Erschien auszugsweise in: Bohl, T. & Kiper, H., Hrsg. (2009). *Lernen aus Evaluationsergebnissen – Verbesserungen planen und implementieren*, S. 61-79. Bad Heilbrunn: Klinkhardt.
Hier ist die vollständige Version einschließlich der Kapitel „Bildung und Evaluation in den USA“ und „Programmevaluation“.

Link zum publizierten Artikel

Amerika als Vorbild?

Erwünschte und unerwünschte Folgen aus Evaluationen[1]

Georg Lind

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

[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. Patricia Knoop und Kay Hemmerling haben mir beim Korrekturlesen wertvolle Hilfe geleistet.

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 als 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 (Deutsches PISA-Konsortium, 2001) ist 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 (z.B. Dewey, 1964, zuerst im Jahr 1915 publiziert).

Bildung hatte seit der Unabhängigkeitserklärung in den USA einen hohen Stellenwert. Thomas Jefferson, einer der Gründungsväter und zweiter Präsident, stellte eine enge Verbindung zwischen Bildung und Demokratie her: "Ich weiß nicht, wo die letzte Gewalt des Volkes besser aufgehoben ist als beim Volk selbst; und wenn wir glauben, dass es nicht aufgeklärt genug ist, um seine Kontrolle mit vollem Verstand auszuüben, ist die Lösung nicht, diese ihnen wegzunehmen, sondern ihren Verstand durch Bildung (*Education*) zu unterrichten."[2] Aber es war den Eltern überlassen, ob sie ihre Kinder auf eine (kirchliche oder lokale) Schule schicken wollten oder es sich finanziell leisten konnten. Erst auf Betreiben des Schulreformers und Abgeordneten *Horace Mann* entstand zunächst in Massachusetts und später in allen US-Staaten ein öffentliches Schulsystem, das zunächst nur lokal finanziert wurde, heute aber auch von den Bundesstaaten und der Bundesregierung mitfinanziert wird. Die Grundidee Manns war, jedem Kind eine vollwertige Schulbildung zu ermöglichen. Die Schulpflicht wurde schon zu Manns Zeiten auf 10 Schuljahre (16. Lebensjahr)

[2] Zitiert nach Boyer, 1990; S. 5, meine Übers., GL.

angehoben (in Deutschland wurde das erst viel später erreicht), bald auf 12 Schuljahre. Hinzu kamen *Kindergartens*, die anders als deutsche Kindergärten bereits akademische Lehraufgaben übernehmen. Die Teilnahmerate liegt heute über 50% (Biedinger & Becker, 2006). Insgesamt hat die USA also ein 13-jähriges Schulangebot (K-12) für alle Kinder.

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.

Die Hauptlast der Bezahlung der Schulgebäude und der Lehrergehälter tragen die Mitglieder der Schulgemeinde. Das Geld für die Schulen wird als Schulsteuer direkt erhoben. In vielen Fällen zahlen nicht alle Bürger diese Steuer, sondern nur die Bürger mit schulpflichtigen Kindern. Oft trauen sich die gewählten Vertreter nicht, die Schulsteuer von den Verweigerern einzutreiben, um ihre nächste Wahl nicht zu gefährden. Der gewählte Schulrat (*School board*) dient politisch ambitionierten Amerikanern oft als erste Karrieresprosse.

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 (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 (Kozol, 1991). Dieser Trend dürfte sich bis heute noch verstärkt haben. Ein immer bedeut-

samerer Grund für die wechselseitige Abhängigkeit von Wohlstand und Schulqualität ist die Koppelung der Grundstückpreise an die (meist öffentlich zugänglichen) Testleistungsniveaus einer Schule (Cullen et al., 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.

Nach und nach fingen auch die Bundesstaaten an, die lokalen Schulen durch Subventionen zu unterstützen. Die Höhe der Subventionen variiert stark zwischen den Bundesstaaten. Mitte der 1960er Jahre hat die US-Bundesregierung angefangen, sich an der Schulfinanzierung zu beteiligen und damit auch, die Bildung vor Ort zu beeinflussen – verbunden mit großflächiger Evaluation. Das war während der Konfrontation zwischen der USA und der ehemaligen Sowjetunion. Die Kuba-Türkei-Krise war gerade überstanden, aber der Vietnam-Krieg weitete sich immer mehr aus und in allen Teilen der so genannten Dritten Welt tobte der Kampf um Einfluss, Rohstoffe und Absatzmärkte. Um die wirtschaftliche und militärische Position der USA zu stärken, hat Präsident Kennedy das Problem der Armut in den USA und das Bildungspotential der armen Bevölkerungsschichten zum Thema gemacht. Sein Nachfolger Johnson führte diese Politik fort und setzte im Rahmen seines "Kriegs gegen die Armut" das "Elementary and Secondary Education" (ESEA) genannte Gesetz durch und damit die Grundlage für Aktivitäten der Bundesregierung im Bildungsbereich, der nach der Verfassung der USA reine Ländersache ist. Dieses Gesetz ermöglichte der US-Bundesregierung das *Project Head Start* (Zigler & Muenchow, 1992; Biedinger & Becker, 2006) und bis heute alle weiteren Schulprogramme der Bundesregierung (Linn, 2008).

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 ESEA 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 durchführen. Inzwischen sind viele Schuldistrikte dazu übergegangen, diese Tests auch von Schulen zu verlangen, die keine NCLB-Mittel beanspruchen (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 nicht direkt genannt wird (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, und wie Evaluation im Bildungsbereich zu gestalten ist, schon lange Gegenstand 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 (Deming, 1995; Kohn, 1999) – vor allem zwischen zwei Zielsetzungen von Bildungsevaluation, zwischen Personen- und

Programm-Evaluation.[3] 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 (Kreitzer et al., 1989; Sacks, 1999; Bracey. 2002; Nichols & Berliner, 2006; Heilig & Darling-Hammond, 2008).[4]

---

[3] Diese Unterscheidung sollte nicht mit Ergebnis- versus Prozessevaluation gleichgesetzt werden, wie manche das tun (Barber, 1999), da Personen- und Programm-Evaluation sowohl *summativ* wie auch *formativ* sein können. Auch sollte sie nicht mit Hierarchie-Ebenen verbunden werden. Wenn z.B. die Schulaufsicht ihre eigene Effektivität überprüfen möchte, betreibt sie ebenso Programm-Evaluation wie der Lehrer, der die Effektivität seiner Unterrichtsmethoden evaluiert. Die (anonyme) Erhebung von Daten bei Lehrern und Schülern ist für sich genommen keine punitive Personen-Evaluation. Dazu wird die Datenerhebung erst, wenn die “Lieferanten” der Daten namentlich identifiziert und für gute Daten belohnt und für schlechte bestraft werden, was bereits dann der Fall ist, wenn ihre Namen aktenkundig gemacht oder gar veröffentlicht werden.

[4] Genau genommen handelt es sich hier um Personen*beurteilung* und nicht um Evaluation im wissenschaftlichen Sinne. Sie sollte daher ebenso wie jede andere Beurteilung von Personen den Anforderungen der Verfassung und der Rechtsprechung genügen. In der Tat sind in den USA bereits mehrere Prozesse zum Einsatz von High-stakes Tests

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 (Bracey, 2005).[5] 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

---

geführt worden, wobei es vor allem um die geringe Prognosefähigkeit dieser Tests (Sacks, 1999; Geiser & Studley, 2001; Nichols & Berliner, 2006) und fehlerhafte Testaufgaben ging (Roades & Madaus, 2003).

[5] Nachdem Forscher der University of California feststellten, dass der meistbenutzte College-Zugangstest, der SAT, eine geringere prognostische Validität besitzt als die Abiturnote (Geiser & Studley, 2001) haben sie den SAT ausgesetzt und vom Testhersteller *College Board* eine Überarbeitung verlangt. Der neue SAT enthält seit 2003 einen 25-minütigen "Essay" als Testaufgabe. Der große Aufwand bei der Auswertung dieser Essays führte wohl dazu, dass die Essays nur nach ihrer Länge bewertet werden; sprachliche Mängel und sachliche Fehler haben keinen Einfluss auf den Testwert (Winerip, 2005). Aber bereits Bridgeman (1992) vom Educational Testing Service fand, dass auch der SAT mit Kurzessay im Vergleich zu Noten und Schulaufsätzen bei der Prognose des College-Erfolgs in Englisch deutlich schlechter abschnitt.

fehlt[6]; und c) die Lehrbedingungen der Schule, z.B. die Ausbildungsqualität der Lehrer und die zur Verfügung stehenden Finanzmittel (Kozol, 1991; Nye et al., 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 (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 (Deming, 1994; Deci, 1995; Deci et al., 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.

Auch die Programmevaluation ist nicht frei von Problemen, wie weiter unten noch eingehender zu diskutieren ist. Die größten Probleme stellt die Verknüpfung der Evaluationsergebnisse mit wirtschaftlichen Interessen dar und die Definition dessen, was man unter 'effektiv' und 'effizient' zu verstehen hat. Sobald bestimmte Unterrichtsmethoden und -programme dem Erzielen von Unternehmensgewinnen dienen, sind auch hier *High stakes* im Spiel, die sich unmittelbar auf die

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

Durchführung, Auswertung und Interpretation von Evaluationsmaßnahmen auswirken (Bracey, 2005). 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?

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 (Afro-Amerikaner, Latinos) und der weißen Mittelschicht-Kinder verringert?

Bislang hat die US-Regierung nicht vorgesehen, 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 und Glass (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." (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." (S. 147)

Immerhin wurden in den vier Jahrzehnten die Bildungsausgaben stark erhöht, die Lehrerbildung ist in den meisten Bundesstaaten von vier auf fünf Jahre erhöht worden, die Bildungs- und Unterrichtsforschung wurden 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 (siehe 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 (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 et al., 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 Evaluations-getriebener 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). Die Politik des NAEP wird von einem überparteilichen *National Assessment Governing Board bestimmt*. Das Gremium setzt sich aus Vertretern der Bundesstaaten, der lokalen Schulbehörden, Wissenschaftlern, Vertretern der Wirtschaft und bekannten Persönlichkeiten zusammen. 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 von der Testteilnahme 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 (Abb. 1), der von der US-

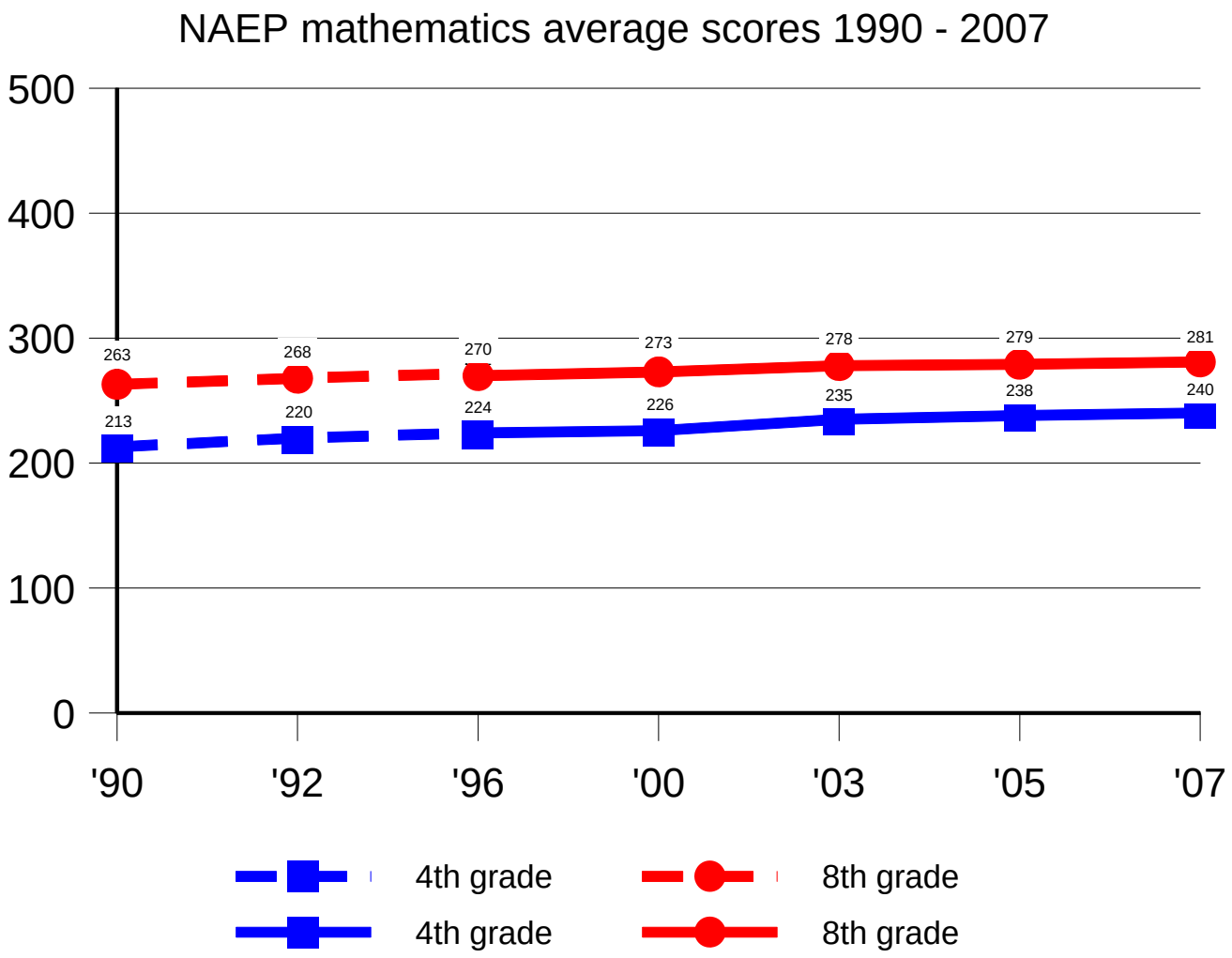

Abb. 1: Daten der NAEP-Erhebungen 1990 bis 2007. Ab 1996 wurden Teilnehmern, bei denen Englisch die zweite Sprache ist, Hilfen gegeben. Quelle: NCES 2007.

Regierung als ein Beleg für die Wirksamkeit der Dauerevaluation von Schülern, Lehrern und Schulleitungen gewertet wird. Es ist allerdings fraglich, ob selbst dieser geringe Anstieg der Testpunktwerte über einen Zeitraum von 17 Jahren als Beleg für den Erfolg des NCLB-Gesetzes angesehen werden kann. Im Mathematik-Test zum Beispiel haben die Testwerte bei den Viertklässlern um 18 (von 500 möglichen) Punkten zugenommen (also um ca. 3%) und bei den Achtklässlern um 27 (von 500 möglichen) Punkten (also um ca. 5%; vgl. Abbildung 1). 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 vermeintlich 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. Heilig und Darling-Hammond (2007) referieren Studien, die zeigen, dass in Texas sehr viele leistungsschwache Schüler der dritten und der neunten Klasse nicht versetzt wurden, obwohl die das Klassenziel eigentlich erreicht hatten, um

sie dann ein Jahr später eine Klasse überspringen zu lassen, nämlich die Klasse, in denen der NAEP-Test durchgeführt wird.[7]

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 (siehe auch Nichols, Glass, & Berliner, 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 ausschlossen als in 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 (Deci, 1995; Deci & Ryan,

[7] Siehe auch Haney (2000); Kreitzer et al. (1989); CBS-News am 25.8.2004; http://www.cbsnews.com/stories/2004/01/06/60II/main591676.shtml; (16.7.08).

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 afro-amerikanischem 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 (Kozol, 1992; Heilig & Darling-Hammond, 2008).

Wenn aber der staatliche Druck mittels sanktionsbewehrter Tests und Personenevaluation nicht zu einem besseren Unterricht führt, weshalb hält ihn die US-Regierung aufrecht und hat ihn durch das NCLB-Gesetz noch verstärkt? Berliner und Biddle (1995) sowie Bracey (2002) vermuten aufgrund zahlreicher Evidenz, dass dieser Druck nicht der Verbesserung des öffentlichen Schulsystems dient, sondern dessen Abschaffung, einem, wie Bracey (2002) es pointiert ausdrückt, "Krieg gegen das öffentliche Schulwesen". Diese Argumentation erhielt vor kurzem Bestätigung durch eine Insiderin der Bush-Regierung, die in einem Interview mit dem Time-Magazin erzählte, dass es zu ihrer Zeit als stellvertretende Bildungsministerin für Primar- und Sekundarschulen einige im Ministerium gab, die das öffentliche Schulwesen unter Druck setzen wollten, um es reif für die Privatisierung zu machen.[8]

[8] TIME (8.6.2008): "Susan Neuman, a professor of education at the University of Michigan who served as Assistant Secretary for Elementary and Secondary Education during George W. Bush's first term, was and still is a fervent believer in the goals of NCLB. […] There were others in the department, according to Neuman, who saw NCLB as a Trojan horse for the choice agenda — a way to expose the failure of public education and 'blow it up a bit,' she says.

“Kollateralschäden”

Sanktionsbewehrte Tests wurden in der Hoffnung eingeführt, 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 (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 (Sacks, 1999; Wuttke, 2007, S. 163-186).
- Die Schulabbrecher- und Sitzenbleiberrate steigen, was heißt, dass der Anteil der Jugendlichen in einem Jahrgang mit einer *High school graduation* infolge dieser Evaluationspolitik sinkt (Kreitzer et al., 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

---

‘There were a number of people pushing hard for market forces and privatization.’”

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 (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 (Goodman, 2008). Die Jugendkriminalitätsrate hat sich in den USA seit 1980 vervierfacht (Lochner & Moretti, 2004).[9]

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.).

(http://www.time.com/time/nation/article/0,8599,1812758,00.html; 20.7.08)

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

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 und Lehrerinnen, 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 (Lankford et al., 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 (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 Qualifikationsanforderungen an junge Lehrer direkt zu besseren Schülerleistungen (White et al., 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 (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 (Popham, 1999; Sacks, 1999; Kohn, 2000; Amrein & Berliner, 2002; Nichols et al., 2006; Nichols & Berliner, 2006; Baker, 2007). Sie verweisen auf die oft mangelhafte Qualitäts-kontrolle bei der Entwicklung von Schulleistungstests (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.

Programmevaluation

Schon viel länger als die Personenevaluation (und fast unbemerkt bei uns) wird in den USA Evaluation als *Programm*evaluation zur Verbesserung des Unterrichts eingesetzt (vgl. u.a. Sanders,

1994). „The most common name in the U.S. for this developing speciality is evaluation research, which now almost always implies program evaluation," schreibt noch Campbell (1976, S. 1).

Man könnte den Start dieser Bewegung auf das Jahr 1928 festlegen, als Hartshorne und May ihre große Evaluation der Charaktererziehung im Religionsunterricht in den USA vorlegten. Diese Nutzung von Evaluation erlebte einen Höhepunkt in den späten 1960er Jahren, als man forderte, staatliche Reformen als Experimente anzulegen und zu evaluieren (Campbell, 1969).

Programm- und Methodenevaluation, so scheint es, bietet eine wirkliche Chance, Schule und Unterricht zu verbessern und damit das Lernen der Schüler zu fördern, indem sie hilft, ineffektive Bildungsmaßnahmen und Unterrichtsmethoden zu identifizieren und zu eliminieren, und neuen, effektiven Maßnahmen und Methoden den Weg in die schulische Praxis ebnet. Wie viele großen und kleine Evaluationsstudien in den USA zeigen, hat sich diese Hoffnung bislang noch nicht richtig erfüllt, da sich auf dem Weg dahin eine Reihe von Problemen auftürmen, die bislang noch nicht ausgeräumt werden konnten. Ich will das im folgenden anhand von vier außergewöhnlich gründlich evaluierten Bildungsprogrammen bzw. Unterrichtsmethoden in den USA illustrieren, die bei uns mehr Beachtung verdienten, als ihnen bislang zuteil wurde. Jede dieser vier Illustrationen bietet Anschauungsmaterial für die Chancen und die Probleme der Programmevaluation.

"Wie wird jungen Menschen Religion beigebracht und was ist der Effekt davon?" − *Studies in the Nature of Character*

Die Beantwortung dieser Frage, war Aufgabe einer der ersten großen experimentellen Programmevaluationen der Geschichte. Auftraggeber war die *Religious Education Association*, eine Vereinigung mehrerer evangelischer Kirchen in den USA. Hintergrund und Ziel dieser Studie beschreibt

Fisher (1928): “Viele Jahre lange wurde auf den Jahresversammlungen [der *Religious Education Association*] das Problem diskutiert, wie man die Ergebnisse der religiösen Erziehung objektiv evaluieren könnte. Auf dem Treffen im Jahr 1922 nahm das Interesse an diesen Studien die Form einer Frage an: 'Wie wird jungen Menschen Religion beigebracht und was ist der Effekt davon?'” (S. v; meine Übers. GL)

Die beiden mit der Durchführung dieser Studie beauftragten Forscher, Hugh Hartshorne und Mark A. May, führten zur Beantwortung der gestellten Frage mehrere experimentelle Studien zum *Täuschungsverhalten* bzw. *Ehrlichkeit* von Kindern und Jugendlichen in Schulsituationen und Fragebogenstudien zum Betrügen und Lügen durch (Hartshorne & May, 1928). In ihrem Vorwort betonen die Forscher, dass ihnen keinerlei Auflagen gemacht wurden: “Außer einer allgemeinen Definition des Hauptgebiets dieser Studie wurde keinerlei Auflagen gemacht, welche Probleme in Angriff genommen werden sollten und welche Techniken benutzt werden sollten.” (S. 3; meine Übers., GL) Sie wählten, vermutlich als erste, experimentelle und quantitative Methoden, um, wie sie schrieben, das „Produkt der Moralerziehung in Sonntagsschulen“ zu überprüfen (der Religionsunterricht in den USA findet sonntags in kirchlichen Einrichtungen statt, da die Trennung zwischen Kirche und Staat keinen schulischen Religionsunterricht zulässt): “Hunderte Millionen Dollar werden vermutlich jedes Jahr von Kirchen, Sonntagsschulen und anderen Organisationen für Kinder ausgegeben, fast ohne jede Prüfung des Produkts.” (S. 5; meine Übers., GL) Dabei stellten sie fest, dass es für diese Evaluation damals noch keinerlei “Tests zur Vorhersage von Lebenserfolg” gab. In den Mittelpunkt ihrer Untersuchung stellten die Autoren “die Gesetze, die die Beziehung zwischen Ideen, körperlichen Bedingungen und Einstellungen regieren, besonders auf dem Feld der sozialen Beziehungen.” (S. 9; meine Übers., GL)

Die Autoren fassen ihre wichtigsten experimentellen Befunde so zusammen (S. 14-15)[10]:

- "Täuschung ballt sich in bestimmten Familien wie Intelligenz und Augenfarbe [...] was kein Beweis für ihre Erblichkeit ist, aber dass diese Dinge zusammen auftreten."
- "Täuschung geht zusammen mit Cliquen und Klassenzimmer. Ein Schüler ähnelt seinem Freund in dieser Hinsicht."
- "Wo die Beziehung zwischen Lehrer und Schüler durch eine Atmosphäre der Kooperation und des guten Willens gekennzeichnet ist, gibt es weniger Täuschung [...]" Diese Atmosphäre, so schreiben die Autoren an anderer Stelle (S. 411), ist am häufigsten in den so genannten *Progressive schools* anzutreffen, den damals in den USA sehr populären Sekundarschulen ohne Noten und ohne Lernzwang, die durch die Pädagogik von John Dewey angeregt waren (siehe unten).
- "Auf der anderen Seite scheint die Teilnahme an der Sonntagsschule oder die Mitgliedschaft in mindestens zwei Organisationen, die als Ziel die Vermittlung von Ehrlichkeit haben, in dieser Hinsicht keine Wirkung zu haben, und es gibt in einigen Fällen Belege, dass diese die Kinder eher weniger als mehr ehrlich macht." Dieses ist wohl der erstaunlichste Befund, werden dafür doch, wie die Autoren im Vorwort betonten, Hunderte Millionen Dollar jährlich aufgewandt.

Daneben haben diese *Studies in the nature of character* einige wichtige psychologische Grundlagenfragen zu beantworten versucht. Die beiden wichtigsten Befunde sind laut dieser Autoren:

a) Das Täuschungsverhalten scheint stärker durch die jeweilige Situation bestimmt zu sein als durch ein durchgängiges Charaktermerkmal oder Ideal der Ehrlichkeit. Kaum ein Kind versuchte in *allen* experimentellen Situationen, in denen sie in Versuchung geführt wurden, sich durch Täuschung einen Vorteil zu verschaffen und kaum ein Kind ist in allen Situationen ehrlich. Die Autoren

[10] Alle folgenden Zitate wurden von mir übersetzt; GL.

machten allerdings keinen Versuch, diesen Befund damit in Einklang zu bringen, dass sie in anderen Kontexten von "Ehrlichkeit" sprachen, als wenn es sich um ein durchgängiges Charaktermerkmal bei Kindern handelt.

b) Täuschungsverhalten lässt sich kaum auf Motive und Einstellungen zurückführen. Das einzige wirksame Motiv für Täuschung, das sich finden ließ, war der Wunsch der Kinder, in der Schule gute Leistungen zu erbringen (S. 15). Die Korrelation zwischen der Fähigkeit, in einem Leistungstest gut abzuschneiden, und dem Umfang des Täuschungsverhaltens in den Testsituationen war hoch negativ (die Korrelationen reichten von r = -0,05 bis -0,51; S. 395). Allerdings drehte sich diese Korrelation bei Tests mit starken Zeitbegrenzungen (*Speed tests*) um. Bei diesen Tests griffen eher die besseren Schüler zu Täuschungen (Band 2, S. 241). Diese Tests, so die Erklärung der Autoren, wurden von den Kindern weniger als Leistungstests wahrgenommen. Diese Erklärung, so scheint mir, ist eine erstaunliche Vorwegnahme der Befunde von Wuttke (2007) bezüglich der Bedeutung von Tricks beim Ausfüllen der *Speed tests* bei PISA.

Diese Studien weisen einige bedeutsame interne Widersprüche auf, die hier schon angesprochen wurden. Obwohl sie meinen, dass ihre Experimente die Vermutung widerlegt hat, es gebe so etwas wie einen Charakterzug der Ehrlichkeit, gehen sie stillschweigend von einer solchen Vermutung aus, wenn sie "Ehrlichkeit" in Beziehung zur familiären Herkunft, Clique, Religionsunterricht, Intelligenz und Leistungswillen setzen. Obwohl die Autoren beabsichtigten, die inneren psychologischen Gesetze aufzudecken, die das Täuschungsverhalten bedingen und regeln, haben sie das Verhalten ihrer Versuchspersonen nur rein äußerlich studiert, anhand von äußeren Normen für moralisches Verhalten und nicht anhand der moralischen Ideale und Normen der beobachteten Kinder selbst. Dieses Manko räumen sie selbst auf der letzten Seite ihres Forschungsberichts ein: "Die Essenz einer Handlung ist ihre Zielsetzung [*The essence of the act is its pretense*.] Demnach kann sie nur beschrieben und verstanden werden, wenn man das subjektive Element in einer

Situation bedenkt. Dabei ist es nicht Intention des Handelnden, die die Täuschung konstituiert, und auch nicht die spezielle Handlung, sondern die Beziehung seiner Handlung zu seinen Intentionen und den Intentionen seiner Mitmenschen." (S. 377, meine Übers., GL)

Vielleicht haben diese Widersprüche dazu beigetragen haben, dass dieses fünfjährige Evaluationsprojekt ziemlich folgenlos blieb. Die religiöse Unterweisung in Sonntagsschulen wird heute noch so betrieben wie vor dieser Studie. Zudem beschloss die US-Regierung vor einigen Jahren ungeachtet der negativen Befunde von Hartshorne und May ein Millionen-teueres Programm zur *Character education*.

Kann die Diskussion moralischer Dilemmas die moralische Urteilsfähigkeit fördern? – *Meta-Analyse der Blatt-Kohlberg-Methode*

Die von Hartshorne und May (1928) angesprochene, aber in ihren eigene Studien vernachlässigte "Beziehung des Verhaltens zu den Intentionen" einer Person und anderer Personen wurde von Piaget (1932/1972) zum zentralen Gegenstand seiner Forschung gemacht. Später griff Kohlberg (1964) diese Forschung auf, indem er – als Bindeglied zwischen moralischen Orientierungen und Verhalten – die moralische Urteilsfähigkeit identifizierte, nämlich die Fähigkeit, moralische Intentionen in moralisches Verhaltens zu übersetzen. Er nahm an, dass vor allem ein Mangel an Fähigkeit, Konflikte zwischen widerstreitenden moralischen und außermoralischen Motiven zu lösen, verantwortlich für unmoralisches Verhalten ist (Lind, 2002). Es lag nahe, eben diese Auseinandersetzung mit solchen moralischen Konflikten oder Dilemmas zur Grundlage einer effektiven Moralerziehung zu machen. Damit war die didaktische Idee der "Dilemmadiskussion" geboren (Blatt & Kohlberg, 1975; Lind, 2003). Der hohe Anspruch Kohlbergs, den er durch eine umfassendere Theoriebildung

("kognitiv-entwicklungstheoretisches Ansatz") untermauerte, forderte geradezu zur Evaluation der Dilemmamethode heraus. Kann die Diskussion moralischer Dilemmas die moralische Urteilsfähigkeit effektiv fördern?

Die ersten Evaluationsstudien stützten sich, wie das damals üblich war, auf eine Auszählung von "signifikanten" Ergebnissen. So fand Leming (1981), dass "of the 27 studies reviewed utilizing the cognitive conflict strategy to stimulate moral development 22, or 81 percent, found significant differences in favor of the treatment groups." (S. 160) So beeindruckend sich dieser positive Befund von dem negativen Ergebnisse des Religionsunterricht abhob, so wenig wurden in diesen Auszählungen von "signifikanten" Befunden aber deutlich, wie groß der Effekt der Dilemmamethode von Blatt und Kohlberg wirklich war. Darum suchten Leonore Link und ich Mitte der 1980er Jahre alle greifbare Interventionsstudien zur Dilemmadiskussion (wir fanden 141) und unterwarfen sie einer Meta-Analyse (Lind, 2003). Leider berichtete keine einzige dieser Studie die Effektstärke der Intervention und nur bei 74 Studien konnten wir (zum großen Teil in schwer zugänglichen Dissertationen und internen Forschungsberichten) genügend Informationen ermitteln, um nachträglich die Effektstärke zu berechnen. Durch die große Zahl von Evaluationsstudien ergaben sich gute Anhaltspunkte für die Generalisierbarkeit der Effekte.

Das Ergebnis war überraschend positiv (siehe Lind, 2002). Wir fanden eine durchschnittliche Effektstärke der Blatt-Kohlberg-Methode von $r = 0{,}40$ (Median). Dieser Wert liegt deutlich über der mittleren Effektstärke von $r = 0{,}30$, die von Lipsey und Wilson (1993) als Mindestmarke für *effektive* Programme bezeichnet wurden. Keine der Interventionen produzierte übrigens eine Abnahme der moralischen Urteilsfähigkeit. Durch diese breite Evaluation konnte auch die Meinung widerlegt werden, dass Moralerziehung erst im Erwachsenenalter zu Erfolgen führen würde. Die höchsten Effektstärken fanden sich in der Gruppe der Zehn- bis Sechzehnjährigen. Schließlich

zeigte sich, dass auch einige andere Unterrichtsmethoden wie Rollenspiel vergleichbare Effekte haben konnten, dass die Effekte der Dilemmadiskussionsmethode aber nachhaltiger waren.

Effektstärkemaße und Meta-Analysen haben es möglich gemacht, den Fortschritt, der auf dem Gebiet der Moralerziehung gemacht wurde, besser sichtbar zu machen, als dies mit Signifikanztests möglich war. Blatt und Kohlberg (1975) gingen sogar noch einen Schritt weiter, indem sie nicht nur Signifikanzwerte und relative Effektstärken (siehe unten) berichteten, sondern auch den absoluten Betrag, um den die moralische Urteilsfähigkeit durch die Dilemmadiskussion gefördert werden konnte. Sie berichteten einen durchschnittlichen Zuwachs an moralischer Urteilsfähigkeit von einer halben Kohlberg-Stufe, das sind umgerechnet auf eine Vergleichsskala von 0 bis 100 ca. *acht* Punkte. In den 74 Interventionsstudien, die wir analysierten, ergab sich insgesamt im Mittel ein Zuwachs von *sechs* Punkten. Dieser Wert liegt deutlich über dem Zuwachs, den Schüler aufweisen, die nicht nach dieser Methode lernen. In deutschen Sekundarschulen fanden wir einen mittleren jährlichen (!) Zuwachs von *drei* Punkten und in eintägigen Berufsschulen wie auch in Sekundarschulen mit eher traditionellem Unterricht sogar einen Rückgang der moralischen Urteilsfähigkeit (Lind, 2002). Damit konnte erstmals gezeigt werden, dass die Entwicklung moralischer Fähigkeiten von dem modernen Unterricht profitiert, wie er seit den 1970er Jahren in (West-)Deutschland praktiziert wird, und dass die Effektivität und auch die Effizienz moralischer Bildung durch systematische Unterrichtsmethoden auf der Grundlage psychologischer Forschung deutlich gesteigert werden kann. Mit der auf Blatt und Kohlberg aufbauenden *Konstanzer Methode der Dilemma-Diskussion* (KMDD) kann diese Effektivität der moralischen Bildung inzwischen auf 15 bis 17 Punkte Zuwachs in wenigen Unterrichtsstunden gesteigert werden. Eine thailändische Forschergruppe hat diese hohe Effektivität der KMDD in einem randomisierten Interventionsexperiment bestätigt (Lerkiatbundit et al., 2006). Allerdings ist dafür eine intensive Ausbildung der KMDD-Lehrer notwendig (Lind, 2003).

Diese erfolgreiche Entwicklung wurde maßgeblich durch zwei Dinge ermöglicht, a) durch eine ständige Selbstevaluation der Methode (Lind, 2004a; o.J.) und b) durch den ständigen Austausch mit der Grundlagenforschung in diesem Bereich (Lind, 2002; 2008). Einen wichtigen Beitrag zur theoretischen Fundierung der KMDD-Evaluation liefern inzwischen auch hirnphysiologische Studien. Mittels bildgebender Verfahren wurde gezeigt, dass die moralische Urteilsfähigkeit deutlich mit Aktivitäten in dem rechten dorsolateralen präfrontalen Kortex korreliert (Prehn et al., 2008). Dieses Hirnareal ist für die motorische Planung, Organisation und Regulation zuständig. Es spielt eine wichtige Rolle bei der Integration von sensorischen und mnemonische Informationen und der Regulation intellektueller Funktionen und Aktionen.

Die Evaluation der Dilemmamethode hat inzwischen auch praktische Folgen. Die KMDD wird seit einiger Zeit weltweit in Schulen, Hochschulen und anderen Bildungsbereichen wie Bundeswehr und Gefängnissen eingesetzt. Noch handelt sich in vielen Fällen um begrenzte Pilotprojekte. Aber in dem Maß, wie die (notwendigerweise) gründliche Ausbildung von KMDD-Lehrern vorankommt, weitet sich auch der Einsatz dieser Methode aus, bei der Selbstevaluation ein integraler Bestandteil ist.

Kann die Änderung des Lernklimas an einer Schule zur Verbesserung akademischer und moralischer Fähigkeiten beitragen? – Die *Eight-Year Study*

Die von Hartshorne und May (1928) angesprochenen *Progressive education*-Schulen haben sich im späten 19. Jahrhundert in den USA gebildet, zeitgleich mit der deutschen Reformpädagogik, von der sie auch wichtige Impulse empfangen hat. Der bekannteste Theoretiker dieser Bewegung war John Dewey (1964/1915), der in seinem Buch *Democracy and education* und anderen Veröffent-

lichungen die Prinzipien dieses schulischen Bildungsansatzes herausgearbeitet hat. Zwar liegen mir keine genauen Zahlen vor, aber man kann davon ausgehen, dass es bis in vor dem Zweiten Weltkrieg in den USA auf allen Schulstufen sehr viel mehr Reformschulen gab als in Deutschland. Akademische und moralische Bildung in einer Demokratie, so Dewey, soll auf eigenen Erfahrungen gründen und durch die eigene Lernbegierde vorangetrieben werden. Die meisten *Progressive schools* hatten daher auch die Noten abgeschafften, selbst in den *High schools*.

Die meisten Hochschulen in den USA waren daher skeptisch und versagten den Abschlüssen an diesen Reformschulen lange Zeit die Anerkennung. War diese Skepsis begründet? Werden die Kinder durch diese Schulen wirklich zu wenig aufs Leben mit all seinen Leistungsanforderungen vorbereitet, und müssen die Absolventen solcher Schulen am Ende im Leben versagen?

Wie bereits berichtet wurde, stellten schon Hartshorne und May (1928) als Nebenbefund ihrer Studie fest, dass die Schüler von *Progressive schools* in ihren Moraltests besser abschnitten als die Schüler der traditionellen Schulen. Im Durchschnitt zeigten sie weniger Täuschungsverhalten in den verschiedensten Experimenten. Im Jahr 1930 etablierte die *Progressive Education Association,* der Zusammenschluss von *progressiven* Schulen, eine Kommission zur Beziehung zwischen Schule und College, die eine Anpassung des Curriculums an den *Progressive schools* und eine Langzeitstudie über die Relevanz des *progressiven* Curriculums und dessen Einfluss auf Erfolg oder Misserfolg beim Studium durchführen sollten. Die Kommission unter ihrem Vorsitzenden W. M. Aiken schaffte es, 300 Colleges und Universitäten in den USA zur Zusammenarbeit zu bewegen. Bis 1933, nachdem 30 öffentliche und private *Progressive schools* ihre Curriculumreform abgeschlossen und sich dem Studienprojekt angeschlossen haben, nahmen die ersten Absolventen dieser Schulen ohne die üblichen Aufnahmetests ein Collegestudium auf und wurden in der Folgezeit von einer Forschergruppe begleitet. Diese Forschergruppe wurde geleitet von Ralph W. Tyler (Smith & Tyler, 1942), der später Chefberater von Präsident Johnson für das *Project Head Start* wurde und

auch maßgeblich an der Einrichtung des *National Assessment of Educational Progress* (NAEP) beteiligt war.

Die zentrale Fragestellung der Acht-Jahres-Studie war: “Is the traditional college-entrance program the only safe and sound plan of preparation for college? Or can boys and girls be equally well – or possibly even better – prepared for college through a considerable variety of widely different programs, devised by competent secondary-school teachers, with their eyes fixed primarily on the conditions and demands of modern life and the individual capacities and interests of particular students, with only incidental reference to the impending college experience?" (Chamberlain et al., 1942, S. xix)

Die Antwort auf diese Frage, so die Autoren, war ein klares “Ja”. (S. xii) Die Absolventen der progressiven Schulen waren den anderen in den meisten Dingen leicht überlegen (S. 207-208; meine Übers., GL):

- “Sie erzielten einen etwas höhere Notendurchschnitt [als die Vergleichsgruppe; GL];
- Sie erzielten besser Noten in alle Fächern außer Fremdsprachen;
- Sie wählten dieselben Fächerschwerpunkte wie die Vergleichsgruppe;
- Sei erhielten jedes Jahr etwas mehr Auszeichnungen;
- Sie wurden öfter positiv beurteilt, was ihre intellektuelle Neugierde und Wissbegierde anging;
- Sie wurden öfter in ihrem Denken als präzise, systematisch und objektiv eingeschätzt;
- Von ihnen wurde öfter gesagt, dass sie klar entwickelte und gut formulierte Ideen haben;
- Sie nahmen an allen organisierte Studenten-Organisationen teil außer in religiösen und “Service”-Aktivitäten;
- Sie zeigten eine bessere Orientierung, was ihre Berufswahl angeht;
- Sie zeigten sich mehr aktiv besorgt um das, was in der Welt vor sich ging.”

Schon damals hat man also herausgefunden, dass es nicht des Drucks durch Noten bedarf, damit Schüler gute Leistungen zeigen, und dass eine integrierte (nicht durch Fächer und Schulformen aufgesplitterte) Schule besseres Lernen ermöglicht. Sie zeigt vor allem, dass auch in der Sekundarschule möglich ist, was heute als "offener Unterricht" in Kindergärten und Grundschulen mit Erfolg praktiziert wird (Peschel, 2002).

Leider ist die Publikation dieser Studie im Jahr 1942 durch die weltpolitischen Ereignisse an den Rand gedrängt worden. Sie fand kaum eine Resonanz und konnte auch nicht verhindern, dass die *Progressive education*-Bewegung – im Zug des scharfen Antikommunismus in den USA nach dem Zweiten Weltkrieg – durch eine beispiellose politische Rufmordkampagne als innerer Feind gebrandmarkt und dadurch fast völlig ausgelöscht wurde (Bracey, 2007). Evaluation erfolgreich – Patient tot, könnte man bilanzieren. Es scheint aber, dass die Acht-Jahres-Studie ebenso wie die Reformpädagogik heute wieder mehr Aufmerksamkeit erhält.

Bekommen benachteiligte Kinder durch acht Wochen Vorschule so viel Vorsprung, dass sie in der Schule mithalten können? – Das *Project Head Start*

Angestoßen durch PISA wird heute bei uns die Frage diskutiert, wie man Kindern aus benachteiligten sozialen Milieus am wirksamsten helfen kann, um den Teufelskreis aus Armut–schlechte Bildung– Armut zu durchbrechen. Das Hauptaugenmerk fällt dabei auf die Vorschule, weil man glaubt, dass man hier am wirkungsvollsten ansetzen kann, um Benachteiligungen auszugleichen. Das Hauptargument ist, das diese Altersphase für die spätere intellektuelle und moralische Entwicklung eine große Bedeutung hat oder gar entscheidend ist (gemäß dem Sprichwort "Was Hänschen nicht lernt, lernt Hans nimmer mehr"). So fiel denn auch das Augenmerk der US-

Regierung auf diese Altersphase, als darum ging, die Armut in den USA wirkungsvoll zu bekämpfen, um im Wettkampf der Supermächte bestehen zu können. Der Sieg der UdSSR (Sowjetunion) im Wettrennen um den Start des ersten Satelliten löste in den USA den "Sputnik-Schock" aus. Man fragte sich, ob am Ende der Kommunismus überlegen ist, weil durch seine Sozial- und Bildungspolitik sich mehr um die Ressource Mensch kümmert?

Die Verantwortlichen in der US-Regierung hofften, mit dem Sofortprogramm, das aus einer achtwöchigen Vorschule für Kinder aus sozial schwachen Milieus bestand, einen Vorsprung (englisch: *Head start*) gegenüber gleichaltrigen Kindern mit besseren Voraussetzungen zu verschaffen, damit sie besser mit den Anforderungen der Schule mithalten können. Die baulichen Voraussetzungen für dieses Programm waren gegeben; die Schulen in den USA stehen im Sommer drei Monate lang leer. Die personellen Voraussetzungen jedoch waren nicht vorhanden. In so kurzer Zeit konnten keine qualifizierten Lehrkräfte für dieses landesweite Projekt gewonnen werden. Man setzte daher meist nur angelerntes Personal, vor allem Mütter, Krankenschwestern und Gemeindehelferinnen ein. Soweit man das heute beurteilen kann, war die Qualität des *Head start*-Programms damals sehr unterschiedlich und insgesamt sehr niedrig (Zigler & Muenchow, 1992; Currie & Thomas, 2000; Biedinger & Becker, 2006).

Man ist sich unter Experten weitgehend einig, dass das *Project Head Start* ein Fehlschlag war. Die teilweise nachweisbaren Vorteile der *Head start*-Teilnehmer in den Schulleistungen sind oft schon am Ende des ersten Grundschuljahres nicht mehr beobachtbar. Hinsichtlich einiger Aspekte zeigten sich sogar Negativeffekte, die aber nicht eindeutig interpretierbar sind. Zwar wurden auch vergleichende Studien durchgeführt, aber man kann nicht ausschließen, dass (Selbst-) Selektionseffekte für diese Unterschiede verantwortlich sind. Es ist denkbar, dass überproportional viele Kinder mit Lern- und Verhaltensproblemen zu den *Head start*-Schulen geschickt wurden – ein Effekt, der sich mit statistischen Mitteln nicht herausrechnen lässt. Sofern längerfristige Effekte

feststellbar waren, waren diese – abgesehen von den qualifiziert durchgeführten Programmen (s. u.) – meist sehr schwach und nicht generell, sondern auf bestimmte Untergruppen (wie Weiße) beschränkt. Die Hoffnung also, dass mit einer solchen Maßnahme in einer vermeintlich sensiblen Phase der kindlichen Entwicklung die Armut in den USA verringert werden könnte, hat sich nicht bewahrheitet. Im Gegenteil, in Bezug auf die spätere Berufskarriere nimmt die Bedeutung des Faktors Fähigkeit und Ausbildung in den USA immer mehr ab, und der Einfluss des elterlichen Einkommens darauf immer mehr zu (Belley & Lochner, 2008).

Dieser Fehlschlag scheint zwei Gründe zu haben. Zum einen wissen wir, dass die Theorie der ‘sensiblen Phasen’ auf Säugetiere und vor allem auf Menschen nicht so eindeutig zutrifft wie auf Gänse. “In der Regel können [...] sensible Phasen bei Säugetieren nicht so genau umrissen werden wie bei den daraufhin untersuchten Vogelarten; vielmehr hat es den Anschein, dass sich die sensible Phase über den größten Teil der Kindheits- und Jugendentwicklung erstreckt. [...] Säugetiere bleiben zeitlebens für Erfahrungen 'offene Systeme', und Verhaltensentwicklung ist bei ihnen eher ein kontinuierlicher Prozess.” (Sachser, 2006, S. 24) Dennoch kann es für viele Kinder wichtig sein, dass schwerwiegende organische und kognitive Störungen frühzeitig erkannt und behandelt werden, weil sie später nur mit einem sehr viel höheren Aufwand behoben werden können und bis dahin viel Leid und viele Kosten erzeugen können (Radigk, 1986). Um solche Diagnosen zu stellen und geeignete Therapien durchzuführen bedarf es jedoch *gut ausgebildeter* Lehrkräfte.

Der zweite Grund für das Scheitern dieses Projekts dürfte darin gelegen haben, dass die Verantwortlichen sich zu wenig Sorge um die Lehrkompetenz des angeheuerten Personals gemacht haben, das meist nicht die dafür notwendige Ausbildung hatte. Wie wichtig das Qualifikationsniveau des Lehrpersonals für den Erfolg solcher Interventionen ist, zeigte sich im “Perry-Projekt”. Dort bekam die Preschool-Gruppe ein *qualitativ hochwertiges* Vorschulprogramm, nämlich täglich 2½ Stunden Unterricht durch *geschulte Lehrer* und zusätzliche Hausbesuche in der Familie, die

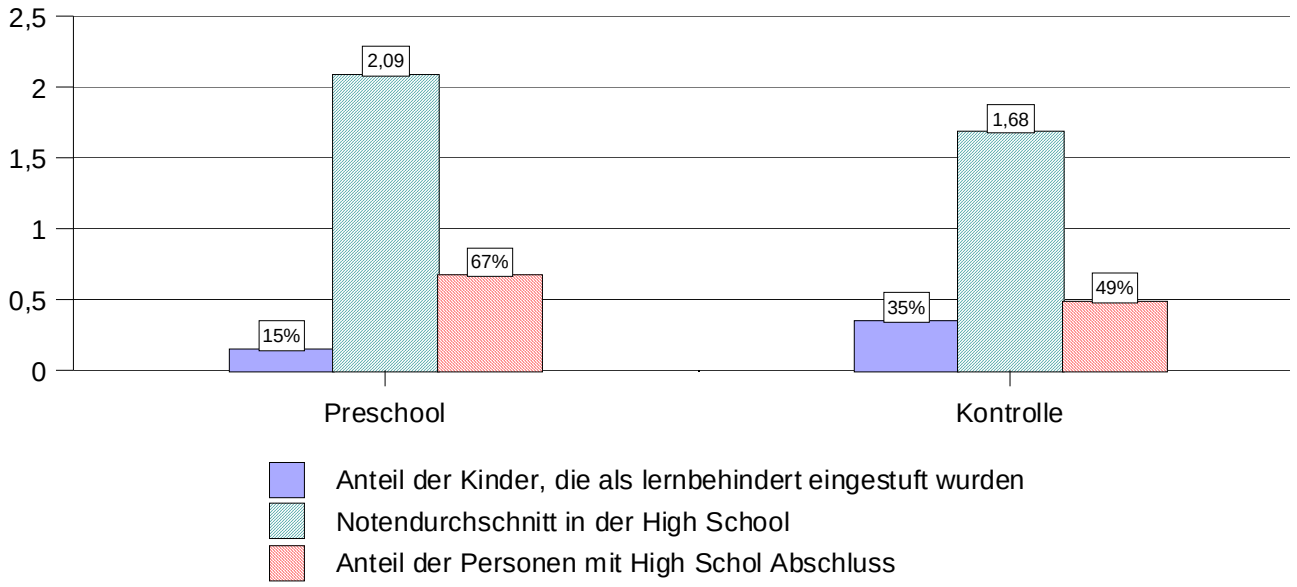

Abb. 2 Langfristiger Bildungserfolg des Perry Preschool-Projekts 1962 - 65 in Ypsilanti, Michigan, mit 123 afro-amerikanischen Kindern im Alter zwischen drei und fünf Jahren (Quelle: Schweinhart et al., 1993; Berrueta-Clement et al., 1984; zitiert nach Biedinger & Becker, 2006, S. 8). Beste Note A = 4, B = 3, C = 2, D = 1, F = 0.

Kontrollgruppe bekam nichts. Die Kinder wurden per Losentscheid auf beide Gruppen aufgeteilt, so dass weitere Ursachenfaktoren weitgehend ausgeschaltet waren. Die Teilnehmer wurden später noch mehrmals befragt, das letzte Mal im Alter von ca. 40 Jahren.

Die Investition in qualifizierte Lehrkräfte hat sich "bezahlt" gemacht. Das Perry-Projekt hatte erstaunlich langfristige Bildungseffekte und Effekte im Sozialverhalten (Abb. 2). Die Kinder, die Gelegenheit bekamen, zwei Jahre lang die personell gut ausgestattete Preschool zu besuchen, wurden später seltener als lernbehindert eingestuft, zeigten einen besseren Notendurchschnitt (in unser System umgerechnet 2,91) gegenüber der Vergleichsgruppe (3,32) und erreichten häufiger den High School-Abschluss. Zudem wurde ihr Sozialverhalten im Alter von 6 bis 9 von den Schullehrern deutlich besser eingestuft und zeigten sie sich später weniger häufig straffällig (2,3 gegenüber 4,6 Verhaftungen). Zigler und Muenchow (1992) rechnen vor, dass sich dieses Projekt allein

durch die geringere Kriminalitätsrate auch volkswirtschaftlich hoch verzinst hat. Andere Vorschulprojekte bestätigten diese langfristigen Erfolge frühkindlicher Förderung, allerdings immer nur, wenn diese Projekte mit qualifizierten und angemessen bezahlten Lehrkräften ausgestattet waren (vgl. die Übersicht bei Biedinger & Becker, 2006).

Lässt sich das Lesenlernen durch "Evidenz-evaluierte" Methoden fördern? – *Reading First*

Alle bisher genannten Programmevaluationen sind Beispiele dafür, wie offen und kontrovers in den USA die wissenschaftliche Diskussionen über schulische Interventionsprogramme geführt wird (auch wenn diese Evaluationen von den politischen Entscheidungsträgern bislang selten beachtet werden). Mit dem wachsenden Vordringen kommerzieller Anbieter von Schulprogrammen und Evaluationen wird die öffentliche Diskussion darüber zunehmend erschwert. Die Methoden der Evaluation und die Tests werden zum Geschäft und zum Geschäftsgeheimnis. Welche Auswirkungen diese Entwicklung auf die Qualität der Evaluation und damit auf die Qualität der Schule und das Lernen der Kinder hat, lässt sich an dem *Reading first*-Programm ablesen, das die Vorlage für viele Programme ist, die den Schulen *kommerziell* angeboten oder von der Regierung aufgezwungen werden.

Reading first beansprucht eine wissenschaftlich evaluierte und effektive Unterrichtsmethode zu sein. Auf der Webseite des US-Bildungsministeriums findet sich dieser Text: "Dieses Programm fokussiert darauf, erprobte Methoden des frühen Lesetrainings in das Klassenzimmer zu bringen. Durch *Reading first* erhalten die Bundesstaaten und die [Schul-]Distrikte Unterstützung, um wissenschaftlich fundierte Leseforschung – und die erprobten Lehr- und Testwerkzeuge, die mit

dieser Forschung übereinstimmen – anzuwenden, um sicher zu stellen, dass alle Kinder bis zum Ende der dritten Klasse gut zu lesen lernen." (U.S. Department of Education, 2008; meine Übersetzung, G.L.).

Viele renommierte Wissenschaftler aus Universitäten und Schul- und Testentwicklungsfirmen haben bei der Entwicklung und Vermarktung als Berater für dieses fast 5 Mrd. Dollar teuere Unterrichtsprogramm fungiert. Viele dieser Berater entwickeln aber selbst Unterrichtsprogramme für die Regierung und ihre nachgeordneten Behörden. In vielen Fällen scheint daher der "wissenschaftliche Rat" dieser Experten nicht ganz frei von Eigennützigkeit gewesen zu sein (siehe auch Bracey, 2005).

Jetzt hat der Bundesrechnungshof der USA anhand unabhängiger wissenschaftlicher Studien festgestellt, dass dieses Programm *wirkungslos* und dass die Bezuschussung des *Reading first*-Programms in Höhe von einer Milliarde Dollar pro Jahr nicht zu vertreten ist. Das Verdikt des Rechnungshofs schaffte es auf die Titelseiten von nationalen Zeitungen wie *USA TODAY* und *New York Times*. *Education Week*, das führende Bildungsmagazin machte diese Entscheidung mit folgenden Worten publik:

"HOUSE APPROPRIATIONS CANCELS READING FIRST"

> Under a fiscal 2009 spending measure approved unanimously last week by a House Appropriations subcommittee, the controversial federal Reading First program would be eliminated [...]. Explaining the decision to zero out the program, Representative David R. Obey (D-Wis.), the chairman of the House Appropriations Committee, cited the results of a preliminary federal evaluation of Reading First released May 1 that found the program has had no impact on students' reading comprehension. Reading First 'has been plagued with mismanagement, conflicts of interest, and cronyism, as documented by the inspector general,' Representative Obey said, referring to a series of reports by the federal watchdog that suggested conflicts of

interest among officials and contractors who had implemented the program in its early years." (Education Week, 2008; siehe auch Toppo, 2008)

Ob die Evaluation von *Reading first* durch den Rechnungshof wirklich Folgen haben wird, ist offen. Zwar hat der zuständige Haushaltsausschuss beschlossen, die jährlich 1 Mrd. Dollar für das Programm auf ein Drittel zu kürzen. Aber Präsident Bush hat bereits sein Veto dagegen eingelegt. Das *High-stakes*-Problem hat inzwischen auch die Programmevaluation erreicht und droht nun auch hier Sinn und Zweck von Evaluation zu untergraben. Ein Hoffnungsschimmer: Ins Rollen gebracht wurde die Untersuchung gegen *Reading first* von einem Konkurrenten, Robert Slavin, der sich gegenüber der *New York Times* beklagte, dass sein Programm/Produkt *Success for all* und andere durch die Präferenz der Regierung für *Reading first* vom Markt ausgeschlossen würde.

Ungelöste Probleme der Programmevaluation

Die oben beschriebenen Evaluationsprojekte sowie viele andere Studien, auf die ich hier aus Platzmangel nicht eingehen kann, lassen die Möglichkeiten, aber auch die noch ungelösten und neuen Probleme der Programmevaluation in den USA erkennen:

- Was ist als "Erfolg" einer Methode zu werten? Ist es ein Erfolg, wie die meisten Autoren offenbar annehmen, wenn sich mit einer bestimmen Unterrichtsmethode statistisch "signifikante" Verbesserungen nachweisen lassen gegenüber herkömmlichen Methoden? Wenn das der Fall wäre, würde der Erfolg einer Methode überwiegend vom Forschungsbudget abhängen, da sich jeder noch so kleine Unterschied durch die Vergrößerung der Stichprobe in ein "signifikantes" Ergebnis verwandeln ließe (siehe u.a. Carver, 1993). In der Tat stellt die weit verbreitete

Art der Anwendung der Signifikanztests dessen Logik auf den Kopf. Sie machen nur Sinn, wenn *zuvor* festgelegt wird, wie groß der Unterschied sein soll, den man als Erfolg ansehen will. Dann kann man entscheiden, wie groß die Stichprobe sein muss, um diesen *zuvor* festgelegten Unterschied fehlerfrei messen zu können (Sedlmeier & Köhlers, 2001).

- Lässt sich der Erfolg einer Methode oder eines Programms in der Schule durch ein so genanntes *Effektstärkemaß*, wie ~~die~~ den Korrelationskoeffizienten *r* oder Cohens *d*, besser ausdrücken, wie es seit 2001 von der *American Educational Research Association* und der *American Psychological Association* (APA, 2001) vorgeschrieben ist? Die Idee dieser Maße ist einfach (Lind, 2007): Eine Methode hat einen umso größeren Effekt, je größer die Differenz zwischen Nachtest- und Vortestwert ist. Damit Messfehler und andere Schwankungen der Messwerte in Rechnung gestellt werden können, wird der Effekt jeder Methode durch mehrmaliges Messen bestimmt (Mittelwerte) und die Differenz durch die Schwankungsbreite (Varianz) der Messwerte geteilt. Diese Maße der Effektstärke haben zweifellos den Vorteil, dass sie nicht, wie die „Signifikanz"-Maße von der Größe der Stichprobe abhängen. Aber auch sie sind nicht optimal. Bei Maßen wie *r* und *d* werden Messfehler zum Schiedsrichter über Erfolg und Misserfolg von Unterrichtsmethoden und Schulprogrammen gemacht, statt dass man sich bemüht, diese Fehler auszuschalten. Sie führen, wie Rosenthal und Rubin (1982) zeigen, oft zu einer Unterschätzung der Effekte. Zudem bieten auch sie eine Möglichkeit der Manipulation. Man kann durch eine geschickte Auswahl der Untersuchungsgruppen den Messfehler klein halten, so dass geringe Effekte groß erscheinen.
- Als Alternative bietet sich das „einfache Ablesen" von Unterschiedswerten an, wie das in den Naturwissenschaften vielfach Brauch ist. Voraussetzung für das einfache Ablesen ist aber, dass wir mit der Messdimension und der Bedeutung der Messwerte vertraut sind, das heißt, dass es

diese Messdimension schon lange gibt.[11] Daher ist es ein großes Manko der Evaluationsforschung im Bildungsbereich, dass es hier meist nur "junge", ad hoc formulierte Evaluationsdimensionen gibt, über deren Validität und Bedeutung wir meist wenig wissen. Die Bedeutung einer Messdimension wie zum Beispiel 'mathematische Grundkompetenz' ergibt sich erst aus der langjährigen Erfahrung mit der Messung und einer gut untermauerten Theorie dieser Kompetenz. Was auf den ersten Blick wie ein guter Test für diese Kompetenz aussieht, kann sich bei genauerer Analyse als ein Test für etwas anderes herausstellen (Wuttke, 2007). Eine Mathematikaufgabe, die einen langen Einleitungstext hat und unter großem Zeitdruck gelöst werden muss, misst in Wahrheit nicht mathematische Fähigkeiten, sondern die Fähigkeit zum Schnelllesen oder Stressresistenz.

- Als neues Problem der Programmevaluation tritt die Vermengung von starken geschäftlichen Interessen mit dem Interesse nach Erkenntnisgewinn und Qualitätssicherung im Bildungsbereich in Erscheinung. Sobald schulische Programme und Unterrichtsmethoden nicht mehr frei ausgewählt und eingesetzt werden können, sondern nach Maßgabe des Staates von Verlagen gekauft werden müssen, bleibt für den Einsatz von Evaluationen und Qualitätssicherung kaum noch Spielraum. Was dem wirtschaftlichen Erfolg der Anbieterfirmen dienen kann, kann durchaus das Lernen an den Schulen ruinieren.

[11] Ein Beispiel für eine gut vertraute Messdimension ist die Temperatur in mittleren Bereichen. Dort lesen wir einfach das Thermometer ab und bewerten dann die abgelesene Differenz der Temperatur. Zwei Grad weniger sind, bei genügend großer Anzahl von 'zufälligen' Messungen statistisch immer signifikant, aber sie sind nicht unbedingt für unser Handeln bedeutsam, etwa für die Frage, ob wir uns wärmer anziehen sollen. Andererseits können fünf Grad weniger für unser Verhalten praktisch signifikant sein, auch wenn wir das Thermometer jeweils nur einmal ablesen, so dass die Differenz im statistischen Sinne nicht "signifikant" werden kann.

Dies sind gewiss nicht alle Probleme, die in den über 80 Jahren Programmevaluation in den USA zutage getreten sind (siehe Smith, 1986; Sanders, 1994; Schoenfeld, 1999; Shepard, 2002; Rhoades & Madaus, 2003; Lind, 2004; Baker, 2007). Sie zeigen, dass nicht nur die Personenevaluation mit großen Problemen zu kämpfen hat. Dennoch erscheinen viele Probleme der Programmevaluation beherrschbar zu sein, so dass sie prinzipiell dazu beitragen kann, die Qualität der Schule mittel- und langfristig anzuheben. Ein Problem allerdings, das bislang kaum in den Griff zu bekommen war, scheint das Problem der Umsetzung von Ergebnissen zu sein. In vier von fünf oben genannten Beispielen für Programmevaluation, spielten die Ergebnisse der zum Teil sehr aufwändigen Evaluationsstudien keine oder eine sehr geringe Rolle bei einschlägigen politischen Entscheidungen. Effektive Projekte wurden zunichte gemacht und schlechte Projekte ungerührt weitergeführt. Nur im Fall der Evaluation der Dilemmadiskussion hat Evaluation wirklich eine entscheidende Rolle bei der Verbesserung und der Einführung einer Unterrichtsmethode gespielt. Hier wurde Programmevaluation konsequent als wissenschaftliche Selbstevaluation durchgeführt (Barber, 1999; Lind, o.J.) und dadurch freigehalten von kommerziellen Interessen und direkten Kontrollansprüchen der Regierung. Solche Selbstevaluation würde sich gute ergänzen mit einer externen Meta-Evaluation (Dubs, 2005), wie dies im Qualitätsmanagement in der Wirtschaft bereits erfolgreich praktiziert wird (Deming, 1994).

Programmevaluation im Bildungsbereich, so zeigen die Erfahrungen in den USA, ist vor allem auf die Entwicklung von validen Tests angewiesen, deren praktisch-pädagogische Bedeutung uns vertraut ist. Die Entwicklung solcher Tests ist jedoch sehr zeit- und forschungsaufwändig, wie sich anhand der Entwicklung eines neuen Tests zur Messung der moralischen Urteilsfähigkeit gezeigt

hat (Lind, 2008). Sobald solche Tests aber in *High stakes*-Umgebungen eingesetzt werden, "verbrennen" sie schnell (Linn, 2000). Alle von testbasierten Sanktionen Betroffenen trachten danach, die Schwierigkeit des Tests durch "Tricks" abzumildern, was unausweichlich zur Ungültigkeit des Tests führt, so dass in immer kürzeren Abständen neue Tests entwickelt und validiert werden müssen. Wirklich valide Tests, so scheint es, können nur in einer Umgebung gedeihen, in denen keine Anreize zum Betrug bestehen.

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 knowledge as well as in new assessment strategies, if the curriculum goals are to be achieved." (S. 364)

References

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

Amrein, A. & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis, 10,* No. 8. Arizona: Education Policy Studies Laboratory. http://epaa.asu.edu/epaa/v10n18/ (12.10.2004).

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. *Education Policy Analysis Archives, 11*(25). http://epaa.asu.edu/epaa/v11n25/ (12.10.2004).

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

Baker, E. (2007). The end(s) of testing. *Educational Researcher, 36* (6), 309-317.

Barber, L. W. (1999) Self-assessment. In: In: J. Millman & L. Darling-Hammond, Hg., *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. Http://economics.uwo.ca/centres/cibc/ (28.7.08)

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

Berliner, D. C. & Biddle, B. J. (1995). *The Manufactured Crisis: Myths, Fraud, and the Attack on America's Public Schools*. New York: Addison-Wesley-Longman.

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. *Journal of Moral, Education, 4,* 129-161.

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

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

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

Bracey. G. W. (2007). The First Time 'Everything Changed'. *The 17th Bracey Report on the Condition of Public Education*. http://www.america-tomorrow.com/bracey/EDDRA/k0710bra.pdf (1.8.2008)

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. *American Psychologist, 24*, 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. *Journal of Experimental Education, 61*(4), 287-292.

Chamberlin, D., Chamberlin, E. S., Drought, N. E. & Scott, W. E. (1942). *Did they succeed in college? Adventures in American education*. Volume 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, http://www.nber.org/papers/w7888 (15.7.2008)

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

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

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

Darling-Hammond, L. & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? *Educational Researcher, December 2002*, 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, 125*, 627-668.

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

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 A. Bartz et al., Hg., *PraxisWissen SchulLeitung*. Nr. 22.15, Neuwied: Luchterhand.

Education Week (2008). http://www.edweek.org/ew/articles/2008/06/24/43senate_web.h27.html (25.6.2008).

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

Fisher, G. M. (1928). Foreword. In: Hartshorne and May, Hg., 1928 (siehe unten), 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. *Mother Jones*, July 21, 2008. http://www.motherjones.com/cgi-bin/print_article.pl?url...nes.com/news/feature/2008/07/slammed-hard-time-out.html (22.7.08).

Haney, W. M. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8* (41). http://epaa.asu.edu/epaa/v8n41/ (1.7.2008)

Haney, W. M. (2002). Ensuring failure: How a state's achievement test may be designed to do just that. *Education Week, 10 July 2002*, 56, 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", Grace Dodge Hall, Teachers College, Columbia University, New York City, Sept. 8-10, 2006.

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. *Educational Evaluation and Policy Analysis, 30*(2), 75-110.

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

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

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

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

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. *Educational Evaluation and Policy Analysis, 24* (1), 37-62.

Leming, J. S. (1981). Curricular effectiveness in moral/values education: A review of research. *Journal of Moral Education, 10*(3), 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. *Journal of Allied Health, 35*(2), 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, Hg., *Standards, Evaluation und neue Methoden. Reaktionen auf die PISA-Studie*, S. 1 - 7. Baltmannsweiler: Schneider Verlag Hohengehren.

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.

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

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

Linn, R. (2000). Assessment and accountability. *Educational Researcher, 29*, 2, 4-16.

Linn, R. (2008). *Toward a more effective definition of Adequate Yearly Progress*. Berkeley Law School, http://www.law.berkeley.edu/centers/ewi-old/research/k12equity/Linn.htm (20.7.2008)

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

Lochner, L. & E. Moretti, E. (2004). The effect of education on crime: evidence from prison inmates, arrests, and self-reports. *The American Economic Review*, *94*(1), 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. G. Orfiled & M.L. Kornhaber, Hg., *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? *Education Policy Anlysis Archives, 14*(1). http://epaa.asu.edu/epaa/v14n1/ (20.7.2008)

Nichols, S. L. & Berliner, D. C. (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? *Educational Evaluation and Policy Analysis, 26*(3), 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 (download http://www.pisa.oecd.org/dataoecd/15/13/39725224.pdf (16.7.08)

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*(6), 8-15.

Prehn, K., Wartenburger,I., Mériau, K., Scheibe1, C., Goodenough, O., Villringer, A., van der Meer, E. & Heekeren, H. (2008). Influence of individual differences in moral judgment com-

petence on neural correlates of socio-normative judgments. *Social Cognitive and Affective Neurosience, 3*(1), 33-46.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R., Hg., (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. Lynch School of Education. Boston College.

Rosenthal, R. & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Sachser, N. (2006). Neugier, Spiel und Lernen: Verhaltensbiologische Anmerkungen zur Kindheit. In: U. Herrmann, Hg., *Neudidaktik*, S. 19-30. Weinheim: Beltz.

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*. 2nd edition. Thousand Oaks, USA: Sage Publications.

Schoenfeld, A. H. (1999). Looking toward the 21th century: Challenges of educational theory and practice. *Educational Researcher, 28*, 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. *Educational Researcher, 29*, 4-14.

Smith, Frank (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*. Harper & Brothers.

Smylie, M. A. (1997). From bureaucratic control to building human capital: The importance of teacher learning in education reform. *Educational Researcher, 26*, 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. http://www.ed.gov/programs/readingfirst/index.html (10.1.2008).

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. (http://www.nytimes.com/2005/05/04/education/04education.html (27.7.08)

Wuttke, J. (2007). Die Insignifikanz signifikanter Unterschiede. In: T. Jahne & W. Meyerhöfer, Hg., *Pisa & Co. Kritik eines Programm*. 2., erweitere Auflage, S. 99-246. Hildesheim: Franzbecker.

Zigler, E. & Muenchow, S. (1992). *Head Start. The Inside Story of America's Most Successful Educational Experiment*. New York: Basic Books.