge, die im weitesten Sinne auf eine Abschaffung gängiger Druckmittel (Sitzenbleiben, Noten) hinauslaufen, nur von einem geringen Anteil der Bevölkerung befürwortet.«

---

286 Petillon (2001, II).

287 Vgl. Gompf/Henrich (2005, 3). Ein Grund für die Abweichung vom allgemeinen Trend könnte sein, dass dieses Fach nicht versetzungsrelevant ist.

288 ... wie sie in Thüringen üblich ist, während in Rheinland-Pfalz nur die Teilnahme bescheinigt wurde (Gompf/Henrich 2005, 13-14).

289 ... bei der Besetzung »schulscharf« ausgeschriebener Stellen.

290 Vgl. Landmesser u.a. (2003, 11-12); die Korrelationen liegen bei .09-.26 (Samson u.a. 1984); .22 (Baron-Boldt u.a. 1988); .27 (Hübner 2003) - gegenüber .45 (Baron-Boldt u.a. 1988) bis .54 (Hübner 2003) für den Zusammenhang von Schulnoten und Zensuren im Studium oder an Berufsakademien.

291 Mit einer Korrelation von .54; a.a.O., 12 und 17, mit Bezug auf Hübner (2003) sowie auf Ghiselli (1966); Reilly/Chao (1982); Hunter/Hunter (1984), die je nach Einsatzbereich Korrelationen von .30 bis .50 berichten.

292 Z.B. bei der Auswahl für Einsatzbereiche mit besonderen Anforderungen, vgl. Landmesser u.a. (2003, 17).

293 Vgl. Kanders/Rolff (2002; 2004); Kanders u.a. (2004).

294 Vgl. ZEPF (2005).

**Heilige Kühe des deutschen Schulsystems**

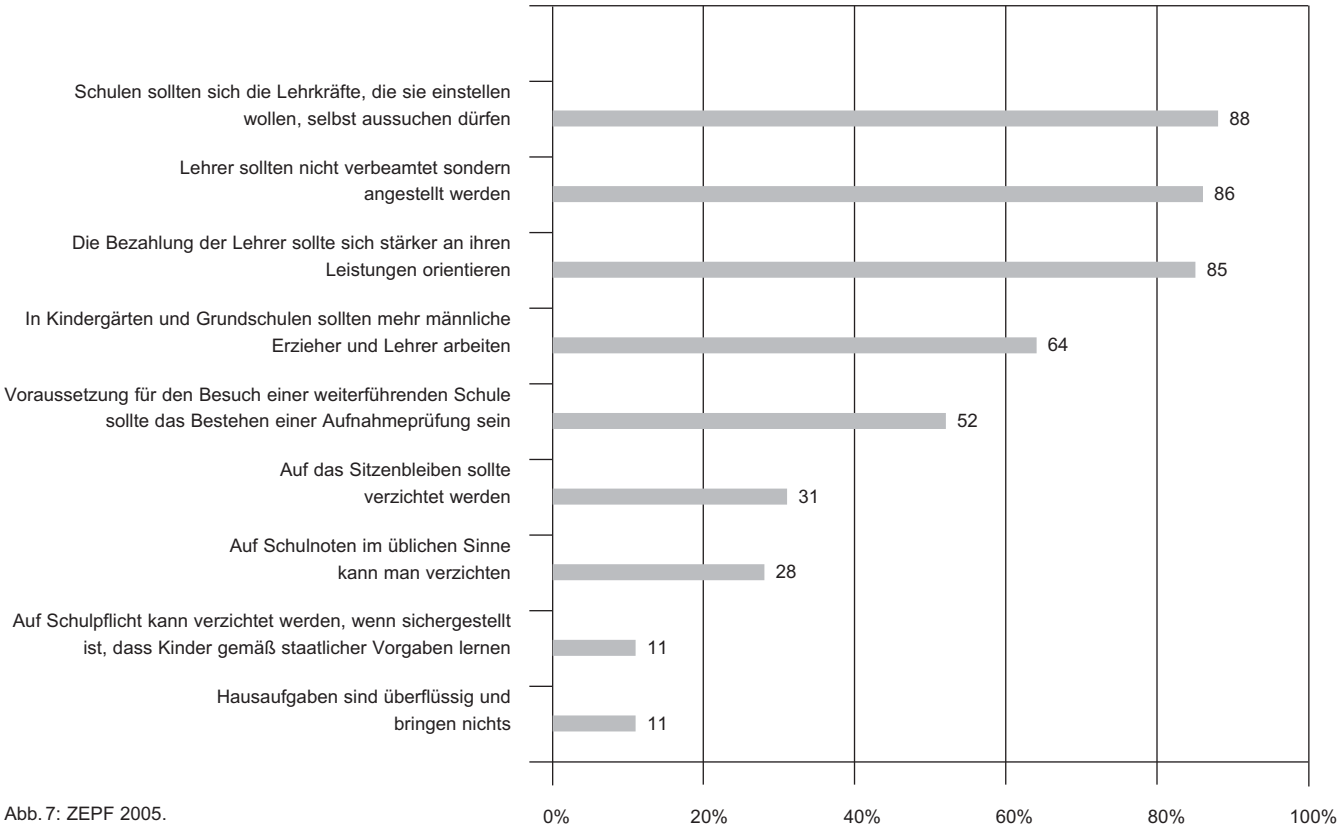



Abb. 7: ZEPF 2005.

Immerhin: Bei einer Befragung durch Schröter (1982) waren es vor 25 Jahren nur etwa 10%, die sich für eine *generelle* Abschaffung aussprachen, während es heute um die 30% sind. Ein differenzierteres Bild vermitteln die Befragungen des IfS in Dortmund, die spezifischer die Leistungsbewertung in der *Grundschule* thematisieren und außerdem die Antworten nach Teilgruppen aufschlüsseln.

Der Verzicht auf Zensuren spaltet das Land, denn insgesamt jeweils etwa 40% sind dafür und dagegen, dabei ist im Westen eine knappe Mehrheit dafür, im Osten dagegen[295]:

»Zumindest in den ersten drei Jahren der Grundschule kann auf Zensuren verzichtet werden« (S. 41).

| Zustimmung | 2004 | 2002 | 1997 | 1993 |
|---|---|---|---|---|
| West | 51% | 43% | 45% | 50% |
| Ost | 26% | 21% | 25% | 29% |

Kopfnoten finden generell eine breite Zustimmung - ob in traditioneller oder in einer den heutigen Erziehungsvorstellungen angepassten Form[296]:

»Alle Schulzeugnisse sollten Noten für Betragen, Fleiß und Ordnung enthalten« (S. 47).

| Zustimmung | 2004 | 2002 |
|---|---|---|
| West | 70% | 66% |
| Ost | 81% | 84% |

»Alle Schulzeugnisse sollten Beurteilungen für Teamfähigkeit, Toleranz und Verantwortungsbewusstsein enthalten« (S. 47).

| Zustimmung | 2004 | 2002 |
|---|---|---|
| West | 75% | 72% |
| Ost | 78% | 78% |

Zentrale Prüfungen finden sowohl in der Gesamtbevölkerung als auch speziell unter Eltern eine überwältigende Zustimmung - nicht nur für das Abitur, sondern auch beim Abschluss der Hauptschule[297]:

»Alle SchülerInnen und Schüler sollten landesweit einheitliche Prüfungen ablegen« (Hauptschule/Abitur) (S.48).

| Zustimmung | 2004 | 2002 |
|---|---|---|
| West: Eltern | 87/91% | 87/91% |
| West: Alle | 84/89% | 84/89% |
| Ost: Alle | 95/97% | 96/97% |
| Ost: Eltern | 96/98% | 95/98% |

295 Kanders u.a. (2004, 41).
296 Kanders u.a. (2004, 47).
297 Kanders u.a. (2004, 48).

In der Öffentlichkeit, vor allem aber in den neuen Bundesländern, herrscht demnach generell eine konservative Haltung vor. Wie beim Sitzenbleiben und bei den Hausaufgaben sprechen sich auch bei den »Schulnoten im üblichen Sinn« nur 30-40% der Befragten für deren Abschaffung aus. Dieser Befund passt zu deutschen und US-amerikanischen Daten, wonach Personen der Schule und vor allem (ungewohnten) Praktiken oft um so kritischer gegenüber stehen, je weniger Kontakt sie zur Schule (über eigene Kinder oder Enkel) haben[298].

## 4.6

## Zwischenbilanz zu »Einschätzungen«

Zunächst ist es wichtig, die deutlichen Differenzen zwischen Teilgruppen wahrzunehmen, die in der Hamburger Studie »LeiHS« so zusammengefasst werden[299]:

▲ »Der pädagogische Dissens unter den Lehrkräften über das Für und Wider von Zensuren findet sich in dieser Schärfe weder bei den Eltern und noch weniger bei den Schüler(innen) wieder.
▲ Die reformpädagogisch motivierte Kritik an Zensuren (Schulangst, geringes Selbstwertgefühl u.ä.) wird von den Befragten geteilt: von den Lehrenden am stärksten, etwas weniger von den Eltern, deutlich weniger von den Schüler(innen) - sogar dann, wenn sie von schlechten Noten betroffen sind.
▲ Die testtheoretischen Mängel der Noten (unzureichende Objektivität, mangelnde »Gerechtigkeit«) werden in allen drei Gruppen nur zum Teil nachvollzogen. Am ehesten findet sich hier an den Grund- und Gesamtschulen eine kritische Einstellung, deutlich weniger im gegliederten Schulsystem.
▲ Von den Schüler(innen) werden die Unzulänglichkeiten von Zensuren am wenigsten reflektiert. Auch ein relativ reformfreudiges Klima unter den Lehrenden der Gesamtschule hat auf die deutliche Befürwortung der Noten seitens der Schüler(innen) nur wenig Einfluss. Anders formuliert: Die Schüler(innen) aller Sekundarschulformen sind die entschiedensten Verfechter der Zensuren. [...]
▲ In der Grundschule gibt es bei Lehrer(innen) und Eltern eine hohe Zustimmung zu den Berichtszeugnissen; in der Haupt- und Realschule sprechen sich Schüler(innen) und Eltern besonders entschieden für Notenzeugnisse aus. Demgegenüber findet sich im Gymnasium eine etwas liberalere Position bei den Eltern, die in einem gewissen Kontrast zu der recht berichtskritischen Position der dortigen Lehrerschaft steht.
▲ In allen Gruppen gibt es eine hohe Wertschätzung der Notenzeugnisse mit Kommentarbogen.«

Insgesamt findet die Abschaffung von Noten in Befragungen nur wenig Zustimmung. Dabei stehen die gegebenen Begründungen (Vergleichbarkeit, Eindeutigkeit) oft in explizitem Widerspruch zu den empirischen Befunden, sind also sachlich nicht gerechtfertigt: »Noten werden positiver bewertet als Verbalgutachten, weil mit ihnen besondere Erwartungen verbunden werden (Eindeutigkeit, Vergleichbarkeit) - die sie aber nicht erfüllen können, was den meisten nicht bewusst zu sein scheint. Wegen ihrer Relativität können sie weder Ausleseentscheidungen rechtfertigen, wegen ihrer Reduktion der komplexen Informationen zum Leistungsstand können sie didaktische Entscheidungen nicht anleiten.«[300]

Nimmt man die in > Kap. 1 bis 3 referierte Vielzahl erdrückender Sachargumente gegen Ziffernnoten und betrachtet man andererseits, wie langsam sie in der Breite wahrgenommen werden, so stellt sich die Frage, ob hier nicht der Gesetzgeber gefordert ist. So wichtig Mitbestimmungsrechte der Betroffenen für die Gestaltung des Schulalltags sind - grundsätzliche Entscheidungen wie die Beurteilung von Leistungen sind gesamtgesellschaftlich zu verantworten, solange sie so nachhaltige Konsequenzen haben wie in unserem selektiven System (> Kap. 7).

Zudem muss die generelle Zustimmung zu Noten differenziert werden. Befragt nach einzelnen Punkten (»macht mir Angst«, »nimmt meinem Kind die Motivation«) äußern sich Mehrheiten in den einzelnen Teilgruppen oft zensurenkritisch. Für Kinder wie LehrerInnen spielt das indirekte Argument der höheren Akzeptanz der Noten bei Eltern, Verwandten und »Abnehmern« eine wichtige Rolle für die eigene Zustimmung. Arbeitgeber dagegen verlassen sich bei der Auswahl von BewerberInnen nicht auf Noten in Abschlusszeugnissen. Insgesamt lassen sich zwei Trends beobachten:
- Personen, die (länger) Erfahrungen mit Verbalgutachten haben, äußern sich generell positiver zu dieser Form der Beurteilung.
- Vor allem die Eltern tendieren zu einer Verbindung von Ziffern und verbalen Aussagen.

Die Vorbehalte gegenüber Verbalbeurteilungen lassen sich auf einen Punkt konzentrieren: »Ein Problemkreis der nahezu alle Arbeiten durchzieht, ist die Kluft zwischen Erwartungen an verbale Zeugnisse und dem oftmals schwierigen Umgang mit dieser Beurteilungsform im schulischen Alltag. Es gibt - vor allem zu Beginn der Phase, in der Grundschulberichte eingeführt werden, Skepsis und Ratlosigkeit bei Lehrern (Schmidt 1991 und Thomas 2001), den Ruf nach administrativen Hilfen und den Wunsch nach Qualifizierung (Schmidt 1981, Schlottke/Speidel 1979/1981) ...«[301].

Damit stellt sich - zusätzlich zu verbreiteten Vorbehalten unter den Betroffenen - die Frage nach der Umsetzbarkeit von Verbalbeurteilungen im Schulalltag.

298 Vgl. etwa Micklos (1982) und einige Wertdifferenzen in den IfS-Umfragen, zuletzt: Kanders u.a. (2004, 28-31).
299 Jachmann (2000, 234, 241).
300 Valtin (2002c, 144).
301 Beutel (2005, 82).

## Rechtfertigt der Ertrag aufwändigere Formen der Erhebung und Bewertung von Leistungen?[302]

Soweit LehrerInnen Vorbehalte gegen die Einführung von Verbalbeurteilungen äußern, wird immer wieder der Zeitaufwand für Beobachtung und Dokumentation genannt[303]. Dieses Argument ist verständlich und problematisch zugleich. Ausführliche, differenzierte und sensible Berichte, wie sie beispielsweise Bambach (1994) publiziert hat, kosten mehr Zeit als die Berechnung eines Notendurchschnitts aus Klassenarbeiten. Allerdings sollten auch Ziffernnoten breiter fundiert sein als nur durch die Ergebnisse in drei, vier punktuellen Leistungsproben. Soweit also der kritisierte Aufwand die begleitende Lernbeobachtung und ihre Dokumentation betrifft, wäre dies ein Argument, das für beide Darstellungsformen gleichermaßen gilt. Sofern die Praxis diesem Anspruch - aus welchen Gründen auch immer - bisher nicht gerecht wird, würde diese Unzulänglichkeit in Verbalbeurteilungen lediglich sichtbar gemacht, das Problem würde aber nicht erst durch sie erzeugt.

In der Hamburger Studie »LeiHS« konnten einige konkrete Daten erhoben werden, die das Bild differenzieren[304]:

▲ GrundschullehrerInnen wenden mehr Zeit für Berichtszeugnisse (aber auch für Notenzeugnisse) auf als ihre KollegInnen in der Sekundarstufe.

▲ Im Durchschnitt wenden sie für ein Berichtszeugnis insgesamt 2,5-2,6 Stunden, für ein Notenzeugnis aber auch schon 2,1 Stunden auf.

▲ In beiden Fällen schwanken die Belastungen zwischen den Lehrkräften erheblich - nicht nur wegen der unterschiedlichen Anzahl an SchülerInnen.

▲ Es besteht wider Erwarten *kein* Zusammenhang zwischen der individuell aufgewandten Zeit und der Einstellung der Person zu Noten vs. Verbalgutachten.

Die Unterschiede sind also nicht so erheblich wie erwartet. Dennoch mahnt Oelkers (2001): »Die meisten Vorschläge, die die Lehrkräfte als ›Diagnostiker‹ (Jäger 2000) aufwerten und ihnen zusätzliche Aufgaben aufbürden, erhöhen nur den Aufwand, ohne die reale Zeitverteilung in Rechnung zu stellen. Nach den vorliegenden Schweizer Daten konzentriert sich die Jahresarbeitszeit der Lehrkräfte mit durchschnittlich zwischen 80 und 90 Prozent auf die unterrichtsbezogenen Tätigkeiten. Den verbleibenden Rest einer stark gestressten Zeit müssen sich Betreuung und Beratung, Weiterbildung oder Gemeinschaftsarbeit und alles Übrige teilen (Landert 1999). Es ist dann ziemlichgrotesk, Listen mit allerlei diagnostischen Tätigkeiten zu lesen, die ungewichtet sind und die zeitliche Belastungen unberührt lassen (Jäger 2000, S. 101). Das Grundproblem von Aufwand und Effekt ist nicht gelöst, zumal nicht in einem Berufsfeld, das vom individuellen Engagement lebt und sich in zeitlicher Hinsicht nicht standardisieren lässt.«[305]

Andererseits stellen Black/Wiliam (1998) in ihrem Forschungsüberblick zur Leistungsbeurteilung im Unterricht fest, dass SchülerInnen von undifferenzierten Noten nicht für ihre Arbeit und ihren Lernerfolg profitieren[306]. Was also ist ein zahlbarer Preis für eine verbesserte Rückmeldung?

Die Arbeitsbelastung von LehrerInnen wird gemeinhin in Zeit gemessen, analog ihre »Leistung« nach Stundendeputat bezahlt. Nimmt man dieses Kriterium als Maßstab, so ergibt sich für verschiedene Tätigkeiten der Leistungsbeurteilung ein unterschiedliches Gewicht, wie die Ergebnisse zweier Studien von Schönwälder (1999) in Bremen und Nordrhein-Westfalen zeigen. Unter 21 Tätigkeiten gab es vier Items aus dem Bereich der Leistungsbeurteilung. Sie fanden sich - geordnet nach zeitlicher Belastung - auf folgenden Plätzen[307]:

| Rang | Wert | Tätigkeit |
|---|---|---|
| 1 | 1,4 | Beurteilen durch Entwicklungsberichte |
| 2 | 1,6 | Unterricht |
| 3 | 1,6 | Planung und Auswertung von Unterricht |
| 4 | 1,8 | Korrigieren von Schülerarbeiten (ohne Benotung) |
| 5 | 1,8 | Zeugnisse geben |
| 6 | 1,9 | Schulveranstaltungen (Wandertage, Schul-/Klassenfeste ...) |
| | | ... |
| 11 | 2,1 | Kooperation mit KollegInnen |
| 12 | 2,1 | Benoten |
| 13 | 2,1 | Planung und Auswertung von Schulveranstaltungen |
| | | ... |
| 20 | 2,5 | Ausschüsse |
| 21 | 2,6 | Beaufsichtigung von SchülerInnen (Pausen, Hausaufgaben usw.) |

302 Vgl. vor allem Schönwälder (1999) und Oelkers (2001).

303 Vgl. Maier (2001, 117).

304 Vgl. Jachmann (2003, 128-137, 141-142).

305 Oelkers (2001, o.S.), der ergänzt: »Eine grosse Untersuchung zur Arbeitszeit der Lehrpersonen in der deutschsprachigen Schweiz (Landert 1999) zeigt unter anderem folgende Befunde:
- Lehrkräfte unterschätzen ihre Arbeitszeit eher als dass sie sie überschätzen.
- Alle Wochentage sind belastet, die Wochenendarbeit variiert nach Schultyp und Schulstufe.
- Die durchschnittliche Arbeitszeit liegt ferienbereinigt höher, als im öffentlichen Dienst verlangt: Zwischen 44,6 und 47,3 Wochenstunden je nach Pensengrösse, zwischen 44,4 und 47,8 Stunden bezogen auf die Schulstufen.
- Die Jahresarbeitszeit konzentriert sich auf das Hauptgeschäft, nämlich Unterrichten, Vor- und Nachbereitung sowie Planung und Auswertung.
- Für Betreuung und Beratung stehen 3% der durchschnittlichen Jahresarbeitszeit zur Verfügung.«

306 Vgl. auch Stiggins (1999, 194).

307 In der zweiten Spalte sind die Mittelwerte der zeitlichen Beanspruchung angegeben, wobei 1 »ganz erheblich« und 3 »geringfügig« bedeutete (vgl. Schönwälder 1999, 119-120).

Die Übersicht zeigt:

▲ Die Teiltätigkeiten der Leistungsbeurteilung gehören zu den aufwändigeren Aufgaben (fast alle in der oberen Hälfte, drei unter den ersten fünf);

▲ das Benoten ist ganz erheblich weniger aufwändig als das Schreiben von Entwicklungsberichten und dieses ist auch noch deutlich aufwändiger als das Korrigieren von Arbeiten oder das Schreiben von Zeugnissen.

Wer fordert, dass Ziffernnoten und -zeugnisse durch gehaltvolle Entwicklungsberichte ersetzt werden, muss LehrerInnen also zeigen, welchen Vorteil sie von dieser zusätzlichen Anforderung haben oder ihre zeitliche Mehrbelastung durch eine Gratifikation ausgleichen - sonst wird Anspruch der Reform zu oft unterlaufen, wie die ernüchternden Ergebnisse der Inhaltsanalysen von Verbalgutachten zeigen (vgl. > Kap. 3.1).

Vor diesem Hintergrund ist ein zweiter Aspekt der Schönwälder-Studie interessant. Die LehrerInnen wurden nämlich zusätzlich gebeten, dieselben Vorgaben nach dem Grad ihrer psychischen »Belastung« einzustufen. Dabei ergab sich folgendes Bild[308]:

| Rang | Wert | Tätigkeit |
|---|---|---|
| 1 (1) | 1,4 | Beurteilen durch Entwicklungsberichte |
| 2 (5) | 1,7 | Zeugnisse geben |
| 3 (6) | 1,8 | Schulveranstaltungen (Wandertage, Schul-/Klassenfeste ...) |
| 4 (7) | 1,8 | Klassenfahrten/Projekte |
| 5 (12) | 1,9 | Benoten |
| | | ... |
| 18 (18) | 2,4 | Fachkonferenzen |
| 19 (4) | 2,4 | Korrigieren von Schülerarbeiten (ohne Benotung) |
| 20 (19) | 2,6 | Fort- und Weiterbildung |
| 21 (11) | 2,6 | Kooperation mit KollegInnen |

Wie Schönwälder (a.a.O., 120) betont, gibt es zwar einen Zusammenhang zwischen zeitlichem Aufwand und psychischer Belastung (r = .58), aber wie die gemeinsame Varianz von nur etwa einem Drittel zeigt, geht letztere nicht in ersterer auf. Deren Besonderheit wird gerade in den Tätigkeiten der Leistungsbeurteilung deutlich. Das sehr zeitaufwändige Korrigieren wird nämlich psychisch als nur geringe Belastung empfunden, während alle Formen der *Bewertung* von Leistungen zu den fünf belastendsten Tätigkeiten zählen[309]. Unter ihnen rangieren die Entwicklungsberichte allerdings wieder eindeutig auf Platz 1, während Noten und Zeugnisse als etwas weniger belastend eingeschätzt werden.

Die beiden Vergleiche machen deutlich, dass entwicklungsorientierte Lernberichte nicht nur einen hohen zeitlichen Aufwand erfordern, sondern von den LehrerInnen auch als besondere Belastung empfunden werden. Einer Person sprachlich differenziert gerecht zu werden, sich nicht hinter einer (scheinbaren) Verrechnung von Daten verstecken zu können, ist offensichtlich eine hohe Anforderung. So berechtigt diese Anforderung nach dem oben Gesagten auch sein mag - ohne entsprechende Ausbildung und Unterstützung (Fortbildung, kollegialer Austausch, Supervision) ist sie in der Breite wohl nicht erfüllbar (vgl. > Kap. 3.1). Außerdem brauchen LehrerInnen alltagstaugliche Verfahren und Aufgaben, die es ihnen ermöglichen, Beobachtung und Förderung im Unterricht enger miteinander zu verknüpfen[310].

Dass der *zeitliche* Mehraufwand sich *psychisch* auf Dauer sogar als Entlastung auswirken kann, zeigen die Ergebnisse der Befragung von 81 LehrerInnen im NRW-Schulversuch mit notenfreien Beurteilungen in Klasse 3 und 4: »Das Statement mit der höchsten Zustimmungsquote dieses Bereichs bezieht sich auf den Arbeitsaufwand der Lehrkräfte. So geben 88% der Befragten an, daß bedingt durch die neue Beurteilungsform der zeitliche Aufwand im Vergleich zu früher erheblich gestiegen ist. ... Mit fast ebenso großer Mehrheit (zu 84%) finden die betroffenen LehrerInnen und Lehrer aber gleichzeitig, dass die pädagogischen Vorteile ohne Noten den Mehraufwand voll und ganz rechtfertigen. Bei 66% (und bei weiteren 18% noch teilweise) ist trotz der beträchtlichen Mehrbelastung die Berufszufriedenheit sogar größer als früher. Dass diese positiven Wirkungen auf die Lehrkräfte nicht gering veranschlagt werden dürfen, zeigt auch der Befund, dass die weitaus meisten der Befragten (80%) meinen, es eigentlich nicht mehr verantworten zu können, noch einmal Zensurenzeugnisse zu schreiben.«[311]

Interessant sind auch die Gründe, die von den LehrerInnen für diese Einschätzungen genannt werden: ein besseres Verständnis der Lernentwicklung der Kinder - und ein größeres Interesse am einzelnen Kind.

Beachtung verdient aber auch der oben genannte Befund, dass alle Formen der Leistungsbeurteilung als besonders belastend wahrgenommen werden, wenn sie der Selektion dienen. Hier wird ein grundsätzlicheres Problem deutlich, das über technische Fragen der Darstellungsform hinausweist (> Kap. 7).

---

308 Vgl. Schönwälder (1999, 115-116): 1 = »sehr belastend« vs. 3 = »kaum belastend«; zum Vergleich sind in der folgenden Tabelle in Klammern die Rangplätze des zeitlichen Aufwands aus der ersten Tabelle mit aufgenommen.

309 Schmude u.a. (2003) berichten aus einer Befragung von Lehramtsstudierenden zu Problemen, die sie im Beruf erwarteten, ebenfalls, dass der Komplex Bewertung/Benotung/Selektionsentscheidungen am häufigsten genannt wurde. Bei der im Rahmen des NOVARA-Projekts durchgeführten Lehrerbefragung kam von Lehrkräften, die überzeugte Vertreter verbaler Beurteilungen waren, auch die Anmerkung, dass für sie die Reduktion ihrer Rückmeldung auf eine Note sehr belastend ist (pers. Mitteilung von Schmude am 20.3.06).

310 Vgl. dazu die umfangreichen Hilfen für die ersten beiden Schuljahre in Bartnitzky u.a. (2005) und für Klasse 3/4 in Bartnitzky u.a. (2006, in Vorb.).

311 Haenisch (1996b, 21-22).

## 6

## Zwischenbilanz und pädagogische Folgerungen[312]

Es gibt kein *Verfahren, Leistungen zu erheben,* das valide, objektiv und verlässlich genug wäre, um Einzelfallentscheidungen über Bildungskarrieren zu rechtfertigen. Sowohl punktuelle Tests als auch Lehrerurteile sind fehleranfällig, methodische Verbesserungen nur begrenzt möglich (vgl. > Kap. 1).

Als Maßstab reicht für sich genommen - keine der drei *Bezugsnormen* - aus, um die Informationsbedürfnisse der verschiedenen Zielgruppen zu befriedigen. Leistungsbeurteilungen haben zu unterschiedliche Funktionen zu erfüllen, als dass eine Form allein genügen könnte (vgl. > Kap. 2).

Die *Wirkungen* sowohl von Ziffernnoten wie auch von Verbalzeugnissen sind aus pädagogischer Sicht kritisch einzuschätzen. Dies hängt (vor allem bei Ziffernnoten) mit den oben genannten methodischen Mängeln zusammen, andererseits (vor allem bei den Verbalgutachten) mit Unzulänglichkeiten ihrer Umsetzung im Schulalltag (vgl. > Kap. 3). Ein besonderes Problem stellt für beide die starke Selektionsorientierung der Schule dar (vgl. > Kap. 7).

Dennoch gibt es hohe Erwartungen an Leistungsbeurteilungen und finden speziell Ziffernnoten immer noch eine breite *Akzeptanz* bei SchülerInnen, Eltern, LehrerInnen, Arbeitgebern und in der Öffentlichkeit generell (vgl. > Kap. 4).

Als Wege aus diesem Dilemma werden verschiedene Vorschläge gemacht, die wir im Folgenden wenigstens kurz kommentieren wollen:

- ein grundsätzlicher, also *genereller* Verzicht auf Leistungsbeurteilungen
- Verzicht auf eine Zertifizierung von schulischen Leistungen *nach außen*
- Ersatz von *Ziffernnoten* durch andere Formate.

### 6.1

### Grundlegende Einwände

Schulkritiker wie die Gruppe der Kinderrechtszänker in Berlin stellen Beurteilungen grundsätzlich in Frage: »Bewertung setzt Kontrolle voraus, die in jedem Fall ein Eingriff in die Privatsphäre des Schülers ist. Ob, wann und in welcher Form bewertet werden soll, soll deshalb nur jeder Schüler selbst entscheiden. Zeugnisse und Bewertungsdokumente jeder Art ignorieren, daß Menschen sich auch nach Erhalt ihres letzten Zeugnisses ändern. Dies trägt dazu bei, daß andere sich Vor- bzw. Fehlurteile anhand des vor Jahren festgeschriebenen Erscheinungsbildes in der Schule bilden, welches wiederum keineswegs objektiv ist. Dieser Umstand kann zu einer lebenslangen Brandmarkung als beispielsweise ›leistungsunfähig‹ führen. Schon aus Gründen des Datenschutzes, der informationellen Selbstbestimmung, sind wir gegen Zeugnisse.«[313]

Dies sind ernst zu nehmende Einwände. Aber es sind unterschiedliche Folgerungen denkbar - die jeweils ihre eigenen Probleme aufwerfen.

#### 6.1.1

#### Genereller Verzicht auf eine Rückmeldung zu Leistungen?

Auftrag der Schule ist Förderung. Aber Kinder können sich nicht entwickeln, ohne dass sie eine Rückmeldung zu ihren Leistungen erhalten[314]. Jeder Mensch, vor allem das Kind, braucht Fremdeinschätzungen, um eine Selbsteinschätzung zu gewinnen. Für Lernende ist wichtig, was Lehrende über ihre Leistung denken. LehrerInnen müssen zu ihrem Urteil stehen - nicht weil es objektiv ist, sondern weil das Kind ein Recht hat zu wissen, was Erwachsene denken. Allerdings ist ein Problem unserer Schule, dass diese Erwachsenen ihren SchülerInnen zugewiesen werden und nicht selbst gewählt werden können. Ein zweites Problem stellt die Betonung der *Bewertungs*funktion von Beurteilungen dar. Würden Rückmeldungen sich darauf beschränken, den Leistungsstand zu *beschreiben,* und zwar mit Bezug auf gemeinsam abgesprochene Ziele und mit dem Fokus auf die Entwicklung der Lernenden, könnten sie leichter angenommen werden.

Damit stellt sich die Frage anders: *Wie* geben wir dem Kind eine Rückmeldung, die sein Lernen fördert? Rückmeldungen zu Leistungen sind lernförderlich, wenn sie sachbezogen erfolgen, d.h. konkrete Stärken und Schwächen benennen und vor allem Hinweise für konkrete Lern- und Fördermöglichkeiten geben. Im Sinne der Forderung von Hartmut v.Hentig (1985) »Die Menschen stärken, die Sachen klären« setzen sie aber auch Sensibilität für mögliche Nebenwirkungen auf die Betroffenen voraus. Ohne Anerkennung der individuellen Fortschritte und ohne Hilfen für die Überwindung von Schwächen sind die - notwendigen - Hinweise auf Fehler kontraproduktiv.

Zwei Relativierungen sind darüber hinaus wichtig: Die Wertschätzung eines Kindes darf in der Schule nicht nur von

312 Mit dem Wechsel zu »Folgerungen« und »Bewertungen« in den folgenden beiden Kapiteln wächst das interpretative Moment in unserer Darstellung. Schon bei der Darstellung der empirischen Befunde sind Gewichtungen und Ausdeutungen durch die AutorInnen unvermeidlich. Im Folgenden nimmt der Einfluss unserer theoretischen Annahmen und Wertvorstellungen auf die Darstellung zu, auch wenn wir uns um eine sachbezogene Argumentation auf der Basis der vorgetragenen empirischen Befunde bemüht haben.

313 Kinderrechtszänker (o.J., 7. Punkt).

314 Vgl. die Metaanalyse von über 600 Studien bei Kluger/de Nisi (1996) zum Lernen mit und ohne Feedback. Im Durchschnitt lag die Effektstärke zugunsten eines Feedbacks bei d = .41. Allerdings streuten die Wirkungen erheblich, und mehr als ein Drittel der Studien zeigte sogar negative Effekte. Insofern scheint die Art des Feedbacks hoch bedeutsam. Wie auch eine Sekundäranalyse von Bangert-Drowns u.a. (1991) bestätigt, ist ein sachbezogenes Feedback, z.B. die Rückmeldung des richtigen Ergebnisses, besonders wichtig, aber auch die persönliche Beziehung kann eine entscheidende Rolle spielen.

seinen Leistungen in einem fachlichen Ausschnitt abhängen. Auch andere Leistungen, beispielsweise seine soziale und emotionale Kompetenz, müssen Aufmerksamkeit und Anerkennung finden. Dies muss ein Kind spüren. Und es darf nicht von *einem* Urteil allein abhängig sein. Lernende müssen sich an verschiedenen »Meistern« orientieren können, um selbstständig zu werden.

### 6.1.2

### Verzicht auf eine Zertifizierung nach außen?

Von denen, die Leistungsbeurteilungen in den Schulen ganz abschaffen wollen[315], ist aber noch eine zweite Frage zu beantworten: Wie verschaffen wir SchülerInnen einen Ausweis ihrer Leistungen, der außerhalb der Schule zählt?

Verzichtet die Schule auf Abschlusszertifikate oder werden Beurteilungen von den AbnehmerInnen nicht ernst genommen, führen diese Eingangsprüfungen durch[316]. Oder es werden statt der Beurteilungen innerhalb der *Klasse* externe Prüfungen auf der zentralen Systemebene eingeführt.

Diese haben ihre eigenen Probleme: Wenn LehrerInnen nicht beurteilen, beurteilen Fremde. Wenn der Markt allein entscheidet, kommen Kriterien stärker zur Geltung, die nicht pädagogisch zu begründen sind. Eine Trennung von Unterricht und Prüfung[317] eröffnet somit Vor- und Nachteile. Sie entlastet das Verhältnis zwischen PädagogInnen und SchülerInnen, sie belastet aber den Unterricht mit Fremdkriterien. Für die Schüler könnte der kontinuierliche Bewährungsdruck entfallen, andererseits dürften sie unter der Bedeutung der punktuellen Prüfungssituation leiden. Vor allem aber stellt sich die Frage, ob die Vorbereitung auf die externe Prüfung den Unterricht nicht stärker prägt als informelle Leistungsproben, die direkt aus konkreten Unterrichtseinheiten erwachsen.

Nur: Soll dies in Form von Noten geschehen? Noten sind eine spezifische Form der Leistungsbeurteilung. Diese ist abgeleitet aus ihrer Funktion bei der Vergabe von Zeugnissen.

### 6.1.3

### Verzicht auf Ziffernnoten als Form der Beurteilung?

Bleibt also die Frage, ob Beurteilungen weiterhin in Form von Noten erfolgen sollen,

- deren Grundlage *informelle* Leistungsnachweise sind,
- die mit Bezug auf den *Rang* in der Klasse bewertet und
- die in Form von *Ziffern* dargestellt werden und
- die *Selektionsentscheidungen* begründen.

Die in diesem Gutachten referierten Studien stellen Noten unter allen vier Gesichtspunkten in Frage: Die Bewertung nach informellen Proben und Beobachtungen ist in hohem Maße fehleranfällig, die soziale Bezugsnorm hat negative Auswirkungen auf die Lernmotivation, und vor allem sind Ziffern wenig aussagekräftig, suggerieren vielmehr eine Genauigkeit, Vergleichbarkeit und Prognosefähigkeit, die sie faktisch nicht gewährleisten können. Schließlich schränkt die Reduktion differenzierter Fähigkeitsprofile und unterschiedlicher Gründe für ein und dieselbe Leistung (als Produkt) die Aussagekraft von Noten zusätzlich ein.

Damit stellt sich die Frage nach den Alternativen, zumal auch Verbalgutachten die in sie gesetzten Erwartungen bisher nicht erfüllt haben.

## 6.2

## Keine Beurteilungsform erfüllt alle Anforderungen - einfache Auswege aus dem Bewertungsdilemma gibt es nicht

Leistungsbeurteilungen haben unterschiedliche Funktionen zu erfüllen. Je nachdem, ob die Förder-, Berichts- oder Selektionsfunktion im Vordergrund steht, und je nach den Adressaten sind verschiedene Formen angemessen.

Zumindest in der Grundschule kann (und sollte) schon heute auf Noten verzichtet werden. Vorrang hat eine möglichst differenzierte Rückmeldung der individuellen Leistung und ihrer Entwicklung. Verbalbeurteilungen dürfen dabei aber nicht bloße Übersetzungen der Ziffernnoten sein, da sie dann eine überflüssige Zusatzbelastung sind. Sie müssen vielmehr durch die Differenzierung von Teilleistungen und durch den Bezug der Bewertung auf die individuelle Entwicklung die Notenbewertung inhaltlich ergänzen.

Die Risiken von Fehlbeurteilungen und die negativen Nebenwirkungen einzelner Formen der Leistungsbeurteilung lassen sich nur reduzieren, nicht gänzlich aufheben. Folgende Maßnahmen sind möglich, wobei ihre Umsetzbarkeit im Schulalltag im Blick zu behalten ist:

▲ Bei der Erhebung von Leistungen sind *verschiedene Verfahren* wie standardisierte Tests, klassenbezogen gestellte oder individuell gewählte Aufgaben sowie informelle Beobachtungen zu kombinieren.
▲ Bei der Rückmeldung sind Leistungen mit Bezug auf *verschiedene Maßstäbe* auszuweisen
- als individuelle Fortschritte gegenüber früheren Leistungen,
- als Grad der Lernzielannäherung und
- als Rangplatz in einer repräsentativen Bezugsgruppe.

▲ In den Beurteilungsprozess sind *verschiedene BewerterInnen* einzubeziehen, einschließlich der Selbsteinschätzung durch die Betroffenen, und eventuelle Differenzen ihrer Urteile explizit auszuweisen.

Diese drei Punkte erfordern noch einige Erläuterungen, die im Folgenden kurz ausgeführt werden.

---

315 Kinderrechtszänker (o.J., 7. Punkt).

316 Tillmann (1997/99) stellt allerdings in Frage, dass die Schule ihre traditionelle »Allokationsfunktion« werde beibehalten können, da sich das Beschäftigungssystem zunehmend vom Bildungssystem abkoppele: Gute Abschlüsse sichern keine Ausbildungsplätze oder gar Berufskarrieren mehr, das Überangebot an Arbeitskräften entwertet schulische Zertifikate zunehmend.

317 S. England, s. Baden-Württemberg und Bayern.

## 6.3

### Daten aus verschiedenen Erhebungsverfahren sind miteinander zu verbinden

Wo immer möglich - und in bei der Nutzung für Selektionsentscheidungen zwingend - sind Leistungsdaten
- ▲ zu mehreren Zeitpunkten
- ▲ anhand verschiedener Aufgaben
- ▲ in unterschiedlichen Situationen

zu erheben.

Schon Christiani/Heller (1981) haben vorgeschlagen,
- ▲ Klassenarbeiten und andere Leistungsproben,
- ▲ prozessbegleitende Lernbeobachtungen und
- ▲ Ergebnisse standardisierter Tests

jeweils zu etwa einem Drittel in die Leistungsbeurteilung einzubeziehen.

Die stärkere Einbeziehung standardisierter *Tests* ist ein plausibler Vorschlag, um die Schwierigkeiten personabhängiger Beobachtungen zu verringern. Im Anschluss an die internationalen Leistungsvergleiche werden beispielsweise landesweite Lernstandserhebungen propagiert. Als Vorteile werden genannt[318]:
- ▲ Vorhersehbarkeit der Anforderungen für die SchülerInnen, wenn der Inhalt von Leistungsproben klar definiert ist (allerdings mit der Gefahr eines bloßen »learning for the test«);
- ▲ Berechenbarkeit der Bewertung unabhängig von personabhängigen Einschätzungen;
- ▲ Einstufung der Leistungen mit Bezug auf repräsentative (Teil-)Stichproben, so dass der Rangplatz eines Schülers unabhängig wird vom Leistungsniveau seiner Klasse.

Über diese Erweiterung der Vergleichsperspektive hinaus eröffnen Tests die Möglichkeit,
- ▲ durch die Zuordnung konkreter Leistungen zu Kompetenzstufen deren Annäherung an gesetzte Lernziele zu bestimmen (Kriteriumsorientierung) und
- ▲ bei Wiederholung der Tests individuelle Lernfortschritte auszuweisen (Entwicklungsorientierung).

Allerdings besteht die Gefahr, die Leistungsfähigkeit von Tests zu überschätzen[319]. Sie erfassen die Leistung nur punktuell, ausschnitthaft und je nach Aufgabe in einer sehr spezifischen Form. Hier hat das zentrale »Institut für Qualität im Bildungswesen« (IQB) in Berlin einige Entwicklungsarbeit zu leisten, um auch nur das Niveau zu erreichen, auf dem etwa das niederländischen CITO-Institut Schulen Instrumente zur Evaluation von Unterricht anbietet (nicht vorschreibt!). Das setzt voraus, dass Fachdidaktik und Unterrichtspraxis gewichtig an der Entwicklung von Aufgaben beteiligt werden. Selbst dann bleibt zu bedenken, dass Tests nur bestimmte Leistungstypen erfassen können und dass ihre Ergebnisse generell interpretationsbedürftig sind: Zahlen sprechen nicht für sich. Insofern können Tests informelle Beobachtungen und vor allem das Lehrerurteil selbst nicht ersetzen. Auch die Daten aus Tests bedürfen der Deutung und Ergänzung.

*Verbale Beurteilungen,* die neben Testwerten weitere Daten einbeziehen und im Zusammenhang interpretieren, können einige der genannten Schwierigkeiten auffangen[320]:
- durch kontinuierliche Sammlung von Daten statt nur punktueller Erhebungen;
- durch Berücksichtigung des jeweiligen Leistungskontexts;
- durch eine situationsbezogene Verständigung über die Bedeutung von Fragen und Antworten (»Was hast du dir dabei gedacht?«);
- durch die Kombination unterschiedlicher Leistungsformen (mündlich vs. schriftlich) und Leistungssituationen (mit/ohne Zeitdruck; selbst gewählte vs. fremd bestimmte Aufgaben; individuelle vs. Gruppen-Arbeit).

Wichtig ist dabei die mehrfach eingeforderte Entwicklung von Schreibstandards[321] und das Gespräch mit SchülerInnen und Eltern über Kriterien und Form der Berichte: »Die Schülerinnen und Schüler - und nicht zuletzt deren Eltern - müssen über die Kriterien der Lernberichte, die Schreibstandards und die Sprachregelungen sowie die Gewichtung der Dimensionen aufgeklärt sein. Sie müssen die Sprache der Lernberichte eindeutig entschlüsseln könne, sollen diese den Anspruch der fördernden Evaluation des Lernprozesses gegenüber dem Lernenden Genüge tun.«[322]

Was die Validität, Objektivität und Reliabilität der Wahrnehmungen und Bewertungen durch eine Lehrperson betrifft unterliegen die beiläufigen Beobachtungen, anders als Tests. Zudem scheint die Versuchung groß, ausformulierte Beurteilungen mit Hilfe von Textbausteinen als bloß verbale Übersetzung von Noten zu gestalten[323].

Beurteilungen müssen deshalb in Gespräche eingebettet werden[324], die Rückfragen erlauben und Gegenperspektiven

---

318 Vgl. Bremerich-Vos u.a. (2005). In dieser Nutzung besteht auch eine sinnvolle Funktion zentraler Lernstandserhebungen - jedenfalls wenn sie LehrerInnen Optionen für die Auswahl von Instrumenten eröffnen, die für ihren Unterricht aufschlussreich sind (vgl. Brügelmann 2003b).

319 Vgl. > Kap. 1 und zur Fehleranfälligkeit des Test»geschäfts« (bezogen auf die langjährigen Erfahrungen in den USA): Rhoades/Madaus (2003).

320 Vgl. dazu die konkreten Hilfen vor allem in Bambach u.a. (1996); Vierlinger (1999); Winter (2004); Bartnitzky u.a. (2005; 2006); Becker u.a. (2006); Brunner u.a. (2006), speziell für die Schriftsprache bei Brinkmann/Brügelmann (1993); Naegele/Valtin (2003); Dehn/Hüttis-Graff (2006); für Mathematik bei Hengarter (1999); Sundermann/Selter (2005; 2006); für die Sekundarstufe bei Beutel/Vollstädt (2000); Winter u.a. (2002); Paradies u.a. (2005).

321 Vgl. Beutel (2005, 41, 110-115).

322 Lübke (1996, 217 f.).

323 Vgl. etwa www.schulbericht.de, aber auch die verbreiteten Konkordanzlisten für Arbeitszeugnisse in der freien Wirtschaft, z.B. bei Weuster/Scheer (2005).

324 Vgl. dazu Gramsch/Krause-Hotopp (2003).

einbeziehen. Dies erfordert auch einen Wechsel von hierarchischen zu dialogischen Formen der Leistungsbewertung (> Kap. 6.4).

**6.4**

### Bewertungen müssen auf unterschiedliche Bezugsnormen bezogen werden

Die verschiedenen Bezugsnormen haben alle ihre Berechtigung - jeweils für spezifische Funktionen[325]. Wenn es darum geht, Lernende zu motivieren[326] und konkrete Vorschläge für ihre Förderung zu entwickeln, müssen allerdings die individuelle und lernzielorientierte Norm im Vordergrund stehen.

Für die Rückmeldung des Lernerfolgs ist es deshalb wichtig, Fortschritte - bezogen auf die individuellen Voraussetzungen - zu dokumentieren (Entwicklungsnorm). Ergänzend sollte der Leistungsstand als Grad der Annäherung an die Lernziele ausgewiesen werden. Leistungsstandards[327] haben dafür eine wichtige Funktion: nicht um *vor*zuschreiben, welche Leistungen von allen SchülerInnen zu einem bestimmten Zeitpunkt zu erbringen sind. Das ist eine illusionäre Erwartung[328]. Hilfreich können sie aber sein, um das Niveau der individuell erworbenen Kenntnisse und Fähigkeiten so zu *be*schreiben, dass Lernfortschritt und Lernanforderungen gleichermaßen sichtbar werden.

Für die Zulassung zu einem Lernangebot oder einer Berufstätigkeit reicht es in der Regel aus zu prüfen, ob BewerberInnen die Anforderungen der aufnehmenden Einrichtung erfüllen. Auch für die Schulfächer lassen sich solche Minimalanforderungen (evtl. in gestufter Form) formulieren. Erbracht werden die entsprechenden Leistungsnachweise von einzelnen SchülerInnen, wenn sie sich sicher fühlen. »Rechenpass«, »Rechtschreibausweis« oder »Computerführerschein« sind Instrumente, um solche Basiskompetenzen auszuweisen.

Muss aus einer zu großen Zahl von KandidatInnen ausgewählt werden, lässt sich durch Vergleich ihrer unterschiedlichen Leistungsniveaus entscheiden. Allerdings kann es auch hier bedeutsam sein, wo jemand mit dem erreichten Niveau im Verhältnis zur Gesamtgruppe steht, so dass ein Ausweis des Prozentrangs bestimmter Kompetenzniveaus (Rangplatz auf einer Skala von 1 bis 100) ergänzend hilfreich wäre.

Prozentränge aus standardisierten Tests dürfen allerdings nicht unkommentiert stehen bleiben. Ähnlich wie Noten könnten sie eine Verbindlichkeit der Einordnung suggerieren, die bei ihrer inhaltlichen und zeitlichen Ausschnitthaftigkeit nicht gerechtfertigt wäre. Auch diese Zahlenwerte bzw. Abweichungen von anderen Einschätzungen müssen also erläutert werden.

Die Gruppennorm sollte im Übrigen eher indirekt wirksam werden: Bei der Bestimmung von Mindestanforderungen für eine Tätigkeit (z.B. Führerschein) wird neben normativen Setzungen eine Rolle spielen, wie viele Personen erfahrungsgemäß welches Leistungsniveau erreichen.

**6.5**

### In dialogischer Form sollten Fremd- durch Selbsteinschätzungen ergänzt werden

Ernest House[329] hat bereits vor 25 Jahren die Fixiertheit der Evaluationsdiskussion auf die klassischen Gütekriterien der Validität, Objektivität und Reliabilität als unzulässige Verkürzung der Anforderungen an eine angemessene Dokumentation und Bewertung von pädagogischen Programmen und Aktivitäten beklagt. Er sieht Evaluation als einen sozialen Prozess, in dem es auch um Macht und um Gerechtigkeit geht. Glaubwürdigkeit, Unparteilichkeit, Fairness sind deshalb für ihn Standards, denen Evaluationen gerecht werden müssen. Es geht um mehr als um technische Präzision. Die Möglichkeiten der Erkenntnis im pädagogischen Feld beschränken sich auf Wahrscheinlichkeiten; ihr Ergebnis sind persönliche Entscheidungen, nicht logische Folgerungen.

Dies gilt besonders für die Beurteilung individueller Leistungen. Die Subjektivität auch des professionellen Urteils lässt sich nicht vermeiden. Im Sinn einer wohlwollenden Empathie ist sie sogar erforderlich, wenn die Rückmeldung zu Leistungen auf deren Förderung zielt. Anders bei Bewertungen, die eine Ausweisfunktion haben. Hier ist eine Kontrolle persönlicher Wahrnehmungen und Deutungen zwingend geboten. Tests können zwar auch dort persönliche Einschätzungen nicht ersetzen, sie sollten aber - als zu *interpretierende* Daten - in diese einbezogen werden. Die auch hierbei nicht vermeidbare Subjektivität lässt sich aber in Verbalgutachten - anders als in Ziffernoten - transparent machen. Und sie kann *sozial* kontrolliert werden, indem mindestens das für Prüfungen übliche »Vier-Augen-Prinzip« angewandt wird[330].

Diese Vorschläge - und ihre Diskussion in der Literatur - nehmen eine Voraussetzung als gegeben hin: SchülerInnen werden durch LehrerInnen beurteilt. Diese Annahme ist aber nicht selbstverständlich[331]. Zudem hat sich der Ansatz einer Beurteilung »von oben« als grundsätzlich problematisch herausgestellt.

---

325 Vgl. oben > Kap. 2 und Klauer (1987).

326 Vgl. Rheinberg (2001, 69) und > Kap. 2.3.

327 Vgl. zur Notwendigkeit, verschiedene Funktionen von Standards genau zu trennen: Klieme u.a. (2003, 81ff.).

328 Vgl. Brügelmann (2005a, Kap. 46-49).

329 Vgl. House (1980, 65 ff.) und neuerdings wieder Winter (2004, 91-95).

330 Allerdings darf nicht übersehen werden, dass eine gemeinsame Beratung von Noten, wie bei mündlichen Prüfungen, eigene gruppendynamische Effekte entfalten kann, in denen sich z.B. der Status einzelner Beteiligter gegen fachliche Kriterien durchsetzen kann (vgl. dazu u.a. die ethnografische Studie von Kalthoff 1996, 118-120). Auch in unserer eigenen Auswertung des Notenniveaus verschiedener Prüfer-Teams zeigte sich diese wechselseitige Beeinflussung durch KollegInnen (vgl. Brügelmann 2000b).

331 Vgl. grundsätzlich Winter (1991) und Wehr (1992) mit vielen konkreten Vorschlägen für die Grundschule.

So sind dialogische Formen wichtig, um den Kindern bei der Entwicklung von sachangemessenen Kriterien zu helfen. Morys (2006) stellte in ihrer Studie Leistungsselbstsicht von Grundschulkindern überraschend fest, dass die befragten Kinder sich in ihrer Selbsteinschätzung einerseits überwiegend an Leistungsrückmeldungen aus der Familie orientierten und andererseits keinen Bezug auf eine differenziertere Rückmeldung von LehrerInnen nahmen (vielleicht nicht nehmen konnten, weil es diese nicht gab?)[332].

Außerdem zeigen Studien zur Wirkung von fremd bestimmten Belohnungen auf Motivation und Leistung von SchülerInnen und anderen Personengruppen höchst negative Effekte. Deutlich wird das in einem Versuch von Grolnick/ Ryan (1987)[333]: Eine Gruppe arbeitete in der Vorstellung, dass anschließend eine externe Leistungskontrolle[334] stattfinden werde, die anderen erhielten die Informationen, die Prüfungen dienten nur zur eigenen Rückmeldung und hätten keinen Einfluss auf die Benotung. Die Gruppe mit Leistungsdruck unterschied sich von der zweiten Gruppe in mehrfacher Hinsicht, und zwar durch

- ▲ Bekundung von weniger Interesse bei der Arbeit,
- ▲ geringere Einschätzung ihrer fachlichen Kompetenz,
- ▲ größere Angst,
- ▲ schlechtere Leistung in drei von fünf Zwischenprüfungen,
- ▲ schlechtere Leistung in der Schlussprüfung.

Für Aufsätze hat schon Merkelbach (1986) ein Verfahren dialogischer Bewertung vorgeschlagen. Gegen Bedenken, Kinder könnten ihre Leistungen nicht beurteilen, sprechen die Daten aus der Studie von Beutel (2004), in der sie fast 150 Kinder zu ihren Zeugnissen befragt hat. Ihr Fazit: »Kinder sind ›Experten‹ ihres Lernens. Sie haben, das erweist die Studie, ein höheres Bewusstsein von ihren Lernfortschritten und -defiziten, als dies die Pädagogik im Schulalltag üblicherweise unterstellt.«[335]

Dieser Befund passt zu den Erfahrungen der Kindheitsforschung der letzten zehn Jahre, die Kinder zunehmend als ExpertInnen ihrer eigenen Lebenswelt wahrnimmt und in die Untersuchungen einbezieht[336]. Parallel dazu wurden Konzepte eines Unterrichts entwickelt, der Kindern als SchülerInnen mehr Verantwortung zugesteht[337] und sie damit auch in die Beurteilung ihrer Arbeiten und Leistungsentwicklung einbezieht[338].

Aber werden Kinder durch eine solche Verantwortung nicht überfordert - zumindest im Grundschulalter? Morys (2006, 331) empfiehlt, die Aufgabe, sich selbst realistisch einzuschätzen, an konkrete Leistungssituationen anzubinden, also die erforderlichen Fähigkeiten mit unmittelbarem »Werkbezug«[339] zu entwickeln. In ihren Interviews mit über 70 Zweit- bis Viertklässlern stellt sie zwar bei leistungsschwachen SchülerInnen fest, dass diese sich oft überschätzen. Das gilt aber nur bei allgemeinen Urteilen[340], während sie bei konkreten Leseaufgaben gelungene Stellen und Schwierigkeiten angemessen wahrnehmen und bewerten können[341].

Inzwischen liegen vielfältige Ideen, Hilfen und Erfahrungen für eine Stärkung der Selbsteinschätzung im Unterricht vor[342]. Zu ihnen zählen selbst erstellte und von der Lehrperson kommentierte Portfolios eigener Arbeiten, Lerntagebücher, Kriterienraster für die Selbstbewertung in einzelnen Fächern usw.[343]. Winter[344] hat in verschiedenen Publikationen Prinzipien und konkrete Alternativen der Leistungsdokumentation vorgestellt, die eine »neue Lernkultur« befördern können. Im Zentrum steht für ihn das Portfolio, das in sehr unterschiedlicher Weise ausgestaltet sein und damit auch verschiedene Funktionen erfüllen kann. Folgende Erträge sieht er als möglichen Gewinn[345]:

- Das Portfolio hilft, die Ziele zu klären und Kriterien zu formulieren
- Portfolioarbeit öffnet den Kreis der Leistungsnachweise
- Fähigkeiten zur Reflexion und Bewertung werden gefördert
- Portfolioarbeit führt zu vielen inhaltlichen Aussagen und Rückmeldungen
- Das Portfolio ist ein aussagekräftiges Leistungsdokument
- Portfolios schaffen Voraussetzungen für öffentliche Leistungswahrnehmung und demokratische Rechenschaftslegung
- Portfolios ermöglichen veränderte Prüfungen und Aufnahmeverfahren.

Die Übersicht macht deutlich, dass die Funktionsüberlast von Noten entzerrt werden kann, wenn man verschiedene Formen der Beschreibung und Bewertung von Leistungen nutzt.

Reich (2003 ff.) hat die Forderung nach einer Kombination von verschiedenen Elementen in sein Konzept für eine systemische Leistungsbeurteilung aufgenommen und im folgenden Schaubild übersichtlich zusammengefasst:

---

332 A.a.O., 310, 329.

333 Zusammengefasst nach Deci/Ryan (1993, 234); vgl. ergänzend > Kap. 3.2.3.1 und zusammenfassend zur Selbstbestimmungstheorie der Motivation den Überblick über verschiedene Studien bei Deci/ Ryan (1993) und Ryan/Deci (2000).

334 Die Ergebnisse von fünf Zwischenprüfungen dienten zur offiziellen Benotung eines wichtigen Schulfachs.

335 Beutel (2004, 231); s.a. Beutel/Vollstädt (2002).

336 Vgl. u.a. Zinnecker (1995); Honig u.a. (1996); Heinzel (2000); Panagiotopoulou/Brügelmann (2003).

337 Vgl. u.a. die Zusammenfassung bei Peschel (2002a+b).

338 Vgl. u.a. Winter (1991); Konrad (1997); Arnold (1999).

339 Mit Bezug auch auf Graf (2004), die in ihrer Studie ebenfalls eine »werkorientierte Selbsteinschätzung« fordert (a.a.O., 314).

340 Zum Beispiel zu der Frage »Wie gut kannst du lesen?« (a.a.O., 304-305, 311, 332-333).

341 A.a.O., 306-307.

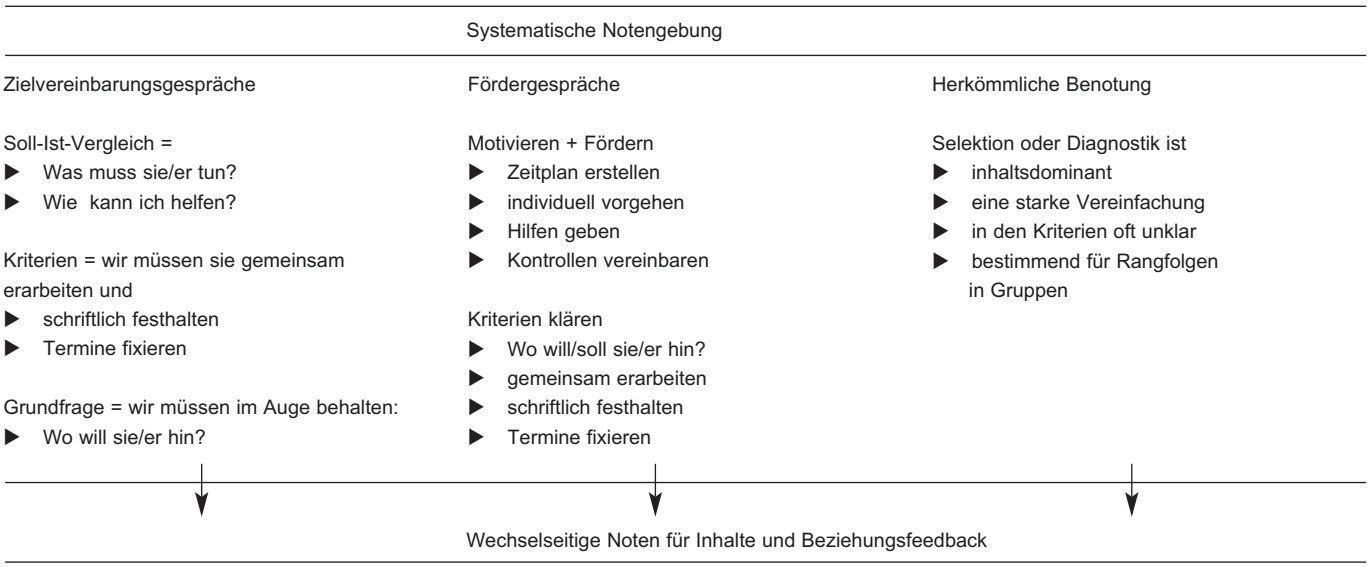
342 Vgl. u.a. Bambach u.a. (1996); Winter u.a. (2002); Bartnitzky (2004); Bartnitzky u.a. (2005; 2006) und zu einer positiven Evaluation in der Schweiz schon Iten/Theiler (1993).

343 Vgl. etwa die Beiträge zu Bartnitzky/Speck-Hamdan (2004).

344 Winter u.a., (2002); Winter (2004); Brunner u.a. (2006).

345 Winter (2006).

Abb. 8 Die Betonung in der konstruktivistischen Didaktik orientiert sich vor diesem Hintergrund an einer systematischen Notengebung, die vor allem aus drei Bereichen besteht:




In der kanadischen Provinz Ontario wird ein solches System bereits seit längerem verwirklicht[346] (Seiten des Erziehungsministeriums in Ontario, Kanada):

▲ Zeugnisse gibt es dreimal im Jahr.

▲ Sie dienen als Gesprächsgrundlage für Lehrer, Eltern, Schüler.

▲ Ergänzend hierzu sollen Aufzeichnungen über Gespräche und Telefonate mit Eltern und Schülern sowie Schülerarbeiten berücksichtigt werden.

▲ Zwei Ziele der Zeugnisse: Berichtfunktion entsprechend den Curriculumvorgaben der Provinz Ontario und Anbahnen der Kompetenz zur Weiterentwicklung auf Seiten des Schülers.

▲ Fester Zeugnisbestandteil ist eine Seite mit Kommentaren und Stellungnahmen zu der Lehrerbeurteilung durch Eltern und den Schüler.

▲ Neben den Zeugnissen werden schriftliche Lernvereinbarungen getroffen. Hier schreiben die Schüler auf, was sie schon gut können, was sie lernen möchten, was sie selbst zur Verbesserung beitragen wollen und welche Hilfe sie sich vorstellen. Eltern und Lehrer schreiben ihre Stellungnahme hierzu auf dasselbe Formular.

Die zentrale Stoßrichtung für eine ernsthafte Reform der Leistungsbeurteilung müsste also sein, dass eine hierarchische Bewertung ersetzt wird durch dialogische Formen der Verständigung über die Qualität von Leistungen[347]. Ermutigen können dabei Veränderungen in der freien Wirtschaft, wie sie Landmesser u.a. (2003, 23) feststellen: »Die Mitarbeiterinnen und Mitarbeiter sind für ihre eigene Entwicklung zunehmend selbst verantwortlich. Sie werden zu Unternehmern ihrer eigenen Talente und Fähigkeiten. Der Personalentwicklungsbereich war früher Vollstrecker, Administrator und Planer, er entwickelt sich heute zunehmend zum Berater für Laufbahn und Lernen.«

In welchem Maße ähnliche Vorstellungen auch in den Schulen politisch durch- und praktisch umgesetzt werden können, hängt von grundlegenden Veränderungen im Bildungswesen ab. Wenn Selektionsentscheidungen nicht - wie in anderen Ländern üblich - zeitlich weiter aufgeschoben werden[348], überlagert der Auslesezwang das zumindest für die Grundschule zentrale Förderinteresse - und Veränderungen der Beurteilungsverfahren bleiben Kosmetik.

Unter dem Gesichtspunkt der Akzeptanz in der Praxis scheint zurzeit eine Kombination von Ziffern und verbalen Kommentaren als Zwischenschritt am ehesten Erfolg zu versprechen. Aber auch dann sind pragmatisch Zwischenlösungen denkbar, die allen Beteiligten Gewinn bringen. So könnte das jährliche Versetzungszeugnis (bestehend aus Ziffernnoten und deren Deutung und Begründung) verbunden werden mit einem Gespräch zum Schulhalbjahr zwischen KlassenlehrerIn, SchülerIn, Eltern. Hier sollte es um Lernbereitschaft/-ergebnisse/-nöte/-defizite und daraus zu ziehende Konsequenzen gehen, die ggf. in einem »Lernvertrag« für das zweite Schulhalbjahr festzuhalten sind. Selbst dann, wenn das Gespräch den schriftlichen Entwicklungsbericht nicht ersetzt, sondern kommentiert würde es seine Abfassung entlasten - weil er nicht für sich allein stünde und weil er durch die Sichtweise des Kindes und der Eltern relativiert werden könnte.

346 Vgl. die Informationen zur Zeugniserstellung und Zeugnisformulare für alle Altersstufen (in französischer, aber auch englischer Sprache > www.edu.gov.on.ca/fre/document/forms/report/1998/report98f.html#elem [Abruf: 27.2.2006].

347 Vgl. die konkreten Vorschläge in Bartnitzky u.a. (2005). Diese Materialien machen deutlich, dass sich Ansprüche wie die Kombination von Selbst- und Fremdbeurteilung oder die Berücksichtigung verschiedener Bezugsnormen auch in alltagstauglichen Formen umsetzen lassen.

348 Diese Forderung wird zunehmend auch von Nicht-Pädagoginnen erhoben, vgl. etwa Sinn (2006).

## 7

## Fazit und bildungspolitische Bewertung

Harlen (2004a, 7) resümiert die angelsächsische Forschung zur Validität und Reliabilität verschiedener Erhebungsverfahren und Bewertungsformen u.a. in den folgenden Punkten[349]:

▲ Wenn über Beurteilungsverfahren entschieden wird, dürfen die Grenzen externer Prüfungen und nationaler Tests nicht übersehen werden.
▲ Die grundlegenden und wichtigen Unterschiede von Lehrerurteil und Test müssen respektiert werden, indem man aufhört, die Qualität des Lehrerurteils über den Grad seiner Übereinstimmung mit Tests zu bestimmen.
▲ Für die Beurteilung sind Kriterien zu entwickeln, die sich auf die Ziele des Unterrichts und nicht nur auf spezifische Aufgaben beziehen. So kann LehrerInnen geholfen werden, ein tieferes Verständnis der Ziele von Unterricht zu gewinnen und die Beurteilung besser auf diese abzustimmen.
▲ LehrerInnen brauchen mehr Aus- und Fortbildung, die sie für die Risiken der Leistungsbewertung sensibilisiert und auf ihre unterschiedlichen Funktionen vorbereitet[350].
▲ Kontinuierliche wechselseitige Abstimmung von Kriterien im Austausch über konkrete Bewertungsversuche hilft LehrerInnen, Klarheit über die Ziele von Unterricht und darauf bezogene Beurteilungskriterien zu gewinnen[351].

Unser Fazit zur Ausgangsfrage »Sind Noten nützlich - und nötig?« fällt ähnlich kritisch aus. Noten erfüllen die Erwartungen ihrer Befürworter nicht:
- Sie sind nicht valider, objektiver und zuverlässiger als andere Beurteilungsformen.
- Die beanspruchte Vergleichbarkeit ist durch den in der Regel üblichen Bezug auf den Klassendurchschnitt und die unvermeidlichen Beurteilungsfehler sehr eingeschränkt.
- Ziffernnoten erfüllen die verschiedenen Funktionen der Leistungsbeurteilung (Motivation, Information) nicht besser, zum Teil sogar schlechter als andere Formen der Rückmeldung.

Wenn Noten im Schulalltag trotzdem so viel Zustimmung finden, hängt dies vermutlich damit zusammen, dass sie SchülerInnen und Eltern vertraut sind. Für LehrerInnen ist ihre Vergabe außerdem mit einem geringeren Arbeitsaufwand verbunden als das Schreiben von Verbalgutachten. Schließlich suggeriert ihre leichte Verrechenbarkeit eine Vereinfachung von Selektionsentscheidungen. Diese haben im deutschen Schulsystem eine hohe und im Vergleich zu anderen Ländern erheblich höhere Bedeutung.

Klaus-Jürgen Tillmann (2004, 10, 16) hat anhand der PISA-Daten vorgerechnet, dass am Ende der Grundschulzeit nur noch rund 80% der SchülerInnen eine Klasse ihres Einschulungsjahrgangs besuchen und dass es unter den 15-Jährigen kaum mehr als 60% sind, die eine »glatte« Schullaufbahn aufweisen können. Fast 40% der SchülerInnen haben also mindestens eine der folgenden Maßnahmen erlebt: Zurückstellung am Schulanfang; Nichtversetzung; Überweisung in die Sonderschule; »Abschulung« in eine niedrigere Schulform. Das bedeutet: »Kinder mit eher schwachen Leistungen machen häufig Misserfolgserfahrungen und werden schließlich in Hauptschulen oder Sonderschulen eingewiesen. Dort treffen sie ganz überwiegend auf Mitschüler/innen mit gleichem Schicksal. Es lässt sich empirisch nachweisen: In solchen Gruppen der Negativauslese ist das Anregungspotential dürftig, ist der Kompetenzerwerb gering (vgl. Schümer 2004), ist eine schul- und lerndistanzierte Haltung weit verbreitet.« (a.a.O., 17).

Man muss insofern eine mehrfache Benachteiligung von Kindern aus anregungsarmen Elternhäusern konstatieren[352]:
▲ Je höher der sozio-ökonomische Status der Eltern ist, umso anregungsreicher sind die Lernmöglichkeiten ihrer Kinder vor der Schule, so dass sie bessere kognitive Voraussetzungen in die Schule mitbringen.
▲ Weil Stadtviertel sich in ihrer sozio-ökonomischen Zusammensetzung stark unterscheiden[353], kommen sie in der Regel auch in eine Lerngruppe, die durch die Herkunft der anderen Kinder ebenfalls ein anregenderes Milieu bietet. Deshalb entwickeln sich auch ihre Leistungen über die Grundschulzeit hinweg besser - und damit ihre Chancen auf den Besuch einer höheren Schulform in der Sekundarstufe.
▲ Selbst wenn Kinder am Ende der Grundschulzeit vergleichbare Leistungen erreichen, ist ihr Zugang zu einer höheren Schulform umso wahrscheinlicher, je höher der soziale Status der Eltern ist: Sie erhalten häufiger eine Empfehlung für das Gymnasium und ihre Eltern folgen dieser Empfehlung auch eher. Diese Entscheidung ist deshalb bedeutsam, weil sich die Leistungen in der Sekundarstufe auch bei gleichen kognitiven Voraussetzungen und gleichem sozialen Status der Eltern umso besser entwickeln, je höher die besuchte Schulform ist.
▲ Aber auch wenn Kinder mit vergleichbaren Grundschulleistungen in dieselbe Schulform wechseln, fällt der Lernerfolg innerhalb dieser Schulform umso besser aus, je höher

349 Auswahl und deutsche Zusammenfassung: Hans Brügelmann.
350 Die sieht Stiggins (1999, 198) auch für die USA als Schlüssel zur Steigerung des Ertrags von Leistungsbewertungen für den Lernprozess der SchülerInnen - vor allem mit Hinweis auf die Mängel der Alltagspraxis, wie sie Crooks (1988) dokumentiert hat (a.a.O., 194). Vgl. analog für Deutschland Jürgens (1998b, 191-192); Valtin (2002c).
351 Zu dem Ergebnis, dass punktuelle Fortbildungen nicht ausreichen, kommt auch Inckemann (2004) aufgrund ihrer Versuche zum Schriftspracherwerb.
352 Vgl. Brügelmann (2005a, 128) und speziell zu den Filtern beim Übergang in die Sekundarstufe, die je nach sozialer Herkunft den in Noten und Tests erfassten Leistungen unterschiedlich stark widersprechen: Elternwunsch > Lehrerempfehlung > Elterentscheidung: Ditton (1992, 132); Lehmann u.a. (1997, 89-102); Bos u.a. (2004b, 211-214); Geißler (2004, 18-19); OECD (2005, 89); zusätzlich wirkt sich der ethnische Hintergrund aus, vgl. Stallmann (1999, 254); Ditton u.a. (2005, 293, 295).
353 Vgl. zur hohen Bedeutung dieser Kontextbedingungen, die wesentlich stärker für Leistungsunterschiede zwischen Schulen verantwortlich sind als schulinterne Bedingungen: OECD (2005, 88).

der sozio-ökonomische Status der Eltern ist, da sie u.a. ihre Kinder besser unterstützen können.

Gesteuert werden die innerschulischen Ausleseprozesse durch Noten. Diese sind offensichtlich nicht in der Lage, unterschiedliche Fähigkeiten zureichend genau auszuweisen. Leistungen *und* ihre Beurteilung werden überlagert durch andere Faktoren, vor allem durch den Einfluss der sozialen Herkunft, den sie doch ersetzen sollen (vgl. > Kap. 0.3).

In Deutschland und Österreich stellt sich dieses Problem wegen der extrem frühen Aufteilung der SchülerInnen auf verschiedene Bildungswege mit besonderer Schärfe. Eine frühe Selektion ist unproduktiv, wie die niedrigeren Durchschnittsleistungen im PISA-Vergleich zeigen[354]. Damit ist sie auch ökonomisch verschwenderisch: Dringend benötigte Kompetenzressourcen werden verschenkt. Die Bindung der Selektion an Noten erweist sich als ineffektiv, weil die beanspruchte Trennung nach Fähigkeiten nicht funktioniert - zumindest wenn man die Testleistung als Maßstab nimmt. Auch dies belegen die PISA-Daten: »So würden - um nur ein Beispiel zu nennen - die 10% Besten in der Hauptschule im Gymnasium zum mittleren Leistungsbereich gehören. Und knapp die Hälfte der 15-Jährigen in Realschulen überschneiden sich in ihren Leistungen mit den Heranwachsenden in den Gymnasien (vgl. Artelt u.a. 2001, S. 121).« (Tillmann 2004, 14).

Damit ist die Gerechtigkeitsfrage gestellt. Denn dass Noten ihre Funktion als Selektionsinstrument nicht wirksam erfüllen, ist nur die eine Seite der Medaille. Zugleich verletzen sie auch das Recht des einzelnen Kindes auf Chancengleichheit und bestmögliche Förderung seines individuellen Potenzials. Die Kritik der »National Coalition für die Umsetzung der UN-Kinderrechtskonvention in Deutschland« (2005) am schulischen Bewertungssystem macht sehr deutlich, dass eine nur systemimmanente Bewertung der Effektivität von Noten zu kurz greift: »Die im Vordergrund internationaler Kritik stehende Bildungsbenachteiligung durch soziale Ungleichheit ist nicht nur Ausdruck eines strukturellen Mangels an Chancengerechtigkeit im gegliederten Schulsystem Deutschlands, sondern untergräbt das Recht auf Bildung jedes einzelnen betroffenen Kindes. [...] Die Leistungsbewertung durch Zensuren als Grundlage eines Berechtigungssystems ist pädagogisch fragwürdig; es verkürzt auch den Anspruch des Kindes auf Würdigung als *eigenständige Persönlichkeit.* Jedes Kind hat Anspruch darauf, dass seine Leistungen an seinem individuellen Vermögen, und nicht an abstrakten Regeln gemessen werden. [...] Einseitige Orientierung an Gesichtspunkten der Verwertbarkeit führt jedoch zu einer Verkürzung der Bildungsziele, die die *Subjektstellung* des Kindes und dessen allseitigen Bildungsanspruch unterminiert. [...] Die Vorgaben der Lehrpläne führen in Verbindung mit dem Bewertungs- und dem gekoppelten Berechtigungssystem in Deutschland zu einer weitgehenden ›Enteignung des Lernens‹ durch Fremdbestimmung.« (a.a.O., 2, 6)

Mit dem letzten Teilsatz nimmt die National Coalition ausdrücklich Bezug auf Bildungsstandards, die keine zureichende »Offenheit« für die individuell unterschiedliche Entwicklung von Kindern gewährleisten. Gleiche Anforderungen für alle zum selben Zeitpunkt verletzen das »Recht auf *Eigenaktivität und Selbstbestimmtheit* des Kindes« (ebda).

Eine Diskussion der Noten nur als »nützliches« oder »nötiges« Mittel der Leistungsbeurteilung greift demnach zu kurz. Problematisch werden sie durch ihre Instrumentalisierung als Auslesefilter. Der Verweis der National Coalition auf die UN-Kinderrechtskonvention macht die gesellschaftspolitische und völkerrechtliche Dimension der Notenfrage unmissverständlich klar: »Die ausdrückliche Hervorhebung, dass das Recht des Kindes auf Bildung ›auf der Grundlage der Chancengleichheit‹ zu verwirklichen sei, unterstreicht, dass Deutschland in diesem Punkt nicht nur bildungspolitisch, sondern auch völkerrechtlich im Abseits steht.« (a.a.O., 2).[355]

Damit wird aber auch deutlich, dass eine »Reparatur« technischer Schwächen von Noten nicht ausreicht, um die Probleme der Leistungsbewertung zu lösen. Sicher: Verbalgutachten können Leistungen, ihre Ursachen und konkrete Fördermöglichkeiten differenzierter ausweisen. Als entwicklungsorientierte Beschreibung von Lernverläufen machen sie Fortschritte und damit die individuelle Leistung des einzelnen Kindes besser sichtbar als eine Benotung im Vergleich mit anderen. Die Einbeziehung verschiedener PrüferInnen und auch standardisierter Aufgaben können helfen, die Validität, Objektivität und Reliabilität von Beurteilungen zu verbessern, indem sie informelle Leistungsproben ergänzen. Der punktuelle Einsatz normierter Tests ermöglicht LehrerInnen zudem, die vergleichende Bewertung von Leistungen über den Durchschnitt der jeweiligen Klasse hinaus auf repräsentative Stichproben zu beziehen und damit ihre eigenen Maßstäbe zu überprüfen.

Eine andere Bedeutung und Wirkung gewinnen Bewertungen - in gleich welcher Form - aber erst, wenn sich ihre Funktion ändert. Solange die Selektionsfunktion im System dominiert, werden eine stärkere Motivation der leistungsschwächeren SchülerInnen und eine differenziertere Förde-

---

354 Vgl. die letzte Auswertung von PISA-2000 durch die OECD (2005, 89, 93, 94) selbst und die dort deutlich formulierte Kritik einer frühen Selektion, auch wegen der auf diesem Weg verstärkten sozialen Selektion.

355 Die gelegentlich umstrittene unmittelbare Geltung der UN-Kinderrechtskonvention für innerstaatliche Maßnahmen ist durch ein Rechtsgutachten von Lorz (2003) geklärt. Danach ist Art 3 der Konvention
- unmittelbar anwendbare Völkerrechtsnorm;
- die nicht nur den Gesetzgeber, sondern auch die Rechtsanwender verpflichtet,
- auch wenn aus ihr keine konkreten Leistungsansprüche herleitbar sind,
- begründet sie eine Klagebefugnis gegen belastende Maßnahmen und
- einen Anspruch auf ermessensfehlerfreie Entscheidung über alle auf innerstaatliches Recht gestützten Anträge (a.a.O., 4).

Vor diesem Hintergrund ist auch der Deutschland-Besuch des Sonderberichterstatters der UN-Menschenrechtskommission, Vernor Muñoz, im Februar 2006 zum Thema »Recht auf Bildung« zu sehen (vgl. Kaube 2006 sowie Spiewak 2006 und die Berichterstattung in den Tageszeitungen vom 22.2.2006 zum abschließenden Pressegespräch des UN-Kommissars).

rung ihres Lernens nicht erreicht werden können. So machen die US-amerikanischen Erfahrungen mit *high-stakes testing* darauf aufmerksam, dass eine Sanktionierung von schlechten Ergebnissen in Leistungsvergleichen pädagogisch kontraproduktiv ist[356]: Einengung des Curriculum auf die »Haupt«-fächer; kurzfristig orientiertes *teaching to the test;* Aussonderung schwacher SchülerInnen, weil sie das Leistungsbild beeinträchtigen. Das gilt nicht nur für Einzelpersonen, sondern auch für Institutionen wie Schulen. Dies haben vor allem die Wirkungen des Gesetzes »No Child Left Behind« gezeigt[357]. Erfahrungen in europäischen Ländern belegen darüber hinaus[358], dass selektive Strukturen alle Versuche einer anderen Beurteilung im Ergebnis außer Kraft setzen. Darum ist auch in Deutschland eine längere gemeinsame Schulzeit geboten, wie sie international längst Standard ist.

Dass und wie eine solche Reform erfolgreich umgesetzt werden kann, wenn sie sich nicht auf Veränderungen der äußeren Struktur beschränkt, zeigt beispielhaft das deutschsprachige PISA-Siegerland Südtirol[359]. Obwohl Italien insgesamt bei PISA-2003 (Lesen) mit 476 Punkten noch schlechter abgeschnitten hat als Deutschland mit durchschnittlich 491, erreichte die autonome Provinz Südtirol bei gleicher Schulstruktur mit Platz 1 im Lesen und Platz 5 in Mathematik ein deutlich besseres Ergebnis als der deutsche Spitzenreiter Bayern. Gleichzeitig arbeitete sich die Provinz gegenüber der IEA-Lesestudie (Anfang der 1990er Jahre) von einem Platz im Mittelfeld an die europäische Spitze vor und schneidet im Lesen noch einen Punkt besser ab als der bildungspolitische Wallfahrtsort Finnland - mit vollständiger Integration aller behinderten Kinder, ohne Sitzenbleiben und ohne Ziffernoten, stattdessen mit individuellen Aufgaben in offeneren Unterrichtsformen und einer Bewertung, die sich am persönlichen Lernfortschritt orientiert[360]. Erfolgreicher Unterricht ist also auch mit weniger Leistungsdruck möglich; und Schulsysteme können lernen, ohne Selektion auszukommen.

356 Vgl. zu den negativen Wirkungen von *high-stakes tests,* also von Bewertungsformen, von deren Ergebnis viel für die Betroffenen abhängt, die breite empirische Evidenz in US-amerikanischen Untersuchungen, zusammengefasst u.a. bei Kohn (2000); Linn (2000); Harlen/Deakin (2002, 4); zusammenfassend mit weiteren Nachweisen: Brügelmann (2005, Kap. 48; 2006).

357 Aktuell berichtet TIME Magazine (Nr. 16 vom 17.4.2006) mit dem Titelbild »Dropout Nation«, dass rund ein Drittel der SchülerInnen die High School ohne Abschluss verlassen.

358 Vgl. etwa die sorgfältige Evaluation des Modellversuchs »Schülerbeurteilung und Schulentwicklung« in Liechtenstein: Roos (2003, 135, 138-139).

359 Vgl. oben > Kap. 0.5 und die bereits dort zitierten: Höllrigl/Meraner (2005); Leitzgen (2005); Meraner (2005); Ratzki (2005; 2006).

360 Das heißt nicht, dass diese Elemente in jeder Klasse in jeder Stunde optimal umgesetzt werden. Aber die pädagogischen Prinzipien weisen deutlich in eine andere Richtung als im deutschen Selektionssystem.

361 In diesem Verzeichnis werden alle im Gutachten genutzten Titel nachgewiesen. Außerdem haben wir Publikationen aufgenommen, die wir zwar in die Vorarbeiten einbezogen, aber im Text nicht ausdrücklich zitiert haben, sowie weitere, z.B. von Dritten zitierte Veröffentlichungen, die uns für vertiefende Analysen relevant erschienen.





## Literaturnachweise, weiterführende Literatur[361] und Abbildungsverzeichnis


Ammann, C.-H. (2002): Subjektive Fehlerquellen in der Beurteilung > www.multimedia-pflege.de/paed/beurteil/ingenk89_67.html [Abruf: 7.3.2006].

Amsbeck, U. (1999): Leistungsbeurteilung ohne Noten im europäischen Ausland. In: Grundschule, 31. Jg., H. 1, 24-26.

Amrein, A.L./ Berliner, D.C. (2002): High-stakes testing, uncertainty, and student learning. In: Education Policy Analysis Archives, Vol.10, No.18. [http://epaa.asu.edu/epaa/v10n18/].

Arnold, K.-H. (1997b): Strukturelemente und Verlauf einer lernförderlichen Leistungsbeurteilung. Schulforschungsprojekt Nr. 87. Senator für Bildung: Bremen.

Arnold, K.-H. (1999): Fairness bei Schulsystemvergleichen.Diagnostische Konsequenzen von Schulleistungsstudien für die unterrichtliche Leistungsbewertung und binnenschulische Evaluation. Waxmann: Münster u.a.

Arnold, K.-H. (2001): Qualitätskriterien für die standardisierte Messung von Schulleistungen. Kann eine (vergleichende) Messung von Schulleistungen objektiv, repräsentativ und fair sein? In: Weinert (2001, 117-130).

Arnold, K.-H./Vollstädt, W. (2001): Arbeits- und Sozialverhalten in der Schule. Möglichkeiten und Grenzen ihrer Beurteilung durch »Kopfnoten«. In: Die Deutsche Schule, 93. Jg., H. 2, 199-209.

Artelt, C., u.a. (2001a): Lesekompetenz: Testkonzeption und Ergebnisse. In: Baumert u.a. (2001, 69-137).

Artelt, C., u.a. (Hrsg.) (2001b): PISA 2000: Zusammenfassung zentraler Befunde. Berlin: Max-Planck-Institut für Bildungsforschung > http://www.mpib-berlin.mpg.de/pisa/ergebnisse.pdf [Abruf: 12.2.06].

Arzberger, K. (1988): Über die Ursprünge und Entwicklungsbedingungen der Leistungsgesellschaft. In: Hondrich u.a. (1988, 23-49).

Backhaus, A. (2005): Beim Lesen stolpern? Vom Stolperwörter-Lesetest zum Siegener Lesetest und der Testung der Leseleistung am PC. In: Hofmann/Sasse (2005, 128-137).

Backhaus, A. (2006): Die Zugehörigkeit zur Klasse oder das Testergebnis? Eine Mehrebenenanalyse zur Vorhersage von Noten für Lesen und Rechtschreibung in der Grundschule. Unveröff. Arbeitspapier des Projekts LUST. FB 2 der Universität: Siegen.

Backhaus, A./Moskopp, M. (2006): Der Siegener Satzlesetest. Ein Vergleich von Papier- und PC-Test. Unveröff. Arbeitspapier des Projekts LUST. FB 2 der Universität: Siegen.

Bambach, H. (1994): Ermutigungen. Nicht Zensuren. Ein Plädoyer in Beispielen. Libelle: CH-Lengwil.

Bambach, H., u.a. (Hrsg.) (1996): Prüfen und beurteilen. Zwischen Fördern und Zensieren. Jahresheft XIV. Friedrich-Verlag: Seelze.

Bangert-Drowns, R.L., et al. (1991) : The instructional effect of feedback in test-like events. In: Review of Educational Research, Vol. 61, 213-238.

Baron-Boldt, J. u.a. (1988). Prädiktive Validität von Schulabschlußnoten: Eine Metaanalyse. Zeitschrift für Pädagogische Psychologie, 2. Jg., 79- 90.

Baron Boldt, J., u.a. (1989): Prognostische Validität von Schulnoten. Eine Metaanalyse der Prognose des Studien und Ausbildungserfolgs. In: Jäger u.a. (1989, 11 39).

Bartnitzky, H. (1995): Stellungnahme zum Zeugniskonzept des Schulversuchs Bern-West. Vervielf. Ms. Bezirksregierung: Düsseldorf.

Bartnitzky, H. (2004): Zeugnisse als Selbstreflexion - mit einem Vorschlag für Schulen. In: Bartnitzky/ Speck-Hamdan (2004, 238-248).

Bartnitzky, H. (2005a): VERA Deutsch 2004: ungeeignet und bildungsfern. In: Grundschule aktuell, H. 89, 10-16.

Bartnitzky, H. (2005b): »Schimpansenkinder müssen laufen lernen« - Lesetest in Bayern. In: Grundschule aktuell, H. 92, 25-27.

Bartnitzky, H./Christiani, R. (1987): Mängelkatalog für Noten. In: Neue Deutsche Schule, H. 11/1987, 4-5.

Bartnitzky, H./Portmann, R. (Hrsg.) (1992): Leistung der Schule - Leistung der Kinder. Beiträge zur Reform der Grundschule, Bd. 87. Arbeitskreis Grundschule: Frankfurt.

Bartnitzky, H./Speck-Hamdan, A. (Hrsg.) (2004): Leistungen der Kinder wahrnehmen - würdigen - fördern. Beiträge zur Reform der Grundschule, Bd. 118. Grundschulverband: Frankfurt.

Bartnitzky, H., u.a. (1999): Zur Qualität der Leistung. 5 Thesen zur Evaluation und Rechenschaft der Grundschularbeit. Grundschulverband - Arbeitskreis Grundschule e.V.: Frankfurt (auch in: Schmitt 1999, 164-198).

Bartnitzky, H. u.a. (Hrsg.) (2005): Pädagogische Leistungskultur: Materialien für Klasse 1/2. Beiträge zur Reform der Grundschule, Bd. 119. Grundschulverband: Frankfurt.

Bartnitzky, H., u.a. (Hrsg.) (2006, i.V.): Pädagogische Leistungskultur: Materialien für Klasse 3/4. Beiträge zur Reform der Grundschule, Bd. 121. Grundschulverband: Frankfurt.

Baumeister, R. F., et al. (2004): Exploding the self-esteem myth. In: Scientific American, December 20, 2004 > www.sciam.com/print_version.cfm?articleID=000CB565-F330-11BE-AD0683414B7F0000 [Abruf: 4.2.2005].

Baumert, J./Schümer, G. (2001): Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In: Baumert u.a. (2001, 323-401).

Baumert, J./Watermann, R. (2000): Institutionelle und regionale Variabilität und die Sicherung gemeinsamer Standards in der gymnasialen Oberstufe. In: Baumert u.a. (2000b, 317-372).

Baumert, J., u.a. (1994): Das Bildungswesen in der Bundesrepublik Deutschland. Max-Planck-Institut für Bildungsforschung - Arbeitsgruppe Bildungsbericht. Rowohlt-Sachbuch 9193: Reinbek.

Baumert, J., u.a. (Hrsg.) (2000b): TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie - Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 2: Mathematische und naturwissenschaftliche Grundbildung am Ende der gymnasialen Oberstufe. Leske + Budrich: Opladen.

Baumert, J., u.a. (Hrsg.) (2001): PISA 2000 - Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich: Opladen.

Baumert, J., u.a. (Hrsg.) (2002): PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich. Leske + Budrich: Opladen.

Baumert, J., u.a. (2003): PISA 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Leske + Budrich: Opladen.

Baumgart, F./Lange, U. (Hrsg.) (1999): Theorien der Schule. Erläuterungen Texte Arbeitsaufgaben. Klinkhardt: Bad Heilbrunn.

Baumgarten, J., u.a. (Red.) (2005): Research Report 2003-2004. Max-Planck-Institut für Bildungsforschung: Berlin.

Baurmann, J. (1975): Aufsatzbenotung und Reihenfolgeeffekt. Beeinflusst die Reigenfolge im Beurteilungsvorgang die Aufsatzbenotung? In: Psychologien in Erziehung und Unterricht, 22. Jg., 181-185.

Baurmann, J. (1977): Der Einfluss von Auswertungsbedingungen, Vorinformationen und Persönlichkeitsmerkmalen auf die Benotung von Deutschaufsätzen. In: Ingenkamp (1977, 117-130).

Baurmann, J./Dehn, M. (2004): Beurteilen im Deutschunterricht. In: Praxis Deutsch, 31. Jg., H. 184, 6-13.

Bayerisches Kultusministerium (2004): Weiterentwicklung der Unterrichtsqualität hat Vorrang. Kultusministerin Monika Hohlmeier zum Schuljahresbeginn 2004/05. Pressemitteilung Nr. 240 vom 13.September 2004.

Beck, O./Hofen, N. (1991): Aufsatzunterricht Grundschule. Schneider Hohengehren: Baltmannsweiler.

Becher, A.L./Maclure, S. (eds.) (1978): Accountability in education. Social Science Research Council. National Foundation of Educational Research: London.

Becker, G./Ramseger, J. (2003): Bewertung des Arbeits- und Sozialverhaltens in den Klassenstufen 3-10 der allgemeinbildenden Schulen in Brandenburg. Inhaltliche Probleme - Weiteres Vorgehen. Aide-mémoire für eine Besprechung im MBJS, Potsdam.

Becker, H./Hentig, H.v. (Hrsg.) (1983): Zensuren. Lüge - Notwendigkeit - Alternativen. Klett-Cotta: Stuttgart.

Becker, G., u.a. (2006): Diagnostizieren und Fördern. Stärken entdecken - Können entwickeln. Friedrich Jahresheft XXIV. Erhard Friedrich Verlag: Seelze.

Behnken, I./Jaumann, O. (Hrsg.) (1995): Kindheit und Schule. Kinderleben im Blick von Grundschulpädagogik und Kindheitsforschung. Juventa: Weinheim/München.

Bellenberg, G., u.a. (2004): Selektivität und Durchlässigkeit im allgemein bildenden Schulsystem. Rechtliche Regelungen und Daten unter besonderer Berücksichtigung der Gleichwertigkeit von Abschlüssen. Arbeitsgruppe Bildungsforschung/Bildungsplanung. Universität Essen/Duisburg: Essen.

Bender, P. (2004): Die etwas andere Sicht auf den mathematischen Teil der internationalen Vergleichsuntersuchungen PISA sowie TIMSS und IGLU. In: GDM-Mitteilungen, H. 78, 101-108.

Benholz, E., u.a. (2005): Wie schwierig sind Texte aus Leistungstests? Textverstehen mehrsprachiger Kinder . In: Grundschule aktuell, H. 92, 21-24.

Benner, D./Ramseger, J. (1985): Zwischen Ziffernzensur und pädagogischem Entwicklungsbericht. In: Zeitschrift für Pädagogik, 31.Jg.,151-74.

Benner, D., u.a. (Hrsg.) (1996a): Pädagogische Eigenlogiken im Transformationsprozeß von SBZ, DDR und neuen Ländern. Freie Universität: Berlin.

Benner, D., u.a. (1996). Bildung und Schule in Transformationsprozess von SBZ, DDR und neuen Ländern - Untersuchungen zu Kontinuität und Wandel. Berlin: Freie Universität: Berlin.

Bennett, R.E, et al. (1993): Influence of behaviour, perceptions and gender on teachers' judgements of students' academic skill. In: Journal of Educational Psychology Vol. 85, 347-356.

Beutel, I. (1998): Berichtszeugnisse anders lesen - Anmerkungen zur eigenen Evaluationsstudie. In: Tillmann/Wischer (1998, 85-95).

Beutel, S.-I. (2000): Grundschulkinder als Experten für Lernberichte - eine Auswertung von Kinderinterviews. In: Beutel u.a. (2000, 155-204).

Beutel, S.-I. (2004): Zeugnisse aus Kindersicht. Habilitation an der Universität: Jena (publ. 2005 in der Schriftenreihe der Max-Traeger-Stiftung. Juventa: Weinheim/München).

Beutel, S.-I.(2005): Zeugnisse aus Kindersicht. Kommunikationskultur an der Schule und Professionalisierung der Leistungsbeurteilung. Juventa, Weinheim und München.

Beutel, S.-I./Vollstädt, W. (Hrsg.) (2000): Leistung ermitteln und bewerten. Bergmann + Helbig: Hamburg.

Beutel, S.-I./Vollstädt, W. (2002): Kinder als Experten für Leistungsbewertung. In: Zeitschrift für Pädagogik, 48. Jg., H. 4, 591-613.

Beutel, S.-I., u.a. (1999): Ermittlung und Bewertung schulischer Leistungen. Behörde für Schule/Freie Hansestadt: Hamburg.

Beutel, S.-I., u.a. (2000): Die schulische Beurteilungspraxis aus der Sicht von Schülern, Lehrern und Eltern. Universitäten: Bielefeld und Jena.

Birkel, P. (1978): Mündliche Prüfungen. Zur Objektivität und Validität der Leistungsbeurteilung. Kamp: Bochum.

Birkel, P. (2003): Aufsatzbeurteilung - ein altes Problem neu untersucht. In: Didaktik Deutsch, 9. Jg., H. 15, 46-63.

Birkel, P./Birkel, C. (2002): Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? In: Psychologie in Erziehung und Unterricht, 49. Jg., 219-224.

Birkhäuser, K. (1999): Mehr fördern, weniger auslesen. Zur Entwicklung der schulischen Beurteilung in der Schweiz. Trendbericht Nr. 3. Schweizerische Koordinationsstelle für Bildungsforschung: Aarau.

Black, P./Wiliam, D. (1998a). Assessment and classroom learning. In: Assessment in Education, Vol. 5, No. 1, 7-71.

Black, P./Wiliam, D. (1998b): Inside the black box. Raising standards through classroom assessment. In: Phi Delta Kappan, Vol . 80, No. 2 (October), 139-148.

Block, R. (2006): Schulrecht vor Elternrecht? Neue empirische Befunde zur Zuverlässigkeit von Übergangsempfehlungen der Grundschulen. Arbeitsgruppe Bildungsforschung/-planung. Universität: Essen.

Block, R./Klemm, K. (2005): Soziale Herkunft entscheidet. PISA E 2003 - NRW im Vergleich. In: nds (GEW-nrw), 57. Jg., H. 12, 18-19.

Block, R./Klemm, K. (2006): PISA 2003: differenzierende Bemerkungen zum neuen Ländervergleich. In: SchulVerwaltung NRW, H. 2/2006, 38-40.

Böhnel, E. (1993): Wirkung von Unterricht in der leistungsheterogenen Gruppe auf Lernleistung, Schulangst, Schulfreude und auf Sozialkontakte zwischen den Schülern - unter besonderer Berücksichtigung des österreichischen Bildungswesens. In: Olechowski/ Persy (1993, 102-120).

Böttcher, W., u.a. (Hrsg.) (1999): Leistungsbewertung in der Grundschule. Beltz: Weinheim/Basel.

Bohl, T. (2003): Aktuelle Regelung zur Leistungsbeurteilung und zu Zeugnissen an deutschen Sekundarschulen. In: Zeitschrift für Pädagogik, 49. Jg , H. 4, S. 550-566.

Bohl, T. (2004): Prüfen und Bewerten im Offenen Unterricht. Beltz: Weinheim/Basel.

Bolscho, D., u.a. (Hrsg.) (1979): Grundschule ohne Noten. Arbeitskreis Grundschule: Frankfurt.

Bos, W./Baumert, J. (1999): Möglichkeiten, Grenzen und Perspektiven internationaler Bildungsforschung: das Beispiel TIMSS/III. In: Aus Politik und Zeitgeschichte, Beilage B 35-36/99 zu »Das Parlament«.

Bos, W./Pietsch, M. (Hrsg.) (2005): KESS 4. Kompetenzen und Einstellungen von SchülerInnen und Schülern Jahrgangsstufe 4. Behörde für Bildung und Sport: Hamburg.

Bos, W., u.a. (Hrsg.) (2004a): Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. Waxmann: Münster.

Bos, W., u.a. (2004b): Schullaufbahnempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe. In: Bos u.a. (2004a, 191-228).

Bos, W., u.a. (Hrsg.) (2005): IGLU. Vertiefende Analysen zu Leseverständnis, Rahmenbedingungen und Zusatzstudien. Waxmann: Münster.

Brammer, P. (1998): Evaluation der Lernentwicklungsberichte an der IGS Göttingen-Geismar. In: Tillmann/Wischer (1998, 96-108).

Breitschuh, G. (1979): Zur Geschichte des Schulzeugnisses. In: Bolscho u.a. (1979, 35-63).

Bremerich-Vos, A., u.a. (2005): Stellungnahme zur Kritik an VERA in »Grundschule aktuell«, H. 89. in: Grundschule-aktuell, H. 90, 3-6. s.a. > www.uni-landau.de/vera/ziele.htm

Brinkmann, E. (2004): Kurz vor den Zeugnissen. In: Grundschule Deutsch, 1. Jg., H. 4, 34-37.

Brinkmann, E. (2006): Bewertung von Aufsätzen - vor und nach einem Seminar. Interne Auswertung. Pädagogische Hochschule: Schwäbisch Gmünd.

Brinkmann, E./Brügelmann, H. (1993): Ideen-Kiste Schriftsprache 1 (mit didaktischer Einführung »Offenheit mit Sicherheit«). Verlag für pädagogische Medien: Hamburg.

Brookhart S.M./DeVoge, J.G. (1999): Testing a theory about the role of classroom assessment in student motivation and achievement. In: Applied Measurement in Education Vol. 12, 409-425.

Brügelmann, H. (1977): Einheitlichkeit durch Operationalisierung - ein Phantom. In: Flitner/Lenzen (1977, 71-87).

Brügelmann, H. (1980): Experimental decision making and responsive accountability. Expert report for »Basic Education Policies Project«. OECD/ CERI: Paris? Reprint der Kurzfassung http://www.agprim.uni-siegen.de/printbrue.htm [14.4.06].

Brügelmann, H. (1994a): Verflixte zweite Halbzeit. Die Länge von Diktaten als Falle für schwache RechtschreiberInnen. In: Brügelmann/Richter (1994, 206 207).

Brügelmann, H (1994b): Zählen LehrerInnen Rechtschreibfehler geschlechtsspezifisch? In: Richter/Brügelmann (1994, 31).

Brügelmann, H. (1998): Leistung auf dem Prüfstand. In: Grundschulverband aktuell, November 1998, 1 und 7f. (auch abgedruckt in Schmitt 1999, 153-156).

Brügelmann, H. (1999): Was leisten unsere Schulen? Qualität und Evaluation von Unterricht in der Diskussion. Kallmeyersche Verlagsbuchhandlung: Seelze.

Brügelmann, H. (2000a): Sind Noten doch nötig? In: Grundschulzeitschrift, 13. Jg., H. 132, 4.

Brügelmann, H. (2000b): Noten im 1. Staatsexamen (Lehramt Primarstufe Siegen) im Überblick (zweite, um weitere Strichproben ergänzte und in Details korrigierte Fassung v,14.4.2000). Vervielf. Ms. Arbeitsgruppe Primarstufe/FB 2 der Universität: Siegen.

Brügelmann, H. (2002): Besserwisser und Alleskönner. Ein erster Kommentar zur Relativierung von Folgerungen aus den Ergebnissen von PISA und zu ihrer Rezeption in den Medien. In: Schulverwaltung (Niedersachsen und Schleswig-Holstein), 12. Jg., H. 2, 36-39 [auch abgedruckt in: Schulverwaltung (Nordrhein-Westfalen), H. 2/2002, und Schulverwaltung (Baden-Württemberg), H. 4/2002, 76, 78-80]

Brügelmann, H. (2003a): Noten abschaffen? Pro. In: Pädagogik, 55. Jg., H. 3, 50.

Brügelmann, H. (2003b) Grundlegende Leseleistungen und der »Karawanen-Effekt« in der Grundschule. Zentrale Befunde aus dem Projekt LUST an der Universität Siegen. In: Grundschulverband Aktuell, Nr. 84 (November 2003), 19-25.

Brügelmann, H. (2003c): Lese-Untersuchung mit dem Stolperwörter-Test. Abschlussbericht des Projekts LUST-1 > www.uni-siegen.de/~agprim/lust/index.htm.

Brügelmann, H. (2004): Lese-/Schreibförderung nach PISA, IGLU und LUST: Was heißt eigentlich »funktional alfabetisiert«? In: Alfa-Forum, Nr. 54-55 (Sommer 2004), 16-18.

Brügelmann, H. (2005a): Schule verstehen und gestalten - Perspektiven der Forschung auf Probleme von Erziehung und Unterricht. Libelle: CH-Lengwil.

Brügelmann, H. (2005b): Der Karawaneneffekt. Eine Zwischenbilanz des Projekts LUST zum Lesenlernen. In: Neue Sammlung, 45. Jg., H. 1, 49-67.

Brügelmann, H. (2005c): Das Prognoserisiko von Risikoprognosen - eine Chance für »Risikokinder«? In: Hofmann/Sasse (2005, 146-172).

Brügelmann, H. (2006): International tests and comparisons in education performance: A pedagogical perspective on standards, core curricula, and the measurement of the quality of schooling. In: Rotte (2006, in print).

Brügelmann, H./Heymann, H.W. (2006): Klärung und Übersetzung von Forschung als Dienstleistung für die pädagogische Praxis. Plädoyer für die Einrichtung einer »Evaluationsstelle für nutzerorientierte Bildungsforschung«. Vervielf. Diskussionspapier (Fassung v. 16.3.06). FB 2 der Universität: Siegen.

Brügelmann, H./Richter, S. (Hrsg.) (1994): Wie wir recht schreiben lernen. Zehn Jahre Kinder auf dem Weg zur Schrift. Libelle Verlag CH Lengwil.

Brügelmann, H., u.a. (Hrsg.) (1998): Jahrbuch Grundschule. Fragen der Praxis - Befunde der Forschung [Schwerpunkte: Offener Unterricht; Mathematik]. Erhard Friedrich Verlag: Seelze.

Brügelmann, H., u.a. (Hrsg.) (1999): Jahrbuch Grundschule. Fragen der Praxis - Befunde der Forschung Bd. 2 [Schwerpunkte: Schulfähigkeit; Sprache]. Erhard Friedrich Verlag: Seelze.

Brunner, I., u.a. (Hrsg.) (2006): Das Handbuch Portfolioarbeit. Kallmeyer: Seelze (im Druck).

Büchner, P./Koch, K. (2002): Von der Grundschule in die Sekundarstufe. In: Die Deutsche Schule, 94. Jg., H. 2, 234-246.

Buff, A. (1988a): Überlegungen zu Reformen in der Schülerbeurteilung. In: Schweizer Schule, H. 4/88, 25-35.

Bundesinstitut für Berufsbildung (1998): Aussagekraft von Prüfungen. Referenz Betriebs System. Information Nr. 12. Bundesinstitut für Berufsbildung: Bonn.

Carter, R. S. (1971): Wie gültig sind die durch Lehrer erteilten Zensuren? In: Ingenkamp (1971, 123-133).

Chamberlin, D., et al. (1942). Did they succeed in college? Adventures in American education. Vol. IV. Harper & Brothers: New York.

Cizek, G.J., et al. (1995/1996): Teachers' assessment practices: preparation, isolation and the kitchen sink. In: Educational Assessment, Vol. 3, 159-179.
Cohen, P.A. (1984): College grades and adult achievement: A research synthesis. In: Research in Higher Education, Vol. 20, 281 293.
Crooks, T. (1988): The impact of classroom evaluation on students. In: Review of Educational Research, Vol. 58, 438-481.
Czerwenka, K., u.a. (1988): Was Schüler von der Schule halten. In: Die Deutsche Schule, 80.Jg., 1988, 132 145.
Czerwenka, K., u.a. (1990). Schülerurteile über die Schule. Bericht über eine internationale Untersuchung. Peter Lang: Frankfurt.
Darge, K., u.a. (2002): Welche Zeugnisarten wünschen sich SchülerInnen und Schüler für ihre Grundschulzeit? In: Valtin (2002a, 61-66).
Deci, E.L./Ryan, R.M. (1993): Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. In: Zeitschrift für Pädagogik, 39. Jg., H. 2, 223-238.
Deci, E.L., et al. (1999): A meta-analysis review of experiments examining the effects of extrinsic rewards on intrinsic motivation. In: Psychological Bulletin, Vol. 125, No. 6, 627-688.
De Groot, A.D. (1971): Fünfen und Sechsen. Beltz: Weinheim/Basel.
Dehn, M. (2001): Leistungsbewertung und -zensierung im Fach Deutsch. In: Pädagogik, 53. Jg., H. 7-8, 74-79.
Dehn, M. (2006): Zeit für die Schrift 1. Lesen lernen und Schreiben können. Cornelsen Scriptor: Berlin.
Dehn, M./Hüttis-Graff, P. (2006): Zeit für die Schrift 2. Beobachtung und Diagnose. Cornelsen Scriptor: Berlin.
Deutscher Bildungsrat (1970): Strukturplan für das Bildungswesen. Empfehlungen der Bildungskommission. Bundesdruckerei: Bonn.
Dicker, H. (1973): Untersuchung zur Beurteilung von Mathematikaufgaben. Diplomarbeit an der Erziehungswissenschaftlichen Hochschule Rheinland-Pfalz: Landau.
Diekmann, A. (1995): Empirische Sozialforschung. Rowohlt Re 55551: Reinbek.
Ditton, H. (1992): Ungleichheit und Mobilität durch Bildung. Theorie und empirische Untersuchung über sozial-räumliche Aspekte von Bildungsentscheidungen. Beltz: Weinheim/Basel.
Ditton, H., u.a. (2005): Bildungsungleichheit - der Beitrag von Familie und Schule. In: Zeitschrift für Erziehungswissenschaft, 2. Jg., 285-304.
Döbert,H./Geißler, G. (2000): Schulleistung in der DDR: Das System der Leistungsentwicklung, Leistungssicherung und Leistungsmessung. Peter Lang: Frankfurt.
Döpp, W., u.a. (2002): Lernberichte statt Zensuren. Erfahrungen von Schülern, Lehrer und Eltern. Klinkhardt: Bad Heilbrunn.
Dohse, W. (1967): Das Schulzeugnis - Sein Wesen und seine Problematik. Beltz: Weinheim/Berlin (2. Aufl.; 1. Aufl., 1963) (S. 39-43 und 62-67 auch in Ingenkamp 1971, 42-51).
Dressel, P.L. (1957): Facts and fancy in assigning grades. In: Basic College Quarterly, Vol. 2, 6-12.
Eells, W.C. (1930): Reliability of repeated grading of essay type examinations. In: Journal of Educational Psychology, Vol. 21, 48-52.
Eells, W.C. (1971): Die Zuverlässigkeit wiederholter Benotung von auf satzähnlichen Prüfungsarbeiten. In: Ingenkamp (1971, 117-122).
Ehmke, T., u.a. (2005): Soziale Herkunft im Ländervergleich. In: Prenzel u.a. (2005a, Kap.9).
Einsiedler, W./Schöll, G. (1995): Pro und contra ziffernfreie Beurteilung in der Grundschule. In: Pädagogische Welt, 49. Jg., H. 3, 120-124.
Elbing, E./Buschmann, S. (1985): Schülerbeurteilung mittels Wortzeugnissen - eine empirische Analyse. Institut für Empirische Pädagogik und Pädagogische Psychologie. Universität: München.
Eurydice (o.J.) Education in Europe, network, comparative studies on education and national education systems > www.eurydice.org [Abruf: 10.02.2006].
Fadsich, F./Steinert, B. (2005): Schulische Rahmenbedingungen im internationalen Vergleich. In: Bos u.a. (2005, 159-186).
Faigel, P. (1973): Die Problematik der Rechtschreibzensur. Überlegungen und Untersuchungsergebnisse. In: Linguistische Berichte, H. 24/1973, 103-108.
Fatke, R./Merkens, H. (Hrsg.) (2006): Bildung über die Lebenszeit. Schriftenreihe der DGfE. VA Verlag für Sozialwissenschaften: Wiesbaden.
Faust, G. (2005): Grundschule nach IGLU. In: Götz/Nießeler (2005, 161-176).
Faust-Siehl, G./Schweitzer, F. (1992): Anstrengung ist alles - Wie Kinder schulische Leistungen verstehen. In: Bartnitzky/Portmann (1992, 50-60).
Fend, H. (2006): Bildungserfahrungen und produktive Lebensbewältigung - Ergebnisse der LifE-Studie. In: Fatke/Merkens (2006, 31-56).
Fend, H., u.a. (1976): Sozialsiationseffekte der Schule. Beltz: Weinheim/Basel.
Ferdinand, W./Kiwitz, H. (1971): Über die Häufigkeitsverteilung der Zeugnisnoten 1 bis 6. In: Ingenkamp (1971, 178-185).
Fiegert, M. (2001): Der Leistungsbegriff in historisch-systematischer Perspektive. In: Solzbacher/Freitag (2001, 19-38).
Fiegert, M./Solzbacher, C. (2001): Alternative Schulen - alternative Leistungsbewertung. In: Solzbacher/Freitag (2001, 289-312).
Finetti, M. (2005): Bessere Noten für Mädchen bei gleicher Leistung. In: Süddeutsche Zeitung v. 8.11.2005.
Finlayson, D.S. (1971): Die Zuverlässigkeit bei der Zensierung von Aufsätzen. In: Ingenkamp (1971, 103-116; engl. 1951).
Flitner, A. (1992): Leistung ist mehr als Schulleistung. In: Bartnitzky/Portmann (1992, 10-14).
Flitner, A./Lenzen, D. (Hrsg.) (1977): Abitur-Normen gefährden die Schule. Piper: München
Fraser, B.J., u.a. (1987). Syntheses of educational productivity research. International Journal of Educational Research, Vol. 11, 145-252.
Frederiksen J./White B. (2004): Designing assessment for instruction and accountability: an application of validity theory to assessing scientific inquiry. In: Wilson (2004, 74-104).
Freitag, C. (2001): Die Schulreform in England und ihre Auswirkungen auf die Leistungsbewertung. In: Solzbacher/Freitag (2001, 59-75).
Fricke, R./Treinies, G. (1985): Einführung in die Metaanalyse. Methoden der Psychologie, Bd. 3. Hans Huber: Bern u.a.
Fuchs, L.S./Fuchs, D. (1986): Effects of systematic formative evaluation: A meta-analysis. In: Exceptional Children, Vol. 53, No. 3, 199-208.
Gaedike, A.-K. (1974): Determinanten der Schulleistung. In: Heller (1974, 46-93).
Gaude, P. (1989): Beobachten, Beurteilen und Beraten von Schülern. Diesterweg: Frankfurt.
Gebert, D. (1983): Zur Aussagekraft von Industrie und Handelskammer Facharbeiter Prüfungen im gewerblich technischen Bereich für die spätere Berufspraxis. In: Zeitschrift für Arbeitswissenschaft, 37. (9.NF) Jg., H 2., 107, 109.
Geißler, R. (2004): Bildung für wen? Die Benachteiligten der Bildungsexpansion. In: Sozialwissenschaften, 33. Jg., H. 2, 12-22.
Ghiselli, E. E. (1966): The validity of occupational aptitude tests. Wiley: New York.
Giest, H./Scheerer-Neumann, G. (Hrsg.) (1999): Jahrbuch Grundschulforschung, Bd. 2. Beltz/Deutscher Studienverlag: Weinheim.
Gipps C./Clarke, S. (1998): Monitoring consistency in teacher assessment and the impact of SCAA's guidance materials at Key Stages 1, 2, and 3. Final Report.Qualifications and Curriculum Authority: London.
Glass, G.V. (1976): Primary, secondary, and meta-analysis of research. In: Educational Researcher, Vol. 11, 3-8.
Glass, G.V. (1977): Integrating findings: The meta-analysis of research. In: Shulman (1977, 351-379).
Glatz, Kell, A. (Hrsg.) (2005): Lernstandserhebungen und Unterrichtsqualität. Siegener Studien Bd. 63. Gesellschaft zur Förderung der Lehrerbildung e.V. (Universität): Siegen, 111-123.
Götz, M. (2005): Verbalzeugnisse in der Grundschule - Anspruch und Realisierung. In: Götz/Nießeler (2005, 78-92).

Götz, M./Nießeler, A. (Hrsg.) (2005): Leistung fördern - Förderung leisten. Auer Verlag: Donauwörth.

Götz, M./Müller, K. (Hrsg.) (2005): Grundschule zwischen den Ansprüchen der Individualisierung und Standardisierung. Wiesbaden: VS Verlag für Sozialwissenschaften.

Gompf, G./Henrich, H. (2005): Englisch ab 3. Grundschuljahr ohne Noten. Wissenschaftliche Untersuchung der Einstellung von Eltern, Schülern und Lehrkräften in Rheinland-Pfalz und Thüringen. Kinder lernen europäische Sprachen e.V. > www.kles.org [Abruf: 9.2.06].

Graf, U. (2004): Schulleistung im Spiegel kindlicher Wahrnehmungs- und Deutungsarbeit. Eine qualitativ-explorative Studie zur Grundlegung selbstreflexiven Leistens im ersten Schuljahr. Dissertation. Pädagogische Hochschule: Ludwigsburg.

Gramsch, A./Krause-Hotopp, D. (2003): Neue Wege in der Leistungsbewertung. Erfahrungen mit Eltern-Kind-Zeugnis-Gesprächen. In: Die Deutsche Schule, 95. Jg., H. 4,.

Greuer-Werner, M., u.a. (Hrsg.) (1985): Berichte aus Schulpsychologie und Bildungsberatung. Deutscher Psychologen Verlag: Bonn.

Grissemann, H. (2000): Deutschnoten als »Ursache« von »Legasthenie«. In: Schweizer Schule, H. 3/2000

Groeben, A.v.d./Lenzen D. (Hrsg.) (1996): Berichten und Bewerten I. Ein Reader zum Beurteilungssystem der Laborschule. Werkstattheft 5. Universität: Bielefeld.

Groeben, A.v.d./Lenzen, D. (Hrsg.) (1997): Berichten und Bewerten II. Ein Reader zum Beurteilungssystem der Laborschule. Werkstattheft 6. Universität: Bielefeld.

Grolnick, W.S./Ryan, R.M. (1987): Autonomy in children's learning: An experimental and individual difference investigation. In: Journal of Educational Psychology, Vol. 81, 143-154.

Grünig, B., u.a. (1999).: Leistung und Kontrolle. Die Entwicklung von Zensurengebung und Leistungsmessung in der Schule. Juventa: Weinheim/München.

Grunder, H.-U./Bohl, T. (2001): Neue Formen der Leistungsbeurteilung in den Sekundarstufen I und II. Schneider Hohengehren: Baltmannsweiler.

Grzesik, J./Fischer, M. (1984): Was leisten Kriterien für die Aufsatzbeurteilung? Theoretische und praktische Aspekte des Gebrauchs von Kriterien und der Mehrfachbeurteilung nach globalem Eindruck. Forschungsbericht Nr. 3192 des Landes NRW. Westdeutscher Verlag: Opladen.

Günther, H./Ludwig, O. (Hrsg.) (1996): Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch. 2. Halbbd. Walter de Gruyter: Berlin/New York.

Haarmann, H. (Hrsg.) (1997): Handbuch elementarer Schulpädagogik. Beltz: Weinheim.

Haas, G. (1999): In der Schule Leistungen bewerten, ohne pädagogische Prinzipien außer Kraft zu setzen. Bewerten und Benoten im offenen Unterricht. In: Praxis Deutsch, 26. Jg., H. 155, 10-19.

Hadley, S. T. (1954): A school mark - fact or fancy. In: Educational Administration and Supervision, Vol. 40, 305-312.

Hadley, S.T. (1971): Feststellungen und Vorurteile in der Zensierung. In: Ingenkamp (1971, 134-141).

Haecker, H. (1971): Subjektive Faktoren im Leistungsurteil der Lehrer. In: Schule und Psychologie, 18. Jg., 74-84.

Haenisch, H. (1991): Erfolgreich unterrichten - Wege zu mehr Schülerorientierung. Forschungsergebnisse und Empfehlungen für die Schulpraxis. Arbeitsbericht No. 17. Landesinstitut für Schule und Weiterbildung: Soest.

Haenisch, H. (1996a): Schulversuch ›Zeugnisse ohne in den Klassen 3 und 4‹. Auswertung der Erfahrungsberichte aus den am Schulversuch beteiligten Grundschulen. Arbeitsberichte zur Curriculumentwicklung Schul- und Unterrichtsforschung, H. 41. Landesinstitut für Schule und Weiterbildung: Soest.

Haenisch, H. (1996b): Beurteilungen ohne Noten auf dem Prüfstand. Ergebnisse einer Befragung von Eltern und Lehrkräften zur Akzeptanz und zu den Wirkungen. Arbeitsberichte zur Curriculumentwicklung Schul- und Unterrichtsforschung, H. 42. Landesinstitut für Schule und Weiterbildung: Soest.

Haas, G. (1999): In der Schule Leistungen bewerten, ohne pädagogische Prinzipien außer Kraft zu setzen. Bewerten und Benoten im offenen Unterricht. In: Praxis Deutsch, 26. Jg., H. 155, 10-19.

Haecker, H. (1971): Subjektive Faktoren im Leistungsurteil der Lehrer. In: Schule und Psychologie, 18 Jg., 74-84.

Hofmann, B./Sasse, A. (Hrsg.) (2005): Übergänge. Kinder und Schrift zwischen Kindergarten und Schule. Bericht über die Jahrestagung der Deutschen Gesellschaft für Lesen und Schreiben, Rauischholzhausen 19.11.2004. Deutsche Gesellschaft für Lesen und Schreiben: Berlin.

Hall, K., et al. (1997): A study of teacher assessment at Key Stage 1. Cambridge Journal of Education, Vol. 27, 107-122.

Hall K./Harding A. (2002): Level descriptions and teacher assessment in England: Towards a community of assessment practice. In: Educational Research, Vol. 44, 1-15.

Hanke, P. (2002): Lehr-Lernkulturen und schriftsprachliche Handlungskompetenzen im Primarstufenbereich. Habilitationsschrift. Universität: Köln (publ. als 2005).

Hanke, P. (2005): Öffnung des Unterrichts in der Grundschule. Lehr-Lernkulturen und orthographische Lernprozesse im Grundschulbereich. Waxmann: Münster (Habil. Universität: Köln 2002)

Hargreaves, D.J., et al. (1996): Teachers' assessments of primary children's classroom work in the creative arts. In: Educational Research, Vol. 38, 199-211.

Harlen, W (2004a): A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In: Research Evidence in Education Library. EPPI-Centre, Social Science Research Unit, Institute of Education: London.

Harlen, W. (2004b) A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes. In Research Evidence in Education Library. EPPI-Centre, Social Science Research Unit, Institute of Education: London.

Harlen, W./Deakin Crick, R. (2002): A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1*). In: Research Evidence in Education Library. Issue 1. EPPI-Centre, Social Science Research Unit, Institute of Education: London.

Hartinger, A./Fölling-Albers, M. (2002): Schüler motivieren und interessieren. Ergebnisse aus der Forschung - Anregungen für die Praxis. Klinkhardt: Bad Heilbrunn.

Hartinger, A., u.a. (2003): Beeinflussen unterschiedliche Übertrittsregelungen an weiterführende Schulen die Leistungsängstlichkeit und die Qualität der Lernmotivation von Grundschüler/innen? Eine vergleichende Studie zwischen Niedersachsen und Bayern. In: Panagiotopoulou/Brügelmann (2003, 115-119).

Hartinger, A., u.a. (2004): »Grundschul-Numerus Clausus« oder Orientierungsstufe? Auswirkungen verschiedener Übertrittsbedingungen auf Motivationsstile und Leistungsängstlichkeit von Grundschulkindern. In: Empirische Pädagogik, 18. Jg. , H. 2, 173-193.

Hartmann, M. (2002): Der Mythos von den Leistungseliten. Spitzenkarrieren und soziale Herkunft in Wirtschaft, Politik, Justiz und Wissenschaft. Campus: Frankfurt/New York.

Hartog, P./Rhodes, E.C. (1971a): Prüfungszensuren in Geschichte und Englisch. In: Ingenkamp (1971, 78-89).

Hartog, P./Rhodes, E.C. (1971b): Die Beurteilung mündlicher Prüfungen. In: Ingenkamp (1971, 142-148).

Haußer, K. (1991): Verbalbeurteilung in Schulzeugnissen. Eine psychologische Inhaltsanalyse. In: Die Deutsche Schule, 83. Jg., H. 3, 348-359.

Heckhausen, H. (1974): Lehrer-Schüler-Interaktion. In: Weinert u.a. (1974, 547-573).

Heinzel, F. (Hrsg.) (2000): Methoden der Kindheitsforschung. Ein Über blick über Forschungszugänge zur kindlichen Perspektive. Juventa: Weinheim u.a.

Hell, B., u.a. (o.J.): Die Validität von Prädiktoren des Studienerfolgs - eine Metaanalyse. Universität: Hohenheim.

Heller, K.A. (Hrsg.) (1974): Leistungsbeurteilung in der Schule. Quelle & Meyer: Heidelberg.
Heller, K.A. (1995): Schulleistungsprognosen. In: Oerter/Montada (1995, 983-989).
Heller, K.A. (1997): Individuelle Bedingungsfaktoren der Schulleistung. In: Weinert/Helmke (1997, 183-201),
Heller, K.A. (1999): Wissenschaftliche Argumente für eine frühzeitige Schullaufbahnentscheidung. In: Schulreport (München), H. 3/99, 10-13.
Heller, K.A./Hany, E.A. (2001): Standardisierte Schulleistungsmessungen. In: Weinert (2001, 87-101).
Heller, K.A./Nickel, H. (Hrsg.) (1982): Modelle und Fallstudien der Erziehungs- und Schulberatung. Huber: Bern.
Heller, K.A., u.a. (1978): Prognose des Schulerfolgs. Eine Längsschnittstudie zur Schullaufbahnberatung. Beltz: Weinheim/Basel.
Helmke, A. (1988): Leistungssteigerung und Ausgleich von Leistungsunterschieden in Schulklassen: unvereinbare Ziele? In: Zeitschrift für Erziehungspsychologie und Pädagogische Psychologie, 20. Jg., H. 1, 45-76.
Helmke, A. (1992): Selbstvertrauen und schulische Leistungen. Hogrefe: Göttingen.
Helmke, A. (1997a): Das Stereotyp des schlechten Schülers: Ergebnisse aus dem SCHOLASTIK-Projekt. In: Weinert/Helmke (1997a, 269-279).
Helmke, A. (1997b): Entwicklung lern- und leistungsbezogener Motive und Einstellungen: Ergebnisse aus dem SCHOLASTIK-Projekt. In: Weinert/Helmke (1997a, 59-76).
Helmke, A. (1997c): Individuelle Bedingungsfaktoren der Schulleistung. Ergebnisse aus dem SCHOLASTIK-Projekt. In: Weinert/Helmke (1997a, 203-216).
Helmke, A. (1998): Vom Optimisten zum Realisten? Die Entwicklung des Fähigkeitsselbstkonzeptes vom Kindergarten bis zur 6. Klassenstufe. In: Weinert (1998, 115-132).
Helmke, A. (1999): Development from optimism to realism? Development of children's academic self-concept from kindergarten to grade 6. In: Weinert/Schneider (1999, 198-221).
Hengartner (1999): Mit Kindern lernen. Standorte und Denkwege im Mathematikunterricht. Klett und Balmer: CH-Zug.
Hentig, H.v. (1985): Die Menschen stärken, die Sachen klären. Reclam: Ditzingen.
Herrlitz, H.-G., u.a. (1998): Deutsche Schulgeschichte von 1800 bis zur Gegenwart. Eine Einführung. Juventa Verlag: Weinheim und München (2. ergänzte Auflage).
Herrmann, U. (2005): Noten abschaffen? Contra. In: Pädagogik, 55. Jg., H. 3, 51.
Hiebert E./Davinroy, K. (1993): Dilemmas and issues in implementing classroom-based assessment for literacy (Technical Report 365). Los Angeles, Centre for Research on Evaluation, Standards and Student Testing (CRESST) > www.cse.ucla.edu/CRESST/Reports/TECH365.PDF
Höllrigl, P./Meraner, R. (2005): Erfreuliche Ergebnisse. Frucht gemeinsamer Arbeit. In: Info (Informationsschrift für Kindergarten und Schule in Südtirol), H. 1 (Jänner)/2005, 2-3.
Hofmann, B./Sasse, A. (Hrsg.) (2005): Übergänge. Kinder und Schrift zwischen Kindergarten und Schule. Bericht über die Jahrestagung der Deutschen Gesellschaft für Lesen und Schreiben, Rauischholzhausen 19.11.2004. Deutsche Gesellschaft für Lesen und Schreiben: Berlin.
Hoge, R.D./Coladarci, T. (1989): Teacher-based judgments of academic achievement: A review of the literature. In: Review of Educational Research, Vol. 59, No. 3 (Fall 1989), 297-313.
Holtappels, H.G., u.a. (Hrsg.) (2004): Jahrbuch der Schulentwicklung, Bd.13. Daten, Beispiele und Perspektiven. Juventa: Weinheim/München.
Hondrich, K.O., u.a. (Hrsg.) (1998): Krise der Leistungsgesellschaft. Westdeutscher Verlag: Opladen.
Honig, M.-S., u.a. (Hrsg.) (1996): Kinder und Kindheit. Soziokulturelle Muster - sozialisationstheoretische Perspektiven. Kindheiten Bd. 7. Juventa: Weinheim.
Hopf, D. (1994): Kindergarten, Vorschule und Grundschule (Elementar- und Primarbereich). In: Baumert u.a. (1994, 292-340).
Hopp, A.-D./Lienert, G.A. (1971): Eine Verteilungsanalyse von Gymnasialzensuren. In: Ingenkamp (1971, 191-204).
Hosenfeld, I. (2002): Kausalitätsüberzeugungen und Schulleistungen. Waxmann: Münster.
Huber, A. (2003). Die Lebensweisheit der 15-jährigen. Warum unsere Jugend besser ist als ihr Ruf. München: Heinrich Hugendubel Verlag: München.
Huber, L. (2002): Leistung in der Schule. Rückblicke in die Geschichte - Fragen an die Gegenwart. In: Winter u.a. (2002, 11-19).
Huberman, M. (1980): Das Selbstkonzept. Eine Untersuchung über die Wirkung von Noten, Ranglisten und Preisen auf Kinder der Genfer Primarschule. FAPSE: Genf.
Hübner, O. (2003): Prognose beruflicher Eignung mittels biographischer Daten. Unveröff. Diplomarbeit. Fb Erziehungswissenschaften und Psychologie. Freie Universität: Berlin (zusammengefasst in: Landmesser 2003, 11).
Hunter, J.E./Hunter, R.F. (1984): Validity and utility of alternative predic tors of job performance. In: Psychological Bulletin, Vol. 96, No.1.,72-98.
Hunter, J., et al. (1982): Meta-Analysis: Cumulating research findings across studies. Sage: Beverly Hills/ Newbury Park, Cal. (new ed. 1990; 2004).
Inckemann, E, (2004): »Dass man aus einer Fortbildung heimgeht und morgen passiert es, geht halt nicht« - förderdiagnostische Kompetenz von Grundschullehrkräften. In: Bartnitzky/Speck-Hamdan (2004, 218-237).
Ingenkamp, K. (1967): Untersuchungen zur Übergangsauslese. Beltz: Weinheim/Berlin.
Ingenkamp, K. (1969): Die Bedeutung objektiver Leistungsbeurteilungen für moderne Grundschularbeit. In: Schwartz (1969, 53-80).
Ingenkamp, K. (Hrsg.) (1971a): Die Fragwürdigkeit der Zensurengebung. Beltz: Weinheim (7. überarb. Aufl. 1977; 9. Aufl. 1995).
Ingenkamp, K. (1971b): Überblick über die prognostische Bewährung der Grundschulgutachten und -zensuren. In: Ingenkamp (1971, 229-232).
Ingenkamp, K. (1971c): Sind Zensuren aus verschiedenen Klassen vergleichbar? In: Ingenkamp (1971, 156-163).
Ingenkamp, K. (1975): Pädagogische Diagnostik. Ein Forschungsbericht zur Schülerbeurteilung in Europa. Trendbericht im Auftrag des Europarats in Straßburg. Beltz: Weinheim/Basel.
Ingenkamp, K. (Hrsg.) (1977): Die Fragwürdigkeit der Zensurengebung. Beltz: Weinheim (7. überarb. Aufl.; 1. Aufl. 1971; 9. Aufl. 1995).
Ingenkamp, K. (Hrsg.) (1981): Wert und Wirkung von Beurteilungsverfahren. Untersuchungen zu den Gütekriterien und der Wirkung diagnostischer Instrumente in der Schule. Beltz: Weinheim/Basel.
Ingenkamp, K. (1989): Diagnostik in der Schule. Beiträge zu Schlüsselfragen der Schülerbeurteilung. Beltz: Weinheim/Basel. S. 95-126 (»Zeugnisse und Zeugnisreformen in der Grundschule aus der Sicht empirischer Pädagogik«)
Ingenkamp, K. (1991): Die Bedeutung von Schultests für moderne Bildungssysteme. Test-Info 1/91. Beltz: Weinheim/Basel.
Ingenkamp K.-H. (1992): Lehrbuch der pädagogischen Diagnostik. Beltz: Weinheim/Basel (2. Auflage).
Ingenkamp, K.-H. (1993): Der Prognosewert von Zensuren, Lehrergutachten, Aufnahmeprüfungen und Test während der Grundschulzeit für den Sekundarschulerfolg. In: Olechowski/Persy (1993, 68-85).
Ingenkamp, K./Jäger R.S. (Hrsg.) (1990): Tests und Trends. Jahrbuch der Pädagogischen Diagnostik., Bd. 8. Beltz: Weinheim/Basel.
Iten, M./Theiler, P. (1993): Ganzheitlich Beurteilen und Fördern. Erziehungsdepartement des Kantons: Luzern.
Jachmann, M. (2000a): Einstellungen von Lehrer, Eltern und Schülern zur Leistungsbeurteilung - ein Vergleich. In: Beutel u.a. (2000, 205-234).
Jachmann, M. (2000b): Zusammenfassung der Ergebnisse. In: Beutel u.a. (2000, 235-241).

Jachmann, M. (2003): Noten oder Berichte? Die schulische Beurteilungspraxis aus der Sicht von Schülern, Lehrern und Eltern. Leske + Budrich: Opladen.
Jachmann, M./Tillmann, K.-J. (2000a): Einführung. In: Beutel u.a. (2000, 9-26).
Jachmann, M./Tillmann, K.-J. (2000b): Leistungsbeurteilung und Zeugnisse aus der Sicht Hamburger LehrerInnen und Lehrer. In: Beutel u.a. (2000, 27-70).
Jacobs, B. (1999): Motivationales Feedback und Lernleistung. > www.phil.uni-sb.de/~jakobs/wwwartikel/feedback/motivation.htm [last update 12.5.05; Abruf: 14.2.2006].
Jäger, S., u.a. (Hrsg.) (1989): Tests und Trends 7. Jahrbuch der Pädagogischen Diagnostik. Beltz: Weinheim.
Jäger, R.S. (1998): Von der Beurteilung zur Notengebung. Verlag Empirische Pädagogik: Landau (2. vollst. überarb. Auflage).
Jäger, R.S. (2000): Von der Beobachtung zur Notengebung. Ein Lehrbuch. Diagnostik und Benotung in der Aus- Fort- und Weiterbildung. Zentrum für empirische pädagogische Forschung: Landau.
Johnston P.H., et al. (1993): Teachers' assessment of the teaching and learning of literacy. In: Educational Assessment, Vol. 1, 91-117.
Jürgens, E. (1997): Das Wortgutachten in der Grundschule. Eine empirische Untersuchung zur Praxis der Verbalbeurteilung. Universität: Bielefeld.
Jürgens, E. (1998a): Leistung und Beurteilung in der Schule. Eine Einführung in Leistungs- und Bewertungsfragen aus pädagogischer Sicht. Academia Verlag: St. Augustin (4. Aufl.).
Jürgens, E. (1998b): Zeugnisse ohne Noten. Die Verbalbeurteilungspraxis in der Grundschule als Gegenstand einer Untersuchung. In: Brügelmann, u.a. (1998, 187-192).
Jürgens, E./Sacher, W. (Hrsg.) (2000): Leistungserziehung und Leistungsbeurteilung: Schulpädagogische Grundlegung und Anregungen für die Praxis. Studientexte für das Lehramt Bd. 6. Luchterhand: Neuwied.
Jung, J. (2005): Formen, Prinzipien und Probleme der Leistungsbeurteilung. In: Götz/Nießeler (2005, 63-77).
Kahlert, J., u.a. (Hrsg.) (2000): Grundschule: Sich Lernen Leisten. Neuwied: Luchterhand.
Kalthoff, H. (1996): Das Zensurenpanoptikum. Eine ethnographische Studie zur schulischen Bewertungspraxis. In: Zeitschrift für Soziologie, 25. Jg., H. 2, 106-124.
Kanders, M./Rolff, H.-G. (2002): Mehr von allem, aber wenig ändern! Ergebnisse der neuen IFS-Repräsentativbefragung zu Schule und Bildung. Pressemitteilung des Instituts für Schulentwicklung. Universität: Dortmund > www.ifs.uni-dortmund.de/Download/Artikel%20zur%20IFS-Umfrage.pdf [Abruf: 16.2.2006].
Kanders, M./Rolff, H.-G. (2004): 13. IFS-Repräsentativumfrage zu Schule und Bildung. Vorlage zur Pressekonferenz am 15. Juni 2004 in Berlin.
Kanders, M., u.a. (1997): Das Bild der Schule aus der Sicht von Schülern und Lehrern. Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie: Bonn.
Kanders, M., u.a. (2004): IFS-Umfrage: Die Schule im Spiegel der öffent lichen Meinung - Ergebnisse der 13. IFS-Repräsentativbefragung der bundesdeutschen Bevölkerung.
Kaube, J. (2006): Der Menschenrechts-Revisor kommt. In: Frankfurter Allgemeine Zeitung, Nr. 32 v. 7.2.2006, 33.
KinderRÄchTsZÄnker (o.J.): Fällt Euch denn nichts besseres ein? Kritik an populärer und oberflächlicher Schulkritik und Pseudo-Alternativen > http://www.kraetzae.de/schule/schulkritik/#7 [Abruf: 27.3.06].
Kirschner, G. (1992): Kinder wollen Zeugnisse - wollen Kinder Noten? Meinungsumfrage über Zeugnisformen. In: Bartnitzky/Portmann (1992, 89-83).
Kirsten, N. (2003): Betragen ins Zeugnis? Verkopfte Debatte. In: Die Zeit, Nr. 37 v. 4.9.03.
Klauer, K.J. (1987): Fördernde Notengebung durch Benotung unter drei Bezugsnormen. In: Olechowski/Persy (1987, 180-206).
Klauer, K.J. (1992): In Mathematik mehr leistungsschwache Mädchen, im Lesen und Rechtschreiben mehr leistungsschwache Jungen? In: Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 24. Jg., H. 1, 48-65.
Klauer, K J. (2001): Wie misst man Schulleistungen? In: Weinert (2001, 103-115).
Key, E. (1992): Das Jahrhundert des Kindes. Pädagogisch Bibliothek Beltz, Weinheim/Basel.
Klieme, E. (o.J.): Abiturnoten, Leistungsstandards und Studierfähigkeit. Validierung von Benotungssystemen anhand von Zulassungsdaten und Ergebnissen des Medizinstudiums. Vervielf. Ms.
Klieme, E., u.a. (2003): Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Deutsches Institut für Internationale Pädagogische Forschung: Frankfurt.
Klieme, E., u.a. (2006): Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie »Deutsch Englisch Schülerleistungen International (DESI)«. Deutsches Institut für Internationale Pädagogische Forschung: Frankfurt > www.dipf.de/desi/DESI_Zentrale_Befunde.pdf [Abruf: 3.3.2006].
Klink, J.G. (1964): Die Schülerleistung im Koordinatensystem der Ziffernzensur. In: Lebendige Schule, 19. Jg., 375-383.
Kluger, A.N./DeNisi, A. (1996): The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. In: Psychological Bulletin, Vol. 119, No. 2, 254-284.
KMK (1970): Empfehlungen zur Arbeit in der Grundschule. Beschluß vom 2.7.1970. Sekretariat der Kultusministerkonferenz: Bonn.
Knoche, W. (1971): Die Noten im Auslesekriterium und der Schulerfolg am Gymnasium. In: Ingenkamp (1971, 236-251).
Köller, O. (2002): Des Schülers Leid, des Lehrers Freud. Schulnoten sind nötig und besser als ihr Ruf. In: Schule - Wissen - Bildung. Klett ThemenDienst Nr. 16: Dezember 2002, 7-10.
Köller, O., u.a. (1999): Wege zur Hochschulreife: Offenheit des Systems und Sicherung vergleichbarer Standards. In: Zeitschrift für Erziehungswissenschaft, 2. Jg., H. 3, 386-422.
Köller, O., u.a. (2000): Zum Zusammenspiel von schulischen Interessen und Lernen im Fach Mathematik: Längsschnittanalysen in den Sekundarstufen I und II. In: Schiefele/Wild (2000, 163-181).
Kohn, A. (1999). Punished by rewards. The trouble with gold stars, incentive plans, A's, praise, and other bribes. Houghton Miflin: Boston.
Kohn, A. (2000). The case against standardized testing. Raising the scores, ruining the schools: Heinemann: Portsmouth, NH.
Konrad, K. (1997): Lernen eigenständig planen, überwachen und bewerten. Explorative Analysen kooperativer Lernsequenzen. Verlag Empirische Pädagogik: Landau.
Koretz, D., et al. (1994): The Vermont Portfolio Assessment Program: findings and implications. In: Educational Measurement: Issues and Practice, Vol. 13, 5-16.
Krampen, G. (1985): Differenzielle Effekte von Lehrerkommentaren zu Noten bei Schülern. In. Zeitschrift für Erziehungspsychologie und Pädagogische Psychologie, 17. Jg., H. 2, 99-123.
Krampen, G. (1987): Effekte von Lehrerkommentaren zu Noten bei Schülern. In: Olechowski/Persy (1987, 297-227).
Krampen, F./Mory, M. (1982): Zur Verarbeitung einer schlechten Mathematikzensur. In: Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 14. Jg., 337-340.
Krapp, A./Mandl, H. (1977): Einschulungsdiagnostik: Eine Einführung in Probleme und Methoden der pädagogisch-psychologischen Diagnostik. Beltz: Weinheim.
Krope, P., u.a. (1999): Ziffernzeugnis versus Berichtszeugnis. Zur Lerneffektivität bei quantitativen und qualitativen Aussagen. In: Giest/Scheerer-Neumann (1999, 299-313).
Kühl, R. (1991a): Berichtszeugnisse in Klasse 1 bis 4. Krach in Schleswig-Holstein. In: Grundschul-Zeitschrift, 5. Jg., H. 49, 2-3.

Kultusministerium Baden-Württemberg (2004): Verordnung des Kultusministeriums über die Notenbildung vom 5. Mai 1983 (GBl. S. 324; K.u.U. S. 449), zuletzt geändert durch: Verordnung vom 23.3.2004. www.leu.bw.schule.de/bild/Notenbildung.pdf [Abruf: 17.3.06].

Lambrou, U. (1989a): Leistungsmessung. Eine Grenzwanderung... In: Päd.extra/Demokratische Erziehung, 2. Jg., H. 3, 36-9.

Landert, C. (1999): Die Arbeitszeit von Lehrpersonen in der Deutschschweiz. Verlag LCH: Zürich.

Landmesser, M., u.a. (2003): Schulleistungen, außerschulische Aktivitäten und Praxiserfolg. Die Bedeutung, Bewertung und Entwicklung von Handlungskompetenz. IBM Deutschland: Stuttgart > http://forum-kritische-paedagogik.de/start/download.php?view.198 [Abruf: 24.2.2006].

Landtag intern (1999): Am Aussagewert von Kopfnoten scheiden sich die Meinungen der Fraktionen im Landtag NRW. In: SchulVerwaltung NRW, 10. Jg., H. 10, 283-284.

Leffelsand, S. (2003): Schullaufbahnempfehlungen: Vergleich diagnostischer Entscheidungen von Grundschullehrer/innen und Lehramtsstudierenden. Poster. Universität: Dortmund > www.ifs.uni-dortmund.de/ifs/download/paeps2003_poster_leffelsend.pdf [24.3.06].

Lehmann, R.H. (1990): Aufsatzbeurteilung - Forschungsstand und empirische Daten. In: Ingenkamp/Jäger (1990, 64-94).

Lehmann, R.H. (1994): Essays, scoring of. In: Postlethwaite/Husén (1994, 2018-2025).

Lehmann, R.H. (1999): Wider die Notenwillkür. Bildungsforscher Rainer Lehmann über die Leistungen deutscher Schüler - und ihrer Schulen. In: Die Zeit, Nr. 41 v. 7.10.99, 38.

Lehmann, R.H. (2001): Messung von Schulleistungen im Primar- und Sekundbereich. In: Weinert (2001, 131-141).

Lehmann, R.H., u.a. (1997): Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen. Behörde für Schule, Jugend und Berufsbildung: Hamburg.

Leitzgen, A. (2005): Neues aus PISA. In: Family & Co, H. 10/2005 v. 15.9.2005.

Lempp, R. (1971): Lernerfolg und Schulversagen. Kösel: München.

Learnline (o.J.) www.learn-line.nrw.de/angebote/gemeinsamerunterricht/leistungsbewertung/index.html (Funktionen und Formen von Leistungsbewertungen, rechtliche Bedingungen) [Abruf: 21.2.2006].

Lenhard, W. (2005): Diagnostische Verfahren zur Schulleistungsfeststellung in der Grundschule. In: Götz/Nießeler (2005, 38 ff.).

Lind, G. (2003): Benoten und Lernen. Vorlesung Pädagogische Psychologie für Lehramtsstudierende. ? http://www.uni-konstanz.de/ag-moral/lernen/15_evaluation/noten.htm#pisa [23.1.2003]

Linn, R.L. (2000): Assessments and accountability. In: Educational Researcher, Vol. 29, No. 2, 4-15.

Lissmann, U. (1977): Gewichtung von Abiturnoten und Studienerfolg. Beltz: Weinheim.

Lissmann, U. (1981): Zur Wirkung verschiedener Rückmeldungstechniken auf Lernende. In: Ingenkamp (1981, 233-289).

Lissmann, U. (1987): Qualität des Unterrichts. Zur Modifikation und Relevanz der Leistungsrückmeldung des Lehrers und ihrer Abhängigkeit von Lernvoraussetzungen. In: Zeitschrift für erziehungswissen schaftliche Forschung, 21. Jg., 195 217.

Lissmann, U./Paetzold, B. (1987): Leistungsrückmeldung, Lernerfolg und Lernmotivation. Beltz: Weinheim/Basel.

Lorz, R.A. (2003). Der Vorrang des Kinderwohls nach Art. 3 der UN-Kinderrechtskonvention in der deutschen Rechtsordnung. Hrsgg. von der National Coalition für die Umsetzung der UN-Kinderrechtskonvention in Deutschland. Arbeitsgemeinschaft für Jugendhilfe: 10178 Berlin (Mühlendamm 3).

Ludwig, P. (1995): Pygmalion im Notenbuch. Die Auswirkung von Erwartungen bei Leistungsbeurteilung und -rückmeldung. In: Pädagogische Welt, 49. Jg., H. 3, 114-119.

Lübke, S.-I. (1996): Schule ohne Noten. Lernberichte in der Praxis der Laborschule. Leske + Budrich: Opladen.

Lütgert, W. (1992): Die Fragwürdigkeit der Zensurengebung und die Berichte zum Lernvorgang der Bielefelder Laborschule. In: Neue Sammlung 32. Jg., H. 3, 387-404.

Lütgert, W. (1999): Leistungs-Rückmeldung, Anforderung, Innovationen, Probleme. Pädagogik, 51. Jg., H. 3, 46-50 (auch in: In: Beutel/Vollstädt (2000).

Lütgert, W. (2002): Die Guten ins Töpfchen, die Schlechten … Zeugnisse und Zensuren: Der vergessene Teil der allgemeinen Didaktik. In: Lütgert/Hallpap (2002, 157-178).

Lütgert, W./Hallpap, X. (Hrsg.) (2002): Didaktik in Jena. Aufgaben zu Beginn des 21. Jahrhunderts. Jena: Friedrich-Schiller-Universität: Jena.

Lütgert, W./Jachmann, M. (2000): Leistungsbeurteilung und Zeugnisse aus der Sicht Hamburger Eltern. In: Beutel u.a. (2000, 71-110).

Lütgert, W./Tillmann, K.-J. (2000): Vorwort. In. Beutel u.a. (2000, 7).

Lütgert, W., u.a. (2001): Leistungsbeurteilung und -rückmeldung an Hamburger Schulen. Bericht über ein Forschungsprojekt. Hrsgg. von der Behörde für Schule, Jugend und Berufsbildung der Freien und Hansestadt: Hamburg.

Maier, M. (2001): Das Verbalzeugnis in der Grundschule. Verlag Empirische Pädagogik: Landau.

Maier, M. (2003): Was leisten Verbalzeugnisse? In: Grundschule, 35. Jg., H. 7-8, 72-75.

Martschinke, S., u.a. (2005): Die ersten Notenzeugnisse und der Übertritt in der Perspektive der Kinder - Ergebnisse aus der KILIA-Studie. In: Götz/ Müller (2005, 85-92).

Meiers, K. (1989a): Nur der Noten wegen schöner schreiben? Offener Brief an das Ministerium für Kultus und Sport Baden-Württemberg. In: Grundschule, 21. Jg., H. 7+8, 92-93.

Meisels S.J., et al. (2001): Trusting teachers' judgements: A validity study of a curriculum-embedded performance assessment in kindergarten to Grade 3. In: American Educational Research Journal, Vol 38, 73-95.

Meraner, R. (2005): Spitze bei PISA. Die Ergebnisse und erste Überlegungen. In: Info (Informationsschrift für Kindergarten und Schule in Südtirol), H. 1 (Jänner)/2005, 12-16.

Merkelbach, N. (1986): Korrektur und Benotung im Aufsatzunterricht. Wissenschaftliche Erkenntnisse und didaktische Konzepte. Frankfurt.

Merkelbach, V. (2005): Die Strukturfrage ist längst gestellt. Schulpolitische Perspektiven der Ländervergleichsstudie PISA 2003. In: PISA-INFO 38/2005 der GEW: Frankfurt. http://user.uni-frankfurt.de/~merkelba/ > Dezember 2005.

Merkelbach, V. (2005b): Schule ohne Noten - wie soll das gehen? Dialogische Leistungsbewertung als Element einer anderen Lernkultur. http://user.uni-frankfurt.de/~merkelba/ > Juni 2005.

Merkens, H. (2005): Schulkarrieren von Kindern mit Migrationshintergrund in den ersten drei Jahren der Grundschule. Ergebnisse aus dem Projekt BeLesen: Berliner Längsschnittsstudie zur Lesekompetenzentwicklung von Grundschulkindern. Berichte aus der der Arbeit des Arbeitsbereichs Empirische Erziehungswissenschaft, Nr. 43. Freie Universität: Berlin.

Metz, H. (1982): Unterrichtsbeurteilungen auf dem Prüfstand. In: Die Deutsche Schule, 74. Jg., H. 1, 44-57.

Micklos, J. (1982): Clouds and silver linings: A realistic look at reading achievement. In: The Reading Teacher, Vol. 35, 644-646.

Minker, U. (2005): Der Übergang von der Grundschule zu den weiterführenden Schulen im Fach Englisch - Fallanalysen im schulischen Kontext. Dissertation im FB 3. Universität: Siegen.

Mount, M., et al. (2000): Incremental vailidity of empirically keyed biodata scales over GMA and the five factor personality constructs. In: Personnel Psychology, Vol. 53, No. 2, 299-323.

Morys, R. (2006): Die Leistungsselbstsicht von Grundschulkindern im Beziehungsgeflecht von Schule und Elternhaus - Schwerpunkt Leseleistung. Dissertation. Pädagogische Hochschule: Ludwigsburg.

Mreschar, R.I. (Hrsg.) (1985): Erzieher und Erzogene. Schüler, Lehrer, Eltern im Blickpunkt der Forschung. Verlag Deutscher Forschungsdienst: Bonn-Bad Godesberg.

Müller, K. (2005): Zeugnisbestimmungen in den Bundesländern. In: Götz/ Nießeler (2005, 93-101):

Müller-Naendrup, B. (Red.) (2005): Lernbeobachtung - Leistungsbeurteilung. Reader zum Seminar. Arbeitsgruppe Primarstufe im FB 2. Universität: Siegen.

National Coalition für die Umsetzung der UN-Kinderrechtskonvention in Deutschland (2005): Die Rechte des Kindes nach der Kinderrechtskonvention der Vereinten Nationen im deutschen Schulwesen. Diskussionspapier. Arbeitsgemeinschaft für Jugendhilfe: 10178 Berlin (Mühlendamm 3).

Naegele, I./Valtin, R. (Hrsg.) (2003): LRS - Legasthenie - in den Klassen 1-10. Handbuch der Lese-Rechtschreib-Schwierigkeiten. Bd. 1: Grundlagen und Grundsätze der Lese-Rechtschreibförderung. Beltz: Weinheim u.a. (6. Aufl.).

Newman, M., et al. (2004): Improving the usability of educational research: Guidelines for the reporting of empirical primary research studies in education. Evidence for Policy and Practice Information and Coordinating Centre (EPPI-Centre)/Social Science Research Unit (SSRU). Institute of Education/University of London.

Nickel, H. (1982): Schuleingangsberatung auf der Grundlage eines ökopsychologischen Schulreifemodells. In. Heller/ Nickel (1982, 81-88).

Nichols, S.L., et al. (2006): High stakes testing and student achievement: Does accountability pressure increase student learning? In: Policy Analysis Archives, Vol. 14, No. 1, 1-180 > epaa.asu.edu/epaa/v14n1/

Nisbet, J. (1978): Procedures for Assessment. In: Becher/ Maclure (1978, 95-112).

Oberholzer, S. (2002): Bedeutung der Schulnoten für den beruflichen Erfolg. Über die Funktionen von Schulnoten, ihre Mängel und ihre Auswirkungen auf den späteren beruflichen Erfolg. FB Wirtschaft und Recht. > http://www.scsh.ch/startseite/themen/maturaarbeiten_05/noten_s_oberholzer.pdf [Abruf: 12.12.2005]

OECD (2005): School factoirs related to quality and equity. Results from PISA 2000. Organization for Economic Co-operation and Development: Paris.

Oelkers, J. (2001): Leistungsbeurteilung als Problem und Chance der Schulentwicklung. > www.impulsmittelschule.ch/themata/noten/2001/leistungsbeurteilung.htm [Abruf: 22.1.2006]

Olechowski, R./Persy, E. (Hrsg.) (1987): Fördernde Leistungsbeurteilung. Jugend und Volk: Wien/München.

Olechowski, R./Rieder, K. (Hrsg.) (1990): Motivieren ohne Noten. Jugend und Volk: Wien/München.

Olechowski, R./Rieder, K. (1991): Verbale Beurteilung in der Schuleingangsstufe - Ergebnisse eine Interventionsstudie. In: Erziehung und Unterricht, 141. Jg., 378-384.

Olechowski, R./Sretenovic K. (Hrsg.) (1983): Schule ohne Angst? Eine empirische Interventionsstudie zur Verminderung der Schulangst. Jugend und Volk: Wien/München.

Osnes (1972) Anm. 106.

Ostrop, G., u.a. (2002): Was denken Kinder über ihre Zeugnisse? In: Valtin (2002a, 49-59).

Ott, U. (2005): Leistungsforderung und Leistungsförderung in Integrationsklassen. In: Götz/Nießeler (2005, 125-160).

Page, E.B. (1992): Ist the world an orerly place? A review of teacher comments and student achievement. In. Journal of Experimental Education, Vol. 60, 161-181.

Panagiotopoulou, A./Brügelmann H. (Hrsg.) (2003): Grundschulpädagogik meets Kindheitsforschung: Zum Wechselverhältnis von schulischem Lernen und außerschulischen Erfahrungen im Grundschulalter. Leske + Budrich: Opladen.

Paradies, L., u.a. (2005): Leistungsmessung und -bewertung. Cornelsen Scriptor, Berlin.

Pekrun, R. (1996): Ziffernzensuren oder Berichtszeugnisse? Drei kritische Anmerkungen zur Annahme unterschiedlicher Wirkungen. In: Benner u.a. (1996b, 253-259).

Persy, E. (1990): Auswirkungen der Leistungsbeurteilung auf Merkmale der Schülerpersönlichkeit. In: Olechowski/Rieder (1990, 129-171).

Peschel, F. (o.J./1999): Leistungsbewertung: Und unsere Beurteilungskriterien stimmen immer noch nicht! Oder: Für eine andere Sichtweise von Produkt- und Prozessorientierung im (offenen) Unterricht. Vervielf. Ms. Universität: Siegen.

Peschel, F. (2002a+b): Offener Unterricht - Idee - Realität - Perspektive und ein praxiserprobtes Konzept zur Diskussion. Teil I: Allgemeindidaktische Überlegungen. Teil II: Fachdidaktische Überlegungen. Schneider Verlag Hohengehren: Baltmannsweiler.

Peschel, F. (2003): Offener Unterricht - Idee, Realität, Perspektive und ein praxiserprobtes Konzept in der Evaluation. Dissertation. FB 2 der Universität: Siegen/Schneider Hohengehren: Baltmannsweiler.

Petersen, P. (1974): Der Kleine Jena-Plan. Beltz: Weinheim/Basel (54./55. Aufl.; 1. Aufl. 1927).

Petillon, H. (2001). Vorwort zu: Maier, M. »Das Verbalzeugnis in der Grundschule«. Verlag Empirische Pädagogik: Landau.

Petzold, K./Woest, V. (Hrsg.) (2003): Leistung und Leistungsbewertung. Beiträge des Zentrums für Didaktik, Bd. 2. Friedrich-Schiller-Universität: Jena.

Pietsch, M. (2005): Schulformwahl in Hamburger Schülerfamilien und die Konsequenzen für die Sekundarstufe I. In: Bos/Pietsch (2005, 255-286).

Pilcher J.K. (1994): The value-driven meaning of grades. In: Educational Assessment, Vol. 2, 69-88.

Pohl, B./Beekmann, A. (2005a): Deutsche Schulen - gut oder ausreichend? Ergebnisse der repräsentativen Lehrer-Befragung durch FORSA. Media-Forschung und -Service für Eltern for Family. Gruner & Jahr: Hamburg.

Pohl, B./Beekmann, A. (2005b): Deutsche Schulen - gut oder ausreichend? Ergebnisse der repräsentativen Eltern-Befragung durch FORSA. Media-Forschung und -Service für Eltern for Family. Gruner & Jahr: Hamburg.

Portmann, R.(1997): Schülerinnen und Schüler beobachten und beurteilen. In: Haarmann 1997, 225-249).

Postlethwaite, T. N./Husén, T. (eds.) (1994): International encyclopaedia of education, Vol. 4. Pergamon Press: Oxford (2nd edition).

Prenzel, M., u.a. (Hrsg.) (2005a): PISA 2003. Der zweite Vergleich der Länder in Deutschland - Was wissen und können Jugendliche? Waxmann: Münster.

Prenzel, M., u.a. (2005b): Vorinformation zu PISA 2003. Zentrale Ergebnisse des zweiten Vergleichs der Länder in Deutschland > http://pisa.ipn.uni-kiel.de [Abruf: 12.02.06]

Preuß, E. (1994): Leistungserziehung, Leistungsbeurteilung und innere Differenzierung in der Grundschule. Bausteine moderner Grundschularbeit - Anregungen und Hilfen. Klinkhardt: Bad Heilbrunn.

Preuß, E. (o.J.): Leistungserziehung und Leistungsbeurteilung in der Grundschule. Ein Lehr- und Arbeitsbuch Medienwerkstatt: Mühlacker.

Preuss-Lausitz, U. (2005): Verhaltensauffällige Kinder integrieren. Zur Förderung der emotionalen und sozialen Entwicklung, Eine empirische Studie und ihre persönlichen Konsequenzen. Beltz: Weinheim/Basel.

Ramseger, J. (1989): Differenzierende Lernerfolgsrückmeldung - eine Chance zur Wiedergewinnung der Pädagogik. In: Die Schleswig-Holsteinische, 43. Jg., Nr. 10, 6-11.

Ramseger, J. (1993a): Für und wider Ziffernbenotung und Verbaleinschätzung. Zwei Wissenschaftler im Meinungsstreit. In: Deutsche Lehrerzeitung, 40. Jg., Nr. 45/1993 (2. Novemberausgabe), 4.

Ramseger, J. (1993b): Ich bleibe dabei: Die Ziffernnoten abschaffen! In: Deutsche Lehrerzeitung, 40.Jg., Nr. 45/1993 (3. Novemberausgabe), 6.

Ratzka, N. (2003): Mathematische Fähigkeiten und Fertigkeiten am Ende der Grundschulzeit - Empirische Studien im Anschluss an TIMSS (Phil. Diss. FB 2 der Universität Siegen). Franzbecker: Hildesheim/Berlin.

Ratzki, A. (2005): »Wir achten die Einzigartigkeit eines jeden Kindes und vertrauen auf sein Potenzial.« Eine Bildungsreise durch Südtiroler Schulen. In: Forum (GEW Köln), November 2005.

Ratzki, A. (2006): Finnland in Südtirol. Die deutschsprachige Region in Italien sorgt für große Überraschung bei PISA 2003. In: e&w, H. 2/2006, 24-25.
Reich, K. (Hrsg.) (2003 ff.): Systemische Benotung. In: Methodenpool. > http://methodenpool.uni-koeln.de [Abruf: 18.12.05]
Reilly, R.R./Chao, G.T. (1982): Validity and fairness of some alternative employee selection procedures. In: Personnel Psychology, Vol. 35, No. 1, 1-62.
Reimers, H. (1991): Länderübersicht zur Leistungsbeurteilung in Zeugnissen der Klassen zwei, drei und vier (Stand: September 1991). In: Grundschul-Zeitschrift, 5. Jg., H. 49, 3.
Reuchlin, M. (1971): Testergebnisse und Zensuren der Klassenlehrer. In: Ingenkamp (1971, 164-167).
Rheinberg, F. (1980): Leistungsbewertung und Lernmotivation. Hogrefe: Göttingen.
Rheinberg, F. (Hrsg.) (1982): Bezugsnormen zur Schulleistungsbewertung. Analyse und Intervention. Jahrbuch für empirische Erziehungswissenschaften. Schwann. Düsseldorf.
Rheinberg, F. (1987): Soziale versus individuelle Leistungsvergleiche und ihre motivationalen Folgen in Lehr-Lernsituationen. In: Olechowsk/ Persy (1987,80-115).
Rheinberg, F. (1998): Bezugsnormorientierung. In: Rost (1998, 39-43).
Rheinberg, F. (1995): Individuelle Bezugsnormen der Leistungsbeurteilung und Motivation im Unterricht. In: Pädagogische Welt 49. Jg., H. 2, 59-62.
Rheinberg, F. (2001): Bezugsnormen und schulische Leistungsbeurteilung. In: Weinert (2001, 59-71).
Rheinberg, F./Peter, R. (1982): Selbstkonzept, Ängstlichkeit und Schulunlust von Schülern. In: Rheinberg (1982, 143-159).
Rhoades, K./Madaus, G. (2003): Errors in standardized tests: A systemic problem. National Board on Educational Testing and Public Policy. Lynch School of Education: Boston. Download > http://www.bc.edu/ research/nbetpp/statements/M1N4.pdf [Abruf: 15.3.06].
Richter, S. (1996): Unterschiede in den Schulleistungen von Mädchen und Jungen. Geschlechtsspezifische Aspekte des Schriftspracherwerbs und ihre Berücksichtigung im Unterricht. S. Roderer: Regensburg > www.uni-regensburg.de/Fakultaeten/phil_Fak_II/ Grundschul_Paedagogik/content/a_sexdif.html
Richter, S./Brügelmann, H. (Hrsg.) (1994): Mädchen lernen ANDERS lernen Jungen. Geschlechtsspezifische Unterschiede beim Schriftspracherwerb. DGLS Reihe »Lesen und Schreiben«. Libelle: CH Lengwil. > www.agprim.uni-siegen.de/maedchenjungen/index.htm
Rieder, K. (Hrsg.) (1990): Motivieren ohne Noten. Wien.
Roeder, P.M. (1997): Entwicklung vor, während und nach der Grundschulzeit. Literaturüberblick über den Einfluss der Grundschulzeit auf die Entwicklung in der Sekundarstufe. In: Weinert/Helmke (1997, 405-421).
Roeder, P.M./Sang, F. (1991): Über die institutionelle Verarbeitung von Leistungsunterschieden. In: Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 23. Jg., H. 2, 159-170.
Röhr, H. (1978): Voraussetzungen zum Erlernen des Lesens und Rechtschreibens. Dissertation. Universität: Münster.
Roos, M. (2000): Evaluationsbericht zum Schulversuch »Erweiterte SchülerInnen- und Schülerbeurteilung«. Befragung der involvierten Gymnasiallehrpersonen, Eltern und SchülerInnen im Auftrag der Luzerner Projektleitung Gymnasialreform. Vervielf. Ms. [am 8.12.2005] direkt über den Verf. bezogen > mroos@dplanet.ch).
Roos, M. (2001): Beurteilen und Fördern in der Primarschule. Eine Untersuchung, wie erweiterte Beurteilungsformen erfolgreich umgesetzt werden können. Rüegger: Chur/Herold: Oberhaching/München.
Roos, M. (2003): Schülerbeurteilung und Schulentwicklung im Fürstentum Liechtenstein. Wissenschaftliche Evaluation. Schlussbericht. Pädagogisches Institut der Universität: Zürich.
Rosemann, B. (1978): Prognosemodelle und Schullaufbahnberatung. Reinhardt: München/Basel.
Rosenfeld, H./Valtin, R. (1997): Zur Entwicklung schulbezogener Persönlichkeitsmerkmale bei Kindern im Grundschulalter. Erste Ergebnisse aus dem Projekt NOVARA. In: Unterrichtswissenschaft, 25. Jg., H. 4, 316-330.
Rosenfeld, H./Valtin, R. (2002): Welche Einstellungen und Erwartungen haben Eltern in Bezug auf die Grundschule? In: Valtin (2002a, 27-36).
Rost, D.H. (Hrsg.) (1998): Handwörterbuch Pädagogische Psychologie. Psychologie Verlags Union: Weinheim.
Roth, P.L., et al. (1996): Meta analyzing the relationship between grades and job performance: A quantitative synthesis. In: Journal of Applied Psychology, Vol. 81, 548 556.
Rotte, R. (ed.) (2006): International perspectives on education policy. Nova Science Publ.: New York (forthcoming).
Ryan, R.M./Deci, E.L. (2000): Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. In: American Psychologist, Vol. 55, 68-78.
Sacher, W. (1996): Prüfen, Beurteilen, Benoten. Klinkhardt: Bad Heilbrunn. Sacks, P. (2004). The Geography of Privilege. In: Encounter: Education for Meaning and Social Justice, Vol. 17, No. 1 (Spring), 7.
Sailer, W. (1998): Lernentwicklungsbericht in der Sekundarstufe I: Abschlussbericht. Schulbegleitforschung Projekt 46. Hrsgg. vom Bremer Landesinstitut für Schule (LIS): Bremen.
Saldern, M.V. (1999): Schulleistung in Diskussion, Schneider Verlag: Hohengehren.
Samson, G. E., et al. (1984): Academic and occupational performance: A quantitative synthesis. In: American Educational Research Journal, Vol. 21, 311 321.
Sauer, J./Gamsjäger, E. (1996): Ist Schulerfolg vorhersehbar? Die Determinanten der Grundschulleistung und ihr prognostischer Wert für den Sekundarschulerfolg. Hogrefe: Göttingen u.a.
Schaub, H. (1993): Weder Noten - noch Berichtszeugnisse: Lernentwicklungsberichte. Von der Zeugnisreform zur pädagogisch-diagnostischen Reform. In: Grundschulzeitschrift, 8. Jg., H. 63, 8-11.
Scheerer, H., u.a. (1985): Verbalbeurteilungen in der Grundschule. Arbeits- und Sozialverhalten in Grundschulzeugnissen in Nordrhein-Westfalen. In: Zeitschrift für Pädagogik, 31. Jg., H. 2, 175-200.
Scheerer-Neumann, G. (1996): Störungen des Erwerbs der Schriftlichkeit bei alphabetischen Schriftsystemen. In: Günther/Ludwig (1996, 2. Hb., 1329-1352).
Scherer, P. (2004): Was »messen« Mathematikaufgaben? - Kritische Anmerkungen zu Aufgaben in den Vergleichsstudien. In: Bartnitzky/ Speck-Hamdan (2004, 270-280).
Schiefele, H. (1960): Sind unsere Noten gerecht? In: Welt der Schule, 12. Jg., 251-257.
Schiefele, U./Wild, K.-P. (2000): Interesse und Lernmotivation. Untersuchungen zu Entwicklung, Förderung und Wirkung. Waxmann: Münster/New York.
Schlattmann, H. (1978): Zur Frage angemessener Methodenstrategien bei der Vorhersage des Studienerfolgs. Phil. Diss. Universität: Saarbrücken.
Schlömerkemper, J. (2001): Leistungsmessung und Professionalität des Lehrerberufs. In: Weinert (2001, 311-321).
Schlottke, P.F./Speidel, E. (1981): Der Schulbericht in der Grundschule. In: Lehren und Lernen, 7. Jg., H. 3, 1-27.
Schmack, E. ( 1978): Zur neuen Schülerbeurteilung in der Grundschule. In: Pädagogische Rundschau 32. Jg., 233-253.
Schmidt, H.-J. (1981): Grundschulzeugnisse unter der Lupe. In: Die Deutsche Schule, 73. Jg., H. 7-8, 486-496.
Schmied, D. (1976): Abiturnoten, Testverfahren und Prognose des Studienerfolgs. Blickpunkt Hochschuldidaktik Nr. 39. Arbeitsgemeinschaft für Hochschuldidaktik: Hamburg.
Schmitt, R. (Hrsg.) (1999): An der Schwelle zum dritten Jahrtausend - BundesGrundschulKongress 1999. Beiträge zur Reform der Grundschule Bd. 105: Grundschulverband - Arbeitskreis Grundschule e.V.: Frankfurt [darin Forum III »Grundschule - Schule der Vielfalt und Gemeinsamkeit. Qualität der Leistung«, 137-196].

Schmitt, R., u.a. (1992): Grundschule in Europa - Europa in der Grundschule. Beiträge zur Reform der Grundschule Bd. 83/84. Arbeitskreis Grundschule: Frankfurt.
Schmitt, R. (Hrsg.) (2001): Grundlegende Bildung in Europa. Beiträge zur Reform der Grundschule Bd. 112. Grundschulverband: Frankfurt.
Schmude, C. (2001): Berichtszeugnisse - unnötiger Aufwand oder auf wendige Notwendigkeit? Evaluation verbaler Leistungsbeurteilungen und differenzielle Entwicklungsverläufe bei Kinder im Grundschulalter. Dissertation an der Humboldt-Universität: Berlin.
Schmude, C. (2002a): Wie werden Berichtszeugnisse realisiert? In: Valtin (2002a, 77-87).
Schmude, C. (2002b): Was ist ein gutes Berichtszeugnis? In: Valtin (2002a, 89-100).
Schmude, C., u.a. (2003). Traumberuf Grundschulpädagoge!? - Beamtenstatus, Freizeit, Versagensängste - Erste Ergebnisse einer Untersuchung bei Studierenden der Grundschulpädagogik an der HU Berlin über die Gründe und Motive ihrer Berufswahl sowie ihrer Ängste und Befürchtungen > http://www2.hu-berlin.de/gsw/downloads/zs_netz.pdf [Abruf: 23.3.06].
Schneider, B. (1985): Lese- und Rechtschreibschwäche. Primäre und sekundäre Ursachen. Dissertation. Fakultät für Biologie der Universität: Freiburg.
Schneider, B. (1985a): Lese- und Rechtschreibschwäche. Primäre und sekundäre Ursachen. Dissertation der Fakultät Biologie. Universität: Freiburg/Hochschulverlag: Freiburg.
Schönwälder, H.-G. (2000): Berufsbelastung von GrundschullehrerInnen. In: Kahlert u.a. (2000, 113-128).
Schrader, F.-W. (1989): Diagnostische Kompetenzen von Lehrern und ihre Bedeutung fr die Gestaltung und Effektivität des Unterrichts. Peter Lang: Frankfurt.
Schrader, F.-W. (1997): Lern- und Leistungsdiagnostik im Unterricht. In: Weinert (1997, 659-699).
Schrader, F.-W./Helmke, A. (2001): Alltägliche Leistungsbeurteilung durch Lehrer. In: Weinert (2003, 43-58).
Schröter, G. (1981a): Zensuren? Zensuren! Allgemeine und fachspezifische Probleme. Burgbücherei Schneider: Baltmannsweiler (3. erw.Aufl.; 1. Aufl.: Henn: Kastellaun 1977).
Schröter, G. (1981b): Zeugnisse muss man richtig lesen - Zensuren richtig beurteilen. In: Schröter (1981c).
Schröter, G. (Hrsg.) (1981c): Schulkinderprobleme. Burgbücherei Schneider: Baltmannsweiler.
Schröter, G. (1982): Was Deutsche von Zensuren halten. In: Westermanns Pädagogische Beiträge, 34. Jg., H. 5, 194-197.
Schröter, G. (1993): Für und wider Ziffernbenotung und Verbaleinschätzung. Zwei Wissenschaftler im Meinungsstreit. In: Deutsche Lehrerzeitung, 40. Jg., Nr. 45/1993 (2. Novemberausgabe), 5.
Schümer, G. (2004): Zur doppelten Benachteiligung von Schülern aus unterprivilegierten Gesellschaftsschichten im deutschen Schulwesen. In: Schümer u.a. (2004, 73-114).
Schümer, G., u.a. (Hrsg.) (2004): Die Institution Schule und die Lebenswelt der Schüler. Vertiefende Analysen der PISA-2000-Daten zum Kontext von Schülerleistungen. Verlag für Sozialwissenschaften: Wiesbaden.
Schuler, H. (1998): Noten und Studien und Berufserfolg. In: Rost (1998, 370 374).
Schuler, H./Stehle, W. (1990): Biographische Fragebogen als Methode der Personalauswahl. Verlag für angewandte Psychologie: Stuttgart (2. unveränderte Aufl.).
Schwark, W., u.a. (Hrsg.) (1991): Beurteilen und Benoten in der Grundschule. Bestandsaufnahme und Anregungen aus der Praxis. Ehrenwirth: München (1. Aufl. 1986).
Schwartz, E. (Hrsg.) (1969): Ausgleichende Erziehung in der Grundschule. Grundschulkongress '69, Bd. 2. Arbeitskreis Grundschule e.V.: Frankfurt.
Schwarzer, R., u.a. (1982): Die Bezugsnorm des Lehrers aus der Sicht des Schülers. Eine Längsschnittstudie zum Einfluß des Klassenlehrers. In: Rheinberg (1982, 161 172).
Schweizerische Koordinationsstelle für Bildungsforschung (1999): Mehr fördern, weniger auslesen. Zur Entwicklung der schulischen Beurteilung in der Schweiz; Trendbericht SKBF Nr. 3, S. 192.
Seel, T. (2002). Studium, Berufseinmündung, beruflicher Werdegang. Ergebnisse einer Befragung von Absolventinnen und Absolventen des Diplomstudiengangs. Diplomarbeit im Fach Psychologie. Universität: Konstanz.
Seidel, B. (2005): Das Risiko punktueller Lernstandserhebungen. Befunde aus einer Fallstudie zur Rechtschreibentwicklung in Klasse 4-6. In: Glatz/Kell (2005, 111-123).
Seidel, B. (Hrsg.) (2006): Einstein, Luke Skywalker und all' die anderen. Kinder und ihre Lernbiografien - Beiträge aus dem Projekt LISA&KO. Universität: Siegen.
Selter, C. (2005): VERA Mathematik 2004. VERbesserungsbedürftige Aufgaben! VERkapptes Ausleseinstrument? In: Grundschule aktuell, H. 89, 17-20.
Severinski, N. (1990): Projekt: Effekte unterschiedlicher Motivierung in der Schuleingangsstufe. Ergebnisse der Untersuchung. In: Olechowski/Rieder (1990, 218-229).
Shepard, L. (1991): Will national tests improve student learning? In: Phi Delta Kappan, Vol. 73, No. 3, 232-238.
Shulman, L. S. (ed.) (1977): Review of research in Education. Vol. 5. Peacock: Itasca, Ill.
Sinn, H.-W. (2006): Alte Ideologien. Über Pisa und die deutsche Drei-Klassen-Gesellschaft. In. Wirtschaftswoche. Nr. 11 von 13.3.06, 250.
Solzbacher, C. (2001): Zwischen Verhalten, Arbeitstugenden und Kompetenzen: Kopfnoten und die »Bewertung« von Schlüsselkompetenzen. In: Solzbacher/Freitag (2001, 77-104).
Solzbacher, C./Freitag C. (Hrsg.) (2001): Anpassen, verändern, abschaffen? Schulische Leistungsbewertung in der Diskussion. Klinkhardt, Bad Heilbrunn.
Sommer, W. (1983): Bewährung des Lehrerurteils. Eine empirische Untersuchung über den Aussagewert des Lehrerurteils über den Bildungs und Berufserfolg. Julius Klinkhardt: Bad Heilbrunn.
Speck-Hamdan, A., u.a. (Hrsg.) (2003): Kulturelle Vielfalt - Religiöses Lernen. Jahrbuch Grundschule, Bd. 4. Kallmeyer: Seelze/Grundschulverband: Frankfurt.
Spiewak, M. (2006): Schlechte Noten. Fehlende Chancengleichheit, verschenktes Bildungspotenzial und die Verlagerung von Kompetenzen auf Länderebene: UN-Sonderberichterstatter Muñoz hat die wunden Punkte unseres Schulsystems benannt. Ein Kommentar. In: Zeit online v. 21.2.2006 > http://zeus.zeit.de/text/online/2006/08/schulsystem [Abruf: 22.2.2006]
Stallmann, M. (1999): Soziale Herkunft und Oberschulübergänge in einer Berliner Schülergeneration. Eine Logit-Analyse von Schülerbögen. In: Zeitschrift für Pädagogik, 36. Jg., H. 2, 241-258.
Starch, D./Elliot, E.C. (1971): Die Verläßlichkeit der Zensuren von Mathematikarbeiten. In: Ingenkamp (1971, 69-77).
Stecher, L. (2003): Schulerleben am Ende der Grundschule. In: Panagiotopoulou/Brügelmann (2003, 55-68).
Steinkamp, G. (1971): Die Rolle des Volksschullehrers im schulischen Selektionsprozeß. In: Ingenkamp (1971, 256-276).
Stepanek, M. (2005): Gute Noten: Schule ködert Schüler mit Geld. Direktor verteidigt Belohnung als leistungs- und motivationsfördernd. pressetext.austria v. 18.11.05.
Stiggins, R. (1999): Assessment, student confidence, and school success. In: Phi Delta Kappan, Vol. 81, No. 3, 191-198.
Strittmatter, A. (2003): Wem Gott ein Amt gibt... Unterrichtsbesuche redlich und hilfreich anlegen. In: Schulmanagement, H. 6/2003, 8-11.
Sundermann, B./Selter, C. (2005): Mathematikleistungen feststellen, beurteilen und fördern. Beschreibung des Moduls 9 für das Projekt SINUS-Transfer Grundschule > www.sinus-grundschule-de/ [Abruf: 13.1.06].

Sundermann, B./Selter, C. (2006): Beurteilen und Fördern im Mathematikunterricht. Gute Aufgaben - Differenzierte Arbeiten - Ermutigende Rückmeldungen. Cornelsen Scriptor: Berlin.

Tent, L. (1998): Zensuren. In: Rost (1998, 580 584).

Textor, A. (2006): Differenzieren und öffnen. Empfehlungen zum Unterricht mit schwierigen Kindern. In: Lernchancen, 9. Jg., H. 49, 19-21.

Theiler, P., u.a. (1987a): Ganzheitliche Schülerbeurteilung. Bericht des Projektleitungsstabes. Erziehungsdepartement: Luzern.

Theiler, P., u.a. (1992): Beurteilen und Fördern. Bericht des Projektsleitungsstabs »Ganzheitlich Beurteilen und Fördern«. Erziehungsdepartement des Kantons: Luzern.

Thiel, O. (2004): Modellierung der Bildungsgangempfehlung in Berlin > http://edoc.hu-berlin.de/dissertationen/thiel-oliver-2005-12-16/PDF/thiel.pdf [Abruf: 24.2.2006].

Thiel, O./Valtin, R. (2002): Eine Zwei ist eine Drei ist eine Vier. In: Valtin (2002a, 67-76).

Thomas, L. (2001): Moderne Kopfnoten - am Beispiel Niedersachsen können erste Ergebnisse und Erfahrungen berichtet werden. In: Schulmanagement, 32. Jg., H. 6, 36-40.

Thüringer Kultusministerium (Hrsg.) (2002a): »Einschätzung zur Kompetenzentwicklung« - ein Beispiel für Schulentwicklung in Thüringen. Kultusministerium: Erfurt.

Thüringer Kultusministerium (Hrsg.) (2003): »Einschätzung zur Kompetenzentwicklung«. Teil II: Praktische Handreichung zum Einschätzungsbogen. Red./Inhalt: Behr, U./Beutel, S.-I./Getschmann, K. u.a. Kultusministerium: Erfurt.

Thurn, S. (1997): Lernen, Leistung, Zeugnisse - eine Schule (fast) ohne Noten. In: Thurn/Tillmann (1997, 63-78).

Thurn, S. (1998): Entwickeln, erstellen, austauschen, reflektieren, vergewissern, bilanzieren, bewerten, weiterentwickeln: 25 Jahre Evaluationsarbeit an Lernberichten. In: Tillmann/Wischer (1998, 74-84).

Thurn, S./Tillmann, K.-J. (1997): Unsere Schule ist ein Haus des Lernens. Das Beispiel Laborschule Bielefeld. Rowohlt: Reinbek.

Tillmann, K. J. (1997): Ist die Schule ewig? Ein schultheoretischer Essay. In: Pädagogik, 49. Jg., H. 6, 6 10 (nachgedruckt in: Baumgart/Lange 1999, 305 314).

Tillmann, K.-J. (2004): Wenig Leistung und viel Selektion: Der PISA-Blick auf deutsche Schulen. Vortrag bei der Jahrestagung der Gesellschaft zur Förderung Pädagogischer Forschung im Mai 2004. Vervielf. als PISA-INFO 02/2006 von der Gewerkschaft Erziehung und Wissenschaft: Frankfurt.

Tillmann, K.-J./Vollstädt, W. (1999): Die Funktion der Leistungsbeurteilung in unterschiedlichen Schulstufen und Bildungsgängen - eine schultheoretische Einordnung. In: Beutel u.a. (1999, 8-39).

Tillmann, K.-J./Vollstädt, W. (2000): Funktionen der Leistungsbewertung. Eine Bestandsaufnahme. In: Beutel/Vollstädt (2000, 27-38).

Tillmann, K.-J./Wischer, B. (Hrsg.) (1998): Schulinterne Evaluation an Reformschulen. Positionen, Konzepte, Praxisbeispiele. Impuls 30. Laborschule an der Universität: Bielefeld.

Travers, C.J./Cooper, C.L. (1996): Teachers under Pressure. Stress in the Teaching Profession. Routledge: London/New York.

Trost, G., u.a. (1998): Evaluation des Tests für medizinische Studiengänge (TMS): Synopse der Ergebnisse. Institut für Test und Begabungsforschung: Bonn.

Trudewind, C./Krohne, W. (1982): Bezugsnorm-Orientierung der Lehrer und Motiventwicklung: Zusammenhänge mit Schulleistung, Intelligenz und Merkmalen der häuslichen Umwelt in der Grundschulzeit. In: Rheinberg (1982, 115-141).

Ubben, L. (1992): Grundschule ohne Noten - Entwicklungslinien zum Entwicklungsbericht in allen vier Grundschuljahren. Vervielf. Ms. Senator für Bildung: Bremen (dazu: Rundverfügung Nr. 65/92).

Ulbricht, H. (1993): Wortgutachten auf dem Prüfstand. Eine empirische Untersuchung zur verbalen Beurteilung in der 1. und 2. Klasse der Grundschule mittels Elternbefragung und Zeugnisanalyse. Münster/New York.

Ullrich, H./Woebcke, M. (1981): Notenelend in der Grundschule. Alternative Beurteilungsformen für die Praxis. Kösel: München.

Ulshöfer, R. (1949): Zur Beurteilung von Reifeprüfungsaufsätzen. In: Der Deutschunterricht, 1. Jg., H. 8, 84-102.

Undeutsch, U. (1971): Die Konstanz des Maßstabes bei Aufnahmeprüfungen. In: Ingenkamp (1971, 233-235).

Valencia, S.W./Au, K.H. (1997); Portfolios across educational contexts: Issues for evaluation, teacher development and system validity. In: Educational Assessment, Vol. 4, 1-35.

Valtin, R. (1999): NOVARA, NOVUS und SABA. Kurzbericht über drei Studien aus der Grundschulforschung. In: Brügelmann u.a. (1999, 110-113).

Valtin, R. (Hrsg.) (2002a): Was ist ein gutes Zeugnis? Noten und verbale Beurteilungen auf dem Prüfstand. Juventa: Weinheim/München.

Valtin, R. (2002b): Die Note als Giftpilz des Haus- und Schullebens? In: Valtin (2002a, 11-16).

Valtin, R. (2002c): Grundschule und Leistungsbeurteilung - Anspruch und Wirklichkeit. In: Valtin (2002a, 139-146).

Valtin, R. (2002d): Informationen zum Projekt NOVARA. In: Valtin (2002a, 147-151).

Valtin, R. (2003): Das Projekt NOVARA. Schulische Sozialisation und Leistungsbeurteilung. In: Speck-Hamdan u.a. (2003, 155-158).

Valtin, R. (2004): »Durch Wiegen wird die Sau nicht fett«. Die Grundschulpädagogin Renate Valtin sagt, warum sie nichts von Schulnoten hält. In: Die Zeit, Nr. 8 v.12.2.04, 71

Valtin, R./Rosenfeld, H. (1997): Zur Präferenz von Noten- oder Verbalbeurteilung - Ein Vergleich Ost- und Westberliner Eltern. In: Zeitschrift für Pädagogik, 37. Beiheft, 293-304.

Valtin, R./Rosenfeld, H. (2002): Welche Erfahrungen, Einstellungen und Wünsche haben Eltern in Bezug auf Notengebung und Verbalbeurteilung? In: Valtin (2002a, 37-47).

Valtin, R./Schmude, C. (2002): Wofür braucht man ein Zeugnis? Zur Funktion von Zeugnissen aus der Sicht von Experten und Betroffenen. In: Valtin (2002a, 17-26).

Valtin, R./Wagner, C. (2002): Wie wirken sich Notengebung und verbale Beurteilung auf die leistungsbezogene Persönlichkeitsentwicklung aus? In: Valtin (2002a, 113-137).

Valtin, R., u.a. (1996): Zeugnisse auf dem Prüfstand. Noten- oder Verbalbeurteilung im Ost-West-Vergleich. In: Benner u.a. (1996a, 122-164).

Valtin, R., u.a. (2004): SchülerInnen und Schüler am Ende der vierten Klasse - schulische Leistungen, lernbezogene Einstellungen und außerschulische Lernbedingungen. In: Bos u.a. (2004, 187-238).

Vierlinger, R. (1999): Leistung spricht für sich selbst. »Direkte Leistungsvorlage« (Portfolio) statt Ziffernzensuren und Notenfetischismus. Dieck: Heinsberg.

Vögeli-Mantovani, U. (1999): Mehr fördern, weniger auslesen: Zur Entwicklung der schulischen Beurteilung in der Schweiz. SKBF/CSRE, Trendberichte Nr. 3. Schweizerische Koordinationsstelle für Bildungsforschung: Aarau.

Vollstädt, W./Jachmann, M. (2000): Leistungsbeurteilung, Zeugnisse und Lernkultur aus der Sicht Hamburger Sekundarschülerinnen und -schüler. In: Beutel u.a. (2000, 111-154).

Wagener, M. (2002): Sind LehrerInnen, die verbal beurteilen, reformorientierter? Zu Unterrichtsorganisation und Rückmeldeverhalten. In: Valtin (2002a, 101-112).

Wagener, M. (2003): Ziffernzensuren oder verbale Beurteilung? Beltz Wissenschaft: Weinheim.

Walcher, U. (1997): Sind Schulnoten und Aufnahmetests Prädiktoren für den weiteren Schulerfolg? Eine empirische Untersuchung. Diplomarbeit. Universität: Wien.

Wallrabenstein, K. (1992): Berichtszeugnisse auch in Klasse 3 und 4 - Erfahrungen aus Hamburg. In: Bartnitzky/Portmann (1992, 120-127).

Wang, M.C., et al. (1993): Toward a knowledge base for school learning. In: Review of Educational Research, Vol. 63, No. 3 (Fall), 249-294.

Wehr, D. (1992): Grundschulkinder schätzen sich und ihre Leistung ein. In: Bartnitzky/Portmann (1992, 61-83).
Weinert, F.E. (Hrsg.) (1997): Psychologie des Unterrichts und der Schule. Hogrefe: Göttingen u.a.
Weinert, F.E. (Hrsg.) (1998): Entwicklung im Kindesalter. Psychologie Verlags Union: Weinheim.
Weinert, F.E. (Hrsg.) (2001): Leistungsmessungen in Schulen. Beltz/ Weinheim.
Weinert, F.E./Helmke, A. (Hrsg.) (1997a): Entwicklung im Grundschulalter. Beltz Psychologie Verlags Union: Weinheim.
Weinert, F.E./Helmke, A. (1997b): Theoretischer Ertrag und praktischer Nutzen der SCHOLASTIK Studie zur Entwicklung im Grundschulalter. In: Weinert/ Helmke (1997a, 457 474).
Weinert, F.E./Schneider, W. (eds.) (1999): Individual development from 3 to 12: Findings from the Munich Longitudinal Study. Cambridge University Press: New York, NY, et al.
Weinert, F.E., u.a. (Hrsg.) (1974): Funk Kolleg Pädagogische Psychologie. Bd. 1 und 2. Fischer Taschenbücher 6115/ 6116: Frankfurt.
Weingardt, E. (1971a): Die Verteilung der Noten von Sexta bis Oberprima. In: Ingenkamp (1971, 205-215).
Weingardt, E. (1971b): Untersuchungen über Korrelationen zwischen Reifeprüfungsnoten und Erfolg auf der Universität. In: Ingenkamp (1971, 252-255).
Weiss, R. (1965a): Über die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen. In: Schule und Psychologie, H. 9/1965, 257-269.
Weiss, R. (1965b): Zensur und Zeugnis. Haslinger: Linz.
Weiss, R. (1966a): Über die Zuverlässigkeit der Ziffernbenotung bei Rechenarbeiten. In: Schule und Psychologie, H. 5/1966, 144-151.
Weiss, R. (1966b): Über die Auswirkung bestimmter Einstellungen auf Zensuren. In: Unser Weg, 166-177.
Weiss, R. (1971): Über die Strenge der Benotung in verschiedenen Unterrichtsgegenständen. In: Ingenkamp (1971, 186-190).
Weiss, R. (1977): Die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen und Rechenarbeiten. Ingenkamp (1977, 104-116).
Weiß, R.A. (1985): Prognostische Validität von Schullaufbahnberatungen in 4. Grundschulklassen. Eine Langzeitstudie. In Greuer-Werner u.a. (1985, 84-107).
Weiß, W.W. (1991): Lehrerbefragung zur Leistungsbeurteilung in der Grundschule. In: Schwark u.a. (1991, 59-102).
Weston, P. (ed.) (1991): Assessment of pupils achievement: Motivation and school success. Swets and Zeitlinger: Amsterdam.
Weuster, A./Scheer, B. (2005): Arbeitszeugnisse in Textbausteinen. Richard Boorberg Verlag: Stuttgart u.a.
Whetton, C., et al. (1991): A report on teacher assessment. School Examinations and Assessment Council: London.
Wilson M (ed.) (2004): Towards coherence between classroom assessment and accountability, 103rd Yearbook of the National Society for the Study of Education. Part II. National Society for the Study of Education: Chicago, Ill.
Winter, F. (1991): Schüler lernen Selbstbewertung. Ein Weg zur Veränderung der Leistungsbeurteilung und des Lernens. Lang: Frankfurt a.M.
Winter, F. (1996): Schülerselbstbewertung. Die Kommunikation über Leistung verbessern. In: Bambach u.a. (1996, 34-37).
Winter, F. (2004): Leistungsbewertung. Eine neue Lernkultur braucht einen anderen Umgang mit den Schülerleistungen. Schneider Hohengehren: Baltmannsweiler 2004.
Winter, F. (2006, im Druck): Wir sprechen über Qualitäten - das Portfolio als Chance für eine Reform der Leistungsbewertung. In: Brunner u.a. (2006, im Druck).
Winter, F., u.a. (Hrsg.) (2002): Leistung sehen, fördern, werten: Neue Wege für die Schule. Klinkhardt: Bad Heilbrunn.
Wolschner, K. (2005): Streit um Zensuren. Die Bildungsdeputation will nur einer von 26 antragsstellenden Grundschulen genehmigen, auf eine Notengebung zu verzichten. In: taz Bremen, Nr. 7826 v. 22.11.05, 22 > www.taz.de/pt/2005/11/22/a0279.nf/text [Abruf: 5.12.05].
Würscher, I./Schmude, C. (1997): Für wen sind Zeugnisse, und zu welchem Zweck werden sie verfasst? Was Zweitkläßler, Lehrkräfte und Eltern darüber denken. In: Deutsche Lehrerzeitung, No. 29-30, 11.
Würscher, I., u.a. (1999): Noten- oder Berichtszeugnisse? Ergebnisse aus dem Forschungsprojekt NOVARA. In: Giest/Scheerer-Neumann (1999, 284-298).
Yung, B. (2002) Same assessment, different practice; professional consciousness as a determinant of teachers; practice in a school-based assessment scheme. In: Assessment in Education, Vol. 9, 97-117.
ZEPF (Hrsg.) (2005): Die wichtigsten Ergebnisse der dritten Befragung des Bildungsbarometers Bildungsbarometer. Newsletter 2/2005. Zentrum für empirische pädagogische Forschung. Universität: Landau. www.bildungsbarometer.de/informationen/downloads.html
Ziegenspeck, J.W. (1999): Handbuch Zensur und Zeugnis in der Schule. Historischer Rückblick, allgemeine Problematik, empirische Befunde und bildungspolitische Implikationen. Klinkhardt: Bad Heilbrunn.
Zielinski, W. (1974a): Die Beurteilung von Schülerleistungen. In: Weinert u.a. (1974, 877, 900).
Zielinski, W. (1974b): Verfahren zur Beurteilung des Unterrichts. In: Weinert u.a. (1974, 901 923).
Zielinski, W. (1980): Lernschwierigkeiten. Ursachen - Diagnostik - Intervention. Kohlhammer: Stuttgart.
Zielinski, W. (1995): Lernschwierigkeiten. Ursachen - Diagnostik - Intervention. Kohlhammer: Stuttgart (1. Aufl. 1980).
Zinnecker, J. (1995): Pädagogische Ethnographie. Ein Plädoyer. In: Behnken/Jaumann (1995, 21-38).

Abbildungen
Abb. 1, S. 10-12: Zeugnisbestimmungen in den Bundesländern. Nach Müller (2005).
Abb. 2, S. 14: Erstellt für diese Expertise von Backhaus (2006).
Abb. 3, S. 32: Bezugsnormen im Vergleich. Zum Modellversuch in zwei Luzerner Gymnasien. In: Roos (2000, 14).
Abb. 4, S. 37: Die größten Ängste der Kinder. In: pro Kids (2004, 25).
Abb. 5, S. 41: Einstellung zu Noten. EFF-Schulbefragung: Ergebnisse der repräsentativen Lehrer-Befragung. In: Pohl/Beekmann (Sept. 2005, 85).
Abb. 6, S. 45: Einstellung zu Noten - nach Klasse des Kindes. EFF-Schulbefragung: Ergebnisse der repräsentativen Eltern-Befragung. In: Pohl/Beekmann (Sept. 2005, 109).
Abb. 7, S. 48: Heilige Kühe des deutschen Schulsystems. In: ZEPF (2005).
Abb. 8, S. 57: Systematische Notengebung. In: Reich (2003).



Niddastraße 52
60329 Frankfurt am Main
Telefon (069) 77 60 06
Fax (069) 7 07 47 80
info@grundschulverband.de
www.grundschulverband.de

Gestaltung
www.hek-design.de
Dr. Helmuth Krieg, Frankfurt am Main
Druck und Bindung
Beltz Druckpartner GmbH & Co. KG, 69502 Hemsbach

3. aktualisierte Auflage 2014

Bestell-Nr. 2040
ISBN 978-3-941649-12-5