Lind, G. (2009). Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. [America as role model? Desired and undesired consequences of evaluation.] In: T. Bohl & H. Kiper., eds., Lernen aus Evaluationsergebnissen – Verbesserungen planen und implementieren, pp. 63-81. Bad Heilbrunn: Julius Klinkhardt.
Der vollständige Aufsatz mit Internetquellen ist hier erhältlich: **Link zum vollständigen Artikel**

Inhaltsverzeichnis

3 Lernen aus Lernstandserhebungen und Schulleistungsstudien

4 Lernen aus Evaluationsergebnissen - Vom Wissen zum Handeln

5 Widerstand, Belastung und Kritik

8 Fazit

Georg Lind

Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen[1]

Vorbemerkungen

Oft genügt ein Blick über den Atlantik, um zu wissen, welche Trends in der nächsten Zeit bei uns angesagt sind und, leider auch, welche Fehler. Sich an einem Vorbild orientieren kann bedeuten, dass man von den Erfahrungen dieses Vorbilds lernt, das heißt, bewährte Maßnahmen übernimmt und Fehler vermeidet. Leider fehlt es uns oft an ausreichender Information über die USA, so dass nur wahrgenommen wird, was die großen Medien bei uns – oft sehr einseitig – berichten.

Dies gilt auch für die Evaluation im Bildungsbereich. Als besonders vorbildlich gilt bei uns die in den USA seit vielen Jahren betriebene Politik, Schulen und Lehrer durch standardisierte Schulleistungstests zu besseren Leistungen anzutreiben. Diese Verwendung von Evaluation hat vor vierzig Jahren in den USA ihren Anfang genommen, 1965, als der US-Präsident Johnson seinen "Krieg gegen die Armut" erklärt und das Head start-Programm per Gesetz eingeführt hat und der Autor, was sein Interesse am amerikanischen Schulsystem begründete, als Austauschschüler in den USA zum ersten Mal einen multiple-choice-Test ausgefüllt hatte. Spätestens mit der Veröffentlichung der Ergebnisse von PISA 2000 (vgl. Deutsches PISA-Konsortium 2001) ist

[1] Für wertvolle Hinweise bei der Erstellung dieses Kapitels möchte ich Gerald W. Bracey, George Madison University, Virginia, und Hans Brügelmann, Universität Siegen, herzlich danken.

Das vollständige Manuskript dieses Beitrags mit dem zusätzlichen Abschnitt ‚Programmevaluation' kann im Internet abgerufen werden: http://www.uni-konstanz.de/ag-moral/b-publik.htm

diese Entwicklung auch bei uns im bildungspolitischen Diskurs angekommen.
Bei uns wird oft übersehen, dass es neben dieser Verwendung von Evaluation noch eine andere Verwendung gibt, nämlich Evaluation als Mittel der Sicherung und Steigerung der Qualität von bildungspolitischen Programmen und Unterrichtsmethoden. Auch diese Verwendung von Evaluation wurde hauptsächlich in den USA entwickelt, und hat dort eine lange Tradition. Absichten und Ziele beider Arten von Evaluation in den USA sind nicht ohne deren historischen und politischen Hintergrund zu verstehen und auf ihre Tauglichkeit für unser Schulsystem hin einzuschätzen.

Bildung und Evaluation in den USA

Historisch-politisch gesehen, gibt es Verbindendes und Trennendes zwischen den USA und Deutschland. *Horace Mann* (1796-1859), der amerikanische Humboldt, hat das Schulsystem der USA nach preußischem Vorbild reformiert. Aber während bei uns zwischen *Bildung* (für die Oberschicht und die Führungselite) und *Ausbildung* (für den arbeitenden, Güter schaffenden Rest der Bevölkerung) strikt getrennt wird, was seinen Niederschlag in unserem weitgehend nach sozialem Status gegliederten Schulsystem gefunden hat, gibt es in den USA für beides nur einen Begriff – *Education* – und für alle sozialen Schichten in der Regel nur eine Schule (eine Regel, von der es natürlich viele Ausnahmen gibt: Privatschulen, *Home schooling* und seit ein paar Jahren öffentlich finanzierte, aber privat gemanagte *Charter schools*).
Während bei uns Bildung am Gymnasium durch einen engen Kanon von Fächern definiert ist, hat die traditionelle amerikanische *Middle School* und *High School* ein breites Lernangebot, das auch Rhetorik- und Führerscheinkurse umfassen kann. Während sich die deutsche Bildungstheorie in der Tradition von Kant, Herbart und der Reformpädagogik bis heute vorwiegend an idealistischen und romantischen Vorstellungen vom gebildeten Menschen orientiert hat (und auch noch immer daran zu orientieren scheint), sehen Amerikaner Bildung entweder als Privatsache (das trifft häufig auf Mitglieder orthodoxer Religionsgemeinschaften zu) oder als ein Mittel der Gesellschaftspolitik (vgl. z. B. Dewey 1964, zuerst im Jahr 1915 publiziert).
Den verschiedenen Interessen und gesellschaftlichen Bedürfnissen für Bildung wird durch innere Differenzierung in Gleise *(Tracks)* entsprochen: Gleise für Collegevorbereitung, Handelsausbildung und gewerbliche Ausbildung. Die soziale Selektion setzte mit dem Zugang zu den Colleges ein, die

mit Zulassungstests und hohen Studiengebühren dafür sorgten, dass die verschiedenen sozialen Schichten und ethnische Minderheiten weitgehend unter sich blieben, auch wenn durch Stipendien an Hochbegabte versucht wurde, die soziale Durchlässigkeit zu erhöhen.

Durch die Struktur der lokalen Finanzierung ergeben sich zum Teil große regionale Unterschiede bei den Bildungschancen der Kinder, die durch ökonomische, ethnische und regionale Ungleichheiten bedingt sind, die nur in wenigen Bundesstaaten durch staatliche Mittelzuweisungen abgemildert wurden (vgl. Darling-Hammond/ Ancess, 1996). Das Geld, das einzelnen Schulen pro Schüler jährlich zur Verfügung steht, schwankt stark. In den 1980er Jahren lagen die Mittel zwischen 2000 Dollar in großstädtischen Minderheiten-Ghettos und 20.000 US-Dollar in vornehmen Vororten, wobei die direkten Spenden reicher Eltern an die Schule ihrer Kinder nicht eingerechnet sind (vgl. Kozol 1991). Dieser Trend dürfte sich bis heute noch verstärkt haben. Ein immer bedeutsamerer Grund für die wechselseitige Abhängigkeit von Wohlstand und Schulqualität ist die Koppelung der Grundstückpreise an die (meist öffentlich zugänglichen) Testleistungsniveaus einer Schule (vgl. Cullen u. a. 2000). Eltern, die um die spätere Karriere ihrer Kinder besorgt sind, kaufen sich ein Haus in dem Schulbezirk, den sie für ihre Kinder ausgesucht haben, um längere Anreisen und Schulgeld zu sparen. Bezirksfremde müssen in der Regel hohe Schulgebühren bezahlen.

Das Project Head Start, das es trotz vieler Kritik heute noch gibt, bietet benachteiligten Kindern und Müttern acht Wochen lang vor der Einschulung Förderkurse an. Heute wird es auch schon Drei- bis Fünfjährigen angeboten. Die Kosten des Programms hatte die Johnson-Regierung den amerikanischen Steuerzahlern dadurch annehmbar gemacht, dass diese Maßnahme mit einer (damals aber noch milden) Kontrolle ihrer Wirksamkeit verbunden wurde.

Head Start war somit nicht nur das erste große Bildungsprojekt der amerikanischen Bundesregierung, sondern auch die erste große, von einer Regierung veranlasste Evaluationsmaßnahme. Es gab auch vorher schon umfangreiche Evaluationsprojekte, aber die Evaluation von Head Start war das erste, das durch politische Vorgaben bestimmt war und das der Durchsetzung und Rechtfertigung eines politischen Ziels diente. Auf der Basis des ‚Elementary and Secondary Education Acr' (ESEA) der Johnson-Regierung folgten bis heute zwei weitere Großprojekte der US-Regierung: Das „Nation-At-Risk" - Programm unter der Ägide von Ronald Reagan und das von den beiden großen Parteien getragene „No-Child-Left-Behind"-Gesetz (NCLB) unter Präsident Bush. Das NCLB-Gesetz macht die Zuteilung von Bundesmitteln davon abhängig, dass die öffentlichen Schulen jedes Jahr Vergleichstests durchfüh-

ren. Inzwischen sind viele Schuldistrikte dazu übergegangen, diese Tests auch von Schulen zu verlangen, die keine NCLB-Mittel beanspruchen (vgl. Bracey 2005). Zudem wurden in den letzten Jahren an vielen Schuldistrikten der normale Abschluss *(High school graduation)* und die Noten durch Schulabgangstests ersetzt.

Das NCLB-Gesetz verlangt von den Schulen, dass ihre Schüler „angemessene jährliche Fortschritte" (*Adequate yearly progress* oder AYP) in Englisch und Mathematik machen, so dass nach zehn Jahren (2013-14) 100% der Schüler in den beiden Fächern „proficient" sind, was bedeutet, dass sie diese Fächer gut beherrschen – ein sehr hoch gestecktes Ziel, das von renommierten Bildungsforschern in den USA als unrealistisch angesehen wird: „Trends on NAEP over the past several years provide ample reasons to doubt that the 100% proficiency goal is obtainable even with the best of efforts or the belief that the rate of improvement would be twice as great in the future as it has been in recent years" (Linn 2008).

Um dieser Norm Nachdruck zu verleihen, sieht das NCLB-Gesetz effektiv hohe Strafen vor, auch wenn das im Gesetz nirgends so genannt wird (vgl. Bracey 2005, S. 7): Wer das Ziel eines "Adequate yearly progress" zwei Jahre hintereinander nicht erreicht, verliert Zuschüsse und muss seinen Schülern ermöglichen, in eine andere Schule zu wechseln und die Schulmittel, die ihm von Staats wegen zustehen, dorthin mitzunehmen. Eine Schule, die das Ziel dreimal hintereinander verfehlt, muss ergänzenden Unterricht anbieten, wozu sie die teuren Dienste von meist kommerziellen Firmen beanspruchen muss. Gerade Schulen in sozialen Brennpunkten mit niedrigem Budget sind hiervon hart betroffen. Im Unterschied zu den Schulen werden diese privaten Tutoren kaum zur Rechenschaft gezogen, wenn sie schlecht arbeiten. Nach viermaligem Zielversagen droht der Schule die Ablösung der gesamten Leitungsstruktur oder gar die Auflösung.

Zwei Zielsetzungen der Evaluation

Evaluation wird in den USA überwiegend als legitim akzeptiert. Diese Akzeptanz beruht auf der weit verbreiteten Überzeugung, dass wissenschaftliche Methoden der Produktion und Evaluation den USA zu einer wirtschaftlichen Vorrangstellung in der Welt verholfen haben und dass dieselben Prinzipien auch in der Bildung zum Erfolg führen müssten. Jedoch ist die Frage, welchen Zielen Evaluation dienen soll und was genau die 'wissenschaftlichen Methoden' sind, die zu diesem Erfolg beigetragen haben, schon lange Ge-

genstand vielfacher Auseinandersetzungen. Bei der Frage, was denn die vierzig Jahre staatlich standardisierter Evaluation im Bildungswesen in den USA erbracht haben, brechen die Gegensätze so stark auf wie nie zuvor.

Die Gegensätze verlaufen – im Bildungsbereich wie in der Wirtschaft (vgl. Deming 1994; Kohn 1999) – vor allem zwischen zwei Zielsetzungen von Bildungsevaluation, zwischen Personen- und Programm-Evaluation. Auf der einen Seite besteht die Vorstellung, dass man durch die Messung von Schulleistungen mittels standardisierter Tests in Verbindung mit Strafen und Belohnungen Schulleitungen zu einer effizienteren Schulverwaltung, Lehrer zu besserem Unterricht und Schüler zu mehr Lernen zwingen kann. „The prevailing theory of action behind accountability ratings and testing is that schools and students who are held accountable to these measures will automatically increase educational output: Educators will try harder; schools will adopt more effective methods; and students will learn more." (Heilig/Darling-Hammond 2008, S. 75) Dieser Ansatz wird in den USA als *High-stakes Testing* (was man etwas umständlich als 'sanktionsbewehrte Leistungstests' übersetzen kann) oder in unserer Systematik als *Personen*evaluation bezeichnet, was besagt, dass letztlich Personen oder Personengruppen das Ziel der Evaluation und der daran geknüpften politischen Entscheidungen sind. Sie werden belohnt oder – im häufigeren Fall – bestraft, wobei die Strafen von einer Abmahnung bis hin zu Schulausschluss, Gehaltskürzungen, Kündigungen, Schließungen oder Verkauf der Schule an private Schulträger reichen können (vgl. Kreitzer u. a. 1989; Sacks 1999; Bracey 2002; Nichols/Berliner 2006; Heilig/Darling-Hammond 2008).

Bei Personenevaluation wird schon aus Kostengründen die *Outcome*-Evaluation mittels standardisierter Tests bevorzugt, bei der schnelle Ergebnisse für politische Entscheidungen zu erwarten sind. Da zur Umsetzung dieser Vorstellung ein gigantisches Testprogramm finanziert werden muss, das jedes Jahr hohe direkte und indirekte Kosten verursacht, müssen die Tests zudem möglichst einfach anzuwenden und auszuwerten sein (vgl. Bracey 2005). Bevorzugt wird das *Multiple choice* (Auswahlantworten-) Format, bei dem die richtige Antwort unter falschen geraten werden muss und das nur richtige oder falsche Lösungen kennt. Ausgeblendet bleiben bei diesem Ansatz der Evaluation a) die Lernvoraussetzungen der Schüler (dieses Manko wird neuerdings durch die Messung von Lern*zuwachs* vermieden – aber nur teilweise, weil der Lernzuwachs nicht individuell, sondern nur auf aggregierter Ebene gemessen wird), b) die Analyse des Lösungswegs, was bei komplexen Aufgaben, zu deren Lösung eine ganze Bedingungskette erfüllt sein muss, zum Verlust aller Punkte führt, wenn nur ein einziges Glied in der

Bedingungskette für eine richtige Lösung fehlt[2]; und c) die Lehrbedingungen der Schule, z. B. die Ausbildungsqualität der Lehrer und die zur Verfügung stehenden Finanzmittel (vgl. Kozol 1991; Nye u. a. 2004).

Auf der anderen Seite besteht die Vorstellung, dass das Bildungssystem eher durch die systematische Evaluation der *Effektivität* und *Effizienz* von Unterrichtsmethoden und bildungspolitischen Programmen verbessert werden kann, die sich auf experimentell überprüfte Theorien stützen (vgl. Schoenfeld 1999; Shepard 2002). Diese Vorstellung beruht auf der Überzeugung, dass Menschen nicht zum Lernen gezwungen werden müssen, da bei allen höheren Lebewesen der Wunsch zu lernen angeboren ist und daher eine äußere 'Motivation' durch finanzielle Anreize, Konkurrenz und Testdruck nicht notwendig, sondern eher kontraproduktiv ist (vgl. Deming 1994; Deci 1995; Deci u. a. 1999; Kohn 1999; auch Spitzer 2002). Von diesem Standpunkt muss Evaluation – je nach Stellung im System Schule – Fragen beantworten wie: „Haben wir die richtige bildungspolitische Entscheidung getroffen?" „Sind die von mir eingesetzten Unterrichtsmethoden und didaktischen Prinzipien effektiv und auch effizient?" „Wende ich die beste Lernmethode an?" Wenn die Programmevaluation frei von äußeren Sanktionen gehalten wird (Selbstevaluation), dann trägt sie nicht nur zur stetigen Verbesserung von Methoden und Programmen bei, sondern auch zur Bereitschaft der Betroffenen, die Ergebnisse der Evaluation umzusetzen.

Umstritten ist zudem, wie die Effektivität einer Methode oder eines Programms definiert werden soll. Sind randomisierte Experimente notwendig und sinnvoll oder eher nutzlos und hinderlich? Reicht es, wenn mit einer Methode statistisch 'signifikante' Ergebnisse erzielt werden oder müssen andere Maße der 'praktischen Signifikanz' und der absoluten Effektstärke herangezogen werden? Reichen überhaupt statistische Analysen von Daten aus einzelnen Studien aus oder müssen diese erst auf der Grundlage einer breiten Erfahrungsbasis und gesicherter Theorien interpretiert und diskutiert werden?

[2] So ein Glied kann zum Beispiel bei einer Mathematikaufgabe ein einfacher Summierungsfehler, ein sprachliches Missverständnis, das Nichtfertigwerden wegen Zeitmangel oder sogar zu viel Wissen sein (vgl. Wuttke 2007, S. 158-186; Jablonka 2006).

40 Jahre Personenevaluation – Versuch einer Zwischenbilanz

Zunächst wollen wir fragen, ob sich nach 40 Jahren Personen-Evaluation und *High-stakes testing* die Erwartungen erfüllt haben, mit denen diese Maßnahmen begründet wurden?

- Erbringen Schüler heute bessere Schulleistungen und hat sich das Bildungsniveau in den USA im Vergleich zu früher und im internationalen Vergleich generell erhöht?
- Wurde, wie die Regierungen immer wieder betonten, die Kluft zwischen den Leistungen der sozial schwachen Schüler sowie der Schüler aus ethnischen Minderheiten (Afroamerikaner, Latinos) und der weißen Mittelschicht-Kinder verringert?

Nicht vorgesehen hat die US-Regierung bislang, diese Politik selbst zu evaluieren oder von unabhängigen Wissenschaftlern evaluieren zu lassen. Man hielt es offenbar für ausreichend, dass entsprechende Gesetze beschlossen wurden, wie Ellwein, Glass und Smith (1988) feststellen: "Planned evaluation efforts were scant or focused on mundane, peripheral questions that could be answered using available technical expertise. The more complex and relevant questions of impact and utility were ignored." (Ellwin u. a. 1988, S. 2) Ähnlich stellten Kreitzer, Madaus und Haney (1989) ein "lack of good evidence on the impact of MCT [Minimum Competency Test] programms" (S. 146) fest und forderten: "The consequences of competency tests must be more thoroughly studied." (Kreitzer u. a. 1989, S. 146-147)

Immerhin wurden in den vier Jahrzehnten die Bildungsausgaben stark erhöht, die Lehrerbildung ist in den meisten Bundesstaaten von vier auf fünf Jahre verlängert worden. Die Bildungs- und Unterrichtsforschung wurde intensiviert (was schon an den steigenden Teilnehmerzahlen an dem Jahreskongress der amerikanischen Bildungsforschungsgesellschaft, AERA, und der Vielzahl von einschlägigen Zeitschriften und Forschungsberichten abgelesen werden kann) und es werden nach einer Schätzung von Bracey (2005) jährlich zwei bis drei Milliarden Dollar für Tests ausgegeben, abgesehen von den indirekten Kosten, die durch diese sanktionsbewehrten Tests verursacht werden (siehe unten).

Inzwischen liegen fundierte Forschungsergebnisse vor, die eine Bilanzierung der "Evaluationsgetriebenen" Bildungspolitik erlauben, vor allem die Daten aus den regelmäßig durchgeführten *National Assessment of Educational Progress* (NAEP), das seit 1969 in den USA durchgeführt wird, und aus

internationalen Vergleichsstudien wie der *Third International Mathematics and Science Study* (TIMSS), der IEA und dem Project of *International Student Assessment* (PISA) der OECD (vgl. Keitel 2007). Wenig ergiebig für eine Bilanzierung sind die Evaluationen auf der Ebene der Bundesstaaten, da Inhalte und Schwierigkeitsgrade der Tests zwischen den Bundesstaaten weit voneinander abweichen (vgl. Linn 2000, S. 10).

Internationale Vergleiche

Dem im internationalen Vergleich unvorstellbar intensiven Einsatz von Tests stehen eher schwache Testleistungen der US-Schüler bei internationalen Vergleichsstudien gegenüber. Beim Naturwissenschafts-Test und im Lese-Test der neuesten PISA-Studie liegen die US-Schüler deutlich unter dem internationalen Durchschnitt, weit abgeschlagen hinter Schülern in Finnland, die bis zur 8. Klasse keine Noten und auch keine *High-stakes* Tests kennen (PISA 2006, 2008; Prenzel u. a. 2007). Sechs Jahre zuvor, bei PISA 2000, lagen die US-Schüler hinsichtlich Lesen, Mathematik und Naturwissenschaften noch etwas besser, aber auch nicht so überragend, wie man aufgrund des langjährigen Vorsprungs an "evaluationsgetriebener" Bildungspolitik erwarten könnte.

National Assessment of Educational Progress (NAEP)

Beim NAEP wird jährlich eine repräsentative Auswahl von Schülern der Klassenstufe 4, 8 und 12 getestet, und zwar in den Fächern Lesen, Schreiben, Mathematik, Naturwissenschaften und Social Studies (politische Gemeinschaftskunde), Geschichte, Geographie und Geisteswissenschaften. Durchgeführt werden die Erhebungen gegenwärtig vom *Commissioner of Education Statistics* in der *Education Commission* der Bundesstaaten (vergleichbar unserer Kultusministerkonferenz). Das jährlich veröffentliche Ergebnis der NAEP-Erhebungen gilt als *Nation's Report Card.* Es stellt gewissermaßen der amtlichen Bildungspolitik ein Zeugnis aus.
Die beim NAEP verwendeten Tests wurden 1990 stark verändert, so dass die Entwicklungstrends davor kaum noch mit den heutigen Testergebnissen vergleichbar sind. Im Jahr 1996 wurden die Regeln für den Ausschluss vom Test auf Grund von Sprachproblemen geändert. Zudem wurden Hilfen für Schüler erlaubt, deren Muttersprache nicht Englisch ist. Die Daten vor und nach dieser Maßnahme sind daher nur bedingt vergleichbar.

Bei der nationalen Dauerbeobachtung NAEP seit 1990 zeigen sich *keine* Verbesserungen beim *Lesetest* und nur beim Mathematiktest ein *leicht ansteigender* Trend (vgl. Abb. 3), von der US-Regierung als ein Beleg für die Wirksamkeit der Dauerevaluation von Schülern, Lehrern und Schulleitungen gewertet wird.

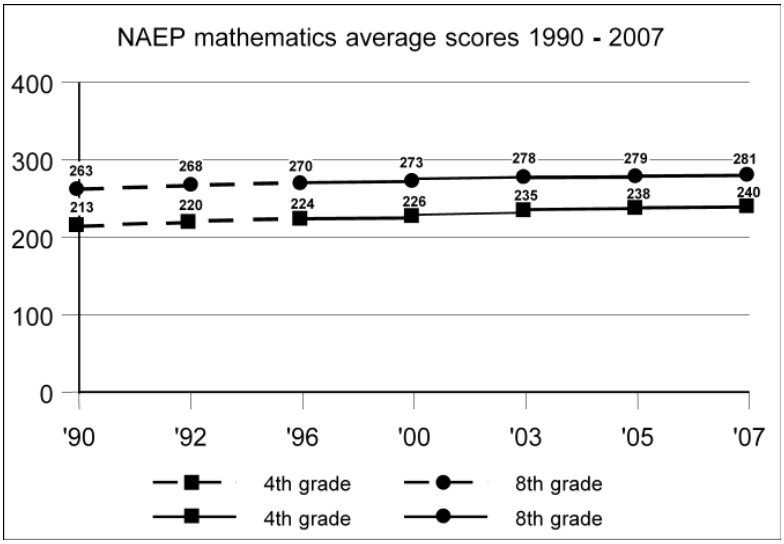

Abb. 3: Daten der NAEP-Erhebungen 1990 bis 2007. Ab 1996 wurden Teilnehmern, bei denen englisch die zweite Sprache ist, Hilfen gegeben. Quelle: NCES 2007.

Es ist allerdings fraglich, ob selbst dieser geringe Anstieg der Testpunktwerte über einen Zeitraum von 17 Jahren als Beleg für den Erfolg sanktionsorientierter Tests angesehen werden kann. Im Mathematik-Test zum Beispiel haben in 17 Jahren die Testwerte bei den Viertklässlern um 18 (von 500 möglichen) Punkte (also ca. 3%) und bei den Achtklässlern um 27 (von 500 möglichen) Punkte (also ca. 5%) zugenommen (vgl. Abb. 3). Die genaue Analyse dieser Daten legt vielmehr die Vermutung nahe, dass selbst dieser (geringe) Anstieg der Testwerte nicht einem besseren Unterricht zu verdanken sind, sondern einer Reihe von anderen Faktoren. Es gibt inzwischen Untersuchungen von unabhängigen Forschern, die zeigen, dass ein Großteil dieses ver-

meintlich positiven Trends durch Verzerrungen der Daten infolge einer wachsenden Korruption im Bildungssystem bedingt ist.

Unter den verschiedenen bekannt gewordenen Betrügereien stellen der Ausschluss leistungsschwacher Schüler von den Tests und das Training im Umgang mit Leistungstests auf Kosten des übrigen Lehrplans die offenbar wirksamsten Tricks dar, um ein positives Ergebnis vorzutäuschen, wo es keinen Leistungszuwachs oder gar eine Leistungsabnahme gibt.

Schon der Bericht des NCES (2007) gibt einen Hinweis auf das Problem des Testausschlusses. Im bevölkerungsreichsten Bundesland Kalifornien wurden zehn und mehr Prozent der Schüler aufgrund von Sprachschwierigkeiten vom NAEP-Test ausgeschlossen. Haney (2006) zeigt, wie in Florida die NAEP-Ergebnisse von Viertklässlern nach oben gedrückt wurden, indem der Staat schwache Schüler in der dritten Klasse verstärkt sitzen ließ.

Amrein und Berliner (2002) zeigen mit umfangreichen Analysen von NAEP-Daten, wie sich in vielen Bundesstaaten nach der Einführung von *High-stakes* Tests die Ergebnisse nicht verbessert, sondern verschlechtert haben (vgl. Nichols u. a. 2006). Sacks (1999) zeigt, dass sich die Schüler in den Staaten besonders stark verschlechtert haben, die an schlechte Testresultate besonders drakonische Strafen für Lehrer und Schulen geknüpft haben, was nicht gerade für den Erfolg sanktionsbewehrter Personenevaluation spricht.

Nach den Analysen von Amrein-Beardsley und Berliner (2003) werden in Bundesstaaten mit sanktionsorientierter Evaluation viel mehr Schüler von der Teilnahme an den NAEP-Erhebungen ausgeschlossen als Vergleichsstaaten ohne solche Evaluationen. Die positiven Korrelationen zwischen Sanktionen und Testleistungen verschwinden aber, wenn man die Ausschlussrate statistisch kontrollierte. Offenbar sind *High-stakes* Evaluationen, bei denen Schulen und Lehrer für das Verfehlen der staatlichen Normen bestraft werden (siehe oben), ein Anreiz für die Schulen, leistungsschwache Schüler von den Tests auszuschließen. Wie erfinderisch Menschen sind, die nicht in der Lage sind, drohende Strafen auf legalem Weg abzuwehren, zeigt sich auch in anderen bekannt gewordenen "Tricks" der Datenschönung (vgl. u. a. Haney 2000; Amrein/Berliner 2002; Bracey 2005; Haney 2006; Nichols/Berliner 2006; Heilig/Darling-Hammond 2008):

– *Teaching to the test*. Auf Kosten des gründlichen Lernens eines Faches und oft auch auf Kosten der curricularen Vielfalt bieten Schulen intensives Test-Training an. Lehrer und Eltern bilden dabei manchmal eine 'unheilige Allianz', um die Karrierechance der Kinder zu erhöhen, auch wenn das wirkliche Lernen und die Lernmotivation darunter leiden (vgl. Deci 1995; Deci u. a. 1999; Kohn 1999; 2000). Es zeigt sich, dass davon vor allem die

Kinder aus benachteiligten sozialen Schichten betroffen sind (Sacks 1999; Madaus/Clarke, 2001; Kohn 2000; Haney 2002).

- Leistungsschwache Schüler werden auf vielfache Weise daran gehindert, an den *High stakes*-Tests teilzunehmen.
- Inzwischen hat der Gesetzgeber dieses Loch verschlossen, indem Tests als ungültig erklärt werden, wenn mehr als ein bestimmter Prozentsatz der Schüler fehlt. Das hat gravierende Folgen für leistungsschwache Schüler. Schulen, deren Leitung besonders ehrgeizig ist, oder bei denen die Voraussetzungen für eine Verbesserung fehlen, weil das dafür notwendige Geld oder Wissen oder beides nicht vorhanden ist, gehen dazu über, die Zusammensetzung ihrer Schülerschaft zu manipulieren. Bekannte Manipulationen sind: a) lernschwache Schüler zum Verlassen der Schule oder ganz zum Schulabbruch zu bewegen (welche Schule will gerade diese Schüler als Wechsler aufnehmen?), und b) diese Schüler in dem Jahr vor dem obligaten Test nicht zu versetzen, um sie dann im nächsten Jahr gleich zwei Jahre weiterkommen zu lassen.
- Ein für die betroffenen Schüler ebenfalls nachteiliger Trick ist die Konzentration der schulischen Förderung auf die Schüler, die nur knapp unter der Norm liegen. Die Schulen kalkulieren, dass es leichter ist, diese auf das geforderte Niveau anzuheben als die "hoffnungslosen" Fälle. Die versucht man am besten aus der Schule heraus zu drängen. So kann man das Erscheinungsbild der Schule in den veröffentlichten sanktionsrelevanten Statistiken verbessern, ohne wirklich etwas zur Förderung lernschwacher Schüler beizutragen. Den Gesetzgeber interessiert nicht, wie hoch der Mittelwert aller Schüler in diesen Tests ist, sondern nur, wie viele Schüler über der festgelegten Norm liegen.
- Schließlich führt die so genannte "Null-Toleranz-Politik" in den Schulen bei Disziplinproblemen und kleineren Vergehen wie Zuspätkommen oder Widerrede gegen den Lehrer oft zu mehrtägigem Schulausschluss und Bedrängen der Schüler, auf eine Sonderschule zu wechseln. Auch hiervon sind wieder die Kinder mit afroamerikanischem und Latino-Hintergrund besonders betroffen (Heilig/Darling-Hammond 2008).

Man ist versucht, solche Tricks und Betrügereien zur Umgehung staatlicher Normen als Ausdruck moralischen Versagens abzutun. Aber tatsächlich resultiert die rasant um sich greifende Korruption im US-Schulsystem aus einem Missverhältnis von überhöhten Anforderungen an Lehrer und Schulleiter einerseits und den verfügbaren Ressourcen für eine tatsächliche Verbesserung des Unterrichts andererseits. Gerade Schulen in den sozial schwachen Bezirken sind vielfach unterfinanziert und viele der Lehrer und Hilfslehrer,

die dort unterrichten, haben eine zu geringe oder gar keine Ausbildung (vgl. Kozol 1992; Heilig/Darling-Hammond 2008).

"Kollateralschäden"

Sanktionsbewehrte Tests wurden eingeführt in der Hoffnung, die Lernleistungen amerikanischer Schüler zu erhöhen und die Kluft zwischen Kindern aus sozial schwachen Schichten und Minoritäten einerseits und Mittel- und Oberschichtkindern andererseits zu schließen ("Kein Kind bleibt zurück"). Aber genau die Kinder aus sozial schwachen Schichten erleiden durch die sanktionsbewehrten Tests und die Evaluation von Schulen und Lehrern die größten Nachteile:

- Ihre Schulen können die staatlichen Normen des NCLB-Gesetzes nicht schaffen (vgl. Linn 2008).
- Durch die Betonung von Lesefähigkeit auch in Mathematik- und Naturwissenschaftstests und die große Rolle, die Stressresistenz und trainierbare Testweisheit bei der Bearbeitung von vielen dieser Speed-Tests mit Auswahlantworten spielen, werden gerade lernschwache Kinder benachteiligt (vgl. Sacks 1999; Wuttke 2007, S. 163-186).
- Die Schulabbrecher- und Sitzenbleiberrate steigt, was heißt, dass der Anteil der Jugendlichen in einem Jahrgang mit einer High school graduation infolge dieser Evaluationspolitik sinkt (vgl. Kreitzer u. a. 1989; Madaus/Clarke 2001; Amrein/Berliner 2002; Nichols/Berliner 2006). Besonders deutlich wurde das in Texas, dessen Bildungspolitik wahre Wunder zu vollbringen schien, wie der damalige Gouverneur von Texas und spätere Präsident George W. Bush in seiner Wahlkampagne im Jahr 2000 immer wieder stolz erklärte. Besonders im Schulbezirk Houston ging laut Pressemeldungen die Zahl der Schulabbrecher stark zurück und nahmen die Testwerte rapide zu. Der für dieses 'Wunder von Texas' verantwortliche Superintendent des Schulbezirks, Rod Paige, wurde dafür von Bush später zum US-Bildungsminister ernannt. Er geriet unter öffentlichen Druck als ans Licht kam, dass dieses 'Wunder' allein der Tatsache zu verdanken war, dass er die Schulabbrecher einfach aus der Statistik verschwinden ließ (vgl. Haney 2000).
- Im Durchschnitt erreichen weniger als 70 Prozent der Schüler in den USA überhaupt noch einen Schulabschluss, in den ärmeren Innenbezirken der US-Metropolen sind es sogar weniger als 50 (!) Prozent, die einen Schulabschluss schaffen (Education Week, vom 22.6.2006). Niemand fragt nach

> den Hunderttausenden Jugendlichen, die aufgrund dieser Politik die Schule vorzeitig, ohne Schulabschluss verlassen müssen und eine geringe Chance haben, ihren Lebensunterhalt später einmal auf 'anständige' Weise zu verdienen. Viele von ihnen werden aus Not straffällig und landen im Gefängnis. Jeder vierte Jugendliche, der weltweit im Gefängnis sitzt, sitzt in den USA im Gefängnis (vgl. Goodman 2008). Die Jugendkriminalitätsrate hat sich in den USA seit 1980 vervierfacht (vgl. Lochner/Moretti 2004).[3]

Die korrumpierenden Folgen der sanktionsbewehrten Personenevaluation, die erst jetzt sichtbar werden, wurden bereits vor mehr als dreißig Jahren von dem renommierten Sozialpsychologen Donald T. Campbell (1976) vorhergesehen. Er stellte schon damals fest, was heute als "Campbell's Law" bezeichnet wird: "Je stärker ein einzelner quantitativer sozialer Faktor dazu benutzt wird, soziale Entscheidungen zu begründen, desto stärker ist er verzerrenden Einflüssen ausgesetzt und je mehr führt er selbst dazu, die sozialen Prozesse zu verzerren und zu verfälschen, die eigentlich untersucht und verbessert werden sollen." (S. 49; meine Übers.).

Die Evaluationen mit Sanktionsfunktion haben nicht nur zweifelhafte Folgen für das Bildungsniveau der Schüler in den USA, sondern auch für die Lehrer. Die Entwicklung der letzten Jahre ist paradox. Einerseits müssten zukünftige Lehrer/innen, in deren Händen das Schicksal vieler Generationen von Kindern liegt, und Schulleiter, die für Personalführung und Schulorganisation verantwortlich sind, vor Eintritt in ihren Dienst gründlich und angemessen auf ihre Lehr- bzw. Leitungsfähigkeit hin überprüft werden. Es ist wichtig, dass sie zeigen, dass sie diese Fähigkeit demonstrieren, bevor sie fest eingestellt werden. Aber obwohl eine solche Personenbeurteilung legitim und sogar dringend geboten erscheint, werden in den USA immer noch Lehrer eingestellt, die eine unzulängliche Ausbildung haben und nicht die für diesen Beruf notwendige Befähigung mitbringen (vgl. Lankford u. a. 2002; Darling-Hammond/Youngs 2002).

Andererseits brauchen Lehrer Eigenverantwortung und Gestaltungsmöglichkeiten, um ihre Fähigkeit im Unterricht voll zu entfalten und Kinder mit unterschiedlichen Lernvoraussetzungen und Interessen optimal fördern zu können (vgl. Smylie 1997). Durch die Ausweitung der Vorschriften und Kontrollen der Lehrerarbeit durch standardisierte Tests wird ihre Verantwortung immer mehr eingeschränkt und behindert. Wie der *Illinois Research Council* am Beispiel des Schulbezirks Chicago zeigt, führt die Anhebung der

[3] Siehe auch U.S. Department of Justice. Bureau of Justice Statistics. http://www.ojp.usdoj.gov/bjs/glance/corrtyp.htm

Qualifikationsanforderungen an junge Lehrer direkt zu besseren Schülerleistungen (vgl. White u. a. 2008). Gleichzeitig ist bekannt, dass viele qualifizierte Lehrer diesen Beruf bald wieder verlassen, weil sie sich durch die ständige Überwachung durch Tests an der Entfaltung ihrer pädagogischen Verantwortung gehindert sehen.

Bilanz: Personenevaluation als Mittel der Bildungspolitik?

Viele Politiker beider großen Parteien in den USA glauben auch heute noch an die qualitätssteigernde Kraft von sanktionsbewehrten Vergleichstests. Sie meinen, dass Menschen (außer ihnen selbst vielleicht) ohne Tests, Strafandrohung und Geldanreize weder lernen noch Leistungen zeigen würden, auch wenn die Forschung heute ziemlich eindeutig zeigt, dass dies ein Irrglaube ist.

Demgegenüber erregen der geringe Nutzen und die immer deutlicher werdenden Schäden dieser Politik gerade für lernschwache Schüler immer mehr Widerstand bei Lehrern und Lehrerverbänden in den USA (vgl. Dillon 2008), worin sie von vielen renommierten Bildungsforschern unterstützt werden. Diese halten die Zielsetzung der Personenevaluation für gescheitert, über Testdruck die Lernleistungen der Schüler zu erhöhen, und haben sich gegen eine Fortführung dieser Politik ausgesprochen (vgl. Popham 1999; Sacks 1999; Kohn 2000; Amrein/Berliner 2002; Nichols u. a. 2006; Nichols/Berliner 2006; Baker 2007). Sie verweisen auf die oft mangelhafte Qualitätskontrolle bei der Entwicklung von Schulleistungstests (vgl. AERA 2003; Rhoades/Madaus 2003; speziell zu PISA siehe Wuttke 2007) und vor allem auf den hohen Preis, den Lehrer, Schüler und Eltern und letztlich auch die amerikanische Gesellschaft für eine Evaluationspolitik zahlen müssen, die selbst nach vierzig Jahren noch nicht den Nachweis ihre Wirksamkeit erbracht hat.

Was kann man von Amerika lernen?

Sofern es hier noch einer Zusammenfassung der im Text bereits angesprochenen Lehren aus fast einem Jahrhundert Evaluation in den USA bedarf, möchte ich als Antwort auf diese Frage eine der renommiertesten Bildungsforscherinnen der USA, Linda Darling-Hammond (1994) zu Wort kommen lassen: “Effective policy strategies will thus need to invest in teacher know-

ledge as well as in new assessment strategies, if the curriculum goals are to be achieved." (Darling-Hammoind 1994, S. 364)

Literatur

AERA (2003): Standards and tests: Keeping them aligned. In: Research Points (American Educational Research Association). Vol. 1/No. 1, S. 1-4.

Amrein, A./Berliner, D. C. (2002): High-stakes testing, uncertainty, and student learning. In: Education Policy Analysis, 10, No. 8. Arizona: Education Policy Studies Laboratory.

Amrein-Beardsley, A./Berliner, D. C. (2003): Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. In: Education Policy Analysis Archives, 11. Jg./Heft 25.

APA (2001): Publication Manual of the American Psychological Association, Fifth Edition. Washington, D.C.: APA.

Baker, E. (2007): The end(s) of testing. In: Educational Researcher, 36. Jg./Heft 6, S. 309-317.

Barber, L. W. (1999): Self-assessment. In: Millman, J./Darling-Hammond, L. (Hrsg.): The new handbook of teacher evaluation. Assessing elementary and secondary school teachers. Newbury Park: Sage, S. 216-227.

Belley, P./Lochner, L. (2008): The changing role of family income and ability in determining educational achievement. Working Paper # 2008-1 January 2008.

Berliner, D. C./Biddle, B. J. (1995): The manufactured crisis. Myths, fraud, and the attack on America's public schools. Reading. MA: Addison-Wesley.

Biedinger, N./Becker, B. (2006): Der Einfluss des Vorschulbesuchs auf die Entwicklung und den langfristigen Bildungserfolg von Kindern. Ein Überblick über internationale Studien im Vorschulbereich. Arbeitspapiere – Working Papers Nr. 97, Mannheimer Zentrum für Europäische Sozialforschung.

Blatt, M./Kohlberg, L. (1975): The effect of classroom moral discussion upon children's level of moral judgment. In: Journal of Moral, Education, Jg. 4, S. 129-161.

Boyer, E. L. (1990): Civic education for responsible citizens. In: Educational Leadership, Nov. 1990, S. 4-7.

Bracey, G. W. (2002): The war against America's public schools. Privatizing schools, commercializing education. Boston: Alyn & Bacon.

Bracey, G. W. (2005): No child left behind: Where does the money go? Education Policy Studies Laboratory, June 2005.

Bracey. G. W. (2007): The First Time 'Everything Changed'. The 17th Bracey Report on the Condition of Public Education.

Bridgeman, B. (1992): Placement validity of a prototype SAT with an essay. Research Report. Research Report No. ETS-RR-92-28. Princeton, NJ: Educational Testing Service, ERIC #ED390893.

Campbell, D. T. (1969): Reforms as experiments. In: American Psychologist, Jg. 24, S. 409-429.

Campbell, D. T. (1976): Assessing the impact of planned social change. Occasional papers # 8. Social research and public policies: The Dartmouth/OECD Conference, Hanover, NH: Dartmouth College, The Public Affairs Center.

Carver, R. P. (1993): The case against statistical significance testing, revisited. In: Journal of Experimental Education, 61. Jg./Heft 4, S. 287-292.

Chamberlin, D./Chamberlin, E. S./Drought, N. E./Scott, W. E. (1942): Did they succeed in college? In: Adventures in American education. Vol. IV. New York: Harper & Brothers.

Cullen, J. B./Jacob, B./Levitt, S. D. (2000). The impact of school choice on student outcomes: an analysis of the Chicago pubic schools, National Bureau for Economic Research (NBER) Working Paper 7888.

Currie, J./D. Thomas (2000): School Quality and the Longer-Term Effects of Head Start. In: The Journal of Human Resources, 35. Jg./Heft 4, S. 755-774.

Darling-Hammond, L. (1994): Policy uses and indicators. In: OECD (Hrsg.): Making education count. Paris: OECD, S. 357-378.

Darling-Hammond, L./Ancess, J. (1996): Democracy and access to education. In: Soder, R. (Hrsg.): Democracy, education, and the schools. San Francisco, CA: Jossey-Bass, S. 151-181.

Darling-Hammond, L./Youngs, P. (2002): Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? In: Educational Researcher, December 2002, S. 13-25.

Deci, E. L. (1995): Why we do what we do: The dynamics of personal autonomy. New York: G. P. Putnam's Sons.

Deci, E. L./Koestner, R./Ryan, R. M. (1999): Examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, Jg. 125, S. 627-668.

Deming, W. E. (1994): The new economics for industry, government, education. Cambridge MA: Massachusetts Institute of Technology, 2nd edition.

Deutsches PISA-Konsortium (2001): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich.

Dewey, J. (1964). Demokratie und Erziehung. Eine Einleitung in die philosophische Pädagogik. Braunschweig: Georg Westermann Verlag (Original 1915).

Dillon, S. (2008): New vision for schools proposes broad role. New York Times, 15.7.2008.

Dubs, R. (2005): Metaevaluation – Anforderungen an Schulaufsicht und Schulleitung. In: Bartz, A./Fabian, J./Huber, S. G./Kloft, C./Rosenbusch, H. S./Sassenscheid, H. (Hrsg.): Praxiswissen SchulLeitung. Basiswissen und Arbeitshilfen zu den zentralen Handlungsfeldern von Schulleitung.

Ellwein, M. C./Glass, G. V./Smith, M. L. (1988): Standards of competence: Propositions on the nature of testing reforms. Educational Researcher, 17. Jg./Heft 8, S. 4-9.

Fisher, G. M. (1928): Foreword. In: Hartshorne and May (Hrsg.), S. v-vii.

Geiser, S./Studley, R. (2001): UC and the SAT. Oakland, CA: University of California Office of the President.

Goodman, D. (2008): Hard Time Out. Five-year-olds in handcuffs, eighth-graders detained for doodling: The prison boom comes to the schools. In: Mother Jones, July 21, 2008.

Haney, W. M. (2000): The myth of the Texas miracle in education. In: Education Policy Analysis Archives, Jg. 8/ Heft41.

Haney, W. M. (2002): Ensuring failure: How a state's achievement test may be designed to do just that. In: Education Week, 10 July 2002, S. 56 und 58.

Haney, W. M. (2006): Evidence on Education under NCLB (and How Florida Boosted NAEP Scores and Reduced the Race Gap). Paper presented at the Hechinger Institute "Broad Seminar for K-12 Reporters", Sept. 8-10, Grace Dodge Hall, Teachers College, Columbia University, New York City.

Hartshorne, H./May, M. A. (1928): Studies in the nature of character. Vol. I: Studies in deceit, Book one and two. New York: Macmillan.

Heilig, J. V./Darling-Hammond, L. (2008): Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. In: Educational Evaluation and Policy Analysis, 30. Jg./Heft 2, S. 75-110.

Jablonka, E. (2006): Mathematical literacy: Die Verflüchtigung eines ambitionierten Testkonstrukts in bedeutungslosen PISA-Punkten. In: Jahnke, T./Meyerhöfer, W. (Hrsg.): Pisa & Co. – Kritik eines Programms. Hildesheim: Franzbecker, S. 155-186.

Keitel, C. (2007): Der (un)heimliche Einfluss der Testideologie auf Bildungskonzepte, Mathematikunterricht und mathematikdidaktische Forschung. In: Jahnke, T./Meyerhöfer, M. (Hrsg.): PISA & Co. Kritik eines Programms. Hildesheim: Franzbecker, S. 25-58., 2. erw. Aufl.

Kozol, J. (1991): Savage inequalities. New York: Crown.

Kreitzer, A. E./Madaus, G. F./Haney, W. M. (1989): Competency testing and dropouts. In: Weis, L./Farrar, E./Petrie, H. G. (Hrsg.): Dropouts from school. Issues, dilemmas, and solutions. Albany, NY: SUNY Press, S. 129-152.

Kohn, A. (1999): Punished by rewards. The trouble with gold stars, incentive plans, A's, praise, and other bribes. Boston: Houghton Miflin.

Kohn, A. (2000): The case against standardized testing. Raising the scores, ruining the schools. Portsmouth, NH: Heinemann.

Kozol, J. (1992): Savage inequalities. Children in America's schools. New York: Harper.

Lankford, H./Loebe, S./Wyckhoff, J. (2002): Teacher sorting and the plight of urban schools: a descriptive analysis. In: Educational Evaluation and Policy Analysis, 24. Jg./Heft 1, S. 37-62.

Leming, J. S. (1981): Curricular effectiveness in moral/values education: A review of research. Journal of Moral Education, 10. Jg./Heft 3, S. 147-164.

Lerkiatbundit, S./Utaipan, P./Laohawiriyanon, C./Teo, A. (2006): Randomized controlled study of the impact of the Konstanz method of dilemma discussion on moral judgement. In: Journal of Allied Health, 35. Jg./Heft 2, S. 101-108.

Lind, G. (2002): Ist Moral lehrbar? Ergebnisse der modernen moralpsychologischen Forschung. Berlin: Logos-Verlag.

Lind, G. (2003:. Moral ist lehrbar. Ein Handbuch zur moralischen und demokratischen Bildung. München: Oldenbourg.

Lind, G. (2004): Jenseits von PISA — Für eine neue Evaluationskultur. In: Institut für Schulentwicklung PH Schwäbisch Gmünd (Hrsg.): Standards, Evaluation und neue Methoden. Reaktionen auf die PISA-Studie. Baltmannsweiler: Schneider Verlag Hohengehren, S. 1-7.

Lind, G. (2007): Effektstärke: Statistische versus praktische und theoretische Bedeutsamkeit. University of Konstanz. http://www.uni-konstanz.de/ag-moral/pdf/Lind-2007_Effektstaerke-Vortrag.pdf.

Lind, G. (2008): The meaning and measurement of moral judgment competence revisited – A dual-aspect model. In: Fasko, D./Willis,W. (Hrsg.): Contemporary Philosophical and Psychological Perspectives on Moral Development and Education. Cresskill. NJ: Hampton Press, S. 185-220.

Lind, G. (o. J.)Verbesserung der Lehre durch Selbstevaluation. Http://www.uni-konstanz.de/itse.

Linn, R. (2000): Assessment and accountability. In: Educational Researcher, 29. Jg./Heft 2, S. 4-16.

Linn, R. (2008): Toward a more effective definition of Adequate Yearly Progress. Berkeley Law School.

Lipsey, M. W./Wilson, D. B. (1993): The efficacy of psychological, educational and behavioral treatment. Confirmation from meta-analysis. In: American Psychologist, Jg. 48, S. 1181-1209.

Lochner, L./Moretti, E. (2004): The effect of education on crime: evidence from prison inmates, arrests, and self-reports. In: The American Economic Review, 94. Jg./Heft 1, S. 155-189.

Madaus, G./Clarke, M. (2001): The adverse impact of high stakes testing on minority students: Evidence from one hundred years of test data. In: Orfiled, G./Kornhaber, M. L. (Hrsg.): Raising standards or raising barriers? Inequality and high-stakes testing in public education. New York: Century Foundation Press.

NCES - National Center for Educational Statistics (2007): The Nation's Report Card. Mathematics 2007. Washington, D.C.: U.S. Department of Education.

New York Times (6.10.2004). Wider gap found between wealthy and poor schools.

Nichols, S. L./Glass, G. V./Berliner, D. C. (2006): High-stakes testing and student achievement: Does accountability pressure increase student learning? In: Education Policy Anlysis Archives, 14. Jg./Heft 1.

Nichols, S.L./Berliner, D. (2006): Collateral damage: How high-stakes testing corrupts schools. Cambridge, MA: Harvard Education Press.

Nye, B./Konstantopoulos, S./Hedges, L. (2004): How large are teacher effects? In: Educational Evaluation and Policy Analysis, 26. Jg./Heft 3, S. 237-257.

Peschel, F. (2002): Offener Unterricht – Idee, Realität, Perspektive und ein praxiserprobtes Konzept zur Diskussion. Hohengehren: Baltmannsweiler.

Piaget, J. (1972): Das moralische Urteil beim Kinde. Frankfurt: Suhrkamp (Original 1932).

PISA-2006 (2008): PISA 2006. Science competencies for tomorrow's world. Executive summary. Paris: OECD.

Popham, W. J. (1999): Why standardized tests don't measure educational quality. In: Educational Leadership, 56. Jg./Heft 6, S. 8-15.

Prehn, K./Wartenburger, I./Mériau, K./Scheibel, C./Goodenough, O./Villringer, A./Meer, E. v. d./Heekeren, H. (2008): Influence of individual differences in moral judgment competence on neural correlates of socio-normative judgments. In: Social Cognitive and Affective Neuroscience, 3. Jg./Heft 1, S. 33-46.

Prenzel, M./Artelt, C./Baumert, J./Blum, W./Hammann, M./Klieme, E./Pekrun, R. (Hrsg.) (2007): PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie. Münster: Waxmann.

Radigk, W. (1986): Kognitive Entwicklung und zerebrale Dysfunktion. Dortmund: Verlag Modernes Lernen.

Rhoades, K./Madaus, G. (2003): Errors in standardized tests: a systemic problem. National Board on Educational Testing and Public Policy.

Rosenthal, R./Rubin, D. B. (1982): A simple, general purpose display of magnitude of experimental effect. In: Journal of Educational Psychology, Jg. 74, S. 166-169.

Sachser, N. (2006): Neugier, Spiel und Lernen: Verhaltensbiologische Anmerkungen zur Kindheit. In: U. Herrmann (Hrsg.): Neurodidaktik. Grundlagen und Vorschläge für gehirngerechtes Lehren und Lernen. Weinheim: Beltz, S. 19-30.

Sacks, P. (1999): Standardized minds. The high prize of America's testing culture and what we can do to change it. Cambridge, MA: Perseus Publishing.

Sanders, J. R. (1994): The program evaluation standards. How to assess evaluations of educational program. Thousand Oaks, USA: Sage Publications, 2nd edition.

Schoenfeld, A. H. (1999): Looking toward the 21th century: Challenges of educational theory and practice. In: Educational Researcher, Jg. 28, S. 4-14.

Sedlmeier, P./Köhlers, D. (2001): Wahrscheinlichkeiten im Alltag. Statistik ohne Formeln. Braunschweig: Westermann.

Shepard, L.A. (2002): The role of assessment in a learning culture. In: Educational Researcher, Jg. 29, S. 4-14.
Smith, F. (1986): Insult to intelligence. The bureaucratic invasion of our classrooms. New York: Arbor House.
Smith, E. R./Tyler, R. W. (1942): Appraising and recording student progress evaluation, records and reports in the Thirty Schools. New York: Harper & Brothers.
Smylie, M. A. (1997): From bureaucratic control to building human capital: The importance of teacher learning in education reform. In: Educational Researcher, Jg: 26, S. 9-11.
Spitzer, M. (2002): Lernen. Gehirnforschung und die Schule des Lebens. Heidelberg: Spektrum.
Toppo, G. (2008): Study: Bush's Reading First program ineffective. USA TODAY, 5.5.2008
U.S. Department of Education (2008): Reading First.
White, B. R./Presley, J. B./DeAngelis, K. J. (2008): Leveling up: Narrowing the teacher academic capital gap in Illinois (IERC 2008-1). Edwardsville, IL: Illinois Education Research Council.
Winerip, M. (2005): SAT Essay test rewards length and ignores errors. New York Times, May 4, 2005.
Wuttke, J. (2007): Die Insignifikanz signifikanter Unterschiede. In: Jahnke, T./Meyerhöfer, T. (Hrsg.): Pisa & Co. Kritik eines Programm. Hildesheim: Franzbecker, S. 99-246, 2. erw. Aufl.
Zigler, E. /Muenchow, S. (1992): Head Start. The Inside Story of America's Most Successful Educational Experiment. New York: Basic Books.

Verzeichnis genutzter Internetseiten

http://economics.uwo.ca/centres/cibc/ [Datum der Recherche: 28.07.2008]
http://edpolicylab.org [Datum der Recherche: 12.10.2008]
http://epaa.asu.edu/epaa/v10n18/ [Datum der Recherche: 12.10.2004]
http://epaa.asu.edu/epaa/v11n25/ [Datum der Recherche: 12.10.2004]
http://epaa.asu.edu/epaa/v14n1/ [Datum der Recherche: 20.07.2008]
http://epaa.asu.edu/epaa/v8n41/ [Datum der Recherche: 01.07.2008]
http://www.america-tomorrow.com/bracey/EDDRA/ k0710bra.pdf [Datum der Recherche: 01.08.2008]
http://www.ed.gov/programs/readingfirst/index.html [Datum der Recherche: 10.01.2008]
http://www.edweek.org/ew/articles/2008/06/24/43senate_web.h27.html [Datum der Recherche: 25.6.2008]
http://www.law.berkeley.edu/centers/ewi-old/research/k12equity/Linn.htm [Datum der Recherche: 20.07.2008]
http://www.motherjones.com/cgiin/print_article.pl?url...nes.com/news/feature/2008/07/slammed-hard-time-out.html [Datum der Recherche: 22.07.2008]
http://www.nber.org/papers/w7888 [Datum der Recherche: 15.7.2008]
http://www.nytimes.com/2005/05/04/education/04education.html [Datum der Recherche: 27.07.08]
http://www.pisa.oecd.org/dataoecd/15/13/39725224.pdf [Datum der Recherche: 16.07.08]
http://www.uni-konstanz.de/ag-moral/pdf/Lind-2007_Effektstaerke-Vortrag.[Datum der Recherche: 12.10.20047
http://www.uni-konstanz.de/itse [Datum der Recherche: 12.10.2008]