Hans Brügelmann

Scharfe Brillen, wache Augen und ein einfühlsamer Blick

Wie Schulen über die Qualität ihrer Arbeit Rechenschaft ablegen können:

Zur Bedeutung von technischer Präzision und sozialer Kontrolle bei der Evaluation pädagogischer Standards[1]

1 Gutachten für den Verbund der "Blick über den Zaun"-Schulen, vorgetragen auf deren Tagung in Hofgeismar am 13.11.2006. Erscheint im Sommer 2007 in dem Band Beobachten, bewerten, beraten – Verfahren und Werkzeuge für eine “andere” Evaluation Eine stark gekürzte Fassung ist veröffentlicht in: PÄDAGOGIK, H. 3/2007. Vgl. ausführlicher zum Hintergrund der folgenden Ausführungen: Brügelmann (2005a, Kap. 4-5, 46-50, 56-63).

0 Überblick: Zehn Ratschläge für Pioniere der Praxis:

Lasst tausend Blumen blühen –
und bindet euch den inhaltlich und methodisch jeweils passenden Strauß!

Meinen Auftrag habe ich so verstanden, dass ich Hilfe auf zwei Ebenen geben soll:

- Einige erwarten ganz praktisch *methodische Ideen*, wie sie ihre Arbeit pädagogisch angemessen *und* mit vertretbarem Aufwand evaluieren können. Die Netzwerke, die im Schulverbund "Blick über den Zaun"[2] aufgebaut worden sind, haben sich als hilfreich erwiesen, um die Entwicklung *in* den Schulen anzuregen. Nun wächst aber das Interesse vieler Schulen, auch nach *außen*, also gegenüber Dritten, überzeugend Rechenschaft über die Qualität ihrer Arbeit ablegen zu können.
- Andere erhoffen sich *methodologische Argumente*, um eine *pädagogisch* begründete Art der Evaluation offensiv und glaubwürdig vertreten zu können, d. h. ohne die inhaltlichen Ansprüche der Blick-über-den-Zaun-Standards[3] gegenüber den heute oft verkürzten und verkürzenden Rechenschaftsforderungen aufgeben zu müssen.

Mein Auftrag ist nicht ganz leicht zu erfüllen. Vor allem fürchte ich, nicht viel Neues bieten zu können. So haben die "Blick über den Zaun"-Schulen selbst schon viele sinnvolle Aktivitäten entwickelt[4]. Überdies wimmelt es derzeit in pädagogischen Zeitschriften und auf dem Buchmarkt nur so von Evaluationsratgebern[5]. Schließlich sind die Ideen, die *ich* im Folgenden vorstellen kann, zum Teil schon über dreißig Jahre alt[6] – verwunderlich, dass man dieselben Dinge immer wieder sagen muss, auch wenn sie oft fast selbstverständlich erscheinen.

Bei meinem Versuch, Evaluation aus verschiedenen Blickwinkeln zu betrachten und Hilfen für ganz unterschiedliche Situationen zu entwickeln, ist eine Fülle von Ideen entstanden. Sie sind als Anregungen gemeint (Entwicklung eines „Baukastens von Werkzeugen“ hieß es in meinem Auftrag), die Vielfalt der Perspektiven und Optionen kann aber auch leicht erschlagend wirken. Darum sei ganz deutlich gesagt: Hier wird kein Komplettprogramm vorgestellt, das als Ganzes übernommen werden soll oder gar kann. Jede Schule, jede Person muss sich für *ihre* Situation überlegen: Was ist *unser* Problem und welche Ideen helfen *uns* weiter, dieses Problem besser zu lösen als mit den bisher verfügbaren Mitteln? Die Vorschläge sollen also vorhandene Repertoires erweitern, diese aber nicht ersetzen.

Zum Inhalt: In meinem Beitrag möchte ich vor allem deutlich machen, dass Evaluation nur dann produktiv wird, wenn sie *funktionsspezifisch* ausgelegt wird und wenn - diesen Funktionen entsprechend - *unterschiedliche* Stile, Rollenverteilungen, Methoden und Informationsquellen

---

[2] ➔ www.BlickUeberDenZaun.de [Abruf: 23.1.2007]

[3] Vgl. Groeben u. a. (2005).

[4] Vgl. etwa: Tillmann/ Wischer (1999) und die Beiträge von Seydel unter ➔ www.blickueberdenzaun.de/10Evaluation.html [Abruf: 20.1.2007].

[5] Empfehlenswert vor allem: Schratz u. a. (2000); Böttcher u. a. (2006).

[6] Dieser Beitrag greift mehrfach zurück auf frühere Arbeiten, die als (zumindest teilweise) Pre- bzw. Reprints verfügbar sind zum Download unter ➔

SAFARI-1974 http://www.agprim.uni-siegen.de/printbrue/brue.checks_and_balances.pdf
EVICIEL-1976 http://www.agprim.uni-siegen.de/printbrue/curriculumevaluation.pdf
OECD-1980 http://www.agprim.uni-siegen.de/printbrue/brue.oecd.evaluation.pdf
EVICIEL-1983 http://www.agprim.uni-siegen.de/printbrue/naturalistischeeval.pdf
STANDARDS-2005 http://www.agprim.uni-siegen.de/printbrue/brue.05i.standards.pdf

genutzt werden. Durch solche grundsätzlichen Überlegungen wird meine Darstellung unvermeidlich anstrengender, als wenn ich nur ein Kochbuch mit methodischen Rezepten bieten würde. Ich habe aber versucht, meine stärker theoretischen Überlegungen möglichst durch konkrete Beispiele plausibel zu machen.

Meine zentrale Botschaft: Wer die Qualität von Bildung erfassen will, kann – und muss – auf sehr verschiedene Rechenschaftsformen und Evaluationsverfahren zurückgreifen. Aktuell leiden wir unter einer Monokultur fachbezogener Leistungstests, die der Breite der Evaluationsanforderungen nicht gerecht wird.

Mit der Durchsetzung der internationalen Leistungsvergleiche sind Hoffnungen geweckt worden, die standardisierte Leistungsmessungen (allein) nicht erfüllen können. Nicht nur in den Kultusverwaltungen glauben viele, mit landesweiten Testprogrammen eine Lösung der Rechenschaftsprobleme auch auf Schul- und Lehrer-Ebene und sogar ein verlässliches Mittel der Lerndiagnose gefunden zu haben. Aber was Politik und Verwaltung hilft - statistische Aussagen zu wenigen Variablen über viele Fälle, Verdichtung in Mittelwerten und Gruppenaussagen, Betonung von Konstanten –, ist für Schulen und LehrerInnen nur sehr begrenzt hilfreich. Für sie ist wichtig, Einzelfälle in ihrem Facettenreichtum zu erfassen und ihre Entwicklung im pädagogischen Kontext zu interpretieren.

Standardisierte Tests erwecken zudem einen Anschein von Objektivität und Präzision, der methodologisch nicht zu rechtfertigen ist[7]. Testdaten sprechen nicht für sich. Ihre Deutung ist immer personabhängig. Technisch-methodisch lässt sich diese Subjektivität nicht kontrollieren. Sowohl die Standardisierung von Konzepten in Form von Indikatoren (z. B. Testaufgaben) als auch deren Bearbeitung durch die ProbandInnen sind Interpretationsakte – und damit auch die Deutung der Ergebnisse am Ende (von den vielfältigen Optionen bei der Auswahl statistischer Verfahren und bei der Bewertung der gewonnenen Kennwerte ganz zu schweigen ...).

0.1 Fazit: Evaluation ist mehr als das Messen von Unterrichtseffekten mit Hilfe von Leistungstests. Produktiv wird sie nur, wenn ihre Formen und Methoden aufgabenbezogen und kontextgerecht ausgelegt werden.

Was kann man konkret tun?

- Man sollte sich nicht jedem beliebigen Evaluationsverfahren unterwerfen, sondern seine Passung auf Gegenstand und Fragestellung prüfen.
- Ebenso wenig sollte man sich aber auch externen Rechenschaftsforderungen nur deshalb verweigern, weil sie oft unzulänglich instrumentiert werden.
- Deshalb: Fordern und nutzen Sie Evaluationsverfahren, die Ihrer Pädagogik inhaltlich und methodisch gerecht werden!

[7] Ausführlicher dazu unten (4) bis (6).

1 Wider ein grundlegendes Missverständnis:
Evaluation ist kein (nur) technisches Problem, sondern ein Politikum!

Als ich Ideen für meinen Vortrag gesammelt habe, hatte ich zunächst folgenden Titel im Sinn:

Wer hat Angst vor Eva-Luation?
Ein „Erste-Hilfe-Kurs“ für den Alltag in Reformschulen

Mein Mentor Wolfgang Harder hat mich zu Recht davon abgebracht, diesen Titel als Motto über den Beitrag als ganzen zu setzen. Er hat mich darauf hingewiesen, dass die "Blick über den Zaun"-Schulen sich schon lange auf Evaluation eingelassen haben. Insofern gehe die Titel-Frage an ihrer Situation vorbei. Mir hat auch eingeleuchtet, dass dies als falsches Signal wahrgenommen werden könnte. Aber die inhaltlichen Probleme, die ich mit diesem Titel im Sinn hatte, liegen auf einer anderen Ebene. Sie von Anfang deutlich zu machen, ist mir auch weiterhin wichtig.

Ich selbst habe vor mehr als dreißig Jahren Evaluation „on the job“ gelernt - im SAFARI-Projekt von Barry MacDonald und im FORD-Teaching Project von John Elliott am Centre for Applied Research in Education in Norwich in England. Damit hatte ich das Glück, bei Praktikern lernen zu dürfen. Und zwar bei Praktikern, denen Methoden zwar wichtig waren, die deren Anwendung aber in einem weiteren Kontext sahen:

- Sie haben Evaluation nicht auf das Testen von Leistungen reduziert, sondern ein sehr breites Spektrum an Fragen aufgenommen und deshalb *vielfältige Instrumente* genutzt, darunter vorrangig qualitative Methoden, ja Alltagsverfahren wie Protokolle von Beobachtungen, Gespräche mit Menschen, Fotografieren von Situationen.
- Sie waren nicht nur an technisch-methodischen Fragen von Evaluation interessiert, sondern vor allem an ihrem sozialen Kontext und damit an normativen Fragen wie den *ethics and politics of evaluation*.

Evaluation wird dagegen heute - im Kontext von PISA & Co - vor allem als methodisch-technisches Problem diskutiert. Zentrale Themen sind Standards der Testentwicklung, Kriterien der Stichprobenziehung, Angemessenheit von statistischen Verfahren wie Faktoren- bzw. Mehrebenen-Analysen usw. Bei Evaluation geht es aber auch um Macht, um Angst und um Glaubwürdigkeit. Evaluation ist deshalb auch ein Beziehungsproblem, ihre Gestaltung verlangt soziologischen Durchblick und psychologische Sensibilität.

Wer diesen Kontext von Evaluation nicht bedenkt, wird mit noch so ausgefeilten Designs und noch so präzisen Messinstrumenten scheitern. Entsprechend hat das US-amerikanische “Joint Committee on Standards for Educational Evaluation” seine Anforderungen an eine “gute” Evaluation formuliert (Sander 2000). Die Standards umfassen die vier Bereiche:

- Genauigkeit, z. B. Validität; Objektivität; Verlässlichkeit
- Nützlichkeit, z. B. Zielgruppen; Fragen; Kriterien
- Machbarkeit, z. B. Akzeptanz; Rechtzeitigkeit; Aufwand
- Integrität, z. B. Unparteilichkeit; Datenschutz; Fairness

Vor allem die deutsche Diskussion beschränkt sich weitgehend auf den ersten Aspekt der methodisch-technischen Präzision. Deshalb ist es wichtig, zwei Probleme von Evaluation im Blick zu halten:

- die Verletzlichkeit der betroffenen Personen und damit ihre Sorge, das Gesicht zu verlieren;
- das politische Störpotenzial von Evaluation, deren Ergebnisse immer Einfluss haben auf die Balance von Macht im untersuchten Feld, denn bei Kontroversen stützen sie bestimmte Positionen und schwächen andere.
- Diese Probleme erhalten ein unterschiedliches Gewicht - je nachdem, vom wem die Evaluation verantwortet wird. Im Bildungswesen werden Evaluationsaufgaben schon heute von verschiedenen Rollenträgern wahrgenommen: von *Experten,* z. B. von Bildungsforschern wie bei PISA und VERA; von*Vorgesetzten,* z. B. von der Schulaufsicht oder Schulleitung; von den *Beteiligten* selbst, z. B. von LehrerInnen bzw. SchülerInnen im Unterricht.

Die Frage der Machtverteilung zwischen verschiedenen an der Schule Beteiligten nehme ich später noch einmal auf (➔ unten S. 28 ff., 46 ff.). Hier nur so viel: Inhaltlich geht es bei Evaluationen immer um die Klärung von Differenzen – und damit auf der Beziehungs- bzw. auf der institutionellen Ebene um eine Intervention in Konflikten. Insofern sind die Werte, die der Evaluator vertritt, von zentraler Bedeutung für die Funktion und den Ertrag seiner Arbeit. Konkret: Was für ein Selbstbild vertritt er im Verhältnis zum Auftraggeber und in den Beziehungen zu anderen Beteiligten?[8]

Für deren Ausgestaltung habe ich aus der Arbeit am CARE zwei Leitideen mitgenommen, die bei der Planung von Evaluationen generell beachtet werden sollten:

- „People own the facts of their lifes“[9] –
  d. h. Vertraulichkeit und Vertrauen sind forschungsethisch hohe Werte und zugleich sind sie Bedingung für eine effektive Durchführung von Evaluation.
- „Wissen gibt Macht“,
  und damit ist Evaluation in demokratischen Institutionen und Kontexten anders zu organisieren als in einem bürokratischen oder in einem expertokratischen Umfeld.

Es spielt insofern eine große Rolle für den Erfolg von Evaluationen, in welcher Konstellation die Verantwortung für ihre Durchführung organisiert ist (➔ unten S. 28 ff., 49 ff.).

[8] Z. B. einen Autoritätsanspruch dank Kompetenzvorsprung als „Experte“(„die Wahrheit vermitteln“); eine Aufklärungsidee als „critical friend „ („Informationen/ Einschätzungen austauschen“) oder ein Hilfsangebot als Unterstützer („Probleme lösen helfen“).

[9] Heute würde ich vorsichtiger sagen: „at least *their views* on their lifes“.

1.1 Fazit: Wie im Unterricht ist auch bei der Evaluation die Beziehungsebene mindestens ebenso wichtig wie die Inhalte und Methoden.

Was kann man konkret tun?

- VertreterInnen aller durch die Evaluation betroffenen Gruppen sollten an ihrer Planung und Durchführung beteiligt werden (z. B. durch Bildung einer „Steuergruppe“).
- Gemeinsam mit den EvaluatorInnen sind klare und verbindliche Absprachen zu treffen über:
  - Auftrag/ Funktion,
  - Fragestellungen,
  - Form der Berichterstattung.
- Die Betroffenen (oder VertreterInnen der wichtigsten Gruppen) sollten ein Kommentierungsrecht zu den Daten und ihrer Deutung haben: Es gibt nicht „die Wahrheit“.

2 Zweck der Evaluation:

Mängel aufdecken, Potenziale ausweisen, Ursachen für Probleme finden?

Eine Evaluation kann unterschiedliche Funktionen haben:

- Zum einen kann sie helfen, im Alltag nicht wahrgenommene Probleme aufzudecken, damit die Schule sich ihnen in der eigenen Arbeit stellen kann. Dann muss das Evaluations-Setting zur Offenlegung der eigenen Schwächen ermutigen und einen geschützten Raum schaffen, in dem Kritik angenommen werden kann.
- Oder sie verspricht Aufklärung für Außenstehende, damit sie sich begründet zwischen Alternativen entscheiden können - sei es für Eltern bei der Wahl einer Schule, sei es für einen Träger bei der Bewilligung von Mitteln. Dann geht es darum, die Besonderheiten der eigenen Institution, eines konkreten Programms oder einer Maßnahme herauszuarbeiten. Den Auftraggebern der Evaluation kommt es in diesem Fall vor allem darauf an, ihre Stärken darzustellen – aber in einer glaubwürdigen Form, z. B. durch die Akzeptanz einer externen Evaluation (➔ unten 28 ff.).

In diesen unterschiedlichen Funktionen deutet sich ein grundlegender Interessenkonflikt an. Also muss man Prioritäten setzen. Dazu stelle ich im Folgenden drei Modelle vor, die sich als Grundmuster für die Planung einer Evaluation anbieten.

2.1 Zurzeit begegnen wir in den Schulen vor allem Beispielen für das *Inspektionsmodell*.

Seine Fragerichtung: Wo sind die Stärken und wo die Schwächen einer Schule? Neben einer allgemeinen Rückmeldung ist das primäre Ziel die *Entdeckung von Problemen*, um Schulentwicklung in Gang zu bringen. Dafür muss die Evaluation *offen* angelegt sein.

Die "Blick über den Zaun"-Standards sind dafür eine gute Suchhilfe – interessant sind aber auch andere Brillen[10], die man sich über Repräsentanten verschiedener Einrichtungen dazuholen kann, wie ich es unten im juridischen Modell vorschlagen werde. Bei Differenzen in den Einschätzungen ist dann wichtig zu prüfen, ob ihr Grund in unterschiedlichen *Kriterien* oder eher in unterschiedlichen *Wahrnehmungen* liegt.

[10] Die „Pädagogischen Entwicklungsbilanzen“ der DIPF-Projekts „Schulentwicklung, Qualitätssicherung und Lehrerarbeit“ können ebenfalls ein hilfreiches Instrument sein, zumindest Anregungen für die Entwicklung eines eigenen Verfahrens liefern:

- Der Fragebogen sondiert eine breite Palette an möglichen Problemfeldern („Schrotschuss“ oder „Omnibus“-Prinzip: alles/s wird mitgenommen)
- Zum Teil werden Ansprüche und Selbsteinschätzungen von deren Umsetzung kontrastiv erfragt.
- Es werden verschiedene Perspektiven (LehrerInnen und SchülerInnen) einbezogen.
- Für hessische Schulen ist die Auswertung kostenfrei.

Für den allgemeinen Teil gibt es Vergleichwerte aus größeren Stichproben – zusätzlich sind daneben schulspezifische Befragungen möglich.

Es gibt inzwischen eine Fülle von Fragekatalogen und Kriterienrastern - und keines, das ich als „das beste“ empfehlen könnte. Mir scheint vielmehr wichtig, dass eine jede Schule über ihre Schwerpunkte neu nachdenkt und entsprechend die Fragen und Kriterien selbst formuliert (wobei bereits erarbeitete Schemata durchaus hilfreich sein können).

(I) analog zum **„Gesundheits-Check“**: Gibt es allgemeine Hinweise auf Probleme?	
Zielgruppe	- Schulleitung - Lehrerkollegium - SchülerInnen/ Eltern
Fragen/ Kriterien	Bezogen auf uns wichtige Ziele und Prozessstandards: - Wo liegen besondere Stärken/ Schwächen bei der Umsetzung? - Welche besonderen Potenziale könnten genutzt werden?
Methodenwahl	- Normdiskussion - Prozessbeobachtung („critical incidents“) - offene & strukturierte Befragung (z. B. Ehemalige) - Erfolgsmessung (informelle Leistungsproben)
Rollenverteilung	- Kritische Freunde in Pro- und Contra-Rolle
Präsentationsform	- schulöffentliches Hearing - visuelle Collage
Folgerungen	- Schul- und Unterrichtsentwicklung - ggf. Auftrag für Expertengutachten (s. III)

2.2 Evaluation kann aber auch auf die Akkreditierung einer Einrichtung angelegt sein (wie für technische Geräte beim TÜV).

Das Ministerium für Schule und Weiterbildung in NRW hat zum Beispiel ein „Gütesiegel Individuelle Förderung“ entwickelt, für das sich Schulen bewerben können. Solche Konstellationen sind für EvaluatorInnen besonders konfliktreich: Ziel der begutachteten Einrichtung ist das *Verstecken von Problemen, Ziel des Auftraggebers ist deren Aufdeckung***.** Für die Planung einer solchen Evaluation bietet sich folgende Checkliste an:

(II) analog zum **„Gesundheitszeugnis“**: Sind vorgegebene Anforderungen erfüllt?	
Zielgruppe	- Träger - „Kunden“ - Abnehmer
Fragen/ Kriterien	- Welche (externen) Ziele und Prozessstandards werden geteilt? - Wie weit werden die Standards tatsächlich umgesetzt? - Wie gut werden die Ziele erreicht? - Mit welchem Aufwand werden die Ziele erreicht?
Methodenwahl	- Kollegiums-Befragung zu Zielen und Standards - Strukturierte Beobachtung und Befragung von SchülerInnen - standardisierte Leistungsmessung - Kosten-Nutzen-Rechnung (finanziell und immateriell)
Rollenverteilung	- intern: Bereitstellung von Informationen - extern: Erhebung von Daten und ihre Bewertung
Präsentationsform	- extern: Bericht, ggf. Zertifizierung - intern: Recht auf Kommentierung vor Publikation
Folgerungen	- extern: Träger; „Kunden“; Abnehmer - intern: Schul- und Unterrichtsentwicklung

In Akkreditierungsverfahren kann der Verbund der Reformschulen die "Blick über den Zaun"-Standards offensiv als Kriterien für die Bewertung einbringen, evtl. in Form eines eigenen Gütesiegels – aber man sollte auch hier nicht übersehen, dass den „*audiences*“ vielleicht andere Kriterien wichtiger sind.

2.3 Evaluation kann drittens auch in Gang gesetzt werden, wenn der Alltag für irgendeinen Beteiligten nicht (mehr) selbstverständlich ist, so dass **spezifische Probleme und ihre Ursachen** untersucht werden sollen:

- KollegInnen sind unsicher, ob bestimmte didaktisch-methodische Maßnahmen erfolgreich sind;
- Eltern wollen wissen, wie zufrieden andere Eltern mit der Schule gewesen sind;
- der Träger will etwas über Auslastung und Kosten-Nutzen wissen.

(III) analog zur **„Facharztdiagnose“**: Was könnte die Ursache für bekannte Probleme sein?	
Zielgruppe	- Schulleitung - einzelnen LehrerInnen
Fragen/ Kriterien	- Unter welchen Bedingungen werden bestimmte Standards unzureichend umgesetzt? - Unter welchen Bedingungen werden bestimmte Ziele nicht erreicht?
Methodenwahl	- interne/ externe Kontrollgruppenvergleiche - Instrumente je nach Problem
Rollenverteilung	- intern: Definition von Fragen (Auftrag) - extern: Dienstleistung durch - SpezialistInnen oder - KollegInnen
Präsentationsform	- Expertengutachten - Außenblick
Folgerungen	- interne Entscheidungen

Eine solche Ursachenforschung ist schwierig. Man sollte sich darüber klar sein, dass die Untersuchung von Einzelfällen (also einer Schule oder gar einzelner Situationen in ihr) nur Hypothesen zur Erklärung von Problemen erbringen kann. Zu Ihrem Trost: Das ist auch in großen Stichproben meist nicht anders. In der Regel kann man weder Versuchs- und Kontrollgruppe nach Zufall ziehen, noch hat man die Zeit für einen Längsschnitt, in dem man nicht nur Korrelationen, sondern auch Kausalität feststellen kann. Langfristige Wirkungen wie Berufs- und Lebenserfolg werden erst recht kaum mehr erfasst, da solche Untersuchungen zu teuer und die intervenierenden Bedingungen schwer zu kontrollieren sind. Dank dieser Probleme kommen dann so verwunderliche Aussagen wie die über die Irrelevanz der Klassengröße für guten Unterricht zustande, weil man z. B. nicht bedenkt/ kontrolliert, dass Schulen dazu neigen, die Lerngruppen bei schwieriger Zusammensetzung kleiner zu machen – so dass die schwächere Leistung oft Ursache, nicht Folge der geringen Schülerzahl ist.
Diese dritte Stufe einer *fokussierten* Evaluation wird noch viel zu selten praktiziert. Wenn die Welle der allgemeinen Inspektionen über die Schule hinweggerollt ist und die Testprogramme

auch für die Politiker ihren ersten Charme verloren haben, werden wir uns auf diesen Typ konzentrieren müssen. Denn er liefert am ehesten Informationen für die Schul- und Unterrichtsentwicklung.

2.4 Fazit: Evaluationen können sehr unterschiedliche Funktionen haben, die sich nicht auf die gleiche Weise umsetzen lassen und sogar wechselseitig stören können.

Was kann man konkret tun?

- Es sind klare Prioritäten zu setzen, welchen Zweck die Evaluation primär haben soll (Vermeidung eines *function overload*).
- Die Erwartungen und Aktivitäten von
  - EvaluatorInnen
  - Betroffene und
  - AdressatInnen

  sind auf diese Fokussierung abzustimmen, um die Gefahr von Missverständnissen und Enttäuschungen gering zu halten.

3 Zielgruppen der Evaluation: Für welche Entscheidungsträger wird sie konkret gemacht?

Eine Evaluation muss sich auf konkrete Entscheidungssituationen beziehen – das unterscheidet sie von Forschung. Forschung zielt auf Erkenntnis – Evaluation auf Handeln. Erstere steigert die Komplexität der Information, letztere muss sie problemorientiert reduzieren. Konkret geht es darum, was Beteiligte wissen müssen, um unter gegebenen Randbedingungen vernünftig entscheiden zu können:

- Was für Fragen haben Eltern, wenn sie überlegen, ihr Kind an Ihrer Schule anzumelden?
- Was für Fragen hat der Träger, wenn es um die weitere Finanzierung Ihrer Schule geht?

Zu bedenken sind verschiedene Entscheidungsebenen und entsprechend unterschiedliche Zielgruppen für die Evaluation. Betrachtet man konkrete Verantwortlichkeiten, so haben die wichtigsten Rechenschaftspflichten:

- Politiker über die Leistungsfähigkeit des Bildungssystems gegenüber der Öffentlichkeit;
- Lehrer und MitarbeiterInnen über die Leistungen der Schule gegenüber Träger und Eltern;
- LehrerInnen über die Ziele und den Erfolg ihres Unterrichts gegenüber den Betroffenen und den Vorgesetzten;
- SchülerInnen als Einzelpersonen gegenüber sich selbst und gegenüber Eltern bzw. LehrerInnen über Fortschritte und Schwierigkeiten bei ihrer individuellen Arbeit.

In diesen Beziehungen stellen sich unterschiedliche Fragen[11], die jeweils auch unterschiedliche Formen der Evaluation erfordern (➔ S. 9 ff., 32 ff.). So können PISA & Co auf keinen Fall Modell für die Evaluation auf Schul- oder gar Schüler-Ebene sein. Solche Erhebungen sind zwar auf der Systemebene nützlich – aber auch dort schon problematisch:

- Einschränkung möglicher Leistungsaspekte durch Standardisierungsanspruch und Zeitknappheit;
- entsprechend unterschiedliche Ergebnisse je nach Fokus und Form des konkreten Tests[12];
- unterschiedliche Bedeutung „derselben" Aufgaben (Textlänge!) und Gruppenkategorien (MigrantInnen!) in verschiedenen Kulturen.

Die Probleme verschärfen sich bei der Evaluation von einzelnen Institutionen oder Personen:

- Ergebnisse punktueller Erhebungen sind zu unzuverlässig;
- Durchschnitts- und Gruppenaussagen vernachlässigen die Streuung von Werten und die Kontextabhängigkeit von Zusammenhängen;

---

[11] Aufgrund der unterschiedlichen Art der Fragen lassen sich in drei Grundtypen der Evaluation abgrenzen:

- die auf allgemeine Aussagen über Erfolg zielende Evaluation, wie sie beispielsweise in der Bildungspolitik benötigt wird – etwa für die Bewertung der Leistungsfähigkeit eines Systems bezogen auf seine Ansprüche („an und für sich gut"). Beispiel: Niveaus der „Literacy" im deutschen Schulwesen bei PISA;
- die vergleichende Einschätzung des Potenzials von Alternativen zur Entscheidung über die allgemeine Einführung eines neuen Programms/ Konzepts („besser als"). Beispiel: Anschaffung des Programms „Lernserver" in einem Bundesland statt Förderung der Fortbildung von LehrerInnen.
- Die auf situationsbezogene Einzelfallentscheidungen zielende Evaluation, die die Passung des Potenzials einer Maßnahme auf einen konkreten Kontext einschätzen soll („gut für hier"). Beispiel: Wahl eines Films für die Unterrichtseinheit in einer Klasse.

[12] Vgl. etwa die häufigen Rangplatzdifferenzen einzelner SchülerInnen in verschiedenen Mathematiktests bei Ratzka (2003).

- auf Einzelfallebene ist der Messfehler, der sich in großen Stichproben ausgleicht, zu groß.

Die Schlüsselfrage: „*who ist the audience*?“ ist deshalb sehr früh zu klären. Meine Vermutung: Ihre Zielgruppen in den Schulen sind vor allem

- Sie selbst, also die unmittelbar beteiligten SchülerInnen, LehrerInnen, Leitungen
- die Eltern
- der Träger und
- die Schulaufsicht.

Besonders ansprechen möchte ich eine Teilgruppe, die in Evaluationen bisher wenig Beachtung gefunden hat.

Exkurs[13] I: Zur Rolle von SchülerInnen bei der Evaluation ➔ 46 ff.

Meine Frage: Wie ernsthaft beteiligen wir eigentlich SchülerInnen (und andere Gruppen, die nicht zur professionellen *in-group* zählen) an unseren Evaluationen – und bemühen wir uns um Formen, die ihnen ermöglichen, ihre Rechte tatsächlich wahrzunehmen?[14]

3.1 Fazit: Pädagogisch relevante Entscheidungen fallen auf verschiedenen Ebenen: Die Handlungsoptionen und Entscheidungskriterien sind jeweils andere – und auf diese konkreten Anforderungen muss sich eine Evaluation einstellen.

Was kann man konkret tun?

- Es ist frühzeitig zu klären, welche Fragen welcher Gruppen durch die Evaluation vorrangig bearbeitet werden sollen, z.B. über eine Befragung zu deren Informationsbedürfnissen vorweg.
- Zusätzlich ist zu bedenken (ggf. stellvertretend für die Betroffenen), welche Handlungsoptionen tatsächlich bestehen und welche Kriterien bei den Entscheidungen relevant sein werden.

Eng verbunden mit der Frage nach den Zielgruppen der Evaluation ist deren konkrete Funktion:

[13] S. Anhang; die Lektüre der Exkurse ist zum Verständnis des fortlaufenden Textes nicht erforderlich, sondern dient nur zur Vertiefung der Argumentation

[14] Vgl. zu konkreten Methoden, die eine Beteiligung von Kindern schon im Grundschulalter ermöglichen, die Hinweise in Cox u. a. (2006).

4 Aufgaben der Evaluation: Beschreibung, Erklärung oder Bewertung pädagogischer Ereignisse?

Evaluation heißt *Bewertung*. Aber eine faire Bewertung setzt eine genaue *Beschreibung* voraus. Vor allem psychologische Studien beschränken sich oft darauf Leistungen von SchülerInnen *vor* und *nach* verschiedenen Interventionen zu vergleichen. Unterschiede in den Zuwächsen werden den verglichenen Konzepten zugerechnet – ohne zu prüfen, in welchem Umfang diese tatsächlich umgesetzt worden sind, wie die Formen der Umsetzung je nach Bedingungen variieren und welche Bedeutung solche situationsbezogenen Auslegungen für die Wirksamkeit der Intervention haben. Dabei besagt eine Grundregel (nicht nur) der Lehr-Lern-Forschung: Die Differenz zwischen den Mittelwerten von zwei Gruppen ist geringer als die Streuung innerhalb dieser Gruppen.

Unterschätzt werden aber auch die *Anforderungen*, pädagogische Ereignisse genau und anschaulich zu beschreiben[15]. Dabei kann es schon ein erster wichtiger Schritt der Evaluation sein, den Beteiligten nur einen Spiegel vorzuhalten, wie ihr Alltag aussieht. Wie leicht erzeugt die Routine blinde Flecke: Probleme werden verdrängt – aber auch Stärken nicht mehr wahrgenommen. Die Bedeutung und der Anspruch einer differenzierten Deskription werden insofern oft unterschätzt.

Oft werden Bewertungen aber auch erst dann praktisch bedeutsam, wenn sie auf *Erklärungen* zurückgreifen, zumindest Hypothesen für bedeutsame Bedingungen anbieten können. Dies wiederum erfordert ein anderes Vorgehen, zu dessen Schwierigkeiten ich oben (S. 12) schon einige Anmerkungen gemacht habe.

Je nach den Informationsbedürfnissen der Zielgruppen sollte eine Evaluation stärker den einen oder den anderen Aspekt in den Blick nehmen[16]. JuristInnen machen eine Unterscheidung, die an dieser Stelle hilfreich sein kann. In streitigen Verfahren haben Richter zwei Fragen zu klären:

- Was ist *wirklich* (gewesen)? ➔ „*Tatbestand*“ = Feststellung des Sachverhalts
  Überprüfung von Wahrnehmungen durch Zeugen und Sachverständige
- Ist das *richtig* (gewesen)? ➔ „*Urteil*“ = Bewertung des Sachverhalts
  Auslegung von Normen durch Anwälte und Richter.

Diese beiden Fragen sind auch die zentralen Probleme in der Evaluation:

- *Tatbestand*: Macht der Lehrer im Unterricht, was er den Eltern erzählt oder was er ins Klassenbuch geschrieben hat, und was können seine SchülerInnen tatsächlich?
- *Urteil*: Was ist „gut“ an seinem Unterricht bzw. wie bewerten wir diese Leistung“?

---

[15] Vielfältige Anregungen finden sich in den beiden Bänden zu Funktionen und Formen von Fallstudien von Fischer (1982; 1983).

[16] Daraus ergeben sich unterschiedliche **Typen von Fragen**,

- wenn man sich über Ziele uneins ist,
- wenn man wissen will, ob der Betrieb tatsächlich wie geplant funktioniert,
- wenn man zweifelt, dass die Ziele wirklich erreicht werden oder
- wenn man den Aufwand für den erzielten Ertrag für zu hoch hält.

Entsprechend unterschiedlich sind auch die **Kriterien** sind, die jeweils zur Bewertung heranzuziehen sind:

- Legitimität der Ziele und Prinzipien,
- Konzepttreue der Umsetzung,
- Wirksamkeit der Maßnahmen und
- Sparsamkeit des Aufwands.

Beide Ebenen sind wichtig für die Evaluation. Aber Entscheidungen über die Interpretation von Normen werden oft einfach vorausgesetzt oder beiläufig gefällt. Dabei erfordert ihre Evaluation einen anderen Zugang.

Es macht z. B. wenig Sinn im Unterricht zu untersuchen, wie oft Gruppenarbeit gemacht wird, wenn die betreffenden LehrerInnen schon den Sinn von Gruppenarbeit bezweifeln. Erst müssen normative Differenzen geklärt werden. Durch die Evaluation kann durchaus in Frage gestellt werden, ob die Ziele und Prinzipien einer Institution bzw. Person überzeugend begründet werden. Aber solche Fragen sind nicht empirisch, sondern argumentativ zu klären.

Empirische Untersuchungen sind durch verschiedene Brillen möglich. Also muss man zuerst mit den Betroffenen klären, welche Brille für sie Sinn macht und anschließend den Unterricht auch durch *diese* Brille betrachten – sonst vertut man Zeit und knappe Ressourcen. Kommt es zu keiner Verständigung, ist der normative Dissens das entscheidende Evaluationsergebnis.

Ein Beispiel: In unserer Studie zur Verbreitung verschiedener Formen offenen Unterrichts haben wir bewusst drei Kontraste untersucht[17]:

- Wir haben die LehrerInnen gefragt, wie wichtig ihnen bestimmte Aktivitäten sind, um ihre Prioritäten mit unseren Vorstellungen von offenem Unterricht abzugleichen (normative Differenzen: INNEN- vs. AUSSEN-Soll).
- Wir haben sie zusätzlich gebeten einzuschätzen, wie oft sie diese Aktivitäten tatsächlich praktizieren (Differenzen zwischen Empirie und normativem Anspruch: IST vs. SOLL).
- Schließlich haben wir LehramtsanwärterInnen, die in diesen Klassen gearbeitet haben, , um *ihre* Einschätzungen gebeten, wie häufig die Aktivitäten in den betreffenden praktiziert wurden (empirische Differenzen: AUSSEN- vs. INNEN-Sicht).

Die Ergebnisse zeigen, dass es – aus der Sicht eines Vertreters von offenem Unterricht - drei verschiedene Probleme gibt, die entsprechend auch mit ganz unterschiedlichen Mitteln zu bearbeiten sind:

- fehlende Zustimmung zu Kernprinzipien offenen Unterrichts
  ➔ hier ist Überzeugungsarbeit nötig;
- unzureichende Umsetzung selbst der eigenen Ansprüche
  ➔ diese LehrerInnen brauchen Implementationshilfen;
- Beschönigung der eigenen Praxis
  ➔ dann ist das Verhältnis von eigenen Ansprüchen und externen Erwartungen zu klären.

4.1 Fazit: In pädagogischen Situationen geht es um empirische *und* normative Fragen – die jeweils andere Formen der Beurteilung erfordern.

Was kann man konkret tun?

- Evaluation hat sich auf empirische *und* auf normative Fragen zu beziehen – aber beide sorgfältig zu trennen.
- Ansprüche sind argumentativ klären und erfordern diskursive Formen der Evaluation.
- Erst anschließend lassen sich ihre Umsetzung in der Praxis und - auf deren konkrete Formen bezogen - ihre Wirkungen prüfen.

[17] Vgl. Brügelmann (2000);. ähnlich die bereits oben (Anm. 10) zitierte SEL-Studie des DIPF zu „Pädagogischen EntwicklungsBilanzen“.

5 Verfahren der Evaluation: Standardisierte oder interpretative Zugänge?

Im Gefolge von PISA und Co. haben Tests und andere standardisierte Verfahren ein hohes Prestige gewonnen. Sie gelten als objektiv, ihre Daten als sicher. Demgegenüber haben die Urteile von Personen an Kredit verloren. Das gilt für die Bewertung von Schüllerleistungen durch LehrerInnen, aber auch für die Beurteilung von LehrerInnen durch Beamte der Schulaufsicht. Der Streit um den Status von standardisierten vs. interpretativen Verfahren fokussiert ein Kernproblem der Evaluation: In der Sozialforschung unterscheiden wir hoch- und niedrig-inferente, d. h. mehr oder weniger interpretations-offene Methoden bzw. Instrumente der Untersuchung. Nehmen wir die Standards der "Blick über den Zaun"-Schulen, dann sind für den Evaluator anspruchsvoll formulierte Standards wie

> Die Lehrenden bedienen individuell verschiedene Lernmöglichkeiten und –wege.

und

> Für Lernprobleme finden sie [die SchülerInnen]geeignete Hilfen.

hoch-inferente Beobachtungskriterien, also in besonderem Maße interpretationsbedürftig, denn sie erfordern Einschätzungen der „individuellen Lernmöglichkeiten“ und der „Eignung“ von Lernhilfen, die unter PädagogInnen und PsychologInnen Glaubenskriege auslösen können.

Dagegen gelten die Standards

> Die Schülerinnen und Schüler werden täglich begrüßt und verabschiedet, wenn sie in die Schule kommen bzw. die Schule verlassen.

und

> Jede Gruppe wird von zwei Erwachsenen betreut.

als *niedrig-inferente* Kriterien, d. h. als interpretationsarm, weil sie sich an Oberflächenmerkmalen festmachen lassen, bei denen verschiedene BeobachterInnen in der Regel zu gleichen Zuordnungen kommen.
Methodisch-technisch kontrollierbar sind die Datenerhebungen also nur im zweiten Fall. Und damit objektiv interpretierbar - behaupten zumindest die Verfechter des Standardisierungsparadigmas. Aber sind sie das wirklich?
Ob die SchülerInnen einzeln begrüßt werden, lässt sich leicht feststellen. Aber pädagogisch gesehen ist entscheidend, *wie* die SchülerInnen begrüßt werden. In der Formulierung der Standards wird ein Oberflächenverhalten als Indikator für eine Haltung genommen, die es nicht notwendig repräsentiert. Äußerungen wie „Guten Morgen, Herr Lehrer“ im Chor oder Verhaltensweisen wie der Handschlag des Lehrers an der Tür mit Prüfung der Schmutzränder und der Abkaulänge der Fingernägel fallen in die Kategorie „tägliche Begrüßung“, *bedeuten* aber etwas Anderes, als in den Standards *gemeint* ist.
Daraus folgt**:** Auch standardisierte Instrumente sind hoch-inferent.

- Das gilt für die *Definition von Indikatoren*, denen von den ForscherInnen eine bestimmte Bedeutung unterstellt wird, die von den BeobachterInnen im Feld oder gar den Beobachteten, den Befragten oder Getesteten keineswegs selbstverständlich geteilt wird.
- Diese Verfahren sind aber auch hoch-inferent in der *Auswertung*, denn *„Zahlen sprechen nicht für sich"*[18]. Der Wert von Studien wie SCHOLASTIK und PISA liegt nur zu einem kleinen Teil in der Qualität der Instrumente begründet – zu einem viel größeren in der Intelligenz und dem Wissen von Forschern wie Franz Weinert, Jürgen Baumert und Manfred Prenzel, die die Daten auf eine anregende Weise interpretiert haben. Aber eben auf *ihre*, und man bekommt nicht minder anregende, aber in wichtigen Punkten ganz andere Deutungen, wenn man Klaus Klemm, Hartmut von Hentig oder Heinrich Bauersfeld an eben diese Daten setzt. Meine feste Überzeugung: Die zu erwartenden Differenzen liegen nicht an geringerer Intelligenz oder unzulänglichem Wissen einzelner Interpreten, sondern an einer *anderen* Sicht.

Deshalb ermutige ich nachdrücklich, an pädagogisch wichtigen Standards festzuhalten, auch wenn sie sich nicht leicht operationalisieren lassen. Manchmal sind *Beispiele* für die Verständigung („das ist wie...") hilfreicher als Versuche, Kategorien abstrakt zu definieren („das fällt unter...")[19]. Inhaltliche Wichtigkeit darf leichter Messbarkeit nicht untergeordnet werden – wir müssen dann andere Verfahren der Evaluation finden (➔ s. unten S. 34 ff.).

Dennoch bieten standardisierte Instrumente bestimmte Vorteile: Die Schulen erhalten Instrumente, die ihnen – z. B. im Fall von Tests - einen anderen Blick auf die Leistungen der SchülerInnen oder – im Fall von Fragebögen oder Checklisten[20] – einen anderen Blick auf den Unterricht und seine Bedingungen eröffnen. Man muss nur die Stärken und die Grenzen dieses Blicks sehen und verstehen.

Auch wenn man sich immer klar sein muss, wie begrenzt der Aussagewert punktueller Leistungsproben ist, will ich zwei Erträge besonders hervorheben, weil Tests von vielen reformpädagogisch orientierten LehrerInnen grundsätzlich abgelehnt werden:

Eine formalisierte Evaluation kann blinde Flecke in der Alltagswahrnehmung erhellen.

Andere bzw. gezielt entwickelte Aufgaben erlauben spezifischere Aussagen als die weithin üblichen Klassenarbeiten oder Leistungsproben, z. B. :

- Erfassung des leisen statt des lauten Lesens beim Stolperwörter-Lesetest[21]
- Gewichtung von Sachtexten neben den üblichen Lesebuchgeschichten bei IGLU und PISA.

Repräsentative Stichproben helfen, die persönlichen Maßstäbe durch Bezug auf Normwerte zu kalibrieren. Die Schulen erhalten Vergleichswerte, die sie zum Nachdenken über ihre Ansprüche und Bewertungskriterien veranlassen können, z. B.:

- Differenzierung nach bestimmten Schülergruppen (Mädchen vs. Jungen)
- Relativierung der Ergebnisse unter bestimmten Bedingungen (60% vs 20% Migrantenanteil).

---

[18] Vgl. zu dieser These ausführlicher: Brügelmann (2005b).

[19] Vgl. zum Wert analogen statt induktiv-deduktiven Denkens ausführlicher mein Plädoyer für einen „Miss-Marple-Stil in der Bildungsforschung" ➔ www.agprim.uni-siegen.de/printbrue/missmarple.pdf

[20] Vgl. etwa das DIPF-SEL oben Anm. 10.

[21] Vgl. zu diesem einfachen und als „Warnlampe" sehr nützlichen Instrument: www.lesetest1-4.de/html/stolperwortertest.html (Abruf: 19.1.2007) und Lust (2004).

5.1 Fazit: Weder standardisierte noch interpretative Verfahren können „Wahrheit“ sichern, keiner der beiden Zugänge ist dem anderen grundsätzlich überlegen.

Was kann man konkret tun?

- Untersuchungspraktisch empfiehlt sich, offene und geschlossene Verfahren zu kombinieren.
- In Interviews kann es ratsam sein, *erst* eine offene Frage zu stellen, um die spontane Einschätzung der Person zu erfassen, *dann* mit standardisierten Auswahlfragen nachzufragen, um eine höhere Vergleichbarkeit der Antworten zu erreichen..
- In Tests kann es hilfreich sein, *erst* standardisierte Antworten vorzugeben und *ergänzend* Möglichkeiten zur Kommentierung zu eröffnen, die wenigstens in offensichtlichen Zweifelsfällen zur Deutung von Lösungen herangezogen werden sollten.

6 Instrumente der Evaluation: Zur Gefahr methodenbedingter Kurzschlüsse

Das Methodenrepertoire pädagogischer Evaluation ist breit: Leistungsmessung, Unterrichtsbeobachtung, Personenbefragung, Dokumentenanalyse, um nur einige zu nennen - und das jeweils in unterschiedlichen Formen, z. B. mehr oder weniger vorstrukturiert – bis hin zur Standardisierung von Fragen und Antwortmöglichkeiten. Einseitigkeiten können durch die Bevorzugung bestimmter Instrumente entstehen.

Vier methodische Sonden des Evaluators

Leistungsproben:	standardisierte Tests <-> informelle Aufgaben Multiple choice <–> freie Lösung falsch/richtig <–> deutend kurzfristig <–> später
Befragung:	persönlich <–> medial vermittelt Auswahlfragen <–> freie Antworten Sachinformation <–> Meinungsäußerung
Beobachtung:	distanziert <-> teilnehmend strukturiert <–> offen zählend <–> verbalisierend
Spurenanalyse:	Primäräußerungen <–> Sekundärprodukte (z. B. Statistiken) verbale Dokumente <–> nonverbale „Produkte“

Es gibt also zwei grundlegend verschiedene Wege, die man bei einer Evaluation gehen kann:

- Man sammelt *Primärdaten*, indem man Leistungen testet, Abgängerstatistiken auswertet oder die Ausstattung und den Zustand der Räume erfasst. Und diese bewertet man anschließend.

Oder:

- Man befragt Personen, die über eigene Erfahrungen verfügen, nach deren Einschätzungen, was gut oder schlecht ist, um über solche *Sekundärdaten*, also eher indirekt ein Bild von den Qualitäten einer Einrichtung zu gewinnen.

Der erste Zugang ist weniger interpretationsgeladen, der zweite ist meist ökonomischer.

Spurenanalysen sind noch aus einem zweiten Grund attraktiv. Es handelt es sich dabei um „unobtrusive measures“, wie die Angelsachsen sagen, also non-reaktive Verfahren, bei denen die Datenerhebung die Personen – anders als bei Befragung und Beobachtung - nicht beeinflusst. Das gilt aber auch nur so lange, wie noch nicht bekannt ist, worauf EvaluatorInnen achten. Meine Vermutung: Nach den ersten Inspektionen werden Toiletten kräftiger geputzt,

Wandschmierereien häufiger entfernt, Klassenbücher stimmiger geführt werden – ohne dass sich der Alltag wirklich verändert

Leistungsproben oder (versteckte) Beobachtungen beeinflussen die Personen im Feld zwar weniger als Befragungen – letztere lassen aber zusätzlich erkennen, wie Situationen von den handelnden Personen eingeschätzt werden. Wahrnehmungen, Einstellungen und Wertungen anderer sind nicht nur als Rückmeldung für die eigene Arbeit wichtig; sie zu kennen hilft auch, das Verhalten der betreffenden Personen zu verstehen, denn sie sind handlungsleitend. Wer nicht nur an einer Beschreibung des Ist-Zustandes interessiert ist, sondern auch an Ursachen, muss Motive und Einstellungen erfragen. Menschliches Handeln ist nicht von objektiven Umständen gesteuert, sondern von deren Wahrnehmung durch die Beteiligten bestimmt.

Ob Eltern eine Schule wählen hängt eben nicht nur von deren Qualität nach fachlich anerkannten Kriterien ab, sondern auch vom *hearsay*, d. h. von den Erfahrungen und den Urteilen anderer Betroffener, in der Regel Laien. Darum sind Befragungen unverzichtbar.

Allerdings: Zufälligkeit persönlicher Erfahrung (ein Schüler hatte einen besonders guten Lehrer), Verzerrungen der Wahrnehmung (ein leistungsschwacher Schüler mag denselben Unterricht anders wahrnehmen als ein leistungsstarker) oder gar bewusst einseitige Urteile (ein Außenseiter lädt seinen Beziehungsfrust auf die Schule ab) sind Risiken dieses Zugangs. Sie lassen sich aber durch vier Maßnahmen minimieren, die ich grundsätzlich, also für eine jede Evaluation, wichtig finde:

Vier Kriterien für die Datenerhebung und -auswertung

➔ **„Sicherung von Mehrperspektivität“**

Befragung unterschiedlicher Personengruppen zu denselben Themen, z. B. Eltern UND LehrerInnen, jüngere UND ältere SchülerInnen; es gibt nicht nur eine Wirklichkeit und es gibt auch nicht die „richtige“ Sicht.

➔ **„Kombination von komplementären Methoden“**

Ergänzung standardisierter Items durch offene (Zusatz-)Fragen, um die Erwartungen/ Kriterien der Befragten zu erfahren, die ihre Bewertungen bestimmen, z. B. indem sie verschiedene (soziale, fachliche, ...) Ziele von Erziehung und Unterricht gewichten.

➔ **„Kontextbezug von Ereignissen durchsichtig machen“**

Variation von Konstellationen, Erhebung der Leistungen über verschiedene Aufgaben, Beobachtung des Verhaltens in verschiedenen Situationen, Deutung der Daten mit Bezug auf die „Sub-Kultur“ der Klasse oder Schule.

➔ **„Verdichtung UND Differenzierung von Ergebnissen“**

Darstellung der Ergebnisse nicht (nur) über bündelnde Mittelwerte, sondern durch bewusste Entfaltung der Streuung: ein Viertel hat sehr positiv, ein Drittel dagegen negativ geurteilt vs. die Mehrheit liegt im Mittelbereich.

Zu den bereits genannten Prinzipien möchte ich noch ein zentrales hinzufügen: Ökonomie. Der Ertrag muss den Aufwand lohnen – auch bei der Evaluation. Man sollte Information auf die

einfachste Art zu gewinnen versuchen – und da liegt es nahe, bereits verfügbare Daten zu nutzen (z. B. Bewerber- oder Verbleibestatistiken; Ausleihedaten in der Schülerbücherei; Buchhaltung über Reparaturkosten für Vandalismusschäden; ...). Oft lassen sich wichtige Daten mit wenig Aufwand gewinnen. Die einfachste Form der Evaluation ist LehrerInnen aus dem alltäglichen Unterrichtsgeschäft vertraut:

Klassenarbeit	intern
Erhebung	**Aufgabe**
Analyse	**Korrektur**
Bewertung	**Zifferennote**
Darstellung	**Notenspiegel**
Entscheiden	**Förderung**

Schon diese simple Form ist interessant für die Beurteilung auch des *Unterrichts*: Eine Leistungsprobe wirft ja nicht nur Noten für die einzelnen Schüler ab, sondern auch einen Durchschnitt und eine Streuung, die Sie mit Ihren Erwartungen abgleichen und deren Diskrepanz Sie außer auf „faule“ oder „dumme“ Schüler auch auf Ihren eigenen Unterricht beziehen können. Das macht vor allem dann Sinn, wenn Sie nicht von vorneherein eine Glockenkurve für die Verteilung unterstellen, sondern von inhaltlich definierten Erwartungen („Lernzielen“ mit neudeutsch „Kompetenzniveaus“) ausgehen.
Ich mache das so bei Klausuren, indem ich z. B. für die Aufgaben zu verschiedenen Bereichen ausrechne, wie hoch die Lösungsquote ist, so dass ich ein Profil des Wissens bzw. Könnens über die Bereiche hinweg und damit eine differenzierte Rückmeldung zu meinem Seminar bekomme[22]. Der große Vorteil dieses Vorgehens ist die Nutzung bereits vorhandener Daten: Prüfungsleistungen muss man sowieso erheben.

6.1 Fazit: Verschiedene Methoden eröffnen unterschiedliche Zugänge zur Wirklichkeit – mit je besonderen Stärken und Schwächen.

Was kann man konkret tun?

- Sichern Sie in der Evaluation, dass verschiedene Sichtweisen bei der Erhebung von Daten und bei ihrer Deutung repräsentiert werden.
- Komplementäre Erhebungsmethoden verringern die Gefahr, dass Schwächen und Einseitigkeiten einzelner Zugänge durchschlagen.

Wie aber lassen sich die Prozesse der Informationsgewinnung und –einschätzung am sinnvollsten organisieren?

[22] Dabei ist unterstellt, die Aufgaben sind gleich schwer... – eine Achillesferse jeder Forschungsmethode: Welcher Anteil des Ergebnisses ist dem Gegenstand, welcher der Erhebungsmethode zuzurechnen?

7 Aufteilung der Zuständigkeiten: Verbindung von Selbst- und Fremd-Evaluation in verschiedenen Konstellationen

Wenn ich die aktuelle Evaluationspraxis der "Blick über den Zaun"-Schulen richtig verstanden habe, lässt sie sich in vereinfachter Form so darstellen:

BüZ-bisher	intern	intime Externe	extern
Erhebung	**Fragen**	**Beobachtungen**	
Analyse	**gemeinsame Diskussion**		
Bewertung	**Selbstkritik**	**Vorschläge**	
Entscheiden	**Entwicklung**		

Das ist die optimale Basis für eine interne Evaluation: Die kritischen Freunde haben zwar einen fremden Blick, aber sie sind nicht Autorität und sie werden erst recht nicht als Feinde wahrgenommen. Vertraulichkeit und Vertrauen fördern die Bereitschaft, sich selbst Probleme einzugestehen. Jeder besucht jeden, wir haben also auch eine symmetrische Machtbeziehung, und dass die Datenhoheit bei der Schule bleibt, stärkt zusätzlich das Gefühl, nicht Opfer fremd bestimmter Maßnahmen zu sein.

Aber nichts ist vollkommen auf dieser Welt: Die wechselseitige Vertrautheit kann leicht gemeinsame Blindheit und das Wissen um den Rollenwechsel eine Beißhemmung erzeugen. Der fremde Blick verschleiert sich oder man sagt nicht so deutlich, was man denkt. Die Gefahr der Beschönigung und zusätzlich die Datenhoheit der Schule verringern außerdem die Glaubwürdigkeit nach außen. Damit befinden wir uns in einem Dilemma zwischen interner Produktivität der Evaluation für eine Verbesserung der Schule und ihrer Glaubwürdigkeit bei Rechenschaft nach außen. Ausgehend von diesem Dilemma will ich im Folgenden Funktionen und Formen von Schulevaluation noch einmal systematischer untersuchen – und zwar im Blick auf alternative Strukturen der Evaluation.

EvaluatorInnen brauchen Abstand – aber sie müssen auch mit der Kultur der untersuchten Schule vertraut sein. Gerade angesichts immer neuer Programme und Institutionen wie PISA, VERA und IQB[23] sollte man sich daran erinnern, dass Evaluation selbstverständlicher Bestandteil eines erfolgreichen Alltags ist – und das nicht nur im Beruf.

In der Küche probieren wir neue Rezepte aus und aufgrund unserer Erfahrungen verfeinern wir unsere Kochkünste zunehmend; wenn wir umziehen, sammeln wir Informationen über die Handwerker und Ärzte vor Ort und bauen uns aufgrund von Empfehlungen eine neue Infrastruktur auf; vor dem Kauf eines neuen Autos reden wir mit Bekannten, werten wir Berichte im Internet aus, machen wir eigene Probefahrten bei verschiedenen Händlern; wenn wir nur zwei Bücher auf eine Wanderung mitnehmen können, wählen wir unsere Ferienlektüre gezielt nach den Erfahrungen der Vorjahre und den Empfehlungen aus Zeitschriften oder von Freunden aus.

JedeR von uns evaluiert also ständig fremdes und sein eigenes Handeln. Die Vielfalt relevanter Kriterien und die wachsende Fülle an Informationen erschwert diese Aufgabe aber zunehmend

[23] IQB ist das von der KMK ins Leben gerufene *Institut zur Qualitätsentwicklung im Bildungswesen* in Berlin

– nicht nur im privaten, sondern auch im Berufsleben. Angesichts der wachsenden Komplexität von Entscheidungen haben wir zwei Optionen:

- Delegation von Evaluationsaktivitäten an SpezialistInnen: an die Stiftung Warentest, den TÜV – an ein Institut für Bildungsforschung oder ein Schulinspektorat.
  Der Preis: Abhängigkeit - und Übersetzungsprobleme bei den Ergebnissen;
- Entwicklung und Unterstützung der Evaluationskompetenz der Entscheidungsträger selbst: um die Diagnosefähigkeit von Lehrern und die Fähigkeit zur Selbsteinschätzung von SchülerInnen zu fördern; um Schulen zu helfen, den Erfolg ihrer Arbeit selbst zu untersuchen und Rechenschaft über ihn abzulegen. Der Preis: Vermittlungsprobleme bei den Verfahren – und Mehrarbeit.

Damit stehen wir im Bildungsbereich vor einem Dilemma: Spezialisierung vs. Demokratisierung. In diesem Kontext ist die gängige Unterscheidung in interne und externe Evaluation zu einfach. Die Interessen und Informationsbedürfnisse sind schon in diesen beiden Kategorien nicht auf einen Nenner zu bringen, wie etwa die folgende Auffächerung zeigt:

intern				**Extern**		
Leitung	**Lehrer**	**Schüler**	**Eltern**	**Experten**	**Autoritäten**	**Nutzer**

Wenn wir z. B. den Bereich „intern“ betrachten, ist eine zentrale Frage, wie wir mit den unterschiedlichen Informationsinteressen und Handlungsoptionen der drei Gruppen Leitung, LehrerInnen und SchülerInnen umgehen – zumal es auch unter Lehrern, unter Schülern und Eltern verschiedene Interessen und Informationsbedürfnisse gibt.
Die Einordnung einer Evaluation als „intern“ oder „extern“ wird weiter dadurch kompliziert, dass Teilaufgaben unterschiedlich zugeordnet werden können. Es lassen sich nämlich mindestens fünf Stufen im Prozess der Evaluation unterscheiden:

Beispiel	intern	extern
Erhebung: Wer sammelt die Daten?		
Analyse: Wer wertet sie aus?		
Darstellung: Wer kommuniziert die Ergebnisse?		
Bewertung: Wer würdigt ihre Bedeutung?		
Entscheiden: Wer zieht die Folgerungen und handelt?		

Die Zuständigkeit kann also auf jeder Stufe wechseln. Allerdings ist es nicht immer nötig, alle fünf Schritte personell oder institutionell zu trennen, und manchmal wechselt auch die *Reihenfolge*. So kann eine Darstellung (z. B. von Schulporträts)*vor* der Analyse (quer über mehrere Schulen hinweg wie bei der Schulinspektion in Bremen) liegen oder aber *nach* der Bewertung erfolgen und diese somit einschließen (wie etwa bei PISA). Dabei ist es attraktiv, eher technische Aufgaben als Dienstleistung nach außen zu vergeben.

Exkurs II: Beispiele für verschiedene Formen der Arbeitsteilung ➔ 49ff.

Der Preis externer Aufträge: *Arbeitsteilung* erleichtert den Alltag, macht aber auch abhängig von Spezialisten. Einen anderen Weg sind wir deshalb im Projekt EVI CIEL[24] Mitte der 1970er Jahre gegangen. Statt den Schulen Evaluationsaufgaben abzunehmen und durch Experten erledigen zu lassen, haben wir versucht, die *Evaluationskompetenz* der LehrerInnen *vor Ort* zu *stärken*, damit sie selbst ihre Arbeit besser evaluieren können. Auch der Grundschulverband versucht mit seinem Konzept „Pädagogische Leistungskultur" (Bartnitzky u. a. 2005; 2006), alltagstaugliche Hilfen für LehrerInnen zu entwickeln, um ihre diagnostische Kompetenz zu stärken und ihre Abhängigkeit von Testprogrammen wie VERA zu mindern. Externe Evaluation kann nur in großen Abständen stattfinden: landesweite Leistungstests jährlich, Inspektionen einzelner Schulen vielleicht alle vier Jahre. Das reicht nicht aus, um Unterricht kontinuierlich auszuwerten und zu verbessern. Evaluation kann sich nicht auf Rückmeldungen in der Sache beschränken. Sie sollte auch die Professionalität von LehrerInnen stärken.
Die Evaluationskompetenz derjenigen zu entwickeln, die Entscheidungen zu treffen und praktisch zu handeln haben, ist auch noch aus einem weiteren Grund wichtig: Verständnis und Nutzung der Ergebnisse von Evaluationsstudien verlangen eine hohe Interpretationskompetenz. Dies gilt für *alle* Verfahren der Untersuchung.

7.1 Fazit: Selbst- und Fremd-Evaluation haben verschiedene Funktionen, lassen sich aber nicht streng voneinander abgrenzen.

Was kann man konkret tun?

- Suchen Sie Entlastungsmöglichkeiten durch SpezialistInnen, z. B. in der Zusammenarbeit mit Forschungsgruppen, die Datenbedarf haben - oder mit anderen Schulen, so dass beide Seiten von einem Rollenwechsel profitieren können.
- Aber sichern Sie der Schule eine Mitsprache bei der Auswahl der Instrumente und die Deutungshoheit über die Daten.
- Entwickeln Sie darüber hinaus die Evaluationskompetenz vor Ort, z. B. durch eine kontinuierliche Nutzung unaufwändiger Formen der Selbstevaluation.

[24] s. ausführlicher EVICIEL-1976/-1983

8 Rollen der Evaluation: Verschiedene Stile aus unterschiedlichen Traditionen

Je nach dem Auftrag und Fokus der Evaluation, bieten sich andere Verfahren an. Damit komme ich zur Bedeutung verschiedener *Stile der Evaluation*[25].

Anders als das dominierende *measurement*-Paradigma unterstellt, kann die Überzeugungskraft von Erkenntnissen und Urteilen auf drei sehr unterschiedliche Weisen begründet werden:

- über die *technische* Qualität der Methoden und Instrumente,
- über das *individuelle* Ansehen, d. h. die Kompetenz bzw. die Autorität der verantwortlichen Personen, oder
- über die Akzeptanz der *sozialen* Regeln, unter deren Anwendung Erkenntnisse gewonnen, gedeutet und bewertet werden.

Die Prototypen dieser drei Traditionen von Evaluation sind in unserer Gesellschaft die Produktkontrolle bzw. der Warentest, die Kunstkritik und die Rechtsprechung.

8.1 Produkttester als Evaluationsmodell

Das technische Methodenverständnis von Evaluation ist aus der Qualitätskontrolle im Rahmen industrieller Fertigungsprozesse, aber auch von Produkttests der Stiftung Warentest oder von Testberichten in Computer-BILD her bekannt.

Das typische Instrument seiner Anwendung in der Bildungsforschung ist der Leistungstest. In dieser Tradition gibt es aber auch andere ergiebige Datenquellen: Fragebögen, statistische Kennwerte zu Kosten, zu Examensnoten und Dropout-Quoten usw.

Typisch sind zwei Merkmale

- die Standardisierung der Datenerhebung und
- die quantitative Auswertung in Form von Gruppenkennwerten.

Der Vorzug dieses Zugangs ist die Fokussierung der Beobachtung. Aber er hat seinen Preis, wie ich an einem – wenn auch überzeichneten - Beispiel verdeutlichen möchte. Die folgende Tabelle ist die Auswertung eines Textes durch Auszählung seiner Elemente:

A 4 x	B 1 x	C 1 x	D 1 x	E 10 x
F 2 x	G 1 x	H 3 x	I 5 x	J -
K 1 x	L 6 x	M 1 x	N 7 x	O -
P 3 x	Q -	R 3 x	S 3 x	T 2 x
U 6 x	V -	W 1 x	X -	Y -
Z -				

Eine solche Reduktion erlaubt eine präzise und objektive Messung von Textmerkmalen. Ihre Beziehung zueinander und damit ihre Bedeutung aber gehen dabei leicht verloren[26]. Versucht man dagegen die inhaltliche Bedeutung eines Textes zu erfassen und seine sprachliche

[25] ausführlicher OECD-1980

[26] Es handelt sich – karikiert – um die „wissenschaftliche Analyse des Goethe-Gedichts ‚Über allen Wipfeln ist Ruh' durch ein Institut für molekulare Poetik" von Dürr (zit. nach Popp 2006, 11-12). Vgl. ähnlich hübsche Beispiele zu „aufgeräumter Kunst" bei Wehrli (2002; 2004).

Qualität zu bewerten, werden die Ergebnisse nicht mehr mit gleicher Verlässlichkeit reproduzierbar sein.

Je nach Interesse und Blickwinkel bekommen wir etwas Anderes in den Blick: das „große Ganze“ mit Verständnis für seine Bedeutung oder die einzelnen Elemente jeweils für sich und mit großer Detailgenauigkeit – aber nie beides gleichzeitig. Wie mit einer Lesebrille oder gar einem Mikroskop können wir kleine Ausschnitte sehr genau betrachten. Der Preis: Wir sehen den Kontext nicht mehr. Wenn man eine neue Wohnung sucht, ist es sinnvoll, den Lack der Türen, die Tapete in den Zimmerecken mit einer Nahbrille genau zu prüfen. Aber ob das Haus ein gutes Zuhause ist, können wir nur entscheiden, wenn wir mit dem Alltagsblick durch die Wohnung gehen, wenn wir das Haus von außen mit Abstand betrachten.

Auch in der Sozialforschung gilt (in einer groben Analogie zur Physik) eine Art Unschärfe-Relation[27]: So kann sie versuchen, additiv viele Details genau zu erfassen – oder ein strukturiertes Gesamtbild zu gewinnen (das „Zoom-Dilemma“). Anders gesagt: Evaluation steht in der Spannung der Anforderungen von technischer Genauigkeit einer Messung und inhaltlicher Bedeutung ihres Ergebnisses.

Zurzeit erleben wir (wieder einmal...) die hohe Zeit des technischen Paradigmas. In der Tat ist die Komplexität der Instrumente zur Datenerhebung und der statistischen Verfahren zur Auswertung der Befunde beeindruckend. Und an dieser Stelle will ich ganz deutlich sagen: Ich halte diese Art von Forschung nicht nur für legitim, sondern für außerordentlich wichtig und fruchtbar. Mehr noch: die Möglichkeiten ihrer Nutzung müssen noch viel breiter bekannt und verfügbar gemacht werden.

Gleichzeitig erschrecken mich die Dominanz dieses Paradigmas in der geförderten Forschung und die Überschätzung ihrer Befunde in der Öffentlichkeit. Das Vertrauen auf die technische Qualität der Datengewinnung und –verarbeitung unterstellt dieser Forschung eine Aussagekraft, die ihren methodologischen Kredit weit übersteigt.

Exkurs III: Erkenntnistheoretische Fallen der Sozialforschung ➔ 51ff.

Aber dieses Paradigma ist nicht unsere einzige Option – und es ist oft nicht die beste. Darauf spielt der Titel meines Vortrags an: Scharfe Brillen – wie die standardisierten Instrumente sie uns bieten - sind wichtig. Aber jede Brille erschließt nur bestimmte Aspekte der Wirklichkeit und sie verschließt andere. Das ist der Preis, den wir für die Fokussierung zahlen. Und zweitens wissen wir aus der Sinnesphysiologie, dass eine Perfektionierung der Peripherie, also der Sinnesorgane gar nichts nutzt, wenn die Zentralverarbeitung im Gehirn nicht richtig funktioniert. Das meine ich mit dem interessierten Blick. Wenn man an das Zoom-Problem und die Analogie zur Unschärferelation denkt, gibt es sogar Situationen, in denen ein waches Auge *ohne* Brille *mehr* sieht – von dem, was wichtig ist

8.2 Kunst-Kritik als Evaluationsmodell

Otto Seydel (2005) hat in einem Kommentar zur Bremer Evaluation die Frage: „Sind Schulen Kunstwerke?“ mit JA, beantwortet, denn Schulen seien Kunstwerke, sie seien u. a.

[27] Diese Analogie verdanke ich Heinrich Bauersfeld (u. a.: 1972, 243), der sie bereits vor 35 Jahren in der Debatte über lernzielorientierten Unterricht immer wieder nachdrücklich ins Spiel gebracht hat.

- einmalig, nie vollständig reproduzierbar (weil jedes Kind einmalig ist, weil jeder Lehrer eine Geschichte durchlebt);
- in der kreativen Kombination der Gelingensbedingungen letztlich keinem (oder nur sehr wenigen) Gesetzen unterworfen
- flüchtig im zeitlichen Verlauf (Unterricht bleibt nie „stehen"!);
- störempfindlich - ein falscher Ton, eine zusätzliche Prise kann alles verderben
- subjektiv - die Schule, die für den einen Schüler genau die richtige ist, kann für den anderen eine Katastrophe sein."

Seydel ist mit diesen Einschätzungen ganz nah bei Horst Rumpfs (1976; 1986) Idee vom Curriculum nicht als Blaupause, sondern als Partitur für eine je neue Inszenierung unter den spezifischen Bedingungen der einzelnen Lerngruppe. Didaktik als Regieanweisung, Unterrichten als Kunst, als eher implizite Fähigkeit zur situativen Interpretation Drehbüchern – all' diese Bilder erinnern an Polanyis Konzept der *tacit knowledge*, das u. a. Neuweg (2006) für das Verständnis des Lehrerhandelns wieder fruchtbar gemacht hat: Man kann mehr, als man weiß. Menschliches Handeln lässt sich nur sehr begrenzt auf explizite Erklärungen zurückführen und durch explizite Anweisungen steuern.

Bedeutet die Anerkennung dieser Aspekte pädagogischer Tätigkeit, dass wir auf Evaluation verzichten müssen? Keineswegs. Im Kunstbetrieb hat Evaluation nur einen anderen Namen: *Kritik*, z. B. über Rezensionen. Kritik, ob in der Literatur, im Theater, in der Musik, setzt auf die Person des Rezensenten als Evaluationsinstrument. Dessen Erfahrung, Sensibilität, Glaubwürdigkeit sind die Basis dafür, dass seine Bewertungen Resonanz finden. Dass Kritiker-Urteile oft sehr umstritten sind, spricht nicht gegen sie. Dass ist bei PISA, IGLU und VERA nicht anders.

Seit langem ist „art criticism" ein gut etablierter Stil von Evaluation (vgl. Eisner, Stake, Stufflebeam u. a. bereits in der 1970er Jahren in den USA[28]). Und dieses Modell wird nicht nur in den Künsten genutzt. Die „Jury für den Deutschen Schulpreis" ist ein aktuelles Beispiel, und auch das Inspektoren-Modell der Schulbesuche, das verschiedene Bundesländer aus den Niederlanden übernommen haben, folgt der Einsicht, dass Wesentliches in pädagogischen Situationen und Prozessen nicht technisch gemessen, sondern nur durch persönliche Sensibilität wahrgenommen werden kann. Man versucht also nicht, Subjektivität als Störung auszumerzen, sondern als Potenzial zu kultivieren[29]. Erfahrung ist die Basis für eine Differenzierung des Urteils.

In der Art und Weise, wie die Schul-Verbünde des "Blick über den Zaun" Evaluationskompetenz nutzen und *on the job* weiter entwickeln, sehe ich eine Stärke des Reformkonzepts, die geschätzt und gepflegt werden sollten.

Die Abhängigkeit von einzelnen Personen, die Subjektivität ihrer Wahrnehmungen, Deutungen und Urteile ist zugleich eine grundsätzliche Schwäche dieses Ansatzes.

Allerdings wird die Gefahr subjektiver Einseitigkeit in den meisten dieser Konzepte durchaus gesehen. Deshalb wird die persönliche Erfahrung und Autorität ja durch Teambildung relativiert – nicht nur bei Ihnen, sondern in der Bremer Evaluation oder beim Deutschen Schulpreis. Damit werden institutionell organisierte *soziale Kontrollen* erforderlich, wie wir sie aus der juristischen Tradition kennen. Evaluation bezieht sich ja nicht nur auf die Frage, ob tatsächlich stattfindet,

[28] Vgl. auch ihre Beiträge in Wulf (1972) und analoge Aufsätze in der Zeitschrift „Thema Curriculum" (1972-1974).

[29] Das gilt aber nur dort, wo nicht immer feinere Kriterienraster die Urteilsfähigkeit der BeobachterInnen ersticken.

was behauptet wird, sondern auch darauf, welche Kriterien angemessen sind, um einen festgestellten Sachverhalt zu bewerten. Genau diese beiden Aufgaben stellen sich auch im Rechtssystem.

8.3 Rechtsprechung als Evaluationsmodell

Wer Schule als Konfliktfeld widersprüchlicher Werte und konkurrierender Wahrnehmungen kennt, dem erscheint das Gerichtsverfahren als Evaluationsmodell nicht fremd. Das juridische Verständnis der Streitschlichtung konzipiert Evaluation als sozialen Prozess, mit einem geregelten Verfahren. Juristen haben die Erfahrung machen müssen, dass es keinen inhaltlichen Maßstab für die Wahrheit von Aussagen und für die Richtigkeit von Urteilen gibt, dass man aber durch ein System von *checks and balances* die Wahrscheinlichkeit verringern kann, dass sich bewusste Manipulationen durchsetzen. Soziale Kontrolle statt technischer Präzision ist die Grundidee des Rechtssystems.
Sie wird realisiert über verschiedene konkrete Prinzipien, die auch für die Evaluationsverfahren im Bildungsbereich hilfreich sein können:

8.3.1 ➔ Mehraugen-Prinzip

Bei Prüfungen ist Teambildung im Bildungswesen schon lange selbstverständlich, und auch bei der Schulinspektion hat sie sich durchgesetzt. Im juridischen Prozess wird dieses Prinzip bewusst differenziert, z. B. durch die Kombination von Volljuristen und Laienrichtern. Für die Evaluation kann das bedeuten, bewusst Personen mit unterschiedlichem Erfahrungshintergrund – und nicht nur KollegInnen aus anderen Schulen – einzuladen. Interessant könnten Vertreter sein...

- aus dem Schulausschuss der Gemeinde
- von der IHK
- für das Jugendamt
- aus der Kirche.

Sie werden jeweils Unterschiedliches wahrnehmen und sie werden evtl. die Wahrnehmung desselben unterschiedlich bewerten

8.3.2 ➔ Funktionale Trennung von Rollen

Im juristischen Prozess werden Zuständigkeiten bewusst aufgefächert, z. B. zwischen

- Richter und Sachverständigen
- Parteien und Zeugen
- Kläger- und Beklagten bzw. Staatsanwalt und Verteidiger.

Eine solche Rollentrennung scheint mir für Evaluationen aus ganz verschiedenen Gründen und auch in unterschiedlichen Formen sinnvoll.
Beispielsweise könnte man bei einem Schulbesuch die „kritischen Freunde“ in eine Pro- und eine Contra-Gruppe teilen: „Versetzt euch in die Rolle eines Kritikers bzw. eines Freundes der Schule und sucht systematisch nach Belegen für euer Vorurteil.“ Die „Verkleidung“ als Kritiker ist gerade für Sympathisanten eine Erleichterung, reduziert die Beißhemmung unter KollegInnen. Umgekehrt fällt es leichter, Kritik anzuhören, wenn sie aus der entsprechenden Rolle kommt – und wenn man weiß, dass die positive Seite auch zu ihrem Recht kommt. Dabei ist die Begriffswahl wichtig: Meist geht es nicht (statisch) um „Stärken“ und „Schwächen“, sondern um „Potenziale“ und „Risiken“. Für dasselbe Problem gibt es unterschiedliche Lösungen. Jede Maßnahme bietet Möglichkeiten, enthält aber auch Gefahren. Nur wer sich der

Ambivalenz bewusst ist, kann eine produktive Balance finden. Oft entscheidet auch der Kontext, welche Seite stärker zur Geltung kommt. Darum ist es wichtig PRO und CON bewusst zu machen und im Bewusstsein zu halten, statt zu raschen Urteilen zu kommen.
Sinn macht aber auch die Trennung nach Phasen der Evaluation, z. B. Erhebung und Analyse vs. Bewertung der Daten. Eine Zuordnung verschiedener Aufgaben zu verschiedenen Personen/ Institutionen hätte z. B. bei den großen Testprogrammen von Vornherein deutlich machen können, dass die empirischen Befunde nicht für sich sprechen, sondern kompetenter Deutung bedürfen, die je nach Hintergrund unterschiedlich ausfallen kann. Die Mehrdeutigkeit von Daten wird transparenter, wenn ihre Erhebung und Interpretation nicht in einer Hand liegen.
Eine klare Definition und Trennung von Rollen entlastet. Ihre Zuweisung zu glaubwürdigen Personen (z. B. im abschließenden Hearing ein Elternvertreter als „Richter“ = Moderator des Verfahrens) steigert die Akzeptanz.

8.3.3 ➔ Stufung von Zuständigkeiten

Das Gerichtsverfahren ist nach Instanzen hierarchisch aufgebaut, um Berufung oder Revision zu ermöglichen. Solche Appellationsmöglichkeiten sind für Evaluationen nur in Ausnahmefällen wichtig. Bedenkenswert ist aber die Stufung von Zuständigkeiten.
Was VERA und andere Testprogramme auf der zentralen Ebene für die Systemevaluation leisten, rechtfertigt deshalb noch lange nicht entsprechende Aktivitäten auf der Schul- oder gar Schülerebene. Schon die unterschiedliche Bedeutung von Messfehlern auf Gruppen- und Individualebene verweist auf die funktionale Begrenztheit von Instrumenten.

8.4 Fazit: Wir müssen im Bildungswesen Rechenschaftspflichten und –formen nach verschiedenen Ebenen und Funktionen differenzieren[30] - und die besonderen Stärken verschiedener Traditionen und Stile aufgabenbezogen nutzen.

Was kann man konkret tun?

- Standardisierte Instrumente lassen sich zur Einordnung von spezifischen Situationen und zur Bestimmung von Randbedingungen nutzen, die für die Fairness der Beurteilung wichtig sind (Vortests!).
- Beziehen Sie Personen in die Evaluation ein, die wegen ihrer Kompetenz / Integrität bei Ihren Zielgruppen anerkannt sind (z. B. Vertreter der Stadt, der Kirche, aus dem Vereinsleben).
- Die Evaluation ist als ein transparentes Verfahren mit klar definierter Rollenverteilung („checks and balances“) zu organisieren, so dass die Gefahr von Informations-Privilegien oder Interpretations-Monopolen verringert wird.

[30] Vgl. etwa den Vorschlag von Bartnitzky u. a. (1999) für den Grundschulverband.

9 Präsentationsformen der Evaluation: Wie lassen sich die Ergebnisse hilfreich und verständlich darstellen – und für wen?

Dieser Schritt misslingt oft schon deshalb, weil vorher keine klaren Fragen formuliert worden sind. Die inflationäre Berufung in sich widersprüchlicher Politikvorschläge „auf PISA“ ist die Folge eines unklaren Auftrags bzw. einer Überforderung der Studie mit Ansprüchen, die sie nicht erfüllen kann.

Grundlage einer erfolgreichen Berichterstattung sind deshalb klare Absprachen. Höchste Priorität hat für mich ein frühes Gespräch der Evaluatoren mit den Beteiligten über die Informationsbedürfnisse der einen und über die Untersuchungsmöglichkeiten der anderen. Evaluationsaufträge sollten also aus den Aktivitäten der Schul- und Unterrichtsentwicklung erwachsen und wieder in sie münden. Ohne Einbindung der Wortführer nützen die besten Argumente nichts – womit wir wieder beim Ausgangspunkt sind: Verletzlichkeit der Personen und Machtverteilung in der Institution ...

Was das bedeutet, habe ich gleich bei der ersten Evaluation erlebt, die ich noch sehr naiv durchgeführt habe. Ich habe damals in England ein Teachers’ Centre erforscht. Meine Idee war, allen am Centre Beteiligten die verschiedenen Sichten der anderen zurückzuspiegeln, indem ich Ausschnitte aus Dokumenten, aus Interviews und Beobachtungen nach *topics* geordnet, aber unkommentiert, also in Originalform, zusammengestellt und veröffentlicht habe[31]. Transparenz und Demokratisierung von Information war meine Idee – heiß laufende Telefondrähte und persönliche Vorwürfe die Folge.

Meine naive Vorstellung war: Alle Beteiligten freuen sich, wenn sie mehr wissen über die Stärken und Schwächen des Zentrums und wenn sie erfahren, wie andere über die Qualität ihrer Arbeit denken. Mit meinem aufklärerischen Impetus hatte ich fälschlich Chancengleichheit in der Nutzung der Information unterstellt und ich hatte die eingangs erwähnten Probleme: persönliche und institutionelle Verletzlichkeit, schlicht unterschätzt.

Hinzu kommt ein zweites Problem. Die aktuelle Entwicklung der Bildungsforschung spiegelt ein Dilemma wie in der Geheimwissenschaft Medizin: Der Gewinn technischer Erkenntniskraft von Expertendiagnosen wird erkauft mit einem Verlust an Sensibilität für die Biografie und das Milieu der Betroffenen. Die Abhängigkeit der PraktikerInnen von der Bildungsforschung wächst mit deren zunehmender Spezialisierung: Diese suggeriert mit ihrer Fachsprache eine Scheinpräzision und –sicherheit der Instrumente und Ergebnisse und sie versteckt die Relativität von Urteilen in statistischen Formeln. Wirksam werden als Praxishilfe können diese nur durch eine Reduktion von Komplexität in handlungsbezogenen Übersetzungen – durch EvaluatorInnen als fachkundige, aber unabhängige Makler.

Insofern beginnt, wenn die Ergebnisse vorliegen, oft erst der schwierigste Teil: Wie kommt man zu umsetzbaren Folgerungen – und mit welchen Reaktionen anderer muss man rechnen? Dabei plädiere ich für eine Relativierung der Position des Evaluators und für Transparenz, indem Evaluation als Lernprozess für alle Beteiligten organisiert und dokumentiert wird:

- **Stufung der Berichterstattung**
  Bericht über die Evaluation als Lernprozess der Evaluatoren

[31] Vgl. Brügelmann (1976).

  - Was habe ich vorher gedacht?
  - Welche Erfahrungen haben mich überrascht?
  - Welche Informationen habe ich daraufhin gesucht?
  - Wie sehe ich die Einrichtung/ Person/ Situation jetzt?
  - Warum sehe ich sie jetzt anders?
- **Darstellung der Ergebnisse aus verschiedenen Blickwinkeln**
  Das kann eine Pro- und Contra-Ordnung sein, aber auch eine mehrperspektivische Einschätzung aus verschiedenen Rollen heraus.
- **Mündliche Verhandlung der Ergebnisse**
  Damit haben die Betroffenen die Möglichkeit zu Fragen und Kommentaren.

Gerade weil es die eine Wahrheit über eine Schule nicht gibt, ist es hilfreich, die Wahrheit der anderen wenigstens zu kennen – schließlich bestimmt sie deren Verhalten. Aus dem politischen Raum sind *hearings* als eine Form bekannt, in der externer Sachverstand – über schriftliche Gutachten hinaus – inhaltlich verarbeitet wird. Schulöffentlich veranstaltet könnte eine solche Befragung dazu beitragen, dass

- Unklarheiten oder gar Fehler frühzeitig bereinigt
- Missverständlichkeiten geklärt
- Informationsmonopole verhindert

werden.

9.1 Fazit: Evaluation sollte sich nicht als *Autorität*, sondern als *Dienstleistung* verstehen.

Was kann man konkret tun?

- Evaluation ist als Lernprozess für alle Beteiligten zu organisieren und entsprechend zu dokumentieren – über ...
- eine Stufung der Berichterstattung als Bericht über den Lernprozess der Evaluatoren;
- über eine Darstellung/ Kommentierung der Ergebnisse aus verschiedenen Blickwinkeln;
- über eine (schul-)öffentlichen Befragung der EvaluatorInnen nach Vorlage des Berichts.

10 Evaluation der Evaluation: Lohnt der Aufwand überhaupt?

Lassen Sie mich mit einer persönlichen Bemerkung zum begrenzten Nutzen von Evaluation schließen:

Die Qualität und der Ertrag von Evaluation sind bisher kaum evaluiert worden. Dabei ist das Verhältnis von Aufwand und Ertrag, von positiven Effekten und negativen Nebenwirkungen durchaus fragwürdig.

Ich habe bereits kurz den vorzüglichen Vortrag von Neuweg zum Thema „Das Schweigen der Könner“ erwähnt. Im Anschluss an Polanyis Konzept der „tacit knowledge“, Kleists „Marionettentheater“ und andere erinnert er an die Ambivalenz unseres ständigen Bemühens um Aufklärung über die Grundlagen und Motive unseres Handelns. Ich will kurz zitieren (Neuweg 2006, 28-29):

> Es gibt ganze Organisationen, die nichts dem Zufall überlassen wollen und Wissensexplikation zum Programm erheben. Ämter gehören natürlich dazu, und auch McDonald's, wenn man Robin Leidner glauben darf, die in den 90er-Jahren davon berichtet hat, dass es dort ein 600-seitiges „Operations and Training Manual" gibt, das von den Managern und Franchisenehmern als „Bibel“ bezeichnet wird - wohl auch wegen der Vorteile biblischen Ausmaßes: McDonald's kann überaus rasch wachsen, weil die an den einzelnen Standorten gelebte Praxis raum-zeit-unabhängig beliebig oft replizierbar ist; der Hamburger ist ebenso standardisiert wie das Lächeln der ihn überreichenden Mitarbeiterin; die Organisation ist von der individuellen Könnerschaft einzelner Mitarbeiter praktisch gänzlich unabhängig.
>
> „Diese Form der Prozessintelligenz“, so der Betriebswirt Rüdiger Reinhardt von der Universität St. Gallen, „basiert somit auf der gleichförmigen Nutzung dokumentierten Wissens durch eine Vielzahl von Anwendern.“
>
> Was Reinhardt nicht dazu sagt: Diese „Prozessintelligenz“ findet sich, wenn überhaupt, nur noch in der Organisation, nicht mehr aber bei ihren Mitgliedern. Denn die werden bei Barbara Garson oder Robin Leidner mit folgenden Worten zitiert:
>
> „And that's what [McDonald's] is, a machine. You don't have to know how to cook, you don't have to know how to think. There's a procedure for everything and you just follow the procedures."30
>
> „Sie haben es so weit herunter gebrochen, dass es praktisch idiotensicher ist."51
>
> Die totale Explikation des Wissens macht die Organisation offenbar vor Idioten sicher, sichert ihr aber zugleich auch Idioten.

Bereits beim Modell der Evaluation als Kunstkritik habe ich darauf hingewiesen: Mehr über etwas zu *wissen* heißt nicht unbedingt, es auch zu *können*. Können lässt sich nur begrenzt bewusst machen und sprachlich fassen. Und das, was Experten sprachlich fassen können, hilft Novizen nicht immer, besser zu werden. Praktisches Handeln ist in hohem Maße auf Erfahrung angewiesen, die nicht durch formelle Evaluation gewonnen werden kann.

Andererseits: Reflexion ermöglicht uns, Abstand vom alltäglichen Handlungsdruck zu gewinnen und uns über seine Rahmenbedingungen klar zu werden, um diese zu verbessern.

Das bedeutet: Evaluatoren müssen nicht selbst gute Pädagogen sein. Sie müssen etwas von Unterricht verstehen, insofern reicht reine Methodenkompetenz nicht, aber selbst unterrichten können müssen sie nicht.
Ich hoffe, das gilt auch für Berater von Evaluatoren: Sie müssen nicht selbst gute Evaluatoren sein, aber sie sollten etwas von Evaluation verstehen, so dass sie beraten können.
Ich selbst tue mich schwer mit Beratern – aber *selbst* bin ich liebend gern aktiv als Berater... und erstaunlicherweise nicht selten auch ein erfolgreicher. Kurioserweise auch dann, wenn ich selbst nicht nach meinen Ratschlägen handele.
Sie sehen, Evaluation ist ein verqueres Geschäft. Allerdings wissen wir: Wie gut auch immer jemand seine Sache macht - Handlungszwänge schwächen die Erkenntnisfähigkeit und Lernbereitschaft. Freiheit vom Handlungsdruck dagegen *kann* zumindest hellsichtig machen – und damit verbinden in der Hoffnung, dann auch offener für neue Erfahrungen zu werden.
Ein interessantes Phänomen – und vielleicht dann doch eine Rechtfertigung für Evaluation durch *andere*...

10.1 Fazit: Die Planbarkeit und die Steuerungsmöglichkeiten menschlichen Handelns sind sehr begrenzt.

Was kann man konkret tun?

- Dringen Sie auf ein ökonomisches Verhältnis von Aufwand und Ertrag in der Evaluation (z. B. Sekundärnutzung anfallender Daten anstelle immer neuer Primärerhebungen).
- Führen Sie ein Blitzlicht am Ende eines jeden Verfahrens – und eine Evaluationskritik einige Monate nach der Evaluation durch: „Was war hilfreich für meine Alltagsarbeit? Was sollten wir beim nächsten Mal anders machen?“

P. S.: Um noch einmal auf die eingangs erwähnte Gefahr zurückzukommen, dass die Fülle der Optionen und Differenzierungen möglicherweise erschlägt statt anzuregen. Diese Sorge hat mein Kollege Horst Rumpf nach einer ersten Lektüre dieses Texts sehr deutlich geäußert. Darum betone ich noch einmal: Evaluation muss *ökonomisch* sein – und sie kann das nur, wenn aus der Fülle der hier entfalteten Möglichkeiten gezielt Elemente ausgewählt werden, um das jeweils konkret anstehende Problem zu lösen. Rumpf selbst hat für das Nachdenken über die Qualität von Unterricht fünf Fragen vorgeschlagen, die ein solches Angebot darstellen, die Aufmerksamkeit auf ausgewählte Probleme zu konzentrieren. Wegen ihrer bewussten Reduktion auf *einen* Aspekt von Schule aus *einer* Perspektive auf Unterricht und damit als Gegenposition zu den üblichen Lernstandserhebungen zitiere ich diese Provokation am Schluss meines Beitrags:

(1) War die Sache, um die es im Unterricht ging, wirklich anwesend (oder war sie nur eine Attrappe für diverse verbale, schriftliche oder handgreifliche Aktivitäten)?

(2) Waren die Menschen, die zusammen den Unterricht bestritten, wirklich beteiligt oder war ihre menschlich-sinnliche Existenz (aufgeladen mit ihren Erinnerungen, Ängsten, Hoffnungen und wilden Ideen) nur ungeschichtliche Prothese für ein äußerliches Mitmachen?

(3) Hatten die am Unterricht teilhabenden Personen Gelegenheit, sich auf die im Unterricht verhandelte Sache und ihre Innenspannung hinreichend intensiv und differenziert einzulassen? Wenn nein, was stand im Wege? (Gab es Symptome der „straffen Flüchtigkeit“ [Martin Wagenschein]?).

(4) Ermöglichte der Unterricht ein gemeinsames Sichvertiefen in einen von ebenso befremdlichen wie herausfordernden Zügen durchzogenen Sachverhalt oder glich er eher einem Hürdenlauf über von vornherein feststehende Aufgaben und Schwierigkeiten? (Hatte der Unterricht Atem für leere Räume, für Brüche und Zwischentöne? Vielleicht auch für Ratlosigkeiten angesichts einer zunehmend unbekannter und ferner werdenden Sache? Sympathisierte er gar mit deren Entstehen – oder tat er alles, sie unkenntlich zu machen oder zu durcheilen).

(5) Wurden die Inhalte ihres Eigengewichts und ihres Nährgehalts für eine nachdenklich-vergegenwärtigende Anschauung entledigt und auf eine notenproduktive Aktivierung zurechtstilisiert?

11 Exkurs I: Zur Rolle von SchülerInnen bei der Evaluation

Üblicherweise sind *SchülerInnen* entweder Objekte der Evaluation (Noten; Leistungstests) oder Datenquelle (Befragung; Interview). Sie spielen aber kaum eine Rolle bei der Entwicklung von Fragestellungen, bei der Analyse und Bewertung von Daten oder gar bei Folgerungen und Entscheidungen

Wenn man die **"Blick über den Zaun" -Standards** nicht nur als inhaltliche Maßstäbe für die Evaluation, sondern auch *als Anspruch an den Evaluationsprozess* selbst ernst nimmt, geht das nicht mehr. In der deutschen Kindheitsforschung (Behnken, Zinnecker u. a.) und in der angelsächsischen und der skandinavischen Schulforschung werden SchülerInnen aus verschiedenen Gründen zunehmend aktiv in die Evaluation von Bildungseinrichtungen einbezogen:

- als Experten ihrer Lebenswelt
- als Handelnde mit eigenen Handlungsoptionen (Kurswahl, Beteiligung am Unterricht)
- als Träger von Rechten (UN-Kinderrechts-Konvention).

11.1 SchülerInnen als Experten ihrer Lebenswelt

Wer seine Schule besser verstehen will, für den müsste das Bild von Schule wichtig sein, dass die SchülerInnen entwerfen, um in ihr zu überleben. Im Projekt „LERNenBILDung“[32] von Behnken/ Zinnecker u. a. an der Universität Siegen wurden Fragen entwickelt, die ich unbedingt in die Selbstevaluation einer Schule einbeziehen würde – und die ich auch für eine Außendarstellung erhellend finde:

- Welche Fächer sind dir wichtig – und warum?
- Was muss man tun, um im Unterricht dieser Schule Erfolg zu haben?
- Wie muss man sich verhalten, um in der Klasse anerkannt zu sein?

SchülerInnen werden also nicht nur als kompetente Quelle für *Sachinformationen* „genutzt“ („wie oft könnt ihr...“), sondern auch ihre *Meinungen* und *Urteile* werden ernst genommen.

11.2 SchülerInnen als Handelnde mit eigenen Handlungsoptionen

SchülerInnen haben ständig Entscheidungen über ihre eigene Bildungsbiografie zu treffen. Fach- und Kurswahlen sind z. B. formelle Entscheidungspunkte. Aber auch tagtäglich müssen sich SchülerInnen entscheiden, welche Ziele ihnen wichtig sind, wie sie ihre Arbeit organisieren wollen, wie sie mit Schwierigkeiten umgehen. Fachnoten sind die traditionelle Form der Evaluation, die SchülerInnen meist als Urteil „von oben“ erleben. Förderdiagnostische Ansätze bedeuten demgegenüber einen Wechsel zur Beratung, die die Eigenverantwortung der Schüler ernster nimmt. Dafür erproben LehrerInnen zunehmend auch Formen der *Selbsteinschätzung* – die sie nicht nur als Medium der Leistungskontrolle nutzen, sondern auch als zu entwickelnde Kompetenz fördern.

Solche Selbsteinschätzungen kombiniert mit Wissensfragen können auch helfen, Lernzuwächse festzustellen, wie dies meine Kollegin Schmidt-Peters tut, von der ich die Idee für mein Seminar „Förderkonzepte“ im kommenden Semester übernommen habe:

[32] Vgl. Behnken u. a. (2004).

Mit den folgenden Fragen möchte ich gerne Ihren Lernzuwachs festhalten

A Was wollen Sie über die/ aus der Reformpädagogik für sich lernen?

B Notieren Sie möglichst viele ReformpädagogInnen, die Sie kennen:...

B Nennen Sie einige Prinzipen/ Arbeitsformen/ Lehr-/Lernmittel, die von ReformpädagogInnen entwickelt worden sind:.....

C Wie weit fühlen Sie sich in der Lage, die Besonderheiten von drei selbst gewählten ReformpädagogInnen zu charakterisieren? (Note)

D Wie groß ist Ihr Interesse, in einer reformpädagogisch arbeitenden Schule zu hospitieren? (Note)

E Wie stark ist Ihre aktuelle Vorstellung von „gutem Unterricht“ durch konkrete reformpädagogische Ansätze beeinflusst? (Note)

F Wie gut vorbereitet fühlen Sie sich, selbst gewählte reformpädagogische Unterrichtselemente in Ihrem zukünftigen Unterricht einzusetzen? (Note)

In einer formellen Evaluation würde ich noch ein Stück weiter gehen. Für mich gehört zu einer jeden Selbst- wie auch Fremdevaluation eine Befragung von ehemaligen SchülerInnen: Wo sehen sie die Stärken und Schwächen der Schule, aber auch: Was denken sie selbst richtig und falsch gemacht zu haben? Welche Ratschläge würden sie jüngeren SchülerInnen geben? Sowohl LehrerInnen als auch SchülerInnen können aus solchen Rückblicken viel lernen.

11.3 Schüler als Träger von Rechten

Die Kinderrechtskonvention ist 1989 von den UN verabschiedet und 1992 vom Deutschen Bundestag ratifiziert worden. In unserem Schulalltag spielt sie praktisch keine Rolle. Ganz anders etwa in England. Nur ein paar Titel aus den letzten Jahren:

- Pollard/Triggs (2000) What pupils say: Changing policy and practice in primary education
- Devine (2003) Voicing Concerns about children's rights and status in schools
- Rudduck/Flutter (2004) How to improve your school: Giving pupils a voice
- Cox u. a. (2006) Children decide: Power, participation and purpose in the primary classroom.

Es gibt also eine Reihe von Gründen, SchülerInnen als PartnerInnen in die Verantwortung für die Evaluation von Schule und Unterricht einzubeziehen – und nicht nur sie. Gegenwärtige läuft der Trend auf Evaluation als Medium der professionsinternen Kontrolle und Fortbildung hinaus. Laien bleiben dabei meist außen vor, obwohl sie oft nachhaltig von der Entwicklung der Schule betroffen sind oder Kompetenzen und Sichtweisen einbringen können, die bildungspolitisch für die Entwicklung der Schule bedeutsam sind.

12 Exkurs II: Beispiele für unterschiedliche Formen der Arbeitsteilung

Als vor einigen Jahren die internationale Grundschulstudie IGLU durchgeführt wurde, haben wir parallel in unserem Projekt LUST (2004) Schulen einen einfachen Lesetest und dessen Auswertung mit Vergleichsdaten angeboten. Erhebung der Daten und Bewertung der Ergebnisse lagen in der Hand der LehrerInnen. Mehr als 1.000 LehrerInnen haben damals dieses Angebot angenommen – für mich ein schlagender Beweis dafür, dass etwa die Kritik vieler KollegInnen an zentralen Lernstandserhebungen wie VERA nicht als grundsätzliche Evaluationsfeindlichkeit der Schulen missverstanden werden darf.
Die Zuständigkeiten waren bei LUST wie folgt verteilt:

LUST	intern	extern
Erhebung	**Test**	
Analyse	**(Fehleranalyse)**	**Statistik**
Bewertung	**Situationsbezug**	**(Vergleichswerte)**
Entscheidung	**Förderung**	

Die Auswertung von Tests und Fragebögen ist oft sehr *aufwändig*. Darum ist es attraktiv, diese Arbeit nach außen zu verlagern. Zwei Bedingungen haben in LUST die Entlastung durch SpezialistInnen interessant gemacht für die Schulen: die freie *Wahl der Instrumente* und der Verbleib der *Datenhoheit* bei der Schule.
Es lohnt, Hochschulen anzusprechen und um Unterstützung für die technische Durchführung von Test- oder Fragebogen-Erhebungen zu werben. Das ist nicht so schwierig, wie man auf den ersten Blick vermuten könnte. Hochschulen haben einen hohen Bedarf an Erprobungsmöglichkeiten für neue Instrumente und an der Gewinnung/ Erweiterung von Stichproben, um Normwerte zu gewinnen[33]. Bei dem Tausch-Geschäft können also beide Seiten gewinnen.
Entsprechend gut eingeschlagen ist das Angebot einer Studentengruppe aus Berlin, den Erhebungs- und Rechenprozess für Hochschulseminare zu automatisieren:

www.meinprof.de	intern		extern		
	Dozent	Aktuelle Studenten	Experten	zukünftige Studenten	Öffent-lichkeit
Erhebung			**Fragen**		
Analyse			**Statistik**		
Bewertung		**Eintrag**			
Entscheidung	**Optimierung**			**Kurswahl**	**Klatsch**

In diesem Fall wird mein Raster erheblich komplizierter. Was für erhebliche Unruhe gesorgt hat, ist vor allem die asymmetrische Öffentlichkeit:

- die Studierenden bleiben anonym

[33] Vgl. auch oben das am *Deutschen Institut für Internationale Pädagogische Forschung* entwickelte PEB.

- die DozentInnen werden nicht nur im Seminar, sondern weltweit Gegenstand von externer Beurteilung.

Wollte man ein solches Verfahren für die "Blick über den Zaun"-Schulen nutzen, müsste man – z. B. passwortgeschützt – eine schul- oder gar nur klassenbezogene Öffentlichkeit sichern, wenn man die Evaluation – zumindest in der ersten Stufe intern halten will. Dieser Gedanke einer *gestuften Öffentlichkeit* kann helfen, einige Probleme zu reduzieren, de ich eingangs mit den Kontexten von Angst und Macht angedeutet habe.
Das Modell sähe dann so aus:

www.meinprof.de **variiert für BüZ**	intern		extern
	LehrerInnen	SchülerInnen	Experten
Planung	**Fragen**		
Erhebung			**Organisation**
Bewertung		**Eintrag**	
Analyse			**Statistik**
Entscheidung	**Optimierung**		

Es gibt aber noch ein zweites Problem :Trotz einer Spalte für freie Kommentare sind die vorgegebenen Kriterien zu eng und auch inhaltlich nicht auf Ihre "Blick über den Zaun"-Standards zugeschnitten. Aber denkbar ist, in Kooperation mit www.meinprof.de ein solches Raster zu entwerfen, das die beteiligten Schulen dann bei Bedarf immer wieder nutzen könnten.
Dass eine *Auslagerung* der Evaluation als „Dienstleistung“ auch misslingen und der Anspruch auf Daten-Hoheit verloren gehen kann, haben die internationalen Leistungsvergleiche wie TIMSS, PISA und IGLU gezeigt. Die Ministerien hatten sich die Konstellation vermutlich so gedacht:

PISA (ideal)	intern	extern
Erhebung		**Tests**
Analyse		**Statistik**
Bewertung	**Vergleiche**	
Entscheidung	**Folgerungen**	

Dieses Modell versprach Qualitätssteigerung der System-Evaluation, für die den Ministerien die Manpower fehlte, und zugleich kapazitative Entlastung der Bildungsverwaltung durch Auslagerung des Projekts an die OECD.
Faktisch ist ein *Autoritätsmodell externer Kontrolle* herausgekommen:

PISA (real)	intern	extern
Erhebung		**Tests**
Analyse		**Statistik**
Bewertung		**Vergleiche**
Entscheidung	**Folgerungen**	**(Folgerungen)**

Die externen Bewertungen und ihre Diskussion in dem Medien werden nun politisch als zusätzliche Belastung erlebt, da sie Druck erzeugen, der zu vielfältigen Auseinandersetzungen

zwingt und damit Ressourcen bindet. An den inhaltlichen Reaktionen der Ministerien kann man auch erkennen, welche Probleme es aufwirft, wenn man den Betroffenen die Bewertung wegnimmt: es wird gemauert, statt zu lernen. Diese Reaktion ist nicht nur für Ministerien typisch...

Das Autoritätsmodell wirkt anders bei „Stiftung Warentest“ oder „Computer Bild“. Hier haben wir grundsätzlich dieselbe Konstellation, aber der Lerneffekt wird indirekt erzwungen über das Kundenverhalten.

Stiftung Warentest	intern	extern	
		Experten	Nutzer
Erhebung		**Tests**	
Bewertung		**Ranglisten**	
Entscheidung	**Verbesserung**		**Kaufen**

Das wollen die Marktliberalen nun auch im *public sector* erreichen. Im Bildungswesen entspricht dem beispielsweise das Ranking-Konzept mit Freigabe der Schulbezirke. Für diejenigen, die an Privatschulen arbeiten, ist das heute schon Alltag. Aber die Folgen sind heikel in einem Bereich, in dem die Erfolgskriterien diffus bzw. schwer zu messen sind.

13 Exkurs III: Erkenntnistheoretische Fallen

Und damit sind wir beim Kern des Problems: Von welchen erkenntnistheoretischen Bedingungen müssen wir ausgehen, wenn wir den Status von konkurrierenden Evaluations- und Daten-Typen vernünftig einschätzen wollen?

Sechs Missverständnisse des Status von "wissenschaftlicher" Erkenntnis[34]

- **der Objektivitätsanspruch von Erkenntnis**, der ausblendet, dass eine jede Beobachtung, Befragung oder Testung verändert, was sie zu erfassen versucht; Beispiel: die „Verbesserung" von PISA-I auf PISA-II kann auch eine Folge davon sein, dass sich die „Objekte" der Forschung besser auf das eingestellt haben, was man an Leistung von ihnen erwartet, und vor allem, *wie* sie diese Leistung demonstrieren können.
- **die Illusion eines archimedischen Punkts,** d. h. die Unterstellung eines Bezugspunkts außerhalb der Forschung, an dem sich Instrumente validieren und damit ihre besondere Autorität begründen ließen; es ist schon paradox, dass man die Unzulänglichkeit von Noten durch die Einführung von Tests überwinden will, die man dadurch „validiert" hat, dass man ihre Ergebnisse mit den Noten korreliert, die sie doch ersetzen sollen;
- **die Spiegel-Illusion der Wahrnehmung**, basierend auf der Annahme, unsere Vorstellungen von der Welt oder von Ausschnitten aus ihr ließen sich mit technischen Mitteln so perfektionieren, dass sie als bloße Abbilder der externen Wirklichkeit gehandelt werden könnten; aber auch sog. „Daten" sind psychisch gefilterte Einheiten;
- **die Induktions-Falle von Verallgemeinerungen**, also die Unmöglichkeit, aus noch so vielen Beobachtungen von „X folgt aus Y" in der Vergangenheit zu folgern, dass auch in Zukunft „X aus Y" folgt;
- **die Deduktions-Falle der Normauslegung**, d. h. die Hoffnung, man könne allgemeine Kriterien oder Aussagen auf besondere Fälle anwenden, ohne dass persönliche Deutungsfolien wirksam würden; sprachliche Äußerungen sind mehrdeutig, ihre Bedeutung vom jeweiligen Kontext abhängig und der wird vor dem Hintergrund der individuellen Erfahrung unterschiedlich interpretiert;
- **die Konstanz-Annahme sozialer Verhältnisse**, die übersieht, dass keine zwei pädagogische Situationen wirklich gleich sind; mit den Worten von einem der bekanntesten Bildungsforscher aus den USA, G. V. Glass (1972, 11):
- „Wenn physikalische Gesetze in ihrer Reichweite so begrenzt wären wie die von Sozialwissenschaftlern bis heute entdeckten Regeln, würden wir morgens sehr zögerlich aus den Bett kriechen, da wir nicht wüssten, ob wir zur Decke schweben oder auf den Boden krachen werden." (Übers. brü)

Nun kann man sagen, diese Einwände seien puristisch, man müsse pragmatisch mit solchen Einschränkungen umgehen, und niemand sei so naiv, diese Annahmen ernsthaft zu vertreten. Letzteres gestehe ich sofort zu, wenn KollegInnen des Measurement-Paradigmas auf einer methodologischen Ebene mit mir diskutieren. Aber ich bezweifele, dass die Forschungspraxis und –politik zurzeit so aufgeklärt „pragmatisch" handeln. Das technische Paradigma hat genau deshalb ein so hohes Ansehen, weil die genannten Annahmen - zumindest unterschwellig - mitgedacht werden.

[34] Ausführlicher ➔ SAFARI-1974

14 Literatur

Bartnitzky, H., u. a. (1999): Zur Qualität der Leistung – 5 Thesen zu Evaluation und Rechenschaft der Grundschularbeit. Grundschulverband – Arbeitskreis Grundschule e. V.: Frankfurt. Auch in: Schmitt, R. (Hrsg.) (1999, 165-196).

Bartnitzky, H., u. a. (Hrsg.) (2005&2006): Pädagogische Leistungskultur: Materialien für Klasse 1/2 und Klasse 3/4. Beiträge zur Reform der Grundschule, Bd. 119 & 121. Grundschulverband: Frankfurt

Bauersfeld, H. (1972): Einige Bemerkungen zum Frankfurter Projekt" und zum "alef"-Programm. In: Schwartz u. a. (1972, 237-246).

Behnken, I., u. a. (2004): Lernen, Bildung, Partizipation. Die Perspektive der Kinder und Jugendlichen. Befragung zum 8. Kinder und Jugendbericht des Landes NRW (im Auftrag des Ministeriums für Schule, Jugend und Kinder). Siegener Zentrum für Kindheits-, Jugend- und Biografieforschung der Universität (SIZE): Siegen/ ProKids: Herten (weitere Informationen ➔ www.size-siegen.de).

Böttcher, W., u. a. (Hrsg.) (2006): Evaluation im Bildungswesen. Eine Einführung in Grundlagen und Praxisbeispiele (Grundlagentexte Pädagogik). Juventa: Weinheim/ München.

Brügelmann, H. (1976): The teachers' centre. Occ. Publ. no. 3. Centre for Applied Research in Education/ University of East-Anglia: Norwich.

Brügelmann, H. (2000): Wie verbreitet ist offener Unterricht? In: Jaumann-Graumann/ Köhnlein (2000, 33-143).

Brügelmann, H. (2005a): Schule verstehen und gestalten – Perspektiven der Forschung auf Probleme von Erziehung und Unterricht. Libelle: CH-Lengwil (fortlaufend aktualisiert unter: www.agprim.uni-siegen.de/schuleverstehen).

Brügelmann, H. (2005b): Von Fall zu Fall. Plädoyer für einen Miss-Marple-Stil in der Bildungsforschung. Publiziert im: Forum Kritische Pädagogik (4.10.2005) ➔ http://forum-kritische-paedagogik.de/start/download.php?view.15

Cox, S., et al. (2006): Children decide: Power, participation and purpose in the primary classroom. School of Education and Lifelong Learning. University of East-Anglia: Norwich.

Fischer, D. (Hrsg.) (1982): Fallstudien in der Pädagogik. Ekkehard Faude: Konstanz.

Fischer, D. (Hrsg.) (1983): Lernen am Fall. Zur Interpretation und Verwendung von Fallstudien in der Pädagogik. Ekkehard Faude: Konstanz.

Glass, G. V. (1972): The wisdom of scientific inquiry on education. In: Journal of Research in Science Teaching, Vol. 9, No. 1, 1-18.

Groeben, A. v. d., u. a. (Red.) (2005): „Blick über den Zaun“ – Bündnis reformpädagogisch engagierter Schulen: Unsere Standards. 2. Vorlage (überarbeitete und erweiterte Fassung der 1. Vorlage nach der Tagung am 29./30. Januar 2005 in Bielefeld) ➔ www.blickueberdenzaun.de/files/UnsereStandards_1206.doc [Abruf: 19.1.2007].

Jaumann-Graumann, O./ Köhnlein, W. (Hrsg.) (2000): Lehrerprofessionalität -- Lehrerprofessionalisierung. Jahrbuch Grundschulforschung, Bd. 3. Klinkhardt: Bad Heilbrunn.

LUST (2004): Berichte unter ➔ www.agprim.uni-siegen.de/lust

Neuweg, G. H. (2006): Das Schweigen der Könner. Strukturen und Grenzen des Erfahrungswissens. Trauner Verlag: Linz.

Popp, F.-A. (2006): Biophotonen – Neue Horizonte in der Medizin. Von den Grundlagen zur Biophotonik. Karl F. Haug Verlag: Stuttgart (3. vollst. überarb. Aufl.; 1. Aufl. 1983).

Ratzka, N. (2003): Mathematische Fähigkeiten und Fertigkeiten am Ende der Grundschulzeit – Empirische Studien im Anschluss an TIMSS (Phil. Diss. FB 2 der Universität Siegen). Franzbecker: Hildesheim/ Berlin.

Rumpf, H. (1976): Unterricht und Identität. Juventa: Weinheim/ München (3. Aufl. 1987).

Rumpf, H. (1986): Die künstliche Schule und das wirkliche Lernen. Ehrenwirth: München.

Sander (Hrsg.) (2000)] Joint Committee on Standards for Educational Evaluation (2000): Handbuch der Evaluationsstandards. Leske+Budrich: Opladen (2. Aufl.; engl. 1994). [s. a. ➔ www.degeval.de & http://web.ansi.org]

Schmitt, R. (Hrsg.) (1999): An der Schwelle zum dritten Jahrtausend. BundesGrundschulKongress 1999. Grundschulverband – Arbeitskreis Grundschule: Frankfurt.

Schratz, M., u. a. (2000): Qualitätsentwicklung. Vorschläge, Methoden, Instrumente. Beltz Verlag: Weinheim.

Schwartz, E., u. a. (Hrsg.) (1972): Materialien zum Mathematikunterricht in der Grundschule. Beiträge zur Reform der Grundschule Bd. 13. Arbeitskreis Grundschule: Frankfurt.

Seydel, O. (2005): „Hilfe! Der Inspektor kommt.“ Oder: Sind Schulen Kunstwerke? In: Pädagogik, 57. Jg., H. 9, 10-15.

Tillmann, K.-J./ Wischer, B. (Hrsg.) (1999): Schulinterne Evaluation an Reformschulen. Positionen, Konzepte, Praxisbeispiele. IMPULS, Bd. 30. Universität: Bielefeld und die Texte von Otto Seydel unter http://www.blickueberdenzaun.de/10Evaluation.html [Abruf: 3.11.06].

Wehrli, U. (2002): Kunst aufräumen. Kein & Aber Verlag: Zürich.

Wehrli, U. (2004): Noch mehr Kunst aufräumen. Kein & Aber Verlag: Zürich.

Wulf, C. (Hrsg.): (1972). Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München: Piper.