Sind Noten nützlich – und nötig?

Ziffernzensuren und ihre Alternativen im empirischen Vergleich

Auszug aus:

Eine Expertise der Arbeitsgruppe Primarstufe an der Universität Siegen[1]
- Hans Brügelmann mit Axel Backhaus, Erika Brinkmann (Gast), Hendrik Coelen,
Thomas Franzkowiak, Simone Knorre, Barbara Müller-Naendrup, Elisabeth Oser, Sara Roth
-
im Auftrag des Grundschulverbands e.V., Frankfurt

[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).

1. Mit welchen Verfahren werden Leistungen erfasst?[2]

Die Fundiertheit von Beurteilungen hängt von ihrer Datengrundlage, diese von den eingesetzten Instrumenten ab. Deren Qualität wiederum wird üblicherweise über drei Gütekriterien bestimmt[3]:

- 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.[4].

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 psycho-

---

[2] 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).

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

[4] 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 brü), „...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;
- should be comprehensive and take account of the wide variety of aspects of education."

logischen 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[5] 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 (subjektiv) unterschiedliche Stressresistenz und Konzentrationsfähigkeit beeinflusst[6].

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[7]. Die Diskussion über die Bildungsstandards zeigt aber ein hohes Maß an Uneinigkeit, was als Mindest- oder Re-

---

[5] Vgl. Schneider (1985); Brügelmann (1994a, 206-207).

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

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

gelleistung 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[8]:

- 58% der SchülerInnen erreichen in anderen Tests als TIMSS andere Ergebnisse. Selbst zwischen zwei verschiedenen Tests des gleichen Grundtyps („Textaufgaben" im 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[9]. Die unterschiedlichen Einschätzungen der Lesefähigkeit deutscher SchülerInnen nach PISA und DESI[10] 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. „Legastheni-

---

[8] 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).

[9] 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".

[10] Vgl. zu den Ergebnissen der Studie „Deutsch Englisch Schülerleistungen International" (DESI) Klieme u. a. (2006).

kerInnen", 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[11].

Aber auch für das Lehrerurteil, das die Noten bestimmt, haben empirische Studien Validitätsprobleme aufgezeigt: Bei Aufsätzen beeinflussen sowohl der Umfang[12] als auch die Zahl der orthografischen Fehler und die Qualität der Handschrift die Note[13].

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 Ergebnissen von Leistungstests in den entsprechenden Fächern übereinstimmten. Damit ist aber ein problematischer Maßstab gesetzt[14], 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[15]. Das einzige unabhängige Kriterium ist ihre Vorhersagekraft, bezogen auf zukünftige Leistungen. Diese aber erweist sich als sehr begrenzt (s. dazu unten ➔ 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 Zif-

---

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

[12] Kürzere Aufsätze werden generell schlechter benotet (Baurmann 1977).

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

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

[15] S. unten ➔ Kap. 1.1.2 .

fer 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. unten ➔ 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[16]. 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[17] 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[18]. Bei PISA-2000 lag die Korrelation zwischen Mathematiknote und curriculumnahem Mathematiktest bei .32 über die verschiedenen Schulformen hinweg und bei .43 innerhalb der Bildungsgänge[19].

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.

---

16 Vgl. etwa Lehmann (1999).
17 U. a. bei Mreschar (1985, 51).
18 Vgl. u. a. Ingenkamp (1971c); Thiel/ Valtin (2002); Brügelmann (2003); Pietsch (2005).
19 Vgl. Baumert u. a. (2003, 325).

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[20]. LehrerInnen differenzieren unterschiedliche Lernstände also weitgehend zutreffend[21], 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[22].

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?

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[23]:

> „Die zur Analyse der Übereinstimmungsvalidität zwischen Verbalzeugnis und Schulleistungsvariablen durchgeführten Korrelationsanalysen belegen insgesamt einen schwachen Zusammenhang, d.h. eine geringe Übereinstimmungsvalidität. Mit der Regressions-

---

[20] 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).

[21] .... unter der Prämisse, dass man die Testergebnisse als Maßstab anerkennt.

[22] Vgl. Merkens (2004).

[23] Maier (2001, 228).

analyse 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)* [...} *Ziffernoten 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[24], 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[25] 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.

---

[24] Vgl. Mreschar (1985, 65).

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

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)[26]. 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%[27]. 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.

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

---

[26] Hier zusammengefasst nach Brügelmann (2005a, 167).

[27] Vgl. zusammenfassend zum Prognoserisiko von Risikoprognosen: Brügelmann (2005c).

„ökologischen Modell" hat Nickel[28] 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[29].

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[30]:

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

Für Schulnoten ergeben sich ähnliche Werte; mittelt man sie über verschiedene Fächer, kommt man sogar auf Werte von .80 und höher[31].

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[32]. 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.

---

[28] Vgl. Nickel (1982).

[29] 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).

[30] Vgl. Weinert/ Helmke (1997b, 467).

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

[32] Brügelmann (2005c, 150).

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[33]. 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[34]. 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[35].

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[36]. 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.

---

[33] 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.

[34] 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 unten ➔ Kap. 1.2.1 und 7.

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

[36] 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).

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 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."[37]

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."[38]

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[39] 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:

---

[37] 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.

[38] Block (2006, 3)

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

„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 Auslese-verfahren 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[40]. 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,

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

Verschiedene ForscherInnen fanden in verschiedenen Ländern[41] Korrelationen von .30 bis .50 für die Vorhersage des Studienerfolgs und Schuler (1998) nennt .41 als Mittelwert[42] 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 und 1. Lehramtsprüfung an Pädagogischen Hochschulen Korrelationen zwischen .29 und .49, zwischen Abiturdurchschnitt und Vordiplom in Physik ein Koeffizient von .37..." [43]

Wer stattdessen auf Tests setzt, sollte aber vorsichtig sein. Gemittelte Abiturnoten sind vorhersagekräftiger als eignungsdiagnostische Verfahren[44]. Und in den USA hat die University of California in Los Angeles nach langen Jahren den fest etablierten SAT als Auswahlinstrument aufgegeben[45]:

> „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."[46]

---

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

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

[43] Zielinski (1974b, 889).

[44] 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.

[45] 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."

[46] Sacks (2004, 7).

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[47]:

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

Die Korrelation zwischen Abiturnote und dem späteren Berufserfolg fällt allerdings auf .10 und liegt damit schon fast im Zufallsbereich[48]. 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[49]. 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[50]:

> „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

---

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

48 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 Stellehierarchie? Einkommen? Zufriedenheit? ...

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

50 Fend (2006, 53).

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[51] 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[52]. 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 .

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

.50+ berufliche 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[54].

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.

---

[51] S. dazu auch die vorhergehende Anmerkung.
[52] Vgl. Schuler (1998, 372).
[53] Auf einer 5er Skala jeweils von 1 bis 5 bewertet.
[54] Vgl. Landmesser u. a. (2003) und unten ➔ Kap 4.4 .

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 Ziffernnoten 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[55]. 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[56], 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[57], 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[58].

---

55 Vgl. u. a. Harlen (2004a, 6-7).
56 S. dazu die Studien im Kap. „Objektivität" (unten ➔ 1.2.3).
57 Baumert u. a. (2001, 43); Artelt u. a. (2001a, 97-101).
58 Artelt u. a. (2001a, 100 vs. 102).

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[59]. 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.

In einer Studie von Weiss (1965)[60] sollten 92 LehrerInnen nur die Rechtschreibung in zwei kleinen Aufsätzen von ViertklässlerInnen benoten. Auch hier streuen die Bewertungen über fünf Notenstufen:

---

[59] Vgl. die Zusammenfassung bei Mreschar (1985, 47). Vgl. zur fehlenden Objektivität von Aufsatzzensuren auch Faigel (1973).

[60] Zusammengefasst bei (Zielinski 1974a, 889).

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[61]:

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 (vgl. unten ➔ 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."

[61] Aktuelle Versuche mit Studierenden bestätigen diese Ergebnisse auch heute, vgl. Mühlhausen/ Wegner (2006, Kap. 14). Für Geometrie und andere Fächer fanden schon Starch/ Elliot (1913) ähnliche Verteilungen.

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[62]:

	Note besser als Testleistung	Note wie Testleistung	Note schlechter Als Testleistung
Beliebteste SchülerInnen	50 %		16 %
Durchschnitt	31 %		34 %
Unbeliebeste SchülerInnen	19 %		50 %

Systematische Verzerrungen wurden auch für die Merkmale: Verhalten, Alter, soziale Herkunft, Geschlecht und ethnische Zugehörigkeit nachgewiesen[63].

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[64]. 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[65] 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 kogniti-

---

[62] 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.

[63] Vgl. außer den im Folgenden zitierten Studien: Baurmann (1971); Bennett u. a. (1993).

[64] Weiss (1965b; 1971, 98-101); Stallmann (1990, 253) und Mühlhausen/ Wegner (2006, Kap. 14) haben diesen Befund erneut bestätigt.

[65] Vgl. Baumert/ Schümer (2001, 357).

> ven 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"[66]

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[67] 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[68]. 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[69]. 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.

---

[66] Bos u. a., (2004, 213).

[67] So auch Hadley (1954), der zugleich feststellte, dass Mädchen auch eher als „sympathisch" eingestuft wurden (s. oben).

[68] Vgl. zusammenfassend: Richter/ Brügelmann (1994) und Richter (1996).

[69] 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).

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 pädagogischen 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[70], 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[71]. Das ist offenkundig bei Übersetzungen, wie Untersuchungen zu PISA belegen. Die

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

[71] Vgl. zum Problem der „Operationalisierung" ausführlicher Brügelmann (1977).

Leistungen von SchülerInnen differieren nämlich je nachdem, ob eine Aufgabe aus dem Testpool des betreffenden Landes stammt oder in deren Sprache übersetzt worden ist[72] .

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.

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[73], 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."[74]

---

[72] Vgl. Baumgarten u. a. (2005, 101-102).
[73] U. a. Lehmann (1990; 1994); Beck/ Hofen (1991).
[74] Grzesik/ Fischer (1984, 193; s. a. 184-185, 215)

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[75], aber immer noch deutliche Differenzen zu erwarten.

Das zeigt sich bei der Beurteilung von pädagogischen Prozessen generell. Metz (1982)[76] 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[77]:

- 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,

---

[75] 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.

[76] Ref. bei Strittmatter (2003, 11).

[77] Hargreaves u. a. (1996); Frederiksen/ White (2004).

dass auch die Leistungen der SchülerInnen deutlich besser werden[78], und zwar profitieren vor allem leistungsschwächere SchülerInnen von einer differenzierteren Rückmeldung[79].

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)

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

[78] 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.

[79] Vgl. Stiggins (1999, 193).

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 Messun-

gen 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. oben 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[80]. Auch der Kontext der Beurteilung spielt eine Rolle: Nach einer guter Arbeit wird eine schlechte noch schlechter beurteilt (Birkel 1978/ 1984)[81].

Aber es sind nicht nur die Aufsätze, deren Beurteilung für den Einfluss von Randbedingungen anfällig ist. Dicker (1973)[82] 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[83].

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.

---

[80] Baurmann (1975).

[81] Man kann solche Reihungs- und Kontrasteffekte auch als Einschränkung der Objektivität interpretieren, s. oben ➔ 1.2.1.

[82] Zusammengefasst bei Zielinski (1974a, 888).

[83] Eells (1930/1971).

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 [84]. Bei der Durchführung in einer anderen Form (PC vs. Papier-und-Bleistift) sank die Korrelation aber schon deutlich auf .70 [85]. 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[86].

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 Situa-

---

[84] Vgl. Brügelmann (2003c, 6).
[85] Vgl. Backhaus/ Moskopp (2006, 4).
[86] Vgl. als ein Beispiel unter vielen Seidel (2005; 2006).

tion zuminimieren (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[87] (➔ Kap. 6.5).

[…]

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[88]:

- 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[89].

---

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

[88] Auswahl und deutsche Zusammenfassung: Hans Brügelmann.

- Kontinuierliche wechselseitige Abstimmung von Kriterien im Austausch über konkrete Bewertungsversuche hilft LehrerInnen, Klarheit über die Ziele von Unterricht und darauf bezogene Beurteilungskriterien zu gewinnen[90].

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 Anregungspoten-

---

[89] 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).

[90] Zu dem Ergebnis, dass punktuelle Fortbildungen nicht ausreichen, kommt auch Inckemann (2004) aufgrund ihrer Versuche zum Schriftspracherwerb.

tial 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[91]:

- 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[92], 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 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.

---

[91] 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).

[92] Vgl. zur hohen Bedeutung dieser Kontextbedingungen, die wesentlich stärker für Leistungsunterschiede zwischen Schulen verantwortlich sind als schulinterne Bedingungen: OECD (2005, 88).

Leistungen *und* ihre Beurteilung werden überlagert durch andere Faktoren, vor allem durch den Einfluss der sozialen Herkunft, den sie doch ersetzen sollen (vgl. oben ➔ 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[93]. 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. [...]

---

[93] 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.

> 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)[94]

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

---

[94] 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).

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 Förderung 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[95]: 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[96]. Erfahrungen in europäischen Ländern belegen darüber hinaus[97], 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[98]. 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 Zifernnoten, stattdessen mit individuellen Aufgaben in offeneren Unterrichtsformen und ei-

---

[95] 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).

[96] 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.

[97] Vgl. etwa die sorgfältige Evaluation des Modellversuchs „Schülerbeurteilung und Schulentwicklung" in Liechtenstein: Roos (2003, 135, 138-139)

[98] Vgl. oben ➔ Kap. 0.5 und die bereits dort zitierten: Höllrigl/ Meraner (2005); Leitzgen (2005); Meraner (2005); Ratzki (2005; 2006).

ner Bewertung, die sich am persönlichen Lernfortschritt orientiert[99]. Erfolgreicher Unterricht ist also auch mit weniger Leistungsdruck möglich; und Schulsysteme können lernen, ohne Selektion auszukommen.

8. Literaturnachweise und weiterführende Literatur[100]

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.

---

[99] 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.

[100] 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.

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.

D. Benner, 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", Heft 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 exmainations. In: Journal of Educational Psychology, Vol. 21, 48-52.

Eells, W. C. (1971): Die Zuverlässigkeit wiederholter Benotung von aufsatzähnlichen 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 zifferfreie 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).

Grace, C./ Shores, E. F. (2005): Das Portfoliobuch für Kindergarten und Grundschule. Verlag an der Ruhr: Mülheim.

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 Nordrhein-Westfalen. Westdeutscher Verlag: Opladen.

Günther, H./ Ludwig, O. (Hrsg.) (1996): Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch. 2. Halbband. 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. M./ 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 Überblick 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. M./ 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, Band 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 predictors 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 Band 6. Luchterhand: Neuwied.

Jung, J. (2005): Formen, Prinzipien und Probleme der Leistungsburteilung. 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 öffentlichen 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. (2004): Konsequenzen von Leistungsgruppierungen. Waxmann: Münster.

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 (GBI. S. 324; K.u.U. S. 449), zuletzt geändert durch: Verordnung vom 23. März 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: Vergleoch diagnsotsocher Entscheidungen von Grundschullehrer/innen und Lehramtsstudierenden. Poster. Universität: Dortrmund =>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-

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 erziehungswissenschaftliche 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 Lesekompetenzentwick-

lung 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.n

Mühlhausen, U./ Wegner, W. (2006): Erfolgreicher Unterrichten?! Eine erfahrungsfundierte Einführung in die Schulpädagogik (Begleit-DVD mit Videoszenen und Online-Übungen zur Unterrichtsanalyse). Schneider Verlag Hohengehren: Baltmannsweiler.

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-Holsteinische43. 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 Shcvulunlust 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, Volume 17, Number 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 aufwendige 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. (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).

Schneider, B. (1985): Lese- und Rechtschreibschwäche. Primäre und sekundäre Ursachen. Dissertation. Fakultät für Biologie der Universität: 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.

Schründer-Lenzen, A. (Hrsg.) (2006): Risikofaktoren kindlicher Entwicklung. Migration, Leistungsangst und Schulübergang. VS Verlag für Sozialwissenschaften: Wiesbaden.

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.).

Schumann-Erny, S. A. (2003):Brauchen wir neue Zeugnisse? Eine empirische Untersuchung zu Zeugnissen der Realschule in Baden-Württemberg – Aussagen und Anforderungen aus der Sicht von Schülern, Eltern und Arbeitgebern. Logos: Berlin.

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 Sc.hulnoten 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.

Zeinz, H./ Köller, O. (2006): Noten, soziale Vergleiche und Selbstkonzepte in der Grundschule. In: Schründer-Lenzen (2006, 177-190).

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).