

Thomas Jahnke (Potsdam)

## Empirische Wünschelrutengängerei und ihre Folgen

In dem Beitrag werden einige Charakteristika der empirischen Bildungsforschung und des Unternehmens PISA und seine verhängnisvollen Folgen diskutiert.

Das Wort Wünschelrute leitet sich her von ahd. wunsiligerta, das in den althochdeutschen Glossen als Übersetzung für lat. caduceus, den geflügelten Stab des Hermes, erscheint. Das Kompositum mit der Verkleinerungsform von ahd. wunsch

„Wunsch“ wird mit der Bedeutungsmöglichkeit „Glück, Heil“ erklärt und wunsiligerta von Jacob Grimm als „gerte, durch deren besitz man alles irdischen heils theilhaftig wird“ gedeutet.<sup>[1]</sup> Im Alt- und Mittelhochdeutschen (mhd. wünschelruote, auch wünschelrîs, wünschelwip, wünschelgerte) bezeichnet das Wort allgemein einen Stab oder eine Gerte mit Wunder- oder Zauberkraft, der dem Besitzer besondere Macht verleiht, so auch im Nibelungenlied erklärt (der daz hêt erkunnen, der möhte meister sîn / wol in aller werlde über iêslichen man: „der das erforscht hatte, der konnte Meister sein wohl in der ganzen Welt über jeglichen Menschen“). Daneben wird es mittelhochdeutsch auch als Umschreibung des männlichen Penis (Konrad



Georgius Agricolas "De re metallica libri XII" Basel 1556.

von Megenberg: schwantz oder wuntzelruot) verwendet, sowie in Anknüpfung an die christliche Umdeutung der Wurzel Jesse als ein Bildmotiv des Marienlobs (Gottfried von Straßburg: „der genade ein wünschelruote“) oder allgemeiner des Frauenlobes (im Jüngerem Titulrel über Sigune: du wünschelruot des grales).<sup>[2]</sup> Die heute übliche Verwendung des Wortes speziell als Bezeichnung einer meist gabelförmigen Rute zum Aufspüren von verborgenen Materialien ist seit Frühneuhochdeutscher Zeit zu belegen.<sup>[2]</sup>

[1] Jacob Grimm: Deutsche Mythologie. 2. überarb. Ausg. 1844, S. 926-929 („Wünschelrute“), hier S. 926

[2] Jacob Grimm und Wilhelm Grimm: Deutsches Wörterbuch. S. Hirzel, Leipzig 1854-1960, Band 30, Spalte 2035ff, <http://www.woerterbuchnetz.de/woerterbuecher/dwb/wbgui?lemid=GW26813>, letzter Zugriff: 10. März 2008.

Text und Bild aus der Internet-Enzyklopädie Wikipedia, <http://de.wikipedia.org/wiki/W%C3%BCnschelrute>, letzter Zugriff: 10. März 2008

In dieser etymologischen Erläuterung des Wortes Wünschelrute – dem Jargon der Zeit folgend würde man heute wohl vom *Konzept* Wünschelrute sprechen – geht es ähnlich wie bei den empirischen Forschungen insbesondere in internationalen Vergleichsuntersuchungen von Schülerleistungen um

- das Wünschen,
- die Macht und - erst an dritter Stelle -
- um die Funktion: das Aufspüren von Verborgenen.

Zum Wünschen

Der Wunsch der empirischen Bildungsforschung, wie wir sie hier betrachten, ist die Vermessung ihrer Welt. Das Unheil, das Unbehagen, das Desaster deutscher Schulbildung soll nicht mehr nur gefühlt

sein, es soll vermessen werden. Das Menetekel soll Dezimale bekommen. Das macht es so die Hoffnung wahrer, realer und macht es vor allem berechenbar. Und in dieser Berechenbarkeit liegt dann auch schon der – von der Bildungspolitik nahezu gierig aufgenommene – Kern der Besserung, der Heilung: Werte, die zu tief sind, müssen erhöht werden, solche, die zu hoch sind, müssen gesenkt werden. So einfach ist das. Am Beispiel

Jugendliche aus Familien der oberen sozialen Schichten haben eine 2,7-mal höhere Chance ein Gymnasium zu besuchen als Kinder eines Facharbeiters. Bei PISA 2000 waren die Chancen noch 4,2-mal größer.

liest man unter der Überschrift „Ergebnisse der PISA 2006-Erhebung“ auf den Webseiten des Bundesministeriums für Bildung und Forschung<sup>1</sup>. Wie diese Daten zustande kommen, was sie bedeuten, ob sie auf einer Zunahme der gymnasialen Aufnahmequote für Facharbeiterkinder oder einer Abnahme derselben für Kinder aus den oberen sozialen Schichten beruhen, wodurch sich letztere überhaupt konstituieren und ob diese sich im betrachteten Zeitraum verändert haben, erfährt man nicht, auch nicht welches Chancenverhältnis anzustreben wäre; offenkundig sind aber die Zahlen ‚besser‘ geworden, was man als Erfolg der Bildungspolitik (welcher eigentlich) auslegen kann.

#### Zur Macht

Zur Macht gehört ja auch das Geld. Unbestätigten Berichten zufolge kostet ein PISA-Durchgang international 300 Millionen US-Dollar. Für die drei Durchgängen 2000, 2003 und 2006 ergäben sich danach fast eine Milliarde Dollar. Wenn man einmal unterstellt, dass ein Drittel dieses Geldes – also 300 Millionen Dollar – auf die in diese Unternehmung involvierten Wissenschaftler entfällt, und beachtet, dass der Tagessatz eines internationalen Experten 300 Dollar beträgt, so ergeben sich eine Million wissenschaftliche Mann-Tage, also für 1000 Wissenschaftler je 1000 Arbeitstage. Da wird es verständlich, dass der OECD-Verantwortliche, Herr Schleicher, jegliche Kritik an PISA gern mengenmäßig zurückweist: Tausende von Wissenschaftler stünden hinter PISA, und dass der Leiter des KMK-Auftragsforschungs-Instituts zur Qualitätsentwicklung im Bildungswesen (IQB), Olaf Köller, sich Kritik an PISA nur so erklären kann, dass die Kritiker „nicht genug vom Kuchen“ abbekommen hätten.

#### Marktmacht

Die Marktmacht der *educational assessment*-Firmen, die PISA entwickelt und an bisher achtundfünfzig Länder verkauft haben, wird in den PISA-Veröffentlichungen gänzlich verschwiegen oder herunter gespielt, als handele es sich allenfalls um Schreib- und Rechenbüros, die den involvierten Wissenschaftlern die technischen Arbeiten abnehmen. Insgesamt entsteht hier eine schwer durchschaubare globalisierte Schulaufsicht, deren Handeln von den einzelnen Staaten nur noch bedingt beeinflusst werden kann. Die Erziehungswissenschaftlerin Elisabeth Flitner schreibt zur deutschen Perspektive:

In Deutschland bestand bis zur Jahrtausendwende kein nennenswerter Testmarkt im öffentlichen Schulsystem; Deutschland beteiligte sich unregelmässig an internationalen educational assessments und deren Ergebnisse wurden nur einem kleinen Kreis von Fachleuten bekannt. Dem deutschen Teil von PISA wurde von der OECD besondere Aufmerksamkeit gewidmet. Als OECD-Projektverantwortlicher für PISA ist ein Deutscher tätig. Während sich die OECD grundsätzlich mit bildungspolitischen Empfehlungen an einzelne Staaten zurückhält, hat dieser sich entschieden, die Debatte in Deutschland immer wieder zu beleben, indem er in öffentlichen Stellungnahmen den Kampf mit der Kultusministerkonferenz um die Deutungshoheit über die PISA-Ergebnisse aufnimmt. Er unterstützt damit die Medienarbeit, die in Deutschland für PISA gemacht wurde. Zugleich aber stört

<sup>1</sup> <http://www.bmbf.de/de/6624.php>

diese Form der Intervention vonseiten der OECD das Arbeitsbündnis zwischen deutschen Kultusministerien und PISA-Direktoren. Diese können kein Interesse daran haben, dass die deutschen Kultusminister ihre bildungspolitische Souveränität durch PISA eingeschränkt sehen. Zwar ist die nationale und regionale kulturpolitische Souveränität überall im Schwinden begriffen. Aber ihr Abbau im Einzelnen ist kein konfliktfreier Vorgang. Die Kultusministerien als Auftraggeber der PISA-Studien willigen mit jeder Fortsetzung und Ausweitung der Test-Projekte in den Aufbau einer zweiten Aufsichtsinstanz über das Schulsystem ein, – einer international orientierten wissenschaftlichen Instanz, die ausserhalb der ministeriellen Weisungshierarchie steht. Sie willigen ein, sich künftig von Aussenstehenden sagen zu lassen, was ihre Verwaltungen und ihre Schulen taugen.

Elisabeth Flitner: Pädagogische Wertschöpfung. Zur Rationalisierung von Schulsystemen durch *public-private-partnerships* am Beispiel von PISA. In: Jürgen Oelkers, Rita Casale, Rebekka Horlacher & Sabina Klee (Hrsg.): Rationalisierung und Bildung bei Max Weber. Beiträge zur historischen Bildungsforschung. Bad Heilbrunn: Klinkhardt, 245-266. Siehe auch: [http://forum-kritische-paedagogik.de/start/e107\\_files/downloads/ef.fend\\_beitrag.pdf](http://forum-kritische-paedagogik.de/start/e107_files/downloads/ef.fend_beitrag.pdf)

Fraglich scheint mir dabei nur, ob die hier angesprochene zweite Aufsichtsinstanz über das Schulsystem tatsächlich wissenschaftlich und nicht eher kommerziell orientiert ist, das heißt, ob ihre Setzungen und Entscheidungen nicht in erster Linie von Verkaufserwägungen und Marktstrategien gekennzeichnet sind.

### Deutungsmacht

Dass die PISA-Daten überhaupt einer Deutung bedürfen und nicht an und für sich und gleichsam von allein das Bild des Unheils einer deutschen Bildungsmisere malen, ist in der deutschen Öffentlichkeit erst anhand der Auseinandersetzung zwischen dem OECD-Koordinator und den deutschen PISA-Verantwortlichen um die Interpretation der Ergebnisse des PISA-Durchgangs 2006 – also im Dezember 2007 – deutlich geworden. Die Stellungnahmen der Kombattanten waren aber so offensichtlich Interessen geleitet, dass sich das Bild von den empirisch gesicherten deskriptiven Befunde, die diese internationalen Vergleichsuntersuchungen zu konstatieren beanspruchen, auch öffentlich kaum noch halten ließ. Der Erziehungswissenschaftler Hans Brügelmann hat daraufhin vorgeschlagen, dass die PISA-Daten interessierten Wissenschaftlern zur Verfügung gestellt werden, die dann ihre – möglicherweise auch konträren – Deutungen in einem gemeinsamen Sammelband veröffentlichen, der dann als Grundlage zu bildungspolitischem Handeln dient, um so einseitigen Deutungen und kurzschlüssigen Reaktionen zu vermeiden.<sup>2</sup> Doch dieser Vorschlag verkennt, dass es sich bei PISA nicht um eine wissenschaftliche Untersuchung sondern um eine politische Auftragsstudie handelt. Die Auftragnehmer werden also schon der Folgeaufträge halber alles daran setzen, ihre Daten und Ergebnisse und Folgerungen als möglichst unproblematisch und zielgerichtet klar erscheinen zu lassen. Von ihrer Seite aus gibt es gar keinen Bedarf, die PISA-Resultate oder Methoden einer wissenschaftlichen Diskussion auszusetzen. Dies gilt sogar für den statistischen Apparat von PISA, den der im Umgang mit großen schlecht konditionierten Datenmengen vertraute Physiker Joachim Wuttke sehr sorgfältig und detailliert untersucht hat mit dem Ergebnis, dass es sich bei PISA um einen sehr teuren Zufallszahlengenerator handelt. Die deutschen PISA-Verantwortlichen

---

<sup>2</sup> Brügelmann, H. (2008a): Fieber genau zu messen ist noch keine Diagnose, Fieber erfolgreich zu senken keine Therapie. Wie Leistungstests in ihren Leistungsmöglichkeiten durch PISA & Co überfordert werden. Beitrag zum Forum „Schule ist mehr als PISA - Zur Bedeutung reformpädagogischer Ansprüche an die schulische Bildung von heute“ der ZEIT-Stiftung in Hamburg am 6./7. März 2008. [http://www.agprim.uni-siegen.de/printbrue/brue.08a.pisa\\_refpaed.pdf](http://www.agprim.uni-siegen.de/printbrue/brue.08a.pisa_refpaed.pdf)

sind auf Wuttkes Analysen<sup>3</sup> allenfalls marginal, verärgert und unterstützt durch höhnische Kommentare ihrer medialen Helfer eingegangen.

Zur Funktion: das Aufspüren von Verborgenen

Das eigenartige an der empirischen Forschung ist, dass sie immer etwas hervorbringt; ob kleinste Examensarbeit oder International Survey, sie kann nicht leer ausgehen. Immer erhält sie Zahlen und damit Skalen, Intervalle, Ranglisten und, was dann der statistische Apparat folglich zur Verfügung stellt, Mittelwerte, Standardabweichungen, Korrelationen und so fort. Man wünscht sich mit der empirischen Rute gleichsam das Ergebnis herbei. Durch das Messen erhält man Daten – eigentlich sogar nur Zahlen, deren Realitätsgehalt (oder genauer Realitätsbeschreibungsvermögen) außer Frage steht, ja dem ganzen Prozess eine Art quasi-naturwissenschaftlicher Dignität verleiht. Die Frage, ob man da tatsächlich etwas misst, was man später ausgraben kann, stellt sich gar nicht, weil der Gegenstand durch den Messprozess hervorgebracht und konstituiert wird. Solche Forschung produziert ihre Begriffe (heute sagt man dann auch *Konzepte*) und Ergebnisse parthenogenetisch, notfalls auch ohne jede Anleihen bei der bemessenen Realität oder Bezüge zu ihr. Und wenn man die Rotzigkeit (Pardon!) des Positivismus zugrunde legt, gilt das dann alles bis zu seiner Widerlegung, die – dessen sind sich die Pisaner und die involvierten Firmen sicher – faktisch wenn nicht sogar prinzipiell ausgeschlossen ist. Die Sache ist versiegelt. Wo Kritik nicht abprallt, freut sich die Forschungsindustrie über Folgeaufträge, die dann in der beschriebenen Art bearbeitet werden.

Wenden wir unsere Aufmerksamkeit einem fassbareren – wenn gleich auch emotional befrachtetem – Gebiet zu: dem Fußball. Der Dortmunder Physiker Metin Tolan<sup>4</sup> hat die etwa 12.000 Bundesligabegegnungen seit deren Anbeginn statistisch untersucht. Zu Vergleichszwecken erstellt er zufällige Bundesligatabellen mit durchweg gleichstarken Vereinen und konstatiert, dass diese den echten Tabellen verblüffend ähnlich sind. Erst in der Betrachtung der Differenz der echten von den zufälligen Tabellen kommt er zu sinnvollen Aussagen über die echte Bundesliga. Ich meine, man sollte den PISA-Auswertern in gleicher Weise ‚Messwerte‘ für einige virtuelle Länder unterschieben, deren Schülerleistungen durchweg, auch was die Verteilung auf Schulen, Schüler, Aufgaben und Aufgabenhefte anlangt, zufällig erzeugt sind. Ich bin mir sicher, dass die Pisaner ohne zu zögern auch ‚empirische‘ belegte Befunde über die Bildungssysteme dieser Länder erschließen würden, sie ihre Rangplätze ermitteln und deuten und *Steuerungswissen*<sup>5</sup> für sie bereit stellen könnten. Auf den ersten Blick mag der Vorschlag für solche Blindversuche möglicherweise hämisch erscheinen, auf den zweiten ist er nur konsequent, und die empirische Bildungsforschung wird hoffentlich auch hier nicht zögern, von der naturwissenschaftlichen Methodik zu lernen.

---

<sup>3</sup> Joachim Wuttke: Die Insignifikanz signifikanter Unterschiede: Der Genauigkeitsanspruch von PISA ist illusorisch. In: Jahnke, Th.; Meyerhöfer, W.: Pisa & Co. Kritik eines Programms. 2., erweiterte Auflage. Franzbecker Verlag. Hildesheim 2007, S. 99-246.

<sup>4</sup> Siehe <http://www.derwesten.de/blogs/querkraft/stories/602/> und <http://www.derwesten.de/blogs/querkraft/stories/735/> .

<sup>5</sup> „Die Auftraggeber, das sind die OECD-Staaten und weitere Staaten (sogenannte Partnerstaaten) , erwarten von dem internationalen Vergleich empirisch fundiertes Steuerungswissen. Die teilnehmenden Staaten möchten von Pisa zum Beispiel erfahren, wo sie im internationalen Vergleich stehen, welche Bildungsergebnisse andere Staaten erreichen und unter welchen Bedingungen das geschieht. (...) Besonderes Interesse findet dabei auch die Frage, inwieweit es in den Staaten gelingt, jungen Menschen unabhängig von ihrer sozialen Herkunft vergleichbare Chancen für die Entwicklung ihrer Kompetenzen zu geben.“, führt Manfred Prenzel in PISA-Konsortium Deutschland (Hrsg.): PISA ´06. Die Ergebnisse der dritten internationalen Vergleichsstudie. Waxmann Verlag. Münster 2007 auf Seite 13 aus.

Zur nicht-virtuellen Funktion von PISA wirft der Erziehungswissenschaftler Volker Ladenthin in einer jüngst erschienenen *bildungstheoretischen Betrachtung* einige grundsätzliche Fragen auf:

Versucht eine bestimmte bildungstheoretische Konzeption mit dem Instrument der PISA-Studie Kriterien für die Zielorientierung von Schule durchzusetzen, die in der deutschen Schulpädagogik, Schulverwaltung oder Schulpolitik (vielleicht ja aus guten Gründen) bisher keine Akzeptanz fanden? Geht es PISA nicht um Empirie, sondern um Normendurchsetzung? Aber welche Normen sind es? Stimmen sie? Wer hat sie als gültig legitimiert? Steht hinter PISA gar nicht eine deskriptive Leistungsmessung anhand von in der deutschen schulpädagogischen Diskussion allseitig anerkannten Kriterien; steht "hinter" PISA vielleicht eher ein über das Vehikel Evaluation implementiertes Programm zur Veränderung, speziell zur Globalisierung der Schulsysteme?

und beantwortet sie dann anhand von Selbstaussagen des PISA-Konsortiums:

Im Verweis auf die zu Grunde gelegte Bildungstheorie wird diese Annahme vom PISA-Konsortium ausdrücklich bestätigt: "Der Gedanke der notwendigen Universalisierung der Basisqualifikationen wird in der angelsächsischen Literacy-Diskussion mit dem Argument neuer und infolge des sich beschleunigenden Wandels von der Industrie- zur Wissensgesellschaft steigender Qualifikationsanforderungen verknüpft. (...) Dieses Konzept von Literalität steht auch im Hintergrund der internationalen Rahmenkonzeption von PISA." Es geht erklärtermaßen der OECD (als Wirtschaftsorganisation) um die Durchsetzung eines zwar als "angelsächsisch" bezeichneten, in Wirklichkeit aber längst auch bei uns bekannten, schlicht funktionalen Bildungsbegriffs; es geht um Universalisierung (also ökonomische Globalisierung), d.h. die Auflösung föderaler und letztlich auch nationaler Bildungssysteme; es geht um die Transformation von Selbstbestimmung und Kulturidentität in funktionale Basiskompetenzen; es geht um Normierung und Vereinheitlichung statt um Individualisierung und Differenzierung. Als Grund und Legitimation für diesen Wandel werden an der zitierten Stelle ausdrücklich und ausschließlich die von der ökonomischen Situation her verursachten und definierten "Qualitätsanforderungen" benannt. Die Schule dient nicht mehr der Personalisation der Person, der Teilhabe an der menschlichen Gesamtpraxis, der Bildung, sondern einer Zurichtung auf das, was das PISA-Konsortium als Kultur definiert und wodurch es diese definiert sieht.

Wenn man diese Selbstaussagen des Konsortiums ernst nimmt, muß die PISA-Studie anders als bisher rezipiert werden. Nicht als Evaluation von Basisqualifikationen, die Bildung oder auch nur Allgemeinbildung anzeigen; nicht als Aussage über das deutsche Bildungssystem im Vergleich zu anderen Bildungssystemen. PISA muß vielmehr als Versuch rezipiert und bewertet werden, Bildungssysteme an funktionalen Kriterien zu messen, um diese - von der bisherigen (besonders von der deutschen) Bildungsdiskussion bisher nicht sehr breit rezipierten oder gar akzeptierten - Kriterien nunmehr unter Umgehung einer breiten Diskussion durchzusetzen. (In der viel zu schnellen Reaktion der Politik ist diese Durchsetzung undiskutierter Kriterien z.T. gelungen.) Es geht der Studie nicht um Messung, sondern um Normierung; nicht um Diagnose, sondern um Steuerung.

Volker Ladenthin: PISA - Recht und Grenzen einer globalen empirischen Studie. Eine bildungstheoretische Betrachtung. Vierteljahrsschrift für wissenschaftliche Pädagogik 84 (2008)

Zu den Folgen

Eine erste Folge einer testorientierten Schulbildung und einer testresultatgesteuerten Bildungspolitik ist, dass das, was in der Philosophie und der Soziologie als Entfremdung bezeichnet, in einer neuen Qualität den Lehrerberuf und das Lehren und Lernen an der Schule gleichsam befällt: es geht nicht mehr um die Sache selbst, um die Schüler, um die Lehrer, um ein gemeinsames Bildungserlebnis, es geht letztlich um Testresultate, anhand derer der Erfolg des Lehrens und Lernens gemessen und beurteilt wird. Verächtlicher und desinteressierter kann man mit einem Schüler kaum umgehen, als dass man sich z.B. in Multiple-Choice-Tests nur noch dafür interessiert, ob er an der richtigen Stelle ein Kreuz macht, und ihm dafür null oder einen Punkt gibt. Auch das hässliche Kompositum *Bildungsstandards* (die überdies auch noch *normiert* werden) bekräftigt aus sich selbst heraus, dass es hier nicht individuelle Lernprozesse oder kulturelle Teilhabe intendiert sind.

Beispiele für die Folgen solcher Entfremdung kann man (nicht nur den amerikanischen) Zeitungen entnehmen.



### **New York: Schüler bekommen für gute Leistungen bis zu 50 Dollar**

Wie die "New York Times" berichtet, will die Stadt New York Schüler zu besseren Leistungen anspornen. Für gute Noten können die Schüler bis zu 50 Dollar erhalten. Insgesamt wurde bereits eine halbe Million US-Dollar ausbezahlt. 200 Schulen beteiligen sich an dem Programm, bei dem auch den Lehrern eine Sonderzahlung am Ende des Jahres winkt, falls Schüler ihre Noten verbessert haben. (...) Ende Februar wurde zudem damit begonnen, Schüler für gute Leistungen und vorbildliches Verhalten mit einem kostenlosen Handy zu belohnen.

Quelle: <http://www.shortnews.de/start.cfm?id=700812>

### **Cheating Scandals Rock Three Top-Tier High Schools - National Survey Finds Two-Thirds of Students Admit Cheating**

By Susan Donaldson James

Sam is a top student in a high-pressure high school just outside New York City who openly admits he "cheats along the way" to academic success. The 16-year-old sees nothing wrong with looking at another student's paper during a quiz or borrowing a classmate's ideas. He insists "90 percent or higher" of the students at his school engage in cheating - from tucking vocabulary crib sheets under their hats to stealing math exams. But Sam insists he has a moral conscience - he won't use his last name for this article - and he swears he will never cheat in college. But he justifies his cheating. "My parents would consider this cheating, but I don't have any major problems with it," Sam told ABCNEWS.com. "It's school, and you're cheating your way through the system."

Sam is typical of most American students. An estimated two-thirds of all high school students admit to "serious" academic cheating, according to a national survey by Rutgers' Management Education Center in New Jersey. A startling 90 percent say they cheat on homework.

Cheating is epidemic, say experts, and recent scandals have rocked - and in some cases divided - both public and private high schools from New Hampshire to California. Some are the highest-performing schools in the nation, where the pressure to get into an Ivy League college is intense and parents buy into the academic game. (...)

Quelle: ABC News, Friday, February 29, 2008. <http://abcnews.go.com/US/story?id=4362510&page=1>

Diese Beispiele erscheinen noch recht oberflächlich, eher als solche für problematische gesellschaftliche Verkehrsformen, fragliche Schulverhältnisse und für eine Verrohung pädagogischer Sitten oder Normen. Es geht viel grundsätzlicher darum, ob die in schulischen Bildungsprozessen angelegte, unmittelbare und authentische Begegnung der Beteiligten miteinander und mit der Sache nun durch eine äußere Zielvorgabe und ein äußeres Maß in ein mittelbares Verhältnis transformiert wird, dessen Zwecke deutlich und sichtbar außerhalb des eigentlichen Bildungsprozesses liegen.

Eine zweite Folge der Orientierung des Schulunterrichts auf Testbildung ist die Abwertung der Fächer, die nicht abgetestet werden. Ihr Wert sinkt herab zu dem einer ‚sozialen Veranstaltung‘, wie man zuweilen in leistungsorientierten Kreisen formuliert. Ästhetische Bildung und moralische Urteilsfähigkeit – um die Termini des bereits oben zitierten Erziehungswissenschaftlers Ladenthin aufzunehmen – sind in dem, was in der Testideologie als Grundbildung bezeichnet wird, offensichtlich verzichtbar. Dieser Verzicht wird sich gerade auch auf den Unterricht in den Grundschulen auswirken, die zunehmend nur als Orte der Präparation für die sogenannten weiterführenden Schulen gesehen werden.

Eine dritte Folge der Testorientierung ist die Zunahme des Drucks auf die Lehrerinnen und Lehrer. Unter der Hand wird gerade von sich fortschrittlich dünkenden Fachdidaktikern – zum Teil wie mir scheint sogar frohlockend – festgestellt, dass man nun endlich ein taugliches Mittel gefunden habe, die Unterrichtenden an der Kandare zunehmen und in die Richtung als modern bezeichneter Didaktik zu zwingen. Dieser Zwang zum Fortschritt lässt die Wünschelrute dann tatsächlich zur Rute werden.

Mit einer vierten Folge von Pisa&Co will ich mich abschließend etwas ausführlicher auseinandersetzen, nämlich dem Teaching to the Test.

Die Kritik an den Testverfahren und an dem Gedanken, die ‚Erträge‘ schulischer Bildung könnten über periodisierte Tests in sinnvoller Weise gemessen und gesteigert werden, hat Deutschland nicht erreicht, und es ist bisher auch kaum zu einer sorgfältigen und redlichen Diskussion dieses *prima vista* selbstverständlich erscheinenden Gedankens hierzulande gekommen ist.<sup>6</sup> Das ist auch nicht weiter erstaunlich, weil von den involvierten Testinstitute und den mit ihnen kooperierenden Wissenschaftler kaum zu erwarten ist, dass sie mit ihrem Test-Knowhow auch die Test-Kritik auf den Markt bringen.

In dem vierzehnteiligen Beschluss der Kultusministerkonferenz zum Bildungsmonitoring vom 02.02.2006 wird auf Seite 13 unter der Zwischenüberschrift „Weiterentwicklung der Bildung, aber kein Teaching to the Test“ auf diese Problematik – wie folgt – kurz eingegangen:

Neben der Funktion der Beschreibung von Leistungsanforderungen und der Leistungsmessung dienen die Bildungsstandards primär der Weiterentwicklung des Unterrichts und vor allem der individuellen Förderung aller Schülerinnen und Schüler. Die Länder sind sich darin einig, dass mit der Setzung der Bildungsstandards als übergreifenden Referenzrahmen eine Entwicklung hin zum „teaching to the test“ oder eine Verengung des Unterrichts auf die Anforderungen der Standards verhindert werden muss. (S. 13)

Diese Kürze ist trotz der beschworenen Einigkeit der Länder erstaunlich. Es liegt nahe, wenn man die ‚Weiterentwicklung des Unterrichts und vor allem der individuellen Förderungen der Schülerinnen und Schüler‘ durch Bildungsstandards befördern oder anordnen will, deren Erreichen im Wesentlichen durch Tests überprüft wird, die Erfahrungen von Ländern und darunter insbesondere der USA zu rezipieren, die seit Jahren oder Jahrzehnten eine solche Politik verfolgen.

For several decades, some measurement experts have warned that high-stakes testing could lead to inappropriate forms of test preparation and score inflation, which we define as a gain in scores that substantially overstates the improvement in learning it implies. (p. 99)

leitet Daniel Koretz, Erziehungswissenschaftler an der Harvard-Universität und assoziierter Direktor des Center of Research, Standards, and Student Testing (CRESST) seinen Aufsatz *Alignment, High Stakes, and the Inflation of Test Scores*<sup>7</sup> ein und beschreibt einen Ausgangspunkt, über den eine öffentliche Diskussion in Deutschland bisher kaum hinausgekommen ist:

On common response to this problem has been to seek „tests worth teaching to“. The search for such tests has led reformers in several directions over the years, but currently, many argue that tests well aligned with standards meet this criterion. If tests are aligned with standards, the arguments runs, they test material deemed important, and teaching to the test therefore teaches what is important. If students are being taught what is important, how can the resulting score gains be misleading? (p. 99)

---

<sup>6</sup> Es ist aufschlussreich und vermutlich nicht folgenlos, dass die Daten von Pisa in Australien aufgearbeitet werden und gleichsam deutschen (oder europäischen) Boden nie betreten. In Zeiten einer sich global verstehenden Forschung scheinen solche räumliche Distanzen ohne jede Auswirkung, u.a. weil der Zugriff auf Datenserver ubiquitär und ohne zeitliche Verzögerung möglich ist. Dennoch ist es von Bedeutung, ob und in welchem Rahmen und geistigen Raum die Verfahren zur Aufbereitung der Daten entwickelt, diskutiert und kritisiert werden, ob sie als die Ergebnisse der Form und dem Inhalt nach prägende Instrumente begriffen werden oder nur - als mehr oder minder schlecht dokumentierte - Routinen in Softwarepaketen erscheinen, ob sie überhaupt wissenschaftlich diskutiert oder schlicht als notwendige und doch arbiträre Essenzen eines ‚State of the Art‘ aufgefasst werden, ob den beteiligten Forschern daran liegt, ihre Verfahren zu verkaufen oder als Erkenntnisinstrumente in die Diskussion einzuführen und zu legitimieren etc.

<sup>7</sup> Koretz, D.: *Alignment, High Stakes, and the Inflation of Test Scores*. Yearbook of the National Society for the Study of Education (2005) 104 (2), 99–118.  
(Online unter: <http://www.blackwell-synergy.com/doi/abs/10.1111/j.1744-7984.2005.00027.x>)

Koretz begründet seinen Widerspruch gegen solche Naivität theoretisch und empirisch unter anderem eindrücklich mit Sägezahnkurven („sawtooth pattern“) für die gemessenen Leistungen der gleichen oder einer vergleichbaren Population, die sich in verschiedenen Erhebungen je nach den verwendeten Tests in unterschiedlichster Weise ergaben. Auch der Hoffnung, solche Effekte seien allein der Testkonstruktion und den Testumständen zuzuschreiben, widerspricht er:

The problem is not confined to commercial, off-the-shelf, multiple-choice tests. It has appeared as well with standards-based tests and with tests using no multiple-choice items. (p. 106)

Die Vorstellung, Schülerleistungen ließen sich in einem Test objektiv oder mit angebbaren Fehlermargen - gleichsam physikalisch - messen, ist schlicht (und) irreführend. Folgerungen aus solcher Vorstellung mehr als fragwürdig. Wird dies in Abrede gestellt, verschwiegen oder das Gegenteil prätendiert, liegen in aller Regel massive Erkenntnisinteressen der Auftraggeber oder -nehmer der Testungen vor.

Auch die Auswirkungen von Testungen auf den Unterricht werden in den USA seit Jahrzehnten untersucht. Koretz zum Beispiel beschreibt und charakterisiert in dem zitierten Papier *Reallocation, Alignment* und *Coaching*:

*Reallocation.* Reallocation refers to shifts in instructional resources among the elements of performance. Research has shown that when scores on a test are important to teachers, many of them will reallocate their instructional time to focus more on the material emphasized by the test. (...) Many observers believe that reallocation is among the most important factors causing the sawtooth pattern (...).

*Alignment.* Content and performance standards comprise material – performance elements, in the terminology used here – that someone (not necessarily the ultimate user of scores) has decided are important. If the material is emphasized in the standards, that implies that users should give this material substantial weight in the inferences they draw about student performance. Alignment gives this same material high weights in the test as well. (...)

*Coaching.* The term “coaching” is used in a variety of different ways in writings about test preparation. Here it is used to refer to two specific, related types of test preparation, called substantive and non-substantive coaching.

*Substantive coaching* is an emphasis on narrow, substantive aspects of a test that capitalizes on the particular style or emphasis of test items. The aspects of the tests that are emphasized may be either intended or unintended by the test designers. For example, in one study of the author’s, a teacher noted that the state’s test always used regular polygons in test items and suggested that teachers should focus solely on those and ignore irregular polygons. The intended inferences, however, were about polygons, not specifically regular polygons. (...)

*Nonsubstantive coaching* refers to the same process when focused on nonsubstantive aspects of a test, such as characteristics of distracters (incorrect answers to multiple-choice items), substantively unimportant aspects of scoring rubrics, and so on. Teaching test-taking tricks (process of elimination, plug-in, etc.) can also be seen as nonsubstantive coaching. In some cases – for example, when first introducing young children to the op-scan answer sheets used with multiple-choice tests – a modest amount of certain types of nonsubstantive coaching can increase scores and improve validity by removing irrelevant barriers to performance. In most cases, however, it either wastes time or inflates scores. (p. 110 – 112)

An anderer Stelle findet sich ähnliche Kritik. So fasst Brian M. Stecher sein Kapitel 4 *Consequences of large-scale, high-stakes testing on school and classroom practices* in dem von ihm mitherausgegebenen Buch *Making Sense of Test-Based Accountability in Education*<sup>8</sup> folgendermaßen zusammen:

The net effect of high-stakes testing on policy and practice is uncertain. Researchers have not documented the desirable consequences of testing – providing more instruction, working harder, and working more effectively – as

<sup>8</sup> Stecher, B. M.: Consequences of large-scale, high-stakes testing on school and classroom practices. In L. S. Hamilton, B. M. Stecher, and S. P. Klein (Eds.): *Making Sense of Test-Based Accountability in Education*. RAND. Santa Monica 2002. P 79 -100  
(Online unter: [http://www.rand.org/pubs/monograph\\_reports/MR1554/index.html](http://www.rand.org/pubs/monograph_reports/MR1554/index.html))



clearly as the undesirable ones – such as negative reallocation, negative alignment of classroom time to emphasize topics covered by a test, excessive coaching, and cheating. More important, researchers have not generally measured the extent or magnitude of the shifts in practice that the identified as a result of high-stakes testing.

Overall, the evidence suggests that large-scale high-stakes testing has been a relatively potent policy in terms of bringing about changes within schools and classrooms. Many of these changes appear to diminish students' exposure to curriculum, which undermines the meaning of the test scores. (p. 99/100)

Der im letzten Absatz angesprochene Antagonismus scheint der deutschen Kultusministerkonferenz möglicherweise von ihren Beratern vorenthalten worden zu sein. Das Gleiche gilt vermutlich für das *Position Statement on High Stakes Testing in PreK-12 Education* der American Evaluation Association (AEA), in dem es heißt:

High stakes testing leads to under-serving or mis-serving all students, especially the most needy and vulnerable, thereby violating the principle of "do no harm." The American Evaluation Association opposes the use of tests as the sole or primary criterion for making decisions with serious negative consequences for students, educators, and schools. The AEA supports systems of assessment and accountability that help education.

Recent years have seen an increased reliance on high stakes testing (the use of tests to make critical decisions about students, teachers, and schools) without full validation throughout the United States. The rationale for increased uses of testing is often based on a need for solid information to help policy makers shape policies and practices to insure the academic success of all students. Our reading of the accumulated evidence over the past two decades indicates that high stakes testing does not lead to better educational policies and practices. There is evidence that such testing often leads to educationally unjust consequences and unsound practices, even though it occasionally upgrades teaching and learning conditions in some classrooms and schools. The consequences that concern us most are increased drop out rates, teacher and administrator deprofessionalization, loss of curricular integrity, increased cultural insensitivity, and disproportionate allocation of educational resources into testing programs and not into hiring qualified teachers and providing sound educational programs. The deleterious effects of high stakes testing need further study, but the evidence of injury is compelling enough that AEA does not support continuation of the practice.

While the shortcomings of contemporary schooling are serious, the simplistic application of single tests or test batteries to make high stakes decisions about individuals and groups impede rather than improve student learning. Comparisons of schools and students based on test scores promote teaching to the test, especially in ways that do not constitute an improvement in teaching and learning. Although used for more than two decades, state mandated high stakes testing has not improved the quality of schools; nor diminished disparities in academic achievement along gender, race or class lines; nor moved the country forward in moral, social, or economic terms. The American Evaluation Association (AEA) is a staunch supporter of accountability, but not test driven accountability. AEA joins many other professional associations in opposing the inappropriate use of tests to make high stakes decisions.

In einer Endnote zu diesem Text wird auf weitere Organisationen verwiesen, die ebenfalls dagegen opponieren, weit reichende Entscheidungen auf Grund von Testergebnissen zu fällen.

AEA joins many other professional associations, teacher unions, parent advocacy groups in opposing the inappropriate use of tests to make high stakes decisions. These include, but are not limited to the American Educational Research Association, the National Council for Teachers of English, the National Council for Teachers of Mathematics, the International Reading Association, the College and University Faculty Assembly of the National Council for the Social Studies, and the National Education Association<sup>9</sup>

Für den deutschen Betrachter ist es kaum nachvollziehbar, mit welchen Besserungs- wenn nicht gar Heilserwartungen gleich in welcher Richtung die hiesige Bildungspolitik umfangreichste Testprogramme einführt, während der sicherlich nicht zimperliche angelsächsische Evaluations-Pragmatismus sich in kaum zu übertreffender Deutlichkeit nach mehr als zwanzigjähriger Erfahrung von solchen Bestrebungen distanziert.

---

<sup>9</sup> American Evaluation Association (AEA): Position Statement on HIGH STAKES TESTING In PreK-12 Education. 2002 (Online unter: <http://www.eval.org/hst3.htm>)

Bei dem vorstehenden Text handelt es sich um das leicht überarbeitete Manuskript zu dem Vortrag *Die empirische Wünschelrutengängerei und ihre Folgen* auf der 42. Jahrestagung der Gesellschaft für Didaktik der Mathematik (GDM) am 16.03.2008 in Budapest.

Literaturhinweise:

Stefan T Hopmann / Gertrude Brinek / Martin Retzl (Hrsg.): PISA zufolge PISA / PISA According to PISA. Wien / Berlin (LIT-Verlag) 2007. ISBN 978-3-7000-0771-5 (Österreich). ISBN 978-3-8258-0946-1 (Deutschland). 432 Seiten. € 34,90

Thomas Jahnke / Wolfram Meyerhöfer (Hrsg.): Pisa & Co – Kritik eines Programms. Zweite, erweiterte Auflage. Franzbecker Verlag. Hildesheim 2007. ISBN 978-3-88120-464-4. 479 Seiten. € 16,80