Lind, G. (2011). Verbesserung des Unterrichts durch Selbstevaluation. Ein Plädoyer für unverzerrte Evidenz. In: Bellmann, J., Hrsg.: Wissen, was wirkt. Kritik evidenzbasierter Pädagogik. Wiesbaden: VS-Verlag für Sozialwissenschaften.

Verbesserung des Unterrichts durch Selbstevaluation

Ein Plädoyer für unverzerrte Evidenz

Georg Lind

Pädagogische Evidenz soll valide Hinweise für effektivere Unterrichtsmethoden und Schulstrukturen und damit für besseres Lernverhalten der Schülerinnen und Schüler geben. In der Pädagogik gibt es erste Versuche, wobei aber kontrovers diskutiert wird, welche Rolle empirische Evidenz bei der Verbesserung von Unterricht und Schulstruktur spielen soll und spielen kann (Brügelmann 2005; Bohl/Kiper 2009; Böttcher/Dicke/Hogreber 2011). Prinzipiell kann sehr Verschiedenes als *Evidenz* angesehen werden: Schulbeteiligung, Wiederholerraten, Übertritt in weiterführende Schulen, Schulzufriedenheit und vieles andere mehr. Wegen medialer Großereignisse wie TIMSS, PISA, VERA & Co. fokussiert die Debatte über pädagogische Evidenz heute aber fast ausschließlich auf vergleichende Tests. Die Kritik an Evidenz als Grundlage pädagogischen Handelns (*data-driven education*) bezieht ihre Argumente zum Großteil, wenn nicht ausschließlich aus der weitgehend berechtigten Kritik an Vergleichstests (Kohn 2000; Jahnke/Meyerhöfer 2007; Hopmann/Brinek/Retzl 2007), mit Ausnahme von vergleichenden Noten, die nach wie vor den Schulalltag dominieren und aus ähnlichen Gründen kritisch diskutiert werden (Kohn 1999; Czerny 2010; Leppert 2010). Dabei werden andere Formen der Evaluation weitgehend ignoriert.

Ich will in diesen Beitrag zweierlei zeigen: Erstens, wie durch die gegenwärtige Praxis der Evidenzbeschaffung in der Pädagogik durch Vergleichstests und die Evaluation von Personen (Schüler, Lehrer, Länder etc.) tatsächlich die Validität und Aussagekraft von Evidenz leidet und gleichzeitig die Kosten nach oben getrieben werden. Zweitens, dass es dazu wenig beachtete Alternativen gibt, die als Grundlage für pädagogische Reformen dienen können. Eine Alternative ist *Programmevaluation*, das heißt die Gewinnung von Evidenz über die Wirksamkeit von bil-

dungspolitischen Programmen, Unterrichtsmethoden etc. (Campbell 1969; Sanders 1994). Die Einrichtung des „What Works Clearinghouse" der US-Regierung im Jahr 2002 hat dieser Art von Evidenz wieder Auftrieb gegeben. Dazu zählen auch die – inzwischen außer Mode gekommenen – lernzielorientierten Tests (Glaser 1963), die unter dem Etikett „Mindeststandards" heute wiederbelebt werden. In beiden Fällen stehen nicht Personen, sondern Programme, Maßnahmen und Methoden im Mittelpunkt. Ich will anhand einiger Beispiele zeigen, dass diese Art der Evaluation mit ähnlichen Problemen zu kämpfen hat wie die Personenevaluation (wenn auch in subtilerer Form), sobald ihre Ergebnisse in Konflikt mit politischen oder kommerziellen Interessen geraten (Schoenfeld 2006; Bracey 2007; Lind 2009a).

Dadurch sind die Möglichkeiten zur Gewinnung pädagogischer Evidenz durch objektive Tests keineswegs erschöpft. Eine weitere, aber meines Erachtens viel zu wenig genutzte Möglichkeit ist *objektive Selbstevaluation*, die bei fachgerechter Durchführung weniger korruptionsanfällig, billiger und unmittelbarer wirksam ist als die beiden anderen Arten der Evaluation. Wie ich unten zeigen werde, liegt der Hauptgrund für die hohen Kosten und die niedrige Validität von Vergleichstests und Programmevaluation in dem hohen Korruptionsdruck, der durch persönliche, politische und kommerzielle Interessen entsteht. Die Betroffenen – Schüler, Lehrer, Administratoren, Bildungsverlage etc. – tun alles, auch dysfunktionale Dinge, um Nachteile durch schlechte Testleistungen zu vermeiden. Wenn Testleistungen mit Sanktionen verbunden sind (*high stakes testing*), aber die für bessere Leistungen erforderlichen Ressourcen nicht bereitgestellt werden oder nicht verfügbar sind, bleibt den Betroffenen – auf allen Ebenen des Schulsystems – oft nichts anderes übrig als zu betrügen. Natürlich hat es Schummeln und Betrug in der Schule schon immer gegeben. Aber mit der Verschärfung der Sanktionen und der Veröffentlichung von Ergebnissen in Form von Ranglisten hat der Korruptionsdruck im Bildungswesen eine neue Qualität erhalten (Berliner/Biddle 1995).

Das muss nicht so sein, wie ein Beispiel aus Kolumbien zeigt. Dort wurden vor einiger Zeit landesweit Englischlehrer auf ihre Sprachfähigkeit getestet, um herauszufinden, wer eine spezielle Förderung nötig hat. Die „Schlechten" wurden aber nicht bestraft, sondern durch ein intensives Fortbildungsprogramm belohnt (persönliche Mitteilung der früheren Bildungsministerin, Cecilia Maria Velez). Damit war kein Anreiz gegeben, in dem Test zu schwindeln. Eine solche konstruktive Verwendung von Personenevaluation ist aber leider ein Einzelfall. Die Regel sieht anders aus.

Die allgegenwärtige Korruptionsgefahr bei Evaluation, die jede evidenzbasierte Pädagogik unterminiert, ließe sich meines Erachtens auf zwei Wegen vermeiden: Erstens, indem man die Ergebnisse nicht mit Strafen für schlechte Werte, sondern, wie in Kolumbien, mit gezielten Fördermaßnahmen verknüpft, und zweitens, indem man Sanktionen ganz ausschaltet. Da aber jedes Bekanntwerden von Testdaten sanktionierenden Charakter hat, ist dies, so scheint mir, nur möglich, wenn Programmevaluation vom jeweiligen Nutzer von Evaluation, also als Selbstevaluation durchgeführt wird und anonym bleibt. Selbstevaluation, so mein Plädoyer, erlaubt am ehesten die Erzeugung von korruptionsfreier, unverzerrter Evidenz für pädagogische Reformen und produktivere Hinweise für Verbesserungen des Unterrichts und des Schulsystems (Barber 1999; Lind 2004).

Personen-Evaluation: *Rankismus*

Wer von evidenzbasierter Pädagogik spricht, meint damit meist eine Evidenz, die auf den Ergebnissen von Vergleichstests beruht. Das Hauptziel der Konstruktion und Auswertung solcher Tests ist es, die Streuung der Werte zwischen den Teilnehmern so groß wie möglich zu machen (oder sie zumindest groß erscheinen zu lassen), damit ein möglichst *eindeutiges Ranking* der Teilnehmer möglich ist. Menschen anhand von Testaufgaben in eine eindeutige Rangreihe zu bringen, ist aber nur möglich, wenn zwei Bedingungen erfüllt sind, nämlich a) wenn die Testwerte weit streuen, d. h. wenn nicht alle Testteilnehmer alle Aufgaben lösen können (was eigentlich das Ziel eines guten Unterrichts wäre), sondern möglichst kein Teilnehmer alle Aufgaben lösen kann, und b) wenn die Aufgaben eines Tests nur eine einzige, eindimensionale Fähigkeit erfordern. Dies ist in der Realität oft nur durch Kniffe und Tricks zu erreichen, die den Wert von Testdaten als Evidenz für pädagogisches Handeln stark einschränken.

Beide Bedingungen können annähernd erfüllt werden, solange der Test für eine kleine, klar umrissene Schülergruppe wie eine bestimmte Schulklasse gedacht ist und die Testaufgaben von Experten konstruiert werden, die den Lehrplan dieser Schüler ebenso kennen wie das Testfach (wie Mathematik) und die affektiv-kognitiven Prozesse, die an der Bearbeitung des Aufgaben beteiligt sind. Vergleichstests werden jedoch für den Vergleich von sehr vielen Personen (z. B. bei PISA für alle Fünfzehnjährigen vieler Nationen) mit unterschiedlichen Lernbiographien und sozialen und kulturellen Hintergründen durchgeführt. Je größer und heterogener die

Zielgruppe eines Tests ist, umso mehr Faktoren kommen ins Spiel, die einen schwer kalkulierbaren Einfluss auf die Testwerte haben. Dadurch werden sie vieldeutig und schwer interpretierbar (Meyerhöfer 2007; Popham 1999; Wuttke 2007). Zudem muss man, um die Testaufgaben trotz sehr unterschiedlicher Interessen, Lehrpläne und Schulqualität für Testteilnehmer aus verschiedenen Klassen, Schulen und Ländern akzeptabel zu machen, ihr fachliches Niveau absenken. Wenn man die Testaufgaben genau anschaut, stellt sich oft heraus, dass ihr eigentlicher Kern die gedachte Zielgruppe vor leichte Aufgaben stellt. Zum Beispiel müssen 15-Jährige bei der Mathematik-Aufgabe „Bauernhäuser" aus der PISA-Studie für eine Fläche von 12 mal 12 Metern den Flächeninhalt zu berechnen. (Was die Aufgabe dann doch wieder schwer macht, sehen wir weiter unten.) Von diesem Nivellierungseffekt sind auch die Zentralabiture betroffen, wie Klein (2010) für das Fach Biologie in einem Versuch nachgewiesen hat. Er hat dazu Schülern der neunten Klasse Abituraufgaben vorgelegt, die sie noch nicht durchgenommen hatten. Die Neuntklässler lösten diese Aufgaben so gut, dass sie bis auf einen alle das Abitur in Biologie geschafft hätten. „Im Gegensatz zu den Abiturprüfungen vor dem Zentralabitur reicht für die neue kompetenzorientierte Aufgabenstellung *Lesekompetenz* aus, um die Aufgabenstellung bearbeiten und lösen zu können. Ein grundlegendes biologisches Fachwissen braucht der Schüler nicht einzubringen. Falsche inhaltliche Darstellungen werden aufgrund des vorgegebenen Erwartungshorizonts und des zugewiesenen Punktesystems nicht mehr bewertet" (Klein 2010, S. 15). Wenn aber die Testaufgaben so leicht sind, dass alle Teilnehmer sie lösen können, ist kein Ranking mehr möglich.

Um dennoch die Bedingungen für Rankings, nämlich breite Streuung und Eindimensionalität zu erfüllen, greifen die Testkonstrukteure zu zwei Kniffen, die zwar den allgegenwärtigen „Rankismus" (Fuller/Gerloff 2008) befriedigen, aber den Wert solcher pädagogischen Evidenz schmälern, wenn nicht sogar aufheben. Kniff 1: Die Testautoren vergrößern die Skalen durch Addition und Multiplikation. Bei PISA und anderen Vergleichstests hat es sich eingebürgert, den Mittelwert willkürlich auf 500 und die Standardabweichung auf 100 zu setzen. Aus der Lösung einer einzigen Aufgabe wird dadurch ein Testwertzuwachs von 18 Punkten (Wuttke 2007, S. 157). Aber das löst nicht das Problem, dass die Aufgaben zu leicht sind und die Testwerte zu eng beieinander liegen, um Rankings zu erstellen. Daher benötigen die Autoren Kniff 2: Sie erschweren die Aufgaben künstlich durch Dinge, die mit der zu messenden Fähigkeit wenig oder gar nichts zu tun haben:

- *Zeitdruck:* Für das Ausfüllen der Tests wird so wenig Zeit bewilligt, dass kein Teilnehmer alle Aufgaben erfolgreich bearbeiten kann. Dies senkt nicht nur die mittleren Testwerte, sondern führt auch zu einer breiteren Streuung der Testwerte, da Menschen unterschiedlich gut mit Zeitdruck fertig werden. Diese Art der Erschwernis wird aber offenbar selbst von den Testautoren nicht als legitim angesehen, sonst würden sie das in ihren Publikationen begründen. Der Umgang mit Zeitdruck wird in kaum einem Unterricht geübt. Unnötiger Zeitdruck untergräbt das Selbstwertgefühl der Schüler und ihre Lernfreude. Sie erhalten dadurch ständig die Rückmeldung, dass sie den Anforderungen nicht genügen (Kohn 2000; Madaus/Clarke 2001).
- *Distraktoren:* Das sind Informationen in einer Testaufgabe, die für ihre Lösung unnötig sind, aber die Testteilnehmer irritieren sollen. Bei der PISA-Aufgabe „Gentechnisch verändertes Getreide“ beispielsweise besteht die Hälfte der Einleitung aus einem Zeitungsartikel, der für die Lösung der Aufgabe völlig irrelevant ist, beim Lesen den Testteilnehmer aber auf falsche Spuren lenkt. Beides kostet Zeit und Testpunkte (Lohner 2008, S. 23f.). Auch hier haben die Erschwernisse nichts mit „biologischer Kompetenz” zu tun und erfordern Fähigkeiten, die nicht auf dem Lehrplan stehen. Dadurch haben jene Schüler einen Vorteil, die aufgrund einer günstigen außerschulischen Lernumwelt schnell lesen können oder sich von Distraktoren nicht irritieren lassen. Man könnte argumentieren, dass die Fähigkeit, Relevantes von Irrelevantem zu unterscheiden, eine wichtige Fähigkeit darstellt. Aber dann müsste man dafür einen eigenen Test konstruieren. Die Vermengung dieser Fähigkeit mit Fachkenntnissen, Lesefähigkeit, Stressresistenz und anderem mehr in einer einzigen Aufgabe führt zu Vieldeutigkeit und Mehrdimensionalität der Tests. Wenn wir aus den Testergebnissen nicht eindeutig ablesen können, welche dieser Teilfähigkeiten sie anzeigen, können wir aus ihnen keine pädagogischen Maßnahmen ableiten.
- *Textlastigkeit der Aufgaben:* Vor allem bei Mathe- und Naturwissenschaftstests, die früher überwiegend aus Zahlen und Schaubildern bestanden, überwiegt heute Text. Angeblich ist dies durch die Bemühung der Testautoren bedingt, die Aufgaben „praxisnah“ zu machen. Aber das ist nicht die ganze Geschichte. Zum einen handelt es sich bei Texten immer nur um vorgestellte, verbalisierte Praxis und nicht um die Einbettung von Mathematik oder Physik in ein wirkliches Praxisproblem. Das heißt, die Testperson muss nicht nur das Problem einer Aufgabe erfassen, sondern sie muss auch einen (zum Teil komplizierten) Text lesen können. Betrachtet man die Auswertungsregeln, so

wird aber schnell klar, dass die Fähigkeit, die Komplexität wirklicher Probleme zu erfassen, gar nicht honoriert wird. Ein Beispiel ist die bei TIMSS verwendete Aufgabe mit dem Stromwiderstand einer Glühbirne, der sich bei Vergrößerung der Stromstärke in der Realität anders verhält als es in der Ohmschen Formel beschrieben wird. Obwohl die Aufgabe „Praxisnähe" suggeriert, wird aber als „richtige" Antwort die Anwendung des Ohmschen Gesetzes verlangt (Hagemeister 1999). Ein weiteres Beispiel ist die bei XXX verwendete Aufgabe mit den Ziegelsteinen, die laut Zeichnung offenkundig verschiedene Größen und daher auch verschiedene Massen haben, die aber nur dann als „richtig" gelöst gilt, wenn man davon ausgeht, dass sie die gleiche Masse haben. Der Grund für die Ungenauigkeit war demnach nicht Absicht, sondern Schlampigkeit. Wer hier das Wort Praxis ernst nimmt, wird mit Verlust von Testpunkten bestraft. Weiter aufgebläht werden die Texte durch Versuche der Autoren, die Aufgaben durch Ergänzungen mit Synonymen und Umschreibungen unterschiedlichen Zielgruppen verständlich zu machen. Das erfordert mehr Lesezeit und führt oft zu noch mehr Verwirrung (Meyerhöfer 2007). Zusätzlich werden Testaufgaben für die Teilnehmer unbeabsichtigt durch Konstruktionsfehler erschwert, die jedem Testautor (wie auch jedem Lehrer) unterlaufen können (Rhoades/Madaus 2003; Wuttke 2006). Nichtsdestotrotz wirken diese Fehler für so manche/n Schüler/in verwirrend und kosten ihn oder sie wertvolle Zeit und Punkte.

- *Weltwissen:* Viele Aufgaben erfordern Wissen, das den Jugendlichen nur außerhalb der Schule, im Elternhaus und im öffentlichen Leben, vermittelt wird. In der PISA-Aufgabe „Mary Montagu" wird gefragt, gegen welche Krankheiten man sich impfen lassen kann. Zur Auswahl stehen Erbkrankheiten, Virenbefall („z. B. Kinderlähmung"), Funktionsschwäche („z. B. Zuckerkrankheit") und alle unheilbaren Krankheiten. Schüler, die selbst schon mit dem Thema Impfen konfrontiert wurden, können auch dann die richtige Antwort ankreuzen, wenn sie sich noch nie mit Biologie befasst haben (Lohner 2008).
- *Testschlauheit:* Angeblich sollen die Aufgaben von Vergleichstests nicht nur Begriffswissen abfragen, sondern auch das Verstehen und Anwenden von Wissen testen. Verstehen und Anwenden benötigen Zeit zum Nachdenken. Manche Testaufgaben sehen tatsächlich so aus, als wenn sie Verständnis und Anwendung prüfen würden und verleiten Testteilnehmer auch dazu, sich intensiv mit ihnen zu befassen. Aber die Auswertungsmaschinerie der Testindustrie honoriert dies überhaupt nicht. Im Gegenteil, wegen der bewusst sehr eng ge-

haltenen Zeitlimits (siehe oben) wird Nachdenken mit Punkteverlust bestraft, weil durch die benötigte Zeit zum Nachdenken weniger Aufgaben bearbeitet werden können. Man schneidet bei diesen Tests also besser ab, wenn man die Ideologie dieser Tests durchschaut hat und einfach versucht, so viele Aufgaben wie möglich zu bearbeiten. Dann gewinnt man zusätzlich PISA-Punkte schon allein dadurch, dass man mehr Aufgaben bearbeitet als der „Denker", ohne dass man dafür eine spezielle Fähigkeit braucht. Man muss nur raten. Da es immer nur vier Alternativen gibt und immer eine davon richtig sein muss, kommen auf viermal Raten eine richtige Lösung und damit 18 zusätzliche PISA-Punkte. Schüler in anderen Ländern mit langer Testtradition wissen das und zeitigen daher bessere PISA-Ergebnisse als testunerfahrene Schüler.

Die hier genannten Aufgabenbeispiele sind keine Einzelfälle. Kein Wunder also, dass man zu absurden Schlussfolgerungen kommen kann, wenn man PISA-Testwerte als pädagogische Evidenz betrachtet. Die PISA-Autoren rechnen PISA-Punkte gern in Unterrichtszeit um. Demnach sollen 18 Punkte Unterschied in dem Test (also eine gelöste Aufgabe mehr oder weniger) einem Unterschied von 4,5 Monaten entsprechen. Spinnt man dieses Kalkül fort, so folgt daraus, dass sich durch Instruktion der Schüler, sie sollten Schulaufgaben durch Raten statt durch Nachdenken lösen, ein ganzes Schuljahr oder mehr einsparen ließe – wenn es uns nur darum ginge, unsere Schulen bei internationalen Rankings gut auszusehen zu lassen.

Viele der fachfremden Zusätze bei den Testaufgaben ermöglichen zwar Rankings, verfälschen aber deren Bedeutung, mindern ihr Validität und machen es fast unmöglich, hieraus Hinweise für pädagogische Reformen zu gewinnen. Einige Zusätze sorgen dafür, dass viele Schüler an den Aufgaben scheitern, bevor sie die Möglichkeit haben, sich an dem eigentlichen Kern der Aufgabe zu versuchen. Andere ermöglichen es, Punkte zu gewinnen, ohne die eigentliche Aufgabe gelöst zu haben. So kann man bei vielen Testaufgaben punkten, wenn man schnell lesen kann, wenn man schnell rät oder man sich nicht von verwirrender Zusatzinformation verunsichern lässt – ohne dass man von dem Fach viel versteht. Oft ist man sogar im Nachteil, wenn man von dem getesteten Fach viel versteht und die Aufgaben ernst nimmt, weil man dadurch Zeit und Punkte verliert (Neuweg 2004).

Die „Beimischung" von Dingen, die mit zu messenden Fähigkeiten nichts zu tun haben, um die Testwerte zu spreizen und so besser eine Rangreihe bilden zu können, führt – ebenso wie die Absenkung der fachlichen Anforderungen bei den Testaufgaben – zu einem Paradoxon. Die-

ses besteht darin, dass entgegen der behaupteten Eindimensionalität das Testinstrument vieldimensional und damit untauglich für ein Ranking wird. Um sagen zu können: Schüler A ist besser als Schüler B oder Land X hat ein besseres Schulsystem als Land Y, müssen sich die Testergebnisse ohne großen Informationsverlust und ohne Willkür auf einer einzigen Dimension anordnen lassen.

Damit sich die Tests auf einer einzigen Dimension anordnen lassen, bedienen sich die Autoren von Vergleichstests erneut eines Kniffs, der die Kosten weiter nach oben treibt und die Test-Validität verringert. Man legt einfach fest, dass die gemessene Fähigkeit eindimensional ist und sucht so lange nach Testaufgaben (mit einem gewissen Bezug zur Testdimension), bis diese Annahme stimmt. Man lässt sehr viele Aufgaben von sehr vielen Testpersonen in mehreren Runden bearbeiten und sondert dann mittels statistischer Analysen (*Item-Response-Analyse*) diejenigen für den Test aus, die den Test eindimensional machen, so dass sich Aufgaben und Personen in eine (möglichst) eindeutige Rangreihe bringen lassen (Wilson 2005). Eindimensionalität ist also keine Eigenschaft der gemessenen Fähigkeit, sondern das Ergebnis eines bestimmten Vorgehens bei der Testkonstruktion. Die Validität solcher Tests lässt sich nicht überprüfen, da ihnen keine psychologische Theorie zugrunde liegt (Schoenfeld 1999). Da solche Konstruktionen nicht falsifizierbar sind, ist mit ihrer Anwendung auch kein Erkenntnisgewinn verbunden (Popper 1968).

Auch hier werden wieder keine Kosten gescheut, die Fiktion von Eindimensionalität und Modellpassung aufrecht zu erhalten. Je größer die untersuchte Stichprobe, so scheint man zu hoffen, desto weniger leicht ändert sich die kunstvoll hergestellte Fiktion. Modellpassung ist aber keine Eigenschaft eines Tests, sondern das Ergebnis der Interaktion zwischen einem bestimmten Test und bestimmten Testpersonen (Allerup 2007). Dass es sich bei der Eindimensionalität nicht um die Eigenschaft der angeblich gemessenen Fähigkeit handelt, sondern um ein Artefakt, zeigt sich schon darin, dass ein Test, der in einer Studie und auf die gesamte Stichprobe bezogen perfekt den Vorgaben der Statistiker entspricht, dies beim nächsten Einsatz meist schon nicht mehr tut, auch nicht in Untergruppen wie Frauen und Männern oder in verschiedenen Ländern.

Wuttke (2007) hat bei seiner Nachanalyse der Mathematikaufgaben in der PISA-Studie so starke Abweichungen von der Eindimensionalität und der lokalen stochastischen Unabhängigkeit gefunden, dass man selbst in diesen hoch selektierten Aufgaben von einer Drei- bis Vierdimensionalität sprechen muss – was die Bildung einer Rangreihe der Testteilnehmer nach ihrer „Mathematik“-Fähigkeit verbietet. Zudem fand er bestätigt,

dass einige Aufgaben bei verschiedenen Gruppen verschiedene Eigenschaften der Testpersonen messen.

Die Tatsache, dass die Selektion von Aufgaben nach ihrer Verträglichkeit mit dem vorgegebenen Modell ihre Gültigkeit (Validität) verringert, hat bereits Cronbach (1960), ein Großmeister der Testpsychologie gewusst: „dropping items with low correlations may reduce content validity“ (S. 157). Die Autoren der Vergleichstests wiegen sich dagegen in dem Fehlglauben, dass die Validität einer Testaufgabe dasselbe sei wie ihre Modellverträglichkeit (Wilson 2005), was von einem großen Unverständnis zeugt.[1]

Der Versuch, eine große, kunterbunte Ansammlung von Drittklässlern (wie bei VERA) oder 15-Jährigen (wie bei PISA) auf einer einzigen Skala aufzureihen, musste schief gehen. Dass dies von der Öffentlichkeit kaum bemerkt wird, liegt wohl daran, dass die Testautoren sich bislang weitgehend gegen Kritik abschirmen konnten. Die Kultusminister haben es versäumt, die Testkonzepte der öffentlichen Diskussion zugänglich zu machen, bevor sie den Auftrag erteilt haben. Durch eine öffentliche Diskussion hätten die oben dargestellten Probleme schneller erkannt und vermieden werden können. So haben sich die großen Vergleichsstudien eine Parallelwelt erschaffen können, in der nicht auffiel, wie gering der Wert der Evidenz ist, die sie erzeugt haben.

Die Vergleichstests scheinen im Endeffekt nichts weiter zu messen als die Fähigkeit, Testaufgaben zu bewältigen (Meyerhöfer 2007; Popham 1999; Wuttke 2006). Vergleichstests haben auch kaum Prognosewert. Sie korrelieren zumeist nur mit Tests von der gleichen Machart (Koretz 2009). Es gibt keine Belege dafür, dass der Einsatz von Vergleichstests, wie von vielen erhofft, zu besseren Lernleistungen führt (Sacks 1999; Amrein/Berliner 2002; Nichols/Glass/Berliner 2006). Entsprechend mehren sich auch die Klagen, dass die vielen aufwändigen Vergleichstests für den Unterricht bislang kaum nutzbar gemacht werden konnten, dass sie sich aber negativ auf den Unterricht auszuwirken beginnen, weil dieser sich immer mehr an diesen Tests ausrichtet, statt dass die Tests lebens- und arbeitsbezogenen Bildungszielen angepasst werden (Jablonka 2006).

1 „Wahrscheinlich lassen Validität und transkulturelle Äquivalenz der Tests in vielen der älteren Schulleistungsuntersuchungen zu wünschen übrig. Wie sich in den technischen Berichten zu PISA nachlesen lässt, hat man aber inzwischen große Anstrengungen unternommen, um die Reliabilität und Objektivität der Tests in allen Teilnehmerländern zu erhöhen und ihre transkulturelle Vergleichbarkeit zu gewährleisten.“ (Schümer 2006, S. 263). Validität scheint keine Anstrengung wert zu sein.

Vergleichstests sind aber nicht nur problematisch als Grundlage für bildungspolitische und pädagogische Entscheidungen. Sie produzieren auch erhebliche *Kollateralschäden*. Sie stellen, wie wir sahen, hohe Anforderungen an die Lesegeschwindigkeit, das Weltwissen und die Fähigkeit der Testteilnehmer, sich nicht durch Zeitdruck, Distraktoren und Konstruktionsfehler der Testautoren irritieren und ängstigen zu lassen. Es liegt auf der Hand, dass dadurch besonders Kinder aus sozial schwachen Elternhäusern (Armut, Arbeitslosigkeit) benachteiligt werden (Kreitzer/Madaus/Haney 1989; Kozol 1992; Sacks 1999; Madaus/Clarke 2001; Cunningham/Sanzo 2002; Nichols/Berliner 2006; Belley/Lochner 2008). Welche negativen Rückwirkungen der Korruptionsdruck von *high-stakes tests* auf das Lernen in der Schule und die Funktion des Schulwesens hat, lässt sich bislang am besten am Beispiel der USA studieren, wo es solche Tests schon sehr lange gibt und seit vierzig Jahren die Bundesregierung Lehrer, Schulen und Schulbezirke mittels Vergleichstest zu einer besseren Pädagogik anzutreiben versucht (Lind 2009a). Über die Korruption auf allen Ebenen des Bildungswesens liegen inzwischen viele Studien vor, die einen Einblick in ihre Ausbreitung und Auswirkungen geben (Amrein/Berliner 2002; Berliner/Biddle 1995; Bracey 2002, 2005; 2007; Kreitzer/Madaus/Haney 1989; Madaus/Clarke 2001; Nichols/Berliner 2006; Sacks 1999). Die Auswirkungen der Vergleichsstudien auf den Unterricht sind auch im Deutschland überall zu spüren. Studien zu diesen Auswirkungen stehen noch weitgehend aus. Aber es gibt sehr gute Analysen zu der Validität von Vergleichsstudien wie PISA (Jahnke/Meyerhöfer 2007; Hopmann/Brinek/Retzl 2007; Wittmann 2010).

Schafft Programmevaluation valide Evidenz?

Kann Evidenz aus Programmevaluation eine brauchbare Grundlage für pädagogisches und bildungspolitisches Handeln sein? Oft auch Methoden- oder Maßnahmen-Evaluation genannt, beruht diese auf der Idee, dass *Effektivität* und *Effizienz* von Unterrichtsmethoden und bildungspolitischen Maßnahmen im Hinblick auf bestimmte Bildungs- und Lernziele

durch Programmevaluation gesteigert werden können (Campbell 1969; Sanders 1994; Schoenfeld 1999).[2]

Entscheidend für die Frage der pädagogischen Relevanz ist, dass die Befunde an alle Betroffenen kommuniziert werden können und dass sie verständlich und vertrauenswürdig sind. Tatsache ist aber, dass die meisten der großen Evaluationsstudien der letzten hundert Jahre, die sich nicht auf Personen richteten, sondern auf pädagogische Maßnahmen und Methoden, selbst in der Fachöffentlichkeit wenig bekannt sind. Die Verbreitung der Ergebnisse wird oft auch durch die Komplexität der Forschungsanlage und der statistischen Auswertung erschwert, auch wenn sie meist transparenter sind als bei der Personenevaluation. Entscheidend für die Brauchbarkeit der damit gewonnenen Evidenz ist aber auch die Frage, ob die Evidenz aus Programmevaluationsstudien vertrauenswürdig ist, also ob sie nicht unter Korruptionsdruck steht.

Bei Evaluation, die auf Methoden und Maßnahmen gerichtet ist, muss man weniger mit plumpem Betrug rechnen als bei Evaluation, die auf Menschen gerichtet ist, deren materielle Interessen direkt von den Daten abhängen, die sie liefern. Sofern in die evaluierten Programme aber ideologische, politische oder wirtschaftliche Interessen investiert wurden (Bracey 2005), ist auch hier mit Korruptionsdruck im Sinne von *Campbell's Law* zu rechnen. Dieses „Gesetz" geht auf den renommierten Sozialpsychologen und Evaluationsexperten Donald T. Campbell zurück, der dies schon vor vierzig Jahren vorhergesagt hatte: „The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell 1975, S. 35).

Der von Campbell vorhergesagte Korruptionsdruck ist inzwischen für Vergleichsstudien gut belegt (Amrein/Berliner 2002; Nichols/Berliner 2005, 2006). Bei Programmevaluation ist der Korruptionsdruck subtiler. Er wirkt über vorgeblich wissenschaftliche Anforderungen an Evaluationsstudien, wie zum Beispiel durch die Anforderung, die Aufteilung der Versuchsteilnehmer zu randomisieren, um alle denkbaren Störfaktoren auszuschalten, und die Vorgabe, Effekte mittels statistischer „Signifikanz" nachzuweisen. Beide Anforderungen machen meist teure Studien notwendig, die sich nur jemand leisten kann, der gute Drähte zu Regierungen oder Sponsoren hat. Randomisierte Experimente, bei denen Testper-

2 Die Frage, ob der Aufwand und Ertrag von solchen Evaluationen und den durch sie ausgelösten Reformen in einem günstigen Verhältnis zueinander stehen, muss aus Platzgründen hier ausgeklammert bleiben.

sonen nach Zufall auf Experimental- und Kontrollgruppen aufgeteilt werden müssen, kosten viel Geld. Anders als in der Agrarwissenschaft stößt die Zufallsaufteilung der Versuchsteilnehmer schnell an ethische, wissenschaftliche und praktische Grenzen. Man kann Schüler nicht beliebig umsetzen, um vermeintliche Störfaktoren zu kontrollieren, und auch nicht zwingen, an einem Versuch teilzunehmen. Das Aufbrechen von natürlichen Bindungen (Schulklasse), wie das einige aus methodischen Gründen für notwendig halten, kann selbst zum Störfaktor werden. Es ist schwer zu verhindern, dass die Kontrollgruppen auch von der Intervention beeinflusst werden, die nur für die Experimentalgruppe gedacht war (*cross-over effects*). All diese potentiellen Störfaktoren auszuschließen oder zu minimieren, bedeutet einen hohen Kostenaufwand. Die teilnehmenden Schulen haben in der Regel keinen Gewinn von solchen Studien und verlangen daher für ihre Teilnahme eine (zum Teil sehr hohe) Entschädigung.

Ein bekanntes Beispiel ist das Leseförderprogramm *Success for All* (SFA), das inzwischen an einigen tausend Grundschulen in den USA von der US-Regierung gefördert wird. Es wird auch von der *Best Evidence Encyclopedia* (BEE) empfohlen. Die BEE wurde vom „Zentrum für datengetriebene Reformen in der Bildung“ im Internet kreiert, das vom US-Bildungsministerium finanziert wird. Dort können Lehrer, so das Versprechen, viele Programme finden, mit denen sie die Leistungen ihrer Schüler verbessern können. Versprochen werden „reliable, unbiased reviews of research-proven educational programs” (CDDRE 2004ff.).

Success for All wurde in verschiedenen Studien evaluiert, unter anderem in einer Studie von den Autoren selbst (Borman u. a. 2005). An dieser Studie lässt sich gut prüfen, wie „reliable“ und „unbiased“ die Evidenz ist, die Reformen in der Bildung antreiben soll. Sie scheint die US-Regierung vor allem deshalb überzeugt zu haben, weil sie sehr viele Teilnehmer hatte (fast 5.000 Schüler) und weil sie randomisiert war, also die Schüler zufällig auf Experimental- und Kontrollgruppe aufteilte. An der Effektivität dieses Programms kann es kaum gelegen haben. In nur einem der vier eingesetzten Lesetests zeigt sich bei dem Vergleich von Vor- und Nachtest ein „signifikanter“ Zuwachs der Lesefähigkeit. Bedeutsam war das Ergebnis allerdings nur statistisch, und das auch nur, weil sich in so großen Stichproben bereits geringfügige Effekte als „statistisch signifikant“ erweisen (Carver 1993; Sedlmeier/Köhlers 2001). Die relative Effektstärke von SFA beträgt nach Borman et al. (2005) $r = .12$. Das ist ein sehr geringer Wert, wenn man bedenkt, dass man üblicherweise erst ab einer Effektstärke von $r = .30$ von „effektiv“ spricht (Lipsey/Wilson 1993). Auch sucht man vergeblich nach Analysen zu der Frage, inwieweit

leistungsschwache Schüler von dem Programm profitieren, denn schließlich verspricht sein Name einen Erfolg für *alle*. Langzeitstudien zur Vorschulerziehung zeigen zudem, dass lehrerzentrierter Unterricht, *teacher-directed instruction* und *scripted teaching*, worauf SFA basiert, meist nur in einem eng definierten Lerngebiet Effekte erzielen können und das oft auch nur kurzfristig: „Direct instruction curriculum prevents children from developing autonomy because the teacher is authoritarian and uses rewards and punishments and [...] the other two curricula encourage children's autonomy because they allow teachers and children to discuss their points of view with one another" (Schweinhardt/Weikart 1988, S. 222). Jedenfalls nimmt sich der „Erfolg" von SFA gemessen an seinen Kosten (ca. 135.000 US-Dollar pro Schule kostet allein die Implementierung von SFA) so bescheiden aus, dass man sich verwundert fragen muss, weshalb die US-Regierung das Programm so überschwänglich empfiehlt und finanziell unterstützt (Poynor/Wolfe 2005).

Ob Randomisierung notwendig oder sinnvoll ist, ist umstritten, aber sie ist auf jeden Fall teuer, sehr teuer, und damit ein möglicher Ansatzpunkt für subtile Formen der Korruption. Borman et al. (2005) berichten, dass erst dann genügend Schulen bereit zur Teilnahme waren, als man die Entschädigung pro Schule von 15.000 auf 75.000 US-Dollar erhöht habe. Allein die Kosten für die ca. 40 teilnehmenden Schulen beliefen sich damit auf ca. 3 Millionen US-Dollar. Die Befunde waren, so die Autoren, „similar to those of earlier matched experiments" (ebd., S. 1), also nicht besser als in Experimenten, in denen die Kontrollgruppe so ausgewählt wird, dass sie der Experimentalgruppe in relevanten Merkmalen entspricht (*matched pair*). Systematische Studien zeigen, dass nichtrandomisierte Experimente, die sehr viel billiger sind, und Meta-Analysen (Rosenthal 1986) meist die gleichen Ergebnisse erbringen wie sehr teure, randomisierte Studien (Cohen et al. 1982; Lipsey/Wilson 1992). Zudem ist Randomisierung immer nur begrenzt möglich und nie können alle Störfaktoren ausgeschaltet werden. Die Glaubwürdigkeit von Evidenz ist auch nicht nur von dieser technischen Vorkehrung abhängig, sondern von dem Gesamt an Wissen, das wir über den psychologischen Prozess haben, der mit einer bestimmten Methode gefördert werden soll. An solchem Wissen mangelt es aber oft bei kommerziell „erfolgreichen" Lehrprogrammen (Schoenfeld 1999).

Der ganze Aufwand für Randomisierung lohnt sich also pädagogisch kaum – aber wirtschaftlich. Nur weil die Methode, die offenkundig keinerlei substantielle Wirkung hat, mit sehr hohen Kosten an randomisierten Samples überprüft wurde und die Ergebnisse durch ebenfalls kostentreibendes Aufblähen der Stichproben statistisch ein bisschen signifikant

wurden, sieht die US-Regierung *Success for All* als „wirksame“ Methode an und fördert inzwischen ihren Einsatz in mehreren tausend Schulen. Wer sich diesen Kostenaufwand nicht leisten kann, weil Regierung und Sponsoren solche teuren Studien nur beschränkt fördern können, kann nicht zeigen, dass seine Unterrichtsmethode wirksamer und vielleicht dazu auch noch billiger ist. Ob beabsichtigt oder nicht, legen die Geldgeber durch ihre Vorgaben (Randomisierung, Signifikanz) und die Auswahl der Mittelempfänger bereits fest, welche Programme die Voraussetzungen für den späteren Einsatz in der Schule und damit eine Förderung erhalten werden und welche nicht. Ob die geförderten Programme das Lernen der Schüler effektiv fördern, scheint dabei kaum eine Rolle zu spielen. In dieser Situation ist der Korruptionsdruck groß. Anbieter von Lernprogrammen können leicht in Versuchung kommen, mit Geld oder Beziehungen die alles entscheidende Unterstützung für teurere Evaluationsstudien zu gewinnen. Tatsächlich hat Bracey (2005) am Beispiel des *No Child Left Behind*-Programms der US-Regierung gezeigt, wie staatliche Auftragsgelder für Lehrmittel und Lernprogramme an Schulverlage teilweise wieder zurückfließen in die Wahlkampfkassen der Politiker, die sie genehmigt haben.

Ein Gegenbeispiel ist die wissenschaftlich sehr viel anspruchsvollere und ergebnisreiche Evaluation der Wirksamkeit von Religionsunterricht auf das (un-) moralische Verhalten von Schülern, publiziert als *Studies in the Nature of Character* (Hartshorne/May 1928). Sie hat kaum Eingang in die Pädagogik gefunden und wird von der Bildungspolitik vollkommen ignoriert. Die Studie wurde von der *Religious Education Association*, einer Vereinigung evangelischer Kirchen in den USA finanziert (Fisher 1928). Die Forscher führten neben umfangreichen Befragungen auch mehrere aufwändige Verhaltensexperimente zum *Täuschungsverhalten* von Kindern in Schulsituationen durch (Hartshorne/May 1928). Der wohl erstaunlichste Befund ist, dass für die Teilnahme am freiwilligen Religionsunterricht – entgegen der Hoffnung der kirchlichen Auftraggeber – keine Moral fördernde Wirkung gefunden werden konnte, allenfalls eine negative. Im Vergleich dazu fanden die Forscher einen positiven Effekt der Teilnahme an reformpädagogischen Schulen (*progressive education*), wo, so die Autoren, die Beziehung zwischen Lehrer und Schülern durch eine Atmosphäre der Kooperation und des guten Willens gekennzeichnet ist und die Kinder ohne Noten und Zwang lernen können. Drittens fanden sie, dass Täuschungsverhalten stärker dadurch motiviert war, gute Leistungen zu erbringen, als durch einen „schlechten Charakter“. Je schwächer die Leistung in fachlichen Tests, umso intensiver war das Täu-

schungsverhalten. Die Korrelationen reichten von r = -.05 bis -.51 (Hartshorne/May 1928, S. 395).

Obwohl diese Studie von einer Kirchenvereinigung finanziert wurde, wissenschaftlich sehr sauber durchgeführt war und eine Menge interessanter Befunde erbrachte, blieb sie für die pädagogische Praxis weitgehend folgenlos. Obwohl sie die Annahme widerlegte, dass das Verhalten von Kindern durch ihren „Charakter" geprägt sei, beschloss die US-Regierung vor einigen Jahren (unter Präsident Clinton) ein Millionenteures Programm zur Charakterbildung (*character education*). Obwohl die Studie gezeigt hat, dass der Religionsunterricht (in den *Sunday schools)* keine fördernde Wirkung auf die Einhaltung sozialer Regeln hat, hält man in den USA (und auch bei uns) unverändert an der Überzeugung fest, dass Religionsunterricht für die Moralentwicklung notwendig sei. Dabei wissen wir heute, dass Regelverletzungen und Straffälligkeit eher durch einen Mangel an Urteils- und Diskurskompetenz bedingt ist, deren Förderung eine ganz andere Lernumgebung verlangt (Schillinger 2006; Lind 2009b).

Ähnlich wie der *Character*-Studie erging es auch der so genannten *Achtjahres-Studie*, mir der systematisch untersucht wurde, wie gut sich Absolventen von reformpädagogischen Sekundarschulen (*progressive high schools*) im Studium zurechtfinden (Chamberlin u. a. 1942). In die Studie wurden 2950 Studierende einbezogen, von denen die Hälfte Absolventen der *progressive schools* war und die andere Hälfte vergleichbare Absolventen von normalen *high schools*. Sie wurden über die gesamt College-Zeit hinweg begleitend untersucht. Die Ergebnisse zeigen, dass „Reform-Schüler", die in der Schulzeit nicht mit Noten zum Lernen gezwungen wurden und keinen Aufnahmetest bestehen mussten, im College besser abschnitten als ihre Altersgenossen und dass sie diesen hinsichtlich sozialer Fähigkeiten deutlich voraus waren (vgl. ebd., S. 207f.): Sie erzielten einen etwas besseren Notendurchschnitt und bessere Noten in alle Fächern außer Fremdsprachen, dabei wählten sie dieselben Fächerschwerpunkte wie die Vergleichsgruppe. Sie erhielten jedes Jahr etwas mehr Auszeichnungen, wurden öfter positiv beurteilt, was ihre intellektuelle Neugierde und Wissbegierde anging, und wurden von ihren Lehrern in ihrem Denken öfter als präzise, systematisch und objektiv eingeschätzt. Von ihnen wurde auch öfter gesagt, dass sie klar entwickelte und gut formulierte Ideen haben. Sie beteiligten sich an allen Studenten-Organisationen außer an religiösen und reinen „Service"-Aktivitäten. Außerdem zeigten sich mehr von ihnen aktiv besorgt um das, was in der Welt vor sich ging. Damit wurde auf beeindruckende Weise gezeigt, dass Notendruck nicht notwendig ist, damit Schüler lernen. Im Gegenteil,

Noten scheinen das Lernen eher zu behindern. In einer Meta-Analyse von verschiedenen Maßnahmen, um das Lernverhalten zu verbessern, zeigte sich nur bei Noten ein negativer Zusammenhang mit objektiv gemessenen Lernfortschritten (Fraser et al. 1987). Diese Ergebnisse werden durch Schulversuche bestätigt, in denen die Kinder sehr viel Freiheit haben, selbst zu bestimmen, wie und was sie lernen (Peschel 2002).

Aber auch die Achtjahres-Studie hatte kaum Folgen für die pädagogische Praxis in den USA oder anderswo. Sie konnte auch nicht verhindern, dass die *progressive education*-Bewegung – im Zuge des Antikommunismus in den USA nach dem Zweiten Weltkrieg – als innerer Feind gebrandmarkt und dadurch fast völlig ausgelöscht wurde (Bracey 2007). Erst heute beginnt man sich wieder für die Befunde dieser Studie zu interessieren. Dominiert wird die pädagogische und bildungspolitische Debatte aber nach wie vor von den Ranking-Spektakeln, die durch Vergleichstests ausgelöst werden.

Wenn wir erfolgreich Pädagogik betreiben wollen, müssen wir *wissen, was wirkt*. Aber welcher Art von Evidenz können wir vertrauen, wenn Korruptionsdruck und politische und wirtschaftliche Interessen die Glaubwürdigkeit von Evaluationsergebnissen unterminieren und die Verbreitung und Anwendung glaubwürdiger Ergebnisse behindern können?

Evidenzbasierte Pädagogik durch Selbstevaluation

Selbstevaluation (SE) ist im Alltag allgegenwärtig, aber in der Pädagogik führt sie noch immer ein Schattendasein, auch wenn sich gelegentlich Handbuchartikel damit befassen (Barber 1999). Selbstevaluation handelt von Fragen wie: Ist mein Unterricht lehrwirksam? Lernen meine Schüler evtl. mehr mit einer alternativen Lehrmethode? Oder: Gelingt es mit meiner Bildungspolitik, auch schwächere Schüler zu einem guten Schulabschluss zu führen und moralische und demokratische Kompetenzen zu fördern? Bei Selbstevaluation handelt es sich um ein wissenschaftliches Projekt, das zum Ziel hat, mittels objektiver, unverzerrter, valider Daten Antworten auf solche Fragen zu liefern. Notwendig dafür sind sorgfältig geplante Wirkungsstudien, möglichst mit Vor- und Nachtests und mit Vergleichsdaten aus zugeordneten Kontrollstudien und/oder aus entsprechenden Großstudien. Selbstevaluation hat nichts mit den weit verbreiteten Rückmeldebögen zu tun, in denen Schüler angeben sollen, wie sie die didaktische Qualität eines Kurses beurteilen. Subjektive Einschätzungen durch Schüler (oder Meinungen von Bürgern) sind sehr wichtig, aber sie

können objektive Selbstevaluation nicht ersetzen. Zudem unterschätzt man oft die Schwierigkeit der Datengewinnung. Zum Beispiel überfordern Fragen nach der didaktischen Qualität einer Lehrperson zumeist die Schüler. Solche Einschätzungen werden meist durch allgemeine Sympathie oder Antipathie gefärbt und durch das Motiv geprägt, der Lehrperson zu gefallen – oder sie durch Kritik zu provozieren. Es wäre besser, Schüler danach zu fragen, ob sie etwas durch den Unterricht gelernt haben. Eine solche Frage kann von Schülern kompetenter beantwortet werden und bringt dem Lehrer oft gute Hinweise. Aber auch diese Informationen sind höchst anfällig für korrumpierende Einflüsse.

Selbstevaluation ist s*elbstbestimmte* Evaluation. Es geht also nicht um Evaluation, die man nur selbst durchführen (und bezahlen) darf, deren Fragestellung und Interpretation aber andere bestimmen. Es geht auch nicht um die Evaluation von einem selbst (die natürlich auch unter Korruptionsdruck steht, da Menschen ein starkes Interesse an einem positiven Selbstbild haben), sondern eben um selbstbestimmte und selbst interpretierte Evaluation konkreter Maßnahmen.

Eine so definierte Selbstevaluation, so die Hauptthese dieses Beitrags, kann vertrauenswürdige Evidenz für pädagogische Reformen liefern. Sie kommt immer auch direkt beim Nutzer an, weil Erzeuger und Nutzer hier identisch sind. Zudem wird sie in der Praxis auch eher kompetent umgesetzt. Eine Lehrerin, der die *eigenen* Daten zeigen, dass eine alternative Unterrichtsmethode die Schüler mit mehr Freude und schneller lernen lässt, wird sich kaum dem Charme dieser neuen Methode entziehen können. Sie kennt diese Evidenz nicht nur oberflächlich vom Hörensagen, sondern sehr intim und in ihrer vollen Bedeutung, da sie die Datengrundlage dafür ja selbst erzeugt hat. Aus demselben Grund vertraut sie dieser Evidenz mehr und fühlt sich stärker ermutigt, ihre pädagogische Praxis entsprechend zu ändern. Wenn sie zum Beispiel selbst Projektunterricht systematisch ausprobiert und evaluiert, wird sie sehen, worauf es bei dieser Methode ankommt, um Erfolg zu haben, und sie häufig einsetzen. Wer hingegen diese nur aus der Literatur oder von Fortbildungsveranstaltungen her kennt, läuft Gefahr, wichtige Dinge nicht mitzubekommen und schlechte Erfahrungen mit ihr zu machen.

Selbstevaluation stärkt die Professionalität und das Selbstvertrauen der Menschen, die sie anwenden, der Lehrpersonen ebenso wie der Kultusminister. Wenn ein Lehrer eine neue pädagogische Praxis und eine Kultusministerin ihr neues Schulgesetz selbst eingehend erprobt und evaluiert hat, werden beide ihre Neuerungen kompetenter und selbstbewusster gegenüber (notwendiger!) Kritik verteidigen und auch andere von ihren Reformen überzeugen können.

Selbstevaluation muss wie jedes wissenschaftliche Projekt bestimmten Regeln folgen, damit sie diese Wirkung entfalten kann und gegen Korruptionsdruck und Verzerrungen gefeit ist:

- Selbstevaluation muss sich auf objektive Daten stützen, die frei von den bekannten Effekten der sozialen Erwünschtheit und des Wunschdenkens sind. Es ist wichtig, sich zu vergewissern, dass die eigene pädagogische Arbeit oder Bildungspolitik Zustimmung und Akzeptanz findet. Aber diese Informationen sind so stark anfällig für Verzerrungen, dass sie nicht als valide Evidenz für die Wirksamkeit bestimmter Maßnahmen angesehen werden dürfen. Ich habe am Anfang meiner Lehrpraxis erfahren, dass Seminare und Vorlesungen, die von den Teilnehmern hoch gelobt wurden, bei ihnen kaum messbare Lerneffekte hatten. Sie hatten hohen Unterhaltungswert, aber kaum Lernwert.
- Selbstevaluation kann sich nur auf Evidenz stützen, die *erreichbar* ist. Je größer die Latenzzeit von erwarteten Effekten ist, umso größer ist der zeitliche und finanzielle Aufwand, sie nachzuweisen. Zum Beispiel kann die Fähigkeit, in wichtigen sozialen Positionen (wie als Eltern, als Chef oder als Wähler) Verantwortung zu tragen, erst dann ermittelt werden, wenn jemand real vor diese Aufgaben gestellt ist. Evidenz dazu ist nur durch Längsschnitt- oder Verbleibstudien erreichbar – wenn überhaupt.
- Selbstevaluation ist nur möglich, wenn Evidenz sich auf *beobachtbare bzw. messbare* Daten stützen kann, also wenn sie objektiv ist. Messbarkeit ist aber kein absolutes Kriterium. Es gibt Lern- und Bildungsziele, die leicht messbar sind, weil sie klar formuliert sind und weil es dafür bereits lang erprobte Messinstrumente gibt, deren Eigenschaften wir gut kennen, und es gibt Ziele, die schwer oder bislang nicht zu messen sind, weil sie (noch) zu vage und kontrovers formuliert sind und für sie (noch) keine vertrauten Messinstrumente vorliegen. Die Wissenschaft von der psychologischen Messung ist noch relativ jung und, wie wir oben gesehen haben, bei reinen Statistikern in den falschen Händen. Die Forschung und Entwicklung guter psychologisch-pädagogischer Messinstrumente müsste intensiviert werden (Schoenfeld 1999; Lind 2004).

Die Möglichkeiten und den Nutzen von Selbstevaluation hat Lind (2009b) über den Zeitraum von acht Jahren anhand von 40 Lehrveranstaltungen systematisch erprobt. Eines der Ziele dieser Lehrveranstaltungen war es, unabhängig von dem spezifischen Stoff die moralische Urteils- und Diskursfähigkeit zu fördern. In mehreren Studien hat sich gezeigt, dass diese Fähigkeit eine wichtige Voraussetzung für effektives Ler-

nen ist (Heidbrink 2010; Lind 2009c) und dass sie durch Gelegenheiten zur Verantwortungsübernahme und angeleiteten Reflexion (Schillinger 2006; Lupu 2009), besonders aber durch den Einsatz der *Konstanzer Methode der Dilemma-Diskussion* (KMDD) gezielt gefördert werden kann (Lind 2009c). Lind hat in einem Teil seiner Lehrveranstaltungen eine 90-minütige KMDD-Sitzung durchgeführt, und die Seminare – aufgrund des guten Erfolgs der KMDD – nach deren didaktischen Prinzipien umgestaltet, um ihre Lehrwirksamkeit zu vergrößern. Die Vorlesungen, die sich gegen eine Reform sperrten, hat er als Vergleich herangezogen. Bei jeder Veranstaltung wurde eine Selbstevaluation mittels Vor- und Nachtests durchgeführt. Die Messung der Veränderung der moralischen Urteilsfähigkeit erfolgte mit der Online-Version des *Moralisches Urteil-Tests* (Lind 2008) mit dem Programm ITSE (Lind o. J.). Dadurch konnte der Lehrende am Ende jedes Semesters sehen, ob seine Maßnahmen Wirkung zeigten und den gewünschten Kompetenzgewinn erbrachten. Zu Beginn dieser Versuche vor etwa 15 Jahren zeigte sich zunächst keine Wirkung. Aber schon nach wenigen Semestern konnten durch einiges Probieren mit der KMDD und den nach der KMDD umstrukturierten Seminaren sehr große Effekte erzielt werden. Der Lernzuwachs betrug zuletzt im Durchschnitt 13 Punkte pro Semester auf einer Skala von 0 bis 100 (Lind 2009b). Das ist sehr viel. Bei einer günstigen Lernumwelt wurde bislang ein durchschnittlicher Gewinn von 5 bis 8 C-Punkten pro Jahr ermittelt (Schillinger 2006; Lupu 2009). Im Vergleich dazu blieben die Effekte bei den Vorlesungen über alle Jahre unverändert bei nahezu null. In anderen Studien wurde gezeigt, dass sich diese Fähigkeit sogar zurückbildet, wenn die Studierenden keine Gelegenheit zur Verantwortungsübernahme bekommen (Lind 2000; Schillinger 2006; Lupu 2009).

Die Selbstevaluation hat in diesem Projekt also einen doppelten Ertrag erbracht. Erstens konnten mit Hilfe der Daten die Seminare mit der Zeit lerneffektiver gemacht werden. Zweitens ließ sich durch die zusammenfassende Analyse der Daten zeigen, dass, unabhängig von der Veranstaltungsform, eine einzige Dilemma-Diskussion etwa die gleiche Lernwirkung entfaltet wie sonst vier Jahre Studium unter günstigen Lernbedingungen.

Resümee

Wenn wir die pädagogische Praxis verbessern wollen, müssen wir wissen, was wirkt. Aber das „Wissen", das wir für pädagogische Reformen benötigen, ist, wie immer deutlicher wird, in hohem Maß von Verzerrungen bis hin zu Korruption bedroht. Selbstevaluation, so meine These, ist weit besser als herkömmliche Formen der Evaluation gegen die beiden starken *Bias*-Faktoren gefeit, an denen Personen- und Programmevaluation heute vielfach kranken:

(1) Der *Korruptionsdruck* ist bei Personenevaluation und zum Teil auch bei Programmevaluation ein großes Problem. Menschen wehren sich verständlicherweise gegen eine öffentliche Demütigung durch Rankings und gegen Sanktionen, die an Vergleichstests und an die Evaluation ihrer Produkte geknüpft sind. Legale und illegale Schwindeleien sind daher an der Tagesordnung. Bei anonym durchgeführter Selbstevaluation hingegen gibt es keine Demütigungen und Sanktionen und daher auch keinen Korruptionsdruck. Wer sich beim Evaluieren selbst betrügen würde, würde sich widersinnig verhalten. Wer selbst die Methoden und Programme evaluiert, auf die er oder sie sich bei der pädagogischen Arbeit stützt, ist an ihrer pädagogischen Wirksamkeit interessiert, nicht an ihrem Verkaufserfolg oder ihrer Wirkung in der Öffentlichkeit.

(2) *Barrieren gegen die Verbreitung* und *Anwendung* fundierter Erkenntnisse aus Evaluationsstudien können sich nur dort bilden, wo zwischen den Erzeugern und Konsumenten von Evidenz eine große kommunikative Distanz besteht. Bei Selbstevaluation kann es keine *Barrieren* zwischen Erzeugern und Nutzern von Evidenz geben, da beide identisch sind.

Selbstevaluation hat weitere Vorteile gegenüber den bisher dominierenden Formen der Evaluation, die hier nur genannt, aber nicht ausgeführt werden können. Der Verzicht auf *Rankismus* (Fuller/Gerloff 2008) fördert die Zusammenarbeit zwischen allen Beteiligten (Schülern, Lehrern, Eltern, Schulverwaltung etc.). Selbstevaluation erfordert meist weniger Finanzmittel als Personenrankings und kommerzielle Programmevaluation, dafür aber mehr Sachverstand und Unterstützung durch die Wissenschaft. Selbstevaluation kann doppelt genutzt werden, zum einen als Instrument der Qualitätskontrolle, um die eigene pädagogische Arbeit auf hohem Niveau zu halten und Innovationen bewerten zu können; zum anderen als Instrument der Grundlagenforschung, um die Wirksamkeit von bestimmten pädagogischen Maßnahmen oder Methoden systematisch zu überprüfen. Selbstevaluation kann auf einfache Weise mit Meta-Evaluation (Dubs 2005) und Qualitätsmanagement verbunden werden. Schließlich ist regelmäßig durchgeführte, kompetente Selbstevaluation auch eine hochwirksame Form der Fortbildung. Wer als Schüler, Lehrer

oder auch als Kultusminister die Wirkungen seines eigenen Handelns regelmäßig systematisch überprüft und sein Handeln entsprechend ändert, zeigt nachhaltiges Lernen. Selbstevaluation ist auch von Fehlern und Verzerrungen bedroht. Aber diese sind meist unabsichtlich und lassen sich m. E. durch entsprechende Ausbildung in den Griff bekommen. Selbstevaluation erfordert gründliche Kenntnisse und eine gute Ausbildung (Lind o. J.).

Demgegenüber muss dringend diskutiert werden, wie herkömmliche Personen- und Programmevaluation verändert werden kann, um brauchbare, unverzerrte Evidenz für pädagogische Reformen zu liefern (Wuttke 2009). Natürlich ist dies nicht ihre alleinige Funktion. Aber verzerrte Daten taugen auch nicht für die Beurteilung von Schülern und die Einstellung und Beförderung von Lehrern, wie das in den USA mit dem *value-added measurement* (VAM) derzeit in Mode kommt (Rothstein 2011). Vorkehrungen gegen Verzerrungen und Korruption sind sehr kostenintensiv. Es wird daher vor allem darauf ankommen, die Zahl von Personenevaluationen stark zu reduzieren, sie dafür gründlicher zu planen, transparenter zu machen und in der Öffentlichkeit einem *Datencheck* zu unterziehen.

Literatur

Allerup, P. (2007): Identification of group differences using PISA scales – considering effects of inhomogeneous items. In: Hopmann, S. T./Brinek, G./Retzl, M. (Hrsg): PISA zufolge PISA. Wien/Berlin: LIT, S. 175-202.

Amrein, A./Berliner, D. C. (2002): High-stakes testing, uncertainty, and student learning. In: Education Policy Analysis, 10. Jg., H. 18. URL: http://epaa.asu.edu/epaa/v10n18/ (Stand: 20.02.2011).

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

Belley, P./Lochner, L. (2008): The changing role of family income and ability in determining educational achievement. URL: http://economics.uwo.ca/faculty/lochner/papers/thechangingrole.pdf (Stand: 20.02.2011).

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

Bohl, T./Kiper, H., (Hrsg.) (2009): Lernen aus Evaluationsergebnissen – Verbesserungen planen und implementieren. Bad Heilbrunn: Klinkhardt.

Borman, G.D./Slavin, R.E./Cheung, A./Chamberlain, A.M./Madden, N.A/Chambers, B. (2005): Success for All: First-year results from the national randomized field trial. In: Educational Evaluation and Policy Analysis, 27. Jg., H. 1, S. 1-22.

Böttcher, W./Dicke, J. N./Hogreber, N. (Hrsg.) (2011): Evaluation, Bildung und Gesellschaft. Steuerungsinstrumente zwischen Anspruch und Wirklichkeit. Münster: Waxmann.

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

Bracey, G. W. (2005): No child left behind: Where does the money go? Tempe, AZ: Arizona State University, Education Policy Studies Laboratory. URL: http://nepc.colorado.edu/files/EPSL-0506-114-EPRU.pdf (Stand: 20.02.2011).

Bracey, G. W. (2007): The First Time 'Everything Changed'. The 17th Bracey Report on the Condition of Public Education. In: Phi Delta Kappan, 89. Jg., H. 2, S. 119-136.

Brügelmann, H. (2005): Schule verstehen und gestalten. Lengwil: Libelle.

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

Campbell, D. T. (1975): Assessing the Impact of Planned Social Change. In: Lyons, G. M. (Hrsg.): Social Research and Public Policies. The Dartmouth/OECD Conference. Hanover, New Hamshire: Dartmouth College, The Public Affairs Center, S. 3-45.

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

Center for Data-Driven Reform in Education (CDDRE) (2004ff.): Best Evidence Encyclopedia. Empowering Educators with Evidence on Proven Programs. URL: http://www.bestevidence.org/ (Stand: 20.02.2011).

Chamberlin, D./Chamberlin, E. S./Drought, N. E./Scott, W. E. (1942): Did they succeed in college? The follow-up study of the graduates of the thirty schools. New York: Harper & Brothers.

Cohen, P./Kulik, J./Kulik, C. (1982): Educational outcomes of tutoring: A meta-analysis of findings. In: American Educational Research Journal, 19. Jg., S. 237-248.

Cronbach, L. (1960): Essentials of psychological testing. London: Harper.

Cunningham, W. G./Sanzo, T. D. (2002): Is high-stakes testing harming lower socioeconomic status schools? In: NASSP Bulletin, 86. Jg., H. 631, S. 62-75.

Dubs, R. (2005): Metaevaluation – Anforderungen an Schulaufsicht und Schulleitung. In: Bartz, A. u. a. (Hrsg.): PraxisWissen SchulLeitung. Neuwied: Luchterhand, Kapitel 22.15.

Fisher, G. M. (1928): Foreword. In: Hartshorne, H./May, M. A. (Hrsg.): Studies in the nature of character. Vol. I: Studies in deceit, Book one and two. New York: Macmillan , S. V-VII.

Fraser, B. J./Walberg, H. J./Welch, W. W./Hattie, J. A. (1987): Syntheses of educational productivity research. In: International Journal of Educational Research, 11. Jg., S. 145-252.

Fuller, R./Gerloff, P. A. (2008): Dignity for All: How to Create a World Without Rankism. San Francisco: Berret-Koehler Publishers.

Glaser, R. (1963): Instructional technology and the measurement of learning outcomes: some questions. In: American Psychologist, 18. Jg., S. 519-521.

Hagemeister, V. (1999): Was wurde bei TIMSS erhoben? Eine Analyse der empirischen Basis von TIMSS. In: Die Deutsche Schule, 91. Jg., S. 160-177.

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

Heidbrink, H. (2010): Moral judgment competence and political learning. In: Lind, G./Hartmann, H.A./Wakenhut, R. (Hrsg.): Moral judgment and social education. Edison: Transaction Publishing, S. 259-271.

Hopmann, S. T./Brinek, G./Retzl, M. (2007): PISA zufolge PISA. Berlin: LIT-Verlag.

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

Jahnke, J./Meyerhöfer, W. (Hrsg.) (2007): Pisa & Co. Kritik eines Programms. 2., erweiterte Auflage. Hildesheim: Franzbecker.

Klein, H. P. (2010): Die neue Kompetenzorientierung: Exzellenz oder Nivellierung? In: Zeitschrift für Didaktik der Biowissenschaften, 1. Jg., S. 15-26.

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: Heinemann.

Koretz, D. (2009): What educational testing really tells us. Cambridge: Harvard University Press

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

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

Leppert, U. (2010): Ich habe eine Eins! Und Du? Von der Notenlüge zur Praxis einer besseren Lernkultur. München: Libres.

Lind, G. (2000): Moral regression in medical students and their learning environment. In: Revista Brasileira de Educacao Médica, 24. Jg., H. 3, S. 24-33.

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

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

Lind, G. (2009a): Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. In: Bohl, T./Kiper, H. (Hrsg.): Lernen aus Evaluationsergebnissen – Verbesserungen planen und implementieren. Bad Heilbrunn: Klinkhardt, S. 61-79.

Lind, G. (2009b): Favorable learning environments for moral development – A multiple intervention study with nearly 3.000 students in a higher education context. Paper presented at the annual meeting of AERA in San Diego, April 13 - 17, 2009. URL: http://www.uni-konstanz.de/ag-moral/pdf/Lind-2009_Favorable_learning.pdf (Stand: 20.02.2011).

Lind, G. (2009c): Moral ist lehrbar. Ein Handbuch zur moralischen und demokratischen Bildung. München: Oldenbourg.

Lind, G. (o. J.): Verbesserung der Lehre durch wissenschaftliche Evaluation. URL: http://www.uni-konstanz.de/itse (Stand: 20.02.2011).

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

Lohner, S. (2008): Kompetenzmessung in der PISA-Studie. Eine Analyse von Biologie-Aufgaben. Examensarbeit im Fach Pädagogik, Universität Konstanz.

Lupu, I. (2009): Moral, Lernumwelt und Religiosität. Die Entwicklung moralischer Urteilsfähigkeit bei Studierenden in Rumänien in Abhängigkeit von Verantwortungsübernahme und Religiosität. Dissertation, Fachbereich Psychologie, Universität Konstanz. KOPS. URL: http://kops.ub.uni-konstanz.de/volltexte/2009/9586/pdf/Diss_Lupu.pdf (Stand: 20.02.2011).

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

Meyerhöfer, W. (2007): Testfähigkeit – Was ist das? In: Hopmann, S. T./Brinek, G./Retzl, M. (Hrsg.): PISA zufolge PISA. Berlin: LIT-Verlag, S. 57-92.

Neuweg, G. H. (2004): Könnerschaft und implizites Wissen. Zur lehr-lerntheoretischen Bedeutung der Erkenntnis- und Wissentheorie Michael Polanyis. 3. Auflage. Münster: Waxmann.

Nichols, S. L./Glass, G. V./ Berliner, D. C. (2006): High-stakes testing and student achievement: Does accountability pressure increase student learning? In: Education Policy Anlysis Archives, 14. Jg., H. 1. URL: http://epaa.asu.edu/ojs/article/view/72/198 (Stand: 20.02.2011).

Nichols, S. L./Berliner, D. C. (2005): The Inevitable Corruption of Indicators and Educators Through High-Stakes Testing. East Lansing: Great Lakes Centre for Education Research & Practice. URL: http://greatlakescenter.org/docs/early_research/g_l_new_doc/EPSL-0503-101-EPRU.pdf (Stand: 20.02.2011).

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

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

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

Popper, K. (1968): The logic of scientific discovery. London: Hutchinson.

Poynor, L./Wolfe, P. M. (Hrsg. 2005): Marketing fear in America's public schools. The real war on literacy. Mahwah: Lawrence Erlbaum Associates.

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. (1986): Meta-analytic procedures for combining studies with multiple effect sizes. Psychological Bulletin, 99. Jg., S. 400-406.

Rothstein, J. (2011): Review of “Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project”. Boulder, CO: National Education Policy Center. URL: http://nepc.colorado.edu/files/TTR-MET-Rothstein.pdf (Stand: 20.02.2011).

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

Sanders, J. R. (1994): The program evaluation standards. How to assess evaluations of educational programs. 2. Auflage. Thousand Oaks: Sage.

Schillinger, M. (2006): Learning environments and moral development: How university education fosters moral judgment competence in Brazil and two German-speaking countries. Aachen: Shaker.

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

Schoenfeld, A. H. (2006): What doesn’t work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics. In: Educational Researcher, 35. Jg., H. 2, S. 13-21.

Schümer, G. (2006): Zur bildungspolitischen Bedeutung internationaler Schulleistungsstudien. In: Brinkmann, C./Koch, S./Mendius, H. G. (Hrsg.): Wirkungsforschung und Politikberatung – eine Gratwanderung? Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit, S. 245-273.

Schweinhardt, L. J./Weikart, D. P. (1988): Education for young children living in poverty: child-initiated learning or teacher-directed instruction? In: The Elementary School Journal, 89. Jg., H. 2, S. 213-225.

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

Wilson, M. (2005): Constructing measures. An item response modeling approach. Mawah: Erlbaum Associates Publishers.

Wittmann, E. (2010): Offener Brief betreffs „Lernstandserhebung VerA 3 Mathematik /2010" an die Kultusminister der deutschen Bundesländer vom 31.5.2010. URL: http://www.mathematik.uni-dortmund.de/ieem/mathe2000/pdf/vera3/KM%2005-10.pdf (Stand: 20.02.2011).

Wuttke, J. (2006): Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung. In: Jahnke, T./Meyerhöfer, W. (Hrsg.): Pisa & Co. Kritik eines Programms. Hildesheim: Franzbecker, S. 101-154.

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

Wuttke, J. (2009): PISA: Nachträge zu einer nicht geführten Debatte. In: Mitteilungen der Gesellschaft für Didaktik der Mathematik, 87. Jg., August 2009, S. 22-34.